Jay Gadhiya

Full-Stack AI Engineer | Backend, Cloud & LLM Systems

Ahmedabad, Gujarat, India

7041570015 | gadhiyajay7676@gmail.com | LinkedIn: linkedin.com/in/jaygadhiya-ai

Professional Summary

Full-Stack AI Engineer with 1.8+ years of experience designing and shipping production-grade backend systems, real-time AI agents, and LLM-powered platforms across healthcare, EdTech, and e-commerce. Hands-on across Python/FastAPI, multi-provider LLM pipelines (OpenAI, Gemini, Bedrock), WebSocket agents, and AWS serverless infrastructure, owning architecture, implementation, and deployment end-to-end.

Technical Skills

Languages & Frameworks: Python, FastAPI, SQLAlchemy, Pydantic, TypeScript, JavaScript, React, Next.js 14/15

AI & LLM: OpenAI (Chat, Responses, Realtime, File Search, Agents SDK), Google Gemini, AWS Bedrock (Converse API), LangGraph, RAG / Vector Stores, Prompt Engineering

Cloud & DevOps: AWS (Lambda, Cognito, S3, DynamoDB, Lightsail, Bedrock), Firebase (Admin SDK, FCM, Hosting), Docker, Docker Compose

Backend & APIs: REST, WebSockets, JWT + Refresh Token Rotation, OAuth2/OIDC, OTP Flows, Rate Limiting (SlowAPI), APScheduler, Redis

Databases & Integrations: PostgreSQL, DynamoDB, SQLite, Supabase, Stripe, Authorize.Net, TaxJar

Professional Experience

Vibgyor Technologies October 2024 – Present

Jr. Software Developer, Full-Stack & AI *Ahmedabad, Gujarat, India*

- Migrated legacy Java Spring Boot services to Python/FastAPI by redesigning database schemas and rewriting core APIs, cutting average API response time by **50%**
- Architected an AI-powered exam grading engine that parses Question Papers and Mark Schemes from PDFs and grades student answer sheets across **3+ exam boards**, with per-criterion rubric evaluation, follow-through logic, and mark dependencies, replacing manual grading workflows that previously took teachers hours per paper
- Engineered a multi-provider LLM dispatcher unifying **3 major providers** (OpenAI, Google Gemini, AWS Bedrock Converse) behind a single interface, enabling provider switching via a single environment variable with **zero code changes**, reducing vendor lock-in and lowering per-request cost by routing to cheaper models for non-critical tasks
- Built WebSocket-first real-time AI agents in healthcare: a diagnostic agent server (OpenAI Agents SDK) that extracts symptoms from natural language and maps them to clinical Finding IDs via a custom Vector Store (RAG), iterated across **4 progressive agent versions**, and a voice-driven clinical intake agent (OpenAI Realtime API) deployed for paediatric nutrition assessments with structured outputs persisted to DynamoDB via AWS Lambda, both with AWS Cognito JWT auth, per-session context, and token cost tracking
- Integrated Authorize.Net and TaxJar across an e-commerce SaaS backend covering **10+ modules** (auth, subscriptions, coupons, reporting, webhooks, PDF generation), with JWT + refresh token rotation, OTP flows, and rate limiting, deployed via Docker Compose to AWS Lightsail

TRooTech January 2024 – August 2024

Python Developer Intern *Ahmedabad, Gujarat, India*

- Engineered end-to-end backend features using Python and FastAPI with strict Pydantic validation and clean service-layer architecture
- Architected secure authentication infrastructure using OAuth2/OIDC and Okta SAML with robust session management and refresh token rotation
- Optimized complex SQL queries and collaborated with senior engineers to define scalable API contracts and database performance strategies

Education

New LJ Institute of Engineering and Technology January 2020 – May 2024

Bachelor of Engineering in Information Technology *CGPA: 7.87*

Certifications

Prompt Design in Vertex AI Skill Badge (Google Cloud) · **JavaScript Algorithms and Data Structures** (freeCodeCamp) · **Responsive Web Design** (freeCodeCamp)