

# Soorya Narayanan V

Los Angeles, CA | +1 (213) 843-8543 | [sv57690@usc.edu](mailto:sv57690@usc.edu) | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

## EDUCATION

### University Of Southern California

*Masters in Applied Data Science. GPA: 3.85/4.0*

- **Coursework:** Data Mining (PhD), Operating Systems, Applied NLP, Parallel and Distributed Computing.
- **Grader** for USC's Data Visualization Course (DSCI 554) – [React, Vue.js, D3.js - Best chart choice and storytelling]

*Los Angeles, CA*

*Expected May 2027*

### Anna University

*Bachelors in Artificial Intelligence and Data Science. GPA: 3.7/4.0*

- **Coursework:** Data Structures & Algorithms, Machine Learning, Deep Learning, Distributed Networks, Data Science
- **Minors:** Business Data Analytics

*Chennai, India*

*May 2025*

Awards & Certifications: <https://inkedsin.wordpress.com/soorya-narayanan-v>

## WORK EXPERIENCE

### Graduate ML Researcher – WiDES lab (Verizon)

*University of Southern California,*

*January 2026 – Current*

- Led a Verizon-sponsored ML initiative to productionize PMNet, a custom 3D autoencoder for RF pathloss prediction across cities, cutting simulation turnaround from months to hours at 98% fidelity dB, multi-node **large scale distributed system** across 2x A40/A100, SLURM allocations (scaling curves) + torchrun, FP16 mixed precision on HPC.
- Ran **large-scale batch inference** across **19,000+ geospatial samples** Engineered an end-to-end ETL pipeline fusing simulation outputs (Wireless Insite + NVIDIA Sionna) with 3D urban geometry into ML-ready pathloss map datasets, profiled and optimized inference throughput and memory footprint across pipeline, reduced test-set **MAE by 57%**, raising Model FLOPs **Utilization from 30% to 37%** over non-deep-learning baseline and delivered recommendations to Verizon.

### AI Research Engineer – (Neurosymbolic, Multi-Agent Negotiation)

*Marshall School of Business, Zage Energy Initiative,*

*May 2026 – Current*

- Architected a multi-turn inference pipeline for a 13-agent swarm, migrating from **vLLM to SGLang to exploit native prefix-tree (Radix)** KV sharing, reduced Time-To-First-Token (TTFT) by **~40%** during complex conversational forking and eliminating per-agent VRAM overhead via shared-history caching, and cost-per-round optimizations.
- Self-hosted serving + evaluation (SGLang, Qwen2.5-7B): tensor parallelism, GPU-adaptive dtype and flash-attention-backend and fail-fast kernel checks on heterogeneous HPC GPU's, ~3,900-call runs/session with atomic checkpointing and crash-resume, shipping statistical ZOPA analysis.

### ML Engineer (Model Evaluation & Computer Vision)

*Mirador Technologies,*

*July 2024 – August 2025*

- Rebuilt **real-time wild face recognition pipeline** processing live video streams, improving accuracy from **42% to 93%** through re-architected embedding, filtering, preprocessing, and async messaging (RabbitMQ) layers, with FAISS vector search to cut query **latency from ~850ms to ~70ms**.
- Benchmarked **14+** open-source face detection and recognition models with systematic **latency-accuracy-cost** tradeoff analysis, fine-tuned selected models to meet client requirements and led technical demos, converted 2 prospects worth ~\$520K revenue.

### Data Science Intern

*Vizza Insurance Broking Services Pvt Ltd,*

*May 2023 – August 2023*

- Engineered production-grade LightGBM lead-scoring model, surfacing 7 KPIs from raw CRM data, cutting lead-to-conversion time from ~3 days to ~7 hours (71% reduction), and improving sales conversion by 65%, with automated retraining triggered deployment on incoming data for the sales operation team.

## PROJECTS

### Nika 2 X 1M LLM – MoE Transformer Built from scratch

*May 2026*

Built a 2x1M-param MoE transformer from scratch (RoPE, attention, SwiGLU, RMSNorm) with torch.export/fx tracing and KV-cached inference. Added DDP + hand-rolled ring all-reduce with failure detection across multi-node workers. Ported to Apple Metal (MPS), profiling to isolate a real communication-bound vs compute-bound bottleneck ([Link](#))

### Immunity – (LLM Routing + Edge Deployment)

*January 2026 – Current*

Built Clinical AI allergy tracker with a routing layer delegating between AWS Nova Micro and MedGemma 7B, 4-bit quantized, stored in DynamoDB, exported to ONNX for fused-graph inference on resource-constrained hardware. ([Link](#))

### RL from Verifiable Rewards (RLVR) – (LoRA + 4bit Quantization)

*December 2025*

Optimized inference cost and accuracy jointly, applying GRPO via HuggingFace TRL to fine-tune Qwen2.5-7B on tau2-bench with a multi-node distributed training setup, improving JSON formatting accuracy to 87% via 4-bit + LoRA ([GitHub](#))

### Intelligent Bistro – Tool-Calling & Recommendation Test Framework

*Mar 2026*

Built a multi-turn, multi-tool agentic system (9 LLM tool functions, MCP server) with a validation framework having 236 automated tests at 94% coverage, ensuring reliability of non-deterministic ranking/recommendation logic - CI/CD. ([Demo Link](#))

## SKILLS

**Languages & Fundamentals:** Python, C++, TypeScript, SQL, Rust, Scala, Git, Linux/Bash

**ML Frameworks:** PyTorch, Keras, OpenCV, Scikit-Learn, JAX, NumPy, Huggingface, CUDA, RAG, MCP, DDS, Triton, torch.compile.

**Systems & Infra:** Flask, Docker, Kubernetes, Spark, Airflow, GitHub Actions, CI/CD Testing, Auto-Retraining, RabbitMQ, MLIR, LLVM

**Machine Learning:** vLLM/SGLang, Pre & Post-Training, Inference, Nsight, Profiling, MoE, SVD, XGBoost, CV, SFT, LoRA, Distillation.