Gaurav Kumar

Founding Engineer
+91 7894421989 | +91 8178916280 | gouravkumar2701@gmail.com
www.gkcs.in | linkedin.com/in/gauravkumar2001 | github.com/Gauravkumar2701

Summary

Founding engineer with 3+ years of experience building AI products and the full-stack systems behind them, including RAG pipelines, LLM integrations, agent tooling, backend architecture and cloud-native deployments.

Skills

AI: RAG, LLM Integration (Gemini, OpenAI, Claude), Tool Calling, Evaluation, Prompt Engineering, LangChain
Languages: TypeScript, JavaScript, Python, SQL
Backend: Node.js, NestJS, FastAPI, Express.js, Prisma, REST APIs
Frontend: React, Next.js, React Query, Tailwind CSS
Data: PostgreSQL, MongoDB, Redis, BigQuery, Milvus, Meilisearch
Cloud and DevOps: Google Cloud (Cloud Run, Cloud Tasks), Azure, Docker, Kubernetes, GitHub Actions, CI/CD

Experience

GalenAI Apr 2026 – Present
Founding Engineer Remote, India

- Built the AI service behind faculty content generation from scratch, including MCQ, mind-map and flashcard pipelines with source-grounded retrieval (FastAPI, Gemini, Milvus).
- Made generation reliable by moving background jobs from arq/Redis to Google Cloud Tasks, with safe retry handling.
- Built AI evaluation tooling: a benchmark harness, golden checklists and a retrieval threshold calibrated on labelled data.
- Owned the core NestJS backend: tenant isolation, invite-only auth, audit logging and cloud uploads (Prisma, PostgreSQL).
- Main author of the institution dashboard: calendar, bulk import, lesson editor and AI workspace (Next.js, React Query).
- Built the licensing system and partner portal, plus analytics and observability using BigQuery and structured logging.
- Built Faculty Sandbox (faculty.galenai.io), a free self-serve version of the platform for faculty.
- Cut frontend CI test time from 507s to 200s and raised backend test coverage from 32.8% to 79.7% across 1,607 tests.

PureCode AI May 2025 – Mar 2026
Senior Software Engineer Hyderabad, India

- Led backend architecture for PureCode Enterprise, defining the Node.js, NestJS and PostgreSQL stack and API standards.
- Engineered an LLM abstraction layer across GPT and Claude models, enabling dynamic model switching and A/B testing.
- Shipped a Terminal Manager and tool calls (file operations, command execution) for the VS Code extension's agents.
- Migrated payments from Paddle to Stripe, including customers and active subscriptions, with zero downtime.
- Built a sales automation pipeline across Apollo, HubSpot and Instantly, plus CI/CD pipelines on Azure App Service and AKS.

LearnLater Apr 2024 – Apr 2025
Software Engineer Remote

- Implemented LLM response streaming on LearnUpAI, sharply reducing perceived latency.
- Built a learning-plan engine for enterprise clients and platform-wide full-text search with Meilisearch.
- Built two-way Outlook and Google Calendar integrations and an email-ingestion webhook.
- Added Redis and HTTP caching on high-traffic endpoints, improving response times and reducing upstream load.

Massive Mobility Nov 2023 – Mar 2024
Backend Developer Intern Gurugram, India

- Built secure REST APIs for the Zecat and Bhamo web apps with JWT auth, input validation and rate limiting.
- Developed a real-time payment analytics dashboard for operations and finance teams.

Projects

Growthsauras (growthsauras.com) Co-founder
Growth agency for founders and startups: outbound sales, marketing automation, web development and launches.

Hotinbox Apr 2024 – Present
AI sales workflow connecting Apollo, Clay, HubSpot and Instantly (TypeScript, NestJS, MongoDB, React, n8n).

VoiceType 2025
macOS dictation app with on-device Whisper transcription and push-to-talk capture (Python, whisper.cpp, VAD).

Voxa 2025
Multilingual text-to-speech app with voice cloning from a short sample (Next.js, FastAPI, PyTorch, XTTS-v2).

Education

Gandhi Institute of Engineering and Technology Aug 2020 – May 2024
B.Tech in Computer Science and Engineering Odisha, India