

Wei Xiong

Member of Technical Staff, OpenAI | RL / Frontier Team
weixiongml@gmail.com | GitHub | Google Scholar

Industry Experience

OpenAI Member of Technical Staff, RL / Frontier Team Develop reinforcement learning algorithms to advance the capabilities of frontier AI models.	Jan 2026 - present
Meta FAIR Research Scientist Intern, Alignment Team Hosts: Sainbayar Sukhbaatar and Jason Weston Developed a stepwise generative judge trained with RL, combining self-segmentation of reasoning trajectories with a process-reward training recipe.	May 2025 - Aug 2025
Google DeepMind Student Researcher, Gemini Post-Training Team Hosts: Tianqi Liu and Bilal Piot - Designed multi-turn DPO for tool-using agents that reason over model-generated tokens and external feedback. - Investigated connections between process rewards and Q-learning, and scaled training for generative process reward models.	May 2024 - Apr 2025
Microsoft Research Asia Research Intern Developed RL-based bandwidth estimation methods for real-time communications in Microsoft Teams; studied distributional reinforcement learning.	Jan 2021 - Jun 2021

Education

University of Illinois Urbana-Champaign Ph.D. in Computer Science Advisors: Tong Zhang and Nan Jiang	2026
The Hong Kong University of Science and Technology M.Phil. in Mathematics Advisor: Tong Zhang	2023
University of Science and Technology of China B.Sc. in Mathematics & Electronic Engineering	2021

Honors & Fellowships

Google PhD Fellowship	2025
Best Paper Award, NAACL Demo Track	2024
Hong Kong PhD Fellowship Scheme	2021 - 2023
Best Teaching Assistant Award, HKUST	2022, 2023
Outstanding Graduate of Anhui Province and USTC	2021
Guo Moruo Scholarship, Finalist	2020

Selected Research & Open Source

RL Algorithms for LLM Reasoning

- Developed rejection-sampling and policy-gradient methods for LLM reasoning, connecting practical post-training recipes with theoretical analysis. [Minimal-RL](#)
- Designed adaptive sampling methods that allocate inference budgets across prompts to reduce gradient variance and recover learning signals. [Reinforce-Ada](#)

Online DPO & RLHFlow

- Co-founded and led [RLHFlow](#), the open-source framework accompanying online DPO, covering supervised fine-tuning, reward modeling, and online preference optimization.
- Established convergence and logarithmic regret analyses for KL-regularized RLHF, connecting mathematical foundations with practical algorithm design.

Reward Modeling

- Developed ARMO, a multi-objective reward model with mixture-of-experts aggregation, and released open-source preference models and generative process reward training recipes. [Code](#)

LMFlow

- Core founding member; led the design and implementation of the RLHF module. Received the Best Paper Award in the NAACL 2024 Demo Track.

Selected Publications

* Equal contribution. Author order follows the cited papers.

Reinforce-Ada: An Adaptive Sampling Framework under Non-linear RL Objectives

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, Tong Zhang
arXiv, 2025 | [Paper](#)

Optimizing Chain-of-Thought Reasoners via Gradient Variance Minimization in Rejection Sampling and RL

Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, **Wei Xiong**, Nan Jiang, Tong Zhang
NeurIPS 2025 | [Paper](#)

A Minimalist Approach to LLM Reasoning: from Rejection Sampling to Reinforce

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, Hanze Dong
arXiv, 2025 | [Paper](#)

Logarithmic Regret for Online KL-Regularized Reinforcement Learning

Heyang Zhao, Chenlu Ye, **Wei Xiong**, Quanquan Gu, Tong Zhang
ICML 2025 | [Paper](#)

Self-rewarding Correction for Mathematical Reasoning

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, Tong Zhang
arXiv, 2025 | [Paper](#)

Building Math Agents with Multi-Turn Iterative Preference Learning

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
ICLR 2025 | [Paper](#)

RLHF Workflow: From Reward Modeling to Online RLHF

Hanze Dong*, **Wei Xiong***, Bo Pang*, Haoxiang Wang*, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, Tong Zhang
Transactions on Machine Learning Research, 2024 | [Paper](#)

Selected Publications (continued)

Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts

Haoxiang Wang*, **Wei Xiong***, Tengyang Xie, Han Zhao, Tong Zhang
EMNLP 2024 | [Paper](#)

Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint

Wei Xiong*, Hanze Dong*, Chenlu Ye*, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, Tong Zhang
ICML 2024 | [Paper](#)

Mitigating the Alignment Tax of RLHF

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, Tong Zhang
EMNLP 2024 | [Paper](#)

A Theoretical Analysis of Nash Learning from Human Feedback under General KL-Regularized Preference

Chenlu Ye, **Wei Xiong**, Yuheng Zhang, Nan Jiang, Tong Zhang
NeurIPS 2024 | [Paper](#)

RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment

Hanze Dong*, **Wei Xiong***, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, Tong Zhang
Transactions on Machine Learning Research, 2023 | [Paper](#)

One Objective to Rule Them All: A Maximization Objective Fusing Estimation and Planning for Exploration

Zhihan Liu*, Miao Lu*, **Wei Xiong***, Han Zhong, Hao Hu, Shenao Zhang, Sirui Zheng, Zhuoran Yang, Zhaoran Wang
NeurIPS 2023 | [Paper](#)

GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond

Han Zhong*, **Wei Xiong***, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, Tong Zhang
Mathematics of Operations Research | [Paper](#)