Last updated in May 2025

Ian Yu

Pragmatic ML Engineer in the Age of Generative AI | 647 972 4689 | ian.yu@arc.com.co | LinkedIn | GitHub

Highlights

- Machine Learning Engineer with 5 years of experience building data enrichment pipelines and deploying production-grade ML systems in the eCommerce industry. Delivered 4 Generative AI projects to production and contributed to 2 open research initiatives on Large Language Models (LLMs).
- Proficient in production-grade Natural Language Processing (NLP), including LLMs, Retrieval-Augmented Generation (RAG), and Information Extraction & Retrieval. Experienced with Google Cloud Platform, Kubernetes, Vertex AI, and vector databases.

Relevant Experience

Machine Learning Engineer, RezolveAI (formerly Groupby Inc) – Toronto, ON — Aug 2022 – Present

- Designed and productionized Enrich AI, an agentic workflow that manages product taxonomy, data strategy, and information enrichment, cutting project timelines from weeks to hours.
- Built autotuning text clustering system with custom loss and outlier handling for efficient AI reasoning reviews, deployed via Ray and Argo on GKE.
- Revamped AI in a patent-pending automotive fitment solution using RAG, achieving >1000 QPS and maintaining <1s latency at 99th percentile.

Data Strategy Analyst, Groupby Inc – Toronto, ON — Apr 2021 – Jul 2022

- Created internal packages for client-specific data enrichment analysis, boosting client satisfaction by 50% and reducing revisions by 40%. Automated NLP/ML processes enhancing data observability and collaboration, improving team productivity by 50%.
- Introduced weakly-supervised text classification system to withhold null tasks from data annotators, saving $5,500 per month throughout its lifetime.

Data Scientist, Stratica X – Toronto, ON — Apr 2021 – Present

- Full-fledged data scientist to a boutique consulting firm for telecommunications clients
- For a utility-telecom client, built an end-to-end market analytics system from data engineering to UI to intersect fiber network objects and business directory data with PostGIS. Increased marketing team productivity by 30% and reduced sales cycle time by 30%.

Independent Consultant, Contract – Toronto, ON — Apr 2024 – Sep 2024

- For a RAG ChatBot client with high-fidelity requirements in legal fields, redesigned agentic workflows to enable faster, more modular, and more accurate output while cutting costs by 3-5 times

Open Contribution

The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset [arxiv] — 2021 – 2022

- Led a team of 10 individual contributors to minimize Personally Identifiable Information leakage. Analyzed and defined data filtering parameters to ensure corpus quality, safeness, and ethics
- NeurIPS 2022, Datasets and Benchmarks Track

SantaCoder: don't reach for the stars! [arxiv] — 2022 – 2022

- Advised on data annotation best practices and framework to facilitate PII detection effort for the Stack corpus

Technologies

Machine Learning: PyTorch, Transformers, Huggingface, Scikit-learn, SciPy, SpaCy, Snorkel, Vector Databases, Ray, Kubeflow, Argo

Geospatial: Geopandas, Shapely, turfpy, folium, PostGIS, googlemaps, geojson

Engineering: Python, JavaScript, TypeScript, SQL, NoSQL, Bash, Linux, Docker, Kubernetes, CI/CD, REST, GraphQL, OpenAPI, Locust, AWS, GCP

Skills: Data Science, Generative AI, Retrieval-Augmented Generation, Agentic Workflow, Model Deployment, Geospatial Analysis, Data Engineering