Amogh Upadhyay

GenAI Engineering Lead • Product, design and full-stack AI

hi@amogh.cloud • amogh.cloud • linkedin.com/in/amoghupadhyay • github.com/The-DarkMatter • Mumbai, India

First prototype to daily use → Solo builder to lead of three engineers
Tested before every release • Product to code to hardware

PROFESSIONAL SUMMARY

GenAI engineering lead with 5+ years building systems people use. I lead engineering for **Neome**, the AI assistant retail investors use daily in the Kotak Neo app: I built its first working version myself and now lead three engineers. I start from the user, design the interface, own the architecture and evaluation, and see it through to real users. Outside work I keep building, from a local voice assistant on one GPU to an Android ride tracker and an F1 app for LG TVs, and I publish the measurements at **amogh.cloud**, including the ones that proved me wrong.

TECHNICAL SKILLS

Languages: Python 3, SQL, TypeScript / JavaScript, Java (Spring), Kotlin

Applied AI & LLM Systems: Agent orchestration, tool / function calling, Text-to-SQL, RAG, hybrid (vector + keyword) retrieval, entity resolution, prompt engineering, context & KV-cache engineering, prompt caching, self-correcting agents, synthetic data generation, LLM fine-tuning (LoRA, QLoRA), local inference (llama.cpp), voice pipelines (VAD, ASR, TTS)

Evaluation & Reliability: Evaluation harnesses and graders (tool-call accuracy, answer relevance, faithfulness), release gating, production monitoring, latency & cost profiling, same-session A/B benchmarking

Product & Design: User research from production behaviour, interface design for AI products, 10-foot TV UI, mobile-first UX

Platform & Data: Docker, Kubernetes, CI/CD, cloud-agnostic architecture, Cloudflare Workers / D1 / R2, FastAPI, MySQL, PostgreSQL, SQLite, Redis, React, Astro, Android, ESP32

PROFESSIONAL EXPERIENCE

Kotak Securities (now Kotak Neo) — Mumbai, India

Innovation Lab: a lean, startup-style team that builds GenAI products fast.

GenAI Engineering Lead (Sr. Manager), Neome — *May 2024 – Present*

- Lead engineering for **Neome**, Kotak Neo's AI assistant for exploring Kotak's research, used daily by retail investors in the Kotak Neo app, and with my team took it from **first prototype to daily use**.
- Built the **first working version myself**, then grew into **leading three engineers**; lead its technical direction and **co-designed our test automation**, with separate checks for answer relevance, faithfulness to tool results and retrieval; we decide every release on its results.
- Drive product thinking from the engineering side: I proposed several of Neome's product ideas, among them the **Morning Brief**, the **F&O payoff tables** and **commodities** coverage, which my team and the product team built together. I read real conversations and talk to customers, and my engineering team runs its **own user-level analysis, production monitoring and model evaluation** alongside the product team.
- Took the assistant through several generations of text-to-SQL design and **retired my own fine-tuned pipeline** for a single prompt-driven step, cutting latency and training cost and letting us change models without a code deploy.
- Built for reliability: queries that correct themselves within a bound, limits that live in code, where a prompt can't argue around them, and **one unified lookup tool** in place of several overlapping ones.

Backend Engineer — *January 2023 – May 2024*

- Built Java / Spring backends for retail investing products from inception, including **corporate-action adjustment algorithms** that keep historical data accurate.
- Automated the monitoring of critical jobs and services, replacing manual checks for the whole team.

Ministry of Agriculture & Farmers Welfare — New Delhi, India

Senior Research Fellow, Office of the Secretary *July 2022 – January 2023*

- Turned crop and programme data into executive Tableau dashboards for the Office of the Secretary of Agriculture; contributed to **Agristack**, India's national farmland-data infrastructure.

Ministry of Rural Development — New Delhi, India

R&D Fellow (IIT Bombay CTARA fellowship), MGNREGA division *July 2021 – June 2022*

- Delivered a **real-time data collection and reporting platform** with dashboards for national, state and district officials; coordinated **12 states and union territories** for the MGNREGA Mid-Term Review.

SELECTED PROJECTS

Piaa: local-first voice assistant *Python, llama.cpp; designed and built solo*

- Talk to it and interrupt it mid-sentence; it handles my mail, WhatsApp and files, with private tools kept on the PC.
- Full-duplex voice assistant with dozens of tools on one **RTX 5080**. Traced a throughput ceiling to one offload flag: **+40% decode for +189 MiB VRAM**, and retracted three of my earlier conclusions in public.
- KV-cache-aware context design after measuring that a new tools array costs a full re-prefill (**69x slower**); agent safety rules live in code, where they can't be argued around.

Slipstream: Android ride tracker *TypeScript, Expo, Cloudflare Workers; in closed testing*

- Records rides with only the phone's GPS, replays them coloured by speed, runs live group rides with friends and family, and tracks fuel spends.
- GPS track codec that bounds position error at **0.556 m** at the same byte cost (rounded deltas drift 73 m); whole backend designed to run on free tiers; live group rides.

makranddubey.com: books recovered from print files *Astro, Android*

- Recovered four Hindi books from legacy typesetting files, validated against independent counts, and built their website and an Android reader, now in the Play Store, designed for readers over fifty.

PlamoGB: group buying for a Gunpla community *Cloudflare Workers, D1, R2; co-built with Vinayak and Yash*

- A web app and Discord bot used daily by a community of over 3,000: it splits each group order's price, shipping and customs down to the rupee, and tracks payments and parcels. It replaced their spreadsheets.
- Members have recorded **47 lakh rupees** in payments to each other through it (12 Nov 2026), and the monthly total has gone up every month since April, **6.7x** from April to October.

Lights Out: an unofficial F1 app for LG TVs *JavaScript, webOS*

- Built for the TV remote: natural navigation, a camera switcher and a live timing panel that stays in sync with the race. Every onboard camera, and seasons back to 2018, from the sofa. Open source (Apache 2.0) and submitted to the webOS Homebrew Channel, awaiting review.

Kursi: a seat-booking demo you can stampede *Java, Spring Boot, MySQL*

- Every seat gets exactly one owner, because each claim is a single conditional database update; load testing found hot seats queueing on a row lock, and a non-locking check first fixed it.
- Live at kursi.amogh.cloud with a public load dashboard: race a friend, or a crowd of bots, for the same seat.

Also: ambient TV lighting on an ESP32 that I soldered myself, and tapoviz, an open-source music visualiser for smart light strips.

EDUCATION

M.Tech, Technology and Development, **IIT Bombay** *August 2019 – July 2021 • CPI: 8.31/10*

Applied research: sensing groundwater use on irrigation pumps with IoT, sending the data over existing powerlines, and low-cost smart irrigation.

B.E., Computer Science Engineering, **UIT, RGPV Bhopal** *August 2014 – July 2018 • Percentage: 73%*

ACHIEVEMENTS & RECOGNITION

MD's Award (2024) and **Innovation Lab Champion (H2 2023-24)** for performance and innovation across backend and AI delivery.