SUMIT JAGTAP

AI Engineer · Generative AI & LLM Systems · AI-Native Builder

Pune, Maharashtra, India • sumitpatil.tech@gmail.com • github.com/Sumit-tech01 • linkedin.com/in/sumitjagtappatil

CSE (Data Science) student and independent AI engineer building production-grade Generative AI, RAG, and agentic systems. Works AI-natively — Claude Code, Cursor, and GitHub Copilot daily — to move fast without sacrificing quality. Strengths: LLM-powered retrieval & RAG, NLP pipelines, model reliability/observability, and end-to-end deployment — built solo, ready for remote AI-company contract work.

Experience

AI & Full-Stack Developer — *Independent* *2025 – Present*

- Designed and deployed multiple end-to-end AI systems across 5+ real-world domains — search, fintech, healthcare-adjacent NLP, and enterprise reliability tooling
- Built and productionized Generative AI / RAG and NLP pipelines, integrating LLM-adjacent retrieval (embeddings, hybrid search) with backend services and user-facing applications
- Focused on production readiness: fault tolerance, model reliability, observability, and scalable REST APIs — the same concerns central to forward-deployment and AI-agent engineering work
- Worked independently end-to-end: architecture, model integration, backend, frontend, and deployment — suited to fast-moving remote contract engagements

Projects

FlipSearch — AI Semantic Search & RAG System | SBERT · FAISS · BM25 · RAG · Docker

- Built hybrid (dense + sparse) retrieval over 205K+ documents — MRR 0.89, NDCG 0.81 using tuned hybrid scoring (α=0.6), the core retrieval pattern behind production RAG and LLM-agent knowledge tools
- Shipped a multilingual (Hindi + English) + voice search pipeline, deployed on Hugging Face; used Cursor to iterate quickly on the retrieval scoring logic

ML Reliability Platform for Text Intelligence | Circuit Breakers · Drift Detection · Observability

- Built fault-tolerant ML orchestration with fallback routing between a primary LLM and deterministic local models — directly relevant to LLM-gateway and agent-reliability engineering
- Implemented drift detection, request tracing, logging, and abuse detection — the observability stack production AI/agent systems require

AI-Powered Bank Statement Analysis Platform | OCR · JWT · React · Full-Stack FinTech

- Delivered a full-stack FinTech app combining OCR + AI for transaction extraction, behind a secure JWT API layer; used Claude Code to scaffold the API layer and accelerate the React dashboard build
- Built an analytics dashboard for financial insights with rate limiting and Redis caching — end-to-end product, not just a model

AI Sentiment Analysis Web App | DistilBERT · Flask · Docker

- Built an NLP system with dual inference (transformer + embedding-based) and confidence scoring, exposed via REST APIs
- Containerized and deployed with Flask + Docker; used GitHub Copilot for boilerplate/test scaffolding

Voice Sentiment Analysis System | Whisper · NLP · Real-Time

- Built a real-time speech pipeline using Whisper + NLP for low-latency emotion classification

Technical Skills

Languages: Python, JavaScript, SQL, C++

Generative AI / NLP: Generative AI, RAG, NLP, Transformers, SBERT, FAISS, DistilBERT, Whisper

AI-Assisted Development: Claude Code, Cursor, GitHub Copilot — used daily for scaffolding, iteration, and code review

Frameworks: Flask, FastAPI, React

Databases: MongoDB, PostgreSQL, Redis

Tools & Deployment: Docker, Git, Hugging Face, Firebase, Supabase

Systems & Practices: MLOps, System Design, REST APIs, Observability, Authentication

Education

B.Tech, CSE — Data Science — *Ajeenkya D Y Patil University × NxtWave* *2025 – 2029*

Certifications & Achievements

- JPMorgan Chase & Co. Software Engineering Virtual Simulation — Forage
- Generative AI Mastery — OpenAI Academy × NxtWave
- NIAT Masterclass — LLM Embeddings & Applications