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 — *Pittsburgh, PA*

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

Experience

ML Research Engineer Intern, ML Systems — *May 2025 – Present*
Scale AI — *San Francisco, CA*

- Designed and implemented a **novel distributed training parallelism** strategy for **online reinforcement learning (RL)** that increases training throughput by 3x. Preparing MLSys submission.
- Built from scratch a new and hackable **RL post-training library** with **GPU co-location** for training using **FSDP** and inference using **vLLM**, and supports **training multimodal models**. Planning to open source.
- Researching **test-time scaling laws** for online RL. Training 7B to 32B models using GRPO on math and video reasoning.
- Implemented a fused **Triton kernel** for calculating post-training losses, decreasing memory usage 3x and enabling **long-context training** up to 96k tokens per GPU.

Research Intern, Multimodal — *Sep. 2024 – May 2025*
Carnegie Mellon University, Ruslan Salakhutdinov — *Pittsburgh, PA*

- Created synthetic data with interleaved **multimodal chain-of-thought reasoning** traces with **visual tool-use** for supervised fine-tuning (SFT) vision-language models (VLMs) for video reasoning.
- 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*

- Built **ML infrastructure**, including data pipelines and observability tools for analyzing petabytes of AWS resource traffic data. Performed **hyperparameter tuning** and **feature engineering**, improving anomaly detection recall by 2%.

Research Intern, LLM Reasoning — *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%.

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. *ICCV 2025 Workshop*

CoLoTa: A Dataset for Entity-based Commonsense Reasoning over Long-Tail Knowledge
Armin Toroghi, **Willis Guo**, Scott Sanner. *SIGIR 2025*

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

Mini-LLMSys

- Implemented important techniques for training LLMs on top of MiniTorch: **fused CUDA kernels** and **distributed training** methods including **data parallel** and **pipeline parallelism**.

Skills

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