

RAJ DESAI

+1 602-582-5573 | desai.raj1807@gmail.com | Fremont, CA | rajdesai.io | linkedin.com/in/rajdesai18/ | github.com/RajDesai-18

SUMMARY

Software Engineer with an MS in Computer Science from UT Dallas (2026) who ships distributed systems and AI-powered tools, from an 8-service contract-review platform with saga orchestration to async code-review pipelines under hard latency limits. Strong in Python, TypeScript, API design, and event-driven architecture, with end-to-end ownership from design through deployment.

EDUCATION

The University of Texas at Dallas | MS – Computer Science Dallas, TX | May 2026
Relevant Courses: Design and Analysis of Algorithms, Machine Learning, NLP, AI, Big Data Analytics, Database Design.
University of Mumbai | BE – Computer Engineering Mumbai, India | May 2024

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, SQL, Dart, Java (coursework), C/C++ (coursework), HTML, CSS
Backend & APIs: FastAPI, Django, Node.js, REST APIs, Server-Sent Events (SSE), Celery, RabbitMQ, BullMQ, Redis, JWT
Data: PostgreSQL, pgvector, Prisma ORM
Frontend: React, Next.js, Tailwind CSS, Framer Motion, Recharts
AI & LLM: LLM APIs (OpenAI, Gemini), LiteLLM, model routing and fallback, semantic caching, Retrieval-Augmented Generation (RAG), Claude Code
Cloud & DevOps: AWS (Lambda, CloudWatch, DynamoDB, SNS), Docker, Docker Compose, Continuous Integration/Continuous Deployment (CI/CD), OpenTelemetry, Vercel, Railway, Git

PROFESSIONAL EXPERIENCE

Full-Stack Software Engineer (Volunteer) - Universal Removal Defense Project Jul 2026 – Present | Remote, US

- Co-own the Django 6 intake platform routing immigrants and asylum seekers to licensed immigration attorneys, serving 10 languages including right-to-left scripts, as one of two engineers on the repo
- Audited architecture, request flow, and code quality before contributing, then converted the open backlog into a sequenced implementation plan the two-person team is executing against
- Restored multilingual support by fixing a silent locale-compilation failure across Docker startup paths

Software Engineer Intern - Hydracoral Technologies Jan 2023 – Dec 2023 | Mumbai, India

- Integrated third-party facial-recognition APIs into a Flutter/Dart onboarding flow, cutting user verification time by 30% in staging
- Reduced app crash rate by 20% by diagnosing and resolving 15+ bugs across state management, API error handling, and memory leaks, using Firebase Crashlytics and structured debugging workflows
- Lifted usability scores by 20% by building responsive cross-platform UI components, validated through structured user-testing
- Managed CI/CD pipelines (Git, Bitbucket) for a 3-person Agile team and context-switched daily across frontend, backend, and DevOps, integrating REST APIs for JWT auth, data sync, and release automation

PROJECTS

ClauseGuard - AI-Powered Contract Review Platform Apr 2026 – May 2026

- Designed and deployed an 8-service distributed system (FastAPI, Celery, PostgreSQL/pgvector, Redis, RabbitMQ, MinIO) on Docker Compose with health checks, graceful degradation, and OpenTelemetry tracing
- Engineered a saga-pattern pipeline across Celery and RabbitMQ with circuit breakers on LLM calls, fan-out parallelism, and Redis Pub/Sub Server-Sent Events (SSE) for live progress
- Cut LLM costs 30% via pgvector semantic caching and model-tier routing through a pluggable LiteLLM fallback layer, grounding every risk flag in retrieval over 50+ curated clause templates

ZeroWall - Serverless Zero-Trust API Gateway on AWS (3-person course project) Spring 2026

- Built the observability and monitoring layer for a 7-service AWS serverless zero-trust API gateway: CloudWatch log groups, custom metrics, and alarms, plus SNS email alerting
- Wrote a scheduled anomaly-detection Lambda (Python) that scans the DynamoDB audit log every 5 minutes for brute-force and credential-stuffing patterns, plus a Postman test suite for the gateway's endpoints

PR Sensei - AI-Powered GitHub Code Review App Nov 2025 – Jan 2026

- Solved GitHub's hard 10-second webhook timeout against 5 to 15-second LLM calls with an async Node.js pipeline on Redis and BullMQ handling retries, rate limiting, and graceful degradation
- Shipped an AI GitHub App posting PR summaries and inline comments in under 30 seconds via a pluggable OpenAI/Gemini provider layer, deployed on Vercel and Railway
- Designed a multi-tenant PostgreSQL schema (Prisma ORM) feeding a React and Recharts dashboard for review analytics and file-hotspot tracking