Injong(Brian) Won

ML Infrastructure Engineer — Cloud Computing — Model Serving & Optimization

✉ injongwbrian@gmail.com   https://injongwon.github.io/

Education

University of Toronto Toronto, ON
B.Sc. in Computer Science — Specialization in ML & Systems *Class of 2025*

- Teaching Assistant: Operating Systems, Computer Networks, Database Management Systems
- Relevant Coursework: Distributed Systems, Machine Learning, Computer Vision, Database Systems

Experience

University of Toronto Engineering Toronto, ON
Research Assistant *Apr 2025 – July 2025*

- Engineered full-stack academic contest platform serving 1,200+ students using **React, Express.js** with **auto-scaling** on **AWS ECS**
- Architected scalable database using **Prisma ORM** with **Redis** caching, reducing query latency by 70% and achieving 99.5% uptime
- Implemented **A/B testing framework** and **feature flags** for Ontario Association of Physics Teachers with real-time analytics

Vector Institute Toronto, ON
Research Assistant *Jul 2024 – Dec 2024*

- Developed hybrid operating system in **C/Rust** with custom packet parsers and **high-performance networking** protocols
- Implemented **lock-free** scheduling algorithms and **memory optimization**, reducing context switch latency by 25%
- Built **performance monitoring** tools using **eBPF** and integrated with **Prometheus** for real-time system metrics

RBC Capital Markets Montreal, QC
Quantitative Developer (Co-op) *Sep 2023 – Dec 2023*

- Architected distributed ETL pipelines using **Apache Kafka** processing 1M+ daily transactions, achieving 2x throughput improvement
- Deployed real-time monitoring dashboard maintaining 99.9% uptime for $500M+ daily trading volume
- Implemented automated alerting system reducing incident response time by 40%

IBM Markham, ON
Software Engineer (Co-op) *May 2019 – Aug 2020*

- Migrated enterprise Angular frontend (15K+ LOC) to microservices architecture, achieving 25% performance improvement
- Developed Watson Voice Agent integrations with Speech-to-Text and Natural Language Understanding APIs, improving call center automation by 40%
- Implemented distributed Watson API gateway with failover logic, achieving 99.9% uptime

Selected Projects

Multi-Head Latent Attention (MLA) Optimization 2025

- Implemented DeepSeek's MLA in **PyTorch** with **ONNX** export, reducing KV cache memory footprint by 60%
- Deployed optimized models using **TensorRT** and **CUDA** kernels, achieving 2.5x speedup on inference benchmarks
- Built **REST API** with **FastAPI** and deployed on **AWS EKS** with horizontal pod autoscaling

Technical Skills

Languages: Python, C/C++, Rust, Java, JavaScript, SQL, CUDA
ML Infrastructure: PyTorch, TensorFlow, ONNX Runtime, TensorRT, Model Serving, MLflow, Ray
Cloud & Orchestration: Kubernetes, Docker, AWS/GCP, Apache Kafka, Istio, Helm, Auto-scaling
Databases & Storage: PostgreSQL, Redis, MongoDB, Vector Databases, Feature Stores, InfluxDB
DevOps & Monitoring: Jenkins, GitHub Actions, Prometheus/Grafana, CI/CD, Performance Profiling