

VANSH

+91 79825 38137 | hi@vimzh.dev | linkedin.com/in/vimzh | github.com/vimzh

EDUCATION

Maharaja Agrasen Institute of Technology

B.Tech in Computer Science & Engineering, CGPA 8.76/10

Delhi, India

Aug. 2024 to May 2028

EXPERIENCE

ArtifexOne

AI Engineer

Feb. 2026 to Present

Remote

- Built **Keishal**, an AI clinic automation platform across **6 clinics, 80K+ patients**. AI receptionist (Whisper, ElevenLabs TTS, Twilio SIP) handles concurrent calls with **400ms p95 STT latency**, writes to PMS via HL7 FHIR, and verifies insurance at **97.8% accuracy**.
- Shipped a dormant patient reactivation engine using cron-based segmentation (PostgreSQL window functions, Twilio SMS) that flags inactive patients and segments **80K+ records** in under **2s** at **97.3% delivery rate**.
- Built a unified ops dashboard processing **1,800+ daily tasks** with **120ms p99 sync latency** via WebSocket pub/sub, priority queue triage, and role-based RBAC views across 3 staff roles.
- Designed an AI outreach engine: Playwright scraper ingests **4,500+ properties/day** across 12 portals, embedding-based matching (text-embedding-3-small) at **89% relevance**, and a LangGraph agent pitches owners and books callbacks at **3s p95 scrape latency**, zero human touch.

Astroraag

Software Intern

Feb. 2026 to Present

Remote

- Built the full Flutter app solo (Riverpod, GoRouter, Drift): Kundli generation, astro matching, puja booking, WebRTC live consultations. Custom Vedic planetary engine. Stripe Connect, shipped on iOS and Android.

Stealth Startup

Founding Backend Engineer

Oct. 2025 to Present

Remote

- Designed v0 platform from scratch on Node.js, Hono, Bun, Drizzle, Postgres (RLS, pgvector) and Better Auth (OAuth2, WebAuthn); shipped MVP in **6 weeks** with **80ms p95 latency** and **99.8% uptime**.
- Architected an event-driven job runner on Cloudflare Queues (DLQ, exponential backoff) + Workers AI (KV, Durable Objects), processing **1.1M tasks/day** at **4.6×** lower cost than AWS Lambda baseline.

PROJECTS

Kensaku | Multi-modal Retrieval Harness *TypeScript, Node.js, Hono, LangChain.js, Vercel AI SDK, Qdrant, Redis*

- Engineered a multi-modal RAG pipeline ingesting PDFs, audio (Whisper-v3) and images (CLIP) into a unified Qdrant index; **2.1M+ chunks** via HNSW + scalar quantization at **38%** memory of float32.
- Cut hallucination from **22% to 4.6%** (RAGAS) using query decomposition, HyDE and cross-encoder rerank; p95 retrieval + generation latency **640ms**.
- Streamed **2,300 tok/s** via SSE on Node.js + Hono, **14×** cheaper than GPT-4o at parity quality.

Karasu | Open-source Agent Harness *TypeScript, Bun, Vercel AI SDK, Zod, Postgres + pgvector, OpenTelemetry*

- Authored an open-source TypeScript agent harness with multi-step tool calling, parallel sub-agent dispatch, Zod-validated structured outputs, persistent memory and resumable workflows.
- Shipped a built-in eval harness with **200+ deterministic tests** per agent, OpenTelemetry traces to Grafana, catching regressions pre-deploy.
- Released on npm with **1.4K stars** and **8,200+ weekly downloads**.

ACHIEVEMENTS

- 80+ merged PRs** across **25+ open-source and production repos**. Core contributor to **fb55/htmlparser2** (62M+ weekly npm downloads), **inokawa/virtua** (3.5K stars), and **aymericzip/intlayer** (711 stars).
- Won **1st prize at ADGIPS hackathon** (1,200+ teams) and **2nd prize at Technohacks NSUT** (1,800+ teams). **Codeforces 1433**, **CodeChef 5**, **LeetCode 1800** (500+ solved).

TECHNICAL SKILLS

Languages: TypeScript, JavaScript, Dart, SQL, Python, Bash

AI & Agents: Vercel AI SDK, LangChain.js, LangGraph, RAG, Qdrant, pgvector, RAGAS, tool calling, eval harnesses

Backend: Node.js, Bun, Hono, Next.js 15, Express, Drizzle, Postgres, Redis, Kafka, Better Auth

Mobile: Flutter 3, Dart, Riverpod, GoRouter, Drift, Platform Channels

DevOps: Docker, Terraform, GitHub Actions, AWS, Cloudflare Workers/Queues, Vercel, Git, Linux, Vitest, Playwright, OpenTelemetry, Grafana, Sentry