

# Jim Wu

jim.wu6@outlook.com  
jimwu.ca  
github.com/jimwu6  
linkedin.com/in/jim-wu-

## SKILLS

---

**Languages:** Python, Go, C++, SQL

**Technologies:** PyTorch, CUDA, JAX, TensorFlow, Kubernetes, Docker, Triton Inference Server, TFServing

## EDUCATION

---

**University of Waterloo**

Sep 2020 - Apr 2025

Bachelor of Computer Science - **3.9/4.0** GPA

## EXPERIENCE

---

**Tesla** [↗](#)

Palo Alto, USA

*Autopilot Intern, Machine Learning Infrastructure*

May 2023 - Aug 2023

- Researching efficient vision neural network training at scale on **GPUs/Dojo** with **PyTorch** and **CUDA**.
- Experimented with various **quantization** schemes and data formats for machine learning by training models with self-written emulators, and evaluating their effectiveness and investigating their impacts.
- Engineered **Quantization Aware Training** to enable INT8 training and export of models on Dojo.
- Reduced streaming data-loading time for Dojo by **5%** by implementing zero-copy shared buffers.

**Cohere** [↗](#)

Toronto, Canada

*Member of Technical Staff, Deep Learning Inference*

Aug 2021 - Apr 2023

- Trained **Large Language Models (LLMs)** (training efficiency and scalably serving finetunes) and improving the model training framework built on top of **JAX, TPUs, and GPUs**.
- Developed model inference/serving infrastructure to serve **multi-GPU large language models** with **billions of parameters** that generate and retrieve embeddings of text on a **low-latency API**.
- Achieved improvements of **latency by 4x** and **throughput by 8x** by engineering new model inference runtime (fork of Nvidia's FasterTransformer [↗](#)) to serve LLMs faster and more efficiently in **C++/CUDA**.
- Sped up the classify endpoint by **3x** while using **30%** less GPUs by replacing it with training classifiers on-the-fly, improving accuracy by **15%** and enabling classification in low-data settings.
- Benchmarked LMs through experiments/profiling to understand their performance characteristics.

*Software Engineering Intern - Machine Learning Infrastructure*

May 2021 - Aug 2021

- Built core backend and infrastructure components to serve LLMs, such as model deployments, autoscaling of services/models, and fractional GPUs (experimental feature) on **Kubernetes**.
- Owned development of **4+** tools of the Cohere Platform including the **Python SDK** [↗](#) (called **8M+** times in 2 months), internal CLI (used by most platform engineers daily), and benchmarking suite.
- Constructed a highly performant **Go** microservice to detect toxicity from user inputs and ML models.

## PROJECTS

---

**Depth Estimation** Trained unsupervised monocular models augmented with semantic segmentation [↗](#)

**CHANSONG** Trained **LSTMs** and **GRUs** to generate lyrics with **TensorFlow** and data from Genius API [↗](#)

## INTERESTS

---

**Academic:** ML, NLP, Medicine, Distributed Systems, PL/Compilers, Education

**Other:** Reading [↗](#), Badminton, Ultimate Frisbee, Guitar