Ilia Alenabi

📱 +1 (778) 708-2776 | @ ialenabi@uwaterloo.ca | in linkedin.com/in/ilia-alenabi/ | 🌐 iliall.com

Education

University of Waterloo Waterloo, Ontario
*Honours **Computer Science**; **Artificial Intelligence** Specialization; **Statistics** Minor* *Sep 2022 – Expected*
GPA: 4.0 – President's Scholarship of Distinction

Awards & Achievements

National Mathematical Olympiad – Silver Medalist
Combinatorics Olympiad (ICO) – Silver Medalist

Skills

Languages: Python, C/C++, Java, JavaScript, TypeScript, Golang, Rust

Frameworks/Libraries: Numpy, Pandas, PyTorch, Tensorflow, LLVM, MLIR, React, Django, Next.js

Tools/Platforms: Git, Docker, AWS, Kubernetes, Redis, BigQuery, PostgreSQL, Pinecone, MongoDB

Experience

Cerebras Toronto, ON
Machine Learning Engineer - Internship *Sep 2025 – Dec 2025*

- Built a **tensor-dump** comparison tool to validate parity between **CPU** and **CSX** inference runs for **GPT-OSS**
- Developed a **checkpoint-conversion** pipeline for all **inference** models with a peak memory usage of only **7%**
- Refactored the **inference pipeline** configs to enable compatibility with custom **tokenizers** and **image encoders**

Huawei Canada Toronto, ON
Compiler Engineer - Internship *Jan 2025 – Apr 2025*

- Designed an **LLVM** pass for automated software cache creation, tuning memory usage in **distributed systems**
- Analyzed **Redis**'s performance, identified hotspots, and implemented **prefetching** to reduce runtime by **20%**
- Developed an **MLIR** pass to annotate attention layers with **sharding** metadata, optimizing **tensor distribution**

Cohere Remote
Data Engineer - Part-time *Sep 2024 – Apr 2025*

- Designed high-quality training and evaluation data used to fine-tune and benchmark **Cohere's production** LLMs
- Evaluated complex **math** and **coding** outputs from LLMs, identifying **reasoning errors** and providing **solutions**

Questrade Financial Group. Remote
Software Engineer - Internship *Sep 2024 – Dec 2024*

- Automated **cloud-based MLOps** pipeline for **fine-tuning** in **Java**, reducing redundant AI inference costs by **8%**
- Integrated **NLP DevOps** solutions, fine-tuned on **1,000+** support tickets, across **50+** internal Slack channels
- Integrated **authentication** checks into the internal pipeline, auditing SQL code from **200+** developers for security

Silverberry Group Vancouver, BC
Data Scientist - Internship *May 2023 – Apr 2024*

- Implemented a **U-Net** for diabetic wound segmentation and a classifier in **PyTorch**, achieving **92%** accuracy
- Developed an **interactive agent environment** to model **medication-use** behavior prior to product release
- Vectorized 200+ hours of doctor-appointment audio in **Pinecone** using **OpenAI Whisper** and **Hugging Face**

Research

Vector Institute Toronto, ON
*Research Intern – **Interpreting Vision-Language Models*** *May 2025 – Aug 2025*

- Developed tooling to extract **intermediate** representations from **vision-language** models for **reasoning** analysis
- Ran **probing experiments** on **CLEVR** to evaluate **VLMs**' internal representations of **primitive concepts**

Pingoo AI Remote
*Research Intern – **Building Trustworthy Generative AI for Diabetes Care*** *May 2024 – Aug 2024*

- Examined the **reliability** of LLMs for diabetes-focused applications using **few-shot learning** and **fine-tuning**
- Created a **self-evaluation** loop to enable models to refine their own outputs, achieving **85%** human-rated accuracy