

Joanna Wang

617-359-8333 | joannawqy@gmail.com | [LinkedIn](#) | [Portfolio](#)

EDUCATION

Harvard University

M.S. Data Science – *Institute for Applied Computational Science* **GPA: 4.0/4.0**

May 2026

University of British Columbia

B.A. Statistics and Economics - **GPA: 4.0/4.0**

May 2024

LLM & MLOps Skills: RAG, multi-agent orchestration (LangGraph/LangChain), evaluation, Python, SQL, R, Transformers, PyTorch, TensorFlow, NumPy, scikit-learn, Pandas, Spark, Airflow, Git, AWS (S3, EMR, DynamoDB, CloudWatch), Redshift, BigQuery, Google Cloud.

PROFESSIONAL EXPERIENCE

Amazon, Alexa AI

Data Scientist

May 2025 - Aug 2025

Seattle, WA

- Processed **18 TB/30-month** utterance logs to support RAG for Alexa's **Semantic Enrichment Pipeline**; split traffic into a **10k head-utterance** deterministic cache (~54% of volume and a tail index of ~10M for retrieval).
- Enabled **Personalized RAG** via identity resolution: cleaned/normalized join keys and applied deterministic earliest-timestamp selection, cutting join mismatches from **3%** to **0.08%** and resolving rare, ambiguous intents.
- Optimized **Spark** pipeline performance for tail processing: mitigated skew with **salted keys** and **AQE**, leveraging the head cache to remove ~54% of rows from heavy stages; backfilled the full **30-month** history and reduced runtime from **2.5 h** to **90 min** (~40%), eliminating OOM failures and enabling regular refreshes.

Harvard & BCH CHIP Lab, Prof. Maimuna Majumder

Machine Learning Engineer, Agentic Unified Review of Unstructured Media — Web Data Validation Cambridge, MA

June 2025 - Present

- Designed a **multi-agent** system to turn static web scraping into structured public health event data with an **LLM-validated** web extraction pipeline.
- Built a modular pipeline with clear roles—**Relevance Gate** to filter noise, **Layout Parser** to read pages/PDFs, **Fact Checker** to confirm details, **Credibility Scorer** to rate sources, and an **Arbiter** to merge updates. Kept it fast and affordable by skipping early when irrelevant, reusing results, batching requests, and capping runtime/cost.
- Validated end-to-end on a **human-labeled** set across **3 datasets** vs. a single-model baseline: **+100 pp** discovery recall, **+13.6%** extra coverage from newly found events, **12%** more new data discovery on average, and **29%** fewer false positives after review—trading off higher cost/latency for broader, cleaner coverage.

Harvard Institute for Applied Computational Science

ChroniCue: AI-driven Diagnostic Assistant for Rare and Common Diseases

Sept 2025 - Present

Cambridge, MA

- Closed the **rare-disease diagnosis** gap (often **4–5 years**) by building a **hybrid engine**: a rare-disease matcher (Phrank) runs alongside a **medical LLM** for common conditions, with a light **meta-learner** that routes cases and blends the two rankings.
- Replaced a slow RAG prototype with **ClinPhen** for fast, reliable symptom extraction: handles **negation** and **family history**, achieves **P=0.84 / R=0.78 / F1=0.80** at ~20 ms per case; the ensemble then uses simple thresholds and a **symptom-specificity score** to weight rare vs. common outputs.
- Set up clear, repeatable **evaluation**: **50** real clinical vignettes (common) plus ~1k synthetic rare-disease patients; report **Top-k accuracy** (1/3/5), **MRR**, extraction **precision/recall/F1**, and **negation accuracy** to track both diagnostic quality and information-extraction quality.

Massachusetts Institute of Technology

Debiasing Pretrained Language Models with Context-Debias

Sept 2024 - Dec 2024

Cambridge, MA

- Tackled **age and disability bias** in BERT by extending the **Context-Debias** framework to new attributes, aiming to remove bias *without* hurting language understanding.
- Implemented a **fine-tuning** strategy using a composite loss function; learn and **orthogonalize** bias directions while adding an **L2** stability term so embeddings preserve the original semantic information (no “forgetting” on GLUE).
- Achieved significant bias reduction on **SEAT9 (Disability) / SEAT 10 (Age)** benchmarks (effect size 0.51 to 0.04) while maintaining **no performance degradation** on the **GLUE** benchmark, matching the original model's 74.2% average score.

British Columbia Lottery Corporation

Data Scientist, Product Analytics

May 2023 - Sept 2023

Vancouver, BC

- Built a **Looker Studio** dashboard on **BigQuery (SQL)** aggregating **300k+** sessions into key performance metrics; the dashboard surfaced funnel bottlenecks that drove a **15%** lift in engagement (**A/B test** result).
- Ran **CTR** analyses across **100** banner placements over **6 months** of historical data; identified a **+10%** position effect to guide content ranking and design.
- Trained a **logistic regression** to predict session-level bounce; **by analyzing model coefficients**, identified high-friction features and informed **UX** changes, reducing bounce by **10%**.