Novelty Assessment Report

Paper: Risk-Sensitive Reinforcement Learning for Alleviating Exploration Dilemmas in Large Language Models
PDF URL: https://openreview.net/pdf?id=7kC8ORye4l
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has proven effective for enhancing Large Language Models (LLMs) on complex reasoning tasks. Yet current methods face an exploration dilemma: standard RL struggles to escape the local optima of pre-trained LLMs' sharply peaked initial policies, boosting single-solution accuracy (pass@1) but suppressing solution diversity and multi-solution performance (pass@k). As a result, RLVR often distills existing capabilities rather than discovering new reasoning strategies. We address this with a Risk-Sensitive Reinforcement Learning framework. By adopting a risk-seeking objective that interpolates between mean and maximum rewards, we derive a novel Risk-Sensitive GRPO (RS-GRPO) algorithm that emphasizes hard prompts to drive exploration. Across six mathematical reasoning benchmarks and five LLMs, RS-GRPO consistently improves pass@k performance while enhancing or maintaing pass@1.

Core Task Landscape

This paper addresses: **Enhancing exploration in reinforcement learning for large language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Exploration Strategy Design and Mechanisms**
- **Knowledge-Guided and Semantic Exploration**
- **Memory and Experience-Based Exploration**
- **Hierarchical and Structured Exploration**
- **Training Efficiency and Sample Optimization**
- **Theoretical Analysis and Empirical Evaluation**
- **Application-Specific Exploration**
- **Surveys and Integrative Frameworks**

Complete Taxonomy Tree

- Enhancing exploration in reinforcement learning for large language models Survey Taxonomy
- Exploration Strategy Design and Mechanisms
  - Intrinsic Motivation and Curiosity-Driven Exploration (4 papers)
  - [4] CDE: Curiosity-Driven Exploration for Efficient Reinforcement Learning in Large Language Models (Song, 2025) View paper
  - [6] Efficient Exploration for LLMs (Dwaracherla, 2024) View paper
  - [36] Navigate the unknown: Enhancing llm reasoning with intrinsic motivation guided exploration (Gao, 2025) View paper
  - [48] Efficient Reinforcement Learning for Large Language Models with Intrinsic Exploration (Sun Yan, 2025) View paper
  - Risk-Sensitive and Objective-Based Exploration ★ (4 papers)
  - [0] Risk-Sensitive Reinforcement Learning for Alleviating Exploration Dilemmas in Large Language Models (Anon et al., 2026) View paper
  - [13] Knapsack rl: Unlocking exploration of llms via optimizing budget allocation (Li, 2025) View paper
  - [18] Outcome-based exploration for llm reasoning (Song Yu-da, 2025) View paper
  - [27] Unlocking reasoning capabilities in llms via reinforcement learning exploration (Deng, 2025) View paper
  - Adaptive and Dynamic Exploration Control (3 papers)
  - [7] LLM-Explorer: A Plug-in Reinforcement Learning Policy Exploration Enhancement Driven by Large Language Models (Song Yiwen, 2025) View paper
  - [25] Know when to explore: Difficulty-aware certainty as a guide for llm reinforcement learning (Li Ang, 2025) View paper
  - [35] On Entropy Control in LLM-RL Algorithms (Shen Han, 2025) View paper
- Knowledge-Guided and Semantic Exploration
  - LLM-Guided Goal and Subgoal Generation (4 papers)
  - [1] Guiding Pretraining in Reinforcement Learning with Large Language Models (Du Yuqing, 2023) View paper
  - [2] ExploRLLM: Guiding Exploration in Reinforcement Learning with Large Language Models (Runyu Ma, 2025) View paper
  - [17] LLM-Guided Reinforcement Learning for Interactive Environments (Fuxue Yang, 2025) View paper
  - [22] Guiding Exploration in Reinforcement Learning Through LLM-Augmented Observations (Vaibhav, 2025) View paper
  - Rubric and Instructional Scaffolding (1 papers)
  - [5] Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning (Zhou Yang, 2025) View paper
  - Semantic and Language-Based Exploration Enhancement (3 papers)

- [20] Accelerating reinforcement learning of robotic manipulations via feedback from large language models (Chu Kun, 2023) View paper
- [44] Language Guided Exploration for RL Agents in Text Environments (Dan, 2024) View paper
- [49] Large Language Model-Enhanced Reinforcement Learning for Diverse and Novel Recommendations (Woo, 2025) View paper

- Memory and Experience-Based Exploration
  - Experience Replay and Trajectory Reuse (2 papers)
  - [14] Reasoning Under 1 Billion: Memory-Augmented Reinforcement Learning for Large Language Models (Le, 2025) View paper
  - [31] Rlep: Reinforcement learning with experience replay for llm reasoning (Zhang Hong-zhi, 2025) View paper
  - Retrospective and Failure-Based Learning (2 papers)
  - [9] Trial and error: Exploration-based trajectory optimization for llm agents (Song, 2024) View paper
  - [10] Improving RL Exploration for LLM Reasoning through Retrospective Replay (Dou, 2025) View paper
- Hierarchical and Structured Exploration
  - Option Discovery and Hierarchical RL (1 papers)
  - [46] Option Discovery Using LLM-guided Semantic Hierarchical Reinforcement Learning (Shek, 2025) View paper
  - Algorithmic and Structured Reasoning Exploration (4 papers)
  - [15] Algorithm of Thoughts: Enhancing Exploration of Ideas in Large Language Models (Sel, 2023) View paper
  - [32] Parallel-R1: Towards Parallel Thinking via Reinforcement Learning (Zheng Tong, 2025) View paper
  - [33] Integrating Large Language Models and Reinforcement Learning for Non-Linear Reasoning (Yoav Alon, 2025) View paper
  - [34] Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning (Li, 2025) View paper
- Training Efficiency and Sample Optimization
  - Few-Shot and Data-Efficient RL (2 papers)
  - [3] Enhancing efficiency and exploration in reinforcement learning for llms (Mengqi Liao, 2025) View paper
  - [11] Reinforcement Learning for Reasoning in Large Language Models with One Training Example (Wang Yi-ping, 2025) View paper
  - Curriculum and Progressive Training (2 papers)
  - [37] Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning (Xi, 2024) View paper
  - [45] Behavior Injection: Preparing Language Models for Reinforcement Learning (Cen, 2025) View paper
- Theoretical Analysis and Empirical Evaluation
  - Exploration Capacity Analysis (5 papers)
  - [16] Disentangling exploration of large language models by optimal exploitation (Tim Grams, 2025) View paper
  - [24] From Trial-and-Error to Improvement: A Systematic Analysis of LLM Exploration Mechanisms in RLVR (Deng Jia, 2025) View paper
  - [42] Can large language models explore in-context? (Dylan Foster, 2024) View paper
  - [43] Evolve: Evaluating and optimizing llms for exploration (Nie, 2024) View paper
  - [47] Large Language Models Think Too Fast To Explore Effectively (Pan Lan, 2025) View paper
  - Comparative Studies and Benchmarking (4 papers)
  - [12] Toward efficient exploration by large language model agents (Arumugam, 2025) View paper
  - [19] Comparing Exploration-Exploitation Strategies of LLMs and Humans: Insights from Standard Multi-armed Bandit Tasks (Zhang Zi-yuan, 2025) View paper
  - [21] Teaching Large Language Models to Reason with Reinforcement Learning (Havrilla, 2024) View paper
  - [38] How Much Backtracking is Enough? Exploring the Interplay of SFT and RL in Enhancing LLM Reasoning (Wang Junlin, 2025) View paper
  - Non-Markovian and Advanced Theoretical Frameworks (1 papers)
  - [50] Beyond Markovian: Reflective Exploration via Bayes-Adaptive RL for LLM Reasoning (Zhang, 2025) View paper
- Application-Specific Exploration
  - Recommendation Systems and Information Retrieval (2 papers)
  - [28] Large Language Model driven Policy Exploration for Recommender Systems (Jie Wang, 2025) View paper
  - [30] Optimizing novelty of top-k recommendations using large language models and reinforcement learning (Amit Sharma, 2024) View paper
  - Multi-Agent and Action Space Pruning (1 papers)
  - [41] Knowing what not to do: Leverage language model insights for action space pruning in multi-agent reinforcement learning (Liu Zhi-hao, 2024) View paper
  - Tool Use and Structured Problem Solving (1 papers)
  - [8] Retool: Reinforcement learning for strategic tool use in llms (Feng, 2025) View paper
- Surveys and Integrative Frameworks (5 papers)
  - [23] A survey on enhancing reinforcement learning in complex environments: Insights from human and llm feedback (Ahmadabadi Majid Nili, 2024) View paper
  - [26] DeepSeek-Inspired Exploration of RL-Based LLMs and Synergy with Wireless Networks: A Survey (Qiao Yu, 2025) View paper
  - [29] Algorithm Discovery With LLMs: Evolutionary Search Meets Reinforcement Learning (Mansouri, 2025) View paper
  - [39] Exploring Advanced Large Language Models with LLMsuite (Roffo, 2024) View paper
  - [40] Real-time integration of fine-tuned large language model for improved decision-making in reinforcement learning (Xiancai Xiang, 2024) View paper

Narrative

Core task: Enhancing exploration in reinforcement learning for large language models. The field organizes itself around several complementary perspectives. Exploration Strategy Design and Mechanisms encompasses foundational techniques such as curiosity-driven methods (Curiosity-Driven Exploration LLM[4]) and risk-sensitive objectives (Risk-Sensitive RL LLM[0]), while Knowledge-Guided and Semantic Exploration leverages linguistic structure and prior knowledge to direct search (Language Guided Exploration[44]). Memory and Experience-Based Exploration focuses on replay mechanisms (Retrospective Replay[10], RLEP Experience Replay[31]) that reuse past trajectories, and Hierarchical and Structured Exploration addresses multi-level decision-making (Algorithm of Thoughts[15], RL of Thoughts[34]). Training Efficiency and Sample Optimization targets reducing computational overhead (Enhancing Efficiency Exploration RL[3], Efficient Exploration LLMs[6]), while Theoretical Analysis and Empirical Evaluation provides rigorous benchmarks (Survey RL Complex Environments[23]). Application-Specific Exploration tailors methods to domains like tool use (Retool Strategic Tool Use[8]) or recommendation (Optimizing Novelty Recommendations[30]), and Surveys and Integrative Frameworks synthesize cross-cutting insights.

Within Exploration Strategy Design, a central tension emerges between intrinsic motivation approaches that reward novelty (Intrinsic Motivation Exploration[36], Intrinsic Exploration LLM[48]) and objective-based methods that explicitly balance risk and reward. Risk-Sensitive RL LLM[0] sits in this latter cluster, emphasizing controlled exploration under uncertainty—contrasting with purely curiosity-driven schemes like Curiosity-Driven Exploration LLM[4] that prioritize information gain without explicit risk constraints. Nearby works such as Outcome-based Exploration[18] and Unlocking Reasoning Capabilities[27] also shape exploration via task-specific objectives, yet Risk-Sensitive RL LLM[0] distinguishes itself by incorporating risk-awareness into the exploration policy. This positions it as a bridge between classical RL safety concerns and modern LLM fine-tuning, addressing how agents can explore efficiently while respecting distributional or worst-case performance criteria.

Related Works in Same Category

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

1. Knapsack rl: Unlocking exploration of llms via optimizing budget allocation

Authors: Li, Ziniu, Chen CongLiang, Yang Tianyun, Ding Tian, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language Models (LLMs) can self-improve through reinforcement learning, where they generate trajectories to explore and discover better solutions. However, this exploration process is computationally expensive, often forcing current methods to assign limited exploration budgets to each task. This uniform allocation creates problematic edge cases: easy tasks consistently succeed while difficult tasks consistently fail, both producing zero gradients during training updates for the widely use...

Relationship Analysis

Both papers belong to the Risk-Sensitive and Objective-Based Exploration category, addressing exploration challenges in RL for LLMs by modifying optimization objectives. They share the goal of improving exploration beyond standard mean-reward optimization, particularly for mathematical reasoning tasks. The original paper introduces a risk-sensitive RL framework using exponential utility functions to interpolate between mean and maximum rewards, while the candidate paper frames exploration as a budget allocation problem using knapsack optimization to dynamically distribute computational resources across tasks based on their learning value and difficulty.

2. 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 Risk-Sensitive and Objective-Based Exploration category, addressing exploration challenges in RL for LLMs by modifying optimization objectives. They share the goal of improving solution diversity (pass@k) while maintaining accuracy (pass@1) in mathematical reasoning tasks. The original paper employs a risk-sensitive RL framework with exponential utility functions to interpolate between mean and maximum rewards, whereas the candidate paper focuses on outcome-based exploration bonuses (UCB-style methods and batch exploration) that directly penalize answer repetition and encourage diverse final outcomes.

3. Unlocking reasoning capabilities in llms via reinforcement learning exploration

Authors: Deng, Wenhao, Wei Long, Wenhao Deng, Yu Chenglei, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reinforcement learning with verifiable rewards (RLVR) has recently enhanced the reasoning capabilities of large language models (LLMs), particularly for mathematical problem solving. However, a fundamental limitation remains: as the sampling budget increases, the advantage of RLVR-trained models over their pretrained bases often diminishes or even vanishes, revealing a strong dependence on the base model's restricted search space. We attribute this phenomenon to the widespread use of the reverse...

Relationship Analysis

Both papers belong to the Risk-Sensitive and Objective-Based Exploration category, addressing exploration challenges in RLVR for LLMs by modifying RL objectives. They share the common goal of improving pass@k performance while maintaining pass@1 accuracy through risk-sensitive formulations. The original paper uses exponential utility functions to derive a risk-sensitive advantage that interpolates between mean and maximum rewards, while the candidate paper (RAPO) focuses on replacing reverse KL divergence with forward KL divergence and reward-aware reference policy reweighting to enable both out-of-distribution and in-distribution exploration.

Contributions Analysis

Overall novelty summary. The paper proposes a risk-sensitive reinforcement learning framework to address the exploration-exploitation dilemma in RLVR for LLMs, introducing the RS-GRPO algorithm that interpolates between mean and maximum rewards to enhance solution diversity. It resides in the 'Risk-Sensitive and Objective-Based Exploration' leaf, which contains four papers total, indicating a moderately sparse research direction within the broader 'Exploration Strategy Design and Mechanisms' branch. This leaf focuses specifically on modifying RL objectives to drive exploration, distinguishing it from intrinsic motivation approaches that rely on curiosity signals or uncertainty estimates.

The taxonomy reveals neighboring leaves addressing exploration through intrinsic rewards (four papers on curiosity-driven methods) and adaptive control mechanisms (three papers on dynamic exploration adjustment). The paper's risk-seeking objective diverges from these by explicitly balancing mean and maximum reward rather than relying on novelty bonuses or adaptive schedules. The broader 'Knowledge-Guided and Semantic Exploration' branch (nine papers across three leaves) represents an alternative paradigm using external knowledge or LLM-generated subgoals, while the paper's approach remains within objective-based exploration without external scaffolding. The scope note clarifies that standard policy gradient methods without exploration-specific modifications belong elsewhere, positioning this work as a deliberate departure from conventional RLVR training.

Among nineteen candidates examined across three contributions, none were identified as clearly refuting the work. The 'Risk-Sensitive Framework' contribution examined eight candidates with zero refutable matches, while 'RS-GRPO Algorithm' examined nine with similar results. The 'Theoretical Analysis' contribution reviewed two candidates, also finding no overlapping prior work. This limited search scope—focused on top-K semantic matches and citation expansion—suggests the specific combination of risk-sensitive objectives and GRPO adaptation for LLM exploration appears novel within the examined literature. However, the analysis does not claim exhaustive coverage of all possible prior work in risk-sensitive RL or exploration methods.

The work appears to occupy a distinct position within a moderately explored research direction, combining risk-aware objective design with practical algorithmic instantiation for LLM reasoning tasks. The absence of refuting candidates among nineteen examined suggests

novelty in the specific technical approach, though the limited search scope means broader connections to risk-sensitive RL literature outside the LLM-exploration context may exist. The taxonomy structure indicates this direction is less crowded than intrinsic motivation or knowledge-guided approaches, potentially offering room for methodological contributions.

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

Contribution 1: Risk-Sensitive Reinforcement Learning Framework for LLMs

Description: The authors propose a risk-sensitive RL framework that uses an exponential utility function to create a risk-seeking objective. This objective interpolates between optimizing mean reward and maximum reward, enabling policies to escape local optima induced by sharply peaked pretrained LLM distributions and discover more diverse reasoning strategies.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Risk-Aware Hierarchical Reinforcement Learning for Long-Range Autonomous Navigation in Off-Road Environments

URL: View paper

Brief Assessment

Risk-Aware Hierarchical Navigation[67] focuses on autonomous navigation for unmanned ground vehicles (UGVs) in off-road environments, not on reinforcement learning frameworks for large language models or exploration strategies in LLM fine-tuning.

---

2. Safe exploration techniques for reinforcement learningâ□□an overview

URL: View paper

Brief Assessment

Safe Exploration Overview[65] is a survey paper on safe exploration techniques in RL, focusing on risk-averse approaches for safety-critical domains. The original paper proposes a risk-seeking framework specifically for LLM exploration using exponential utility to interpolate between mean and maximum rewards, which is a different application domain and objective.

---

3. Risk-Sensitive RL for Alleviating Exploration Dilemmas in Large Language Models

URL: View paper

Brief Assessment

Risk-Sensitive RL Exploration[51] presents the same risk-sensitive RL framework with exponential utility for LLMs. Both papers use identical formulations and objectives for addressing exploration in LLM fine-tuning.

---

4. State-aware perturbation optimization for robust deep reinforcement learning

URL: View paper

Brief Assessment

State-aware Perturbation Optimization[61] focuses on adversarial robustness in deep RL for robotic control, not risk-sensitive RL frameworks for LLM reasoning exploration.

---

5. Bayesian robust optimization for imitation learning

URL: View paper

Brief Assessment

Bayesian Robust Imitation[64] focuses on imitation learning with uncertain reward functions in traditional MDPs, not on reinforcement learning for large language models or exploration in sharply peaked pretrained distributions.

---

6. Bridging Distributional and Risk-Sensitive Reinforcement Learning: Balancing Statistical, Computational, and Risk Considerations

URL: View paper

Brief Assessment

Bridging Distributional Risk-Sensitive[66] focuses on tabular MDPs with exponential utility functions for risk-sensitive control in traditional RL settings, not on large language model fine-tuning or exploration dilemmas in pretrained models. The candidate addresses computational efficiency in distributional RL for known MDP structures, while the original work targets LLM-specific challenges like escaping sharply peaked pretrained distributions.

---

7. Revisiting domain randomization via relaxed state-adversarial policy optimization

URL: View paper

Brief Assessment

Domain Randomization Relaxed[63] focuses on robust RL for sim-to-real transfer via state-adversarial perturbations in continuous control tasks, not on risk-sensitive objectives for LLM reasoning or exploration in language models.

---

8. Inductive biases in machine learning for robotics and control

URL: View paper

Brief Assessment

Inductive Biases Robotics[62] focuses on robotics and control applications with adversarial worst-case optimization, not on LLM fine-tuning with risk-seeking objectives that interpolate between mean and maximum rewards for exploration in language model reasoning tasks.

Contribution 2: RS-GRPO Algorithm

Description: The authors instantiate their risk-sensitive framework as RS-GRPO, a simple algorithm requiring only minor code modifications to existing GRPO implementations. It uses a risk-sensitive advantage function that dynamically re-weights optimization to emphasize hard prompts where the model performs poorly.

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. Risk-Aware Financial Portfolio Management with Distributional Deep Deterministic Policy Gradient

URL: View paper

Brief Assessment

Risk-Aware Portfolio Management[58] applies distributional DDPG to financial portfolio optimization with continuous state/action spaces, while the original paper's RS-GRPO targets discrete language model outputs for mathematical reasoning tasks. The domains, problem formulations, and algorithmic approaches are fundamentally different.

---

2. Risk-aware Direct Preference Optimization under Nested Risk Measure

URL: View paper

Brief Assessment

Risk-aware Direct Preference[55] focuses on preference optimization with nested risk measures and Bradley-Terry models for alignment, not on policy gradient algorithms with risk-sensitive advantage functions for exploration in RLVR settings.

---

3. Risk-sensitive policy optimization via predictive CVaR policy gradient

URL: View paper

Brief Assessment

Risk-sensitive CVaR Gradient[56] focuses on CVaR policy optimization in general RL settings with predictive tail probabilities, not specifically on LLM fine-tuning with risk-sensitive advantage functions for challenging prompts. The technical approaches and application domains differ substantially.

---

4. AdaRisk: risk-adaptive deep reinforcement learning for vulnerable nodes detection

URL: View paper

Brief Assessment

AdaRisk Vulnerable Nodes[54] addresses vulnerable node detection in uncertain graphs using risk-adaptive deep RL for graph mining, not policy gradient algorithms for language model alignment with risk-sensitive advantage functions for challenging prompts.

---

5. Risk-Sensitive RL for Alleviating Exploration Dilemmas in Large Language Models

URL: View paper

Brief Assessment

Risk-Sensitive RL Exploration[51] presents the same RS-GRPO algorithm with identical advantage function formulation and implementation as a drop-in replacement for GRPO.

---

6. Catastrophic-risk-aware reinforcement learning with extreme-value-theory-based policy gradients

URL: View paper

Brief Assessment

Catastrophic-risk-aware RL[59] focuses on minimizing catastrophic risk in sequential decision processes using extreme value theory for tail risk estimation, not on amplifying learning from challenging prompts in LLM training. The technical approaches and problem domains are fundamentally different.

---

7. A risk-sensitive approach to policy optimization

URL: View paper

Brief Assessment

Risk-sensitive Policy Optimization[57] focuses on general risk-sensitive RL using cumulative distribution functions over full-episode rewards in continuous control tasks, not specifically on amplifying learning from challenging prompts in LLM fine-tuning via risk-sensitive advantage functions as described in the original paper's RS-GRPO.

---

8. Policy Gradient Bayesian Robust Optimization for Imitation Learning

URL: View paper

Brief Assessment

Policy Gradient Bayesian[60] focuses on robust optimization under reward function uncertainty in imitation learning, not on risk-sensitive advantage functions for amplifying learning from challenging prompts in RLVR settings. The candidate addresses epistemic uncertainty over reward hypotheses, while the original addresses exploration dilemmas in LLM fine-tuning with verifiable rewards.

---

9. DSAC: Distributional Soft Actor-Critic for Risk-Sensitive Reinforcement Learning

URL: View paper

Brief Assessment

DSAC Distributional[53] focuses on risk-sensitive RL in continuous control tasks (e.g., MuJoCo, robotic navigation) using distributional value functions and quantile regression, not on policy gradient methods for LLM training with risk-sensitive advantage functions that amplify learning from challenging prompts.

Contribution 3: Theoretical and Empirical Analysis of Exploration Dilemma

Description: The authors provide both theoretical proofs (showing standard policy gradient can decrease optimal action probability while risk-sensitive gradient guarantees improvement) and empirical demonstrations (bandit experiments) that standard RL fails to escape local optima from sharply peaked initial policies, while their risk-sensitive approach succeeds.

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. Risk-Sensitive RL for Alleviating Exploration Dilemmas in Large Language Models

URL: View paper

Brief Assessment

Risk-Sensitive RL Exploration[51] provides the same theoretical proofs (Lemmas 2-4) and bandit experiments demonstrating that standard policy gradient fails while risk-sensitive gradient succeeds in escaping local optima.

---

2. AEAP: A Reinforcement Learning Actor Ensemble Algorithm with Adaptive Pruning

URL: View paper

Brief Assessment

AEAP Actor Ensemble[52] focuses on actor ensemble methods with adaptive pruning for continuous control tasks, not on risk-sensitive policy gradients or escaping local optima from sharply peaked initial policies in LLMs. The exploration mechanisms and theoretical frameworks are fundamentally different.

Appendix: Text Similarity Detection

Textual similarity detection checked 20 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Risk-Sensitive RL for Alleviating Exploration Dilemmas in Large Language Models

Detected in: Contribution: contribution_1, Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Risk-Sensitive Reinforcement Learning for Alleviating Exploration Dilemmas in Large Language Models View paper
- [1] Guiding Pretraining in Reinforcement Learning with Large Language Models View paper
- [2] ExploRLLM: Guiding Exploration in Reinforcement Learning with Large Language Models View paper
- [3] Enhancing efficiency and exploration in reinforcement learning for llms View paper
- [4] CDE: Curiosity-Driven Exploration for Efficient Reinforcement Learning in Large Language Models View paper
- [5] Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning View paper
- [6] Efficient Exploration for LLMs View paper
- [7] LLM-Explorer: A Plug-in Reinforcement Learning Policy Exploration Enhancement Driven by Large Language Models View paper
- [8] Retool: Reinforcement learning for strategic tool use in llms View paper
- [9] Trial and error: Exploration-based trajectory optimization for llm agents View paper
- [10] Improving RL Exploration for LLM Reasoning through Retrospective Replay View paper
- [11] Reinforcement Learning for Reasoning in Large Language Models with One Training Example View paper
- [12] Toward efficient exploration by large language model agents View paper
- [13] Knapsack rl: Unlocking exploration of llms via optimizing budget allocation View paper
- [14] Reasoning Under 1 Billion: Memory-Augmented Reinforcement Learning for Large Language Models View paper
- [15] Algorithm of Thoughts: Enhancing Exploration of Ideas in Large Language Models View paper
- [16] Disentangling exploration of large language models by optimal exploitation View paper
- [17] LLM-Guided Reinforcement Learning for Interactive Environments View paper
- [18] Outcome-based exploration for llm reasoning View paper
- [19] Comparing Exploration-Exploitation Strategies of LLMs and Humans: Insights from Standard Multi-armed Bandit Tasks View paper
- [20] Accelerating reinforcement learning of robotic manipulations via feedback from large language models View paper
- [21] Teaching Large Language Models to Reason with Reinforcement Learning View paper
- [22] Guiding Exploration in Reinforcement Learning Through LLM-Augmented Observations View paper
- [23] A survey on enhancing reinforcement learning in complex environments: Insights from human and llm feedback View paper
- [24] From Trial-and-Error to Improvement: A Systematic Analysis of LLM Exploration Mechanisms in RLVR View paper
- [25] Know when to explore: Difficulty-aware certainty as a guide for llm reinforcement learning View paper
- [26] DeepSeek-Inspired Exploration of RL-Based LLMs and Synergy with Wireless Networks: A Survey View paper
- [27] Unlocking reasoning capabilities in llms via reinforcement learning exploration View paper
- [28] Large Language Model driven Policy Exploration for Recommender Systems View paper
- [29] Algorithm Discovery With LLMs: Evolutionary Search Meets Reinforcement Learning View paper
- [30] Optimizing novelty of top-k recommendations using large language models and reinforcement learning View paper
- [31] Rlep: Reinforcement learning with experience replay for llm reasoning View paper
- [32] Parallel-R1: Towards Parallel Thinking via Reinforcement Learning View paper
- [33] Integrating Large Language Models and Reinforcement Learning for Non-Linear Reasoning View paper
- [34] Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning View paper
- [35] On Entropy Control in LLM-RL Algorithms View paper
- [36] Navigate the unknown: Enhancing llm reasoning with intrinsic motivation guided exploration View paper
- [37] Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning View paper
- [38] How Much Backtracking is Enough? Exploring the Interplay of SFT and RL in Enhancing LLM Reasoning View paper
- [39] Exploring Advanced Large Language Models with LLMsuite View paper
- [40] Real-time integration of fine-tuned large language model for improved decision-making in reinforcement learning View paper
- [41] Knowing what not to do: Leverage language model insights for action space pruning in multi-agent reinforcement learning View paper
- [42] Can large language models explore in-context? View paper
- [43] Evolve: Evaluating and optimizing llms for exploration View paper
- [44] Language Guided Exploration for RL Agents in Text Environments View paper
- [45] Behavior Injection: Preparing Language Models for Reinforcement Learning View paper
- [46] Option Discovery Using LLM-guided Semantic Hierarchical Reinforcement Learning View paper
- [47] Large Language Models Think Too Fast To Explore Effectively View paper
- [48] Efficient Reinforcement Learning for Large Language Models with Intrinsic Exploration View paper
- [49] Large Language Model-Enhanced Reinforcement Learning for Diverse and Novel Recommendations View paper
- [50] Beyond Markovian: Reflective Exploration via Bayes-Adaptive RL for LLM Reasoning View paper
- [51] Risk-Sensitive RL for Alleviating Exploration Dilemmas in Large Language Models View paper
- [52] AEAP: A Reinforcement Learning Actor Ensemble Algorithm with Adaptive Pruning View paper
- [53] DSAC: Distributional Soft Actor-Critic for Risk-Sensitive Reinforcement Learning View paper
- [54] AdaRisk: risk-adaptive deep reinforcement learning for vulnerable nodes detection View paper
- [55] Risk-aware Direct Preference Optimization under Nested Risk Measure View paper
- [56] Risk-sensitive policy optimization via predictive CVaR policy gradient View paper
- [57] A risk-sensitive approach to policy optimization View paper

- [58] Risk-Aware Financial Portfolio Management with Distributional Deep Deterministic Policy Gradient View paper
- [59] Catastrophic-risk-aware reinforcement learning with extreme-value-theory-based policy gradients View paper
- [60] Policy Gradient Bayesian Robust Optimization for Imitation Learning View paper
- [61] State-aware perturbation optimization for robust deep reinforcement learning View paper
- [62] Inductive biases in machine learning for robotics and control View paper
- [63] Revisiting domain randomization via relaxed state-adversarial policy optimization View paper
- [64] Bayesian robust optimization for imitation learning View paper
- [65] Safe exploration techniques for reinforcement learningâ□□an overview View paper
- [66] Bridging Distributional and Risk-Sensitive Reinforcement Learning: Balancing Statistical, Computational, and Risk Considerations View paper
- [67] Risk-Aware Hierarchical Reinforcement Learning for Long-Range Autonomous Navigation in Off-Road Environments View paper