SAURABH KISHORE

+1 (647) 572-2197 | saurabhkishore.com | gksaurabh@outlook.com | github.com/gksaurabh

SUMMARY

Machine Learning Engineer building production-grade LLM and retrieval systems for large-scale intelligence and analytics applications. Experienced across the full ML lifecycle including data pipelines, model fine-tuning, inference infrastructure, and evaluation. Passionate about building scalable AI systems that bridge research innovation with production deployment.

EXPERIENCE

AI/ML Engineer Dec 2024 – Present
JSI, Ottawa, Ontario — *IRIS - AI enabled search platform for law enforcement and intelligence agencies*

- Architected and deployed a production **agentic Retrieval-Augmented Generation (RAG)** platform enabling analysts to query large-scale intelligence datasets using natural language (Agno).
- Built scalable **LLM inference pipelines** integrating vector search, custom knowledge retrievers, and multi-agent reasoning workflows using agentic frameworks such as Agno.
- Developed high-throughput **data pipeline infrastructure** using Flink, LanceDB (Vector store), Amazon S3 and ANN indexing techniques (IVF-PQ, HNSW) to support large-scale semantic retrieval.
- Trained and fine-tuned LLMs using LoRA (PEFT), and built end-to-end pipelines for automated evaluation, deployment, and scalable inference with vLLM.

Software Developer / Cyber Auditor Oct 2023 – Dec 2024
Department of National Defence, Canada

- Developed backend services for operational logistics systems using **C#, .NET, and SQL** supporting Canadian Armed Forces operations.
- Implemented **CI/CD pipelines and unit testing workflows** using Azure DevOps, improving reliability of mission-critical systems.
- Conducted technical cybersecurity audits and analyzed enterprise datasets to identify system vulnerabilities and operational anomalies.

EDUCATION

Georgia Institute of Technology 2026 – Present
M.Sc. Computer Science — Machine Learning.
Courses: Machine Learning, Deep Learning

Carleton University 2019 – 2024
B.Sc. Computer Science
Thesis: Semantic Retrieval Engine using LLM Embeddings and Ranking Algorithms

B.Cog.Sc. Cognitive Science (Honours) — AI Specialization.
Courses: AI and Cognitive Science, NLP, Cognitive Modeling

TECHNICAL STRENGTHS

Machine Learning and GenAI: LLM routing / RAG / Agentic Orchestration / LoRA Fine-Tuning / Knowledge retrieval systems / NLP / Agno / LangGraph / DeepEval / AI enrichement pipelines

Data & Infrastructure: LanceDB / Apache Flink / Distributed Systems / vLLM / LLM inferencing and routing / LambdaLabs / Azure Foundry / RunPod

Programming: Python / C# / JavaScript / SQL / Java

DevOps: Docker / Kubernetes / Helm / CI/CD / GitHub / Bundle deployments (Hauler)