

Harsh Upadhyay

Software Engineer | Full-Stack Development & AI/ML Engineering

Jaipur, India | mail2harshx@gmail.com | 7389850145 | linkedin.com/in/codelikeharsh | www.harshcodes.me

Professional Summary

Final-year B.Tech Information Technology student (Manipal University Jaipur, graduating 2026) with hands-on experience building and shipping full-stack, AI-powered systems – from a multi-service SaaS platform with retrieval-augmented generation (RAG) and serious security engineering, to a live, production-deployed voice AI application with a rigorously fine-tuned and measured NLP evaluation model. Comfortable across the full stack (Java, Python, React, Node.js/FastAPI, PostgreSQL/MongoDB), with real testing and CI/CD discipline (400+ automated tests across projects) and genuine security-conscious engineering (OAuth/OIDC, dependency vulnerability scanning). Co-founder of a small web development startup delivering real client solutions end-to-end. Strong in data structures, algorithms, and problem-solving, and a fast, self-directed learner.

Technical Skills

Languages: Java, Python, JavaScript, TypeScript, SQL

AI/ML: Gemini API, OpenAI-Compatible APIs, RAG, Vector Search (Qdrant), LLM Fine-Tuning (LoRA), TensorFlow

Development: FastAPI, React, Node.js, Express.js, REST APIs, WebSockets

Databases & Cloud: PostgreSQL, MongoDB, Redis, AWS, Docker

Tools & Practices: Git, GitHub Actions (CI/CD), Unit Testing, OAuth/OIDC Security

Certifications

Oracle Cloud Infrastructure 2025 Certified AI Foundations Associate (Oracle University)

Work Experience

Co-Founder & Full-Stack Developer — zero2one labs

2026–Present

- Co-founded a full-stack web development startup, owning end-to-end product delivery across 9 completed client projects – from requirements gathering through development, deployment, and support.
- Selected and implemented the right backend, database, and hosting stack per client requirement, applying clean, maintainable coding practices across relational and NoSQL databases.
- Managed direct client communication and feedback across approximately 1–3 iterative revision rounds, translating business requirements into working technical solutions.

Technical Projects

Codenaut | AI-Powered Codebase Intelligence Platform

GitHub

Tech Stack: Python, FastAPI, Gemini API, Qdrant, PostgreSQL, Redis, React, TypeScript

- Designed and built a multi-tenant AI platform end-to-end – FastAPI backend, PostgreSQL/Redis-backed job processing, and a React 19 + TypeScript frontend – that indexes codebases and answers natural-language questions grounded in retrieval-augmented generation (RAG).
- Integrated Gemini and OpenAI-compatible LLM APIs behind one provider-agnostic interface, with vector search over Qdrant (768-dim embeddings) and a bounded, tool-calling AI agent producing evidence-grounded, hallucination-resistant answers.
- Designed the authentication and authorization layer – OAuth with PKCE and OIDC login, short-lived scoped tokens, rotating refresh-token cookies with reuse detection – and ran static analysis and dependency vulnerability scans in CI with non-root Docker images.
- Verified the system with 332 backend tests at 91.04% branch coverage plus a Vitest/Playwright frontend and browser test suite, wired into a CI pipeline that runs on every change.

Anteroom | AI-Powered Mock Interview Platform (Live)

GitHub

Tech Stack: Python, FastAPI, Groq API (Llama 3.3 70B, Whisper), Apple MLX, PostgreSQL

- Built and shipped a real-time, voice-based AI mock interview platform – live at anteroom-sigma.vercel.app – integrating Groq-hosted Llama 3.3 70B and Whisper APIs over WebSockets for adaptive, voice-driven interview practice.
- Fine-tuned a quantized Qwen2.5-1.5B model via LoRA and knowledge distillation, designing a leakage-free train/validation/test split across 720 labeled examples spanning 7 categories and 3 difficulty levels.
- Cut mean-absolute scoring error from 5.3–7.4 to 0.6–1.9 points across four evaluation dimensions on a held-out test set, and improved JSON-output validity from 93% to 100%.
- Diagnosed and resolved multiple real production incidents – an out-of-memory crash from an eagerly loaded model, a cross-origin cookie bug, a WebSocket idle-timeout bug – through systematic root-cause isolation testing.

Education

Manipal University Jaipur, B.Tech in Information Technology

2022–2026

- CGPA: 7.29
- Relevant Coursework: Data Structures & Algorithms, Object-Oriented Programming, Machine Learning/AI, DBMS, Software Engineering

Problem Solving

- Solved 180+ algorithmic problems on LeetCode covering arrays, recursion, dynamic programming, and trees — demonstrating strong analytical and logical problem-solving aptitude.