ANDREW BAI

andrewbai@cs.ucla.edu

I am currently a fifth year PhD student at UCLA Computer Science Department advised by Prof. Cho-Jui Hsieh. My research interest involves high quality **data selection** and studying how data impacts novel skill **acquisition** and existing skill **forgetting**. Recently I am focused on language model (LLM) post-training, with my latest project investigating why training LLM with reinforcement learning leads to less forgetting compared to supervised fine-tuning. For previous project, I investigated the forgetting of context-awareness when pretrained LLMs are supervised fine-tuned on instruction data. I also engage in concurrent collaboration projects spanning divserse topics, including reward modeling, video generation, LLM agents, diffusion model data memorization, and prompt optimization.

EDUCATION

University of California, Los Angeles Sep 2021 – Dec 2025 (expected)
Ph.D. in Computer Science (advised by Prof Cho-Jui Hsieh)

National Taiwan University Sep 2016 – Jan 2021
B.S. in Computer Science and Information Engineering (GPA: 4.2/4.3)
Minor in Mechanical Engineering

SELECTED PUBLICATIONS

- A. Bai, J. Cui, R. Wang, and C.-J. Hsieh. **Concepts or Skills? Rethinking Instruction Selection for Multi-modal Models.** Under submission review.
- Y. Wang*, A. Bai*, N. Peng, and C.-J. Hsieh. **On the Loss of Context-awareness in General Instruction Fine-tuning.** Under submission review.
- L. Lan, A. Bai, M. Cheng, C.-J. Hsieh, and T. Zhou. **Exploring Expert Failures Improves LLM Agent Tuning.** Under submission review.
- A. Bai, C.-K. Yeh, C.-J. Hsieh, and A. Taly. **An Efficient Rehearsal Scheme for Catastrophic Forgetting Mitigation during Multi-stage Fine-tuning.** In *Findings of the Association for Computational Linguistics: 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, Apr 2025.
- A. Bai, C.-K. Yeh, P. K. Ravikumar, N. Lin, and C.-J. Hsieh. **Concept Gradient: Concept-based Interpretation Without Linear Assumption.** In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, May 2023.

WORK EXPERIENCE

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

- Investigated the severity of catastrophic forgetting when fine-tuning LLMs with different reinforcement learning algorithms, expert demonstrations, and reward models for reasoning tasks.

Google DeepMind (Mountain View, CA) Jun 2025 – Sep 2025
Research Intern

- Created an end-to-end prototype for supporting tool calls in the Gemini App using small language models (SLM) loaded on mobile devices. Synthesized high quality tool call instruction-response pairs for model fine-tuning and improved the accuracy from 70% to 98%. Implemented tool calls in Kotlin.
- Created a challenging evaluation benchmark for LLM response formatting inspired by pedagogical principles. Leveraged LLM persona role-play to enable multi-turn conversational evaluation.

Google Bard (Remote) Jun 2024 – Oct 2024
Student Researcher

· Investigated early-stopping metrics for supervised fine-tuning (SFT) on instruction data to maximize downstream preference alignment (specifically DPO) performance.
· Benchmarked "alignment tax" – the deterioration of instruction following capabilities after preference alignment (RLHF) on LLMs, and identified it as an artifact caused by longer generated responses.

Google Cloud (Remote) Apr 2023 – Aug 2023
Student Researcher

· Designed a computationally-free rehearsal scheme to mitigate catastrophic forgetting for multi-stage training by increasing the likelihood of sampling "useful" samples online. Achieved equivalent performance to baselines with up to 50% less computation.
· Investigated targeted active learning setting on improving specific data slice while maintaining performance on the overall data distribution.

Amazon (Palo Alto, CA) 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).
· Investigated the impact of replacing inner product search with cross-attention methods (e.g. factorization machine) in two-tower deep neural network retrieval models.

RESEARCH EXPERIENCE

Dept. of Computer Science, UCLA (Los Angeles, CA) Sep 2021 – Present
Graduate Student Researcher (advised by Prof. Cho-Jui Hsieh)

· Identified the loss of context-awareness after instruction-tuning LLMs, traced the problem source to the bias in instruction data, and proposed straighforward solutions utilizing attention-steering and conditional supservied fine-tuning (CSFT).
· Designed concept-based interpretability methods for general differentiable models (e.g. neural networks) by propagating gradients through shared input feature representation.

Dept. of Computer Science and Engineering, NTU (Taipei, Taiwan) Jun 2018 – Jan 2021
Research Assistant (advised by Prof. Hsuan-Tien Lin)

· Collaborated with Kaggle (now a subsidiary of Google) on generative modeling metric design and held the first-ever public large-scale generative modeling competition with 900+ participating teams.
· Designed the first deep neural network model to predict tropical cyclone rapid intensification using satellite image data and establish strong baseline alongside our proposed dataset and benchmark.

TEACHING AND MENTORSHIP

Undergraduate Research Mentorship Jan 2023 – Present

· Volunteered mentoring 20+ undergraduate students on machine learning research projects. The mentorship led to **Data Attribution for Diffusion Models: Timestep-induced Bias in Influence Estimation.** being published in *Transactions on Machine Learning Research (TMLR)*, Jun 2024.
· Volunteered instructing *Algorithms* to 50+ high school students (LACC 2024).
· Assisted teaching *Introduction to Algorithms and Complexity* and *Introduction to Programming* for 2 academic quarters each.