




SHOU-YI (RAY) HUNG

 <https://www.shouyihung.com>

 437-349-9099

 syhung0927@gmail.com

 <https://www.linkedin.com/in/shouyihung/>

 [Lei-Tin](#)

Education

University of Toronto St. George | GPA 4.0/4.0

Sept 2021 - June 2026

Honours Bachelor of Science in Computer Science, Statistics Minor

Toronto, Ontario, Canada

- Arts & Science Internship Program (Co-op)
- Relevant Coursework: Data Structures and Analysis, Algorithm Design and Complexity, Intro to Relational Database, Web Programming, Intro to Machine Learning, Software Design, Systems Programming, Operating Systems, Computer Networking System, Deep Learning, Probabilistic Learning, Linear Programming, Computer Vision with Deep Learning
- **Awards:** Woodsworth College Scholarship, Dean's List, University of Toronto Excellence Award (UTEA)
- **Teaching Assistant:** Lead TA CSC369H1 (Operating Systems), TA CSC369H1
- **Relevant Coursework:** Data Structures and Analysis, Relational Database, Web Programming, Computer Networking System, Operating Systems, Machine Learning, Deep Learning, Linear Programming, Computer Vision

Publications

- Hannah Liu, Ethan Yue Heng Cheung, **Shou-Yi Hung**, et al. 2025. Datasheets Aren't Enough: DataRubrics for Automated Quality Metrics and Accountability. In 2nd Conference on Language Modeling (COLM WMDQS Workshop 2025, Poster)
- Genta Indra Winata*, David Anugraha*, Emmy Liu*, Alham Fikri Aji*, **Shou-Yi Hung**, et al. 2025. Datasheets Aren't Enough: DataRubrics for Automated Quality Metrics and Accountability. In Advances in Neural Information Processing Systems (under review for NeurIPS 2025)
- Syed Mekael Wasti*, **Shou-Yi Hung***, En-Shiun Annie Lee. 2025. TranslationCorrect: A Unified Framework for Machine Translation Post-Editing with Predictive Error Assistance. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations) (ACL Demo 2025)
- Yun-Hsin Chu, Shuai Zhu, **Shou-Yi Hung**, Bo-Ting Lin, En-Shiun Annie Lee, and Richard Tzong-Han Tsai. 2025. ATAIGI: An AI-Powered Multimodal Learning App Leveraging Generative Models for Low-Resource Taiwanese Hokkien. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations) (NAACL Demo 2025)
- En-Shiun Annie Lee, Luki Danukarjanto, Sadia Sharmin, **Shou-Yi (Ray) Hung**, Sicong Huang, and Tong Su. 2024. Exploring Student Motivation in Integration of Soft Skills Training within Three Levels of Computer Science Programs. In Proceedings of The Technical Symposium on Computer Science Education (SIGCSE 2024). **Paper presenter.**

★ Equal contribution


Experience

Software Development Engineer Intern

May 2025 - Present

Amazon Web Services (AWS)

Vancouver, BC, Canada

- Led the development of a CloudFormation (CFN) compatible package on resource provisioning that enables more than **500+** EventBridge (EB) partners and customers to provision EB resources with CFN, deployed to more than **35** AWS Regions 
- Developed the function to provision AWS resources through Cloud Development Kits (CDKs), resulting in enhanced tool interoperability and enabling more than **1000+** active developers using AWS CDKs to provision EB resources
- Implemented an advanced throttling management system for the EB backend to prevent the noisy neighbors problem when scheduling invocations
- Enabled support for EB Archive resources with Customer Managed Key Management Service on CDK constructs

Machine Learning Research Assistant

May 2025 - Present

University of Toronto, supervised by Prof. Maryam Mehri Dehnavi

Toronto, ON, Canada

- Researched on methods to achieve 2:4 model sparsity with quantization by applying **Sprase Marlin** kernels with **VLLM** for increased efficiency in LLM inference
- Researched on methods with blockwise training for quantization with sparsity with MaskLLM to improve downstream task performance

- Researched on pretrained model distillation by optimizing transformer blocks in Large Language Models (LLMs) to train a compressed model with fewer training tokens and retaining the model's capabilities

Machine Learning Research Assistant

University of Toronto, supervised by Prof. En-Shiun Annie Lee

May 2023 - Present

Toronto, ON, Canada

- Researched on multilingual translation models and LLMs applications in the real world by building language learning and educational frameworks
- Managed a team of Research Assistants, organized regular touch-point meetings to ensure project delivery, facilitated collaboration with other research teams
- Built a low-resource language learning framework for **100+** users by developing an interactive UI using **React Native** and **NoSQL (Google Firebase)**
- Deployed training pipelines with **NVIDIA A100 GPU** for numerous machine translation models such as NLLB and xComet, conducted statistical analyses and visualization to evaluate performance

Machine Learning Researcher Intern

Huawei Canada

May 2024 - Apr 2025

Markham, ON, Canada

- Finetuned LLMs with various methods like Low-Rank Adaptation (**LoRA**) to adapt to downstream tasks such as input classification and text generation tasks, increasing classification accuracy from **76% to 85%**
- Finetuned a 34M Llama-like model with knowledge distillation to work with Sequoia Speculative Decoding on a 1.5B model, increased the acceptance rate from **54% to 65%**, released on HarmonyOS 5.1
- Deployed and launched multi-node, multi-card distributed training to accelerate the training process using PyTorch's **Distributed Data Parallel** and HuggingFace's **Accelerate** library, increasing training efficiency **up to 4x**
- Enhanced the inference performance of multiple Llama-like LLMs on edge devices with specific tasks such as dialog summary through knowledge distillation and model quantization with **Python (PyTorch, Transformers)**
- Developed **Bash** scripting with **Docker** to systematically train and evaluate LLMs with **NVIDIA V100 GPU**

Software Developer

University of Toronto, supervised by Prof. Kuei (Jack) Sun

Jan 2024 - Dec 2024

Toronto, ON, Canada

- Collaborated with **10+** developers to create "KidneyOS", a prototype operating system written in **Rust** for teaching operating system concepts such as memory allocation and file systems
- Curated tutorial materials for "Buffer Overflow" attack, used by **200+** students taking the operating systems course
- Written in **C**, demonstrated how shell codes can be ran when an unsafe function is used
- Prepared course materials on **HTML/CSS** and **React**, used by **250+** students taking the web programming course

AI Arena Reinforcement Learning Competition | 4th place

Tencent

Jan 2023 - Apr 2023

ChengDu, China

- Trained an off-policy actor-critic **deep reinforcement learning model**
- Deployed CNN, LSTM, and Multi-Head attention to enhance model performance
- Introduced Dropout and Lookahead Optimizer, and applied other techniques to increase model stability
- Achieved model performance within the **top 15th** percentile among human players

Skills

Programming Languages: Python, Java, C/C++, R, \LaTeX , HTML/CSS, JavaScript, SQL, Ruby, Rust, NoSQL

Technologies and Frameworks: Unix/Linux, REST API, React, Flask, Pandas, Matplotlib, Selenium, OpenCV, Slurm

Machine Learning Frameworks: PyTorch, NumPy, HuggingFace, Distributed Training (Accelerate, FSDP, DDP), DeepSpeed, Llamafactory

Developer Tools: Jupyter Notebook, Conda, Shell (sh, bash, zsh), Git, GitHub, GitHub Actions, CI/CD, Docker, AWS, Microsoft Azure

Volunteer Work

Conference Student Volunteer | ACM SIGCSE 2024

March 2024