Allan Zhang

Last updated in October 2025

Los Angeles, CA allanzhang440@gmail.com
609-943-8429 https://giyushino.github.io giyushino

Education

University of California, Los Angeles *BS in Computational Mathematics* *Sept 2024 – June 2028*

- **Relevant Coursework:** Multivariable Calculus, Linear Algebra, Differential Equations, Discrete Structures, Probability Theory, Data Structures and Algorithms, Large Scale Machine Learning (graduate coursework)

Experience

Undergraduate Research Assistant | BigML @UCLA *Nov 2024 – Present*

- Conducting research on **data-efficient** training for LLMs, focusing on data selection for reinforcement learning and synthetic data generation. Advised by Prof. Baharan Mirzasoleiman
- Co-designed a data-efficient variant of GRPO, achieving 30% faster convergence and a 40% reduction in required training samples relative to vanilla GRPO
- Managed multi-GPU research servers equipped with NVIDIA A6000 and A40 GPUs; reduced CPU overhead by writing Bash scripts to automatically detect and kill orphaned processes
- Participate in weekly reading groups to stay up-to-date with the latest ML/LLM publications

Projects

Thinking Constrained GRPO *grpo-thinking-budget*

- Implemented constrained reasoning for LLMs by enforcing thinking token budgets during GRPO training, enabling fine-grained control over model inference costs
- Designed a two-phase generation system with vLLM that separates thinking and response phases, with automatic delimiter injection when thinking budget is exhausted
- Created fast inference pipeline using vLLM with two-phase generation to control thinking vs. response token allocation for evaluations

Neural Network Evolution *NeuroEvolution*

- Designed and implemented a project to explore the effectiveness of evolutionary principles (sexual reproduction, mutation, and natural selection) in producing accurate CNNs and ViTs
- Developed a flexible Python framework to apply custom transformations to any PyTorch or Jax neural networks, including layer-wise merging and controlled injection of randomness into model weights
- Compared neural networks trained via gradient descent and evolution-based optimization, found that both approaches converged to models with highly similar weight configurations

Compositional Generalization *Ongoing*

- Studied different techniques to improve LLM performance on compositional tasks, tested performance of prefix scorers and hidden Markov models for controlled language generation
- Implemented Google's paper *Controlled Decoding from Language Models* from scratch in PyTorch, collected synthetic data to train prefix scorer.
- Engineered an automated pipeline for deploying custom merged LLMs with SGLang and vLLM, accelerating benchmarking workflows

Skills

Programming Languages and Frameworks: Python, PyTorch, JAX, NumPy, Matplotlib, TensorFlow, scikit-learn, OpenCV, Hugging Face, LaTeX, C++, Bash, vLLM, SGLang

Languages: English, Korean