# Gary Qiao

Data Scientist | Machine Learning Engineer | Software Developer
365-357-2992 | qiaoy33@mcmaster.ca | LinkedIn | Personal Website

## Highlights of Qualifications

- **3 years development, design and architecture** experience of high-performance, **large-scale softwares**.
- Hands on **experience** in **LLM fine-tuning**, and **RAG** system building **from scratch** using LangChain.
- Deep expertise in **agile methodologies**, **software development life cycle**, and **problem-solving skills**.
- **Strong passion** for AI with **solid ML/DL/NLP expertise** and practical project experience.

## Education

**Master of Engineering, System and Technology (Co-op)** Jan 2025 - Expected Apr 2027
McMaster University, Hamilton, ON

- Focus on Machine Learning, Deep Learning, NLP, Applied AI and Data Systems.

**Bachelor of Engineering, Software Engineering** Aug 2017 – Jun 2021
Xi'an University of Technology, Xi'an, China

- Awards: 2019 ACM-ICPC (International Collegiate Programming Competition) - Bronze Medal

## Skills

| | |
|---|---|
| **Programming & Systems**: | Python, SQL, C++, Linux, Bash Shell, JavaScript |
| **DL & Optimization**: | PyTorch, LLMs, LangChain, RAG, LoRA(PEFT), Quantization, Prompt Engineering |
| **Data & Parallel Compute**: | Numpy, Pandas, Scikit-Learn, Chroma, ETL, PySpark, Matplotlib |
| **MLOps & Deployment**: | AWS SageMaker, Azure (DevOps, Databricks), Docker, Kubernetes, CI/CD, Git/GitHub |
| **Backend & APIs**: | FastAPI, RESTful Services, Hadoop, Spark |
| **Frontend & UX**: | React, TypeScript, HTML/CSS |
| **Math & Architecture**: | Linear Algebra, Numerical Methods, Statistics, Data Structure & Algorithms |

## Project Experience

### Llama-2 7B Script Generation

*Fine-tuned Llama-2 7B to generate multi-character TV scripts.*

- Curated 223 episodes (**1.5M tokens**), cleaned speaker-stage directions and split to 90% train and 10% validation.
- Applied **QLoRA** (LoRA and 4-bit quantization), **reduced** peak **VRAM** by **over 35%** (over 60GB to 39 GB).
- **Designed** scene-character **conditioned prompts**, **human testing** achieves **over 75% win rate** of dialogue coherence vs base Llama2-7b.
- Developed a responsive **web frontend** enabling external **access** and script generation from the fine-tuned model, forming a complete deployable **AI system**.

### RAG-based Course Knowledge QA System

*Built a RAG pipeline for searching and summarizing course materials and question answering.*

- Developed the pipeline using **LangChain**, processed and indexed PDF lecture slides with **PyPDF based parsing**, **recursive chunking** and **Chroma** for clean semantic segmentation.
- Implemented **RAG-Fusion** scoring to re-rank retrieved passages and select the top-3 chunks for context injection.
- Applied **step-back prompting** and **multi-query rewriting** to enhance retrieval accuracy.

### Ensemble Classifiers Analysis on Multiclass Prediction

*Comparison between Random Forest and XGBoost models for smartphone price-range classification.*

- Built a complete **ML pipeline** for data preprocessing, feature selection, and model tuning on a 2000×21 dataset.
- Improved Random Forest accuracy from 0.87 to 0.92 by **feature selection**, and achieved 0.94 with fine-tuned XGBoost by **early stopping** and **grid search**.
- Explained model decisions via feature importance and **SHAP** analysis, and visualized comparative performance using **confusion matrices** and **loss curves**.

## Work Experience

**Data Scientist Intern** Jan 2026 – Expected Aug 2026
Government of Ontario North York, ON, CA

- **Led** the team's **first AI-driven project** by leveraging **Python, LLM API, and prompt engineering** to perform parallelized validation between Curam database tables and Microsoft FHIR resources, **reducing over 50%** manual validation effort.
- **Designed** and **implemented** the team's **first QA automation framework**, utilizing **ThreadPoolExecutor** for multi-threaded task orchestration, applying **OOP principles** and **primary–replica architecture**, and integrating **PySpark**-based row-level data validation, **achieving 100% coverage** of table validation tasks.

**Senior iOS Software Developer** Jul 2023 – Feb 2024
Baidu (China's Google, pioneers AI, search engine, and autonomous tech.) Shenzhen, China

- Maintained application **stability** by resolving **several OOM issues** and **monitoring performance.**
- Proficient in **Requirements Analysis**, system design, feature development, showcase and test, debug and release **Agile Development** and **Cross-team Workflow**.

**iOS Software Developer** Dec 2020 – Jun 2023
Kuaishou (China's earliest and second largest short video platform, over 400 millions' DAU.) Shenzhen, China

- **Refactored** the live streaming **widgets system**, significantly **increased** the **businesses exposure** by **over 50%** for e-commerce, gaming and recruiting.
- **Refactored** the **Unit Tests** for one of the **most important live streaming frameworks** and improved line coverage from **40%** to **over 97%**.
- **Achieved patent** (ID: US20220400137) for specialized a business widget carousel component by **Design Thinking**, improving content interaction and visibility by **over 30%**.