William Glazer-Cavanagh

Applied ML Engineer — Bridging Research & Production

Machine Learning Engineer specializing in deploying Transformer-based foundation models into enterprise environments. I focus on the engineering reality of AI: optimizing large-scale inference, validating product viability with internal customers, and building robust human-in-the-loop workflows for high-stakes data.

Contact Information

Phone: 438-389-8650 | **Email**: willcavanagh@hotmail.com | **Location**: Montreal, QC / Vancouver, BC
LinkedIn: linkedin.com/in/williamglazercavanagh | **GitHub**: github.com/williamGlazer | **Website**: williamglazer.github.io

Skills

Modeling: PyTorch | Transformers (BERT/ESM Foundation Models) | Fine-tuning | NLP | Computer Vision
Infrastructure: Docker | AWS | Modal (Serverless) | Dask | Flask | Distributed Systems
Core Stack: Python | SQL | Bash | NumPy | ML System Design | Retrieval-Augmented Architectures

Work Experience

AbCellera, Montreal/Vancouver | Machine Learning Scientist II | 05/2023 - Current

- **Delivered a sequence modeling product (Ideation to Production):** Led the full lifecycle: scoping requirements with internal customers, designing data normalization pipelines, and calibrating Transformer likelihoods. Shipped a tool that enables users to filter high-risk sequences (30% failure rate).
- **Engineered high-fidelity training datasets:** Collaborated with experts to design domain-specific normalization logic. Enforced strict evaluation rigor using **temporal splits** and **sequence identity clustering**, ensuring zero test-set leakage and realistic performance estimates.
- **Consulted on technical strategy for domain experts:** Validated product feasibility by demonstrating that proposed Deep Learning solutions would fail due to stochastic noise (Low SNR). Saved months of development time by redirecting resources to viable statistical methods.
- **Accelerated Transformer inference by 30x:** Re-engineered pipelines for BERT-style Foundation Models (ESM) to remove serial bottlenecks, leveraging **serverless fan-out (Modal)** and distributed volume caching to minimize cold starts.
- **Orchestrated cross-functional deployment:** Reduced processing latency by 50% by deploying a custom MobileNet model across **Robotics, Backend, and Science** workflows, hitting hardware speed limits to process 100,000 images/day.
- **Mitigated hallucinations via Fine-tuning:** Doubled precision (0.4 to 0.8 mAP) by fine-tuning a custom model on hard-negative examples. Built a human-in-the-loop labeling interface to capture expert feedback on the top 20 OOD cases per class.

Croesus, Montreal | Research Intern | 05/2021 - 08/2021

- **Evaluated Retrieval-Augmented Generation (RAG) feasibility:** Assessed early methods for injecting Knowledge Graphs into BERT-based architectures to define the company's roadmap for grounding NLP models in factual data.

Merck, Montreal | Data Engineer Intern | 05/2021 - 08/2021

- **Cut reporting latency by 2 days:** Automated manual ETL pipelines using AWS and Python, delivering critical operational data 48 hours faster.

Education

MSc. Machine Learning, Professional Masters | Graduation Year 2023 | University of Montreal (MILA)

BEng. Software Engineering, Artificial Intelligence Profile | Graduation Year 2022 | Polytechnique Montreal

- **Co-authored ML systems research:** Published analysis on *Change Taxonomy for ML Systems* (accepted at EMSE).
- **Benchmarked RL frameworks:** Conducted comparative performance analysis of **JAX** vs. PyTorch for Deep Reinforcement Learning. [GitHub/jax-4-deeprl]
- **Scientific Contributor (IVADO/MILA):** Served as technical reference and annotator for the *Deep Learning Essentials* course, validating scientific accuracy of CNN, GAN, and RNN content for content teams.