

Naman Kumar

New Delhi, India
naman@outlook.in
+91 85277 05775
linkedin.com/in/namaneng

Senior Software Engineer with 6+ years building production AI systems — agents, context retrieval, evaluations, and infrastructure. Scaled the AI coding assistant Cody from idea to multi-million ARR across 250+ enterprises, and built autonomous agents at Railway that diagnose and fix production failures. M.S. in Machine Learning & AI.

EXPERIENCE

Railway

Dec 2025 – Jun 2026

Senior Product Engineer

- Designed, built, and launched an **autonomous deployment diagnosis** and **error monitoring agent** that analyzes runtime logs, failed deployments, code, and configuration across **millions of services** deployed on Railway.
- Architected the reliable execution and evaluation infrastructure for Railway Agents using **Temporal workflows**, enabling **autonomous code fixes** and continuous quality improvements.
- Shipped billing, usage metering, and rate limiting for Railway Agents using **Stripe Meters**, supporting **300K+ paid subscribers** consuming **~\$2M annualized** in agent tokens.
- Designed **log fingerprinting** for a **ClickHouse**-based runtime log ingestion pipeline processing **1M+ RPS**, enabling scalable detection of new, regressed, and high-volume production errors.

Sourcegraph

Apr 2022 – Dec 2025

Senior Fullstack Engineer

- Scaled **Sourcegraph's Cody** from prototype to **multi-million ARR** AI coding assistant serving **250+ enterprises** and **15K+ daily developers** including Netflix, Uber, Canva, Grab and Booking.com.
- Designed Cody's **hybrid multi-repo context retrieval** pipeline using **BM25**, **vector embeddings**, and **reciprocal rank fusion (RRF)**, enabling reasoning across **thousands of repositories**. Evaluated OpenAI and Sentence Transformer embeddings.
- Implemented **OpenCtx** before the emergence of **MCP**, then later adopted the industry-standard **Model Context Protocol (MCP)** to enable Cody to retrieve context and execute actions across GitHub, Slack, Sentry, Datadog and other developer tools.
- Built monetization infrastructure with **Stripe subscriptions**, usage-based rate limits, and abuse prevention, enabling Cody's PLG-driven growth to **multi-million ARR**.
- Tech Lead for Cody Web**, launching standalone Cody on the web, embedding inline AI assistance across code search, and shipping prompt libraries and a unified AI + code search experience across **VS Code** and **JetBrains**, powered by intent detection.
- Tech Lead for Permission Center**, rebuilding code-host permission sync across thousands of repositories and users per customer, and giving admins a faster way to audit repo access and sync status.
- Built in-product analytics dashboards quantifying developer time saved and **customer ROI**, enabling major enterprise expansion deals.

Product Hunt

Sep 2019 – Mar 2022

Fullstack Engineer

- Built, launched, and A/B tested **10+ product growth experiments**, including personalized home feeds, content gating, promo codes, and ad placement optimizations, driving user acquisition and retention.
- Built rule-based spam detection services that automatically flagged **80% of fake accounts**, significantly reducing moderation efforts by the support team.
- Improved page load times and cut origin load by building a **Cloudflare CDN edge-worker caching layer** that served cached pages to logged-out users while conditionally bypassing the cache for users in **active A/B experiments**.
- Built and operated the platforms for Product Hunt's flagship community events, including the annual **Golden Kitty Awards** and quarterly **Maker Fest** hackathons, handling nomination and voting flows at peak traffic.

INTERNSHIPS

GraphQL Foundation

May 2020 – Aug 2020

Google Summer of Code

- Developed a fast, maintainable **streaming GraphQL parser** adopted in **graphql-js** and **GraphiQL**, improving performance and enabling syntax-aware highlighting.

Grofers

May 2019 – Aug 2019

Fullstack Engineer Intern

- Automated customer support at Grofers by building a self-serve chat system that handled **~12K daily tickets** during peak sales, cutting manual workload by **65%** and reducing mean response time by **25 minutes**.

Redacumen

Jan 2019 – Aug 2019

Contract Fullstack Engineer (Part-time)

- Built an **accounts-payable automation platform** that cut AP operations cost by up to **85%**, including an **OCR-based** document aggregation and auto-indexing pipeline (email, scanner, cloud) and an ERP sync agent.
- Built a **cloud kitchen food-ordering platform** spanning three apps: a consumer app for **subscription-based ordering**, a cook-facing app for menu and order management, and a delivery-hub app for dispatch coordination.

HackerEarth

Jun 2018 – Apr 2019

Frontend Engineer Intern

- Led frontend development of **FaceCode**, a remote interview platform with real-time coding (**WebSockets**), live video, and code execution.
- Built enterprise analytics, billing, and team management features, including **RBAC**, invitations, and admin workflows.

Coding Blocks

Jun 2017 – Nov 2018

Software Engineer Intern

- Built an online IDE supporting **13 programming languages** with real-time collaboration features.
- Wrote a **static site generator** that cut code redundancy by **80%** and saved hours of daily content-management effort for the website.
- Taught **Data Structures & Algorithms** and **React.js** to **100+ students**.

EDUCATION

Liverpool John Moores University

2022 – 2024

M.S. in Machine Learning & AI

- Thesis: Optimizing Information Retrieval in Conversational Context via LLM Query Rewriting

Guru Gobind Singh Indraprastha University

2016 – 2020

B.Tech in Computer Science & Engineering

- Project: Distributed Key-Value Fault-Tolerant In-Memory Database

SKILLS

AI Engineering

Coding Agents, AI Agents, RAG, MCP, Context Engineering, Browser Automation, Token Compression, LLM Evaluations, LLM-as-Judge, NLP

Search & Retrieval

Code Search, Hybrid Search, Information Retrieval System Evaluations, BM25, Vector Search, Embeddings, Sentence Transformers

Backend & Observability

Temporal, ClickHouse, Postgres, Kafka, Redis, Docker, Kubernetes, OpenTelemetry, Datadog, Sentry, AWS, GCP, Cloudflare, Vercel

Languages & Tools

TypeScript, Python, Go, Rust, SQL, React, Next.js, GraphQL, gRPC, tRPC, Playwright, Vitest, Ruby