

VAIBHAV CHAUBEY

+91-7874696877 | vaibhavrm6667@gmail.com | [LinkedIn](#) | github.com/vaibhav449 | [LeetCode](#)

EDUCATION

B.Tech in Computer Science & Engineering

IIIT Raichur, India
Expected May 2027
GPA: 7.3 / 10.0
[Joining Letter ↗](#)

TECHNICAL SKILLS

- Languages:** JavaScript (ES2022), TypeScript, Python, C, C++, SQL, GraphQL
- AI & LLM:** Anthropic Claude API, OpenAI API, RAG pipelines, Hybrid vector + BM25 retrieval, Qdrant, Chroma, LangGraph, MCP, SSE Streaming, Token optimization, Prompt engineering, Conversation evals, Ollama
- Frontend:** React.js (18/19), React Router 7, Redux, Vite (IIFE SDK), Tailwind CSS 4, Axios
- Backend:** Node.js, Express.js, FastAPI, Prisma ORM, JWT, Webhooks, Producer-consumer pipelines, Socket.IO, WebSocket, Cron jobs
- Databases & Infra:** MySQL, MongoDB, Qdrant, Chroma, Docker, Bitbucket Pipelines, Linux, pm2, Shopify Admin/Billing API, Vapi, Twilio, Cloudinary
- Architecture:** Multi-tenant SaaS, Dual-layer caching, Provider-agnostic LLM abstraction, Rate limiting, State machines, IIFE widget SDKs, WebSocket gateways

ACHIEVEMENTS & INTERESTS

- Relevant Coursework:** Data Structures & Algorithms, DBMS, Operating Systems, Computer Networks, Software Engineering
- Problem Solving:** Solved 200+ DSA challenges across [LeetCode](#), HackerRank, and GeeksforGeeks
- Interests:** Agentic AI Architecture, LLM Systems Design, Voice AI Systems, Competitive Programming

EXPERIENCE

Full-Stack Developer Intern (MERN + AI/LLM)

Dec 2025 – Present

Setubridge Technolabs, Ahmedabad

AI Shopping Assistant — Production Multi-Tenant Shopify SaaS (Sole Engineer)

React Router 7 · Node.js · TypeScript · Prisma/MySQL · Qdrant · Socket.IO · Anthropic Claude · OpenAI · Docker · DigitalOcean

- Shipped a 3-service production Shopify app — Remix backend, a JWT-authenticated WebSocket gateway (Socket.IO), and a separate React agent dashboard — serving merchants on DigitalOcean with per-shop row-level isolation across 26 Prisma models and 66 REST routes.
- Cut LLM cost per conversation ~55% by replacing a monolithic prompt with a 7-state intent engine that loads tier-based prompt modules on demand; classification uses cosine-scored embedding prototypes across 3 confidence bands with an LRU-cached GPT-4o-mini fallback (500 entries, 1500 ms timeout) bounding cost on repeat ambiguous queries.
- Built end-to-end bot-to-human handoff over WebSockets — JWT-authenticated agents, atomic SQL-level conversation claim preventing double-claim races, shared inbox across all agents in a shop, multi-tab presence dedupe, and reconnect catchup of messages missed during 2-min disconnect windows.
- Designed a provider-agnostic LLM gateway over Anthropic Claude and an OpenAI-compatible bus (OpenAI, Groq, Together, Gemini) with unified SSE streaming, tool-schema normalization, prompt-cache hit-rate monitoring tied to microdollar-precision cost accounting, and recovery from malformed failed_generation tool calls.
- Integrated Shopify Storefront MCP behind a two-tier cache (10-min tool-discovery + 5-min result TTL) saving ~250 ms/message, and a Qdrant RAG pipeline over 5 source types (page/url/pdf/text/faq) with batched multi-provider embeddings.

OpsPanel — Internal Shopify Pricing & Subscription Platform

React 19 · Express 5 · MySQL · Vite · Shopify Billing API · JWT · Bitbucket Pipelines

- Built an internal full-stack platform to manage pricing plans, subscriptions, and billing for company Shopify apps via Shopify Billing API (GraphQL); multi-tenant architecture with dual-auth (JWT + stateless Shopify token validation), per-shop rate limiting (100 req/min), and a 5-min token cache reducing external API calls by ~80%.
- Developed embeddable pricing widgets (React + Vite IIFE bundles) with MutationObserver auto-mount and cross-widget communication, enabling 10 composable components to deploy on any Shopify storefront via a single <div> tag; integrated Shopify Billing API with a 6-state subscription machine and automated trial-expiry cron.
- Built an admin dashboard covering campaign management, coupon engine, CMS, trial tracking, and automated merchant provisioning — backed by 16 MySQL tables across 15 REST modules (~80K LOC).

AI Demo-Video Generator — Narrated Screen Walkthrough to Auto-Voiced Marketing Video (Built End-to-End)

Chrome MV3 (vanilla JS) · Node/TypeScript/Express · OpenAI GPT-4o (structured outputs) + Whisper · Python · FFmpeg · edge-tts · pm2

- Engineered a Chrome MV3 recorder that turns screen understanding into a JSON merge instead of computer vision, synchronizing tab video, mic audio, live transcript, and DOM events (each with a fallback-chain selector, label, and URL) across a service worker, offscreen document, and content script onto one absolute clock via an explicit readiness handshake, not timing guesses.
- Built a token-gated Express pipeline that fuses events and transcript into an 8-12-scene timeline, then drives GPT-4o structured outputs on a hallucination-guarded prompt grounded only in real UI labels, with a timestamp/scene-parity validator that re-prompts once before falling back deterministically, and server-side Whisper backfilling empty in-browser transcripts.
- Architected a Python/FFmpeg compositor with scene-level time-warp: each scene's video is stretched (clamped 0.5x-3x) toward its voiceover, with residual mismatch absorbed as trailing silence or a held-frame freeze, eliminating chipmunk audio and drift by construction. Deployed under pm2 with a sequential FFmpeg queue so teammates self-serve a finished MP4 in under a minute at cents/video.

TECHNICAL PROJECTS

Multi-System Voice AI Platform — solo project, 4 interconnected systems incl. localization to 2 markets (11K LOC Python)

Python · FastAPI · OpenAI (gpt-4o-mini, text-embedding-3-small) · Chroma/BM25 hybrid retrieval · Vapi · Twilio · WebSocket

- Built a RAG voice agent with hybrid vector+BM25 retrieval and confidence gating as a structural anti-hallucination guardrail — facts come only from retrieval, business rules only from backend code, never the LLM
- Applied a cost-aware deterministic-first cascade (rules → LLM only on ambiguous cases) to both PII redaction and real-time compliance monitoring, cutting LLM calls without sacrificing coverage
- Built 127 conversation evals (100% pass) plus P50/P95 latency and false-positive-rate benchmarking; testing caught 3 real bugs, including one exposed only by live system integration
- Engineered graceful degradation across every external dependency — Twilio-verified call transfers, ASR/TTS fallback modes, KB-outage handling — so no vendor failure breaks a live call

HireScope — AI Interview Assistant

May – Oct 2025

React · Redux · Node.js/Express · Gemini API

- Built a two-tab AI interview app — interviewee chat tab and a live-synced interviewer dashboard — with PDF/DOCX resume parsing (Name/Email/Phone extraction) and auto-prompting for missing fields; state persisted across reloads via redux-persist.
- Gemini-powered interview engine generates 6 role-specific questions with configurable auto-submit timers (20s/60s/120s); scores answers, produces summaries, and surfaces candidates ranked by score with per-candidate chat history and search/sort in the recruiter dashboard.

Placement Website — Placement Management Portal

2025

React · Node.js · MongoDB · Cloudinary · JWT

- Full-stack placement management platform with JWT auth, Cloudinary resume upload/download, and bulk admin workflows — CSV student import and ZIP export of resumes.
- Company CRUD, application tracking with advanced filters (on/off-campus, remote/hybrid), and a coordinator dashboard with key placement statistics.