Aditya Vaish

+1-203-988-8816 | adityavaish846@gmail.com | U.S. Permanent Resident | linkedin.com/in/aditya-vaish | github.com/vaishcodescape

Summary

AI Engineer focused on production LLM systems, RAG, and agentic workflows with end-to-end delivery. Built and deployed GPT-4, Llama 3/3.1, Claude 3, and Mixtral pipelines using LoRA/PEFT with emphasis on evaluation, reliability, and latency. Shipped tool-augmented agents (LangChain, LlamaIndex, Codex, Claude Code) and LLM services on Docker, Kubernetes, and AWS.

Technical Skills

Languages: Python, C/C++, JavaScript, TypeScript

LLM & AI: GPT-4, Llama 3/3.1, Mixtral, Claude 3 (Anthropic API), Hugging Face Transformers, PyTorch, RAG, Embeddings, Fine-tuning (LoRA/PEFT), Instruction Tuning, RLHF, Prompt Engineering, Evaluation, NLP, Inference Optimization

Agentic AI: Claude Code, OpenAI Codex, LangChain Agents, LlamaIndex, Tool Use, Function Calling

MLOps/LLMOps: FastAPI, Docker, Kubernetes, AWS (EC2, S3), CI/CD (GitHub Actions), MLflow, Weights & Biases (W&B), Logging

Data & Vector DB: Pinecone, FAISS, PostgreSQL, MongoDB, Vector Search

Experience

Datacurve.ai (Y Combinator Backed) — Remote

Core Code Reviewer & OSS Developer — *Jan 2026 – Present*

- Reviewed 100+ OSS repositories using LLM reasoning and benchmark scoring for correctness, performance, and maintainability cut review cycle time by 40%.
- Built Docker sandbox environments with structured logging to validate 50+ code submissions via automated tests.

Superr.ai — Remote

Product Engineer Intern — *Aug 2025 – Oct 2025*

- Integrated GPT-4 inference into Next.js + Go APIs reduced latency 25% via prompt caching and backend tuning.
- Built Python ML pipelines on AWS EC2 with MLflow tracking containerized with Docker + GitHub Actions, cutting deploy time 30%.

Projects

OpenX MCP | *Autonomous Agentic AI Developer* | GitHub

- Built multi-agent system using **Claude 3 (Anthropic API)** for GitHub workflows (PR generation, code review, CI/CD recovery) reduced manual toil 60%.
- Implemented RAG with LangChain + FAISS over Claude backends achieved sub-200ms retrieval and shipped a Rust TUI.

pwn-guard | *AI Scam Detection & Threat Intelligence API* | GitHub

- Shipped NLP + LLM inference API with Llama 3 + spaCy NER achieved F1=0.91 and ROC-AUC=0.94.
- Fine-tuned LoRA adapter with W&B FastAPI + Docker served 500+ requests/day at p95 <300ms.

Education

Dhirubhai Ambani University — Gandhinagar, India

B.Tech — Information and Communication Technology — *Expected 2028*

- Coursework: Data Structures, Object-Oriented Programming, Algorithms, Operating Systems, Computer Networks

Achievements & Leadership

IIM Ahmedabad CTDP 2025 | AI Accessibility

- Fine-tuned open-source computer vision models on handwriting datasets for dyslexia detection using transfer learning recognized among top AI accessibility projects.

IIM Ahmedabad IMRC 2025 | ML Forecasting

- Built XGBoost model for agricultural price forecasting evaluated with RMSE and F1 on structured government datasets.

Google Developer Groups (GDG) | Technical Lead

- Led workshops on agentic AI, LLM engineering, RAG, and LangChain for 50+ students shipped internal platforms serving 300+ students.