GAURAV KASHYAP

Full Stack AI Engineer

gaurav404.gk@gmail.com

www.linkedin.com/in/gaurav-kashyap-909504172/

+12368650534

https://www.gauravkashyap-portfolio.com/

EDUCATION

University of Toronto: *MEng, Computer Engineering* (Emphasis: ML, Data Analysis) — GPA 4.0 **Sep 2024–Sep 2025**
(Toronto, Canada)

Simon Fraser University: *BASc, Mechatronic Systems Engineering* (with Distinction) - GPA 3.76 **Sep 2018–Sep 2023**
(Burnaby, Canada)

EXPERIENCE

Senior AI Engineer — Makalu Health *(Sept 2025–Present)*

https://makaluhealth.com/

- Built a multi-agent provider enrichment pipeline (TypeScript, Firecrawl, GPT-4o, Postgres/pgvector, Prisma) that scopes the url to be scraped, performs iterative gap analysis, normalizes fields, embeds profiles, and persists to a clinician graph → fresher data and faster sourcing.
- Engineered job-side enrichment & schema parity so RAG compares like-for-like across roles and providers (OpenAI embeddings, pgvector) → higher-precision candidate–job alignment and consistent downstream Q&A.
- Implemented Fusion RAG: LLM job analysis → dynamic weighted query gen → BM25 + vector retrieval → RRF re-rank → LLM-driven shortlist (LangChain, pgvector, BM25) → improved top-K quality and reduced recruiter screening time.
- Architected a Multi-Agent Residency Intelligence System (Navigator / Content Analyzer / Decision Coordinator + smart context manager, section-aware extraction) to autonomously explore residency sites and extract residents, curriculum, sourcing with 80%+ success rates across 500+ medical residency programs across the United States.
- Operated a privacy-first, scalable ingestion stack (BullMQ/Redis concurrency, role-scoped access, PII handling aligned with HIPAA practices) → reliable crawls, auditable access, and compliance ready workflows.

AI Engineer— The Delivery Company ***(Founder)*** *(Jan - Sept 2025)*

https://thedeliverycompany.net/

- Engineered an AI delivery platform using LangChain, RAG, and MCP (Model Context Protocol) for memory aware logistics automation.
- Orchestrated a multi-agent LLM workflow (LangChain + MCP) where voice/chat transcripts are analyzed for parameter extraction and executed through schema-validated MCP tool calls that invoke webhook endpoints backed by serverless functions; integrated STT/TTS for hands-free route and fleet management.
- Enhance AI models with LoRA fine-tuning on LLaMA pre trained models for agent tasks, achieving more consistent domain-specific performance compared to external LLM APIs, while reducing latency and cost.
- Engineered a dual-server real-time optimization engine (OSRM + Concorde/OR Tools on EC2) that outperformed Mapbox in all Ontario scenarios, leveraging custom Ontario traffic profiles for near-Google-level accuracy.
- Implemented real-time WebSockets to provide customers with live driver locations and dynamic ETAs, enhancing user experience.
- Managed backend infrastructure using Typescript, Java, and Supabase, handling complex workflows, scheduled cron jobs, and multiple serverless functions for robust operational efficiency.

Full-Stack Engineer — Vivvion *(Dec 2023 – Sep 2024)*

- Migrated a hardware ad network to event-driven microservices with AWS Lambda, API Gateway, S3, DynamoDB, cutting real-time sync delays by ~40% and delivering ~99.9% devic - cloud uptime.
- Built secure ingestion/analytics pipelines and BLE→cloud bridges, improving low-latency content delivery across fleets; added CI/CD + observability and least-privilege IAM.

Embedded Software Engineer — Illumina Technology *(May 2022 – Oct 2023)*

- Developed ARM Cortex-M4 (FreeRTOS, C/C++) firmware for ADC/DSP, CCD calibration, and LED control; hardened UART/SPI/I²C stacks with error detection to reduce field failures by ~30%.
- Wrote Python calibration & test tooling; applied DSP/ML for signal quality and pattern recognition in diagnostic probes to speed validation and improve reliability.

PROFESSIONAL PROJECTS

All About RAG — RAG Evaluation Platform *(Sept 2025)* **https://all-about-rag.vercel.app/**

- Built an interactive RAG comparison system covering 12+ architectures (Self-RAG, CRAG, HyDE, Agentic RAG, etc.) with real-time side-by-side evaluation.
- Implemented end-to-end pipeline: multi-format doc parsing → chunking → OpenAI embeddings → hybrid pgvector/HNSW search.

NextGenEduCoder — AI LeetCode Agent *(Aug 2025)* **https://b2-codegen.vercel.app/**

- Built a multi-agent system (Analyzer, Planner, Coder, Reviewer) with real-time reasoning stream and secure Judge0 execution over SSE.
- Added self-improvement loop (failure-pattern mining + adaptive strategy selection) to refine solutions across sessions
- Built Conversation + Knowledge Base style experience: embeddings, chunking, semantic search, streaming UI; deployed on Vercel + Supabase.

Realtor X - AI Media Studio + CRM + Voice/SMS *(May 2025)* **https://realtorx.app/**

- Built an end-to-end AI media & ops platform (Next.js, Postgres/pgvector, AWS Lambda, LangChain, Gemini/OpenAI) consolidating toolchain and automating lead → booking.
- Implemented conversational memory via RAG for cross-session continuity in voice workflows; voice orchestration with STT/TTS and agent tools.
- Orchestrated 2D→3D tour pipeline (diffusion + SAM segmentation + vector retrieval + auto-voiceover + stitching) to produce interactive property tours.
- Added LLMOps: version control, evals (correctness/latency), tracing, usage caps, and caching for predictable unit costs.

HelloGenie AI — Voice-Powered Automation *(Sept 2024)* **https://hellogenieai.com/**

- Built inbound call agent with Deepgram STT + OpenAI TTS, RAG/pgvector memory, and interruption-aware multi-agent orchestration.
- Orchestrated tasks/scheduling and human handoff; enforced PII controls, prompt/versioning, and usage caps.