

JAY HEMNANI

Gujarat, India (open to relocate) • jayhemnani992000@gmail.com • +91-9925164910

LinkedIn • GitHub • jayhemnani.me

Summary

Engineer who scopes messy customer problems, ships production integrations, and hands them over without leaving a fire behind. Built and deployed **6-plus client engagements** end to end as the sole engineer: discovery with incomplete specs, architecture, implementation, and rollout. Led with **agentic systems**: 5-agent **LangGraph** research pipeline built in 10 days, **MCP** contributor to the Anthropic Python SDK, WebMCP integration on a production site. Daily driver of **Cursor + Claude Code + Codex** for 1+ year. **3 merged OSS PRs** (vLLM, Anthropic MCP SDK, Google A2UI) show ability to debug and fix integrations inside large codebases that are not mine.

Technical Skills

Languages	Python (primary, daily), TypeScript, Go, SQL, Bash
AI / Agents	LangGraph, LangChain, MCP (Anthropic SDK contributor, 2 servers shipped), RAG pipelines, Ollama local inference, OpenAI / Anthropic / Groq APIs with fallback and cost routing, vLLM (LoRA contributor)
Backend & APIs	FastAPI, async REST, Pydantic contracts, multi-provider API integration, webhook and streaming ingestion
Deployment & Infra	Docker / Docker Compose (7-service and 14-service stacks shipped), AWS (EC2, S3, Lambda, RDS), GCP, on-prem model serving via Ollama, GitHub Actions CI/CD
Data Systems	Apache Kafka, Apache Airflow (18-DAG pipeline), PostgreSQL / TimescaleDB, pgvector, Redis, Elasticsearch

Selected Builds, Deployments & OSS

- CAG Deep Research**: 5-agent LangGraph system, 4 model providers, built in 10 days · GitHub
- Orchestrated **5 LangGraph agents** (search, analyzer, fact-checker, synthesizer, editor) producing 20-page reports from raw queries. Integrated **4 model-provider APIs** (OpenAI, Anthropic, Groq, Ollama) behind one routing layer with local-first fallback and cost accounting; cut inference cost about **70 percent**. Shipped solo as **FastAPI** + Docker in 10 days with **Pydantic** guardrails on every provider boundary.
- Agentic OS**: self-hostable agent runtime, zero cloud dependency · GitHub
- FastAPI** + **PostgreSQL** + **pgvector** + local **Ollama**, with goal / plan / execute tracking and a shell, filesystem, git, browser, and OCR toolchain. Packaged as **Docker Compose** so it stands up identically on a laptop or a customer's on-prem server.
- WebMCP / OSS Contributions**: MCP in production + fixes in codebases I did not write
- Registered **8 structured MCP tools** on jayhemnani.me via W3C WebMCP, one of the earliest production implementations (**89 percent** token-efficiency gain over screenshot-based agents). Merged **modelcontextprotocol** / **python-sdk** #1826 (typing fix on `call_tool`), **vllm-project** / **vllm** #31513 (LoRA attach-path fix in 200k-plus LOC codebase), and **google** / **A2UI** #407 (secret-setup scaffolding).

Experience

- Independent Software & AI Engineer (Forward-Deployed)**, *Small Businesses and Early-Stage Startups, Remote* 2022 to Present
- Sole engineer across **6-plus customer engagements**: discovery with incomplete requirements, architecture, implementation, Docker Compose deployments, and handover.
 - Integrated third-party and model-provider APIs, stood up observability and CI, documented data contracts. Also diagnosed and fixed integration failures in vLLM, Anthropic MCP SDK, and Google A2UI under maintainer review.
- Data Analyst, Multi-property Analytics**, *Elite Hotel Group, India* Summer 2025
- Integrated **4 source systems** into one reporting pipeline, eliminating manual reconciliation of about **1,000 rows/week**. Designed metric SLAs with finance and ops stakeholders, shipping **12-plus** reports and a metrics glossary. **Airflow** alerting caught **3** data-quality issues before leadership dashboards saw them.
- AI / ML Intern**, *Amnex Infotech, Gujarat, India* Jan to May 2022
- Shipped production credit-fraud detection (**XGBoost**, **Random Forest**, **SMOTE**), **94 percent precision** on **180k-plus transactions**, with PSI/KS drift monitoring and handover data contract.

Education

Pandit Deendayal Energy University (PDEU)
B.Tech in Computer Engineering · GPA: 8.7 / 10

2018 to 2022 · Gandhinagar, India