

Hiu Yan Kwok

(647) 885-1284 | hiuyan.kwok@mail.utoronto.ca | [linkedin.com/in/hiu-yan-kwok-14435625a](https://www.linkedin.com/in/hiu-yan-kwok-14435625a) | github.com/hiuyear
hiuyankwok.com | Toronto, ON

EXPERIENCE

The Matter Lab, University of Toronto

Toronto, ON

AI Engineer & AI Research Intern, Software Engineering on EL AGENTE

May 2026 – Present

- **Sole software engineer** on the agent-memory subsystem of a production multi-agent LLM platform (UofT / Vector Institute / Acceleration Consortium), supervised by Prof. Alán Aspuru-Guzik; **owned the full lifecycle** from design to shipped service to its evaluation benchmark
- Built the typed storage layer in async Python, validating every write (schema, references, file paths) so bad data fails fast, never silently, then a record/replay **evaluation harness** with swappable adapters; **red/green TDD**, distributed tracing

Department of Computer Science, University of Toronto

Toronto, ON

Software Developer, CAVE Project (Volunteer)

May 2026 – Present

- Ship features to CAVE, an **open-source** CLI + web tool **used by 500+ students** and maintained by the UofT CS Department (TypeScript, Node.js, React); it maps architecture dependency graphs for Java repos
- Shipped a **full-stack** feature end to end in an unfamiliar codebase, through code review; building the project's **CI/CD** pipeline in GitHub Actions (Jest, Vitest, Playwright, lint and type-checks on every push and PR)

Royal Bank of Canada (RBC)

Toronto, ON

Data Science & Analytics Intern, Customer Data Management

Jan. 2026 – Apr. 2026

- Traced a **200K** client-count discrepancy across **42M+** records with PySpark and SQL, proving it was a definition mismatch between divisions, not a pipeline bug; mapped the platform's SOAP-to-REST API migration endpoint by endpoint and caught contracts that would have broken in production

PROJECTS

Bedrock Prompt Router: Cost-Optimizing LLM Routing Microservice | *Python, AWS CDK, DynamoDB* 2026

- Designed, **deployed and operated** a microservice on AWS (API Gateway → Lambda → DynamoDB → Bedrock) as **infrastructure-as-code** in CDK, shipped by GitHub Actions via a short-lived OIDC role, not stored keys
- Load-tested the live API with **500** skewed-traffic requests, found the key that would throttle a single DynamoDB partition, and proposed **write-sharding**; monitored cost and cache-hit metrics in CloudWatch

Agora Village / Agora Evals: Multi-Agent LLM Sim & Eval | *Next.js, TypeScript, Python* | [\(code\)](#) · [\(code\)](#) 2026

- Built and deployed a **full-stack** app (Next.js, Postgres) where LLM agents trade resources; an async 202-Accepted API runs each simulation in a **durable, restart-safe workflow** traced with OpenTelemetry
- Companion harness evaluates it as a **black box** through its public HTTP API and Postgres only, never its code; a pilot across 2 model conditions found **non-overlapping 95% bootstrap CIs** on wealth inequality

EDUCATION

University of Toronto

Toronto, ON

Hon. BSc, Mathematics, Computer Science & Statistics

2024 – 2028 (Expected)

- **Relevant Coursework: Data Structures & Algorithms** (CSC263), Software Design in Java (CSC207), Systems Programming in C / Unix (CSC209), Machine Learning (CSC311), Probability with Proofs (STA257)

TECHNICAL SKILLS

Languages: Python, TypeScript, Java, C, JavaScript, SQL, Solidity, HTML/CSS

Backend & Web: REST APIs, Node.js, Express, FastAPI, Next.js, React, Postgres, DynamoDB, PySpark

Cloud & CI/CD: AWS (Lambda, API Gateway, DynamoDB, CloudWatch, CDK), GitHub Actions, Linux

AI / ML: Multi-agent systems, LLM evaluation & benchmarks, RAG, PydanticAI, LangGraph, scikit-learn, NumPy

Software Engineering: TDD (pytest, Jest, Vitest, Playwright), code review, load testing, OpenTelemetry, monitoring

Spoken: English, Cantonese, Mandarin (**all fully fluent**)