

Suraj Shah

Mumbai, Maharashtra, India | surajshah.bombay@gmail.com | +91 98929 30800 | github.com/suraj2906 | linkedin.com/in/suraj-p-shah

SUMMARY

AI Engineer specializing in agentic and LLM systems — multi-agent orchestration (CrewAI), retrieval-augmented generation, and LLM-driven document pipelines — with the full-stack range to ship them end to end (FastAPI, PostgreSQL, Next.js). Currently building a healthcare-credentialing platform; previously published NLP research (IEEE) and built an AI voice-clone classifier at 92.5% accuracy.

EXPERIENCE

Junior AI Engineer — Vitco Software Solutions

May 2025 – Present

Backend engineer on a 3-person team building a healthcare-credentialing platform.

- Engineered a declarative, field-driven routing layer for a CrewAI multi-agent document pipeline, replacing LLM-based classification with deterministic tool composition — cutting agent runtime ~25% and eliminating redundant LLM reasoning calls (and their token cost) on documents resolvable from structured fields.
- Hardened the FastAPI/PostgreSQL backend against concurrency failures, diagnosing and fixing connection-pool exhaustion and transaction-rollback errors that were breaking batch uploads under load.
- Added a pre-scan step that extracts expiry and reference data from structured fields before invoking agents, avoiding unnecessary LLM calls on already-structured inputs.
- Designed a re-run endpoint to re-execute agents on existing documents without re-triggering full ingestion, cutting reprocessing and iteration time.

SKILLS

AI / LLM: CrewAI (multi-agent orchestration), LangChain, RAG, TensorFlow, OpenCV, scikit-learn

Backend: Python, FastAPI, PostgreSQL, SQLAlchemy, REST API design

Cloud: AWS (Bedrock, Textract, S3)

Frontend: Next.js, TypeScript, React, Tailwind, React Native

Also: JavaScript, C++, SQL · Git/GitHub · Pandas, NumPy

EDUCATION

Vellore Institute of Technology (VIT), Chennai — B.Tech, CSE (AI & Robotics) · CGPA: 8.7

Relevant coursework: Machine Learning, NLP, Data Structures

PROJECTS

AI-Generated Voice Classifier — TensorFlow, Librosa, Python

Trained a classifier to detect AI-generated voices and voice clones at 92.5% accuracy on a self-curated dataset of 5,000 audio samples (3,000 AI-generated).

AutoResearcher — LangChain, GPT-4o, FastAPI, RAG

Agentic system that autonomously researches a topic, summarizes findings, and generates structured papers using RAG over retrieved sources with streaming FastAPI output.

PUBLICATION

Automated Tweet Sentiment Analysis Using Machine Learning — IEEE Xplore, 2023

Co-authored a published conference paper introducing an XLNet-based transformer pipeline with a sentiment lexicon for multi-class tweet sentiment classification and trend forecasting, validated via cross-validation.

CERTIFICATIONS

AWS Certified Cloud Practitioner · Generative AI with Large Language Models (DeepLearning.AI)

LEADERSHIP & INTERESTS

- Hackathons:** Finalist — Rakuten Code the Future and TetraFlip; competed across blockchain, healthcare, and edtech (4 hackathons).
- Android Club, VIT Chennai:** built React Native / Node apps for the technical team and ran logistics for a national in-campus hackathon.
- Community:** social media & outreach for Carter Cleanup, a citizen-led beach-cleanup movement in Mumbai.

Languages: English (full professional), Hindi (native), Gujarati (native). **Interests:** piano, origami, badminton, chess.