

SHEKHAR KUMAR

Bengaluru, IN | 8409209072 | shekharkumar235838@gmail.com | linkedin.com/in/shekhar-kumar1 | github.com/hitmandark07

SUMMARY

Software engineer with 4+ years building AI infrastructure and backend systems — agent evaluation environments, LLM orchestration platforms, RAG pipelines, and distributed GPU inference. Works across TypeScript, Python, Go, Node.js and PostgreSQL, with end-to-end ownership from architecture through production deployment.

EXPERIENCE

Pure Billion Technologies

Mumbai, IN · Remote | Aug 2022 – Present

Senior Software Engineer

NeatCode — Repository Intelligence Engine for AI Coding Agents (Open Source · 3 npm packages)

- Built a TypeScript engine (**20.5k LOC**) that indexes a codebase into a SQLite code graph and serves it to AI coding agents over MCP — **131k LOC indexed in 3.8s** (~35k LOC/s, single-threaded ts-morph with batched transactions), answering structural queries in **<300ms**.
- Designed a content-hash incremental sync layer cutting re-index cost from **3.8s to 0.49s** on an unchanged 1,821-file repository.
- Shipped **11 MCP tools** for graph traversal, rule validation, and prompt-driven convention capture, returning **12–60x less context** than a grep-and-read agent baseline for the same query.
- Maintained **510 tests** at a 0.68:1 test-to-code ratio, full suite green in 12s.

Fleet — Agent Evaluation & Reinforcement Learning Platform

- Led development of a platform that recreated production applications such as GitHub and Booking.com as interactive UI environments for agent training and evaluation.
- Enabled agents to execute multi-step tasks, observe application state, and receive task-level feedback for reinforcement learning.

Generative AI Platform

- Architected a production-grade LangChain platform implementing model routing, categorized prompt orchestration, context management, and tool-calling across multiple fine-tuned models, plus a custom MCP server for real-time web search and extensible tool integration.
- Architected Retrieval-Augmented Generation pipelines over vector databases, improving AI response relevance and accuracy by **25%**.

Distributed Speech-to-Image Inference Pipeline

- Architected a real-time pipeline orchestrating multiple GPU-backed pods, routing ComfyUI workflow requests across parallel queues with single-concurrency execution per worker for scalable parallel inference.

Platform Engineering

- Designed and owned the backend architecture of **Ivorie**, a marketplace serving **5k+ users** (NestJS, Expo React Native, Next.js), with real-time chat and notifications handling **10k+ messages/day**, Stripe Connect payment flows, and an automated multi-tenant migration system.
- Designed an LLM-assisted testing strategy combining user-story-driven test generation, randomized edge-case scenarios, and conventional unit/integration tests, increasing test coverage by **35%**.

Invest With Tribe

Chandigarh, IN · Remote | Feb 2022 – Aug 2022

Software Engineer I

- Built a stocks investment platform enabling experts to publish strategies for retail investors, with analytics dashboards and leaderboards used by **1k+ users**; integrated AWS SQS and Twilio WhatsApp for real-time trade notifications.

SKILLS

AI & ML	RAG, Vector Databases, LangChain, LangGraph, MCP, LLM Evaluation, Agent Evaluation, RL Environments
Languages	TypeScript, JavaScript, Python, Go, C++, Solidity (basic)
Backend	Node.js, NestJS, Express, REST APIs, WebSockets, Prisma
Data	PostgreSQL, MongoDB, Redis, DynamoDB
Cloud	AWS (EC2, S3, SQS), Docker, CI/CD, Git
Frontend	React, React Native, Expo, Next.js

EDUCATION

Bachelor of Computer Applications, Techno Main Salt Lake

2020 – 2023 · Kolkata, IN

ACHIEVEMENTS

EV Charging App recognized among the **Top 30 Startups in India** (Inc42, 2022). · Mentored **50+ students** in React and React Native development at PrepBytes.