YINING WANG

☎ 437-986-6343 ✉ yning.wang@mail.utoronto.ca

EDUCATION

University of Toronto, Toronto, ON, Canada

M.A.Sc. in Edward S. Rogers Sr. Department of Electrical and Computer Engineering *Sep 2024 – Present*

- Research Interests: Large Language Models Post-Training, Distributed Machine Learning, Datacenter Networks

B.A.Sc with Honours in Division of Engineering Science *Sep 2019 – Jun 2024*

- Focus: Machine Intelligence & Maths, Stats, and Finance

SUPPORTING ABILITIES

LLM Experience: Fine-tuning (RLVR, LoRA, SFT, etc.), RAG, Multi-Agent Collaboration, Model Evaluation, MoE
Programming Languages: Python, C++, C, Rust, Java, JavaScript, SQL, HTML, Bash, Pearl, Tcl, CSS
Machine Learning Libraries: PyTorch, Scikit-learn, TensorFlow, TVM, Keras, Spacy
Full-stack development skill: React, Node.js, PostgreSQL, SQLite

AWARDS & SCHOLARSHIP

The Edward S. Rogers Sr. Graduate Scholarship, University of Toronto *Sep 2024 - Present*
Awarded to the research-based students in ECE

School of Graduate Studies (SGS) Conference Grant, University of Toronto *Aug 2024*
Awarded to students with a paper accepted by a conference

Honours upon Graduation of B.A.Sc in Division of Engineering Science, University of Toronto *June 2024*
Awarded to students with a cumulative weighted average over 80%

Dean's Honours List, University of Toronto *2019 - 2024*
Recognized for academic achievement

Euclid Contest World's Top 25%, University of Waterloo *2018*
Scored in the top 25% of all competitors

Canadian Senior and Intermediate Mathematics Contests World's Top 25%, University of Waterloo *2017*
Scored in the top 25% of all competitors

PUBLICATIONS & PREPRINTS

1. **Yining Wang**, Jinman Zhao, Gerald Penn, and Shinan Liu
   λ-GRPO: Unifying the GRPO Frameworks with Learnable Token Preferences
   Submitting to ACL 2026

2. **Yining Wang**
   Towards Optimizing Bandwidth Allocation in Distributed Machine Learning with Flow Dependencies
   Submitting to ICDCS 2026

3. Jinman Zhao, Xueyan Zhang, and **Yining Wang**
   UORA: Uniform Orthogonal Reinitialization Adaptation in Parameter Efficient Fine-Tuning of Large Models
   ACL 2025, July 2025

4. Jinman Zhao, Yitian Ding, Chen Jia, **Yining Wang**, and Zifan Qian
   Gender Bias in Large Language Models across Multiple Languages
   Proceedings of the 5th Workshop on Trustworthy Natural Language Processing (TrustNLP 2025), March 2025

5. **Yining Wang**, Jinman Zhao, and Yuri Lawryshyn
   GPT-Signal: Generative AI for Semi-automated Feature Engineering in the Alpha Research Process
   Proceedings of Financial Technology and Natural Language Processing @ IJCAI (collected in ACL Anthology), September 2024

6. Jinman Zhao, Zifan Qian, Linbo Cao, **Yining Wang**, and Yitian Ding
   Bias and toxicity in role-play reasoning
   Preprint, September 2024

7. Savanna Blade, Zongyan Yao, Yuhan Hou, Yinfei Li, Sihan Zhou, **Yining Wang**, and Xilin Liu
   An sEMG-Controlled Prosthetic Hand Featuring a Tiny CNN-Transformer Model and Force Feedback
   IEEE Biomedical Circuits and Systems Conference (BioCAS), October 2023

PROFESSIONAL EXPERIENCES

University of Toronto *Sept 2024 - Present*
Graduate Research Assistant Toronto, ON, Canada

- Developed λ-GRPO, a learnable token-weighting frame based on GRPO that lets models learn their own token preferences and boosts Qwen2.5 1.5B/3B/7B by +1.9/+1.0/+1.7 points on math reasoning with no extra data or compute
- Achieved equivalent performance on MT-Bench with LLaMA 2 using quantization-based PEFT, while training with only 1% of LoRA's trainable parameters on the Alpaca dataset
- Developed a series of optimization algorithms for bandwidth allocation with flow dependencies in inter-datacenter networks, improving resource efficiency and network performance
- Optimized the information loading of flow simulation, making it adaptable to any optimization methodology and accelerating implementation speed by 15× compared to the original setting
- Developed and implemented a distributed training framework across multiple GPUs and servers in both intra-datacenter and inter-datacenter environments
- Integrated dynamic bandwidth allocation into NCCL's Ring AllReduce (RAR) algorithm in PyTorch by analyzing and optimizing the interplay between algorithmic bandwidth and bus bandwidth, enhancing overall communication efficiency during model training
- Researched on designing of New Collective Communication Libraries (CCLs) to optimize communication and maximize link utilization in Multi-GPU Distributed Systems
- Developed expertise in RAG, enhancing content relevance and contextual depth within a multi-agent pipeline

Canada Pension Plan (CPPIB) and University of Toronto *Sept 2023 - May 2024*
Thesis Research Assistant Toronto, ON, Canada

- Conducted research on leveraging LLM to automate feature generation for financial datasets (alpha research)
- Acquired and processed historical financial ratios and return data for multiple stocks from public sources, including Yahoo Finance and FactSet. Ensured data cleanliness and transformation before input into LLMs
- Built Langchain applications around Gpt-3/3.5 models, implementing various prompting techniques such as In-Context Learning, Chain-of-Thought, and Step-by-Step explanations
- Designed evaluation methods and assessed the performance of LLM-generated features, demonstrating their capability for alpha mining

University of Toronto *Mar 2022 - Sep 2023*
Undergraduate Research Assistant Toronto, ON, Canada

- Researched on FoG (Freezing of Gait) detection of Parkinson's disease using machine learning models (CNN and LSTM)
- Trained models and did parallel testings with single modal sensor data of the patients and achieved an accuracy of 85.6%
- Processed, filtered, and segmented data for multimodal sensor data using packages such as numpy and pandas
- Fit the ACC and EEG segmented multimodal data to different models (CNN2D, CNN3D etc.)
- Researched on sEMG-Controlled Prosthetic Hand Featuring a Tiny CNN-Transformer Model for classifying 21 gestures from sEMG signals, achieving a high average accuracy of 81.39%

Alphawave Semi *May 2022 - Aug 2023*
Software Engineer Co-Op Toronto, ON, Canada

- Wrote scripts in Python/SQL/Bash/Pearl to monitor and adjust the server usages, and license utilization, create databases, pull data from the database, and generate graphs of login/compute/remote servers in different aspects
- Created bind key logic functions in TCL to simplify the process of designing for Samsung and TSMC technologies, improving the efficiency of designing chips in Custom Compiler for the layout team
- Integrated the Atlassian Confluence and Jira platforms with the company's ticketing system by using Python API
- Wrote scripts in Python/HTML to create, modify and upload content to the company's confluence (wiki) page
- Modified, debugged for errors, and improved the efficiency of existing scripts by approximately 50% of the original runtime

TEACHING EXPERIENCES

University of Toronto Toronto, ON, Canada
Teaching Assistant for CSC209: Software Tools and System Programming (in C) *Winter 2025*

- Provided support to students with their lecture worksheets, labs, and assignments
- Wrote Automaker scripts and graded assignments on coding correctness
- Marked midterm and final exams
- Collaborated with course instructors to refine course materials and improve the overall learning experience for students

Teaching Assistant for ECE1778: Creative Applications for Mobile Devices *Fall 2025*

- Provided support to students with their assignments
- Wrote Autograder scripts and graded assignments on coding correctness

Teaching Assistant for ECE1779: Introduction to Cloud Computing *Winter 2026*

SERVICE

Reviewer
ACL (2025), MMSys (2025), IWQoS (2025), IEEE INFOCOM (2025, 2026)