Chieh-An (Andy) Chang

Toronto, ON | c84chang@uwaterloo.ca | linkedin.com/in/chiehanchang | github.com/ChiehAnChang
Canadian Citizen | Seeking Data Engineer / Machine Learning Engineer

EDUCATION

University of Waterloo Waterloo, ON
*Master of **Data Science and Artificial Intelligence** (Co-op)* *09/2025 - 04/2027 (Expected)*
- CGPA: **4.00**/4.00 (**91.0**/100); Recognition: **Nominated for the Vector Institute Scholarship**.
- Relevant Coursework: Data Visualization, Machine Learning, Data Engineering (Distributed Systems, Big Data, Vector Database).

University of Toronto Toronto, ON
*Honours Bachelor of Science, Double Major in **Computer Science and Statistics**, with **High Distinction*** *09/2019 - 06/2025*
- CGPA: **3.81**/4.00 (300/400-level Average: **89.8**/100); Awards: Dean's List (2021, 2022, 2024, 2025).
- **Transitioned** from a **Computer Science focus (2019 - 2023)** to a **Computer Science & Statistics Double Major (2023 - 2025)** to bridge AI engineering and data science **by integrating machine learning and probabilistic inference**.
- Relevant Coursework: Econometrics, Survey Sampling, Regression Analysis, Bayesian Statistics.

WORK EXPERIENCE

Investment Management Corporation of Ontario (IMCO) Toronto, ON
Data Engineering (Databricks, Snowflake, Microsoft Azure, Python, SQL, ETL Processing) *05/2026 - 08/2026 (Expected 4 mos)*
- To deliver data solutions for **front-office** teams by gathering **requirements from analysts** to streamline **investment workflows**. (**Expected)**
- To build **investment data solutions** on **Azure** using **Databricks** and **Snowflake** to enhance enterprise-wide data accessibility. (**Expected)**
- To **automate data workflows** using **Python** and **SQL** to improve operational efficiency for downstream financial data analytics. (**Expected)**

University of Toronto Toronto, ON
Teaching Assistant (Probability and Statistics I & II, Bayesian Statistics, Introduction to Databases) *01/2024 - 08/2025 (1 yr 8 mos)*
- Mentored **240+ students** to boost academic performance by delivering tutorials on **SQL** and **statistical inference**.
- Improved student satisfaction, achieving **90%+ positive feedback**, by demystifying research ideas via **35+ office hours**.
- Ensured fair assessment for 2,000+ students with **100% on-time delivery** via consistent, **rubric-based marking schemes**.

PROJECT EXPERIENCE

Human-in-the-Loop Email Agent via LangChain Waterloo, ON
Python, AI Safety, LangChain, LangGraph, Human-in-the-Loop (HITL), Middleware, Tool Calling, OpenAI API *02/2026 - 02/2026*
- Built a **state-aware** email agent, securing tool access by **80%+**, by implementing **LangChain** dynamic routing and prompt middleware.
- Enforced strict execution safety, preventing **100%** of unauthorized emails, by integrating **LangChain Human-in-the-Loop** middleware.
- Managed secure session contexts, tracking user authentication state seamlessly, by engineering custom **LangGraph State** schemas.

Multi-Agent Event Coordinator Waterloo, ON
Python, Multi-Agent Architecture, LangChain, LangGraph, LangSmith (LLMOps), MCP, Text-to-SQL, Tavily, OpenAI API *01/2026 - 02/2026*
- Built a **multi-agent AI** coordinator, automating **3 event workflows**, by integrating **LangChain**, **LangGraph** schemas and **LangSmith** tracing.
- Enabled dynamic real-time vendor sourcing, successfully querying **2 external** web **APIs**, by integrating **MCP** and **Tavily** Search tools.
- Automated structured DB extraction, resolving **more than 80%** of user constraints, by building a **self-correcting SQLite Text-to-SQL** agent.

YouTube Summary RAG Video Analyzer Waterloo, ON
Python, LLMs, Google Gemini, LangChain, FAISS, Streamlit *01/2026 - 02/2026*
- Built a **RAG** video analyzer, achieving high query accuracy, by engineering **Gemini Flash** LCEL chains, **FAISS**, and AI vector embeddings.
- Optimized retrieval context, measured by a **4**-document retrieval window, by chunking transcripts via **LangChain** recursive text splits.
- Deployed a responsive web interface, measured by **3** functional input sections, by programming a **Streamlit** UI for user video queries.

BC PM2.5 Short-term Forecasting Report Waterloo, ON
Python, Pandas, GeoPandas, Matplotlib, Scikit-learn, SciPy, Statsmodels, Selenium *01/2026 - 02/2026*
- Forecasted short-term PM2.5, as measured by **73.5%** accuracy, by developing a statistical learning model via **SciPy** and **Statsmodels**.
- Retrieved public data, capturing **200 million** records across **300** sources, by automating downloads via **Python** and **Selenium** scripts.
- Optimized data storage, reducing total file sizes from **600MB** to **80MB**, by executing robust **ETL pipelines** via **Pandas** and **GeoPandas**.

Customer Segmentation and Strategic Recommendation Waterloo, ON
Python, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn *12/2025 - 01/2026*
- Increased investment propensity by **10% vs. baseline** by developing targeted strategies via customer segmentation in **Python**.
- Engineered **5 distinct customer segments** using **Scikit-learn K-Means** unsupervised learning, validated via **Silhouette scores**.
- Produced actionable insights for business reporting by performing multivariate EDA on **1,000+** records using **Pandas** and **Seaborn**.

High-Dimensional Data Imputation via Group Lasso Waterloo, ON
Python, Pandas, Scikit-learn, NumPy, Matplotlib *09/2025 - 12/2025*
- Enhanced imputation quality by **9x** RMSE on **Communities & Crime data** via **Python**, **Group Lasso** & **KNN**.
- Reduced feature space by **85%** on **synthetic data** using **Pandas** & **Group Lasso** on **MNAR** to cut noise.

- Optimized **robustness** measured by **Scikit-learn** Decision Tree utility via **dimension reduction** before **imputation**.

Data Career Analytics Dashboard Waterloo, ON
PowerBI, M Language, DAX, Power Query, Excel/CSV *10/2025 - 11/2025*
- Visualized data job trends across **479K** records by building a **PowerBI** dashboard with field parameters for dynamic axis toggling.
- Streamlined data workflows across **12** monthly datasets by building automated **ETL pipelines** via **Power Query** and **M Language**.
- Quantified market demand with skill popularity percentages per position by programming dynamic aggregation measures using **DAX**.

Cancellation Policy for Ridesharing Waterloo, ON
Python, Pandas, A/B Testing, SciPy, Statsmodels, Matplotlib *09/2025 - 10/2025*
- Saved an estimated **$6M** in annual revenue, as measured by **A/B test** performance, by designing a new platform cancellation policy.
- Identified **key cancellation drivers**, as measured by **Root Cause Analysis** on a **1M+ row dataset**, using **Pandas** and **Matplotlib**.
- Validated policy impact across **3 treatment groups**, as measured by **statistical significance**, using **SciPy** and **Statsmodels**.

Predicting Falcon 9 Reusability Toronto, ON
Python, Pandas, Scikit-learn, Requests, REST API *06/2025 - 08/2025*
- Predicted rocket reusability outcomes by engineering an automated **machine learning** workflow using a **Scikit-learn pipeline**.
- Automated data ingestion and transformed raw **SpaceX API** data by performing **ETL** using **Python Requests** and **Pandas**.
- Boosted performance, improving **accuracy (86% to 98%)** and **F1 (96%)**, by applying **stratified sampling** to address **class imbalance**.

Music Generation with Deep Learning Toronto, ON
Python, PyTorch, Pandas *10/2024 - 12/2024*
- Built a generative music model, as measured by a **3.7x performance gain**, by implementing an **LSTM** model using **PyTorch**.
- Improved model training stability, as measured by significantly reduced test loss, by integrating **Batch Normalization** and **Dropout**.
- Engineered features, as measured by sequence accuracy, by encoding music notes into **129-dimensional** one-hot vectors via **Pandas**.

Modeling Equity Market Trends (Research Under Professor Jazi) Toronto, ON
R, MCMC, Bayesian inference, Yahoo Finance API *09/2024 - 12/2024*
- Forecasted JNJ prices within **95% Credible Interval** and **40-day accuracy** by engineering a **Bayesian GBM** via **HMC**.
- Quantified market risk by estimating **drift (0.1166)** and **volatility (0.1460)** via **100 stochastic simulations** on **252 days**.
- Automated data ingestion for **252 trading observations** via **Yahoo Finance API** to support Bayesian inference.

Student Learning Preference Analysis (Research Under Professor Labadi) Toronto, ON
R, Regression Analysis, ANOVA, Excel, Google Forms, stratified sampling, EDA *09/2024 - 12/2024*
- Determined LLM impact on study efficiency, as measured by **p-value > 0.05**, by conducting **Regression Analysis** in **R**.
- Secured **60 balanced samples**, as measured from **100 initial responses**, by designing a **Google Forms** survey with **stratified sampling**.
- Validated **100 observations** for integrity, measured by **data normality**, by performing **Data Cleaning** and **EDA** from **Excel** to **R**.

Student Collaboration Platform Toronto, ON
JavaScript, React, Docker, GitHub CI/CD, Node.js, Express.js, RESTful API, Neo4j, AWS EC2, Selenium, Unit Tests *05/2024 - 08/2024*
- Engineered a scalable backend API, measured by **5+** robust endpoints, by utilizing **Node.js**, **Express.js**, **RESTful API** protocols, and **Neo4j**.
- Automated testing and deployment, measured by **3 GitHub CI/CD stages**, by configuring **Selenium**, **unit tests**, **Docker**, and **AWS EC2**.
- Built a student platform, measured by **2+** core features like calendars, by coding the frontend in **React** and **JavaScript** with **5 peers**.

Ridesharing Experience Design for Campus Safety Toronto, ON
User-Centered Design, Figma, Agile Methodology, Survey Research *05/2024 - 08/2024*
- Designed **high-fidelity Figma prototypes** resolving **100% of pain points** from low-fidelity research via **agile design iterations**.
- Identified safety pain points via **research from 50+ users** by conducting **structured surveys** and **stakeholder interviews**.
- Improved system alignment by synthesizing research into **User Personas** and **Journey Maps** through **user-centered design**.

Rental and Lending Platform for GTA Housing Toronto, ON
JavaScript, Python, Microsoft SQL, React, Django, RESTful APIs *01/2023 - 04/2023*
- Optimized platform scalability, handling **3+ core listing features**, by engineering normalized **SQL** schemas and **Django** APIs.
- Delivered a rental solution, enabling property listing workflows, by co-architecting a **React** and **Django** REST architecture.
- Enhanced code reusability, deploying a modular day scheduler, by developing dynamic **React** components and **Bootstrap** layouts.

CERTIFICATIONS & COMPETITIONS

AWS Certified Cloud Practitioner | *Amazon Web Services (AWS)* Expected 06/2026
Databricks Certified Data Engineer Associate | *Databricks* Expected 05/2026
SQL (Advanced, Intermediate, Basic) | *HackerRank* *Issued 11/2025*
Python for Data Science, AI & Development | *IBM* *Issued 07/2025*
GenAI Genesis 2026 |*UTMIST & Google Developer Group* *03/2026 - 03/2026*
AI & Data Science for Good Hackathon | *Waterloo.AI* *03/2026 - 03/2026*
Statistical Modelling (Supervised Learning) | *UofT (Kaggle)* *11/2024 -12/2024*
- **Ranked 6th** out of 59 teams, as measured by an **RMSE of 0.39**, by using XGBoost, CatBoost, and Deep Learning across **100** features.