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)
- **Awards**: Woodsworth College Scholarship, Dean's List, University of Toronto Excellence Award (UTEA)
- **Teaching Assistant**: 2x CSC369H1 (Operating Systems)
- **Relevant Coursework**: Data Structures and Analysis, Relational Database, Web Programming, Computer Networking System, Operating Systems, Machine Learning, Deep Learning, Linear Programming, Computer Vision

Publications

- 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 | EventBridge | Java *Vancouver, BC, Canada*

- Implemented EventBridge (EB) resource provisioning by providing a CloudFormation (CFN) compatible package, enabling more than **500+** EB partners to provision EB resources with CFN
- Extended support for EB resources to be provisioned through Cloud Development Kits (CDKs) to enable more than **1000+** active developers using AWS to provision EB resources
- Implemented unit tests, integration tests, and contract tests to ensure that the deployed service will be error-free

Machine Learning Research Assistant **May 2025 - Present**
University of Toronto, supervised by Prof. Maryam Mehri Dehnavi *Toronto, ON, Canada*

- Explored possibilities of model distillation in a block-wise manner among transformer based models (Llama and Qwen)
- Retained **98%** of the model's pretrained capabilities by model distillation by using only **10%** of the training tokens involved in the pertaining phase

Machine Learning Research Assistant **May 2023 - Present**
University of Toronto, supervised by Prof. En-Shiun Annie Lee *Toronto, ON, Canada*

- Researched on Multilingual translation models and LLMs applications in the real world
- Collaborated in building an interactive UI using **React Native** and **NoSQL (Google Firebase)**, enabling **100+** users to access a low-resource language learning framework
- Conducted statistical analyses and built data visualizations using **Python (matplotlib, seaborn, pandas)**
- Deployed training with **NVIDIA A100 GPU** for numerous machine translation models such as NLLB and xComet

- Facilitated and coordinated data collection pipelines for low resource languages used to train translation models and error detection models
- Co-authored two papers published at top NLP conferences

Machine Learning Researcher Intern **May 2024 - Apr 2025**
Huawei Canada *Markham, ON, Canada*

- Fine-tuned LLMs with various methods like Low-Rank Adaptation (**LoRA**) to adapt to downstream tasks like input classification and generation tasks, increasing classification accuracy from **76% to 85%**
- Finetuned a 34M Llama 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**, increasing training efficiency **up to 4x**
- Enhanced the inference performance of multiple Llama-based LLMs on edge devices with specific tasks like dialog summary through knowledge distillation and model quantization with **Python (PyTorch, Transformers)**
- Utilized **Bash** scripting with **Docker** to systematically train and evaluate LLMs with **NVIDIA V100 GPU**

Software Developer **Jan 2024 - Dec 2024**
University of Toronto, supervised by Prof. Kuei (Jack) Sun *Toronto, ON, Canada*

- Collaborated with **10+** developers to create "KidneyOS", a prototype operating system written in **Rust** for teaching Operating Systems concepts such as memory allocation and file systems
- Created 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 **Jan 2023 - Apr 2023**
Tencent *ChengDu, China*

- Implemented an off-policy actor-critic **deep reinforcement learning model**
- Deployed CNN, LSTM, and Multi-Head attention to enhance performance of model
- Introduced Dropout, NoisyNet, 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, 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)
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**