ANDREW BAI

andrewbai@cs.ucla.edu | Github | Google Scholar

PhD Candidate in Computer Science at UCLA, advised by Prof. Cho-Jui Hsieh. Research focuses on improving individual stages of the model training pipeline through a holistic, system-level perspective, with recent emphasis on **post-training** for large language models (e.g., SFT, RLHF, multimodality).

Other projects: reward modeling, LLM jailbreaking, continual learning, data attribution, interpretability

WORK EXPERIENCE

Nvidia Sep 2025 – Dec 2025 (expected)
LLM Technologist Intern

- Achieve the best of both worlds by bridging off-policy (fast and simple) and on-policy (better generalization and less forgetting) RLHF techniques for LLM post-training.

Google
Research Intern @ **DeepMind** Jun 2025 – Sep 2025

- Architected an end-to-end prototype with cross-team collaboration for tool calling in the Gemini App using SLMs loaded on Android devices. Achieved an average of 98% accuracy and sub-1s E2E latency.
- Designed a benchmark for LLM response formatting inspired by pedagogical principles. Leveraged LLM persona role-play for online multi-turn auto-eval. Identified 60%+ losses with LLM-as-a-judge.

Student Reseacher @ **Bard** Jun 2024 – Oct 2024

- Developed novel early-stopping metrics for supervised fine-tuning on instruction data to maximize downstream DPO performance. Validated hypothesis across 7 instruction and 3 alignment datasets.

Student Reseacher @ **Cloud** Apr 2023 – Aug 2023

- Designed a computationally-free rehearsal scheme to mitigate catastrophic forgetting by increasing the likelihood of sampling “useful” samples. Achieved equal performance with up to 50% less computation.

Amazon Jun 2022 – Sep 2022
Applied Scientist Intern

- Implemented and optimized factorization machine training and inferencing in C++, increasing the training speed by 43x compared to `libffm` (see open-source code PECOS for details).

EDUCATION

University of California, Los Angeles Sep 2021 – Dec 2025 (expected)
Ph.D. in Computer Science

National Taiwan University Sep 2016 – Jan 2021
B.S. in Computer Science and Information Engineering

SELECTED PUBLICATIONS

- On the Loss of Context-awareness in General Instruction Fine-tuning Under review.
- When More Instruction-Tuning Hurts: Rethinking the Path to Better Pairwise Alignment Embargoed.
- Concepts or Skills? Rethinking Instruction Selection for Multi-modal Models Under review.
- An Efficient Rehearsal Scheme for Catastrophic Forgetting Mitigation during Multi-stage Fine-tuning In *Findings of the Association for Computational Linguistics* (**NAACL**), Apr 2025.
- Concept Gradient: Concept-based Interpretation Without Linear Assumption In *Proceedings of the 11th International Conference on Learning Representations* (**ICLR**), May 2023.