Charles Yuan

437-988-6789 | charlesyuan59@gmail.com | Website | Github | LinkedIn

TECHNOLOGIES AND LANGUAGES

- Python, GoLang, SQL (postgreSQL, HiveQL, Presto), noSQL (mongoDB), HTML, CSS, Javascript
- Git, Bash, Django, Flask, Spark, React, Tensorflow, Pytorch, OpenCV
- Machine Learning, Deep Learning, Computer Vision, Natural Language Processing, Quantization

EDUCATION

University of Toronto - Bachelor of Applied Science Sept 2020 – May 2025

Engineering Science - Machine Intelligence (with Engineering Business Minor)

Relevant Courses: Algorithms & Data Structures, Digital and Computer Systems, Computer Organization

EXPERIENCE

Uber | New York, New York Sept 2023 - Apr 2024

Software Engineering Intern | Delivery Matching Optimization Team

- Spearheaded unification of regression target for **XGBoost** models with **Presto** queries and data analysis, reducing MAE by 33% and RMSE by 37% while increasing data coverage by 46%
- Designed **Spark SQL ETL pipeline** to update **Hive** tables, integrating 500 million records from Kafka
- Created **automated testing suites** for **Golang** backend endpoints, achieving 96% code coverage
- Integrated **structured logging** of errors and metrics to improve observability and troubleshooting
- Leveraged Kafka client library to encode, verify, and publish waypoint predictions to **Kafka clusters**

Qualcomm | Markham, Ontario May 2023 – Aug 2023

Machine Learning Engineering Intern | Low Power/Embedded AI R&D Team

- Spearheaded development of quantization pipeline for **multimodal vision transformer**
- Leveraged Qualcomm's AI Model Efficiency Toolkit **(AIMET)**, **Pytorch**, **ONNX** quantization techniques to achieve 75% reduction in model size and **lossless per-tensor quantization**
- Utilized **automatic mixed-precision** to optimize model for **on-device inference** across multiple tasks
- Presented results to the SVP of Engineering, garnering approval for new product release

Content Turbine | Toronto, Ontario Jul 2022 – Sept 2022

Software Engineering Intern

- Created and deployed data extraction pipeline using a CRON-scheduled **web scraper** and Scrapy
- Developed **Flask REST API** with **PostgreSQL (Supabase) database,** deployed on **Digital Ocean**
- Designed and implemented billing feature with custom webhook using **Stripe API**
- Utilized scraped data to fine tune **GPT-3 NLP model** using **OpenAI**'s API for blog generation

PROJECTS

Portfolio.io | React + Django Full-Stack Web App

- Developed an web-based platform with **React** and **Django** for users to share stock portfolios
- Integrated **Auth0** for secure user authentication and **Recharts** for dynamic data visualization
- Optimized API call efficiency from **PostgreSQL (Supabase) database** w/ caching, reducing server load
- Leveraged **yfinance** for real-time financial data retrieval and **Axios** for seamless API calls to backend

RAG Financial Report Chatbot | MongoDB GenAI Hackathon Project

- Developed a **Retrieval Augmented Generation (RAG)** financial report chatbot to aid retail investors in digesting company 10Qs and 10Ks
- Utilized **LangChain** & **Nomic AI** for vector search and embedding, storing results on **MongoDB Atlas**
- Leveraged OpenAI's **GPT-3.5 Turbo** for retrieval-based QA, displaying results on Gradio frontend

AWARDS AND GRANTS

- 2022 T-CAIREM Summer Research Studentship ($6400), 2021 Dean's Innovation Fellowship ($7500)