Ting-Yun (Charlotte) Chang

Homepage: https://terarachang.github.io
Email: tingyun@usc.edu

RESEARCH INTERESTS

Inference Efficiency Reducing LLM serving costs through model quantization and KV cache compression
Actionable Interp. Diagnosing model behavior and translating scientific insights into actionable methods

EDUCATION

University of Southern California, USA *2021 - Spring 2026*
PhD student in Department of Computer Science
(All but dissertation)

National Taiwan University, Taiwan *2018 - 2020*
M.S. in Department of Computer Science and Information Engineering

National Tsing Hua University, Taiwan *2014 - 2018*
B.S. in Department of Computer Science Rank 2/41; GPA 4.14/4.3

Tsinghua University, China *Fall 2015*
Exchange Student in Department of Computer Science and Technology

RESEARCH EXPERIENCE

University of Southern California California, USA
Research Assistant *2021 - Present*

- Advisors: Robin Jia and Jesse Thomason
- KV cache compression on reasoning models with eviction, sparse attention, and quantization (ongoing)
- Studying why low-bit model quantization affects certain examples disproportionately [1]
- Improving LLMs' consistency to prompt variants with model decomposition and efficient weight updates [3]
- Localizing where memorized data resides in LLM parameters and unlearning them [4]
- Improving the stability of in-context learning to prompt choices via data valuation of demonstrations [5]
- Benchmarking continual learning abilities of vision-language models with a suite of multimodal tasks [6]

Google DeepMind New York, USA
Research Intern *Summer 2025*

- Hosts: Xiang Zou, Zun Li, Jiao Sun, and Shyam Upadhyay
- Studying the interplay between RL algorithms and post-training model quantization

Amazon AWS AI California, USA
Applied Scientist Intern *Summer 2024*

- Improving the safety of LLMs against jailbreaking attacks with steering

Academia Sinica Taipei, Taiwan
Research Assistant *2020 - 2021*

- PI: Chi-Jen Lu
- Studying which pre-finetuning tasks can improve pretrained language models [7]
- Compressing large image generators and stabilizing GANs training with knowledge distillation [10]

Amazon Alexa AI California, USA
Applied Scientist Intern *Spring 2020*

- Hosts: Yang Liu and Dilek Hakkani-Tür
- Improving models' commonsense knowledge with specialized loss terms and learnable knowledge base [8, 9]

National Taiwan University Taipei, Taiwan
Research Assistant *2018 - 2020*

- Advisor: Yun-Nung (Vivian) Chen
- Probing the degree of contextualization in ELMo and BERT embeddings using multisense word definitions [11]
- Automated clinical note diagnosis [12]

PUBLICATIONS

[1] **Ting-Yun Chang**, Muru Zhang, Jesse Thomason, and Robin Jia. **Why Do Some Inputs Break Low-Bit LLM Quantization?** EMNLP 2025.

[2] Ming Zhong, Xiang Zhou, **Ting-Yun Chang**, Qingze Wang, Nan Xu, Xiance Si, Dan Garrette, Shyam Upadhyay, Jeremiah Liu, Jiawei Han, Benoit Schillings, and Jiao Sun. **Vibe Checker: Aligning Code Evaluation with Human Preference**. arXiv 2025. Internship Project.

[3] **Ting-Yun Chang**, Jesse Thomason, and Robin Jia. **When Parts Are Greater Than Sums: Individual LLM Components Can Outperform Full Models.** EMNLP 2024.

[4] **Ting-Yun Chang**, Jesse Thomason, and Robin Jia. **Do Localization Methods Actually Localize Memorized Data in LLMs? A Tale of Two Benchmarks.** NAACL 2024.

[5] **Ting-Yun Chang** and Robin Jia. **Data Curation Alone Can Stabilize In-context Learning.** ACL 2023.

[6] Tejas Srinivasan, **Ting-Yun Chang**, Leticia Leonor Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. **CLiMB: A Continual Learning Benchmark for Vision-and-Language Tasks.** NeurIPS Datasets and Benchmarks Track 2022.

[7] **Ting-Yun Chang** and Chi-Jen Lu. **Rethinking Why Intermediate-Task Fine-Tuning Works.** Findings of EMNLP 2021.

[8] **Ting-Yun Chang**, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tür. **Go Beyond Plain Fine-tuning: Improving Pretrained Models for Social Commonsense.** IEEE SLT 2021.

[9] **Ting-Yun Chang**, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tür. **Incorporating Commonsense Knowledge Graph in Pretrained Models for Social Commonsense Tasks.** DeeLIO Workshop at EMNLP 2020 (**best paper award**).

[10] **Ting-Yun Chang** and Chi-Jen Lu. **TinyGAN: Distilling BigGAN for Conditional Image Generation.** Asian Conference on Computer Vision 2020.

[11] **Ting-Yun Chang** and Yun-Nung Chen. **What Does This Word Mean? Explaining Contextualized Embeddings with Natural Language Definition.** EMNLP-IJCNLP 2019.

[12] Shang-Chi Tsai, **Ting-Yun Chang**, and Yun-Nung Chen. **Leveraging Hierarchical Category Knowledge for Data-Imbalanced Multi-Label Diagnostic Text Understanding.** LOUHI Workshop at EMNLP-IJCNLP 2019.

[13] Chao-Chun Liang, Shih-Hong Tsai, **Ting-Yun Chang**, Yi-Chung Lin, and Keh-Yih Su. **A Meaning-based English Math Word Problem Solver with Understanding, Reasoning and Explanation.** COLING 2016: System Demonstrations.

TEACHING EXPERIENCE

Teaching Assistant
USC CS 544 Applied Natural Language Processing, Fall 2024. Instructor: Swabha Swayamdipta.
USC CS 467 Introduction to Machine Learning, Spring 2023. Instructor: Robin Jia.
NTU CS Applied Deep Learning, Spring 2019. Instructor: Yun-Nung (Vivian) Chen.

PROGRAMMING

Languages: Python, C/C++, Java
Frameworks: PyTorch, TensorFlow, Triton, vLLM, verl