Tianhao Wu

thw@berkeley.edu

EDUCATION

University of California, Berkeley - PhD 08/2021-Present
Department of Electrical Engineering and Computer Sciences

Peking University - Undergrad 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
Gold Medal in Russian Mathematical Olympiad (**RMO**) 08/2015
Gold Medal in Chinese Mathematical Olympiad (**CMO**) 01/2015

CURRENT INTEREST

My research focuses on LLM post-training. I'm interested in designing scalable training techniques to improve models' reasoning and planning capabilities, as well as developing multi-agent systems capable of self-evolution.

WORK EXPERIENCE

AI @ Meta 05/2024-Present
– My research focuses on AI self-improvement and enhancing the reasoning abilities of AI models

Nexusflow 02/2024-05/2024
– We trained *Starling-7B-LM-beta*, a small model surpassing *Mixtral-8x7b* and *Gemini Pro*

TikTok 09/2023-02/2024
– My work includes applying RL in optimizing performance and trustworthiness of recommendation system

RESEARCH

Thinking LLMs: General Instruction Following with Thought Generation (Post)
Under Review
- We found that optimizing a model's internal thought process yields gains not just in reasoning and math, but across *all instruction-following tasks*
- We employed RL and search to *incentivize* the model to enhance its own thought process, rather than explicitly teaching it how to think

Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge (Post)
Under Review
- We proposed Meta-Judge, which provide feedback for self-improving model's judging abilities
- Via self-play, we significantly improve the win-rate of *Llama-3-8B-Instruct* on Arena-Hard from 20.6% to **29.1%** and AlpacaEval from 22.9% to **39.4%**, approaching Claude-Opus without human supervision

Starling-7B: Increasing LLM helpfulness & Harmlessness with RLAIF (Blog)
Accepted by COLM 2024
- We introduce *Starling-LM-7B*, an open-source language model trained by RLAIF. The model harnesses the power of the innovative GPT-4 labeled ranking dataset - *Nectar*, along with our advanced reward training and policy tuning pipeline
- *Starling-7B-LM* ranked 1119 on Chatbot Arena, surpassing *Mixtral-8x7b* and *Gemini Pro*

Pairwise Proximal Policy Optimization: Harnessing Relative Feedback in LLM Alignment (Blog)
Accepted by COLM 2024
- 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 in RLHF, unifying them through comparative training
- P3O exhibited a high win-rate against established algorithms such as PPO and DPO

From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline (Blog)
Under Review

- We introduce **Arena-Hard**, a data pipeline to build high-quality benchmarks from live data in Chatbot Arena

RouteLLM: Learning to Route LLMs with Preference Data (Blog)
Under Review
- We train routers to route between strong and weak model pairs
- We achieve cost reductions of over 85% on MT Bench, 45% on MMLU, and 35% on GSM8K as compared to using only GPT-4, while still achieving 95% of GPT-4's performance

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.