Fandi Yi

7 Mabelle Ave, Toronto, ON M9A 0C9 | H: 514-550-9528 | E: fandiyi2333@gmail.com | www.linkedin.com/in/fandi-yi-/ | GitHub: Bubbletea98

PERSONAL HIGHLIGHT

I am a builder and problem solver with over **6 years' Experience** in Machine learning. Skilled in **multimodal LLMs**, **agentic frameworks**, and **RAG**, with end-to-end expertise from model design to deployment and optimization.
At **RBC**, I lead a team to develop **LLM-driven agentic chatbot** that streamline enterprise operations and projected 30% reduction in operational time for entire enterprise change management team. Outside of work, I co-authored a **MODELS 2023 paper on LLM taxonomy generation** and contribute to **Sherpa**, an open-source LLM agentic framework.

EDUCATION

McGill University | Master of Management in Analytics **2020/08 to 2021/12**

CGPA: 3.89/4.00 (Top 5 in the program), Entrance Scholarship
Related courses: Deep Learning, Database Distrib. Syst., NLP, Large Language Model, A/B Testing

McGill University | Bachelor of Electrical Engineering **2016/09 to 2020/05**

CGPA: 3.35/4.00, Final Year: 3.93/4.00
Related courses: Algorithm Design, Computer Vision, Applied Machine Learning, Numerical Methods

SKILLS

Languages & Tools: Python (8+ yrs), Java, SQL, JavaScript, MATLAB, C
Frameworks: PyTorch, TensorFlow, HuggingFace, LangGraph, LangChain, LangFuse
ML Focus: LLM Post-training, RAG, Reinforcement Learning, Model Quantization, A/B Testing
Cloud & Infra: GCP (Vertex AI, BigQuery), Azure, Docker, OCP, Kafka, Jenkins, CI/CD
Data Systems: ElasticSearch, Neo4j, MongoDB, MySQL
Specialties: Multimodal Agents, Conversational AI, Model Deployment, Performance Optimization

LANGUAGES

English (Fluent), Chinese (Native), French (Beginner)

WORK EXPERIENCE

RBC (Toronto, Canada) **Senior Machine Learning Engineer, Intelligent Ops Team** **2024/10 – current**

- Won RBC employee **performance awards twice.**
- **Lead** a **5 members** team developed and deployed the **first agentic RAG based MCP server in RBC** by leveraging **FastMCP**, **LangGraph**, **Elasticsearch** (VectorDB) and **self-hosted LangFuse** (monitoring), which consumed by 100+ employees under RBC change management team, projected **30% reduction in operational time by the end of 2026**.
- Deployed and Optimized **4 enterprise-level models.** Designed Models' **architecture** to optimize their performance on OCP: Successfully reduced alert triage model's **80% memory** consumption by refactoring script.
- Delivered a real-time incident alerting system using Java **Spring Boot,** Python **FastAPI** and **Kafka**. Reduced incident alert noise and saved estimated **~500K** CAD annually.

McGill University (Remote) **Machine Learning Research Assistant (Part-time)** **2022/12 - 2023/05**

- Worked with Prof. Emine Sarigollu's team to explore customers' interaction for circular economy (CE) topic on Twitter
- **Scraped 500k+ tweets** and built NLP pipelines for emotion, sentiment and topic analytics using fine-tuned **BERT-based classifiers** using PyTorch with CUDA accelerators.

CIBC (Toronto, Canada) **NLP & ML Engineering, Data & Analytics Team** **2022/01 – 2024/10**

- **Developed entity linking API** using scikit-learn and Spacy for daily contract analytics, which are used for matching entities in contract documents with CIBC internal suppliers' name.
- Created a **hierarchical structure for contracts' parent-child relationship**, helping identify contracts' property.
- **Enabled large-scale contract documents search** (>100k files) by building an automated document metadata extraction pipeline to clean, extract and update data from unstructured text files by using Python and SQLite.

Alibaba Group (Hangzhou, China) **ML Engineering Intern, Alibaba Brain Team** **2021/10 - 2021/12**

- Worked in the Alibaba Brain team to analyze and develop a product for **Objectives Key Results** (OKR) project management tool.
- Partnered with product managers and developers from other teams to Define the indicators measuring the **synergy effect** among different business units using the ORK data (**10 million+ records**) from the AlibabaCloud database.
- Applied **Neo4j** to build a **graph database** for OKR data to visualize collaboration among departments.

CIBC (Toronto, Canada) **Data Scientist intern, Data & Analytics Team** **2021/05 - 2021/09**

- Developed an end-to-end generative system to automate upplier profile slides generation process by querying Microsoft Access **databases** with **VBA**
- Created a customized named entity recognition **(NER) model** to extract key information from contract documents.

Allianz SE Insurance (Montreal, Canada) **Data Scientist Coop, Capstone project** **2020/11 - 2021/05**

- Built a **semi-supervised model** to predict the intentison of Canadian small business buying insurance products, and built an **LSTM model** to predict the Google Trend for insurance products to provide insights for their market team.
- Deployed **real-time ML application** for social media analytics with **Google Cloud Function and Google Data Studio**.

PROJECT/RESEARCH EXPERIENCE

Dream Journal App (GitHub) **2025/10 - Current**

- Practicing my **vibe coding** skill by publishing a dream journal application to **apple app store** from scratch
- Vibe coding frontend and backend with **Cursor** IDE, set up database and edge functions in Supabase, Build and submit app with Expo Application Services.

LLM framework – Sherpa (GitHub) **2024/05 - Current**

- Joined as one of the core contributors in Sherpa developer community.
- Built **search refinement** and **chain-of-action** tools for this agentic LLM framework.

MODELS Conference 2023 (GitHub) **2023/02 - 2023/09**

- Published on **MODELS conference**: Prompting or Fine-tuning? A Comparative Study of Large Language Models for Taxonomy Construction as **a co-first author**
- Compared different **LLMs'** prediction performance with prompting and fine-tuning methods for various taxonomy datasets.

Stock Signal Bot (Discord) **2023/01 – Present**

- Created a taxonomy dataset specific to computing classification system domain.
- Deployed real-time Discord bot by using Heroku for stock signal alerts using MACD/RSI analytics.

Advanced AI analytics for Airbnb hosts (GitHub) **2021/02 – 2021/04**

- Built an application powered by a **polynomial regression model** to help hosts to adjust their prices.
- Applied **AutoML** with ML Flow on Databricks to choose the best performance ML model and hyperparameter tuning
- Applied **Docker** to containerize ML models and the application orchestrated with **Kubernetes**.

Continuous Testing And Validation of Jamscript (GitHub) **2019/09 - 2020/05**

- Worked in Prof. M.Maheswaran's lab to **test and validate** a programming language for Edge-Oriented mobile IoT.
- Developed a continuous integration pipeline for an open-source programming language: Jamscript (a polyglot language that combines C and JavaScript) with **Travis CI**.

Face Recognition and Tagging (GitHub) **2019/09 - 2019/12**

- Developed a **face recognition system from scratch** in a team of 5 using Python Sklearn.
- Compared the face recognition performance on **PCA** and **bag-of-words** methods.

McGill Rocket Club (McGill University) **Aero-Structure and Propulsion Sub-Team Member** **2018/10 - 2020/04**

- Participated in designing different parts of the rocket model; Won **Spaceport America Cup 2018 champion**.