Chenliang Li

chenliangli@tamu.edu

LLM Post-Training | Reinforcement Learning | Long-Horizon Agents

(+1) 979-344-3365

Google Scholar

Research Profile

Ph.D. candidate developing algorithms and training systems for reliable, efficient LLM post-training. Research spans fully asynchronous RL, multi-turn agent optimization, and joint reward–policy learning, with experience in dense and mixture-of-experts models. Additional work addresses robust and inverse reinforcement learning. Selected publications: **ICLR 2025 Spotlight**, **NeurIPS 2023 Oral**, **COLM 2026**, and **TMLR 2026**.

Education

Texas A&M University — **Aug. 2023 – Present**
Ph.D. in Data Science; Advisor: Prof. Alfredo Garcia — *College Station, TX*

The Chinese University of Hong Kong, Shenzhen — **Sep. 2021 – Jul. 2023**
Master of Data Science; First-Class Merit Scholarship (top 3%) — *Shenzhen, China*

Xi'an Jiaotong University — **Sep. 2017 – Jul. 2021**
Bachelor of Engineering in Computer Science — *Xi'an, China*

Research Internships

ByteDance | Research Intern — **May 2026 – Aug. 2026**
Large Language Model Systems Team; PACE — *San Jose, CA*

Where Does Staleness Accumulate? Pool Aware Effective Staleness Control for Asynchronous RL in LLM Post-Training (Preprint, 2026)

- Developed **PACE** to preserve the throughput gains of fully asynchronous RL while mitigating policy mismatch caused by rollout backlog. Separated trajectory staleness into generation and waiting components, converted excess pool occupancy into an adaptive rejection budget, and ranked trajectories using prefix-policy drift. This admission mechanism reduces avoidable waiting staleness while retaining long-running reasoning trajectories whose prefixes remain compatible with the updated policy.
- Built a **post-training recipe** with `verl` Async RL Trainer, `vLLM`, and **DAPO** on **DAPO-MATH-17K**; extended it to sandbox-based multi-turn reasoning with **GRPO/REINFORCE++** and **Qwen3-30B-A3B-Base** with **Megatron** tensor/expert parallelism. On Qwen3-8B, achieved an **18.7% relative improvement** in average mean@32 accuracy over unfiltered asynchronous RL at the **same wall-clock budget** across **AIME 2023, AIME 2024, AIME 2025, BeyondAIME, BrUMO 2025, and HMMT 2025**; matched synchronous RL performance with **47.1% less GPU time**. Tool-integrated Qwen3-4B improved average accuracy by **63.9% relative to unfiltered asynchronous RL** and **6.3% relative to synchronous RL**.

GE HealthCare | Research Intern — **May 2025 – Aug. 2025**
Long-Horizon LLM Agent Post-Training; SORL — *Bellevue, WA*

Stabilizing Off-Policy Training for Long-Horizon LLM Agents via Turn-Level Importance Sampling and Clipping-Triggered Normalization (**COLM 2026**)

- Developed **SORL** to address optimization collapse when multi-turn LLM agents reuse trajectories for repeated off-policy updates. Identified a mismatch between token-level optimization and turn-structured decisions, compounded by extreme importance ratios and unreliable advantage estimates. Designed **SO-PPO** with **turn-level importance sampling** for credit assignment and **clipping-triggered normalization (CTN)** to suppress unreliable updates; extended CTN to **SO-GRPO** while retaining GRPO's token-level importance weighting.
- Established an **agent post-training recipe** integrating interactive retrieval, PPO/GRPO optimization, and distributed training with `verl`, FSDP, and `vLLM`. Evaluated **Qwen2.5-7B** on **Natural Questions and HotpotQA**, and trained **Llama-3.1-8B-Instruct** on **AlphaMed19K** for evaluation on **MedQA, MedMCQA, PubMedQA, MMLU-M, and MedXpertQA**. The resulting COLM study demonstrated stable learning where standard baselines degraded; SO-GRPO improved average accuracy across the five medical QA benchmarks by **20.1% relative to the best Search-R1 checkpoint**.

Selected Research Projects

AIHF: Joint Reward and Policy Learning **ICLR 2025 | Spotlight**

Learning Reward and Policy Jointly from Demonstration and Preference Improves Alignment [arXiv]

- Developed **Alignment with Integrated Human Feedback (AIHF)** to improve alignment when demonstrations and preference annotations are imbalanced and conventional sequential RLHF underutilizes their complementary information. Formulated joint reward–policy learning as a **bilevel optimization** problem, combining demonstration likelihood with preference supervision. Designed alternating reward estimation and policy optimization, established finite-time performance guarantees, and developed DPO and self-play variants.
- Built a **Mistral-7B alignment recipe** using **UltraChat200K** demonstrations and **UltraFeedback** preferences: sample responses from the current policy, construct preference pairs using the learned reward model, apply **online DPO**, and alternate with reward-model updates. Used **DeepSpeed ZeRO-3** and `vLLM` for distributed training and generation. Improved average performance by **2.5% relative to DPO** across **ARC-Challenge, TruthfulQA-MC2, WinoGrande, GSM8K, HellaSwag, and MMLU**; additionally evaluated Pythia-160M/1B/2.8B and reward-model quality.

Offline ML-IRL: Learning Rewards with World Models **NeurIPS 2023 | Oral**

When Demonstrations Meet Generative World Models: A Maximum Likelihood Framework for Offline Inverse Reinforcement Learning (Co-first author) *[arXiv]*

- Developed **Offline ML-IRL** to recover reward functions and expert behavior from fixed offline datasets without online environment interaction. Addressed the propagation of world-model estimation error into reward learning when transition coverage is limited. Formulated maximum-likelihood reward estimation as a bilevel problem with conservative policy optimization, and established statistical and computational guarantees for reward estimation.
- Designed a two-stage learning procedure: fit an **ensemble of probabilistic dynamics models**, then alternate reward learning with **uncertainty-penalized model-based policy optimization**. Evaluated on **D4RL HalfCheetah, Hopper, and Walker2d** across medium, medium-replay, and medium-expert datasets. Achieved the highest mean performance among the compared methods in **six of nine configurations** with 5,000 expert samples, including comparisons with behavioral cloning, ValueDICE, and CLARE; demonstrated transfer of learned rewards to separate offline datasets for subsequent policy learning.

Additional Publications

[1] **Reinforcing Multi-Turn Reasoning in LLM Agents via Fine-Grained Reward Structure and Credit Assignment**. *Preprint.* [arXiv]

[2] **Stabilizing Off-Policy Training for Long-Horizon LLM Agents via Turn-Level Importance Sampling and Clipping-Triggered Normalization**. *COLM 2026.* [arXiv]

[3] **Where Does Staleness Accumulate? Pool Aware Effective Staleness Control for Asynchronous RL in LLM Post-Training**. *Preprint, 2026.* [arXiv]

[4] **Understanding Inverse Reinforcement Learning under Overparameterization: Non-Asymptotic Analysis and Global Optimality**. *AISTATS 2025.* [arXiv]

[5] **Adaptive Rank Control for Robust Reinforcement Learning**. *TMLR 2026.* [arXiv]

Technical Skills

LLM training: PPO, GRPO, DAPO, REINFORCE++, DPO, RLHF; asynchronous RL; multi-turn tool use.

Frameworks and systems: PyTorch, `verl`, `slime`, `vLLM`, Megatron, FSDP, DeepSpeed ZeRO-3, TRL, TRLX.

Programming and tools: Python, C++, MATLAB, NumPy, Linux, LaTeX.

Awards and Talks

ICCOPT talk: *Adaptive Low-Rank Representation in Reinforcement Learning*	2025
Ai4 invited speaker: *When Demonstrations Meet Generative World Models*	2024
First-Class Merit Scholarship, The Chinese University of Hong Kong, Shenzhen	2023
First Prize, National RoboMaster Robotics Competition, XJTU Computer Vision Team	2020
Third-Class Merit Scholarship, Xi'an Jiaotong University	2018