WEI XIONG

Computer Science, University of Illinois Urbana-Champaign
wx13@illinois.edu | Website | GitHub | Scholar

RESEARCH INTERESTS

My research focuses on *reinforcement learning* and its applications in *LLM post-training*. Beyond algorithm design, I am also interested in understanding *training dynamics* and the *mathematical foundations* behind these methods.

Some of my past work include: *iterative rejection-sampling fine-tuning*[1][2][12], *online iterative DPO*[9], *regret analysis of KL-regularized RL*[3][9], and exploring the advantages of *GRPO*[1] from the viewpoint of online data filtering.

I **co-founded and lead** the open-source **RLHFlow**[6][7][8] project (**2,000+ GitHub stars**, **∼500 academic citations**), and have trained many widely used open-source reward models for RLHF, including the **first open-source implementation** of *(generative) process reward* (**1M+ downloads** on Hugging Face). Among them, the *multi-head reward models* with *MoE-style aggregation* (**ARMO**[7]) have contributed to the open-source research community with **200+ citations**.

EDUCATION

University of Illinois Urbana–Champaign — *Urbana, USA*
Ph.D. Candidate in Computer Science, GPA: 4.0/4.0 — *Aug. 2023 – Present*
Advisors: Prof. Tong Zhang, Prof. Nan Jiang

The Hong Kong University of Science and Technology — *Hong Kong*
M.Phil. in Mathematics — *Aug. 2023*
Advisor: Prof. Tong Zhang

University of Science and Technology of China — *Hefei, China*
B.Sc. in Mathematics & Electronic Engineering — *Jun. 2021*
Ranking: 1/72 in Statistics; 2/352 in EE

Shanghai Jiao Tong University — *Shanghai, China*
Exchange Student, School of Electronic Information and Electrical Engineering — *2018*

EXPERIENCE

Research Scientist Intern, Meta FAIR, Alignment Team — *May 2025 – Present*
Hosts: Dr. Sainbayar Sukhbaatar, Dr. Jason Weston
Developed a stepwise generative judge trained via RL, incorporating a self-segmentation technique for trajectory splitting and a complete RL recipe. Achieved significant improvements over SFT-trained classification baselines across all evaluation axes.

Student Researcher, Google DeepMind, Gemini Post-training Team — *May 2024 – Mar. 2025*
Hosts: Dr. Tianqi Liu, Dr. Bilal Piot
Designed and implemented multi-turn DPO for tool-using agents capable of reasoning over self-decoded tokens and external messages. Also investigated the theoretical connection between process reward and Q-learning, and scaled training for generative process rewards.

University of Illinois Urbana–Champaign — *Feb. 2024 – Present*
Advisors: Prof. Tong Zhang, Prof. Nan Jiang
Lead developer of RLHFlow, providing a complete pipeline for online iterative RLHF, including SFT, reward/preference modeling, and RLHF/DPO. Released state-of-the-art open-source reward/preference

models, contributing to 200+ research projects with 2000+ GitHub stars, ∼500 academic citations and 1M+ downloads on Hugging Face. The final LLMs match or surpass LLaMA3-8B-instruct.

The Hong Kong University of Science and Technology *Jan. 2023 – Present*
Advisor: Prof. Tong Zhang
Core founding member of LMFlow, an LLM development framework (8K+ GitHub stars, ranked #2 on GitHub trending). Led RLHF module design and implementation and won Best Paper Award in Demo Track, NAACL 2024.

Microsoft Research Asia, Networking Research and Machine Learning Group *Spring 2021*
Advisors: Dr. Wenxue Cheng, Dr. Li Zhao
Developed RL-based bandwidth estimation methods for real-time communications of Microsoft Teams.

SELECTED AWARDS AND FELLOWSHIPS

Google PhD Fellowship, Finalist	2025
Best Paper Award, Demo Track, NAACL	2024
Hong Kong PhD Fellowship Scheme (HKPFS) (approx. $90,000 USD over two years)	2021 - 2023
Best Teaching Assistant Award, HKUST (Awarded twice)	2022, 2023
Outstanding Graduate of Anhui Province and USTC	2021
Guo Moruo Scholarship, Finalist (Highest honor for undergraduates at USTC)	2020
Yuanqing Yang Scholarship, USTC	2020
National Scholarship (Awarded by Ministry of Education, PRC)	2017

PROFESSIONAL ACTIVITY

Conference Reviewer

- ICLR (2024-2025), NeurIPS (2022-2024, **Top Reviewer Award (Top 8%) 2023**), ICML (2022-2023, 2025), AISTATS (2023-2025), ARR (2024-2025)

Journal Reviewer

- Journal of Machine Learning Research (JMLR), Transactions on Machine Learning Research (TMLR), Journal of the American Statistical Association (JASA)

Invited Talks

Frequently invited to speak at leading academic institutions and industry labs on my research in LLM alignment, reinforcement learning theory, and building agentic AI systems. Selected venues include:

- **Industry Labs:** Google (DeepMind, Learning Theory Seminar, RLHF Workshop), Microsoft Research (MSR Asia), Amazon.
- **Top Universities & Institutes:** Stanford, Yale, UChicago (TTIC), UCLA, UIUC, UWaterloo, U of Toronto, UCSB, UW-Madison, Peking University, MBZUAI, HKU.
- **Leading Research Centers:** Simons Institute for the Theory of Computing, Mila (Alignment Seminar), INFORMS Annual Meeting.

Guest Lectures

Delivered guest lectures for graduate-level courses at the University of Virginia (CS 6501 & CS 4501) and the University of Wisconsin-Madison (CS 760).

SELECTED PROJECTS

(α, β) denotes random or alphabetical order, * denotes equal contribution.

[1] Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, Hanze Dong, "A Minimalist Approach to LLM Reasoning: from Rejection Sampling to Reinforce", Preprint.

[2] Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, Wei Xiong, Nan Jiang, Tong Zhang, "Optimizing Chain-of-Thought Reasoners via Gradient Variance Minimization in Rejection Sampling and RL", Preprint.

[3] Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, Tong zhang, "Logarithmic Regret for Online KL-Regularized Reinforcement Learning", [ICML 2025].

[4] Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang, "Self-rewarding Correction for Mathematical Reasoning", Submitted.

[5] Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, Chi Jin, Tong Zhang, Tianqi Liu, "Building Math Agent by Iterative Preference Learning", [ICLR 2025] [Code].

[6] (α, β) Hanze Dong*, Wei Xiong*, Bo Pang*, Haoxiang Wang*, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, Tong Zhang, "RLHF Workflow: From Reward Modeling to Online RLHF", [Transactions on Machine Learning Research (TMLR)] [Code].

[7] Haoxiang Wang*, Wei Xiong*, Tengyang Xie, Han Zhao, Tong Zhang, "Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts", [EMNLP 2024] [Code].

[8] Wei Xiong*, Hanning Zhang, Nan Jiang, Tong Zhang, "An Implementation of Generative PRM", [Code and Blog].

[9] Wei Xiong*, Hanze Dong*, Chenlu Ye*, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, Tong Zhang, "Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint", [ICML 2024] [Code].

[10] (α, β) Yong Lin*, Hangyu Lin*, Wei Xiong*, Shizhe Diao*, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Yuan Yao, Heng Ji, and Tong Zhang, "Mitigating the Alignment Tax of RLHF", [EMNLP 2024].

[11] Zhihan Liu*, Miao Lu*, Wei Xiong*, Han Zhong, Hao Hu, Shenao Zhang, Sirui Zheng, Zhuoran Yang, Zhaoran Wang, "Maximize to explore: One objective function fusing estimation, planning, and exploration", submitted to [Operation Research], a short version accepted to [NeurIPS 2023].

[12] (α, β) Hanze Dong*, Wei Xiong*, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum and Tong Zhang, "RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment" [Transactions on Machine Learning Research (TMLR)] [Code].

[13] (α, β) Han Zhong*, Wei Xiong*, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang, "GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond", [Mathematics of Operation Research (MOR)] [Slide].