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 designing and deploying production-grade ML systems in the eCommerce industry. Deployed 5 agentic projects to production and contributed to 2 open research initiatives on LLMs.
- Proficient in production-grade NLP, agentic workflows, RAG, tool-use, MCPs, OCR, MLOps, analytics data engineering, excel at designing-prototype to implement-at-scale transition
- Experienced in consultancy, communication with various levels of stakeholders, product thinking and strategize with users, specialized in utilizing subject matter knowledge with AI

Relevant Experience

LLM Engineer, Pixomondo (Sony Subsidiary) – Toronto, ON — Nov 2025 – Present

- Assigned to a high impact innovation platform project with most artifacts built from scratch
- Built custom system discovery and registry flow for internal tools on an event-driven platform, enabling agent tool use over 20+ system services
- Designing and implementing event-driven internal tool use evaluation, modification suggestion, and continual learning

Machine Learning Engineer, Aggregate Intellect, Contract – Toronto, ON — Apr 2024 – Present

- For a RAG ChatBot client with high-fidelity requirements in legal fields, redesigned agentic workflows with 10x lower latency, 3-5x lower cost, while maintaining same output quality
- Architect and implemented agentic systems fora legal tech company, including document intelligence (Wordx and PDFs), RAG conversations, and memory.

Machine Learning Engineer, RezolveAI (formerly Groupby Inc) – Toronto, ON — Aug 2022 – Oct 2025

- Led, designed, and productionized Enrich AI, a major product with agentic workflow that manages product taxonomy, data strategy, and information enrichment, cutting full-service 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. Lowered orders of magnitude compute while increasing quality by 70%
- 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, saved $100,000 USD troughout its lifetime.

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, Arrows, Polars, SageMaker, Vertex AI, Ray, Kubeflow, MLflow, DSPy

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