Novelty Assessment Report

Paper: Policy Likelihood-based Query Sampling and Critic-Exploited Reset for Efficient Preference-based Reinforcement Learning

PDF URL: https://openreview.net/pdf?id=ITeuGb2bYg

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Preference-based reinforcement learning (PbRL) enables agent training without explicit reward design by leveraging human feedback. Although various query sampling strategies have been proposed to improve feedback efficiency, many fail to enhance performance because they select queries from outdated experiences with low likelihood under the current policy. Such queries may no longer represent the agent's evolving behavior patterns, reducing the informativeness of human feedback. To address this issue, we propose a policy likelihood-based query sampling and critic-exploited reset (PoLiCER). Our approach uses policy likelihood-based query sampling to ensure that queries remain aligned with the agent’s evolving behavior. However, relying solely on policy-aligned sampling can result in overly localized guidance, leading to overestimation bias, as the model tends to overfit to early feedback experiences. To mitigate this, PoLiCER incorporates a dynamic resetting mechanism that selectively resets the reward estimator and its associated Q-function based on critic outputs. Experimental evaluation across diverse locomotion and robotic manipulation tasks demonstrates that PoLiCER consistently outperforms existing PbRL methods.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Efficient Preference-Based Reinforcement Learning with Human Feedback**

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:

- **Reward Model Learning and Estimation**
- **Policy Optimization and Training Algorithms**
- **Feedback Collection and Query Strategies**
- **Personalization and Multi-Objective Alignment**
- **Foundations and Survey Literature**
- **Domain-Specific Applications and Extensions**

Complete Taxonomy Tree

- Efficient Preference-Based Reinforcement Learning with Human Feedback Survey Taxonomy
- Reward Model Learning and Estimation
  - Robust Reward Learning from Noisy Preferences (5 papers)
  - [2] Adaptive preference scaling for reinforcement learning with human feedback (Hong, 2024) View paper
  - [4] Rime: Robust preference-based reinforcement learning with noisy preferences (Cheng Jie, 2024) View paper
  - [26] Adaptive Confidence-aware Preference-based Reinforcement Learning with Noisy Feedback (Yuhao Gong, 2025) View paper
  - [32] Mixing corrupted preferences for robust and feedback-efficient preference-based reinforcement learning (Jongkook Heo, 2025) View paper
  - [40] Robust Reinforcement Learning from Corrupted Human Feedback (Bukharin, 2024) View paper
  - Dynamics-Aware and Model-Based Reward Learning (2 papers)
  - [5] Sample-efficient preference-based reinforcement learning with dynamics aware rewards (Metcalf, 2024) View paper
  - [7] Efficient preference-based reinforcement learning using learned dynamics models (Yi Liu, 2023) View paper
  - Theoretical Foundations of Preference-Based Reward Learning (2 papers)
  - [11] Principled reinforcement learning with human feedback from pairwise or k-wise comparisons (Zhu, 2023) View paper
  - [27] Online iterative reinforcement learning from human feedback with general preference model (Hanze Dong, 2024) View paper
  - Alternative Feedback Modalities for Reward Learning (5 papers)
  - [12] Rlaif: Scaling reinforcement learning from human feedback with ai feedback (H Lee, 2023) View paper
  - [21] Online Preference-based Reinforcement Learning with Self-augmented Feedback from Large Language Model (Tu Songjun, 2024) View paper
  - [22] RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback (Lee, 2023) View paper
  - [35] RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback (Wang Yu-Fei, 2024) View paper
  - [43] A Self-Supervised Reinforcement Learning Approach for Fine-Tuning Large Language Models Using Cross-Attention Signals (Kiruluta, 2025) View paper
- Policy Optimization and Training Algorithms
  - Direct Preference Optimization Without Reward Models (3 papers)
  - [19] Contrastive preference learning: learning from human feedback without rl (Hejna, 2023) View paper
  - [42] Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model (Kai Yang, 2023) View paper
  - [48] Mixed Preference Optimization: Reinforcement Learning with Data Selection and Better Reference Model (Gou Qi, 2024) View paper

  - Reinforcement Learning-Based Policy Training (3 papers)
  - [3] A minimaximalist approach to reinforcement learning from human feedback (Swamy, 2024) View paper
  - [33] Nash learning from human feedback (Munos, 2024) View paper
  - [47] MA-RLHF: Reinforcement Learning from Human Feedback with Macro Actions (Chai, 2024) View paper
  - Parameter-Efficient and Scalable Training (3 papers)
  - [9] Training a helpful and harmless assistant with reinforcement learning from human feedback (Bai Yuntao, 2022) View paper
  - [10] Parameter efficient reinforcement learning from human feedback (Sidahmed, 2024) View paper
  - [44] Rlhf workflow: From reward modeling to online rlhf (Dong, 2024) View paper
  - Risk-Aware and Safe Policy Learning (2 papers)
  - [1] Safe rlhf: Safe reinforcement learning from human feedback (Josef Dai, 2023) View paper
  - [25] Ra-pbrl: Provably efficient risk-aware preference-based reinforcement learning (Zhao Yu-jie, 2024) View paper
- Feedback Collection and Query Strategies
  - Active Query Selection and Sampling ★ (4 papers)
  - [0] Policy Likelihood-based Query Sampling and Critic-Exploited Reset for Efficient Preference-based Reinforcement Learning (Anon et al., 2026) View paper
  - [14] Sequential preference ranking for efficient reinforcement learning from human feedback (M Hwang, 2023) View paper
  - [34] Towards Efficient Online Exploration for Reinforcement Learning with Human Feedback (Li Gen, 2025) View paper
  - [41] Reward uncertainty for exploration in preference-based reinforcement learning (Liang Xinran, 2022) View paper
  - Label Smoothing and Experience Alignment (3 papers)
  - [6] Efficient preference-based reinforcement learning via aligned experience estimation (Bai, 2024) View paper
  - [28] LEASE: Offline Preference-based Reinforcement Learning with High Sample Efficiency (Liu Xiao-yin, 2024) View paper
  - [39] Learning state importance for preference-based reinforcement learning (Guoxi Zhang, 2024) View paper
- Personalization and Multi-Objective Alignment
  - Pluralistic and Personalized Preference Modeling (3 papers)
  - [13] Multi-turn reinforcement learning with preference human feedback (Daniele Calandriello, 2024) View paper
  - [17] Personalizing reinforcement learning from human feedback with variational preference learning (Poddar, 2024) View paper
  - [45] Strategyproof Reinforcement Learning from Human Feedback (Buening, 2025) View paper
  - Multi-Objective Preference Learning (2 papers)
  - [24] Preference-based multi-objective reinforcement learning (Ni Mu, 2025) View paper
  - [31] Structured preference modeling for reinforcement learning-based fine-tuning of large models (Zhu Lin, 2025) View paper
- Foundations and Survey Literature
  - Survey and Review Papers (3 papers)
  - [8] A survey of reinforcement learning from human feedback (Timo Kaufmann, 2024) View paper
  - [15] Advances in preference-based reinforcement learning: A review (Youssef Abdelkareem, 2022) View paper
  - [30] Human-centered reinforcement learning: A survey (Guangliang Li, 2019) View paper
  - Critical Analysis and Foundational Studies (3 papers)
  - [16] Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms (Shreyas Chaudhari, 2025) View paper
  - [20] Deep reinforcement learning from human preferences (Christiano, 2017) View paper
  - [23] Efficient Preference-based Reinforcement Learning (Wirth, 2025) View paper
- Domain-Specific Applications and Extensions
  - Large Language Model Alignment (2 papers)
  - [18] Transforming human interactions with AI via reinforcement learning with human feedback (RLHF) (Liu, 2023) View paper
  - [38] BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset (Ji, 2023) View paper
  - Robotics and Autonomous Systems (1 papers)
  - [36] Preference-Based Reinforcement Learning Framework for Autonomous Vehicles (Dake Ding, 2024) View paper
  - Combinatorial and Scheduling Problems (2 papers)
  - [49] Large Language Model-Assisted Deep Reinforcement Learning from Human Feedback for Job Shop Scheduling (Yuhang Zeng, 2025) View paper
  - [50] Combinatorial Reinforcement Learning with Preference Feedback (Oh, 2025) View paper
  - Theses and Comprehensive Frameworks (3 papers)
  - [29] Efficient and Robust Reinforcement Learning from Human Feedback (Huazheng Wang, 2025) View paper
  - [37] Towards safe, aligned, and efficient reinforcement learning from human feedback (Daniel, 2025) View paper
  - [46] Curiosity-Driven Reinforcement Learning from Human Feedback (Sun Hao-Ran, 2025) View paper

Narrative

Core task: efficient preference-based reinforcement learning with human feedback. The field has matured into several interconnected branches that address complementary challenges in learning from human preferences. Reward Model Learning and Estimation focuses on building accurate representations of human values from comparison data, often grappling with noise and distributional robustness (e.g., RIME Robust Preferences[4], Minimaximalist RLHF[3]). Policy Optimization and Training Algorithms develops scalable methods for aligning agent behavior with learned reward models, including parameter-efficient approaches (Parameter Efficient RLHF[10]) and online iterative schemes (Online Iterative RLHF[27]). Feedback Collection and Query Strategies tackles the sample efficiency bottleneck by intelligently selecting which queries to pose to human annotators, encompassing active learning and exploration-driven methods (Reward Uncertainty Exploration[41], Efficient Online Exploration[34]). Personalization and Multi-Objective Alignment extends the framework to handle diverse user preferences and safety constraints (Safe RLHF[1], Multi-objective Preference RL[24]), while Foundations and Survey Literature provides theoretical grounding (RLHF Survey[8], Preference-based RL Review[15]) and Domain-Specific Applications demonstrates practical deployment across robotics, language models, and autonomous systems.

A particularly active research direction centers on reducing the number of human queries required for effective learning, balancing exploration with exploitation under uncertainty. Policy Likelihood Query Sampling[0] sits squarely within this active query selection cluster, proposing to prioritize queries based on policy likelihood to maximize information gain. This approach contrasts with uncertainty-driven methods like Reward Uncertainty Exploration[41], which explicitly model epistemic uncertainty in the reward function, and with exploration-focused strategies such as Efficient Online Exploration[34], which emphasize discovering informative state-action regions. While Sequential Preference Ranking[14] addresses query efficiency through structured ranking formats, Policy Likelihood Query Sampling[0] leverages the policy's own trajectory distribution to guide sampling. The central trade-off across these works involves computational overhead versus sample efficiency: more sophisticated query selection can dramatically reduce annotation burden but may introduce additional modeling complexity or assumptions about the underlying preference structure.

Related Works in Same Category

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

1. Sequential preference ranking for efficient reinforcement learning from human feedback

Authors: M Hwang, G Lee, H Kee, CW Kim | **Year/Venue**: 2023 | **URL**: View paper

Abstract

â□¦ from human feedback, ie, preference-based reinforcement learning, is to learn the reward â□¦ [5] present the human-in-the-loop reinforcement learning framework for robotic tasks. Lee et â□¦

Relationship Analysis

Both papers belong to the Active Query Selection and Sampling category, focusing on strategies to select informative trajectory pairs for preference-based reinforcement learning. The original paper (PoLiCER) proposes policy likelihood-based query sampling to select queries aligned with the current policy's behavior, while the candidate paper (SeqRank) introduces sequential preference ranking that iteratively compares trajectories to augment preference data through transitivity. The key difference is that PoLiCER focuses on selecting policy-relevant queries from the replay buffer using log-likelihoods, whereas SeqRank emphasizes feedback efficiency by constructing preference rankings through sequential comparisons to generate multiple preference labels from fewer human feedbacks.

2. Towards Efficient Online Exploration for Reinforcement Learning with Human Feedback

Authors: Li Gen, Yan Yu-ling, Gen Li, Yuling Yan | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reinforcement learning with human feedback (RLHF), which learns a reward model from human preference data and then optimizes a policy to favor preferred responses, has emerged as a central paradigm for aligning large language models (LLMs) with human preferences. In this paper, we investigate exploration principles for online RLHF, where one seeks to adaptively collect new preference data to refine both the reward model and the policy in a data-efficient manner. By examining existing optimism-ba...

Relationship Analysis

Both papers belong to the Active Query Selection and Sampling category, focusing on strategies to select informative trajectory pairs for preference-based reinforcement learning. The original paper (PoLiCER) proposes policy likelihood-based query sampling to select queries aligned with the current policy's behavior in continuous control tasks, while the candidate paper addresses online exploration in RLHF by proposing uncertainty-based sampling that compares actions from consecutive policies to reduce reward estimation uncertainty. The key difference is that PoLiCER focuses on policy likelihood as a relevance metric for query selection in offline preference datasets, whereas the candidate paper develops an adaptive online exploration scheme with theoretical regret bounds for sequential preference data collection in bandit settings.

3. Reward uncertainty for exploration in preference-based reinforcement learning

Authors: Liang Xinran, Xinran Liang, Shu, Katherine, Katherine Shu, et al. (13 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Conveying complex objectives to reinforcement learning (RL) agents often requires meticulous reward engineering. Preference-based RL methods are able to learn a more flexible reward model based on human preferences by actively incorporating human feedback, i.e. teacher's preferences between two clips of behaviors. However, poor feedback-efficiency still remains a problem in current preference-based RL algorithms, as tailored human feedback is very expensive. To handle this issue, previous method...

Relationship Analysis

Both papers belong to the Active Query Selection and Sampling category, focusing on strategies to select informative trajectory pairs for preference-based RL. While the original paper (PoLiCER) proposes policy likelihood-based query sampling to ensure queries align with the current policy's behavior, the candidate paper (RUNE) takes a different approach by using reward uncertainty (disagreement across ensemble reward models) as an exploration bonus rather than as a query selection criterion. The key distinction is that PoLiCER directly addresses query-policy misalignment through likelihood-based sampling, whereas RUNE focuses on exploration-driven sample efficiency through intrinsic rewards derived from reward model uncertainty.

Contributions Analysis

Overall novelty summary. The paper proposes PoLiCER, combining policy likelihood-based query sampling with a critic-exploited reset mechanism to improve feedback efficiency in preference-based reinforcement learning. It resides in the 'Active Query Selection and Sampling' leaf, which contains four papers total including this one. This leaf sits within the broader 'Feedback Collection and Query Strategies' branch, indicating a moderately populated research direction focused on minimizing human annotation effort. The taxonomy reveals this is an active area with multiple competing approaches to query selection, suggesting the problem space is well-explored but not saturated.

The paper's immediate neighbors include uncertainty-driven methods (Reward Uncertainty Exploration) and exploration-focused strategies (Efficient Online Exploration), both addressing query efficiency through different principles. The sibling 'Label Smoothing and Experience Alignment' leaf contains three papers tackling related overfitting issues through regularization rather than dynamic resetting. The broader 'Reward Model Learning and Estimation' branch (nine papers across four leaves) addresses complementary challenges in handling noisy preferences and model robustness, while 'Policy Optimization and Training Algorithms' (nine papers) focuses on downstream policy updates. PoLiCER bridges query selection with policy training dynamics, positioning it at the intersection of these branches.

Among eleven candidates examined, the policy likelihood-based sampling contribution shows overlap with prior work: two papers appear to provide refutable evidence from nine candidates reviewed. The critic-exploited reset mechanism examined two candidates with no clear refutations, suggesting greater novelty in this component. The combined PoLiCER framework was not directly compared against candidates. The limited search scope (eleven papers, not exhaustive) means these statistics reflect top-K semantic matches rather than comprehensive field coverage. The reset mechanism addressing overestimation bias appears less explored in the examined literature than policy-aligned sampling strategies.

Based on the limited search of eleven semantically similar papers, the work appears to offer incremental contributions in query sampling with potentially stronger novelty in the dynamic reset mechanism. The taxonomy context shows this sits in a moderately active research area with established competing approaches. The analysis cannot definitively assess novelty beyond the examined candidates, and a broader literature review would be needed to confirm whether the critic-exploited reset represents a substantive departure from existing overfitting mitigation strategies in preference-based RL.

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

Contribution 1: Policy likelihood-based query sampling (PLS)

Description: A query sampling strategy that selects trajectory pairs for human feedback based on their likelihood under the current policy rather than temporal recency. This ensures queries remain relevant to the agent's evolving behavior throughout training, addressing query-policy misalignment in preference-based reinforcement learning.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. DAPPER: Discriminability-Aware Policy-to-Policy Preference-Based Reinforcement Learning for Query-Efficient Robot Skill Acquisition

URL: View paper

Brief Assessment

DAPPER Policy-to-Policy[58] focuses on discriminability-aware sampling across multiple policies trained from scratch, not likelihood-based sampling within a single evolving policy as in the original paper's PLS approach.

2. SENIOR: Efficient Query Selection and Preference-Guided Exploration in Preference-based Reinforcement Learning

URL: View paper

Brief Assessment

SENIOR Query Selection[54] focuses on motion-distinction-based selection (MDS) using kernel density estimation of robot states and motion directions, not policy likelihood. The candidate's query selection evaluates physical motion characteristics rather than policy alignment through likelihood computation.

3. S-EPOA: Overcoming the Indistinguishability of Segments with Skill-Driven Preference-Based Reinforcement Learning

URL: View paper

Brief Assessment

Skill-Driven Segments[56] addresses segment indistinguishability through skill-based query selection mechanisms, not policy likelihood-based sampling. The candidate focuses on balancing information gain and distinguishability over learned skill spaces, which is technically distinct from selecting queries based on their likelihood under the current policy.

4. S-EPOA: Overcoming the Indivisibility of Annotations with Skill-Driven Preference-Based Reinforcement Learning

URL: View paper

Brief Assessment

Skill-Driven Annotations[57] addresses segment indistinguishability through skill-based query selection that balances information gain and distinguishability in skill space, rather than selecting queries based on policy likelihood under the current policy as in the original paper's PLS approach.

5. Query-Policy Misalignment in Preference-Based Reinforcement Learning

URL: View paper

Prior Art Analysis

Query-Policy Misalignment[60] demonstrates that policy-aligned query selection was proposed and implemented prior to the original paper's PLS contribution. Both methods address the same core problem of query-policy misalignment by selecting queries based on their relevance to the current policy rather than temporal recency. Query-Policy Misalignment[60] explicitly introduces 'policy-aligned query selection' that ensures queries align with the on-policy distribution by selecting from recent trajectories, which is fundamentally the same approach as PLS but uses recency as a practical proxy for policy likelihood rather than computing explicit likelihood values.

Evidence

Evidence 1 - **Rationale**: Both papers propose methods to align query selection with the current policy distribution to address query-policy misalignment, demonstrating prior work on this concept. - **Original**: we propose a policy likelihood-based query sampling and critic-exploited reset (policer). our approach uses policy likelihood-based query sampling to ensure that queries remain aligned with the agent's evolving behavior. - **Candidate**: To address this issue, we propose an elegant method: qpa (query-policy alignment). The key idea of qpa is that rather than learning the reward model across the entire global state-action space inadequately as existing methods do, it is more effective to focus on learning the reward model precisely w...

Evidence 2 - **Rationale**: Query-Policy Misalignment[60] explicitly introduces policy-aligned query selection that ensures queries align with the current policy's distribution, addressing the same problem as PLS before the original paper. - **Original**: to address this issue, we propose policy likelihood-based query sampling (pls), which selects queries based on their likelihood under the current policy rather than temporal recency. our method ensures that selected queries remain relevant throughout training by continuously adapting to the evolving... - **Candidate**: In contrast to existing query selection schemes, we highlight that the segment query selection should be aligned with the on-policy distribution. In particular, it is crucial to ensure that the pairs of segments (σ0, σ1) selected for preference queries stay close to the current policy's visitation d...

Evidence 3 - **Rationale**: Query-Policy Misalignment[60] implements policy-aligned query selection by selecting from recent trajectories, which ensures queries remain relevant to the evolving policy, the same goal as PLS. - **Original**: our method ensures that selected queries remain relevant throughout training by continuously adapting to the evolving policy. - **Candidate**: a simple yet effective way to perform policy-aligned query selection is to choose (σ0, σ1) from the most recent trajectories stored in d. for the sake of clarity, we refer to the buffer that stores the most recent trajectories as the policy-aligned bufferdpa.

Evidence 4 - **Rationale**: Both papers identify and analyze the same fundamental problem: existing query selection methods fail to align with the current policy's distribution, leading to inefficient feedback. - **Original**: figure 1: policy likelihoods of replay buffer episodes in meta-world sweep into (10,000/50), categorized into all past episodes (all), recent 30 episodes (last 30), and top 30 episodes into current policy likelihood (top 30). temporal recency becomes an increasingly poor indicator of policy relevanc... - **Candidate**: we observe that the selected segments of existing query selection schemes typically fall outside the scope of the visitation distribution dπ (marked with green circles). we refer to this phenomenon as query-policy misalignment, as illustrated in figure 3. such misalignment wastes valuable feedback b...

Evidence 5 - **Rationale**: The original paper cites Query-Policy Misalignment[60] (hu et al. 2024) as prior work that introduced policy-aligned sampling, demonstrating this is not a novel contribution. - **Original**: hu et al. (2024) argued that query-policy misalignment occurs when queries are selected from longpast experiences that no longer represent the current policy. this reliance on outdated information diminishes feedback effectiveness and limits its ability to guide policy updates. to address this issue... - **Candidate**: By contrast,policy-aligned selection selects fresh segments that are recently visited by the current rl policy, enabling timely feedback on the current status of the policy and leading to significant performance gain, as shown in figure 2(c).

6. VARIQuery: VAE Segment-Based Active Learning for Query Selection in Preference-Based Reinforcement Learning

URL: View paper

Brief Assessment

VARIQuery VAE Segments[59] focuses on VAE-based query selection for diversity and reward model uncertainty in preference-based RL, not on policy likelihood-based sampling strategies.

7. Improving Reward Models with Proximal Policy Exploration for Preference-Based Reinforcement Learning

URL: View paper

Brief Assessment

Proximal Policy Exploration[61] focuses on expanding preference buffer coverage through out-of-distribution exploration and mixture distribution queries, not on policy likelihood-based query sampling. The candidate's query selection is based on in-distribution vs. out-of-distribution metrics using Morse neural networks, which is fundamentally different from selecting queries based on their likelihood under the current policy.

8. DUO: Diverse, Uncertain, On-Policy Query Generation and Selection for Reinforcement Learning from Human Feedback

URL: View paper

Prior Art Analysis

DUO Query Generation[53] demonstrates that similar policy-likelihood-based query sampling approaches existed prior to the original paper. Both papers address query-policy misalignment by prioritizing queries based on their likelihood under the current policy rather than temporal recency. DUO Query Generation[53] explicitly computes trajectory log-likelihood under the current policy (equation 4) and uses this for prioritized sampling (equation 5), which is conceptually identical to the original paper's PLS approach. The candidate paper was published before the original submission and presents the same core insight: using policy likelihood to ensure queries remain aligned with the agent's evolving behavior.

Evidence

Evidence 1 - **Rationale**: Both papers propose using policy likelihood for query sampling to ensure relevance to the current policy, addressing the same query-policy misalignment problem. - **Original**: to overcome this limitation, we propose policy likelihood-based query sampling (pls), which selects queries based on their likelihood under the current policy rather than temporal recency. our method ensures that selected queries remain relevant throughout training by continuously adapting to the ev... - **Candidate**: to compare trajectories that are more relevant to the current policy, duo uses prioritized sampling to favor on-policy trajectories during query generation. to accurately capture the query uncertainty, duo evaluates the epistemic uncertainty of a query in a principled way

Evidence 2 - **Rationale**: DUO Query Generation[53] explicitly implements priority sampling based on log probability under current policy π, which is the same technical approach as PLS. - **Original**: we propose policy likelihood-based query sampling (pls), which selects queries based on their likelihood under the current policy rather than temporal recency. - **Candidate**: instead, we propose to implement ξo(b) via priority sampling over all trajectories (generated by π and past policies) in the replay buffer b, favoring trajectories that have a higher probability of being generated by π. formally, the priority is derived from the following on-policiness measure o(τ) ...

Evidence 3 - **Rationale**: Both papers use nearly identical mathematical formulations for computing trajectory log-likelihood under the current policy, demonstrating the same technical implementation. - **Original**: at each feedback session, we begin by uniformly sampling 2 x l x k trajectories, where k is the number of queries to extract and l is a scaling factor, following prior weighted sampling approaches (christiano et al., 2017; lee et al., 2021a;b). we then compute the average log-likelihood of each traj... - **Candidate**: formally, the priority is derived from the following on-policiness measure o(τ) = t-1x t=0 log π (at | st) , (4) where τ = (s0, a0, s1, . . . , st) is a t-length trajectory. this measure computes the log probability of a trajectory being generated by π, ignoring the unknown transition probabilities...

9. Dueling Posterior Sampling for Preference-Based Reinforcement Learning

URL: View paper

Brief Assessment

Dueling Posterior Sampling[55] addresses preference-based RL through posterior sampling over dynamics and rewards, using trajectory-level preferences for Bayesian inference. It does not propose query sampling strategies based on policy likelihood; instead, it samples two policies from posteriors to generate trajectory pairs for comparison, which is fundamentally different from selecting queries based on their likelihood under the current policy.

Contribution 2: Critic-exploited reset (CER)

Description: A dynamic resetting mechanism that monitors critic outputs to detect reward overestimation and strategically resets both the reward estimator and Q-function when necessary. This approach mitigates primacy bias while maintaining computational efficiency through adaptive thresholding.

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. Mad-td: Model-augmented data stabilizes high update ratio rl

URL: View paper

Brief Assessment

Model-augmented Data Stabilizes[51] addresses value function misgeneralization in high update-to-data ratio RL through model-generated on-policy data, not through dynamic resetting mechanisms. The candidate focuses on stabilizing training by augmenting off-policy data with synthetic transitions from learned world models, which is a fundamentally different approach from monitoring critic outputs to trigger adaptive resets of reward estimators and Q-functions.

2. DARLR: Dual-Agent Offline Reinforcement Learning for Recommender Systems with Dynamic Reward

URL: View paper

Brief Assessment

Dual-Agent Dynamic Reward[52] focuses on dynamically updating reward functions in recommender systems through user selection and aggregation, not on resetting mechanisms based on critic outputs to mitigate overestimation bias in preference-based RL.

Contribution 3: PoLiCER framework combining PLS and CER

Description: An integrated framework that combines policy likelihood-based query sampling with critic-exploited reset to address both query-policy misalignment and primacy bias in preference-based reinforcement learning. The two components work synergistically to improve feedback efficiency and learning stability.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

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

References

- [0] Policy Likelihood-based Query Sampling and Critic-Exploited Reset for Efficient Preference-based Reinforcement Learning View paper
- [1] Safe rlhf: Safe reinforcement learning from human feedback View paper
- [2] Adaptive preference scaling for reinforcement learning with human feedback View paper
- [3] A minimaximalist approach to reinforcement learning from human feedback View paper
- [4] Rime: Robust preference-based reinforcement learning with noisy preferences View paper
- [5] Sample-efficient preference-based reinforcement learning with dynamics aware rewards View paper
- [6] Efficient preference-based reinforcement learning via aligned experience estimation View paper
- [7] Efficient preference-based reinforcement learning using learned dynamics models View paper
- [8] A survey of reinforcement learning from human feedback View paper
- [9] Training a helpful and harmless assistant with reinforcement learning from human feedback View paper
- [10] Parameter efficient reinforcement learning from human feedback View paper
- [11] Principled reinforcement learning with human feedback from pairwise or k-wise comparisons View paper
- [12] Rlaif: Scaling reinforcement learning from human feedback with ai feedback View paper
- [13] Multi-turn reinforcement learning with preference human feedback View paper
- [14] Sequential preference ranking for efficient reinforcement learning from human feedback View paper
- [15] Advances in preference-based reinforcement learning: A review View paper
- [16] Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms View paper
- [17] Personalizing reinforcement learning from human feedback with variational preference learning View paper
- [18] Transforming human interactions with AI via reinforcement learning with human feedback (RLHF) View paper
- [19] Contrastive preference learning: learning from human feedback without rl View paper
- [20] Deep reinforcement learning from human preferences View paper
- [21] Online Preference-based Reinforcement Learning with Self-augmented Feedback from Large Language Model View paper
- [22] RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback View paper
- [23] Efficient Preference-based Reinforcement Learning View paper
- [24] Preference-based multi-objective reinforcement learning View paper
- [25] Ra-pbrl: Provably efficient risk-aware preference-based reinforcement learning View paper
- [26] Adaptive Confidence-aware Preference-based Reinforcement Learning with Noisy Feedback View paper
- [27] Online iterative reinforcement learning from human feedback with general preference model View paper
- [28] LEASE: Offline Preference-based Reinforcement Learning with High Sample Efficiency View paper
- [29] Efficient and Robust Reinforcement Learning from Human Feedback View paper
- [30] Human-centered reinforcement learning: A survey View paper
- [31] Structured preference modeling for reinforcement learning-based fine-tuning of large models View paper
- [32] Mixing corrupted preferences for robust and feedback-efficient preference-based reinforcement learning View paper
- [33] Nash learning from human feedback View paper
- [34] Towards Efficient Online Exploration for Reinforcement Learning with Human Feedback View paper
- [35] RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback View paper
- [36] Preference-Based Reinforcement Learning Framework for Autonomous Vehicles View paper
- [37] Towards safe, aligned, and efficient reinforcement learning from human feedback View paper
- [38] BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset View paper
- [39] Learning state importance for preference-based reinforcement learning View paper
- [40] Robust Reinforcement Learning from Corrupted Human Feedback View paper
- [41] Reward uncertainty for exploration in preference-based reinforcement learning View paper
- [42] Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model View paper
- [43] A Self-Supervised Reinforcement Learning Approach for Fine-Tuning Large Language Models Using Cross-Attention Signals View paper
- [44] Rlhf workflow: From reward modeling to online rlhf View paper
- [45] Strategyproof Reinforcement Learning from Human Feedback View paper
- [46] Curiosity-Driven Reinforcement Learning from Human Feedback View paper
- [47] MA-RLHF: Reinforcement Learning from Human Feedback with Macro Actions View paper
- [48] Mixed Preference Optimization: Reinforcement Learning with Data Selection and Better Reference Model View paper
- [49] Large Language Model-Assisted Deep Reinforcement Learning from Human Feedback for Job Shop Scheduling View paper
- [50] Combinatorial Reinforcement Learning with Preference Feedback View paper
- [51] Mad-td: Model-augmented data stabilizes high update ratio rl View paper
- [52] DARLR: Dual-Agent Offline Reinforcement Learning for Recommender Systems with Dynamic Reward View paper
- [53] DUO: Diverse, Uncertain, On-Policy Query Generation and Selection for Reinforcement Learning from Human Feedback View paper
- [54] SENIOR: Efficient Query Selection and Preference-Guided Exploration in Preference-based Reinforcement Learning View paper
- [55] Dueling Posterior Sampling for Preference-Based Reinforcement Learning View paper
- [56] S-EPOA: Overcoming the Indistinguishability of Segments with Skill-Driven Preference-Based Reinforcement Learning View paper
- [57] S-EPOA: Overcoming the Indivisibility of Annotations with Skill-Driven Preference-Based Reinforcement Learning View paper
- [58] DAPPER: Discriminability-Aware Policy-to-Policy Preference-Based Reinforcement Learning for Query-Efficient Robot Skill Acquisition View paper
- [59] VARIQuery: VAE Segment-Based Active Learning for Query Selection in Preference-Based Reinforcement Learning View paper
- [60] Query-Policy Misalignment in Preference-Based Reinforcement Learning View paper

- [61] Improving Reward Models with Proximal Policy Exploration for Preference-Based Reinforcement Learning View paper