Novelty Assessment Report

Paper: FlowRL: Matching Reward Distributions for LLM Reasoning
PDF URL: https://openreview.net/pdf?id=lObnTKbm9U
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We propose FlowRL: matching the full reward distribution via flow balancing instead of solely maximizing rewards in large language model (LLM) reinforcement learning (RL). Recent advanced reasoning models adopt reward-maximizing methods (e.g., PPO and GRPO), which tend to over-optimize dominant reward signals while neglecting less frequent but valid reasoning paths, thus reducing diversity. In contrast, we transform scalar rewards into a normalized target distribution using a learnable partition function, and then minimize the reverse KL divergence between the policy and the target distribution. We implement this idea as a flow-balanced optimization method that promotes diverse exploration and generalizable reasoning trajectories. We conduct experiments on both math and code reasoning tasks: FlowRL achieves a significant average improvement of 10.0% over GRPO and 5.1% over PPO on math benchmarks, and performs consistently better on code reasoning tasks. These results highlight reward distribution-matching as a key step toward efficient exploration and diverse reasoning in LLM reinforcement learning.

Core Task Landscape

This paper addresses: **Matching Reward Distributions for Large Language Model Reasoning**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Reward Distribution Modeling and Optimization**
- **Reward Maximization Approaches**
- **Process and Dense Reward Models**
- **Self-Rewarding and Inverse RL Frameworks**
- **Open-Ended and Domain-Specific Reasoning**
- **Reward Model Design and Theoretical Foundations**
- **Multi-Agent and Structured Reasoning Systems**
- **Inference and Sampling Techniques**
- **Alignment Frameworks and Practical Considerations**

Complete Taxonomy Tree

- Matching Reward Distributions for Large Language Model Reasoning Survey Taxonomy
- Reward Distribution Modeling and Optimization
  - Flow-Based and Distribution-Matching Methods ★ (3 papers)
  - [0] FlowRL: Matching Reward Distributions for LLM Reasoning (Anon et al., 2026) View paper
  - [16] Cal-dpo: Calibrated direct preference optimization for language model alignment (Vasant Honavar, 2024) View paper
  - [31] Enhancing reasoning for diffusion llms via distribution matching policy optimization (Zhu Yuchen, 2025) View paper
  - Adversarial and Moment-Matching Techniques (1 papers)
  - [9] Adversarial moment-matching distillation of large language models (Jia, 2024) View paper
  - Ensemble and Weighted Reward Models (3 papers)
  - [3] Routing to the expert: Efficient reward-guided ensemble of large language models (Keming Lu, 2024) View paper
  - [25] Warm: On the benefits of weight averaged reward models (Rame, 2024) View paper
  - [47] Reward-robust rlhf in llms (Yan, 2024) View paper
- Reward Maximization Approaches
  - Outcome-Based Reinforcement Learning (4 papers)
  - [2] Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? (Yue Yang, 2025) View paper
  - [5] Exploring the limit of outcome reward for learning mathematical reasoning (Lyu, 2025) View paper
  - [7] Offline reinforcement learning for llm multi-step reasoning (HUAIJIE WANG, 2025) View paper
  - [21] Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms (Liu Zihan, 2025) View paper
  - Policy Optimization Algorithms for Reasoning (4 papers)
  - [4] R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization (Zhang Jingyi, 2025) View paper
  - [13] Teaching large language models to reason with reinforcement learning (Havrilla, 2024) View paper
  - [39] Reinforcement Pre-Training (Dong, 2025) View paper
  - [42] Reinforcing multi-turn reasoning in llm agents via turn-level reward design (Wei Quan, 2025) View paper
  - Inference-Time and Test-Time Scaling (3 papers)

- [11] RFG: Test-Time Scaling for Diffusion Large Language Model Reasoning with Reward-Free Guidance (Chen Tianlang, 2025) View paper
- [12] InfAlign: Inference-aware language model alignment (Balashankar, 2024) View paper
- [45] Your Reward Function for RL is Your Best PRM for Search: Unifying RL and Search-Based TTS (Jin Can, 2025) View paper

- Process and Dense Reward Models
  - Process Reward Model Learning (4 papers)
  - [15] Reward reasoning model (Guo Jiaxin, 2025) View paper
  - [28] Learning a Dense Reasoning Reward Model from Expert Demonstration via Inverse Reinforcement Learning (Fanconi, 2025) View paper
  - [29] Boosting Policy and Process Reward Models with Monte Carlo Tree Search in Open-Domain QA (Chi-Min Chan, 2025) View paper
  - [34] Process reward models for llm agents: Practical framework and directions (Choudhury, 2025) View paper
  - Dense Reward Redistribution and Shaping (2 papers)
  - [22] R3hf: Reward redistribution for enhancing reinforcement learning from human feedback (Jiahui Li, 2024) View paper
  - [43] Align to Structure: Aligning Large Language Models with Structural Information (Kim, 2025) View paper
  - Tree Search and Reasoning Path Ensembling (1 papers)
  - [17] Ensembling large language models with process reward-guided tree search for better complex reasoning (Sungjin Park, 2025) View paper
- Self-Rewarding and Inverse RL Frameworks
  - Self-Rewarding and Meta-Rewarding Models (4 papers)
  - [1] Self-rewarding language models (Yuan, 2024) View paper
  - [14] Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge (Tianhao Wu, 2025) View paper
  - [36] Co-rewarding: Stable Self-supervised RL for Eliciting Reasoning in Large Language Models (Zhang, 2025) View paper
  - [37] Consistent Paths Lead to Truth: Self-Rewarding Reinforcement Learning for LLM Reasoning (Yao Qi, 2025) View paper
  - Inverse RL and Demonstration-Based Reward Learning (4 papers)
  - [20] Introspective reward modeling via inverse reinforcement learning for llm alignment (Zhiqiang Wang, 2025) View paper
  - [38] Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment (Alfredo GarcÃa, 2024) View paper
  - [49] From demonstrations to rewards: Alignment without explicit human preferences (Zeng Si-liang, 2025) View paper
  - [50] Learning Reward and Policy Jointly from Demonstration and Preference Improves Alignment (Li, 2024) View paper
- Open-Ended and Domain-Specific Reasoning
  - Open-Ended Reasoning Without Verifiable Rewards (1 papers)
  - [8] Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks (Xu Yifei, 2025) View paper
  - Multimodal and Vision-Language Reasoning (3 papers)
  - [24] Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding (Wang Ye, 2025) View paper
  - [35] Mixed-R1: Unified Reward Perspective For Reasoning Capability in Multimodal Large Language Models (Xu, 2025) View paper
  - [44] Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning (Li Ming, 2025) View paper
  - Domain-Specific Applications (3 papers)
  - [23] Breaking Reward Collapse: Adaptive Reinforcement for Open-ended Medical Reasoning with Enhanced Semantic Discrimination (Liu Yizhou, 2025) View paper
  - [30] Learning reward for robot skills using large language models via self-alignment (Zeng Yu-wei, 2024) View paper
  - [33] L2M-AID: Autonomous Cyber-Physical Defense by Fusing Semantic Reasoning of Large Language Models with Multi-Agent Reinforcement Learning (Preprint) (Xu Tianxiang, 2025) View paper
- Reward Model Design and Theoretical Foundations
  - Reward Model Architecture and Training (2 papers)
  - [27] Generative Reward Models (Dakota Mahan, 2024) View paper
  - [32] Simultaneous reward distillation and preference learning: Get you a language model who can do both (Nath, 2024) View paper
  - Theoretical Foundations and Risk Bounds (2 papers)
  - [19] Rethinking reward modeling in preference-based large language model alignment (H Sun, 2025) View paper
  - [26] On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization (Jiancong Xiao, 2024) View paper
  - Fairness and Bias Mitigation in Reward Models (1 papers)
  - [41] Guiding LLM decision-making with fairness reward models (Subbiah, 2025) View paper
- Multi-Agent and Structured Reasoning Systems
  - Multi-Agent Reinforcement Learning Systems (1 papers)
  - [10] Reso: A reward-driven self-organizing llm-based multi-agent system for reasoning tasks (Zhou Heng, 2025) View paper
  - Structured and Hierarchical Reasoning (2 papers)
  - [18] Interleaved Reasoning for Large Language Models via Reinforcement Learning (Qiu, 2025) View paper
  - [48] Distilling the Implicit Multi-Branch Structure in LLMs' Reasoning via Reinforcement Learning (Xu Shicheng, 2025) View paper
- Inference and Sampling Techniques (1 papers)
  - [6] Amortizing intractable inference in large language models (Hu, 2023) View paper
- Alignment Frameworks and Practical Considerations (2 papers)
  - [40] Rule based rewards for language model safety (Joshua Achiam, 2024) View paper
  - [46] Tuning for LLM alignment (Uday Kamath, 2024) View paper

Narrative

Core task: Matching reward distributions for large language model reasoning. The field addresses how to design, learn, and optimize reward signals that guide LLMs toward improved reasoning capabilities. The taxonomy reveals a rich structure spanning nine major branches. Reward Distribution Modeling and Optimization explores flow-based and distribution-matching techniques, exemplified by FlowRL[0] and Distribution Matching Policy[31], which aim to align policy outputs with target reward distributions rather than simply maximizing scalar rewards. Reward Maximization Approaches and Process and Dense Reward Models focus on outcome-based versus step-level feedback, with works like Outcome Reward Limit[5] and Dense Reasoning Reward[28] investigating the trade-offs between coarse and fine-grained supervision. Self-Rewarding and Inverse RL Frameworks, including Self-rewarding[1] and Meta-rewarding[14], enable models to generate their own training signals, while Open-Ended and Domain-Specific Reasoning branches address generalization

across tasks. Additional branches cover theoretical foundations, multi-agent systems, inference techniques, and practical alignment considerations, forming a comprehensive landscape of reward-driven reasoning research.

A particularly active line of work contrasts distribution-matching methods with traditional reward maximization. While many studies pursue direct optimization of scalar rewards, a smaller cluster emphasizes matching entire distributions to avoid reward collapse and mode-seeking behavior, as highlighted by Cal-DPO[16] and FlowRL[0]. FlowRL[0] sits squarely within the Flow-Based and Distribution-Matching Methods branch, sharing conceptual ground with Distribution Matching Policy[31] in treating reasoning as a probabilistic flow problem. Compared to outcome-focused approaches like Outcome Reward Limit[5], FlowRL[0] emphasizes aligning the generative process itself rather than merely optimizing terminal rewards. This distinction reflects broader tensions in the field: whether to rely on sparse outcome signals, dense process supervision, or distributional objectives that preserve diversity. Open questions remain about scalability, sample efficiency, and how these distribution-matching techniques interact with self-rewarding frameworks and inference-time search methods.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Cal-dpo: Calibrated direct preference optimization for language model alignment

Authors: Vasant Honavar, Mingxiao Li, Teng Xiao, Yige Yuan, Huaisheng Zhu | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ implicit reward to match the ground-truth reward through â¦ theoretical framework from a distribution matching perspective. We â¦ data in alignment for large language model? a self-imitation â¦

Relationship Analysis

Both papers belong to the Flow-Based and Distribution-Matching Methods category, employing techniques that optimize reward distributions rather than solely maximizing rewards. While FlowRL uses flow balancing and trajectory balance objectives from GFlowNets to match the full reward distribution for LLM reasoning, Cal-DPO focuses on calibrating implicit rewards in direct preference optimization by ensuring learned rewards match ground-truth reward scales through squared loss regression. The key difference is that FlowRL addresses mode collapse through flow-balanced exploration across diverse reasoning paths, whereas Cal-DPO addresses the issue of decreasing chosen response likelihoods by constraining implicit rewards to match actual reward values.

2. Enhancing reasoning for diffusion llms via distribution matching policy optimization

Authors: Zhu Yuchen, Guo Wei, Yuchen Zhu, Choi, Jaemoo, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Diffusion large language models (dLLMs) are promising alternatives to autoregressive large language models (AR-LLMs), as they potentially allow higher inference throughput. Reinforcement learning (RL) is a crucial component for dLLMs to achieve comparable performance with AR-LLMs on important tasks, such as reasoning. However, RL algorithms that are well-suited for dLLMs'unique characteristics have yet to be developed. This paper proposes Distribution Matching Policy Optimization (DMPO), a princ...

Relationship Analysis

Both papers belong to the Flow-Based and Distribution-Matching Methods category, employing distribution matching via KL divergence minimization for reward optimization in LLM reasoning. They share the core approach of matching policy distributions to reward-induced target distributions rather than pure reward maximization. However, FlowRL focuses on autoregressive LLMs using flow balancing from GFlowNets with trajectory balance objectives for long chain-of-thought reasoning, while the candidate paper (DMPO) specifically targets diffusion LLMs with cross-entropy optimization and addresses unique challenges of small batch training through weight baseline subtraction techniques.

Contributions Analysis

Overall novelty summary. The paper proposes FlowRL, a method that matches reward distributions via flow balancing rather than maximizing scalar rewards in LLM reinforcement learning. It resides in the Flow-Based and Distribution-Matching Methods leaf, which contains only three papers including this one. This is a relatively sparse research direction within the broader taxonomy of 50 papers across nine major branches, suggesting that distribution-matching approaches remain less explored compared to traditional reward maximization methods that dominate neighboring branches.

The taxonomy reveals that FlowRL sits within Reward Distribution Modeling and Optimization, adjacent to Reward Maximization Approaches containing outcome-based RL and policy optimization algorithms. The sibling papers in the same leaf (Distribution Matching Policy and one other) share the conceptual foundation of treating reasoning as a probabilistic flow problem. Neighboring branches like Process and Dense Reward Models and Self-Rewarding frameworks pursue different supervision strategies—step-level feedback versus self-generated rewards—highlighting how FlowRL's distributional objective diverges from both sparse outcome signals and dense process supervision paradigms.

Among 18 candidates examined across three contributions, the FlowRL algorithm contribution shows 2 refutable candidates out of 10 examined, while the theoretical equivalence between KL minimization and trajectory balance shows 4 refutable candidates out of 7 examined. The length normalization contribution examined only 1 candidate with no refutations. These statistics indicate that the core algorithmic and theoretical contributions face more substantial prior work overlap within the limited search scope, while the technical implementation details appear less contested. The search scale of 18 candidates suggests this analysis captures prominent related work but may not be exhaustive.

Based on the limited literature search of 18 candidates, FlowRL appears to occupy a relatively sparse research direction with meaningful but not overwhelming prior work overlap. The taxonomy structure confirms that distribution-matching methods remain a minority approach compared to scalar reward maximization, though the contribution-level statistics reveal that specific technical elements—particularly the flow balancing formulation and KL-trajectory balance equivalence—have notable precedents among the examined candidates. A broader search might uncover additional related work in adjacent optimization or probabilistic inference communities.

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

Contribution 1: FlowRL algorithm for reward distribution matching

Description: The authors introduce FlowRL, a policy optimization algorithm that shifts from reward maximization to reward distribution matching. It transforms scalar rewards into normalized target distributions using a learnable partition function and minimizes reverse KL divergence between the policy and target distribution, promoting diverse exploration and generalizable reasoning trajectories.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Transforming and combining rewards for aligning large language models

URL: View paper

Brief Assessment

Transforming Rewards[54] focuses on transforming scalar rewards using log-sigmoid functions for preference-based alignment, not on matching full reward distributions via flow balancing as in FlowRL.

---

2. R3hf: Reward redistribution for enhancing reinforcement learning from human feedback

URL: View paper

Brief Assessment

R3HF[22] focuses on token-level reward redistribution within traditional reward-maximizing RL frameworks, treating reward prediction as regression to allocate fine-grained rewards. This differs fundamentally from FlowRL's approach of transforming scalar rewards into normalized target distributions via flow balancing and minimizing reverse KL divergence for distribution matching.

---

3. Amortizing intractable inference in large language models

URL: View paper

Prior Art Analysis

Amortizing Inference[6] demonstrates prior work on using GFlowNets for distribution matching in LLMs through diversity-seeking reinforcement learning. The candidate explicitly describes using 'amortized bayesian inference' and 'diversity-seeking reinforcement learning algorithms: generative flow networks (gflownets)' to sample from intractable posterior distributions, which directly addresses the same core problem as FlowRL: moving from reward maximization to distribution matching. Both papers frame the problem as matching distributions rather than maximizing rewards, and both employ GFlowNets-based approaches to achieve this goal in language model fine-tuning contexts.

Evidence

Evidence 1 - **Rationale**: Both papers propose using GFlowNets-based approaches to shift from reward maximization to distribution matching in LLM fine-tuning. The candidate explicitly describes 'diversity-seeking reinforcement learning algorithms' using GFlowNets, which is the same fundamental approach as FlowRL's 'reward distribution matching via flow balancing.' - **Original**: we propose flowrl, a policy optimization algorithm that shifts from reward maximization to reward distribution matching via flow balancing, encouraging diverse reasoning path exploration while addressing the inherent mode-collapse limitations of existing rl methods. - **Candidate**: we address this limitation by using amortized bayesian inference to sample from these intractable posteriors. such amortization is algorithmically achieved by fine-tuning llms via diversity-seeking reinforcement learning algorithms: generative flow networks (gflownets).

Evidence 2 - **Rationale**: Both papers explicitly frame their contribution as moving away from reward maximization toward distribution matching in LLM contexts. The candidate describes 'distribution-matching paradigm' as an alternative to 'reward-maximizing policy optimization,' which directly parallels FlowRL's claim of matching distributions 'instead of solely maximizing rewards.' - **Original**: we propose flowrl: matching the full reward distribution via flow balancing instead of solely maximizing rewards in large language model (llm) reinforcement learning (rl). - **Candidate**: we empirically demonstrate that this distribution-matching paradigm of llm fine-tuning can serve as an effective alternative to maximum-likelihood training and reward-maximizing policy optimization.

Evidence 3 - **Rationale**: Both approaches use GFlowNets to transform reward signals into distributions for sampling. The candidate's 'amortized bayesian inference' to sample from 'intractable posteriors' using GFlowNets is conceptually equivalent to FlowRL's use of a partition function to normalize rewards into a target distribution, as both leverage GFlowNets' distribution-matching capabilities. - **Original**: the core idea of flowrl is to introduce a learnable partition function that normalizes scalar rewards into a target distribution, and to minimize the reverse kl divergence between the policy and this reward-induced distribution. - **Candidate**: we address this limitation by using amortized bayesian inference to sample from these intractable posteriors. such amortization is algorithmically achieved by fine-tuning llms via diversity-seeking reinforcement learning algorithms: generative flow networks (gflownets).

---

4. Generalist Reward Models: Found Inside Large Language Models

URL: View paper

Brief Assessment

Generalist Reward Models[56] focuses on extracting reward signals from pre-trained LLMs without additional training, using inverse reinforcement learning theory. This differs fundamentally from FlowRL's flow-balanced optimization approach that transforms scalar rewards into normalized distributions via a learnable partition function and minimizes reverse KL divergence for diverse exploration.

---

5. Human-centric reward optimization for reinforcement learning-based automated driving using large language models

URL: View paper

Brief Assessment

Human-centric Reward[53] focuses on using LLMs to optimize reward functions for automated driving tasks, not on reward distribution matching in RL for language models. The candidate addresses reward design for vehicle control, while the original paper proposes a policy optimization algorithm that matches reward distributions via flow balancing for LLM reasoning tasks.

---

6. Reward collapse in aligning large language models

URL: View paper

Brief Assessment

Reward Collapse[58] addresses reward model training for preference learning, focusing on how ranking-based objectives lead to identical reward distributions across prompts. FlowRL addresses policy optimization for reasoning tasks, transforming scalar rewards into distributions via flow balancing to prevent mode collapse during RL training—fundamentally different problem domains and technical approaches.

---

7. Guiding pretraining in reinforcement learning with large language models

URL: View paper

Brief Assessment

Guiding Pretraining[52] focuses on using LLMs to generate exploratory goals for task-agnostic pretraining in RL, not on reward distribution matching via flow balancing. The candidate rewards agents for achieving LLM-suggested goals through semantic similarity, which is fundamentally different from FlowRL's approach of transforming scalar rewards into normalized target distributions using a learnable partition function and minimizing reverse KL divergence.

---

8. Direct preference optimization: Your language model is secretly a reward model

URL: View paper

Brief Assessment

DPO[57] focuses on direct policy optimization from preferences using a Bradley-Terry model reparameterization, bypassing explicit reward modeling. It does not address reward distribution matching or flow balancing for diverse exploration in RL.

---

9. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting

URL: View paper

Prior Art Analysis

Distribution Matching[55] demonstrates that the core concept of transforming scalar rewards into normalized target distributions and minimizing reverse KL divergence was previously established. The candidate paper explicitly formulates distribution matching through reverse KL minimization between policy and reward-weighted distributions using a learnable partition function, presenting the same fundamental approach. Both papers address the shift from reward maximization to distribution matching in RL for language models, with Distribution Matching[55] providing the theoretical foundation that FlowRL builds upon.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly describe the paradigm shift from reward maximization to distribution matching, establishing this as a known approach rather than a novel contribution. - **Original**: we propose flowrl, a policy optimization algorithm that aligns the policy model with the full reward distribution, encouraging mode coverage. flowrl achieves more efficient exploration by fundamentally shifting from reward maximization to reward distribution matching - **Candidate**: here we explore the theoretical connections between the two paradigms, and show that methods such as kl-control developed for rm can also be construed as belonging to dm. we further observe that while dm differs from rm, it can suffer from similar training difficulties

---

10. Self-rewarding language models

URL: View paper

Brief Assessment

Self-rewarding[1] focuses on iterative self-improvement where the language model acts as its own reward model through LLM-as-a-judge prompting, rather than matching reward distributions via flow balancing. The candidate does not address distribution matching or flow-based optimization methods.

Contribution 2: Theoretical equivalence between KL minimization and trajectory balance

Description: The authors establish theoretical equivalence (Proposition 1) showing that minimizing the KL objective is equivalent to minimizing the trajectory balance loss from GFlowNets. This provides a practical surrogate for reward-guided KL minimization that can be integrated into existing RL frameworks.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A variational perspective on generative flow networks

URL: View paper

Prior Art Analysis

Variational GFlowNets[61] demonstrates that the theoretical equivalence between KL divergence minimization and trajectory balance objectives was established prior to the original paper. The candidate paper explicitly proves that 'variational inference in gfns is equivalent to minimizing the trajectory balance objective when sampling trajectories from the forward model,' which directly addresses the same theoretical relationship claimed as novel in Proposition 1 of the original paper. Both papers establish the mathematical connection between KL-based objectives and trajectory balance formulations in the context of generative flow networks, with the candidate providing this equivalence result before the original submission.

Evidence

Evidence 1 - **Rationale**: Both papers establish the same theoretical equivalence between KL-based objectives and trajectory balance. The candidate explicitly states this equivalence result, demonstrating that this theoretical connection was known prior to the original paper's submission. - **Original**: proposition 1. in terms of expected gradients, minimizing the kl objective in eq. 2 is equivalent to minimizing the trajectory balance loss used in gflownet - **Candidate**: we show that variational inference in gfns is equivalent to minimizing the trajectory balance objective when sampling trajectories from the forward model

Evidence 2 - **Rationale**: The original paper's claim that trajectory balance serves as a surrogate for KL minimization is directly addressed by the candidate's demonstration of equivalence between variational inference (KL-based) and trajectory balance objectives in GFlowNets. - **Original**: remark 2 (trajectory balance as a practical surrogate for kl minimization). given the equivalence established in proposition 1, the kl-based distribution matching objective can be reformulated as the trajectory balance loss. - **Candidate**: in this work, we define variational objectives for gfns in terms of the kullback-leibler (kl) divergences between the forward and backward distribution. we show that variational inference in gfns is equivalent to minimizing the trajectory balance objective

---

2. Amortizing intractable inference in diffusion models for vision, language, and control

URL: View paper

Prior Art Analysis

Amortizing Diffusion[60] demonstrates that the theoretical equivalence between KL divergence minimization and trajectory balance was established prior to the original paper. The candidate paper presents Proposition 1, which proves that minimizing KL divergence is equivalent to minimizing trajectory balance loss, and explicitly states this provides 'a practical surrogate for reward-guided KL minimization that can be integrated into existing RL frameworks.' This directly refutes the novelty claim, as the candidate paper published this equivalence result before the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers describe converting the trajectory balance constraint into a practical loss function for optimization, establishing this as a known technique prior to the original paper. - **Original**: remark 2 (trajectory balance as a practical surrogate for kl minimization). given the equivalence established in proposition 1, the kl-based distribution matching objective can be reformulated as the trajectory balance loss. this reformulation provides a practical optimization approach by using a st... - **Candidate**: relative trajectory balance as a loss. analogously to the conversion of the tb constraint (7) into a trajectory-dependent training objective in [46, 38], we define the relative trajectory balance loss as the discrepancy between the two sides of (8), seen as a function of the vector □that parametrize...

---

3. Relative Trajectory Balance is equivalent to Trust-PCL

URL: View paper

Prior Art Analysis

Relative Trajectory Balance[63] demonstrates that the theoretical equivalence between KL-regularized RL and trajectory balance objectives was established prior to the original paper's work. The candidate explicitly states that it establishes 'an equivalence between

rtb and trust-pcl, an off-policy rl method with kl regularization' and that 'this equivalence situates rtb within the broader theoretical landscape of kl-regularized rl.' This directly challenges the novelty claim of Proposition 1 in the original paper, which presents the equivalence between KL minimization and trajectory balance as a new theoretical contribution.

Evidence

Evidence 1 - **Rationale**: The candidate paper explicitly establishes the same theoretical equivalence between KL-regularized RL methods and trajectory balance objectives that the original paper claims as novel. The candidate's work on 'relative trajectory balance' demonstrates this equivalence was known prior to the original submission. - **Original**: proposition 1. in terms of expected gradients, minimizing the kl objective in eq. 2 is equivalent to minimizing the trajectory balance loss used in gflownet (malkin et al., 2022; 2023; lee et al., 2024; bartoldson et al., 2025): min θ dkl πθ(y | x) exp(βr(x, y)) zϕ(x) ⟺ min θ (log zϕ(x) + logπθ(y... - **Candidate**: building on prior work linking gflownets and maximum-entropy rl, we establish in this paper an equivalence between rtb and trust-pcl, an off-policy rl method with kl regularization. this equivalence situates rtb within the broader theoretical landscape of kl-regularized rl, and clarifies its relatio...

Evidence 2 - **Rationale**: The candidate describes RTB as serving the role of KL-regularized RL for fine-tuning, indicating that the connection between trajectory balance and KL regularization was already established in prior work, contradicting the original paper's claim of novelty for this theoretical insight. - **Original**: remark 2 (trajectory balance as a practical surrogate for kl minimization). given the equivalence established in proposition 1, the kl-based distribution matching objective can be reformulated as the trajectory balance loss. this reformulation provides a practical optimization approach by using a st... - **Candidate**: recent progress in generative modeling has highlighted the importance of reinforcement learning (rl) for fine-tuning, with kl-regularized methods in particular proving to be highly effective for both autoregressive and diffusion models. complementing this line of work, the relative trajectory balanc...

---

4. On divergence measures for training gflownets

URL: View paper

Prior Art Analysis

Divergence Measures GFlowNets[59] demonstrates that the theoretical equivalence between KL divergence minimization and trajectory balance loss was already established in prior work. The candidate paper explicitly proves this equivalence (Proposition 1) and extends it beyond discrete spaces to arbitrary topological spaces, showing that minimizing KL divergence is equivalent to minimizing trajectory balance loss in terms of expected gradients. This directly refutes the novelty claim, as the candidate establishes the same theoretical relationship before the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers recognize that the KL divergence provides a practical alternative to trajectory balance loss, with the candidate explicitly noting this relationship and its practical implications before the original paper. - **Original**: Proposition 1. In terms of expected gradients, minimizing the KL objective in eq. 2 is equivalent to minimizing the trajectory balance loss used in gflownet (malkin et al., 2022; 2023; lee et al., 2024; bartoldson et al., 2025): min θ dkl πθ(y | x) exp(βr(x, y)) zϕ(x) ⟺ min θ (log zϕ(x) + logπθ(y... - **Candidate**: this proposition shows that minimizing the on-policy tb loss is theoretically comparable to minimizing the kl divergence between pf and pb in terms of convergence speed. since the tb loss requires estimating the intractable r(x), the kl divergence, which avoids this estimation, can be a more suitabl...

Evidence 2 - **Rationale**: The candidate paper explicitly extends the KL-trajectory balance equivalence to arbitrary topological spaces, demonstrating prior establishment of this theoretical relationship beyond the discrete case that the original paper addresses. - **Original**: Remark 2 (trajectory balance as a practical surrogate for kl minimization). Given the equivalence established in proposition 1, the kl-based distribution matching objective can be reformulated as the trajectory balance loss. This reformulation provides a practical optimization approach by using a st... - **Candidate**: malkin et al. [56] demonstrated that, for discrete target distributions, the tb loss in (1) aligns with the kl divergence in terms of expected gradients. extending this, our proposition 1 establishes that this relationship also holds for distributions over arbitrary topological spaces.

---

5. KL DIVERGENCE OPTIMIZATION WITH ENTROPY-RATIO ESTIMATION FOR STOCHASTIC GFLOWNETS

URL: View paper

Brief Assessment

KL Divergence Optimization[65] focuses on stochastic GFlowNets with entropy-ratio estimation for exploration-exploitation trade-offs in stochastic environments, rather than establishing equivalence between KL minimization and trajectory balance for LLM reasoning tasks.

---

6. Streaming Bayes GFlowNets

URL: View paper

Brief Assessment

Streaming Bayes GFlowNets[62] focuses on streaming Bayesian inference over discrete spaces using GFlowNets, not on establishing the fundamental equivalence between KL minimization and trajectory balance for RL frameworks. The candidate builds upon existing GFlowNet theory rather than proposing this equivalence as a novel contribution.

---

7. FlowHF: Generative Flow Networks for RLHF

URL: View paper

Brief Assessment

FlowHF[64] applies GFlowNets to RLHF for language models but does not establish the theoretical equivalence claimed in the original paper. The candidate focuses on implementing trajectory balance and forward-looking objectives for RLHF without proving gradient equivalence between KL minimization and trajectory balance.

Contribution 3: Length normalization and importance sampling techniques

Description: The authors develop two technical solutions for long chain-of-thought training: length normalization to prevent gradient explosion from variable-length sequences, and importance sampling to correct distribution mismatch between generated rollouts and the current policy.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Remaining Useful Life Prediction of Aircraft Engines with Variable Length Input Sequences

URL: View paper

Brief Assessment

RUL Prediction[51] addresses variable-length input sequences in aircraft engine monitoring through data normalization and sampling techniques for sensor data preprocessing, not gradient stabilization or distribution mismatch correction in reinforcement learning training for chain-of-thought reasoning.

Appendix: Text Similarity Detection

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

References

- [0] FlowRL: Matching Reward Distributions for LLM Reasoning View paper
- [1] Self-rewarding language models View paper
- [2] Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? View paper
- [3] Routing to the expert: Efficient reward-guided ensemble of large language models View paper
- [4] R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization View paper
- [5] Exploring the limit of outcome reward for learning mathematical reasoning View paper
- [6] Amortizing intractable inference in large language models View paper
- [7] Offline reinforcement learning for llm multi-step reasoning View paper
- [8] Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks View paper
- [9] Adversarial moment-matching distillation of large language models View paper
- [10] Reso: A reward-driven self-organizing llm-based multi-agent system for reasoning tasks View paper
- [11] RFG: Test-Time Scaling for Diffusion Large Language Model Reasoning with Reward-Free Guidance View paper
- [12] InfAlign: Inference-aware language model alignment View paper
- [13] Teaching large language models to reason with reinforcement learning View paper
- [14] Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge View paper
- [15] Reward reasoning model View paper
- [16] Cal-dpo: Calibrated direct preference optimization for language model alignment View paper
- [17] Ensembling large language models with process reward-guided tree search for better complex reasoning View paper
- [18] Interleaved Reasoning for Large Language Models via Reinforcement Learning View paper
- [19] Rethinking reward modeling in preference-based large language model alignment View paper
- [20] Introspective reward modeling via inverse reinforcement learning for llm alignment View paper
- [21] Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms View paper
- [22] R3hf: Reward redistribution for enhancing reinforcement learning from human feedback View paper
- [23] Breaking Reward Collapse: Adaptive Reinforcement for Open-ended Medical Reasoning with Enhanced Semantic Discrimination View paper
- [24] Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding View paper
- [25] Warm: On the benefits of weight averaged reward models View paper
- [26] On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization View paper
- [27] Generative Reward Models View paper
- [28] Learning a Dense Reasoning Reward Model from Expert Demonstration via Inverse Reinforcement Learning View paper
- [29] Boosting Policy and Process Reward Models with Monte Carlo Tree Search in Open-Domain QA View paper
- [30] Learning reward for robot skills using large language models via self-alignment View paper
- [31] Enhancing reasoning for diffusion llms via distribution matching policy optimization View paper
- [32] Simultaneous reward distillation and preference learning: Get you a language model who can do both View paper
- [33] L2M-AID: Autonomous Cyber-Physical Defense by Fusing Semantic Reasoning of Large Language Models with Multi-Agent Reinforcement Learning (Preprint) View paper
- [34] Process reward models for llm agents: Practical framework and directions View paper
- [35] Mixed-R1: Unified Reward Perspective For Reasoning Capability in Multimodal Large Language Models View paper
- [36] Co-rewarding: Stable Self-supervised RL for Eliciting Reasoning in Large Language Models View paper
- [37] Consistent Paths Lead to Truth: Self-Rewarding Reinforcement Learning for LLM Reasoning View paper
- [38] Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment View paper
- [39] Reinforcement Pre-Training View paper
- [40] Rule based rewards for language model safety View paper
- [41] Guiding LLM decision-making with fairness reward models View paper
- [42] Reinforcing multi-turn reasoning in llm agents via turn-level reward design View paper
- [43] Align to Structure: Aligning Large Language Models with Structural Information View paper
- [44] Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning View paper
- [45] Your Reward Function for RL is Your Best PRM for Search: Unifying RL and Search-Based TTS View paper
- [46] Tuning for LLM alignment View paper
- [47] Reward-robust rlhf in llms View paper
- [48] Distilling the Implicit Multi-Branch Structure in LLMs' Reasoning via Reinforcement Learning View paper
- [49] From demonstrations to rewards: Alignment without explicit human preferences View paper
- [50] Learning Reward and Policy Jointly from Demonstration and Preference Improves Alignment View paper
- [51] Remaining Useful Life Prediction of Aircraft Engines with Variable Length Input Sequences View paper
- [52] Guiding pretraining in reinforcement learning with large language models View paper
- [53] Human-centric reward optimization for reinforcement learning-based automated driving using large language models View paper
- [54] Transforming and combining rewards for aligning large language models View paper
- [55] On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting View paper
- [56] Generalist Reward Models: Found Inside Large Language Models View paper
- [57] Direct preference optimization: Your language model is secretly a reward model View paper
- [58] Reward collapse in aligning large language models View paper
- [59] On divergence measures for training gflownets View paper
- [60] Amortizing intractable inference in diffusion models for vision, language, and control View paper
- [61] A variational perspective on generative flow networks View paper
- [62] Streaming Bayes GFlowNets View paper
- [63] Relative Trajectory Balance is equivalent to Trust-PCL View paper
- [64] FlowHF: Generative Flow Networks for RLHF View paper

- [65] KL DIVERGENCE OPTIMIZATION WITH ENTROPY-RATIO ESTIMATION FOR STOCHASTIC GFLOWNETS View paper