Novelty Assessment Report

Paper: TROLL: Trust Regions Improve Reinforcement Learning for Large Language Models
PDF URL: https://openreview.net/pdf?id=X9D5MVpPJ9
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Reinforcement Learning (RL) with PPO-like clip objectives has become the standard choice for reward-based fine-tuning of large language models (LLMs). Although recent work has explored improved estimators of advantages and normalization, the clipping mechanism itself has remained untouched. Originally introduced as a proxy for principled KL-based trust regions, clipping is a crude approximation that often causes unstable updates and suboptimal performance. We replace the clip objective with a novel discrete differentiable trust region projection, which provides principled token-level KL constraints. The projection operates on a sparse subset of the model's most important token logits to balance computational cost and projection effectiveness. Our approach, Trust Region Optimization for Large Language Models (TROLL), serves as a direct replacement for PPO-like clipping during training and does not alter the model's inference behavior. Across mathematical reasoning and code generation tasks, model families, as well as advantage-estimation methods, TROLL consistently outperforms PPO-like clipping in terms of training speed, stability, and final success rates.

Core Task Landscape

This paper addresses: **reinforcement learning for large language model fine-tuning**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Core RL Algorithms and Optimization Methods**
- **Reward and Preference Learning**
- **Application Domains and Task-Specific Adaptations**
- **Training Infrastructure and Efficiency**
- **Theoretical Foundations and Surveys**
- **Auxiliary Techniques and Complementary Methods**

Complete Taxonomy Tree

- reinforcement learning for large language model fine-tuning Survey Taxonomy
- Core RL Algorithms and Optimization Methods
  - Policy Optimization and Trust Region Methods ★ (4 papers)
  - [0] TROLL: Trust Regions Improve Reinforcement Learning for Large Language Models (Anon et al., 2026) View paper
  - [5] Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models (Li, 2023) View paper
  - [8] Secrets of rlhf in large language models part i: Ppo (Zheng Rui, 2023) View paper
  - [37] Nested-ReFT: Efficient Reinforcement Learning for Large Language Model Fine-Tuning via Off-Policy Rollouts (Cui, 2025) View paper
  - Offline and Regularized RL Approaches (2 papers)
  - [4] Offline regularised reinforcement learning for large language models alignment (Richemond, 2024) View paper
  - [31] On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting (Korbak, 2022) View paper
  - Online and Interactive RL (3 papers)
  - [35] Rlhf workflow: From reward modeling to online rlhf (Dong, 2024) View paper
  - [41] Grounding large language models in interactive environments with online reinforcement learning (Carta, 2023) View paper
  - [47] Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning (Xi, 2025) View paper
  - Specialized RL Techniques (4 papers)
  - [1] Revolutionizing reinforcement learning framework for diffusion large language models (Wang Yinjie, 2025) View paper
  - [9] Multi-Objective Reinforcement Learning for Large Language Model Optimization: Visionary Perspective (Kong Ling-xiao, 2025) View paper
  - [11] Hierarchical continual reinforcement learning via large language model (C Pan, 2024) View paper
  - [43] MetaEvo-Rec: Self-Evolving Meta-Reinforcement Learning Recommendation with Large-Language-Model Guided Policy Adaptation (Alamdari, 2025) View paper
- Reward and Preference Learning
  - Reward Model Design and Training (3 papers)
  - [7] Inverse reinforcement learning meets large language model post-training: Basics, advances, and opportunities (Sun Hao, 2025) View paper
  - [16] Secrets of rlhf in large language models part ii: Reward modeling (Wang Bing-hai, 2024) View paper

- [26] Structured preference modeling for reinforcement learning-based fine-tuning of large models (Zhu Lin, 2025) View paper
- Direct Preference Optimization (2 papers)
- [13] Active Preference Learning for Large Language Models (Hayes, 2024) View paper
- [18] Direct Preference Optimization: Your Language Model is Secretly a Reward Model (Rafailov, 2023) View paper
- Robust and Adversarial Reward Learning (3 papers)
- [21] Learning diverse attacks on large language models for robust red-teaming and safety tuning (Lee, 2024) View paper
- [22] Adversarial Reinforcement Learning for Large Language Model Agent Safety (Wang, 2025) View paper
- [24] Robust Reinforcement Learning from Human Feedback for Large Language Models Fine-Tuning (Ye Kai, 2025) View paper

- Application Domains and Task-Specific Adaptations
  - Reasoning and Mathematical Problem Solving (3 papers)
  - [6] Reinforcement Learning for Reasoning in Large Language Models with One Training Example (Wang Yi-ping, 2025) View paper
  - [39] Teaching large language models to reason with reinforcement learning (Havrilla, 2024) View paper
  - [42] Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning (Liu Zhaowei, 2025) View paper
  - Code Generation and Security (1 papers)
  - [34] Code Security Vulnerability Repair Using Reinforcement Learning with Large Language Models (Islam, 2024) View paper
  - Factuality and Alignment (3 papers)
  - [10] Fine-tuning language models for factuality (Tian, 2023) View paper
  - [15] OpenAssistant Conversations - Democratizing Large Language Model Alignment (Andreas Kopf, 2023) View paper
  - [40] Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback (Viet Lai, 2023) View paper
  - Domain-Specific Knowledge Integration (3 papers)
  - [28] A Framework for Domain-Specific Dataset Creation and Adaptation of Large Language Models (Georgios Balaskas, 2025) View paper
  - [33] Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue (Songhua Yang, 2023) View paper
  - [45] Adaptive Reinforcement Learning Framework for Enterprise Data Integration in LLM Training (Vemulapalli, 2025) View paper
  - Agent Behavior and Decision-Making (4 papers)
  - [3] Efficient reinforcement learning with large language model priors (Yan Xue, 2024) View paper
  - [19] Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning (Park Chanâ□□Woo, 2025) View paper
  - [30] Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents (Deng Yang, 2023) View paper
  - [50] MAGELLAN: Metacognitive predictions of learning progress guide autotelic LLM agents in large goal spaces (Loris Gaven, 2025) View paper
  - Vision-Language and Multimodal Models (1 papers)
  - [25] Fine-tuning large vision-language models as decision-making agents via reinforcement learning (Hao Bai, 2024) View paper
- Training Infrastructure and Efficiency
  - Scalable Training Frameworks (2 papers)
  - [14] Real: Efficient rlhf training of large language models with parameter reallocation (Mei, 2025) View paper
  - [44] Openrlhf: An easy-to-use, scalable and high-performance rlhf framework (Hu Jian, 2024) View paper
  - Parameter and Data Efficiency (2 papers)
  - [29] Self-Adapting Language Models (Pari, 2025) View paper
  - [48] Mixture-of-Skills: Learning to Optimize Data Usage for Fine-Tuning Large Language Models (Haffari, 2024) View paper
- Theoretical Foundations and Surveys
  - Comprehensive Surveys and Reviews (5 papers)
  - [2] A technical survey of reinforcement learning techniques for large language models (Aggarwal, 2025) View paper
  - [12] Reinforcement learning enhanced llms: A survey (Wang, 2024) View paper
  - [20] The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models (Moschoula Pternea, 2024) View paper
  - [27] Reinforcement Learning for Large Language Model Fine-Tuning: A Systematic Literature Review (L Kong, 2025) View paper
  - [49] Refining the giants: A comprehensive review of fine-tuning strategies for large language models (Kunal Singh, 2024) View paper
  - Specialized Theoretical Topics (4 papers)
  - [23] Reinforcement learning in large language models (llms): The rise of ai language giants (Baihan Lin, 2024) View paper
  - [32] Reinforcement learning: Advanced techniques for llm behavior optimization (Hariharan, 2025) View paper
  - [38] Reinforcement Learning for Prompt Optimization in Language Models: A Comprehensive Survey of Methods, Representations, and Evaluation Challenges (Zhangqi Liu, 2025) View paper
  - [46] Group Fairness in Reinforcement Learning and Large Language Models (Song, 2024) View paper
- Auxiliary Techniques and Complementary Methods (2 papers)
  - [17] Query Rewriting for Retrieval-Augmented Large Language Models (Ma, 2023) View paper
  - [36] Reflexion: language agents with verbal reinforcement learning (Shinn, 2023) View paper

Narrative

Core task: reinforcement learning for large language model fine-tuning. The field has organized itself into several major branches that reflect both algorithmic innovation and practical deployment concerns. Core RL Algorithms and Optimization Methods focus on policy optimization techniques—including trust region methods like those explored in Secrets PPO[8] and Remax[5]—that ensure stable updates when fine-tuning large models. Reward and Preference Learning addresses how to elicit and model human preferences, with works ranging from direct preference optimization (Direct Preference[18]) to active querying strategies (Active Preference[13]) and robust reward modeling (Secrets Reward[16]). Application Domains and Task-Specific Adaptations demonstrate the breadth of deployment scenarios, from conversational assistants (OpenAssistant Conversations[15]) to specialized domains like finance (Fin-R1[42]) and code security (Code Security[34]). Training Infrastructure and Efficiency tackle the computational challenges of scaling RL to billion-parameter models, while Theoretical Foundations and Surveys (Technical Survey[2], Enhanced Survey[12]) provide conceptual grounding, and Auxiliary Techniques explore complementary methods such as prompt optimization and self-adaptation.

Within the policy optimization landscape, a central tension emerges between sample efficiency and stability: some methods prioritize tight trust regions to prevent catastrophic forgetting, while others explore offline regularization (Offline Regularised[4]) or efficient prior incorporation (Efficient Priors[3]) to reduce the need for extensive online rollouts. TROLL[0] situates itself squarely in this trust region

tradition, emphasizing controlled policy updates akin to the principles underlying Remax[5] and the practical insights from Secrets PPO[8]. Compared to Nested-ReFT[37], which explores nested representations for parameter-efficient tuning, TROLL[0] focuses more directly on the optimization dynamics that govern how far a policy can safely deviate from its initialization. This positioning reflects an ongoing debate in the community: whether the key to effective LLM fine-tuning lies in algorithmic safeguards during optimization or in architectural choices that constrain the hypothesis space from the outset.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models

Authors: Li, Ziniu, Ziniu Li, Xu Tian, Tian Xu, et al. (16 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Reinforcement Learning from Human Feedback (RLHF) is key to aligning Large Language Models (LLMs), typically paired with the Proximal Policy Optimization (PPO) algorithm. While PPO is a powerful method designed for general reinforcement learning tasks, it is overly sophisticated for LLMs, leading to laborious hyper-parameter tuning and significant computation burdens. To make RLHF efficient, we present ReMax, which leverages 3 properties of RLHF: fast simulation, deterministic transitions, and t...

Relationship Analysis

Both papers belong to the Policy Optimization and Trust Region Methods category, focusing on alternatives to standard PPO for LLM fine-tuning. While TROLL replaces PPO's clipping mechanism with a differentiable trust region projection that enforces token-level KL constraints between successive policies, ReMax simplifies RLHF by building on REINFORCE without requiring a value model and exploits properties like fast simulation and trajectory-level rewards. The key difference is that TROLL maintains PPO's importance sampling framework but improves the trust region enforcement, whereas ReMax abandons PPO's architecture entirely for a simpler REINFORCE-based approach.

---

2. Secrets of rlhf in large language models part i: Ppo

Authors: Zheng Rui, Dou, Shihan, Rui Zheng, Gao Song-yang, et al. (67 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Large language models (LLMs) have formulated a blueprint for the advancement of artificial general intelligence. Its primary objective is to function as a human-centric (helpful, honest, and harmless) assistant. Alignment with humans assumes paramount significance, and reinforcement learning with human feedback (RLHF) emerges as the pivotal technological paradigm underpinning this pursuit. Current technical routes usually include \textbf{reward models} to measure human preferences, \textbf{Proxi...

Relationship Analysis

Both papers belong to the Policy Optimization and Trust Region Methods category, focusing on improving PPO-based algorithms for LLM fine-tuning. They overlap in addressing the limitations of PPO's clipping mechanism and exploring trust region constraints to stabilize policy updates during RLHF. The key difference is that the original paper (TROLL) proposes a novel differentiable trust region projection that directly enforces token-level KL constraints as a replacement for clipping, while the candidate paper provides an empirical analysis of PPO's inner workings and introduces PPO-max, which refines existing PPO implementations through careful calibration of policy constraints rather than replacing the clipping mechanism entirely.

---

3. Nested-ReFT: Efficient Reinforcement Learning for Large Language Model Fine-Tuning via Off-Policy Rollouts

Authors: Cui, Yufei, M. Heuillet, Chen, Boxing, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Advanced reasoning in LLMs on challenging domains like mathematical reasoning can be tackled using verifiable rewards based reinforced fine-tuning (ReFT). In standard ReFT frameworks, a behavior model generates multiple completions with answers per problem, for the answer to be then scored by a reward function. While such RL post-training methods demonstrate significant performance improvements across challenging reasoning domains, the computational cost of generating completions during training...

Relationship Analysis

Both papers belong to the Policy Optimization and Trust Region Methods category, addressing how to improve policy updates during RL-based LLM fine-tuning beyond standard PPO clipping. While TROLL focuses on replacing PPO's clipping mechanism with a principled KL-based trust region projection that enforces token-level constraints, Nested-ReFT addresses computational efficiency by using nested behavior models (via layer skipping) for off-policy rollouts and explores variance reduction techniques for importance sampling. The key difference is that TROLL targets the policy update mechanism itself (projection vs. clipping), whereas Nested-ReFT optimizes the rollout generation process while maintaining standard importance-weighted updates.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: TROLL: differentiable trust region projection for discrete distributions

Description: The authors introduce TROLL, a method that replaces PPO-style clipping with a fully differentiable trust region projection. This projection enforces per-token KL divergence constraints between successive policies by solving a convex optimization problem, providing a more principled alternative to heuristic clipping.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. PPO, GAE, and KL Control for RLHF in Large Language Models: A Mathematical Reference

URL: View paper

Brief Assessment

PPO GAE KL[60] is a mathematical reference document covering standard PPO, TRPO, and KL control methods. It does not present novel differentiable trust region projections for discrete distributions or token-level KL constraints as proposed in TROLL.

---

2. Multi-Agent Constrained Policy Optimization for Conflict-Free Management of Connected Autonomous Vehicles at Unsignalized Intersections

URL: View paper

Brief Assessment

Multi-Agent Constrained[59] focuses on safe multi-agent reinforcement learning for autonomous vehicle coordination at intersections, using KL divergence constraints in a multi-agent setting. The original paper addresses trust region projections for discrete token distributions in large language models, which is a fundamentally different application domain and technical approach.

---

3. Self-alignment of large video language models with refined regularized preference optimization

URL: View paper

Brief Assessment

Self-alignment Video[53] focuses on preference optimization for video language models using sub-sequence-level rewards and token-wise KL regularization, not on differentiable trust region projections with per-token KL constraints for general RL policy optimization.

---

4. Adaptive Cruise Control Based on Safe Deep Reinforcement Learning

URL: View paper

Brief Assessment

Adaptive Cruise[56] focuses on continuous control for autonomous vehicles using trust regions with KL divergence constraints in a safety-constrained MDP setting, not on discrete token-level distributions for language model optimization.

---

5. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping

URL: View paper

Brief Assessment

ARES[52] focuses on adaptive reasoning via entropy-based exploration control for multimodal models, not on trust region projections for policy optimization. The technical approaches are fundamentally different.

---

6. Reinforcement Learning based Hovering Control of a Buoyancy Driven Unmanned Underwater Vehicle with Discrete Inputs

URL: View paper

Brief Assessment

Buoyancy Hovering[55] applies standard TRPO with KL divergence constraints to underwater vehicle control with discrete actions, not a novel differentiable trust region projection method for token-level distributions in language models.

Contribution 2: Sparsification scheme for scaling to large vocabularies

Description: The authors develop a sparsification approach that retains only the most probable tokens (typically 5-10 tokens capturing over 99.999% probability mass), making the trust region projection computationally feasible for modern LLMs with vocabularies exceeding 100,000 entries while maintaining theoretical guarantees.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A multilevel proximal trust-region method for nonsmooth optimization with applications

URL: View paper

Brief Assessment

Multilevel Proximal[73] addresses sparsification in the context of multilevel optimization for nonsmooth problems, not specifically for trust region methods in large vocabulary language models. The paper focuses on PDE-constrained optimization and neural network training rather than LLM token distribution sparsification.

---

2. Recursive Gradient Perturbation through Hyperspatial Token Inversion in Large Language Models

URL: View paper

Brief Assessment

Recursive Gradient[74] uses a fixed vocabulary size of 50,000 tokens with sparsification for embedding drift, which differs from TROLL's dynamic sparsification approach that retains 5-10 tokens capturing 99.999% probability mass for vocabularies exceeding 100,000 entries while maintaining trust region projection guarantees.

Contribution 3: Empirical validation across methods, models, and tasks

Description: The authors provide comprehensive experimental evidence showing that TROLL consistently outperforms PPO-style clipping across multiple advantage estimation methods (GRPO, Dr.GRPO, GSPO, REINFORCE++), model families (Qwen, LLaMA, SmolLM, Apertus), and tasks (mathematical reasoning and code generation), achieving 3-10 percentage point improvements in success rates.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Maximum Entropy Softmax Policy Gradient via Entropy Advantage Estimation

URL: View paper

Brief Assessment

Entropy Advantage[70] focuses on maximum entropy RL with entropy regularization for policy optimization, not on trust region methods improving reward and stability across advantage estimation methods like TROLL does.

---

2. Trust region methods

URL: View paper

Brief Assessment

Trust Region Methods[68] appears to focus on general trust region optimization techniques rather than specifically validating improvements across multiple advantage estimation methods (GRPO, Dr.GRPO, GSPO, REINFORCE++), diverse model families, and LLM-specific tasks like mathematical reasoning and code generation that the original paper demonstrates.

---

3. An implicit trust region approach to behavior regularized offline reinforcement learning

URL: View paper

Brief Assessment

Implicit Trust[63] focuses on offline RL with behavior regularization through reward shaping, not online RL for LLMs. The candidate addresses different advantage estimation methods and trust regions in a fundamentally different setting (offline vs. online RL).

---

4. Separated trust regions policy optimization method

URL: View paper

Brief Assessment

Separated Trust[67] focuses on continuous control tasks in MuJoCo with Gaussian policies, not on LLM post-training with discrete token distributions across multiple advantage estimation methods and model families.

---

5. AoI-Aware Joint Spectrum and Power Allocation for Internet of Vehicles: A Trust Region Policy Optimization-Based Approach

URL: View paper

Brief Assessment

AoI-Aware[69] focuses on trust region policy optimization for Internet of Vehicles resource allocation, not on comprehensive validation across multiple advantage estimation methods, model families, and reasoning/code generation tasks as in the original paper.

---

6. Twin Trust Region Policy Optimization

URL: View paper

Brief Assessment

Twin Trust[61] focuses on continuous control tasks in classical RL environments (MuJoCo benchmarks), not LLM post-training. The paper addresses step size bounds in TRPO for policy gradient methods in robotics, whereas the original contribution validates improvements across LLM-specific advantage estimation methods (GRPO, Dr.GRPO, GSPO, REINFORCE++), model families (Qwen, LLaMA), and language tasks (math reasoning, code generation).

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] TROLL: Trust Regions Improve Reinforcement Learning for Large Language Models View paper
- [1] Revolutionizing reinforcement learning framework for diffusion large language models View paper
- [2] A technical survey of reinforcement learning techniques for large language models View paper
- [3] Efficient reinforcement learning with large language model priors View paper
- [4] Offline regularised reinforcement learning for large language models alignment View paper
- [5] Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models View paper
- [6] Reinforcement Learning for Reasoning in Large Language Models with One Training Example View paper
- [7] Inverse reinforcement learning meets large language model post-training: Basics, advances, and opportunities View paper
- [8] Secrets of rlhf in large language models part i: Ppo View paper
- [9] Multi-Objective Reinforcement Learning for Large Language Model Optimization: Visionary Perspective View paper
- [10] Fine-tuning language models for factuality View paper
- [11] Hierarchical continual reinforcement learning via large language model View paper
- [12] Reinforcement learning enhanced llms: A survey View paper
- [13] Active Preference Learning for Large Language Models View paper
- [14] Real: Efficient rlhf training of large language models with parameter reallocation View paper
- [15] OpenAssistant Conversations - Democratizing Large Language Model Alignment View paper
- [16] Secrets of rlhf in large language models part ii: Reward modeling View paper
- [17] Query Rewriting for Retrieval-Augmented Large Language Models View paper
- [18] Direct Preference Optimization: Your Language Model is Secretly a Reward Model View paper
- [19] Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning View paper
- [20] The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models View paper
- [21] Learning diverse attacks on large language models for robust red-teaming and safety tuning View paper
- [22] Adversarial Reinforcement Learning for Large Language Model Agent Safety View paper
- [23] Reinforcement learning in large language models (llms): The rise of ai language giants View paper
- [24] Robust Reinforcement Learning from Human Feedback for Large Language Models Fine-Tuning View paper
- [25] Fine-tuning large vision-language models as decision-making agents via reinforcement learning View paper
- [26] Structured preference modeling for reinforcement learning-based fine-tuning of large models View paper
- [27] Reinforcement Learning for Large Language Model Fine-Tuning: A Systematic Literature Review View paper
- [28] A Framework for Domain-Specific Dataset Creation and Adaptation of Large Language Models View paper
- [29] Self-Adapting Language Models View paper
- [30] Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents View paper
- [31] On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting View paper
- [32] Reinforcement learning: Advanced techniques for llm behavior optimization View paper
- [33] Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue View paper
- [34] Code Security Vulnerability Repair Using Reinforcement Learning with Large Language Models View paper
- [35] Rlhf workflow: From reward modeling to online rlhf View paper
- [36] Reflexion: language agents with verbal reinforcement learning View paper
- [37] Nested-ReFT: Efficient Reinforcement Learning for Large Language Model Fine-Tuning via Off-Policy Rollouts View paper
- [38] Reinforcement Learning for Prompt Optimization in Language Models: A Comprehensive Survey of Methods, Representations, and Evaluation Challenges View paper
- [39] Teaching large language models to reason with reinforcement learning View paper
- [40] Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback View paper
- [41] Grounding large language models in interactive environments with online reinforcement learning View paper
- [42] Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning View paper
- [43] MetaEvo-Rec: Self-Evolving Meta-Reinforcement Learning Recommendation with Large-Language-Model Guided Policy Adaptation View paper

- [44] Openrlhf: An easy-to-use, scalable and high-performance rlhf framework View paper
- [45] Adaptive Reinforcement Learning Framework for Enterprise Data Integration in LLM Training View paper
- [46] Group Fairness in Reinforcement Learning and Large Language Models View paper
- [47] Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning View paper
- [48] Mixture-of-Skills: Learning to Optimize Data Usage for Fine-Tuning Large Language Models View paper
- [49] Refining the giants: A comprehensive review of fine-tuning strategies for large language models View paper
- [50] MAGELLAN: Metacognitive predictions of learning progress guide autotelic LLM agents in large goal spaces View paper
- [51] Token-level direct preference optimization View paper
- [52] Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping View paper
- [53] Self-alignment of large video language models with refined regularized preference optimization View paper
- [54] The Path Not Taken: RLVR Provably Learns Off the Principals View paper
- [55] Reinforcement Learning based Hovering Control of a Buoyancy Driven Unmanned Underwater Vehicle with Discrete Inputs View paper
- [56] Adaptive Cruise Control Based on Safe Deep Reinforcement Learning View paper
- [57] Pairwise Proximal Policy Optimization: Large Language Models Alignment via Comparative RL View paper
- [58] Latent Space Exploration and Trajectory Space Update in Temporally-Correlated Episodic Reinforcement Learning View paper
- [59] Multi-Agent Constrained Policy Optimization for Conflict-Free Management of Connected Autonomous Vehicles at Unsignalized Intersections View paper
- [60] PPO, GAE, and KL Control for RLHF in Large Language Models: A Mathematical Reference View paper
- [61] Twin Trust Region Policy Optimization View paper
- [62] Maximum Entropy On-Policy Actor-Critic via Entropy Advantage Estimation View paper
- [63] An implicit trust region approach to behavior regularized offline reinforcement learning View paper
- [64] Risk-averse trust region optimization for reward-volatility reduction View paper
- [65] High-Dimensional Continuous Control Using Generalized Advantage Estimation View paper
- [66] Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation View paper
- [67] Separated trust regions policy optimization method View paper
- [68] Trust region methods View paper
- [69] AoI-Aware Joint Spectrum and Power Allocation for Internet of Vehicles: A Trust Region Policy Optimization-Based Approach View paper
- [70] Maximum Entropy Softmax Policy Gradient via Entropy Advantage Estimation View paper
- [71] A fast, performant, secure distributed training framework for large language model View paper
- [72] Enhancing multiple dimensions of trustworthiness in LLMs via sparse activation control View paper
- [73] A multilevel proximal trust-region method for nonsmooth optimization with applications View paper
- [74] Recursive Gradient Perturbation through Hyperspatial Token Inversion in Large Language Models View paper