YILIN (LARRY) LI

☎ (416)-834-8954 ⋄ ✉ larryli1999@gmail.com ⋄ 🌐 larryli1999.github.io

TECHNICAL SKILLS

Languages	Python, C++, C#, SQL, Java, Matlab, R
Frameworks	PyTorch, TensorFlow, Hugging Face (Transformers), CrewAI, Scikit-learn
Cloud & MLOps	GCP (Vertex AI, BigQuery), AWS (Textract), Docker, Databricks, ONNX Runtime

WORK EXPERIENCES

Arteria AI May 2024 - Present
Data Scientist *Toronto, ON*

- Led the end-to-end development of production document extraction systems for **major financial institutions and consulting firms**, processing **10K+ pages daily** via **AWS Textract**.
- Evaluated **Zero-shot Vision Language Models (VLMs)** (including **Qwen** and **Idefics**) to solve complex layout analysis and table extraction tasks, enhancing pipeline capabilities.
- Fine-tuned **Transformer-based encoder architectures** for latency-critical environments; applied **post-training quantization** (ONNX) to reduce inference latency by **50%**.
- Architected **multi-agent workflows** using **CrewAI** for synthetic data generation and automated document reasoning, utilizing **Label Studio API** to integrate model-in-the-loop pre-annotation.
- Translated complex model performance metrics into actionable business insights for stakeholders, directly influencing product roadmap and client delivery timelines.

Telus Oct 2022 - May 2024
Data Scientist *Toronto, ON*

- Designed and deployed an **Offline RL agent** (Q-learning based) utilizing historical data for HVAC optimization, achieving a **20% reduction** in energy costs via safe policy iteration.
- Engineered production ML pipelines on **GCP Vertex AI**, automating training, validation, and drift monitoring using **BigQuery** for real-time system health tracking.
- Led a 4-person squad to develop an unsupervised clustering system for fiber network fault detection, reducing **Mean Time To Resolution (MTTR)** by **40%** across the network.

Huawei Canada Sep - Dec 2020
Machine Learning Engineer Intern *Montreal, QC*

- Executed **8-bit Quantization Aware Training (QAT)** on BERT encoders using PyTorch, compressing model size by **75%** while retaining **98%** of FP32 performance on GLUE benchmarks.
- Implemented **knowledge distillation** techniques to stabilize low-precision training and optimized Feed Forward Networks via structured pruning during pre-training.

RESEARCH EXPERIENCES

University of Waterloo (Data Systems Group) May - Dec 2021
Research Assistant (Advisor: Prof. Jimmy Lin) *Waterloo, ON*

- Co-developed a multi-stage retrieval system utilizing **Neural Query Synthesis (NQS)** with T5-3B to decompose complex EHRs into atomic search queries, improving recall for clinical trial matching.
- Solved the **quadratic inference bottleneck** of cross-encoding long document pairs by implementing a decoupled field scoring pipeline; system achieved **1st place at TREC 2021** (0.71 nDCG).

PROJECTS

Smart Kitchen Multi-Agent System (SKMS) Python, Google ADK, Gemini, Arize Phoenix
Lead Developer

- Architected a hierarchical agent system with routing and negative constraints to filter invalid requests; designed a custom persistence layer using **Google Sheets API** to bridge unstructured intents with structured inventory.
- Engineered a **stateful "Soft-Lock" transaction protocol** directly over the Sheets API to handle concurrency, ensuring data consistency and preventing hallucinated inventory deductions during planning.
- Built a hybrid normalization engine achieving **98% unit conversion accuracy** via deterministic fallbacks; integrated **Arize Phoenix** for LLM trace observability.

EDUCATION

University of Waterloo Sep 2017 - Apr 2022
B.A.Sc. Mechatronics Engineering, AI Option GPA: 90.6/100 (Dean's List)