Lirong Yao

E-Mail: ly373@cornell.edu. Tel: (607)-379-5333. Website: Yao-Lirong.github.io LinkedIn: linkedin.com/in/yao-lirong

Education

Cornell University **Ithaca, NY**

Master of Engineering in Computer Science **Aug 2025 - May 2026**

Bachelor of Arts in Computer Science, College of Arts & Sciences **Aug 2019 - May 2023**

Distinction in All subjects Magna Cum Laude - GPA: 3.944/4.300, CS GPA: 4.000/4.300
Arts & Sciences Extraordinary Senior (50 selected from 4,800+ Class of 2023 graduates)

Relevant Courses: Large Scale ML Principles & Systems, ML Hosting, Reinforcement Learning, Robot Learning

Relevant Experiences

Machine Learning Engineer, Content Understanding, Xiaomi Corp **Sept 2023– June 2025**

- Applied SOTA models across company's product to provide structured data for search & recommendation and experimented with the boundaries of newly emerged agentic AI to liberate human labor.
- Deployed NLP, CV, and multi-modal models on Triton and Docker with PyTorch models exported to ONNX or TensorRT, processing at most 1.2M+ data entries per day; validated impact by conducting 10+ A/B tests
- Mentored 2 new graduates and improved their GPT-4o based classifier by designing a multi-stage pipeline that cut costs by 83% ($100K/month savings) with only a 2% accuracy trade-off
- Actively participated in internal events such as Hackathon, where won 2nd place with an AutoGPT powered automate app testing bot for Xiaomi App Store, potentially replacing the need for 20+ manual testers
- Contributed to long-term team growth by interviewing 10+ applicants and regularly holding internal workshop on new technology, including prompt engineering before industry widely promoting ChatGPT

Lead Machine Learning Engineer, Edge AI, Xiaomi Corp **Apr 2024– Oct 2024**

- Led a 5-member team to tackle department annual key project to build on-device AI that fully utilizes NPU in flagship phone chipset to execute latency-critical tasks or process sensitive data
- Pioneered the company's first full pipeline of deploying language and vision models on mobile using TensorFlow Lite, producing models that achieve same accuracy with 2% of original size and 50% latency after distillation, quantization, and hyper-parameter search with Optuna
- Engineered an AI gaming assistant via a hybrid of RAG-enhanced LLM and on-device CNN model, custom-implementing buggy or missing operators in C++; presented as the first Internet Business Department product featured in a next-gen flagship phone launch with 150M+ online viewers

Additional Experiences

Online Multiplayer Mobile Game: Aphelion Defense, Cornell University **Jan 2023 – May 2023**

- Collaborated in an 8-person team on an online RTS C++ mobile game, focusing on WebRTC-based AdHoc server, offline enemy AI, user experience research, and solo sound design/engineering using Logic Pro

Teaching Assistant, Cornell University **Aug 2020 – May 2023**

- Proactively reached out to students facing challenges with tailored 1-on-1 assistance, presented complex materials clearly to over 100 students in exam review sessions, and mentored junior TAs across 6 semesters

Independent Researcher on Abstraction Reasoning Corpus, Cornell University **Sept 2021 – Mar 2023**

- Demonstrated limitations of both symbolic and neural program synthesis approaches on Abstraction and Reasoning Corpus - a benchmark designed to measure machine and human fluid intelligence

Skills

Language & Tools: Python, Java, C++, OCaml, Scala, SQL, Redis, Hadoop, Docker, C, Bash, Lisp, MATLAB
Machine Learning: PyTorch, TensorFlow1, TensorFlow2, Keras2, Optuna, TensorFlow Lite, MediaPipe, TensorRT, ONNX, Triton, HF Transformers, GGML & LLaMA.cpp, vLLM, LangChain, LlamaIndex, Dify