

Muru Zhang

Los Angeles, CA (206) 245-7454
muruzhan@usc.edu - <https://nanami18.github.io/>

Education

University of Southern California

Ph.D. in Computer Science

Advisors: Swabha Swayamdipta, Robin Jia

Los Angeles, CA

August 2024 – Present

University of Washington

M.S. in Computer Science

Advisor: Noah A. Smith

Seattle, WA

September 2022 – June 2024

University of Washington

B.S. in Computer Science, B.A. in Mathematics

Honor and Award: James Hewitt Endowed Scholarship

Seattle, WA

September 2020 – June 2022

Publication

LLM Unlearning Without an Expert Curated Dataset

Xiaoyuan Zhu, [Muru Zhang](#), Ollie Liu, Robin Jia, Willie Neiswanger

Conference on Language Modeling (COLM) 2025, URL: <https://arxiv.org/abs/2508.06595>

Why Do Some Inputs Break Low-Bit LLM Quantization?

Ting-Yun Chang, [Muru Zhang](#), Jesse Thomason, Robin Jia. (2025)

Empirical Methods in Natural Language Processing (EMNLP) 2025 Main, URL: <https://arxiv.org/abs/2506.12044>

Large-Scale Data Selection for Instruction Tuning

Hamish Ivison, [Muru Zhang](#), Faeze Brahman, Pang Wei Koh, Pradeep Dasigi. (2025)

URL: <https://arxiv.org/abs/2503.01807>

Ladder Residual: Redefining Tensor Parallelism in Transformers for Accelerated Inference

[Muru Zhang*](#), Mayank Mishra*, Zhongzhu Zhou, William Brandon, Jue Wang, Yoon Kim, Jonathan Ragan-Kelley, Shuaiwen Leon Song, Ben Athiwaratkun, Tri dao. (2024)

International Conference on Machine Learning (ICML) 2025, URL: <https://arxiv.org/pdf/2501.06589>

Toward a More Complete OMR Solution

Guang Yang, [Muru Zhang](#), Lin Qiu, Yanming Wan, Noah A. Smith. (2024)

International Society for Music Information Retrieval (ISMIR) 2024, URL: <https://arxiv.org/abs/2409.00316>

Learning to Build by Building Your Own Instructions

Aaron Walsman, [Muru Zhang](#), Adam Fishman, Ali Farhadi, Dieter Fox. (2024)

European Conference on Computer Vision (ECCV) 2024, URL: <https://www.arxiv.org/abs/2410.01111>

How Language Model Hallucinations Can Snowball

[Muru Zhang](#), Ofir Press, William Merrill, Alisa Liu, Noah A. Smith. (2023)

International Conference on Machine Learning (ICML) 2024, URL: <https://arxiv.org/abs/2305.13534>

STOW: Discrete-Frame Segmentation and Tracking of Unseen Objects for Warehouse Picking Robots

Yi Li, [Muru Zhang](#), Markus Grotz, Kaichun Mo, Dieter Fox. (2023)

Conference on Robot Learning (CoRL) 2023, URL: <https://sites.google.com/view/stow-corl23>

Impossibly Good Experts and How to Follow Them

Aaron Walsman, [Muru Zhang](#), Sanjiban Choudhury, Dieter Fox, Ali Farhadi. (2023)

International Conference on Learning Representations (ICLR) 2023, URL: https://openreview.net/forum?id=sciA_xgYofB

Measuring and Narrowing the Compositionality Gap in Language Models

Ofir Press, [Muru Zhang](#), Sewon Min, Ludwig Schmidt, Noah A. Smith, Mike Lewis. (2022)

Empirical Methods in Natural Language Processing (EMNLP) 2023 Findings, URL: <https://arxiv.org/abs/2210.03350>

Break and Make: Interactive Structural Understanding Using LEGO Bricks

Aaron Walsman, [Muru Zhang](#), Klemen Kotar, Karthik Desingh, Ali Farhadi. (2022)

European Conference on Computer Vision (ECCV) 2022, URL: <https://arxiv.org/abs/2207.13738>

Industry Experience

Google Deepmind, Student Researcher

San Francisco, CA

Mentors: Yann N. Dauphin, Michael Mozer

May 2025 – August 2025

- Proposed a new kind of residual module that improves scaling law curve of Transformer-based language model.

Together AI, Research Scientist Intern

San Francisco, CA

Mentors: Ben Athiwaratkun, Tri Dao

May 2024 – May 2025

- Designed Transformer architecture variants that allow better overlapping of communication and computation in Tensor Parallelism for faster inference.
- Exploring and comparing various post-training approaches to adapt a pretrained model for various downstream use cases.

Amazon Web Services (AWS) CodeWhisperer, Applied Scientist Intern

New York, NY

Supervisor: Haifeng Qian

June 2023 – September 2023

- Designed and implemented a new transformer architecture that aims to model long-context more efficiently. Trained a series of models up to 1.3B parameters, achieved up to 6x speedup at the inference time compared with vanilla transformer on the range of 16k context length and is able to maintain up to 92% of the performance.

Amazon/UW Robotics and State Estimation Lab, Perception Team Intern

Seattle, WA

Supervisors: Dieter Fox, Maya Cakmak, Joshua Smith

June 2022 – March 2023

- Worked on the perception part of a robotics project on developing a grasping pipeline in warehouse settings, replicating tasks performed by Amazon workers. Built a transformer-based video segmentation and tracking model based on Detectron2 that better utilizes memory information to handle ambiguous detection scenarios and allow object re-identification even with large movements and occlusion.

Teaching Experience

Teaching Assistant, University of Washington – Computer Science and Engineering Department

- CSE 312: Foundations of Computing II** March 2024 – June 2024
 - CSE 447/517: Natural Language Processing** December 2023 – March 2024
 - CSE 421: Introduction to Algorithms** September 2022 – December 2023
 - CSE 311: Foundations of Computing I** January 2022 – June 2022
 - CSE 417: Algorithms and Computational Complexity** September 2021 – December 2021
-