Mahdiar Khodabakhshi

Data Scientist, Canada, Toronto

mahdiar.edu@gmail.com ☎ +1 647-394-3016 LinkedIn GitHub Portfolio

Technical Skills

Programming: Python(scikit-learn, Numpy), PyTorch, SQL, React, REST APIs, Linux, Bash/CLI automation, R, OOP.

Machine Learning: RAG/RALM-style pipelines, Prompt Engineering, Retrieval systems, Similarity & Embedding models & Rankers, Agentic Systems, Vector Databases, Knowledge graphs (KG), Model Context Protocol (MCP) servers/Tools.

Tools and Platforms: Cload Services(AWS, S3, Lambda, Azure ML), Docker, Chroma DB, MongoDB, Reids, Kubernetes.

Work Experiences

Ross Video | *Machine Learning Engineer* **Toronto, CA ◆ *Sep. 2025 – Present***

- Built **SEA(structure-aware candidate filtering)** using feature schema constraints to reduce the candidate-set by **92% (12× fewer candidates)**, then applied **HTS (Hybrid Term-Search)** in Amazon OpenSearch to rank targets by combining **term frequency** with **semantic similarity** to the query, improving **Hit@5** by **18%**.
- Deployed an **on-prem** Knowledge Graph term retriever integrated with **AWS S3** and **AWS Lambda**, boosting tagging throughput **5×** and **reducing** p95 latency from **15.0s** to **2.8s** (**81% faster**) versus the LLM Tag Generator.
- Implemented a **generator–critic** (review-and-critique) **multi-agent** loop with specialized critic agents that score separate criteria (structure, semantics, and rule compliance) and provide **iterative feedback**.
- Built automation for configuring and running **A/B tests** on **model variants** within the AWS hosted customer demo.

Vector Institute | *Machine Learning Researcher* **Toronto, CA ◆ *Fall 2024 & Jan. 2026-Present***

- Optimized VideoDiffHMR inference by **cutting unnecessary tokens** during inference**(step pruning)** and compressing token representations **(token merging)**, which reduced latency by 90% **(10× faster)** with a small accuracy reduction.
- Built a resumable evaluation pipeline to benchmark **layer-wise embeddings** from Pythia checkpoints across **MTEB** tasks, generating per-layer and per-task metrics to quantify **representation quality and efficiency trade-offs**.
- Built a **LinUCB bandit** to select optimal Pythia checkpoint×layer configurations, shrinking the search from **4,650** candidates (155 checkpoints × (6 layers for 70M + 24 layers for 410M)) to **200** bandit-selected evaluations **(- 95.7%)**.
- Achieved up to **40× faster** experimentation, by running 10 × L40 nodes (Including parallel finetuning & evaluating).

Toronto Metropolitan University | *Machine Learning Researcher* **Toronto, CA ◆ *May. 2024 – Sep. 2025***

- Designed and implemented a **dual-stage KG retriever** (first & second-hop candidate generation + hybrid re-ranking with **ColBERT + BM25**), achieving a **10× improvement** in **gold-triple retrieval**.
- Finetuned SOTA models (e.g., CodeLlama 34B) to generate accurate SPARQL in a RAG system, improving F1 by 11%.
- **Fine-tuned on-prem SLMs** (Phi-3 Mini, Gemma 2B) for SPARQL structure motif recognition, boosting accuracy **27%**, **saving $\tilde{\$}$700** in OpenAI usage(gpt-4o-mini), and **tripling** end-to-end processing **speed (3×)**.
- Designed a DPR+Cross Encoder Retriever which outperformed Adaptive-RAG, Self-RAG, IRCoT by $\tilde{1}$4% in top-k.

Cyberometrics | *Data Scientist* **Toronto, CA ◆ *May. 2024 – Sep. 2024***

- Achieved **20%** reduction in processing latency by utilizing **Azure ML** for model training and deployment, **Azure Kubernetes Service(AKS)** for scalable containerization, and **Azure Blob** Storage for efficient data handling.
- Implemented **Tokenization**, **NLU-based privacy filters**, and secure data storage solutions to enable chatbot's ability to handle privacy-sensitive queries and ensuring full compliance with **GDPR standards**.
- Implemented our **Vector Database** using **ChromDB** to enhance semantic search performance by **25%**, improving the chatbot's ability to retrieve contextually relevant and accurate responses based on recognized tokens of input query.

Publications

IPM 2026 **Improving SPARQL Query Generation using Triple-Augmented Generative Language Models and Dynamic Chain-of-Thought Prompting**

***Mahdiar Khodabakhshi**, Fattane Zarrinkalam, Faezeh Ensan*

Proposed **DyCoT-TAGLM**, a dynamic reasoning framework for **text-to-SPARQL** that couples a **triple-augmented** language model with **dynamic** Chain-of-Thought prompting and a **Hybrid Example Search** combining **semantic** and **SPARQL-structural** similarity to retrieve exemplars and refine queries for robust generation.(Submitted)

Education

Toronto Metropolitan University ***Sep 2023 – Aug 2026 (Expected)***

*B.Eng. in **Software Engineering** | Minor in **Applied Mathematics*** *Toronto, ON, Canada*

*Minor in **Applied Mathematics** — 2× Bridging Divides Research Internship — Mitacs Research Internship at Germany*