Novelty Assessment Report

Paper: Differential Smoothing Mitigates Sharpening and Improves LLM Reasoning
PDF URL: https://openreview.net/pdf?id=2RWf359T0p
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

It is widely recognized that reinforcement learning (RL) fine-tuning of large language models often leads to \textit{diversity collapse}, where outputs lack variety. Prior work has proposed a range of heuristics to counteract this effect, but these methods are ad hoc: they frequently trade off correctness for diversity, their effectiveness varies across tasks, and in some cases they even contradict one another. In this work, we place these observations on a rigorous foundation. We first provide a formal proof of why RL fine-tuning exhibits diversity collapse. Building directly on this analysis, we introduce a principled method—\textit{differential smoothing}—that provably improves both correctness and diversity, outperforming vanilla RL as well as widely used entropy-based heuristics. Our theory precisely characterizes why differential smoothing outperform vanilla RL and RL with direct entropy maximization. Extensive experiments with models from 1B to 7B parameters, across domains including CountDown and real-world mathematical reasoning, demonstrate consistent gains. Differential smoothing improves both Pass@1 (correctness) and Pass@k (diversity), with up to 6.7\% improvements on AIME24 dataset.

Core Task Landscape

This paper addresses: **Mitigating Diversity Collapse in RL Fine-Tuning of LLMs**
A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

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

- **Diversity-Aware Optimization Methods**
- **Theoretical Analysis and Mechanistic Understanding**
- **Preference Learning and Reward Model Design**
- **Data and Training Strategies**
- **Task-Specific Applications**
- **Evaluation and Measurement**
- **Related Methods and Techniques**

Complete Taxonomy Tree

- Mitigating Diversity Collapse in RL Fine-Tuning of LLMs Survey Taxonomy
- Diversity-Aware Optimization Methods
  - Joint Quality-Diversity Optimization Frameworks (3 papers)
  - [4] Diversity-Aware Policy Optimization for Large Language Model Reasoning (Yao Jian, 2025) View paper
  - [11] Jointly reinforcing diversity and quality in language model generations (Li Tianjian, 2025) View paper
  - [29] Post-training Large Language Models for Diverse High-Quality Responses (Chen Yi-lei, 2025) View paper
  - Adaptive Regularization Techniques (2 papers)
  - [8] Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models (Fan Jiajun, 2025) View paper
  - [17] The Choice of Divergence: A Neglected Key to Mitigating Diversity Collapse in Reinforcement Learning with Verifiable Reward (Li Long, 2025) View paper
  - Exploration-Driven RL Approaches (3 papers)
  - [15] Curiosity-Driven Reinforcement Learning from Human Feedback (Chai, 2025) View paper
  - [28] EDGE-GRPO: Entropy-Driven GRPO with Guided Error Correction for Advantage Diversity (Zhang Xingjian, 2025) View paper
  - [38] Entropy-guided sequence weighting for efficient exploration in RL-based LLM fine-tuning (Vanlioglu, 2025) View paper
  - Diverse Sampling and Decoding Strategies (3 papers)
  - [16] Diverse preference optimization (Lanchantin, 2025) View paper
  - [26] Group-Aware Reinforcement Learning for Output Diversity in Large Language Models (Oron Anschel, 2025) View paper
  - [34] Semantic-guided Diverse Decoding for Large Language Model (Shi Weijie, 2025) View paper
- Theoretical Analysis and Mechanistic Understanding
  - Formal Characterization of Diversity Collapse ★ (2 papers)
  - [0] Differential Smoothing Mitigates Sharpening and Improves LLM Reasoning (Anon et al., 2026) View paper
  - [10] Outcome-based exploration for llm reasoning (Song Yu-da, 2025) View paper
  - Empirical Attribution Studies (2 papers)
  - [2] Attributing mode collapse in the fine-tuning of large language models (L O'Mahony, 2024) View paper
  - [44] Reinforcement Learning Fine-Tuning Enhances Activation Intensity and Diversity in the Internal Circuitry of LLMs (Zhang Hong-lin, 2025) View paper
  - RL Algorithm Analysis for LLM Planning (2 papers)
  - [6] Benefits and pitfalls of reinforcement learning for language model planning: a theoretical perspective (Wang, 2025) View paper

    - [20] Rewarding the unlikely: Lifting grpo beyond distribution sharpening (He, 2025) View paper
- Preference Learning and Reward Model Design
    - Diverse Preference Modeling (3 papers)
    - [5] Diverse preference learning for capabilities and alignment (Stewart Slocum, 2025) View paper
    - [23] MaxMin-RLHF: Alignment with diverse human preferences (Chakraborty, 2024) View paper
    - [25] Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences (Chakraborty, 2024) View paper
    - Uncertainty-Aware Reward Ensembles (1 papers)
    - [45] Uncertainty-penalized reinforcement learning from human feedback with diversified reward LoRA ensembles (Y Zhai, 2026) View paper
- Data and Training Strategies
    - Synthetic Data Diversity (1 papers)
    - [31] Synthetic Eggs in Many Baskets: The Impact of Synthetic Data Diversity on LLM Fine-Tuning (Gatt, 2025) View paper
    - Supervised Fine-Tuning for Diversity Preservation (2 papers)
    - [1] Preserving diversity in supervised fine-tuning of large language models (Ze-yu, 2024) View paper
    - [49] Enhancing Large Language Model Reasoning via Selective Critical Token Fine-Tuning (Ruan Zhi-wen, 2025) View paper
    - Multi-Stage Post-Training Paradigms (2 papers)
    - [12] Multi-modal preference alignment remedies regression of visual instruction tuning on language model (Shengzhi Li, 2024) View paper
    - [48] MindGPT-4ov: An Enhanced MLLM via a Multi-Stage Post-Training Paradigm (Wei Chen, 2025) View paper
- Task-Specific Applications
    - Mathematical Reasoning and Theorem Proving (1 papers)
    - [37] GFlowNet Fine-tuning for Diverse Correct Solutions in Mathematical Reasoning Tasks (Takase, 2024) View paper
    - Subjective Reasoning and Creative Generation (2 papers)
    - [3] Beyond Quality: Unlocking Diversity in Ad Headline Generation with Large Language Models (Wang Chang, 2025) View paper
    - [7] Diversity-enhanced reasoning for subjective questions (Wang Yumeng, 2025) View paper
    - Red-Teaming and Safety Applications (5 papers)
    - [18] Learning diverse attacks on large language models for robust red-teaming and safety tuning (Lee, 2024) View paper
    - [27] Diversity Seeking Techniques for Red-Teaming Large Language Models (Seokhan Lee, 2025) View paper
    - [30] Adversarial Reinforcement Learning for Large Language Model Agent Safety (Wang, 2025) View paper
    - [35] Diverse and effective red teaming with auto-generated rewards and multi-step reinforcement learning (Beutel, 2024) View paper
    - [42] Active Attacks: Red-teaming LLMs via Adaptive Environments (Yun Taeyoung, 2025) View paper
    - Specialized Domain Applications (4 papers)
    - [21] RecLLM-R1: A Two-Stage Training Paradigm with Reinforcement Learning and Chain-of-Thought v1 (Xie Yu, 2025) View paper
    - [36] Large Language Model-Enhanced Reinforcement Learning for Diverse and Novel Recommendations (Woo, 2025) View paper
    - [41] RL-TweetGen: A Socio-Technical Framework for Engagement-Optimized Short Text Generation in Digital Commerce Using Large Language Models and Reinforcement Learning (C. S, 2025) View paper
    - [43] Foundation Models at Work: Fine-Tuning for Fairness in Algorithmic Hiring (Korkmaz, 2025) View paper
- Evaluation and Measurement
    - Diversity Metrics and Evaluation Frameworks (1 papers)
    - [13] Evaluating the diversity and quality of llm generated content (Shypula, 2025) View paper
    - Conceptual Diversity Studies (1 papers)
    - [22] One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity (Hu, 2025) View paper
- Related Methods and Techniques
    - Amortized Inference and Posterior Sampling (1 papers)
    - [33] Amortizing intractable inference in large language models (Hu, 2023) View paper
    - Mode Collapse in Other Generative Models (1 papers)
    - [39] Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning (Barcelo, 2024) View paper
    - Demonstration Selection and In-Context Learning (1 papers)
    - [19] Demonstration selection for in-context learning via reinforcement learning (Wang Xubin, 2024) View paper
    - Agent Training and Decision-Making (2 papers)
    - [9] Fine-tuning large vision-language models as decision-making agents via reinforcement learning (Hao Bai, 2024) View paper
    - [24] Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning (Xi, 2025) View paper
    - Peripheral Applications and Surveys (6 papers)
    - [14] An Automated Reinforcement Learning Reward Design Framework with Large Language Model for Cooperative Platoon Coordination (Peng Yi, 2025) View paper
    - [32] LLM challenges and solutions (Uday Kamath, 2024) View paper
    - [40] Human-AI Interaction in the Era of Large Language Models (LLMs) (Behnam, 2025) View paper
    - [46] Ethical Concerns and Mitigation Strategies in AI-Driven Language Models (R AGRAWAL, 2024) View paper
    - [47] Reinforcement Learning for Safe LLM Code Generation (Huang, 2025) View paper
    - [50] Uncertainty-Driven Adaptive Sampling for Resource-Efficient Language Model Inference (Abdulloh, 2025) View paper

Narrative

Core task: Mitigating diversity collapse in reinforcement learning fine-tuning of large language models. The field addresses a critical challenge that arises when RL methods optimize LLMs toward narrow reward signals, causing models to lose their ability to generate varied, creative responses. The taxonomy organizes research into several complementary branches: Diversity-Aware Optimization Methods develop algorithmic techniques that explicitly encourage varied outputs during training, while Theoretical Analysis and Mechanistic Understanding seeks to formalize why and how collapse occurs. Preference Learning and Reward Model Design examines how reward signals themselves can be structured to preserve diversity, and Data and Training Strategies explores curriculum design and data selection approaches. Task-Specific Applications demonstrate these principles in domains like red-teaming, reasoning, and recommendation, while Evaluation and Measurement provides metrics to quantify diversity loss. Related Methods and Techniques connects this work to broader ideas in generative modeling and exploration.

Particularly active lines of work contrast algorithmic interventions with diagnostic analysis. Many studies propose explicit diversity regularizers or multi-objective formulations—such as Diversity-Aware Policy[4] and Diverse Preference Optimization[16]—that balance reward maximization with entropy or coverage objectives, while others like Preserving Diversity Fine-tuning[1] and Adaptive Divergence Regularization[8] adjust KL penalties dynamically. Meanwhile, works such as Mode Collapse Attribution[2] and Alignment Reduces Diversity[22] investigate the underlying mechanisms, revealing how standard RL objectives systematically favor mode-seeking behavior. Differential Smoothing[0] sits within the theoretical branch alongside Outcome-Based Exploration[10], offering a formal characterization of how diversity collapses under gradient-based updates. Compared to purely algorithmic fixes like Diversity Quality Reinforcement[11] or Curiosity-Driven RLHF[15], Differential Smoothing[0] emphasizes mechanistic insight, aiming to understand the collapse phenomenon rigorously before prescribing remedies. This positioning complements empirical mitigation strategies by providing foundational principles that can guide the design of more robust training procedures.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Outcome-based exploration for llm reasoning

Authors: Song Yu-da, Kempe, Julia, Yuda Song, Munos, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reinforcement learning (RL) has emerged as a powerful method for improving the reasoning abilities of large language models (LLMs). Outcome-based RL, which rewards policies solely for the correctness of the final answer, yields substantial accuracy gains but also induces a systematic loss in generation diversity. This collapse undermines real-world performance, where diversity is critical for test-time scaling. We analyze this phenomenon by viewing RL post-training as a sampling process and show...

Relationship Analysis

Both papers belong to the 'Formal Characterization of Diversity Collapse' category, providing theoretical frameworks to explain why RL fine-tuning causes diversity collapse in LLMs. The original paper (Differential Smoothing) provides formal proofs identifying selection bias and reinforcement bias as the two compounding mechanisms driving diversity collapse, then derives a principled differential smoothing reward modification. The candidate paper (Outcome-based Exploration) analyzes diversity collapse by framing RL as a sampling process and demonstrates transfer of diversity degradation across questions, then proposes outcome-based exploration methods (UCB-Con, Batch) that add exploration bonuses to final outcomes rather than modifying rewards based on log-probabilities.

Contributions Analysis

Overall novelty summary. The paper contributes a formal proof explaining why RL fine-tuning causes diversity collapse, alongside a principled method called differential smoothing that provably improves both correctness and diversity. It resides in the 'Formal Characterization of Diversity Collapse' leaf under 'Theoretical Analysis and Mechanistic Understanding', which contains only two papers total. This represents a sparse research direction within the broader taxonomy of 50 papers, indicating that rigorous mathematical characterizations of diversity collapse remain relatively underexplored compared to empirical mitigation techniques.

The taxonomy reveals that most work concentrates in 'Diversity-Aware Optimization Methods' (13 papers across four leaves) and 'Task-Specific Applications' (13 papers across four leaves), emphasizing algorithmic interventions and domain-specific solutions. The paper's theoretical branch sits adjacent to 'Empirical Attribution Studies' and 'RL Algorithm Analysis for LLM Planning', which investigate collapse mechanisms through controlled experiments rather than formal proofs. While neighboring branches like 'Joint Quality-Diversity Optimization Frameworks' and 'Adaptive Regularization Techniques' propose heuristic solutions, this work provides foundational analysis that could inform those algorithmic designs.

Among 25 candidates examined across three contributions, none were found to clearly refute the paper's claims. The formal proof contribution examined 10 candidates with zero refutations, the differential smoothing method examined 10 candidates with zero refutations, and the theoretical characterization examined 5 candidates with zero refutations. This suggests that within the limited search scope, the formal proof of diversity collapse and the universal superiority characterization of differential smoothing over entropy-based heuristics appear relatively novel. The algorithmic contribution (DS-GRPO) also shows no substantial prior overlap among examined candidates.

Based on top-25 semantic matches and citation expansion, the analysis indicates the work occupies a sparsely populated theoretical niche. The formal characterization and provable superiority claims appear distinctive within the examined literature, though the limited search scope means potentially relevant theoretical work outside these candidates remains unassessed. The taxonomy structure confirms that rigorous mathematical foundations for diversity collapse constitute a minority research direction compared to empirical method development.

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

Contribution 1: Formal proof of diversity collapse in RL fine-tuning

Description: The authors formally prove that RL fine-tuning causes diversity collapse through two mechanisms: selection bias (correct high-probability trajectories are more likely reinforced) and reinforcement bias (these trajectories receive disproportionately larger updates). This theoretical analysis explains why RL amplifies existing proficiencies rather than rectifying deficiencies.

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. Evolutionary reinforcement learning: A survey

URL: View paper

Brief Assessment

Evolutionary RL Survey[60] is a broad survey of evolutionary reinforcement learning methods across various domains (hyperparameter optimization, policy search, exploration, reward shaping). It does not address diversity collapse in RL fine-tuning of language models through selection bias and reinforcement bias mechanisms, which is the specific theoretical contribution of the original paper.

---

2. The Choice of Divergence: A Neglected Key to Mitigating Diversity Collapse in Reinforcement Learning with Verifiable Reward

URL: View paper

Brief Assessment

Choice of Divergence[17] focuses on divergence term selection (forward-KL, JS-divergence) as a solution mechanism, not on formally proving the underlying causes of diversity collapse through selection bias and reinforcement bias as the original paper does.

---

3. Epistemic diversity and industrial selection bias

URL: View paper

Brief Assessment

Epistemic Diversity Selection[66] focuses on industrial selection bias in scientific funding using reinforcement learning to model funding decisions, not on diversity collapse mechanisms in RL fine-tuning of language models through selection and reinforcement biases.

---

4. Federated Learning for All: A Reinforcement Learning-Based Approach for Ensuring Fairness in Client Selection

URL: View paper

Brief Assessment

Federated Learning Fairness[61] addresses client selection fairness in federated learning systems using reinforcement learning, not diversity collapse in language model fine-tuning. The technical domains and problem formulations are entirely distinct.

---

5. Evolving language models without labels: Majority drives selection, novelty promotes variation

URL: View paper

Brief Assessment

Evolving Without Labels[62] focuses on label-free self-improvement and diversity collapse in that specific context, whereas the original paper provides formal proofs of diversity collapse mechanisms (selection bias and reinforcement bias) in standard RL fine-tuning with verifiable rewards. The candidate does not present formal proofs of these specific mechanisms.

---

6. Diversity oriented Deep Reinforcement Learning for targeted molecule generation

URL: View paper

Brief Assessment

Diversity Molecule Generation[64] focuses on molecular generation using reinforcement learning for drug design, not on formal theoretical analysis of diversity collapse mechanisms in language model fine-tuning. The candidate addresses diversity in chemical compound generation through exploration strategies, while the original paper provides theoretical proofs about selection bias and reinforcement bias in LLM reasoning tasks.

---

7. Training Diffusion Models Towards Diverse Image Generation with Reinforcement Learning

URL: View paper

Brief Assessment

Diverse Image Generation[53] focuses on diversity in image generation using diffusion models with set-based reward functions, not on proving mechanisms of diversity collapse in RL fine-tuning for language models through selection and reinforcement bias.

---

8. Diversity-Driven Exploration Strategy for Deep Reinforcement Learning

URL: View paper

Brief Assessment

Diversity-Driven Exploration[63] addresses diversity collapse in traditional deep RL settings (Atari, MuJoCo) through exploration strategies, not through formal proofs of selection and reinforcement bias mechanisms in LLM fine-tuning with verifiable rewards.

---

9. Evolutionary diversity optimization with clustering-based selection for reinforcement learning

URL: View paper

Brief Assessment

Evolutionary Diversity Optimization[65] focuses on evolutionary algorithms for behavior diversity in general RL settings, not on the specific mechanisms of diversity collapse (selection bias and reinforcement bias) during LLM fine-tuning that the original paper formally proves.

---

10. Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning

URL: View paper

Brief Assessment

Skewness-Aware RL[59] addresses racial bias in face recognition using reinforcement learning to adaptively adjust margins for different demographic groups. This is fundamentally different from analyzing diversity collapse mechanisms in language model fine-tuning through selection and reinforcement biases.

Contribution 2: Differential smoothing method (DS-GRPO algorithm)

Description: The authors propose differential smoothing, a novel reward modification approach that applies distinct pressures to correct and incorrect trajectories. For correct trajectories, it subtracts a log-probability term to enhance diversity; for incorrect ones, it adds the log-probability to improve correctness. This is implemented as the DS-GRPO algorithm.

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. Toolrl: Reward is all tool learning needs

URL: View paper

Brief Assessment

ToolRL[52] focuses on reward design for tool selection and application tasks in LLMs, not on differential reward modification for correctness-diversity tradeoffs in general RL reasoning tasks.

---

2. Text2Reward: Reward Shaping with Language Models for Reinforcement Learning

URL: View paper

Brief Assessment

Text2Reward[55] focuses on automated reward function generation using LLMs for robotic tasks, not on differential reward modification methods for RL training. The candidate addresses reward design/shaping, while the original contribution concerns training dynamics and diversity collapse mitigation.

---

3. Learn to reason efficiently with adaptive length-based reward shaping

URL: View paper

Brief Assessment

Adaptive Length Reward[57] focuses on length-based reward shaping for reasoning efficiency, not differential reward modification for correctness vs. diversity trade-offs in RL fine-tuning.

---

4. Jointly reinforcing diversity and quality in language model generations

URL: View paper

Brief Assessment

Diversity Quality Reinforcement[11] focuses on jointly optimizing quality and diversity through a learned semantic classifier and multiplicative reward fusion, rather than differential reward modification based on trajectory correctness. The candidate's approach partitions responses into semantic clusters and multiplies diversity scores with quality rewards, which differs fundamentally from the original paper's differential smoothing that applies distinct log-probability adjustments to correct versus incorrect trajectories.

---

5. Enhancing deep reinforcement learning for stock trading: a reward shaping approach via expert feedback: A. Orra et al.

URL: View paper

Brief Assessment

Reward Shaping Stock[54] focuses on stock trading with expert feedback for reward shaping, not on differential reward modification methods for LLM reasoning that apply distinct pressures to correct vs. incorrect trajectories.

---

6. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers

URL: View paper

Brief Assessment

Verifiable Noisy Rewards[58] addresses reward noise from imperfect verifiers (false positives/negatives) through backward/forward corrections to policy gradients, whereas the original paper tackles diversity collapse in RL fine-tuning through differential reward modification based on trajectory correctness. These are distinct technical problems with different mechanisms.

---

7. Toward Diverse Text Generation with Inverse Reinforcement Learning

URL: View paper

Brief Assessment

Diverse Text IRL[56] focuses on text generation using inverse reinforcement learning with entropy regularization to address mode collapse and reward sparsity. The original paper's differential smoothing applies distinct reward modifications to correct vs. incorrect trajectories in mathematical reasoning tasks, which is a different technical approach and application domain.

---

8. Training Diffusion Models Towards Diverse Image Generation with Reinforcement Learning

URL: View paper

Brief Assessment

Diverse Image Generation[53] proposes a diversity-oriented RL fine-tuning method for diffusion models using set-based diversity rewards, which is fundamentally different from the differential smoothing approach that modifies rewards based on trajectory correctness with log-probability terms in language model reasoning tasks.

---

9. Diversity-enhanced reasoning for subjective questions

URL: View paper

Brief Assessment

Diversity-Enhanced Reasoning[7] focuses on subjective reasoning tasks with role-based perspectives and multi-role reasoning paths, not on differential reward modification for correctness vs. diversity in general RL settings.

---

10. Process vs. Outcome Reward: Which is Better for Agentic RAG Reinforcement Learning

URL: View paper

Brief Assessment

Process Outcome Reward[51] focuses on process-level rewards for agentic RAG systems (query generation, evidence extraction, answer generation), while the original paper proposes differential smoothing for general RL fine-tuning that applies distinct reward modifications to correct vs. incorrect trajectories to balance correctness and diversity.

Contribution 3: Theoretical characterization of existing heuristics and universal superiority proof

Description: The authors provide formal theoretical guarantees proving that differential smoothing outperforms vanilla RL and entropy-based heuristics in both correctness and diversity. They also clarify the contradictory effects of global entropy regularization, explaining when entropy maximization or minimization helps based on task characteristics.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Statistical analysis of Inverse Entropy-regularized Reinforcement Learning

URL: View paper

Brief Assessment

Inverse Entropy-Regularized[69] focuses on inverse reinforcement learning (inferring reward functions from expert demonstrations) rather than forward RL optimization. The candidate addresses statistical estimation of expert policies and reward recovery, not the correctness-diversity trade-offs in RL fine-tuning that the original paper analyzes.

---

2. Convergence of softmax policy gradient: incorporating entropy regularization and handling linear function approximation

URL: View paper

Brief Assessment

Softmax Policy Convergence[68] focuses on convergence analysis of softmax policy gradient methods in bandits and tabular MDPs with entropy regularization, not on comparing reward smoothing versus entropy regularization heuristics in LLM reasoning tasks or proving universal superiority of differential smoothing.

---

3. EPO: Entropy-regularized Policy Optimization for LLM Agents Reinforcement Learning

URL: View paper

Brief Assessment

EPO Entropy-Regularized[70] focuses on multi-turn agent environments with sparse rewards and addresses exploration-exploitation cascade failures through entropy smoothing. The original paper addresses diversity collapse in single-turn RL fine-tuning of LLMs with

verifiable rewards, providing theoretical guarantees for differential smoothing versus vanilla RL and entropy regularization. These are fundamentally different problem settings and theoretical contributions.

---

4. Reward Shaping via Diffusion Process in Reinforcement Learning

URL: View paper

Brief Assessment

Reward Shaping Diffusion[71] focuses on thermodynamic interpretations of reward shaping in RL using diffusion processes and stochastic thermodynamics, not on theoretical guarantees comparing entropy regularization versus reward smoothing methods for LLM reasoning tasks.

---

5. Utilizing Prior Solutions for Reward Shaping and Composition in Entropy-Regularized Reinforcement Learning

URL: View paper

Brief Assessment

Prior Solutions Reward[67] focuses on reward shaping and task composition in entropy-regularized RL using prior solutions, not on proving superiority of differential smoothing over vanilla RL and entropy-based heuristics in LLM reasoning tasks.

Appendix: Text Similarity Detection

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

References

- [0] Differential Smoothing Mitigates Sharpening and Improves LLM Reasoning View paper
- [1] Preserving diversity in supervised fine-tuning of large language models View paper
- [2] Attributing mode collapse in the fine-tuning of large language models View paper
- [3] Beyond Quality: Unlocking Diversity in Ad Headline Generation with Large Language Models View paper
- [4] Diversity-Aware Policy Optimization for Large Language Model Reasoning View paper
- [5] Diverse preference learning for capabilities and alignment View paper
- [6] Benefits and pitfalls of reinforcement learning for language model planning: a theoretical perspective View paper
- [7] Diversity-enhanced reasoning for subjective questions View paper
- [8] Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models View paper
- [9] Fine-tuning large vision-language models as decision-making agents via reinforcement learning View paper
- [10] Outcome-based exploration for llm reasoning View paper
- [11] Jointly reinforcing diversity and quality in language model generations View paper
- [12] Multi-modal preference alignment remedies regression of visual instruction tuning on language model View paper
- [13] Evaluating the diversity and quality of llm generated content View paper
- [14] An Automated Reinforcement Learning Reward Design Framework with Large Language Model for Cooperative Platoon Coordination View paper
- [15] Curiosity-Driven Reinforcement Learning from Human Feedback View paper
- [16] Diverse preference optimization View paper
- [17] The Choice of Divergence: A Neglected Key to Mitigating Diversity Collapse in Reinforcement Learning with Verifiable Reward View paper
- [18] Learning diverse attacks on large language models for robust red-teaming and safety tuning View paper
- [19] Demonstration selection for in-context learning via reinforcement learning View paper
- [20] Rewarding the unlikely: Lifting grpo beyond distribution sharpening View paper
- [21] RecLLM-R1: A Two-Stage Training Paradigm with Reinforcement Learning and Chain-of-Thought v1 View paper
- [22] One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity View paper
- [23] MaxMin-RLHF: Alignment with diverse human preferences View paper
- [24] Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning View paper
- [25] Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences View paper
- [26] Group-Aware Reinforcement Learning for Output Diversity in Large Language Models View paper
- [27] Diversity Seeking Techniques for Red-Teaming Large Language Models View paper
- [28] EDGE-GRPO: Entropy-Driven GRPO with Guided Error Correction for Advantage Diversity View paper
- [29] Post-training Large Language Models for Diverse High-Quality Responses View paper
- [30] Adversarial Reinforcement Learning for Large Language Model Agent Safety View paper
- [31] Synthetic Eggs in Many Baskets: The Impact of Synthetic Data Diversity on LLM Fine-Tuning View paper
- [32] LLM challenges and solutions View paper
- [33] Amortizing intractable inference in large language models View paper
- [34] Semantic-guided Diverse Decoding for Large Language Model View paper
- [35] Diverse and effective red teaming with auto-generated rewards and multi-step reinforcement learning View paper
- [36] Large Language Model-Enhanced Reinforcement Learning for Diverse and Novel Recommendations View paper
- [37] GFlowNet Fine-tuning for Diverse Correct Solutions in Mathematical Reasoning Tasks View paper
- [38] Entropy-guided sequence weighting for efficient exploration in RL-based LLM fine-tuning View paper
- [39] Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning View paper
- [40] Human-AI Interaction in the Era of Large Language Models (LLMs) View paper
- [41] RL-TweetGen: A Socio-Technical Framework for Engagement-Optimized Short Text Generation in Digital Commerce Using Large Language Models and Reinforcement Learning View paper
- [42] Active Attacks: Red-teaming LLMs via Adaptive Environments View paper
- [43] Foundation Models at Work: Fine-Tuning for Fairness in Algorithmic Hiring View paper
- [44] Reinforcement Learning Fine-Tuning Enhances Activation Intensity and Diversity in the Internal Circuitry of LLMs View paper
- [45] Uncertainty-penalized reinforcement learning from human feedback with diversified reward LoRA ensembles View paper
- [46] Ethical Concerns and Mitigation Strategies in AI-Driven Language Models View paper
- [47] Reinforcement Learning for Safe LLM Code Generation View paper
- [48] MindGPT-4ov: An Enhanced MLLM via a Multi-Stage Post-Training Paradigm View paper

- [49] Enhancing Large Language Model Reasoning via Selective Critical Token Fine-Tuning View paper
- [50] Uncertainty-Driven Adaptive Sampling for Resource-Efficient Language Model Inference View paper
- [51] Process vs. Outcome Reward: Which is Better for Agentic RAG Reinforcement Learning View paper
- [52] Toolrl: Reward is all tool learning needs View paper
- [53] Training Diffusion Models Towards Diverse Image Generation with Reinforcement Learning View paper
- [54] Enhancing deep reinforcement learning for stock trading: a reward shaping approach via expert feedback: A. Orra et al. View paper
- [55] Text2Reward: Reward Shaping with Language Models for Reinforcement Learning View paper
- [56] Toward Diverse Text Generation with Inverse Reinforcement Learning View paper
- [57] Learn to reason efficiently with adaptive length-based reward shaping View paper
- [58] Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers View paper
- [59] Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning View paper
- [60] Evolutionary reinforcement learning: A survey View paper
- [61] Federated Learning for All: A Reinforcement Learning-Based Approach for Ensuring Fairness in Client Selection View paper
- [62] Evolving language models without labels: Majority drives selection, novelty promotes variation View paper
- [63] Diversity-Driven Exploration Strategy for Deep Reinforcement Learning View paper
- [64] Diversity oriented Deep Reinforcement Learning for targeted molecule generation View paper
- [65] Evolutionary diversity optimization with clustering-based selection for reinforcement learning View paper
- [66] Epistemic diversity and industrial selection bias View paper
- [67] Utilizing Prior Solutions for Reward Shaping and Composition in Entropy-Regularized Reinforcement Learning View paper
- [68] Convergence of softmax policy gradient: incorporating entropy regularization and handling linear function approximation View paper
- [69] Statistical analysis of Inverse Entropy-regularized Reinforcement Learning View paper
- [70] EPO: Entropy-regularized Policy Optimization for LLM Agents Reinforcement Learning View paper
- [71] Reward Shaping via Diffusion Process in Reinforcement Learning View paper