KAUSHAL KUMAR SUBRAMANI

Waterloo, Ontario | kaushalkmrgr8@gmail.com | github.com/rkhexed | LinkedIn

TECHNICAL SKILLS

Languages: Python, SQL, DQL, TypeScript, JavaScript, Java, C
AI/ML: PyTorch, Hugging Face Transformers, scikit-learn, Pandas, NumPy, LoRA/PEFT fine-tuning, quantization
LLM & Agents: LangChain, LangGraph, LangSmith, CrewAI, MCP, n8n, RAG, model evaluation & benchmarking, semantic chunking
Databases: Qdrant, Milvus, Postgres, Supabase, MongoDB, Neo4j, Dgraph
Infrastructure & Web: Docker, GCP Cloud Run, Terraform, FastAPI, Flask, REST APIs, Git, Linux, React, Next.js

EXPERIENCE

Kinara Systems (Co-op) | *AI/ML Software Engineer* January – April 2026

- Deployed, benchmarked, and **fine-tuned Qwen 2.5 3B** on edge hardware, improving telemetry and log classification by **10%** over baseline while cutting external LLM API cost and keeping customer data on-device.
- Built a self-hosted **OpenAI-compatible inference service** for on-device SLMs behind an Apache reverse proxy with **API key** authentication, letting local models drop into existing LLM-based workflows.
- Redesigned the **knowledge ingestion pipeline** for PDF and PPT sources, chunking with **semchunk** under a hierarchical-plus-semantic strategy and a max token cap, improving **retrieval precision** across base and user knowledge stores in **Qdrant**.

Dawson Partners (Co-op) | *Machine Learning Engineer* May – August 2025

- Built a **LangSmith** evaluation system with retry logic and multimodal support that invokes **LangGraph** agents in parallel, making benchmarking fast and repeatable across environments and raising **output reliability 33%** on time-sensitive financial calculations.
- Developed a community prompt gallery with **FastAPI** and folded contributions into the evaluation pipeline, creating a feedback loop between prompt usage and agent benchmarking that improved response consistency.
- Deployed the first custom **MCP** Excel server on **GCP Cloud Run** with **Terraform**, automating releases via volume mounts and schedulers, then extended the tooling to generate formatted reports, charts, and pivot tables.

IGS (Co-op) | *Junior ML/AI Engineer* September – December 2024

- Increased **RAG** accuracy by **13%** and **ROUGE** scores by 6% with an auto-tester built to score retrieval quality.
- Achieved **80% classification accuracy** across user story domains, platforms, and software stacks via exploratory data analysis and experiments with classical ML and **Hugging Face** models.

Wilfrid Laurier University | *Quantum Machine Learning Research Assistant* June 2026 – Present

- Built and tested **quantum-classical ML pipelines** for a **MARL** framework driving decentralized multi-drone navigation.
- Replaced the **DQN**'s classical value network with a **parameterized quantum circuit (PQC)** as the Q-function approximator, benchmarking returns and trainable parameter counts against the classical baseline.

[case]HACKS | *VP of Technology* | GitHub September 2025 – May 2026

- Shipped the platform running a **200-hacker** event, spanning an organizer control plane (**Next.js**, **Supabase**) and participant portal (**React**, **Flask REST API**, JWT auth) with constraint-based **team matching** and **realtime QR check-in**.
- Built a sponsor dashboard with live event metrics, filtering, and bulk signed-URL resume exports.

PUBLICATIONS

Multi-Agent Email Security System | *23rd L&T Conference, IEEE* | Paper Published 2026

- Achieved **84.1% zero-shot phishing detection accuracy** on a dataset the system had never seen, architecting a **5-agent system** in **CrewAI** that combines **BERT** classification, WHOIS validation, and threat intelligence to autonomously analyze and quarantine emails.
- Engineered **parallel agent execution** with **n8n** middleware and SQLite storage, using weighted coordination to fuse linguistic, technical, and threat-intelligence assessments concurrently.

PROJECTS

LLM-Guided Semantic Path Planning for ROS2 Nav2 | *Python, PyTorch, ROS2* | GitHub May – September 2026

- Ran a research **RRT*** planner as a drop-in **Nav2 global planner** via a costmap adapter, then embedded a **class-conditioned semantic cost** at all four RRT* cost sites so tree rewiring makes the clearance-versus-distance trade-off.
- Diagnosed that first-solution termination silently disabled rewiring, converted to **anytime RRT***, and benchmarked against Nav2's **NavFn** baseline across matched goals: **2–7% shorter paths**, **97% goal-reach**, **zero collisions**.

EXTRACURRICULARS

AI/ML Club, Wilfrid Laurier University | *President* January 2026 – Present

- Lead the university's first AI/ML club backed by AI faculty, building and running project teams, organizing workshops that bring junior students into the field, and forming research groups and hackathon teams.

EDUCATION

Wilfrid Laurier University | *Honours BSc Computer Science, Co-op* Expected December 2026