Willis Guo

willisg@cs.cmu.edu • (878) 834-9154 • willisguo14.github.io • linkedin.com/in/willisguo • github.com/willisguo14

Education

MS Machine Learning — *Dec. 2025*
Carnegie Mellon University (CMU) — *Pittsburgh, PA*

BASc Machine Intelligence — *Apr. 2024*
University of Toronto — *Toronto, CAN*

Experience

Research Intern, Multimodal — *Sep. 2024 – Present*
Carnegie Mellon University, Ruslan Salakhutdinov — *Pittsburgh, PA*

- **Post-training vision-language models (VLMs)** by **fine-tuning** and aligning with human preferences via **RLHF** to build **multimodal dialogue agents**.
- Developed an inference algorithm for video understanding with **VLMs** that reduces inference costs 5x by leveraging **video diffusion models** as a **world model**.

Software Engineer Intern — *June 2024 – Aug. 2024*
Amazon Web Services — *Vancouver, CAN*

- Designed and implemented infrastructure and tools for analyzing PBs of AWS resource traffic data. Performed **hyperparameter tuning** and **feature engineering**, increasing recall of detected distributed denial-of-service attacks by 3%.

Research Intern, Large-Language Models (LLMs) — *Sep. 2023 – Apr. 2024*
University of Toronto, Scott Sanner — *Toronto, CAN*

- Designed a neuro-symbolic, **inference-time search** algorithm for **logical reasoning with LLMs** that improves LLM commonsense reasoning accuracy by 13%.
- Designed an **LLM agent** for **knowledge graph question answering** (KGQA) with planning and active **retrieval augmentation (RAG)**, reducing hallucinations by 79% and outperforming existing KGQA methods by 8%.

Machine Learning Engineer Intern — *Sep. 2021 – Apr. 2022*
aUToronto (University of Toronto Self-Driving) — *Toronto, CAN*

- Led the development of a **machine learning pipeline** for traffic light detection and classification: collected a dataset with 10,000 training examples and **finetuned** 50M parameter CNNs, achieving 89% accuracy.

Research Intern, Machine Learning — *May 2021 – Aug. 2021*
University of Toronto, Shurui Zhou — *Toronto, CAN*

- Implemented and trained **CNNs** and **attention-based RNNs** for code vulnerability detection, improving recall by 5%.
- Created a data pipeline for scraping code vulnerabilities from open-source projects, creating a dataset with >5,000 examples.

Publications

Active Perception for Efficient Inference-Time Long-Form Video Understanding in Vision-Language Models
Martin Ma, **Willis Guo**, Aditya Agrawal, Ankit Gupta, Paul Liang, Russ Salakhutdinov, Louis-Philippe Morency. *In submission*

Verifiable, Debuggable, and Repairable Commonsense Logical Reasoning via LLM-based Theory Resolution
Armin Toroghi, **Willis Guo**, Ali Pesaranghader, Scott Sanner. *EMNLP 2024*

Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering
Armin Toroghi, **Willis Guo**, Mohammad Mahdi Abdollah Pour, Scott Sanner. *EMNLP 2024*

Projects

Deep Learning Library

- Built a deep learning library from scratch with **GPU-accelerated operations**, ND array, automatic differentiation, optimizers, etc.
- Implemented and trained **transformers, CNNs** and **RNNs** using the custom-built deep learning library.

Transformer to State Space Model Distillation

- **Distilled vision transformers (ViT)** to **state space models (SSMs)**, retaining 80% of the ViT's performance on image classification.

Skills

Languages	Python, Java, Scala, C, C++, SQL, JavaScript, TypeScript, MATLAB
Machine Learning	PyTorch, Hugging Face, CUDA, TensorFlow, Apache Spark
Software Engineering	AWS, PostgreSQL, Docker, React