Lin (Bill) Qi

+1 514-746-5515 | bill9555@hotmail.com | LinkedIn | Google Scholar

PROFESSIONAL EXPERIENCE

CGI | Montreal, QC
Data Scientist (AI Engineering) June 2024 – Present

- Led the architecture, research, and development of a **multi-agent AI assistant** for *Public Services and Procurement Canada*, achieving 95%+ response accuracy (link) and reducing manual search by hours).
- Developed an **Agentic Mixture of Experts** approach with a router agent delegating questions to specialized agents. Utilized **ReAct** and **dynamic few-shot prompting** technique to improve routing decisions.
- Implemented the AI assistant backend application using **FastAPI** for concurrent request handling and **SGLang** for batched inference using quantized **Llama 3.1** models on GPU VMs
- Prototyped an information extraction system using **Prompt Flow** and **OpenAI Agents** framework combining **Azure Document Intelligence** with a **Vision-Language Model** (VLM).
- Developed a special image chunking/scaling approach to maximize tokens used to encode images, resulting in **reduced hallucinations** and 99%+ accuracy in extracting key information from complex forms with handwritten text.
- Automated the provisioning of secure generative AI architectures using **Terraform** integrated with **Azure DevOps pipelines**.
- Developed a real-time compliance monitoring feature leveraging Azure **OpenAI**, **Snowflake** SQL and **FastAPI** for an internal generative AI platform at AT&T.

McGill University | Montreal, QC
PhD Candidate, Department of Human Genetics September 2018 – February 2024

- Leveraged domain-specific normalization techniques for preprocessing high-throughput biological data (genomic, RNA-seq, metabolomic) to remove technical and batch effects and enhance biological signals for downstream analysis.
- Performed Genome Wide Association (GWAS) analyses with multiple testing correction to identify genetic loci significantly associated with disease; and **Bayesian fine-mapping** methods to identify causal genetic variants.
- Implemented **Principal Component Analysis** (PCA), for correction of population structure confounding effects in machine learning analysis of high-dimensional genetic data.
- Experimented with **Variational Autoencoders** in Tensorflow with specialized genetic chromosome encoders to improve parameter efficiency, enhanced the identification of distinct human populations when clustering with latent representation.
- Investigated a range of machine learning techniques (e.g., **boosted trees, clustering, deep learning**) to discover novel biomarkers and predict disease risk from high-dimensional biological data.
- Created novel **Graph Representation Learning** approaches for integrating knowledge graphs and linkage-disequilibrium graphs with biological data for prediction of medication usage and cardiovascular disease, leading to two patent applications with McGill University.
- **Github code samples:** Bayesian fine-mapping of causal genetic variants (link); Gibbs sampling algorithm for learning a Latent Dirichlet Allocation topic model (link)

Ericsson Canada | Montreal, QC
Software Engineer (Machine Learning) July 2017 – September 2018

- Developed a classification model for engineer assignment for 1000+ support engineers. Backend written in Python, deployed as a **Kubernetes** microservice pod. Frontend developed using Javascript (AngularJS).
- Prototyped a question-answering system using a neural network model for **answer span classification** for information retrieved from product documentation.

PROJECT HIGHLIGHTS

Document Agent Application (https://talk-to-billy.fly.dev):

- Designed an efficient and scalable AI system for answering questions grounded on multiple documents.
- Engineered an **agentic workflow**, leveraging Pydantic for constrained LLM output to enable effective agent task delegation, prompt-chaining, tool usage, and answer synthesis.
- Implemented the core RAG system using OpenAI embedding models with **Elastic Cloud Serverless** as a scalable vectorstore, and Azure Blob Storage for storing raw documents, images, and tables.
- Developed and containerized the full-stack application using **Python, Reflex, Redis,** and **Docker**; scaled using fly.io

Graph Representation Learning for the Prediction of Medication Usage in the UK Biobank Based on Pharmacogenetic Variants (link):

- Designed and implemented a novel Graph Neural Network (GNN) in TensorFlow to overcome challenges in integrating high-dimensional genomic data with structured knowledge graphs.
- Engineered scalable data pipelines on a high-performance computing cluster to process terabyte-scale genomic data into a format ready for model ingestion.
- Validated the model's design through ablation studies, showing that the knowledge graph integration was a critical component for improving predictive performance.

Patents:

- **US-20240404700-A1:** SYSTEM AND METHOD FOR PERSONALIZED TREATMENT PRIORITIZATION (link)
- **US-20240404623-A1:** SYSTEM AND METHOD FOR PERSONALIZED INTERPRETATION OF GENETIC VARIANTS (link)

Ubiquant Market Prediction (Rank 48/2893) (link): Created an ensemble of neural networks using **TensorFlow** for stock market prediction (code).

RSNA Breast Cancer Detection (Rank 60/1687) (link): Fine-tuned pretrained computer vision models in **TensorFlow** and created optimized inference pipelines using **CuPy and NVIDIA TensorRT** (code).

A full list of my publications is available on Google Scholar

EDUCATION

- **PhD in Human Genetics (Statistical & Machine Learning focus)** | *McGill University, Montreal, QC (2018-2024)*
- **AI in Healthcare Nanodegree** | *Udacity, Online (2022-2023)*
- **Bachelor of Science, Microbiology & Immunology** | *McGill University, Montreal, QC (2013-2017)*

TECHNICAL SKILLS

- **AI Research & Modeling**: TensorFlow, Keras, PyTorch, TRL, CuPy, TensorRT, ONNX, Unsloth, LoRA, GRPO
- **RAG & Agentic AI**: LangChain, LangGraph, CrewAI, Prompt Flow, OpenAI Agents, SGLang, vLLM
- **Cloud & MLOps**: Azure, AWS, Docker, Kubernetes, Terraform, Databricks, MLflow, Wandb
- **Databases & Backend**: Python, FastAPI, Reflex, Vector Databases (Azure AI Search, Elasticsearch), SQL, Git