Kausik D

+91 63695 75179 kausikdevanathan@gmail.com
linkedin.com/in/kausik-devanathan github.com/Kaus1kC0des kausik.codes

Education

St. Joseph's College of Engineering Nov 2022 – May 2026
B.Tech Artificial Intelligence and Machine Learning (GPA: 8.06 / 10.0) *Chennai, IN*

Experience

InfactIQ (via Qik Office) Oct 2026 – Present
AI Engineer *Remote, US*
DepoIQ – AI Legal Deposition Analysis Platform (FastAPI, Postgres + pgvector, GCP)

- Architected and engineered **hybrid search** for Ask AI, the platform's legal research assistant, combining **BM25** keyword and **pgvector** semantic search with a **Jev**-based relevance ranker to surface the most relevant testimony first.
- Designed and built **subagent-driven search** that splits whole-case questions across **parallel LLM subagents**, so every document is covered and each quoted passage is verified against its source.
- Built a **cross-cloud integration** that lets a partner upload depositions to **AWS S3** and have them processed automatically on **GCP** (Lambda, Cloud Workflows, Cloud Run), with status updates sent back at each stage.

Qik Office July 2026 – Present
AI Product Engineer *Chennai, India*

- Engineered a speech/NLP transcription and diarization pipeline (Whisper + Pyannote), evaluated at **95% speaker-attribution accuracy** and **7–9 min** per audio hour; adopted by **95% of enterprise customers**.
- Wrote a custom **MCP server** exposing Qik Office services (meetings, email, notes, tasks) as LLM tools; built agents on the **OpenAI Agents SDK**, writing agent prompts and tool descriptions for cross-service workflow orchestration.
- Shipped an **OAuth 2.0** (PKCE) server and public **REST API** (**56 endpoints**, OpenAPI 3) for a multi-tenant SaaS without modifying existing controllers; validated via **27 E2E suites** and mutation testing.

Qik Office Aug 2024 – June 2026
Software Developer Trainee *Chennai, India*

- Architected real-time transcription service with per-meeting dynamic port allocation; SFU routes audio via **FFmpeg** to **Faster-Whisper**, streaming live transcripts over **WebSockets**.
- Implemented GPU auto-scaling via **DigitalOcean API**, provisioning new instances at **85% VRAM utilization** and sustaining **8–10 concurrent meetings per GPU**.
- Sole-owned CPU-to-GPU migration on **DigitalOcean** – provisioned instances, configured CUDA and Python runtimes, and deployed all inference services, cutting processing time from **40+ min to 8–10 min** (**$5\times$ speedup**).

Projects

Takkaru – E-commerce Platform | *Next.js*, *TypeScript*, *PostgreSQL*, *Prisma*, *Razorpay*, *AWS*, *Terraform* takkaru.com

- Built and run a live full-stack e-commerce site for a local apparel brand: backend, payments, infra and **React** pages (checkout, account, auth, CMS blocks); **Razorpay HMAC-verified webhooks**, atomic stock updates in payments.
- Provisioned **AWS** (VPC, EC2, RDS, S3 + CloudFront, Route 53) as modular **Terraform**; **GitHub Actions** CI/CD gates deploys on 945 tests and ships via **GitHub OIDC**; **Postgres full-text + trigram search**.

Technical Skills

Generative AI: LLMs (OpenAI, Gemini), RAG, Embeddings, Vector Search, Prompt Engineering, LangChain, Agents, MCP
ML / NLP: PyTorch, NLP/Speech (Whisper, Pyannote), Computer Vision (YOLO, CNNs), Model Evaluation
Languages: Python, JavaScript, SQL
Backend & APIs: FastAPI, REST (OpenAPI), Node.js, WebSockets, OAuth 2.0/PKCE, Microservices
Cloud & DevOps: AWS, GCP, DigitalOcean, Docker, Terraform, CI/CD (GitHub Actions), CUDA, GPU Inference
Databases: PostgreSQL, MySQL, MongoDB, Redis
Practices: Git/GitHub, Testing (E2E, Mutation), System Design, API Security