HARSH KESHARI

+91 9598642369 | Delhi NCR, India | harshkeshari100@gmail.com | Portfolio | LinkedIn | GitHub

Education

KIET Group of Institutions — **Ghaziabad, India**
Bachelor of Technology, Computer Science & Engineering (AI) — Aug 2021 – Jun 2025

Skills

Languages & Frameworks: TypeScript, JavaScript, Python, Node.js, Express, React, Next.js, Prisma
AI/ML: RAG (embeddings, Pinecone) · LLM agents & tool-calling (LangGraph, LangChain) · evals (LLM-as-judge, golden datasets, regression) · LLMs (OpenAI, Google Gemini) · speech & realtime voice (Deepgram, GCP STT)
Cloud, Infra & DevOps: AWS (EC2, Lambda), GCP (Cloud Tasks, KMS), Docker (multi-stage), Vercel, GitHub Actions CI/CD, nginx; Bull + Redis queues, node-cron
Databases & Architecture: PostgreSQL, MongoDB, Redis, Pinecone; Microservices, Multi-tenancy, Vector Search

Experience

Founding Engineer — **Nov 2024 – Present**
Experiment Labs — *Remote*

- **Built the recommendation engine** on OpenAI embeddings + Pinecone vector search: the platform's primary revenue driver, serving 3,500+ personalized activity recommendations to students across 8 partner orgs including Leverage Edu and PhysicsWallah.
- **Shipped an agentic AI learning copilot**: LangGraph + Gemini agent with realtime voice (WebSocket); grounds replies on each student's activity, progress, and knowledge gaps plus RAG over a 100+ skill library; tool-calling (GitHub, web search, tasks); regression-tested in CI by an LLM-as-judge grader calibrated to human ratings.
- **Architected the platform's 10 production microservices** spanning core (Auth, CRM, Payments, Notifications), AI services (Recommendation Engine, Copilot, Smart Meeting Assistant, AI Interviewer), and product modules (Internship Portal, Evaluator, Ops Dashboard); built the central identity provider behind all of them (auth-core) with SSO, JWT + refresh sessions, Google/Zoom OAuth, service-to-service tokens, and a cached RBAC engine enforcing multi-tenant org isolation.
- **Engineered Smart Meeting Assistant**: recorded and transcribed 500+ counselor–student sessions with AI summaries auto-pushed to learner profiles; async Bull + Redis queue with a background worker batch-processing transcription across orgs.
- **Developed AI Interviewer**: Google Cloud Speech-to-Text for student answers, Vertex AI for adaptive question generation and scoring, async video transcoding via Cloud Tasks.
- **Implemented Payments and Notifications services**: Razorpay subscription billing with webhook reconciliation; Brevo transactional email with scheduled reminders.
- **Migrated long-running services (transcription, video processing, recommendation regeneration) from AWS Lambda to EC2**, eliminating serverless timeouts; added cron monitors for payment reconciliation and meeting-link retries.

Open Source

Oppia Foundation (Quality Team Lead, Nov 2023 – Apr 2024): led LaCE, Oppia's frontend quality team; merged 10 PRs across admin acceptance tests and UI fixes.
Tailcall: resolved search modal focus bug; earned $150 bounty from the maintainer team.
Also contributed to **CircuitVerse** (digital circuit simulation) and **Zowe** (Linux Foundation's open mainframe platform).

Projects

Crelyzor: Demo | **PERN stack · TypeScript · Prisma · Bull/Redis · Next.js · Deepgram Nova-3 · Google Gemini · Recall.ai · Google Cloud KMS/GCS · Docker · GitHub Actions**

- **Built an all-in-one productivity OS for teams** (cards + scheduling + meeting AI + tasks; replaces HiHello, Cal.com, Otter.ai, Todoist): live in production with usage-metered billing, across a 4-repo system (API, dashboard, SSR public site, admin) with Docker multi-stage builds and GitHub Actions CI/CD.
- **Engineered the meeting-intelligence pipeline**: Deepgram Nova-3 diarized transcription, Gemini AI summaries and task extraction, SSE-streamed Ask AI with history, and Recall.ai bots that auto-join Meet/Zoom calls; plus human-in-the-loop transcript and summary editing.
- **Shipped multi-user Teams**: shared cards, team booking pages, invites and roles, and team-scoped access control enforced across every meeting, task, and card.
- **Designed a timezone-aware scheduling engine**: slot calculation across availability, GCal busy-time, and buffers; serializable transactions prevent double-booking races; bidirectional GCal push webhooks.
- **Launched an internal admin and ops portal**: audit logs, system health (queue depth, Redis, KMS), per-team DEK rotation and crypto-shred, and an encryption-coverage dashboard.
- **Hardened security and encryption**: field-level encryption at rest (Google Cloud KMS, per-user/team AES-256-GCM DEKs) kept searchable via a blind index, key rotation, GDPR crypto-shredding; fixed XSS/SSRF, httpOnly JWT, CSP, rate-limited admin auth.