GAURAV KASHYAP

AI Systems Engineer

✉ *gaurav404.gk@gmail.com*

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

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

EDUCATION

University of Toronto: *MEng, Computer Engineering* (Emphasis: ML, Data Analysis) GPA 4.0 **Sep 24–Sep 25**

Simon Fraser University: *BASc, Mechatronic Systems Engineering* (Distinction) GPA 3.76 **Sep 18–Sep 23**

HIGHLIGHTS

- End-to-end experience building and operating LLM-based production systems, including multi-agent platforms processing 500K+ entities with comprehensive observability, evaluation frameworks, and guardrails.
- Expert in LLMOps infrastructure: LangSmith tracing/evaluation, prompt versioning, A/B experimentation, cost tracking, and automated quality monitoring in high-throughput environments.
- Extensive experience with conversational AI and agentic systems: voice agents (STT/TTS), chat interfaces, RAG pipelines, and multi-channel orchestration (SMS, email, voice) with human-in-the-loop escalation.
- Strong DevOps/MLOps foundation: CI/CD pipelines, Docker, Kubernetes, AWS (Lambda, EC2, S3), distributed task queues (BullMQ/Redis), and infrastructure-as-code.
- Proven track record optimizing LLM costs via fine-tuning (LoRA on Azure AI Foundry, AWS Bedrock), caching strategies, and usage caps while maintaining quality benchmarks.

TECHNICAL SKILLS

- **Languages:** Python, TypeScript/JavaScript, C/C++, SQL
- **LLM/AI Frameworks:** LangChain, OpenAI Agents SDK, CrewAI, MCP, RAG (BM25, pgvector), Hugging Face
- **LLMOps & Observability:** LangSmith (tracing, evals), prompt versioning, A/B testing, cost tracking, guardrails, PII filtering
- **Infrastructure:** Docker, Kubernetes, AWS (Lambda, EC2, S3, DynamoDB, Bedrock), Azure AI Foundry, Supabase, Vercel
- **Data & Queuing:** PostgreSQL, Redis, BullMQ, pgvector, DynamoDB, event-driven architectures

EXPERIENCE

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

- Architected multi-agent autonomous recruitment system using OpenAI Agents SDK with models deployed via Azure AI Foundry for enterprise-grade reliability, achieving 75% recruiter handoff rate without human intervention and reducing cycle time from 1 month to <1 week.
- Built comprehensive LLMOps observability layer with LangSmith for end-to-end tracing, evaluation pipelines, latency monitoring, and automated regression detection across all agent workflows.
- Implemented production guardrails including deterministic output validation, PII filtering middleware, and schema enforcement preventing LLM generated database corruption, reducing incidents to near-zero.
- Designed playbook driven execution engine with adaptive branching and node-based state machines, orchestrating multi-channel outreach (SMS, email, voice) with real-time intent tracking and conversation context.
- Integrated AI voice screening via CSM model with dynamic prompt assembly and intelligent human-in-loop escalation, reducing unnecessary recruiter interruptions by 95%.
- Engineered hybrid RAG pipelines combining BM25 lexical search with vector similarity and LLM re-ranking for candidate-job matching; built automated evaluation framework measuring retrieval precision and answer quality.

AI Engineer: <u>The Delivery Company</u> *(Founder)* ***(Jan - Sept 2025)***

- Built conversational AI delivery platform using LangChain and MCP where voice/chat transcripts trigger autonomous multi-step workflows with schema-validated tool execution.
- Fine-tuned LLaMA models using LoRA on Azure AI Foundry for domain-specific agent tasks, reducing latency by 40% and API costs by 60% while maintaining quality benchmarks.
- Engineered dual-server route optimization combining OSRM (real-time) and Concorde TSP (optimal sequencing), achieving near-Google-level accuracy and outperforming Mapbox in all tests in Ontario.
- Deployed real-time WebSocket infrastructure for live tracking with graceful degradation; managed CI/CD pipelines on Supabase with scheduled jobs and serverless functions.

Full-Stack Engineer: <u>Vivvion</u> ***(Dec 2023 – Sep 2024)***

- Migrated hardware ad network to event-driven microservices (AWS Lambda, API Gateway, DynamoDB), reducing real-time sync delays by 40% and achieving 99.9% device-cloud uptime.
- Built secure data ingestion/analytics pipelines with CI/CD, observability dashboards, and least-privilege IAM policies for production reliability.

Embedded Software Engineer: <u>Illumina Technology</u> ***(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.

SELECT PROJECTS

<u>All About RAG: RAG Evaluation Platform</u> ***(Sept 2025)***

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

<u>NextGenEduCoder: Multi-Agent Reasoning System</u> ***(Aug 2025)***

- Built multi-agent system with specialized roles (Analyzer, Planner, Coder, Reviewer) featuring real-time reasoning stream and secure code execution over SSE.
- Added self-improvement loop with failure-pattern mining and adaptive strategy selection — automated evaluation driving continuous model improvement.

<u>Realtor X: AI Media Studio + CRM + Voice/SMS</u> ***(May 2025)***

- Built end-to-end AI platform (Next.js, Postgres/pgvector, AWS Lambda, LangChain) with conversational memory via RAG for cross-session continuity in voice workflows.
- Added comprehensive LLMOps layer: prompt versioning, PII redaction, correctness/latency evals, distributed tracing, usage caps, and response caching for predictable unit economics.

.

<u>HelloGenie AI: Voice-Powered Conversational Agent</u> ***(Sept 2024)***

- Built production inbound call agent with Deepgram STT + OpenAI TTS, RAG/pgvector memory, and interruption-aware multi-agent orchestration for natural conversations.
- Implemented LLMOps controls: prompt versioning, PII redaction, human handoff protocols, and usage caps for cost management.