

Hargun Mujral

 github.com/hargunmujral |  [linkedin.com/in/hargun-mujral](https://www.linkedin.com/in/hargun-mujral) |  x.com/hargunmujral

 hmujral@uwaterloo.ca |  +1 (604) 440-0712

EDUCATION

University of Waterloo

Bachelor of Computer Science (BCS), Minor in Combinatorics & Optimization + AI

Sept 2021 - Apr 2026

cGPA: **92%**

Dean's Honour List · Software Eng. Society Director of Finance · Academic Representative · Waterloo AI Institute researcher

SKILLS

Languages & Frameworks: Python, C++, TypeScript, Scala, Java, SQL, GraphQL, Bash, Node, React, Flask, Django, Spring

ML & Data Analysis: PyTorch, Tensorflow, Pandas, Numpy, SciPy, Sklearn, Databricks, Matplotlib, OpenCV, Kaggle

Technologies & Infra: Linux, Git, Docker, Jenkins, Kubernetes, CUDA, AWS, GCP, Azure, PostgreSQL, Nginx, Mixpanel

EXPERIENCE

Hudson River Trading

Algorithm Engineer (Co-op)

May 2025 - Aug 2025

New York, USA

Databricks Mosaic Research

ML Engineer (Co-op)

Jan 2025 - Apr 2025

San Francisco, USA

- Created a Python framework for asynchronous GPU scheduling, enabling hyperparameter sweeps on Databricks Notebook
- Successfully rolled out feature updates to over 2,000 research scientists and engineers, directly unblocking contracts worth **\$500,000+**, and reducing end-to-end latency by **>50%** for all asynchronous workloads
- Re-engineered the Databricks Notebook cell output buffer by implementing stream size estimation, downscaling and adaptive truncation algorithms, to increase cell output limits tenfold, improving UX for over **10,000** customers

X (Twitter)

ML Platform Engineer (Co-op)

May 2024 - Aug 2024

San Francisco, USA

- Designed and scaled auto-generated captions for **2 million+** videos daily, with batching and load balancing between GPUs, eliminating external API dependencies to save over **\$1 million** annually; presented directly to **Elon** 
- Built internal ASR framework to support pluggable runtimes and models for inference, with fallback heuristics and hyperparameter shuffling to reduce speech-to-text hallucination by **90%**, and decrease relative word error rate by **35%**
- Scaled low-latency vLLM inference server to handle **10+ QPS** traffic, with flash attention, quantization and KV caching

Wat.AI - Medical AI Team

Team Lead & ML Researcher (University Affiliate)

Aug 2022 - Aug 2024

Waterloo, Canada

- Fine-tuned ImageNet RCNN on FracAtlas & MURA datasets, with **85%** accuracy in fracture + appendage detection 
- Trained ResNet-34 on Correlated Diffusion MRIs from Prof Alexander Wong, to detect prostate cancer from lesions 
- Represented Waterloo AI Institute at the Canadian Undergraduate Conference in AI, raising over **\$10,000** in support

Cognite

ML Ops & Software Engineer (Co-op & Part-time)

Sep 2023 - Apr 2024

Oslo, Norway

- Customized Postgres DB to store **8 million+** vector embeddings, powering document similarity search for **10,000+** users
- Built a multi-cloud, FastAPI LLM inference & fine-tuning server with JSON streaming, cutting API costs by **\$330,000**
- Demoed Cognite GenAI tools with CTO to Microsoft Zurich and Google Nordics, sealing partnership deals of **\$5 million+**

Stratum AI

Machine Learning Engineer (Co-op)

Jan 2023 - May 2023

Toronto, Canada

- Shipped "**auto-ensembling**" Python library, eliminating manual intervention and reducing ensembling runtime by **90%**
- Researched applying derivative-free optimization techniques in **SciPy** to scale predictions in multi-model ensembles, optimize training hyperparameters and prune low-performing artifacts to increase relative precision/recall by **20%**
- Formulated novel soft scoring formulae as **loss functions** in training, improving models' absolute accuracy by **2-5%**

Smart Waterloo Innovation Lab

ML Engineering Lead (Contract)

Sep 2022 - Dec 2023

Waterloo, Canada

- Architected a **GraphRAG + Function Calling** chatbot, with ingestion pipeline to automate webscraping vectorization
- Fine-tuned T5, BERT and GPT with mock data, and built NLP eval benchmarks to boost response satisfaction by **10%**
- Engineered GraphQL APIs in Flask, and built custom CI/CD pipeline with feature testing and **100% uptime** on AWS

PROJECTS

3Brown1Blue

Python, Manin, FastAPI,

- Created **multimodal AI tutor** with live emotion detection + streaming for adaptive learning, fine-tuned with voice cloning
- Employed **Asyncio multiprocessing** for real-time speech, and visualization generation with feedback-driven code execution

Scala Compiler

Scala, MIPS Assembly

- Implemented scanner and CYK parsing, to convert string tokens into parse trees abstracting MIPS procedure code
- Wrote assembler from MIPS to binary, executing instructions in JVM, with garbage collection using **Cheney's Algorithm**