

MAXIMILIAN MILLER

Software Engineer (ML / AI) | US Citizen | CA Work Permit Holder

☎ 382-885-3134 ✉ mtmlr101@gmail.com 🌐 maxtmiller.dev 📄 maximilianmiller 📄 maxtmiller

Education

University of Waterloo | 80% avg

Sep 2024 – Present

Bachelor of Mathematics (Co-op) | Applied Mathematics, Scientific Computing & Machine Learning

Waterloo, ON

- **Coursework:** Probability, Statistics, Linear Algebra 2, Optimization, Multivariable Calculus, Data Structures

Technical Skills

Languages: Python, C++, TypeScript, SQL, JavaScript

ML: PyTorch, Hugging Face, scikit-learn, LeRobot, MuJoCo, IsaacSim, W&B, MLflow

LLM: RAG, LLM Agents, LangSmith, Pinecone

Data & Infra: pandas, NumPy, Snowflake, dbt, PostgreSQL, Docker, Azure, Git

Experience

WAT.ai (Large Manipulation Model Team)

May 2026 – Present

ML Research Member

Waterloo, ON

- Raised sim pick-and-place success from **30%** to **99.8%** by replacing an overfit DINO ViT policy with a CNN + MLP
- Fine-tuning **pi0.5** and **SmolVLA** vision-language-action models with flow matching for task-specific manipulation
- Added a diffusion policy head and a pluggable frozen-encoder layer (DINO, Dyna) to the team's imitation-learning stack
- Engineered a custom **MuJoCo** environment and async real-time inference stack for live closed-loop robotic evaluation

Carta

May 2026 – Aug 2026

Data Intern

Waterloo, ON

- Built a **Claude**-powered data agent, shipped as an MCP tool to **118** users and validated by a **20**-question eval suite
- Designed a semantic layer of **11** Cortex Analyst views that lets the agent answer plain-English product questions
- Built a Claude skill and Airflow DAG to safely deprecate dbt models, dropping **53** models and reclaiming **549 GB**
- Incrementalized the heaviest dbt models (up to **1.69B** rows), cutting daily DAG runtime from **495** to **302** minutes

Jenly.ai

May 2025 – Aug 2025

Software Engineer Intern

Munich, DE

- Built a **RAG** chatbot for regulatory docs with source traceability and quality reports, cutting review time by **35%**
- Shipped a Word Add-in in **TypeScript** and **React** that drafts regulatory documents, cutting first-draft time by **60%**
- Automated Azure infrastructure as code with **Bicep**, cutting environment setup from hours of work to one command

ArteMed Stiftung

Sep 2023 – Present

Software Engineer Volunteer

Remote

- Developed a cross-platform patient data collection app with **Electron** and **PostgreSQL**, reducing data entry by **60%**
- Deployed the app offline for Burmese doctors to track **200+** patients daily across **16** villages on the Irrawaddy river

Projects

QaNN 📄 📄 | C++20, Python, SIMD, BLAS

- Built a C++20 vector search library with a NumPy API, using IVF, product quantization (**32x**) and exact re-ranking
- Reached **93%** recall@10 at **18K QPS** on SIFT1M, **92x** faster than brute force and **1.7x** faster than plain IVF
- Parallelized batch queries, PQ encoding and k-means training with configurable thread counts, scaling **4.9x** on 8 threads
- Wrote AVX2/NEON distance kernels and cut index training from **37s** to **2.5s** by recasting k-means as a BLAS matmul

AlphaPoisson 📄 | Python, PyTorch, ONNX Runtime, MLflow

- Trained a residual CNN chess policy-value net in **PyTorch** on 400K elite-rated games, reaching **79%** top-3 accuracy
- Built a training pipeline with gradient clipping, cosine LR scheduling and MLflow logging of loss and top-k accuracy
- Cut per-move latency from **30 ms** to **0.4 ms (76x)** with cached **ONNX Runtime** INT8 sessions and batched search

UW Grad Graph 📄 📄 | TypeScript, Next.js, LLM Agents

- Developed an **LLM** planning assistant with structured tool actions and full app-state context that edits students' plans
- Designed a degree audit engine using Hopcroft-Karp $O(E\sqrt{V})$ matching to optimally assign courses to requirements
- Modeled **400+** courses as an interactive graph with BFS-based topological layering and real-time prerequisite evaluation

Lean @ Ramp Hackathon 📄 📄 | TypeScript, Next.js, LLM Agents

- Built a SaaS spend optimizer with tool-calling **LLM agents** that audit vendor seat usage, plan tiers and SLA uptime
- Wrote a custom tool-calling agent loop, swappable across three LLM providers, that streams each step live to the UI