# Vijay Ram Enaganti

+1 (951) 347-5576 | San Francisco, CA, USA (Open to relocation) | OPT EAD / STEM OPT
vijayram.enag2002@gmail.com | linkedin.com/vijay-ram-enaganti | github.com/VjayRam | vijay-ram.vercel.app

## Professional Summary

AI / ML engineer with production experience in LLM orchestration, prompt optimization, embedding based retrieval, and agentic systems using LangGraph and LangChain. Shipped AI features at LivePerson for enterprise customers and architected AI inference infrastructure on AWS as a founding engineer. Strong across Python, FastAPI, React/TypeScript, and cloud (AWS, GCP).

## Skills

**Languages:** Python, TypeScript, JavaScript, SQL, Bash
**AI / ML:** PyTorch, Scikit-learn, HuggingFace, LangChain, LangGraph, ChromaDB, ElasticSearch, OpenCV, RAG, MLFlow, LangSmith
**Backend & Infra:** FastAPI, Git, Celery, Redis, PostgreSQL, SQLite, Firebase, Nginx, Docker, Kubernetes, Pydantic, Kafka
**Cloud & MLOps:** GCP (Vertex AI, Cloud Run), AWS (S3, EC2, Lambda, DynamoDB, Kinesis, ECS Fargate, SageMaker, Redshift)
**Frontend:** React, Next.js, Vue.js, Tailwind CSS, Vite, shadcn

## Work Experience

**LivePerson Inc.** | *AI/ML Engineer (Internship & Co-op)* New York City, USA | Jun 2025 - Dec 2025

- Implemented **Automatic Prompt Optimization** (based on **Textual Gradient Descent research by Google & Microsoft**) as a self-serve feature in LivePerson's AI Studio, enabling all enterprise customers to **migrate between LLM providers** for customer support chatbots and agents without manual prompt retuning.
- Designed the **end-to-end prompt optimization pipeline** - from **Vue.js frontend** through **Cloud Run job orchestration**, event-driven monitoring, and **Firebase checkpointing** for crash recovery, with **real-time progress updates** to UI and database.
- Reduced model migration timelines from **14 days to 3 days (78%)** with **automated benchmarking** across customer-facing agent configurations before production deployment.

**C-ISFCR, PES University** | *Machine Learning Research Intern* Bengaluru, India | Jan 2024 - Jul 2024

- Engineered **high-throughput model training pipelines** with **PyTorch** to process and analyze over **100 GB of video data** for real-time prediction model training across **transformer architectures** like ViT, Swin, C3D, and I3D.
- Achieved **94.06% accuracy** and **0.938 F1-score** with the best-performing architecture (outperforming baseline by 4% across 8 combinations), with **60% faster training** throughput via mixed-precision (FP16) pipelines and **2838 ms inference latency** at 30 FPS on GPU, meeting real-time processing requirements.

**The Param Science Experience Center** | *Founding Software Engineer* Bengaluru, India | Jul 2022 - Dec 2023

- Architected the **digital infrastructure** for a science museum, building the backend platform that powered **18 interactive exhibits** serving **1000+ daily visitors** with **visitor authentication**, session persistence, interaction logging, and real-time analytics.
- Built an **offline-first data platform** using **SQLite, DynamoDB, and Kinesis Data Streams** with an outbox sync pattern, ensuring **zero data loss** during network outages and **sub-500ms session restore** via **ElastiCache Redis** as a hot cache layer.
- Designed **auto-scaling infrastructure** on **ECS Fargate, DynamoDB** on-demand, and **SageMaker** endpoints that handled **5x weekend traffic surges** while scaling down during off-hours, **reducing cloud compute costs by 62%** compared to fixed-capacity provisioning.

## Projects

**Project Hawkeye - Autonomous AI QA Testing Platform** Demo | GitHub
*LangGraph | FastAPI | Next.js | PostgreSQL | Redis | Docker | Playwright MCP*

- Built a **full-stack SaaS platform** deploying **autonomous LLM-driven browser agents** for goal-driven QA testing. Agents reason over live screenshots and accessibility trees each step instead of replaying recorded scripts, with a **cost of less than $2/run.**
- Designed the **Python** backend serving 85 REST API endpoints with **parallel job execution, real-time test streaming**, and reverse proxy routing built on **FastAPI, Celery, Redis Pub/Sub, NoVNC Stream and Nginx.**
- Implemented **GitHub CI/CD integration, Docker sandboxed execution**, **authentication (JWT/OAuth)**, **encrypted secrets vault**, Stripe billing with usage metering with container pooling, and multi-tenant organization management.

**Project Sensei - Privacy-First Local AI Agent** Demo | GitHub
*LangGraph | FastAPI | Ollama | React | SQLite | ChromaDB*

- Designed a local **privacy-first AI agent** platform using **LangGraph** and Ollama with **27 tools** for file I/O, ChromaDB semantic search, AST code parsing, git inspection, math, conversation recall with **zero cloud dependency**.
- Built a **FastAPI backend** (30 endpoints) with **real-time SSE streaming, background execution** for **concurrent sessions**, and **response performance tracking (TTFT, throughput, per-tool latency)** traced via **LangSmith.**
- Consolidated **5 data categories** (sessions, checkpoints, conversation archive, metrics, file index) into a single **SQLite persistence** layer with automatic **context summarization** at 75% token window usage to **prevent overflow**.
- Created an **offline evaluation pipeline** to pull LangSmith traces and tool results to monitor **retrieval quality, tool output quality, and agent trajectory**.

## Education

**University of California, Riverside** | *MS, Computer Science* Riverside, CA | Sep 2024 - Dec 2025
***Coursework***: Data Mining, Foundations of ML, GPU Architecture & Parallel Programming ***GPA***: 3.70 / 4.00

**PES University** | *B.Tech, Computer Science & Engineering* Bengaluru, India | Aug 2020 - Jun 2024
***Coursework***: Algorithms, Database Management, Operating Systems, Cloud Computing ***GPA***: 3.32 / 4.00

## Open Source Contributions

- **Sktime** | *Added comprehensive, standardized docstrings for 4 core classes in sktime* [#7540, #7563]