Shirui (Carl) Chen

Linkedin Google Scholar sc256@uw.edu +1 206-822-8936 GitHub

Ph.D. candidate from the University of Washington (UW) with 5+ years of experience in deep learning, generative modeling, and large-scale AI systems. Early research examined how intelligent systems represent uncertainty and generalize; current work applies these ideas to diffusion language models, vision-language models, and robotics. Skilled in PyTorch, large-scale training, and mathematical modeling, focusing on building scalable and robust AI systems.

Education

University of Washington *2021 – June 2026 (expected)*
Ph.D. Candidate, Applied Mathematics GPA: 4.0
Awards: Boeing Research Award (2024)

University of Wisconsin–Madison *2017 – 2021*
B.Sc., Mathematics and Computer Science GPA: 4.0
Awards: Dean's Prize Nominee (top 0.1%); International Collegiate Programming Contest (ICPC) World Finalist; Mathematical Contest in Modeling (MCM) Meritorious Winner (top 10%)

Industry Experience

Ai2 Seattle, WA
Student Researcher Intern Nov 2025 – Now

- Curated and standardized **1,500+** heterogeneous LeRobot robotics datasets into a unified corpus that consists of **19M** frames and **40k** episodes.
- Auto-annotated robot instructions with a vision-language model (Qwen-2.5VL). Pretraining $\pi 0$ model on this annotated dataset reduced pretraining loss by **30%** compared to training on the unannotated dataset.
- Proposed cold-start pretraining with small high-quality data: lowered pretraining loss by **50%** and improved pick-up success rate by **67%** vs. training from scratch.

Meta New York, NY
Research Scientist Intern Jun 2024 – Dec 2024

- Scaled neural interface foundation model to **1B+** parameters (4× increase), improving handwriting success rate by **39%**.
- Learned discretized EMG representations via Gumbel–Softmax; eliminated null-behavior false positives.
- Implemented global–local attention masking with Torch FlexAttention to expand context window, improving throughput in long-sequence modeling by **40×**.

Selected Research Projects

dUltra: Ultra-Fast Diffusion Language Models via Reinforcement Learning

- **Outperformed all state-of-the-art baselines:** achieved **2.3–4× faster inference** than Fast-dLLM across mathematical reasoning and code generation benchmarks while simultaneously improving accuracy
- Designed novel reinforcement learning framework using GRPO to learn optimal token unmasking strategies, demonstrating learned policies significantly outperform fixed heuristics and offline distillation methods
- Built complete training pipeline with multi-component reward system and distributed training infrastructure using PyTorch and HuggingFace Transformers; evaluated across multiple block sizes and inference strategies
- Achieved consistent improvements across all benchmarks (GSM8K, MATH500, HumanEval, MBPP); authored research paper advancing diffusion language models toward competitive performance with autoregressive models

Deep Neural Network Generalization & Robustness

- Established theoretical links between sharpness (loss Hessian trace) and robustness of neural representations, proving new bounds on volume compression and sensitivity.
- Validated theory through experiments on VGG-11, MLP, and ViT architectures, demonstrating that sharpness

directly affects generalization and adversarial robustness.
- Highlighted mathematical principles to guide design of robust and generalizable deep learning models.

Selected Publications

1. **Shirui Chen**, Jiantao Jiao, Lillian J. Ratliff, Banghua Zhu. *dUltra: Ultra-Fast Diffusion Language Models via Reinforcement Learning.*
2. **TMLR** 2025 – **Shirui Chen**, Stefano Recanatesi, Eric Shea-Brown. *A simple connection from loss flatness to compressed representations in neural networks.*
3. **NeurIPS** 2023 – **Shirui Chen**, Linxing Preston Jiang, Rajesh P. N. Rao, Eric Shea-Brown. *Expressive probabilistic sampling in recurrent neural networks.*
4. **COLM** LM4Sci Workshop 2025 – Linxing Preston Jiang, **Shirui Chen**, Emmanuel Tanumihardja, Xiaochuang Han, Weijia Shi, Eric Shea-Brown, Rajesh P. N. Rao. *Data Heterogeneity Limits the Scaling Effect of Pretraining Neural Data Transformers.*
5. **iScience** 2023 – **Shirui Chen**, Qixin Yang, Sukbin Lim. *Efficient inference of synaptic plasticity rule with Gaussian process regression.*

Technical Skills

- **Programming & Tools:** Python, C++, Go, MATLAB, NumPy, Pandas, Linux, Git
- **Frameworks & Libraries:** PyTorch, Accelerate, TRL, Torch FlexAttention
- **Expertise Areas:** Deep Learning, Generative Models, Reinforcement Learning, Computational Neuroscience, Bayesian Inference & Statistics