# Tianhao Wu

341-766-8125 | thw@berkeley.edu

## EDUCATION

**University of California, Berkeley - PhD** 08/2021-Present

*Department of Electrical Engineering and Computer Sciences*

**Peking University - Undergraduate** 09/2017-07/2021

*School of Mathematical Sciences*

Major in Statistics and Probability, Rank: 5%

*Awards and Honors*

| | |
|---|---|
| Merit Student, Peking University | 09/2017-07/2021 |
| 1st Prize in Russian Mathematical Olympiad | 08/2015 |
| 1st Prize in Chinese Mathematical Olympiad | 01/2015 |

## CURRENT INTEREST

At present, my research focuses on fine-tuning LLMs with RLHF. I'm particularly drawn to the prospect of developing an AI agent capable of self-evolution, with minimal human supervision. I'm also exploring the concept of decentralized intelligence. I envision a system where individual AI agents can link together in modular fashion to form a more capable collective intelligence. Such framework could mitigate the huge memory and computing demand that limit centralized AI systems.

## WORK EXPERIENCE

**TikTok Inc** 09/2023-Present

- Serving as a member of the Data-US team. My primary focus lies in enhancing the safety and robustness of the recommendation system.
- My work includes leveraging off-the-shelf LLMs in annotation and RL methods in optimizing performance and trustworthiness of the system.

## RESEARCH

**Starling-7B: Increasing LLM helpfulness & Harmlessness with RLAIF (Blog Post)**

- We release Starling-7B, an open-source language model trained by RLAIF. The model harnesses the power of the innovative GPT-4 labeled ranking dataset, along with our advanced reward training and policy tuning pipeline
- Starling-7B-alpha scores 8.09 in MT Bench with GPT-4 as a judge, outperforming every model to date on MT-Bench except for GPT-4 and GPT-4 Turbo

**Pairwise Proximal Policy Optimization: Harnessing Relative Feedback in LLM Alignment (Blog Post)**

*Under Review*

- Propose the new RL with relative feedback framework for optimizing reward trained from comparative loss. In contrast to traditional RL approaches that rely on absolute feedback - not suited for LM alignment
- Develop P3O, a new policy gradient algorithm that features comparative update. This approach addresses the discrepancy between the reward learning stage and the RL fine-tuning stage, unifying them through comparative training.
- Empirical evaluations demonstrate P3O's superior in terms of KL-Reward trade-off. Evaluated by GPT-4, it exhibited a high win-rate against established algorithms such as PPO and DPO.

**A Reduction-based Framework for Sequential Decision Making with Delayed Feedback**

*Accepted by NeurIPS 2023*

- Propose a reduction-based framework which turns any multi-batched algorithm for sequential decision making with instantaneous feedback into a sample-efficient algorithm that can manage stochastic delays in sequential decision making.
- Prove the corresponding sharp upper bound, as a result, provide the first guarantee in sequential decision making with function approximation.

**Nearly Optimal Policy Optimization with Stable at Any Time Guarantee**
*Accepted by ICML 2022*

- Propose a novel Online Mirror Descent type algorithm that eliminate the data coverage assumption.
- Show that by using a novel reference V estimator and l2 regularization term in OMD, we can ensure the stability of the V estimation.
- Further show that the stability property is the key to achieve nearly optimal regret and obtain SOTA regret upper bound in the policy optimization setting.

**A unified framework for conservative exploration**
*Accepted by ICLR 2021*

- Propose a unified framework for conservative bandits and RL, in which the core is to calculate the necessary and sufficient budget obtained from running baseline policy.
- Based on the framework, we can turn a certain nonconservative algorithm into a conservative one
- Obtain SOTA regret upper and lower bound in tabular and low-rank settings.

**On Reinforcement Learning with Adversarial Corruption and Its Application to Block MDP**
*Accepted by ICML 2021*

- Propose a new algorithm for RL that can be robust to adversarial corruption with intensity lower than C.
- Prove that the regret incurred by the corruption can be upper bounded by SAC in the tabular setting and establish a corresponding lower bound to show that the algorithm is optimal.
- Apply our algorithm to BMDPs yields the first sqrt T regret bound in this setting.

**Sanity-checking pruning methods: Random tickets can win the jackpot**
*Accepted by NeurIPS 2021*

- Our experiments show that the conventional beliefs that ‘Pruning methods exploit information from training data to find good subnetworks’ and ‘the architecture of the pruned network is crucial for good performance’ are wrong.
- Propose a simple zero-shot random pruning method that outperform or attains similar performance compared to SOTA.