# Fengyuan Liu

Toronto, Canada | 437-340-9507 | fy.liu@mail.utoronto.ca | linkedin.com/in/fyliu/ | fy-liu.github.io/

## Education

**University of Toronto** Toronto, ON

*Bachelor of Applied Science in Engineering Science, Major in Machine Intelligence* *Sep. 2020 – Jun. 2025 (PEY)*

- cGPA: 3.79, Year-3 Annual GPA: 4.00, Dean's Honour List for all terms
- Relevant courses: Natural Language Processing, Computational Linguistics, Artificial Intelligence, Machine Learning, Matrix Algebra & Optimization, Neural Networks and Deep Learning, Computer Algorithms & Data Structure, Computer Programming, Systems Software, Probabilistic Reasoning, Decision Support Systems, Calculus.

## Experience

**Research Intern** May 2024 – Ongoing

*Vector Institute, Supervisor: Prof. Colin Raffel* *Toronto, ON*

**AttriBoT: Context Attribution for Interpreting Large Languange Model Generations**

- Developed an efficient interpretation tool to efficiently attributing an LLM's predictions to spans of text in its input context.
- Innovated a set of optimization techniques, boosting context attribution speed by up to 300×, while remaining more faithful than prior methods for efficient context attribution.
- **Fengyuan Liu**, Nikhil Kandpal, Colin Raffel. "AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution.", https://arxiv.org/abs/2411.15102
- An efficient and user-friendly implementation empowering real-world attribution at scale: github.com/r-three/AttriBoT

**Grafting Finetuned Adaptors: Modular Knowledge Transfer Across Architectures**

- Finetuned LoRA adaptors on different base models and benchmark by interchanging the adaptors.
- Design and develop algorithm to transfer LoRA adaptors without access to any dataset.

**Machine Learning Engineer** May 2023 – Apr. 2024

*Huawei Canada Toronto Research Center* *Toronto, ON*

**Super large-scale AI training and inference**

- Pioneered a precise workload and communication traffic pattern modeling for popular large language models.
- Applied PyTorch and Python to analyze and mitigate communication bottlenecks in LLM training and inference via trace-driven simulations, improving training efficiency.
- Developed a super large-scale distributed training architecture by optimizing collective communication algorithms, employing multi-dimensional parallelism strategies, and executing meticulous searches for optimal accelerator topology, to harness cutting-edge NPU clusters with over 10K accelerators from Huawei.

**High Performance Network Simulator**

- Developed a multithread high performance network simulator with Rust, significantly outperforms ns-3.

**Fair Allocation of Network Resources with Multi-Agent Reinforcement Learning**

- Proposed, implemented, and experimented a multi-agent deep reinforcement learning method to optimize resource allocation for wide area networks and data center networks, using TensorFlow and C++.
- Experimented and evaluated the proposed method, achieving 56%, 35%, and 25% reductions in flow completion time for high, medium, and low-priority flows compared to state-of-the-art methods.
- Wrote sections of design, evaluation, and related work of the paper submitted.

**Optimization on Network Utilization**

- Developed and deployed algorithms to enhance flow control, congestion control, multi-path routing, and packet loss retransmission for networks serving Huawei cloud, data centers, regionless services, and cloud edge collaboration, using C.
- Collected, analyzed, and processed massive data from Huawei's global backbone network connections. Leveraged machine learning and Bayesian optimization techniques to predict and enhance real-time communication quality, using Python.

**Student Researcher** May. 2022 – Aug. 2022

*National Research Council Canada (NRC), supervisor: Dr. Guocheng Liu* *Ottawa, ON*

**Machine Intelligence in Digital Signal Processing for Optical and Wireless Communication**

- Improved the efficiency of optical communication systems with combining end-to-end customized simulation and experimental measurement.
- Developed neural-network-based nonlinear equalizers and classifiers for QAM signals, significantly outperform conventional DSP methods (50 times lower bit error rate) in high noise and strong nonlinearity scenarios.
- Utilized pruning and quantization to optimize computational efficiency for real-time implementations.
- Delivered as a part of the 'National Challenge Program: High Throughput and Secure Networks' seminar.

## Extra Criculum

**Project Director** Sep. 2022 – Feb. 2023

*University of Toronto Machine Intelligence Student Team (UTMIST)* *Toronto, ON*

**Personalized Photography Culling System**

- Developed a personalized photography culling system to help professional photographers select images that reflect their unique style.
- Led a team of 7 machine learning developers, and collaborated with a start-up, PhotoML.
- Identified key features of images, by deploying various state-of-the-art computer vision models.
- Designed a cloud service architecture for personalized model fine-tuning and inference.

## Projects

**MealplanIQ | *LLM, RAG, LLM Agents*** Sep. 2024 - Dec. 2024

- Designed and implemented a retrieval-augmented generation (RAG) system to generate personalized recipes, integrating user preferences, existing recipe data, and nutritional guidelines.
- Engineered LLM agents to enforce complex constraints, including dietary restrictions and user-specific requirements. Code available at: https://github.com/Soniazdp/MealPlanning.

**The Change of Taste in Music | *LLM, Transformer, Creativity*** Jan. 2023 – Apr. 2023

- Advisor: Prof. Michael Guerzhoy
- Employed five models with BERT-base transformer architecture to analyze whether a classical piano composition was acclaimed or not before five time nodes: classical, early Romantic, late Romantic, World War II, and 2023.
- Pretrained and finetuned models with over 10000 collected music compositions with emsembling, reached at most 83% test accuracy.
- Investigated the change of taste in music, and the relationship between a music piece's entropy and its creativity or popularity. Code available at: https://github.com/fengyuanL/AI-music-critic.

**Retail Store Layout Optimization & Segment Marketing | *Computer Vision*** Jan. 2023

- Daisy Intelligence Hackathon 2023 First Place.
- Performed customer path tracking by deploying object detection (YOLOX) and multiple object tracking (ByteTrack) on store surveillance videos.
- Diagnosed and optimized customer paths / store layout with association rule learning to maximize profit.
- Incorporated recommendar system (Tensorrec) to cluster, segment and target custormer groups with dynamic store layout across different days.

## Skills

**Programming:** Python, C/C++, Rust, PyTorch, TensorFlow, Numpy, Matplotlib, Scikit Learn, JAX, git, Linux, Verilog (FPGA), ARM Assembly

**Softwares:** MATLAB (Deep learning / Communication / DSP system Toolboxes), AutoCAD, Fusion 360

**Fields:** Machine Learning, Deep Learning, Large Language Models, Interpretability, Quantization, Reinforcement Learning, Statistical Learning, Distributed and Hybrid Parallelized Machine Learning, Natural Language Processing, Collective Communication, Federated Learning, Computer Vision, Generative Models, Data Analysis, Independent Research, Signal Processing, Engineering Design, Drawing and Prototyping