Razin Chinwala

+91 6353050964 | razin59185@gmail.com | linkedin.com/razin | github.com/rxzin

Summary

Full-Stack AI Engineer building production TypeScript/Python backends and LLM agent systems: multi-tenant SaaS, AI sales automation, RAG, durable workflows with PostgreSQL, Redis, and Docker.

Skills

Languages: TypeScript, JavaScript, Python, SQL, HTML, CSS

Backend and Frontend: Node.js, NestJS, Fastify, FastAPI, Express.js, Prisma, TypeORM, Temporal.io, BullMQ, React, Next.js

AI and LLM: LangChain, LangGraph, RAG, AI Agents, OpenAI, Gemini, Mistral, AWS Bedrock, pgvector, Pinecone, Hugging Face, PyTorch

Databases and DevOps: PostgreSQL, Redis, Docker, AWS (EC2, RDS, S3, CloudFront), Railway, Vercel, Git

Experience

Associate Fullstack AI Consultant April – Present 2026

Code Colonies | *Nextjs, NestJS, Fastify, Prisma, PostgreSQL, Temporal, Redis, Docker* *Ahmedabad*

- Led development of WhatsAuto, a multi-tenant WhatsApp automation SaaS: chose the tech stack, designed the database schema (98 of 113 migrations), and deployed it for internal use on a Contabo VPS.
- Built WhatsAuto's tenant-isolated RBAC using PostgreSQL row-level security and a Meta WhatsApp Cloud API integration with webhook ingestion, an idempotent outbound queue, and template sync.
- Built an AI auto-reply agent that answers from each tenant's knowledge base with human takeover, plus a visual flow builder for message automation, campaigns scheduling
- Migrated an internal sales CRM from a legacy Express.js backend (68+ routes) to Fastify, Prisma, and PostgreSQL with JWT RBAC in under a week with zero regressions.
- Built an autonomous AI sales agent that reads and replies to email, qualifies leads, books meetings, and drafts proposals, with human approval gates.
- Designed a multi-LLM gateway with distributed rate limiting, priority lanes, and token metering, used by Temporal lead-discovery and campaign workflows.
- Built a distributed multi-agent scheduling engine for a community platform in 1 day for a stakeholder demo, with probabilistic, activity-aware scheduling.
- Implemented two-phase action reservations with row locks, and crash-recovery reconciliation to prevent duplicate AI actions and budget leaks.
- Built search-by-image for an e-commerce platform using AWS Bedrock embeddings, a pgvector HNSW index, and an SSRF-hardened image proxy.
- Built an event-driven "frequently bought together" recommendation engine (Jaccard similarity to reduce popularity bias), plus order, shipping, tracking notifications through WhatsAuto.

Internship Feb – May 2025

BISAG-N | *QGIS, PyTorch, NumPy, Rasterio, Matplotlib, Scikit-Learn* *Gandhinagar*

- Designed a rooftop detection pipeline on satellite imagery comparing YOLOv11 (S/M), YOLOv8 (S/L), and U-Net (ResNet-152), with U-Net reaching the best F1 score of **0.89** (YOLO variants **0.73–0.76**).
- Improved feature extraction with classical and deep-learning filters, and regularized predictions by generating minimum bounding rectangles from model and HED edge outputs and refining them with geospatial queries.

Internship June – July 2024

IBM SkillsBuild | *IBM WatsonX AI, IBM Cloud, Data Preprocessing* *Remote*

• Developed a disease diagnosis model with IBM AutoAI on 257K+ rows, 371 features, and 757 diseases, reaching **86%** accuracy.

• Integrated IBM watsonx Assistant for conversational specialist recommendations, and placed in the **winning top 6 of 650+ teams**] across India.

PROJECTS

Agentic AI Chatbot with RAG 2026

Web app | Nextjs, Fastapi, Redis, Docker, Pinecone, Huggingface, Cohere, Langchain, Langgraph

• Architected a full-stack agentic chatbot using a 4-path state machine and a hybrid router (regex fast path + structured classification) across chat, RAG, weather, and bug triage, cutting redundant LLM calls. Streams tokens over SSE with disconnect safety and keeps dual Redis/PostgreSQL memory with token-budget compaction and rolling sliding-window summarization.

• Built auth from scratch with OAuth2/JWT, automatic refresh token rotation, SHA-256 hashed token storage, multi-session device pruning, and RBAC on admin actions. Added async structured JSON logging (QueueHandler/QueueListener) that tracks request IDs and latency.

• Engineered a conversation-scoped RAG pipeline with Heirarchical chunking, local ONNX Runtime BGE embeddings on CPU, Pinecone multi-tenant namespaces, and Cohere Rerank v3 for ranked, verified citations.

• Built a human-in-the-loop bug-reporting workflow that turns chats into structured Jira tickets, using row-level locks and atomic Redis rate limiting to prevent race conditions and duplicate tickets on admin approval. Added browser voice input/output (Web Speech API) and deployed multi-stage Docker images built with uv on AWS (EC2, RDS PostgreSQL, S3/CloudFront) for a live pilot with 8 users.

Sentiment Analysis of Movie Reviews 2024

Natural Language Processing Project | Python, PyTorch, Transformers

• Fine-tuned BERT and RoBERTa on IMDB, reaching **93.86 %** and **94.84 %** accuracy.

EDUCATION

Adani Institute of Infrastructure Engineering 2021 – 2025
B.E. in Computer Science and Engineering, AI and ML *CGPA: **8.94** /10*

Infocity Junior Science College 2019 – 2021
XII *Grade: **86%***

CERTIFICATIONS

• Introduction to Node.js, v3

• TypeScript: From First Steps to Professional

• Professional CSS: Build a Website from Scratch