

Vraj Patel

San Jose, CA | (714) 726-4605 | vraipatel00222@gmail.com
linkedin.com/in/vraj-patel-csuf | github.com/vraj00222 | vraipatel.xyz

SUMMARY

AI Engineer (M.S. CS, May 2026) with **2 years** of software engineering experience spanning production full-stack systems (**50K+ users**), **IEEE-published** ML research, and LLM products deployed end-to-end: multi-agent orchestration, RAG pipelines, MCP tool integrations, local model inference. **Two-time solo hackathon winner** (Citrus Hack XI · HackDartmouth X). Open-source contributor to **Caveman**, an LLM token compressor used by **10M+ professionals**, and **CareerOps**, an AI job-search agent (63K+ stars).

EDUCATION

M.S. Computer Science, California State University, Fullerton, GPA: **3.75/4.00** Aug 2024 to May 2026

B.Tech. Information & Communication Technology, Marwadi University, GPA: 3.70/4.00 Jul 2019 to Apr 2023

TECHNICAL SKILLS

Languages & Frameworks: Python, TypeScript, JavaScript, SQL, C++, Next.js, React, Node.js, FastAPI, Django REST Framework, Express

LLM & Agents: Anthropic Claude API & Agent SDK, Model Context Protocol (MCP), OpenAI API, Gemini API, Vercel AI SDK, LangGraph, LangChain, Ollama, agentic workflows, function calling / tool use, prompt engineering, NLP, streaming (SSE, WebSockets)

ML & MLOps: Machine learning, deep learning (TensorFlow, Keras), RAG (Retrieval-Augmented Generation), fine-tuning (LoRA, QLoRA), embeddings, semantic search, vector databases (pgvector, Pinecone, FAISS), quantization (GGUF), on-device inference (PyTorch, ExecuTorch), LLM evals (LangSmith, Langfuse), Hugging Face

Cloud & Infra: AWS (EC2, S3, Lambda), Docker, Kubernetes, Linux, microservices, GitHub Actions, CI/CD, PostgreSQL, MongoDB, Redis

Practices: System design, object-oriented programming (OOP), distributed systems, unit & integration testing, TDD, code review, observability

EXPERIENCE

Teaching Associate · California State University, Fullerton Aug 2025 to May 2026

- Built automated Python grading pipelines on Gradescope, cutting grading turnaround **~40%** across 200+ weekly submissions; taught Data Structures and Algorithm Engineering to **200+ students**.

Deep Learning Research Assistant · Yuan-Ze University, Taiwan Jul 2023 to Feb 2024

- Built unsupervised LSTM-Autoencoder models in TensorFlow/Keras flagging ECG anomalies via reconstruction error on the MIT-BIH dataset with **zero labeled training data** (selected as **1 of 3 students** university-wide).
- Co-authored **peer-reviewed IEEE paper** (ICCE 2024, Las Vegas) after benchmarking LSTM-AE against Transformer and Vision Transformer (ViT) variants; **cited** in follow-on cardiac deep learning research.

Software Developer · Lucent Innovation, Ahmedabad, India Feb 2023 to Jul 2023

- Delivered full-stack features across **3 production e-commerce platforms** serving **50K+ users**, shipping TypeScript/React modules on Django REST Framework services backed by PostgreSQL and MySQL.
- Designed **REST APIs** with JWT authentication, role-based access control (RBAC), and pagination; integrated payment and logistics providers with **webhook handling, retry logic, and structured logging** for faster incident triage.
- Reduced critical query **p95 latency 40%** by tuning MySQL query plans and indexes; shipped **GitHub Actions CI/CD** gating every merge with linting, tests, and Docker image builds.

Software Developer Intern · RKIT Software, Rajkot, India May 2022 to Jul 2022

- Built ERP modules with C#, ASP.NET Core MVC, and SQL Server; cut page load times **30%** via optimized stored procedures and idempotent retry/timeout handling for third-party APIs.

PROJECTS

IntelMaxxing: Multi-Agent Career Intelligence Platform · intelmaxxing.tech

- Built and deployed a Next.js + TypeScript platform that **fans out 5 parallel AI agents** with dual-provider LLM routing and automatic **cloud-to-local failover**, generating ranked company dossiers in **under 30 seconds** from Hacker News, GitHub, Reddit, and YC signals.

PocketLLM: Zero-Install Local LLM Runtime on a USB Drive · pocketllm-site.vercel.app · github.com/vraj00222/pocketllm

- Built a USB-portable Ollama inference stack running **quantized GGUF models** (Llama 3.1, Mistral, Phi-3, Gemma) with **zero host install**: memory-mapped weight loading, redirected model storage, token streaming to a web chat UI, zero-footprint eject.
- Profiled cold-start vs decode on an M4 MacBook Pro: model load is **storage-bound** (7× slower from USB) while decode is **compute-bound**, with the same **54 tok/s** (Gemma) and **21 tok/s** (Llama 3.1 8B) from USB or internal SSD.

Agent-Farm: Parallel AI Coding-Agent Orchestrator · agent-farm-website.vercel.app · github.com/vraj00222/agent-farm

- Solo-built** an open-source npm CLI + macOS app on the Claude Code SDK (**same space as YC-backed Conductor and Emdash**): runs N parallel coding agents in **isolated git worktrees**: structured JSON event streaming, conflict-aware diff review and cherry-pick.

Claudeware: AI Text/Voice-to-3D-Print Pipeline · claudewaresite.vercel.app

- Built an **agentic text/voice-to-3D-print pipeline**: Claude **tool-calling** routes designs between OpenSCAD (parametric) and Blender (organic) engines, auto-fixes printability (tolerances, supports, fit), and exports print-ready models via Bambu Studio.

PUBLICATION

I. Farady, **V. Patel**, C.-C. Kuo, C.-Y. Lin. "ECG Anomaly Detection with LSTM-Autoencoder for Heartbeat Analysis," IEEE ICCE 2024. DOI: [10.1109/ICCE59016.2024.10444327](https://doi.org/10.1109/ICCE59016.2024.10444327)