Qianyi Wang

617-359-8333 | joannawqy@gmail.com | LinkedIn | Portfolio

Education

Harvard University May 2026

M.S. Data Science – *Institute for Applied Computational Science*

Notable Coursework: Machine Learning, Advanced Practical Data Science MLOps, Quantitative Methods for Natural Language Processing, Multilevel and Longitudinal Model, Causal Inferences, Time Series

University of British Columbia May 2024

B.A. Economics and Statistics

Technical Skills: Python, SQL, R, PyTorch, Tensorflow, NumPy, Scikit-Learn, Pandas, MapR, AWS, Google Cloud, Git, Spark, Scripting, Google Analytics, BigQuery, Excel, Tableau, Power BI

Professional Experience

Data Engineer May 2025 - Aug 2025

Amazon *Seattle, WA*

- Architected and built a **Spark**-based ETL pipeline on a custom framework to process 10M+ daily requests, optimizing job partitions and join keys to manage in-memory computation and ensure cluster stability under heavy load.
- Designed a deterministic **caching architecture** to prepare for a 50x traffic increase, using **Redshift** Spectrum to analyze historical data and pinpoint the 20% of request types responsible for over 70% of the system's daily load.
- Deployed a low-latency **DynamoDB key-value store** and integrated ”cache-first” logic into the Spark ETL pipeline, bypassing expensive model inference for 70% of traffic through exact-string matching on high-frequency requests.

Data Scientist, Product Analytics May 2023 - Sept 2023

British Columbia Lottery Corporation *Vancouver, BC*

- Created **Looker Studio dashboard** through **SQL Queries** in **BigQuery**, monitored over 300,000 user sessions , identifying key trends and **conversion rate** bottlenecks that led to a 15% improvement in **user engagement**.
- Performed **hypothesis testing** on banner CTR in **Python**, analyzing 100 banner placements over 6 months; found correlations between banner position and a 10% higher CTR, informing content and design optimizations.
- Built and validated a logistic regression model to predict user bounce probability based on session behavior, identifying key friction points that informed UI/UX improvements contributing to a 10% reduction in bounce rate

Research Experience

Wildfires and Distribution of Risk for Commercial Properties | *Prof. Rachel Meltzer, Harvard GSD* Ongoing

- Employing **Event Study Analysis** to quantify wildfire risk distribution across commercial properties.
- Leveraging GEOS's **point-in-polygon algorithms** to create fire-impact buffers and **Difference-in-Differences (DID)** for **causal inference** on consumer behaviors with **Safegraph** data.

Modeling Scholarly Influence with LLM-Based Agents | *Prof. Mengyu Wang, Harvard, HMS* Ongoing

- Designing an **LLM-based agent** to approximate academic mentorship, incorporating adaptive weighting of scholarly outputs by recency and citation impact.
- Developing an **agent framework** for modeling researcher networks through co-publication overlap, enabling analysis of collaborative structures and influence patterns.

Network Meta-analyses on Energy Consumption for Climate Mitigation | *Prof. Tarun Khanna, UBC* 2024 - 25

- Implementing **Bayesian hierarchical model** for **network meta-analysis (NMA)** using **Markov Chain Monte Carlo (MCMC)** sampling via **JAGS** to synthesize insights on energy conservation and generate counterfactual effectiveness estimates for unimplemented policies, guiding evidence-based policy recommendations.

Projects

RareMind: LLM-Powered HPO Phenotyping Platform for Rare Disease Diagnostics

- Implemented an LLM-driven chatbot using GPT-4 to interactively identify HPO terms from patient symptom descriptions, enabling layperson-friendly rare-disease phenotyping.
- Integrated ClinPhen's ML-powered NLP pipeline to extract HPO codes from unstructured clinical notes for downstream medical analytics and phenotype-driven diagnostics.
- Orchestrated FastAPI backend and React frontend within Docker containers to deliver a scalable medical ontology API platform for HPO term identification.

Debiasing Pretrained Language Models with Auto-Debias

- Applied **Auto-Debias**, an **orthogonal projection-based method** to mitigate age and disability biases in **BERT** embeddings, reducing **SEAT** effect sizes from 0.51 to 0.04 (disability) and 0.51 to 0.13 (age).
- Implemented layer-wise debiasing across token- and sentence-level embeddings, preserving semantic integrity while maintaining NLP performance on **GLEU** benchmark tasks.
- Evaluated debiasing effectiveness using **cosine similarity metrics** on **News-Commentary v15 corpus** and achieved bias reduction without sacrificing model accuracy in downstream tasks.