Karanvir (Karan) Khanna

Machine Learning Engineer · Cloud Architect (GCP/AWS) · E-Commerce Optimization

416-347-3761 | karanvir.khanna@mail.utoronto.ca | Personal Website | GitHub | LinkedIn

Education & Professional Profile

Profile: Hack the North Winner (2025) and ML Engineer with deep expertise in **GCP-native AI solutions** and **distributed systems**. Proven track record in building high-availability e-commerce platforms using **Cassandra**, **Redis**, and **Kubernetes**. Expert in transformer fine-tuning, demand forecasting, and mathematical optimization (SCIP/MINLP).

Availability: Graduating April 2026 — Available for immediate start (Full-time).

University of Toronto Apr. 2026
Honours Bachelor of Science in Computer Science CGPA: 3.7/4.0 · Dean's List

- **Key Coursework:** Neural Networks & Deep Learning (CSC413), Fundamentals of ML (CSC311), Distributed Systems, Algorithm Design, Linear Algebra, Probability & Statistics, Operating Systems.

Technical Skills

Cloud & Infrastructure: GCP (BigQuery, Pub/Sub, Vertex AI), AWS (EC2, Lambda, S3), Docker, Kubernetes, Terraform, NGINX, GitHub Actions (CI/CD)
Databases & Caching: Cassandra (Wide-column), Redis (In-memory), Firestore (NoSQL), PostgreSQL, MongoDB, Pinecone (Vector DB)
Machine Learning: PyTorch, TensorFlow, Hugging Face, CodeT5, Gemini 1.5 Pro, Prophet, Scikit-learn, XGBoost
E-Commerce Stack: SAP Business One, Moneris API, OMS/WMS Logic, Auth0, Shopify API, Stripe, Amazon SP-API
Optimization: SCIP, CPLEX, Pyomo, Mixed-Integer Nonlinear Programming (MINLP), Gurobi
Languages: Python, TypeScript/Node.js, SQL, C/C++, Java, GDScript, Bash, Go

Experience

Machine Learning Engineer & E-Commerce Architect Jun. 2024 – Sept. 2025
IPPINKA Toronto, ON

- Architected an **event-driven** inventory and order sync platform on **GCP** using **Pub/Sub** and **Cloud Functions** to propagate real-time updates across OMS/WMS, Shopify, and Amazon marketplaces for 5,000+ SKUs.
- Engineered a **BigQuery** analytics pipeline over 10+ years of sales and operations data; trained and deployed **demand forecasting** models to reduce stockouts and improve inventory turnover by 23%.
- Built **semantic product search** using **Vertex AI embeddings** and ranking heuristics; improved query relevance and increased search-to-cart conversion by 18%.
- Implemented a low-latency metadata layer with **Firestore** and caching patterns to replace a legacy Sheets-based backend, cutting API latency by 35% and improving reliability under burst traffic.
- Developed personal **AI automation workflows** (including a Gemini-based documentation agent) that enabled a sustained **average productivity of 7+ hours daily active coding time (3.5x average dev)**, significantly accelerating delivery.
- Built an automated **purchase-order optimizer** in Pyomo using **SCIP/CPLEX** to solve supplier constraints (MOQ, lead time, budget) and generate feasible, cost-efficient bulk orders.
- Developed Python middleware for **SAP Business One** and **Moneris** to automate reconciliation, invoicing, and near real-time financial reporting.
- Designed and shipped a **customer loyalty and points system** integrated with **Auth0** for secure identity and personalized marketing triggers.
- Mentored 3 developers on ML and production readiness, and authored technical documentation for data pipelines and solver orchestration.

AI/ML Engineer – Data Governance May 2023 – Oct. 2023
KPMG LLP Toronto, ON

- Developed **LLM-powered assistants** using OpenAI and Gemini APIs to automate data governance workflows, reducing manual compliance review time by 60%.
- Designed **GCP and Azure ML architectures** for enterprise-scale PII detection and automated classification across multi-cloud data lakes.
- Implemented **similarity matching pipelines** for Master Data Management (MDM) using fuzzy matching and Levenshtein distance to create Golden Records from noisy data.
- Authored KPMG's Data Governance Playbook with ML-enhanced automation strategies; presented findings to 50+ enterprise clients.

Machine Learning Projects

Tarazoo (Hack the North Winner 2025) [Repo] | *GCP, FastAPI, PyTorch, SCIP, Redis, Cassandra* 2025–2026

- Co-built an **AI-powered inventory optimization SaaS** for Shopify merchants, live at tarazoo.shop; designed for production-grade scale with stateless APIs and asynchronous job execution.
- Implemented **demand forecasting** pipeline (Prophet, ARIMA) with automated backtesting and retraining triggers; produced SKU-level forecasts feeding downstream optimization and reorder recommendations.
- Built **constraint-based purchase order generation** with **SCIP** to solve real-world procurement constraints (MOQ, lead times, budget, stockout risk), generating feasible POs optimized for cash flow and service level.
- Integrated **NLP review intelligence** using transformer embeddings for sentiment/topic signals to adjust demand priors and surface product-quality drivers in merchant analytics.
- Deployed on **GCP Cloud Run** with containerized microservices; used **Pub/Sub**-style eventing patterns for decoupled ingestion, forecasting, and optimization stages, enabling horizontal scaling under burst traffic.
- Engineered low-latency state and feature access with **Redis** caching and high-write telemetry storage patterns (clickstream and events), enabling near real-time dashboards and model monitoring.

Python-to-Java Code Translator [Repo] | *PyTorch, Hugging Face, CodeT5, CUDA* 2025

- Fine-tuned a **CodeT5 transformer** on curated Python→Java pairs with a reproducible training pipeline (tokenization, batching, mixed precision, checkpointing) for neural code translation.
- Implemented evaluation harness using **BLEU/CodeBLEU** plus compile-and-run checks; achieved **78% compilation rate** on translated outputs and documented failure modes (types/imports/APIs).
- Built inference workflow for scalable serving: batched generation, deterministic decoding options, and post-processing for syntax/format normalization to improve compilation stability.

GuardianCruise – Driver Safety AI [Repo] | *OpenCV, Python, ML Classification, Cohere* 2024

- Built a **real-time CV pipeline** (OpenCV) for driver monitoring: eye aspect ratio tracking, yawn detection, and distraction cues with on-frame temporal smoothing for stability.
- Trained and integrated an **ML classifier** for driver state and impairment risk; used an **LLM-based alerting layer** (Cohere) to generate context-aware interventions and escalation logic.

ML Boilerplates Library [Repo] | *PyTorch, scikit-learn, Reproducible Training* 2025

- Created a **24+ model** implementation library covering supervised, unsupervised, and embedding workflows with consistent training loops, metrics, and experiment structure.
- Authored technical guides and decision trees for model selection, feature engineering, and evaluation, optimizing for rapid iteration and production handoff.

Tormented by Lights – Adaptive AI Game [Game] | *Godot, GDScript, Object Pooling* 2024

- Built a **Game Jam winner** (2D platformer) featuring a custom **Dynamic Difficulty Adjustment (DDA)** engine that modulates obstacle speed and spawn density using a PID controller based on player reaction time.
- Implemented efficient **object pooling** and a deterministic **physics state machine** in GDScript to handle collision detection and memory management, ensuring consistent 60FPS on web builds.

Research Experience

Matroid Theory & Polytope Optimization [Repo] | *University of Toronto* May 2023 – Sep. 2023

- Conducted research on **volume estimation of matroid polytopes** under Prof. Ahmed Ashraf, establishing links between combinatorial geometry and **loss landscapes** in high-dimensional optimization.
- Modeled **submodular set functions** as polytope constraints to analyze **greedy algorithm convergence**; findings have theoretical implications for **active learning** and **feature selection** in efficient ML.
- Applied **Hepp bounds** to estimate volumes, validating conjectures in SageMath that bridge discrete structures with **continuous relaxation methods** used in neural architecture search.

Certifications & Awards

Hack the North Winner (2025) — Shopify Challenge: AI Shopping with Computer Vision & Optimization.

Invited to YCombinator (YC) Dinner (2025) — Selected for exclusive founder networking event.

Halton Game Jam Winner (2025) — Best Game

DAMA Certified Data Management Professional (2023) — Associate Level (80% Score).

UofT Hacks Runner-up (2023) — Best use of gen AI

DeerHacks Winner (2022) — Best Usage of UiPath: Created NLP-powered Essay Generator.

Dean's List Scholarship (2021–Present) — Awarded for academic excellence.