

Vishanth Hari Raj Balasubramanian

US | 📞 (310)-254-6562 | vishanth.hari.raj.ucla@gmail.com | [Website](#) | [LinkedIn](#) | [GitHub](#)

SUMMARY

Staff Machine Learning Engineer & **M.S @UCLA** graduate with **5+ years' of SDE, MLOps & Applied Machine Learning** experience at **F500 & Startup**, leading **AI Guardrails / LLM Security R&D** at Zscaler, proven track record of innovation impacting **\$5M+** in business value, expertise in **LLM inference optimization**, GenAI, AI Agents, RAG, multimodal, NLP, LLMs, MLOps, ETL data pipelines, Agile (SDLC), CI/CD. Proficient in **SQL, Python, C++, AWS, Azure**. **STEM: M.S @UCLA, B.Tech @NITT**

SKILLS

Programming: Python (PyTorch, TensorFlow, MLflow, NLTK), C++, JavaScript, SQL, NoSQL.
Inference: TensorRT, TensorRT-LLM, ONNX Runtime, vLLM, llama.cpp, CUDA, Tensor/Pipeline Parallelism, FSDP, Quantization (FP8)
Tools: Git, Jira, Airflow, Hadoop, Spark, Snowflake, AWS, LangChain, LangGraph, Kubernetes, Docker, Streamlit
Certification: **Lean Six Sigma Green Belt**, Machine Learning using AWS, Apache Spark, DevOps Leadership
Model: LLM, AI Guardrails, LoRA Fine-tuning, XGBoost, RAG, NLP, CNN, RNN, GAN, Transformers, Random Forest, A/B Test.

EXPERIENCE

ZSCALER

Staff Machine Learning Engineer – AI Guard
Senior Machine Learning Engineer – AI Guard

San Jose, USA
Oct 2026 – Present
Dec 2025 – Oct 2026

- **Leading AI Guard ML R&D:** built & **fine-tuned AI guardrail models** (encoder & decoder LLMs, **LoRA**) to detect **jailbreak & prompt-injection** attacks and LLM data leakage for enterprise AI applications
- **Optimized GPU inference** for production guardrail models using **TensorRT, ONNX & vLLM**, achieving **2x RPS** and **~50% lower latency**, saving **~\$2M per year** in serving cost
- **Benchmarked GPUs** (NVIDIA, AMD) for self-hosted models; **built MLOps** taking AI Guard detectors to production

CORNAMI INC

AI/ML Software Engineer – GenAI

San Jose, USA
Feb 2025 – Dec 2025

- Developed & optimized **LLM inference engines**, leveraging flash attention, **KV cache & quantization FP8** to reduce latency & improve TPS by **30%** for inference on **LLM Models** (Llama 3.3 70B, Llama 4, DeepSeek)
- **Implemented distributed computing techniques** with **Tensor Parallelism, Pipeline Parallelism**, scaling inference across multiple GPUs (DGX H100 servers) & proprietary fractal cores, replicating low-level algorithms through CUDA
- **Optimized & validated LLM inference** on H100 GPUs and Fractal cores, increasing **60% TPS** and reducing **10% latency** using continuous batching
- **Benchmarked LLMs** on TensorRT-LLM, Ollama & vLLM (prefill, memory, latency, TPS) to drive optimization strategy

TRAVERSAAL AI

Machine Learning Engineer Intern

Los Angeles, USA
June 2024 – Sep 2024

- Developed **Enterprise LLM Search Engine** Chatbot E2E application integrated with **RAG (Retrieval-Augmented Generation)** multimodal model format, *used LLM models (Llama 3.1, BART, nomic-v3)* for clients, *enhanced information retrieval by 30%*.
- **Trained/Fine-tuned** pre-trained LLM using **LoRA** with domain knowledge & designed a **knowledge graph** with semantic chunking, improved data retrieval efficiency by 20%, reduced search time by **500s**
- **Stack:** LangChain, Hugging Face, Llama, BERT, AWS SageMaker, SQL, APIs, FAISS, LangGraph, OpenAI API, Vertex AI

CITIGROUP

Software Engineer 2 – AI/ML

Chennai, India
July 2021 – Sep 2023

- **Innovation Lead** for HNI AI/ML: spearheaded ML models to *forecast incidents (in Citi Payments) and analyse downtime impact*, *reduced incidents impact by 15%, saved \$700K+*
- **Integrated LLM/NLP models** into *customer chatbots* for automatic incident creation, LoB assignment & sentiment analysis; **improved resolution time by 30%**
- Developed **ML Models** using **XGBoost, Random Forest** to predict *app login & payment failures* from real-time logs & **Citi 3DS** data, prevented **50+** major incidents, boosting revenue by **~\$1.1M**
- **Implemented recommendation system (Random Forest, TF-IDF)** suggesting incident-resolution actions from past incidents, improving resolution time by **10%**
- **Architected** robust **MLOps** pipeline & infrastructure; **deployed APIs** integrating ML endpoints with dashboards, **reducing latency by 5 mins** on SNOW & Splunk data streams
- **Stack:** ML models (TensorFlow, Scikit-learn), NLP, LLM, Kubernetes, ReactJS, REST APIs, SQL, microservices, Linux

EDUCATION

UNIVERSITY OF CALIFORNIA, LOS ANGELES (UCLA)

Los Angeles, USA
Dec 2024

Master of Science, Business Analytics – Data Science/Machine Learning (STEM)

NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPALLI (NIT-T)

Trichy, India
June 2021

Bachelor of Technology, Computer Science/Industrial Engineering (Institute Rank Holder)

RESEARCH/PROJECTS

Multimodal Hotel Search Engine Chatbot : **fine-tuned BART** for review summarization; embed pipeline reduced search time by 20%

GenAI Agents Auto-Interviews: Built **AI Agent** (crewai) using **whisper-large-v3-turbo** for an audio interview chatbot with ratings

Detect fake images generated by AI: Built **multi-modal GAN** combining image/text features (ResNet, BERT), **accuracy: 97.6%**

Detection of Cars/Traffic Signal : Ensemble model (ResNet, ConvNext) with additional dense layers achieving **MSE of 0.02**

Weather Prediction, LSTM/GRU: Built a **bidirectional LSTM model** predicting weather trends across cities, achieving an **MSE of 0.012**