Novelty Assessment Report

Paper: ExGRPO: Learning to Reason from Prior Successes

PDF URL: https://openreview.net/pdf?id=701tjQXWVk

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-07

Abstract

Reinforcement learning from verifiable rewards (RLVR) is an emerging paradigm for improving the reasoning ability of large language models. However, standard on-policy training discards rollout experiences after a single update, leading to computational inefficiency and instability. While prior work on RL has highlighted the benefits of reusing past experience, the role of experience characteristics in shaping learning dynamics of large reasoning models remains underexplored. In this paper, we are the first to investigate what makes a reasoning experience valuable and identify rollout correctness and entropy as effective indicators of experience value. Based on these insights, we propose ExGRPO (Experiential Group Relative Policy Optimization), a framework that organizes and prioritizes valuable experiences, and employs a mixed-policy objective to balance exploration with experience exploitation. Experiments on five backbone models (1.5B-8B parameters) show that ExGRPO consistently improves reasoning performance on mathematical/general benchmarks, with an average gain of +3.5/7.6 points over on-policy RLVR. Moreover, ExGRPO stabilizes training on both stronger and weaker models where on-policy methods fail. These results highlight principled experience management as a key ingredient for efficient and scalable RLVR.

Disclaimer

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

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

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

Core Task Landscape

This paper addresses: **Reinforcement Learning from Verifiable Rewards for Language Model Reasoning**

A total of **41 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

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

- **Core RLVR Algorithms and Training Dynamics**
- **Reasoning Capability Analysis and Evaluation**
- **Verification and Reward Design**
- **Domain-Specific Applications and Extensions**
- **Beyond Verifiable Domains**
- **Multi-Objective and Multi-Domain Training**
- **Training Challenges and Mitigation Strategies**
- **Safety and Alignment**
- **Surveys and Methodological Reviews**
- **Peripheral and Tangential Work**

Complete Taxonomy Tree

- Reinforcement Learning from Verifiable Rewards for Language Model Reasoning Survey Taxonomy
- Core RLVR Algorithms and Training Dynamics
  - Policy Optimization Methods and Theoretical Analysis (3 papers)
  - [3] Reinforcement Learning with Verifiable Rewards: GRPO's Effective Loss, Dynamics, and Success Amplification (Mroueh, 2025) View paper
  - [29] Random Policy Valuation is Enough for LLM Reasoning with Verifiable Rewards (He, 2025) View paper
  - [34] Sharpness-Controlled Group Relative Policy Optimization with Token-Level Probability Shaping (Tue Le, 2025) View paper
  - Experience Management and Replay Strategies ★ (1 papers)
  - [0] ExGRPO: Learning to Reason from Prior Successes (Anon et al., 2026) View paper
  - Exploration Strategies and Diversity Mechanisms (2 papers)
  - [10] Diversity-incentivized exploration for versatile reasoning (Hu, 2025) View paper
  - [38] Low-probability Tokens Sustain Exploration in Reinforcement Learning with Verifiable Reward (Huang Guan-hua, 2025) View paper
- Reasoning Capability Analysis and Evaluation
  - Reasoning Capability Emergence and Boundaries (3 papers)
  - [1] Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? (Yue Yang, 2025) View paper
  - [2] Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs (Liu Zihan, 2025) View paper
  - [37] The Reasoning Boundary Paradox: How Reinforcement Learning Constrains Language Models (Nguyen Phuc Minh, 2025) View paper
  - Long Chain-of-Thought Reasoning Mechanics (2 papers)
  - [5] Demystifying long chain-of-thought reasoning in llms (Tong, 2025) View paper
  - [21] UloRL:An Ultra-Long Output Reinforcement Learning Approach for Advancing Large Language Models' Reasoning Abilities (Du, 2025) View paper

  - Measurement and Benchmarking Challenges (1 papers)
  - [36] Position: The Hidden Costs and Measurement Gaps of Reinforcement Learning with Verifiable Rewards (Xuan, 2025) View paper
- Verification and Reward Design
  - Process Reward Models and Step-Level Verification (1 papers)
  - [6] Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning (Setlur, 2024) View paper
  - Outcome-Based Verification and Ranking (2 papers)
  - [7] Generative Verifiers: Reward Modeling as Next-Token Prediction (Zhang Lun-jun, 2024) View paper
  - [26] Learning to Rank Chain-of-Thought: An Energy-Based Approach with Outcome Supervision (Luo, 2025) View paper
  - Self-Verification and Critique Mechanisms (2 papers)
  - [20] Trust, But Verify: A Self-Verification Approach to Reinforcement Learning with Verifiable Rewards (Liu Xiao-yuan, 2025) View paper
  - [32] Critique-RL: Training Language Models for Critiquing through Two-Stage Reinforcement Learning (Xi, 2025) View paper
  - Knowledge-Grounded Verification (1 papers)
  - [31] Brittleness and Promise: Knowledge Graph Based Reward Modeling for Diagnostic Reasoning (Cheng He, 2025) View paper
  - Composite and Anti-Hacking Reward Design (1 papers)
  - [17] Reward Hacking Mitigation using Verifiable Composite Rewards (Tarek, 2025) View paper
- Domain-Specific Applications and Extensions
  - Formal Theorem Proving (2 papers)
  - [14] Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction (Lin Yong, 2025) View paper
  - [16] Leanabell-prover-v2: Verifier-integrated reasoning for formal theorem proving via reinforcement learning (Liu Yahui, 2025) View paper
  - Mathematical and Coding Tasks (1 papers)
  - [13] REASONING GYM: Reasoning Environments for Reinforcement Learning with Verifiable Rewards (Stojanovski, 2025) View paper
  - Visual and Chart Reasoning (1 papers)
  - [4] Chart-RVR: Reinforcement Learning with Verifiable Rewards for Explainable Chart Reasoning (Sinha, 2025) View paper
  - Logical Reasoning and Symbolic Feedback (2 papers)
  - [28] Rlsf: Reinforcement learning via symbolic feedback (Piyush Jha, 2024) View paper
  - [41] RLSF: Reinforcement Learning from Self-feedback for improved logical reasoning (M Sutton, n.d.) View paper
- Beyond Verifiable Domains
  - Unverifiable Data and Proxy Rewards (3 papers)
  - [9] Beyond Verifiable Rewards: Scaling Reinforcement Learning for Language Models to Unverifiable Data (Tang, 2025) View paper
  - [18] RLPR: Extrapolating RLVR to General Domains without Verifiers (Yu Tianyu, 2025) View paper
  - [25] Language Models that Think, Chat Better (Bhaskar, 2025) View paper
  - Instruction Following and Content Moderation (2 papers)
  - [22] IFDECORATOR: Wrapping Instruction Following Reinforcement Learning with Verifiable Rewards (Guo, 2025) View paper
  - [27] Scaling Reinforcement Learning for Content Moderation with Large Language Models (Hamed Firooz, 2025) View paper
  - Forecasting and Noisy Outcomes (1 papers)
  - [33] Outcome-based Reinforcement Learning to Predict the Future (Hewitt, 2025) View paper
- Multi-Objective and Multi-Domain Training
  - Multi-Domain Reasoning and Transfer (1 papers)
  - [23] Can one domain help others? a data-centric study on multi-domain reasoning via reinforcement learning (Li Yu, 2025) View paper
  - Simultaneous Multi-Objective Alignment (1 papers)
  - [39] Simultaneous Multi-objective Alignment Across Verifiable and Non-verifiable Rewards (Shen, 2025) View paper
- Training Challenges and Mitigation Strategies
  - Weakness-Driven Problem Synthesis (1 papers)
  - [30] SwS: Self-aware Weakness-driven Problem Synthesis in Reinforcement Learning for LLM Reasoning (Liang Xiao, 2025) View paper
  - Truthfulness and Hallucination Mitigation (1 papers)
  - [11] TruthRL: Incentivizing truthful LLMs via reinforcement learning (Wei, 2025) View paper
- Safety and Alignment (1 papers)
  - [24] Breaking the Safety-Capability Tradeoff: Reinforcement Learning with Verifiable Rewards Maintains Safety Guardrails in LLMs (Dongkyu Derek Cho, 2025) View paper
- Surveys and Methodological Reviews (2 papers)
  - [8] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle (Liu Ke-liang, 2025) View paper
  - [12] Trust but verify! a survey on verification design for test-time scaling (Rathee, 2025) View paper
- Peripheral and Tangential Work (4 papers)
  - [15] Writing-Zero: Bridge the Gap Between Non-verifiable Problems and Verifiable Rewards (Lu, 2025) View paper
  - [19] Shielded learning for resilience and performance based on statistical model checking in simulink (Julius Adelt, 2023) View paper
  - [35] A Critical Analysis of the Proposed Recursive Logic Subsystem for Self-Learning LLMs in Scientific Discovery (Ghosh, 2025) View paper
  - [40] Knowledge-to-Verification: Unlocking Reinforcement Learning with Verifiable Rewards for LLMs in Knowledge-Intensive Domains (Z Yuan, n.d.) View paper

Narrative

Core task: reinforcement learning from verifiable rewards for language model reasoning. The field has organized itself around several major branches that reflect different facets of this challenge. Core RLVR Algorithms and Training Dynamics focuses on the mechanics of policy optimization and experience management, exploring how to efficiently leverage verifiable feedback signals. Reasoning Capability Analysis and Evaluation examines what models learn and how well they generalize, while Verification and Reward Design addresses the construction of reliable reward signals across domains. Domain-Specific Applications and Extensions targets particular problem settings such as mathematical theorem proving and code generation, whereas Beyond Verifiable Domains and Multi-Objective and Multi-Domain Training consider settings where clean verification is unavailable or multiple objectives must be balanced. Training Challenges and Mitigation Strategies tackles practical issues like reward hacking and distribution shift, and Safety and Alignment ensures that capability gains do not compromise model safety. Surveys and Methodological Reviews provide broader perspectives, with works like RL LLM Survey[8] and Verification Design Survey[12] synthesizing key themes.

Within this landscape, a particularly active line of work centers on experience management and replay strategies, where ExGRPO[0] sits. This branch investigates how to make the most of collected trajectories during training, contrasting with approaches that emphasize success amplification such as GRPO Success Amplification[3] or those that focus on diversity and exploration like Diversity Exploration[10]. ExGRPO[0] emphasizes efficient reuse of experience through replay mechanisms, addressing the sample efficiency challenges that arise when verifiable rewards are expensive to obtain. Nearby works such as Rewarding Progress[6] and Long Chain Thought[5] explore complementary themes around credit assignment and extended reasoning chains, while RL Incentivize Reasoning[1] and Verifiable Rewards Reasoning[2] provide broader algorithmic perspectives on how reinforcement learning can be tailored to reasoning tasks. The positioning of ExGRPO[0] reflects an ongoing tension between maximizing data efficiency and maintaining exploration breadth, a trade-off that remains central to scaling RLVR methods.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on data management techniques for improving sample efficiency through experience replay and prioritization, while its siblings address complementary aspects of the RLVR training pipeline. Exploration Strategies handles diversity-driven exploration mechanisms that generate varied experiences, and Policy Optimization covers the algorithmic and theoretical foundations of how those experiences are used to update models. Together, these subtopics form a natural division between data generation (exploration), data management (experience replay), and data utilization (policy optimization).

Similarities: - All three subtopics aim to improve training efficiency and stability in RLVR systems - Each addresses a distinct phase of the rollout-to-update cycle in reinforcement learning - All are concerned with sample efficiency, though through different mechanisms (diversity, reuse, or optimization)

Differences: - Experience Management focuses on organizing and reusing collected data, while Exploration Strategies focuses on generating diverse data through probability shaping and entropy objectives - Policy Optimization emphasizes algorithmic updates and theoretical guarantees, whereas Experience Management emphasizes data storage, prioritization, and replay mechanisms - Exploration operates at rollout generation time, Experience Management at the storage/retrieval phase, and Policy Optimization at the gradient update phase

Suggested Search Directions: - Interactions between exploration diversity and replay prioritization (e.g., how diversity metrics inform experience selection) - Theoretical analysis of how experience replay affects policy gradient convergence guarantees - Joint optimization of exploration bonuses and replay buffer management for RLVR

Sibling Subtopics

- **Exploration Strategies and Diversity Mechanisms** (leaves: 1, papers: 2)
- Scope: Techniques for incentivizing exploration through diversity objectives, entropy management, and token-level probability shaping.
- Exclude: Experience prioritization methods belong under Experience Management; training stability issues belong under Training Challenges.
- **Policy Optimization Methods and Theoretical Analysis** (leaves: 1, papers: 3)
- Scope: Research on GRPO variants, policy gradient methods, and theoretical foundations of RLVR optimization.
- Exclude: Experience replay and data management strategies belong under Experience Management below.

Contributions Analysis

Overall novelty summary. The paper proposes ExGRPO, a framework for managing and replaying reasoning experiences in RLVR training, identifying rollout correctness and entropy as indicators of experience value. According to the taxonomy, this work sits in the 'Experience Management and Replay Strategies' leaf under 'Core RLVR Algorithms and Training Dynamics'. Notably, this leaf contains only one paper—the original submission itself—indicating a relatively sparse research direction within the broader RLVR landscape. The taxonomy shows 41 total papers across the field, with most concentrated in verification design, domain applications, and policy optimization methods.

The taxonomy reveals that neighboring leaves focus on policy optimization theory (3 papers) and exploration strategies (2 papers), suggesting that experience management has received less direct attention than algorithmic foundations or diversity mechanisms. The scope note for this leaf explicitly excludes diversity-focused exploration, positioning ExGRPO as complementary to works like Diversity Exploration that incentivize broad sampling. The broader 'Core RLVR Algorithms' branch contains multiple active directions, but experience replay specifically appears underexplored compared to verification design (5 sub-categories) and domain applications (4 sub-categories).

Among 30 candidates examined, the contribution-level analysis shows mixed novelty signals. The identification of valuable experience characteristics (Contribution 1) examined 10 candidates with 3 appearing to provide overlapping prior work. The ExGRPO framework itself (Contribution 2) and performance improvements (Contribution 3) each examined 10 candidates with 1 refutable match apiece. These statistics suggest that while some aspects of experience characterization have precedent in the limited search scope, the integrated framework and empirical validation may offer incremental advances. The relatively low refutation counts should be interpreted cautiously given the 30-candidate search scale.

Based on the limited literature search, the work appears to address a gap in experience management for RLVR, though the search scope (30 candidates from semantic matching) cannot confirm exhaustive novelty. The taxonomy structure suggests this direction is less crowded than verification design or domain applications, but the contribution-level statistics indicate that key ideas around experience value and replay have some precedent among examined candidates. A more comprehensive search would be needed to definitively assess originality.

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

Contribution 1: Identification of valuable reasoning experience characteristics

Description: The authors systematically analyze reasoning experiences in RLVR and identify two key properties that determine experience value: rollout correctness for questions (with medium-difficulty questions being most valuable) and trajectory entropy (with lower entropy indicating better reasoning quality). This analysis provides empirical guidelines for experience selection in reinforcement learning for reasoning models.

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. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning

URL: View paper

Brief Assessment

Uncertainty Advantage Shaping[59] focuses on uncertainty-aware advantage shaping at response and token levels to prevent entropy collapse, not on identifying valuable experience characteristics based on rollout correctness and trajectory entropy for experience selection in replay buffers.

2. EFRame: Deeper Reasoning via Exploration-Filtering-Replay Reinforcement Learning Framework

URL: View paper

Brief Assessment

EFRame[56] focuses on sample quality categorization (high-quality vs. low-quality samples) for filtering purposes, not on systematic analysis of rollout correctness and trajectory entropy as indicators of experience value for selection in RLVR.

3. Staying in the Sweet Spot: Responsive Reasoning Evolution via Capability-Adaptive Hint Scaffolding

URL: View paper

Brief Assessment

Capability Adaptive Scaffolding[54] focuses on dynamically adjusting hint length to control problem difficulty for optimal rollout accuracy (~50%), not on identifying experience value through rollout correctness buckets and trajectory entropy as quality indicators for experience replay.

4. First return, entropy-eliciting explore

URL: View paper

Prior Art Analysis

Entropy Eliciting Explore[55] demonstrates prior work that systematically identifies valuable reasoning experience characteristics using entropy-based metrics. The candidate paper explicitly analyzes token-level entropy to identify high-uncertainty decision points in reasoning trajectories and uses these entropy signals to guide experience selection. This directly overlaps with the original paper's claim of being the first to identify rollout correctness and trajectory entropy as effective indicators of experience value.

Evidence

Evidence 1 - **Rationale**: This pair shows that Entropy Eliciting Explore[55] already used entropy as an indicator of valuable reasoning positions, challenging the original paper's claim of being first to identify entropy as an effective indicator of experience value. - **Original**: we are the first to investigate what makes a reasoning experience valuable and identify rollout correctness and entropy as effective indicators of experience value. - **Candidate**: To identify positions in the trajectory that exhibit high uncertainty - and thus are suitable for exploration - we compute the token-wise entropy at each positionk. specifically, let: πθ(v|q,t<k) (4) denote the softmax-normalized probability distribution over the vocabulary at stepk, conditioned on ...

Evidence 2 - **Rationale**: This demonstrates that Entropy Eliciting Explore[55] systematically analyzed entropy patterns in trajectories to identify valuable reasoning segments, which overlaps with the original paper's contribution of identifying trajectory entropy as an effective metric for experience quality. - **Original**: Through systematic analysis, we identify rollout correctness (for questions) and trajectory entropy (for trajectories) as effective online proxy metrics for characterizing experience quality. specifically, tasks of intermediate difficulty and their associated low-entropy trajectories tend to be bene... - **Candidate**: As shown in figure 2, frequent tokens with high average entropy can be identified and used as key segmentation points. this figure is adapted from prior work [33], which analyzes token-level uncertainty across model trajectories. these entropysensitive positions serve as natural breakpoints for segm...

Evidence 3 - **Rationale**: Both papers analyze experience properties from trajectories, with Entropy Eliciting Explore[55] explicitly identifying high-uncertainty decision points through entropy analysis, which demonstrates prior work on characterizing valuable reasoning experiences. - **Original**: we hypothesize that experience utility varies with measurable properties. 1in the context of rlvr, the experience refers to a state-action-reward trajectory during rollout of the reasoning chain. we will use the terms experience/trajectory/rollout interchangeably throughout the paper. under review a... - **Candidate**: we proposefr3e (first return, entropy-eliciting explore), a structured exploration framework that identifies high-uncertainty decision points in reasoning trajectories and performs targeted rollouts to construct semantically grounded intermediate feedback. our method provides targeted guidance witho...

5. ExGRPO: Learning to reason from experience

URL: View paper

Prior Art Analysis

ExGRPO Experience[53] demonstrates that the identification of valuable reasoning experience characteristics using rollout correctness and trajectory entropy was already established prior to the original paper's submission. The candidate paper explicitly states they are 'the first to investigate what makes a reasoning experience valuable and identify rollout correctness and entropy as effective indicators of experience value.' Both papers systematically analyze reasoning experiences using the same two key properties: rollout correctness for categorizing question difficulty and trajectory entropy for assessing reasoning quality. The candidate paper's preliminary study (Section 3.2) presents identical findings that medium-difficulty questions (based on rollout correctness) and low-entropy trajectories are most valuable for training.

Evidence

Evidence 1 - **Rationale**: Both papers claim to be 'the first' to identify rollout correctness and entropy as indicators of experience value, using identical language. This demonstrates the candidate established this contribution before the original paper. - **Original**: we are the first to investigate what makes a reasoning experience valuable and identify rollout correctness and entropy as effective indicators of experience value. - **Candidate**: we are the first to investigate what makes a reasoning experience valuable and identify rollout correctness and entropy as effective indicators of experience value.

Evidence 2 - **Rationale**: The candidate paper presents the exact same systematic analysis identifying rollout correctness and trajectory entropy as proxy metrics, with identical conclusions about intermediate difficulty and low-entropy trajectories. - **Original**: Through systematic analysis, we identify rollout correctness (for questions) and trajectory entropy (for trajectories) as effective online proxy metrics for characterizing experience quality. specifically, tasks of intermediate difficulty and their associated low-entropy trajectories tend to be bene... - **Candidate**: Through systematic analysis, we identify rollout correctness (for questions) and trajectory entropy (for trajectories) as effective online proxy metrics for characterizing experience quality. specifically, tasks of intermediate difficulty and their associated low-entropy trajectories tend to be bene...

Evidence 3 - **Rationale**: Both papers derive identical guidelines from their analysis, using the same terminology and conclusions about medium-difficulty questions and entropy minimization for trajectory selection. - **Original**: we highlight two guidelines:medium-difficultyquestionsprovide the most valuable optimization signals, andentropy minimization is an effective heuristic fortrajectoryselection.

- **Candidate**: we highlight two guidelines:medium-difficultyquestionsprovide the most valuable optimization signals, andentropy minimization is an effective heuristic fortrajectoryselection.

Evidence 4 - **Rationale**: The candidate paper presents the identical methodology for categorizing questions by rollout correctness, using the exact same difficulty buckets and thresholds as claimed in the original contribution. - **Original**: are all questions equally useful for training?during training, we categorize each question q into one of three difficulty buckets based on its online (rollout) correctness rate:easy[75%,100%), medium (25%,75%] , andhard (0,25%] - **Candidate**: are all questions equally useful for training?during training, we categorize each question q into one of three difficulty buckets based on its online (rollout) correctness rate:easy[75%,100%), medium (25%,75%] , andhard (0,25%]

6. Exploring Multi-Temperature Strategies for Token-and Rollout-Level Control in RLVR

URL: View paper

Brief Assessment

Multi Temperature Strategies[57] focuses on temperature-based exploration strategies during token generation and rollout sampling, not on identifying valuable experience characteristics through rollout correctness and trajectory entropy for experience replay buffer management.

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

URL: View paper

Brief Assessment

Evolving Without Labels[61] focuses on label-free self-improvement using majority voting and novelty rewards during inference, not on analyzing rollout correctness and trajectory entropy as indicators of experience value for RLVR training as in the original paper.

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

URL: View paper

Brief Assessment

ARES[52] focuses on adaptive reasoning via window-entropy tokens to control exploration depth based on difficulty, not on identifying valuable experience characteristics using rollout correctness and trajectory entropy for experience replay in RLVR.

9. PEAR: Phase Entropy Aware Reward for Efficient Reasoning

URL: View paper

Brief Assessment

PEAR[58] focuses on entropy as a control signal for response length efficiency in reasoning models, not on identifying valuable experience characteristics for reinforcement learning experience replay as in the original paper.

10. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning

URL: View paper

Prior Art Analysis

Entropy Performance Exchange[60] demonstrates that prior work has already systematically analyzed reasoning experiences using rollout correctness and entropy as key indicators of experience value. The candidate paper divides training into stages based on entropy dynamics and identifies that entropy reduction in negative samples and low-entropy trajectories are critical for learning - findings that directly overlap with the original paper's claimed novel identification of rollout correctness and trajectory entropy as valuable experience characteristics.

Evidence

Evidence 1 - **Rationale**: Both papers systematically analyze entropy as a key indicator for understanding valuable reasoning experiences in RLVR, suggesting the original paper was not the first to investigate this relationship. - **Original**: we are the first to investigate what makes a reasoning experience valuable and identify rollout correctness and entropy as effective indicators of experience value. based on these insights, we proposeexgrpo(experiential grouprelativepolicyoptimization), a framework that organizes and prioritizes val... - **Candidate**: we conduct a systematic empirical analysis of the entropy-performance exchange mechanism of rlvr across different levels of granularity. specifically, we first divide the training process into two distinct stages based on entropy dynamics, i.e., rising stageand plateau stage, and then systematically...

Evidence 2 - **Rationale**: Both papers identify entropy characteristics of trajectories as indicators of experience quality for RLVR optimization, with the candidate demonstrating similar systematic analysis of entropy's role in learning. - **Original**: we hypothesize that experience utility varies with measurable properties. we study experience properties from these two components. through systematic analysis, we identify rollout correctness (for questions) and trajectory entropy (for trajectories) as effective online proxy metrics for characteriz... - **Candidate**: our analysis reveals that, in the rising stage, entropy reduction in negative samples facilitates the learning of effective reasoning patterns, which in turn drives rapid performance gains. moreover, in the plateau stage, learning efficiency strongly correlates with high-entropy tokens present in lo...

Evidence 3 - **Rationale**: Both papers analyze the relationship between entropy and reasoning quality, identifying that lower entropy correlates with better reasoning outcomes, demonstrating prior work on this characteristic. - **Original**: does lower entropy imply better reasoning?we then test which metric can serve as an online proxy for reasoning quality. While outcome-based rewards check the final answer, they do not capture whether the reasoning cot is logically correct or valid. as shown in figure 1b, correct reasoning trajectori... - **Candidate**: entropy reduction mainly stems from negative samples. as shown in fig. 2a, negative samples consistently exhibit higher average policy entropy than positive samples. More importantly, their entropy declines at a substantially more rapid rate during the rising stage. also, tokens that appear exclusiv...

Contribution 2: ExGRPO framework for experience management and replay

Description: The authors introduce ExGRPO, a novel framework that maintains a replay buffer of reasoning trajectories, organizes them into buckets by correctness levels, and uses a sampling strategy that prioritizes beneficial experiences with lowest entropy trajectories. The framework combines on-policy exploration with strategic experience replay through a mixed-policy optimization objective.

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. Anti-jamming routing for internet of satellites: a reinforcement learning approach

URL: View paper

Brief Assessment

Anti Jamming Routing[70] focuses on routing optimization for satellite networks using experience replay buffers in a traditional RL context, not on reasoning trajectory management or mixed-policy optimization for large language models with verifiable rewards.

2. A prioritized objective actor-critic method for deep reinforcement learning

URL: View paper

Brief Assessment

Prioritized Objective Actor[65] focuses on prioritized experience replay for actor-critic methods in general deep RL settings, not specifically on reasoning trajectory management with correctness-based bucketing and entropy-driven selection for large language model reasoning tasks.

3. Sample-efficient LLM Optimization with Reset Replay

URL: View paper

Prior Art Analysis

Reset Replay[68] demonstrates that prior work exists on experience replay mechanisms for LLM optimization. The candidate paper introduces LORR (LLM Optimization with Reset Replay), which maintains a replay buffer of experiences, employs strategic sampling from this buffer, and uses a mixed-policy optimization objective combining on-policy and off-policy data. These core mechanisms—maintaining replay buffers, strategic experience selection, and mixed-policy objectives—directly overlap with the original paper's claimed novel contributions in ExGRPO. Both papers address the same fundamental problem of improving sample efficiency through experience replay in LLM reasoning tasks.

Evidence

Evidence 1 - **Rationale**: Both papers propose frameworks that use experience replay mechanisms to enhance sample efficiency in LLM optimization, challenging the novelty claim that ExGRPO was the first to introduce such a framework. - **Original**: we proposeexgrpo(experiential grouprelativepolicyoptimization), a framework that organizes and prioritizes valuable experiences, and employs a mixed-policy objective to balance exploration with experience exploitation. - **Candidate**: we introduce llm optimization with reset replay (lorr), a general and powerful plugin designed to enhance sample efficiency in any preference-based optimization framework. lorr's core mechanism enables training at a high replay number, maximizing the utility of each collected data batch.

Evidence 2 - **Rationale**: Both papers describe maintaining replay buffers with strategic sampling mechanisms based on experience quality, demonstrating prior work on this concept before the original paper's submission. - **Original**: exgrpo maintains a replay buffer of reasoning trajectories derived from partially correct rollouts and organizes them into buckets according to their correctness levels. To manage the buffer effectively, it uses a sampling strategy that prioritizes experiences from the most beneficial buckets - **Candidate**: we develop a replay strategy into the reset loop. specifically, we first sample multiple trajectories {y1, · ··, yk} for each prompt x, and then re-annotate them with a reward verifier r that selects the highest-scoring one as yw and the lowest-scoring one as yl

Evidence 3 - **Rationale**: Both papers describe mechanisms for reusing collected experiences through multiple updates, demonstrating that the concept of experience replay with mixed objectives was not novel to ExGRPO. - **Original**: The overall exgrpo objective is a combination of two components: theon-policy objective jon(θ) follows grpo in eq. 3: for each query q~ bon, k trajectories {oi}k i=1 are sampled from policy πθold to form an advantage group gq - **Candidate**: the core idea is to update the llm parameters multiple times for each batch finetuning with the dataset, which is achieved by setting a high replay number while keeping the batch size constant.

4. Hybrid attention-oriented experience replay for deep reinforcement learning and its application to a multi-robot cooperative hunting problem

URL: View paper

Brief Assessment

Hybrid Attention Replay[66] focuses on multi-agent cooperative tasks using MADDPG with attention mechanisms for experience selection, not on reasoning trajectory management for large language models with GRPO-based optimization.

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

URL: View paper

Brief Assessment

Query Policy Misalignment[63] focuses on preference-based RL with human feedback and query selection for reward learning, not on experience replay buffers for reasoning trajectories in verifiable reward settings.

6. Experience Replay-based Deep Reinforcement Learning for Dialogue Management Optimisation

URL: View paper

Brief Assessment

Dialogue Management Replay[67] applies experience replay to dialogue policy optimization in spoken dialogue systems, not to reasoning trajectory management in large language models. The domains, objectives, and technical implementations differ fundamentally.

7. Cooperative Traffic Scheduling in Transportation Network: A Knowledge Transfer Method

URL: View paper

Brief Assessment

Traffic Knowledge Transfer[64] focuses on traffic scheduling in transportation networks using knowledge transfer methods, not on experience replay buffers or mixed-policy optimization for reinforcement learning reasoning tasks with language models.

8. Experience Consistency Distillation Continual Reinforcement Learning for Robotic Manipulation Tasks

URL: View paper

Brief Assessment

Experience Consistency Distillation[69] focuses on continual reinforcement learning for robotic manipulation tasks with experience distillation and compression, not on reasoning trajectory replay with correctness-based bucketing and entropy-guided selection for large language models.

9. Relay Hindsight Experience Replay: Continual Reinforcement Learning for Robot Manipulation Tasks with Sparse Rewards

URL: View paper

Brief Assessment

Relay Hindsight Replay[71] focuses on robot manipulation tasks with hindsight experience replay for goal-conditioned RL, not on reasoning trajectory management for large language models with mixed-policy optimization.

10. Experience Replay for Continual Learning

URL: View paper

Brief Assessment

Experience Replay Continual[62] focuses on continual learning across different tasks to prevent catastrophic forgetting, not on optimizing reasoning trajectories within RLVR. The candidate addresses task-switching scenarios, while the original targets reasoning quality improvement within single-domain mathematical problems.

Contribution 3: Consistent performance improvements and training stabilization

Description: The authors demonstrate that ExGRPO achieves substantial performance gains across five backbone models (1.5B-8B parameters) on both in-distribution mathematical reasoning and out-of-distribution benchmarks. Notably, ExGRPO successfully stabilizes training on models where standard on-policy RLVR collapses, such as Llama-3.1 8B base model.

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. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning

URL: View paper

Brief Assessment

Logic RL[46] focuses on rule-based RL for logic puzzles with different training dynamics and evaluation metrics. The candidate does not demonstrate prior work on stabilizing RLVR training across multiple model sizes (1.5B-8B) or preventing training collapse on specific models like Llama-3.1 8B base.

2. Training Language Models to Reason Efficiently

URL: View paper

Brief Assessment

Efficient Reasoning Training[49] focuses on reducing inference costs by training models to generate shorter chain-of-thoughts while maintaining accuracy, not on stabilizing RL training across different model sizes or preventing training collapse as in the original paper's ExGRPO method.

3. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning

URL: View paper

Brief Assessment

L1 Thinking Control[45] focuses on controlling reasoning length through LCPO for test-time compute allocation, not on stabilizing RL training across different model sizes or preventing training collapse as in the original paper's ExGRPO method.

4. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models

URL: View paper

Brief Assessment

ProRL[42] focuses on prolonged RL training (2k+ steps) with KL divergence control and reference policy resetting to expand reasoning boundaries. The original paper addresses training stabilization through experience replay and bucketing strategies for different model sizes (1.5B-8B). These are distinct technical approaches to different aspects of RL training stability.

5. Reason-rft: Reinforcement fine-tuning for visual reasoning

URL: View paper

Brief Assessment

Reason RFT[48] focuses on visual reasoning tasks for vision-language models using a two-stage SFT+GRPO approach, while the original paper addresses mathematical reasoning with experience replay mechanisms. The candidate does not demonstrate prior work on stabilizing RL training across different model sizes for reasoning tasks through experience management.

6. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models

URL: View paper

Brief Assessment

Entropy Mechanism[47] focuses on entropy collapse prevention through theoretical analysis of entropy dynamics and covariance-based token clipping. The ORIGINAL paper addresses training stabilization through experience replay and trajectory selection mechanisms, representing different technical approaches to the stability problem.

7. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model

URL: View paper

Brief Assessment

Open Reasoner Zero[44] focuses on scaling RL training on base models using PPO with GAE, while the original paper addresses stabilizing RLVR training across different backbone models using experience replay mechanisms. These represent different technical approaches to training stability.

8. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr

URL: View paper

Prior Art Analysis

Dual Token Constraints[51] demonstrates that their method achieves substantial performance gains across multiple model sizes and successfully stabilizes training where standard on-policy methods fail. Specifically, they report stabilizing training on the llama-3.1 8b base model where on-policy rlvr collapses, and show consistent improvements across different model architectures (1.5b-8b parameters). The paper presents evidence of training stabilization through entropy control and dual-token constraints, addressing the same core challenge of training stability across different model sizes that the original paper claims as novel.

Evidence

Evidence 1 - **Rationale**: Both papers claim to stabilize RLVR training through differentiated treatment of tokens, with the candidate demonstrating this capability before the original paper's submission. - **Original**: exgrpo stabilizes training on both stronger and weaker models where on-policy methods fail - **Candidate**: our method applies weaker kl regularization and higher clipping thresholds to reasoning tokens to encourage exploration, while using stronger constraints on knowledge tokens to maintain factual knowledge. experimental results on several mathematical reasoning and code generation benchmarks show that...

Evidence 2 - **Rationale**: The candidate paper demonstrates consistent performance improvements across benchmarks and model sizes, addressing the same training stabilization challenge claimed as novel by the original paper. - **Original**: exgrpo stabilizes rlvr training on the weaker llama-3.1 8b model (grattafiori et al., 2024) and continual learning on the stronger luffy model (yan et al., 2025), where on-policy optimization collapses - **Candidate**: we evaluate our approach on challenging mathematical reasoning and code generation benchmarks. our experiments show significant performance improvements across different tasks. compared to the standard dapo algorithm (yu et al., 2025), our dual-token constraints method achieves notable gains: +6.6 p...

Evidence 3 - **Rationale**: Both papers evaluate on models in the 1.5b-8b parameter range and demonstrate consistent improvements, showing that the approach of stabilizing training across different model sizes was already demonstrated in prior work. - **Original**: experiments on five backbone models (1.5b-8b parameters) show that exgrpo consistently improves reasoning performance on mathematical/general benchmarks, with an average gain of +3.5/7.6 points over on-policy rlvr - **Candidate**: we adoptdeepseek-r1-distill-qwen-1.5b as the base model, which is distilled from deepseek-r1 (deepseek-ai et al., 2025) using qwen2.5-1.5b (yang et al., 2024) as the backbone and fine-tuned on 800k high-quality reasoning data

9. Reasoning-table: Exploring reinforcement learning for table reasoning

URL: View paper

Brief Assessment

Reasoning-Table[43] focuses on applying RL to table reasoning tasks (table QA, fact verification, text-to-SQL), not general mathematical reasoning or stabilizing training across different model sizes for reasoning tasks as in the original paper.

10. Efficient reasoning models: A survey

URL: View paper

Brief Assessment

Efficient Reasoning Survey[50] is a survey paper that reviews existing methods for efficient reasoning. It does not present original experimental results on training stabilization across different model sizes, and therefore cannot refute the novelty of ExGRPO's empirical findings on stabilizing RLVR training.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 2 similarity segment(s) across 2 paper(s).

The following **2 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. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping

Detected in: Contribution: contribution_1

⚠ **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.

2. ExGRPO: Learning to reason from experience

Detected in: Contribution: contribution_1

⚠ **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] ExGRPO: Learning to Reason from Prior Successes View paper
- [1] Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? View paper
- [2] Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs View paper
- [3] Reinforcement Learning with Verifiable Rewards: GRPO's Effective Loss, Dynamics, and Success Amplification View paper
- [4] Chart-RVR: Reinforcement Learning with Verifiable Rewards for Explainable Chart Reasoning View paper
- [5] Demystifying long chain-of-thought reasoning in llms View paper
- [6] Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning View paper
- [7] Generative Verifiers: Reward Modeling as Next-Token Prediction View paper
- [8] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle View paper
- [9] Beyond Verifiable Rewards: Scaling Reinforcement Learning for Language Models to Unverifiable Data View paper
- [10] Diversity-incentivized exploration for versatile reasoning View paper
- [11] TruthRL: Incentivizing truthful LLMs via reinforcement learning View paper
- [12] Trust but verify! a survey on verification design for test-time scaling View paper
- [13] REASONING GYM: Reasoning Environments for Reinforcement Learning with Verifiable Rewards View paper
- [14] Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction View paper
- [15] Writing-Zero: Bridge the Gap Between Non-verifiable Problems and Verifiable Rewards View paper
- [16] Leanabell-prover-v2: Verifier-integrated reasoning for formal theorem proving via reinforcement learning View paper
- [17] Reward Hacking Mitigation using Verifiable Composite Rewards View paper
- [18] RLPR: Extrapolating RLVR to General Domains without Verifiers View paper
- [19] Shielded learning for resilience and performance based on statistical model checking in simulink View paper
- [20] Trust, But Verify: A Self-Verification Approach to Reinforcement Learning with Verifiable Rewards View paper
- [21] UloRL:An Ultra-Long Output Reinforcement Learning Approach for Advancing Large Language Models' Reasoning Abilities View paper
- [22] IFDECORATOR: Wrapping Instruction Following Reinforcement Learning with Verifiable Rewards View paper
- [23] Can one domain help others? a data-centric study on multi-domain reasoning via reinforcement learning View paper
- [24] Breaking the Safety-Capability Tradeoff: Reinforcement Learning with Verifiable Rewards Maintains Safety Guardrails in LLMs View paper
- [25] Language Models that Think, Chat Better View paper
- [26] Learning to Rank Chain-of-Thought: An Energy-Based Approach with Outcome Supervision View paper
- [27] Scaling Reinforcement Learning for Content Moderation with Large Language Models View paper
- [28] Rlsf: Reinforcement learning via symbolic feedback View paper
- [29] Random Policy Valuation is Enough for LLM Reasoning with Verifiable Rewards View paper
- [30] SwS: Self-aware Weakness-driven Problem Synthesis in Reinforcement Learning for LLM Reasoning View paper
- [31] Brittleness and Promise: Knowledge Graph Based Reward Modeling for Diagnostic Reasoning View paper

- [32] Critique-RL: Training Language Models for Critiquing through Two-Stage Reinforcement Learning View paper
- [33] Outcome-based Reinforcement Learning to Predict the Future View paper
- [34] Sharpness-Controlled Group Relative Policy Optimization with Token-Level Probability Shaping View paper
- [35] A Critical Analysis of the Proposed Recursive Logic Subsystem for Self-Learning LLMs in Scientific Discovery View paper
- [36] Position: The Hidden Costs and Measurement Gaps of Reinforcement Learning with Verifiable Rewards View paper
- [37] The Reasoning Boundary Paradox: How Reinforcement Learning Constrains Language Models View paper
- [38] Low-probability Tokens Sustain Exploration in Reinforcement Learning with Verifiable Reward View paper
- [39] Simultaneous Multi-objective Alignment Across Verifiable and Non-verifiable Rewards View paper
- [40] Knowledge-to-Verification: Unlocking Reinforcement Learning with Verifiable Rewards for LLMs in Knowledge-Intensive Domains View paper
- [41] RLSF: Reinforcement Learning from Self-feedback for improved logical reasoning View paper
- [42] Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models View paper
- [43] Reasoning-table: Exploring reinforcement learning for table reasoning View paper
- [44] Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model View paper
- [45] L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning View paper
- [46] Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning View paper
- [47] The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models View paper
- [48] Reason-rft: Reinforcement fine-tuning for visual reasoning View paper
- [49] Training Language Models to Reason Efficiently View paper
- [50] Efficient reasoning models: A survey View paper
- [51] Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr View paper
- [52] Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping View paper
- [53] ExGRPO: Learning to reason from experience View paper
- [54] Staying in the Sweet Spot: Responsive Reasoning Evolution via Capability-Adaptive Hint Scaffolding View paper
- [55] First return, entropy-eliciting explore View paper
- [56] EFRame: Deeper Reasoning via Exploration-Filtering-Replay Reinforcement Learning Framework View paper
- [57] Exploring Multi-Temperature Strategies for Token-and Rollout-Level Control in RLVR View paper
- [58] PEAR: Phase Entropy Aware Reward for Efficient Reasoning View paper
- [59] Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning View paper
- [60] Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning View paper
- [61] Evolving language models without labels: Majority drives selection, novelty promotes variation View paper
- [62] Experience Replay for Continual Learning View paper
- [63] Query-Policy Misalignment in Preference-Based Reinforcement Learning View paper
- [64] Cooperative Traffic Scheduling in Transportation Network: A Knowledge Transfer Method View paper
- [65] A prioritized objective actor-critic method for deep reinforcement learning View paper
- [66] Hybrid attention-oriented experience replay for deep reinforcement learning and its application to a multi-robot cooperative hunting problem View paper
- [67] Experience Replay-based Deep Reinforcement Learning for Dialogue Management Optimisation View paper
- [68] Sample-efficient LLM Optimization with Reset Replay View paper
- [69] Experience Consistency Distillation Continual Reinforcement Learning for Robotic Manipulation Tasks View paper
- [70] Anti-jamming routing for internet of satellites: a reinforcement learning approach View paper
- [71] Relay Hindsight Experience Replay: Continual Reinforcement Learning for Robot Manipulation Tasks with Sparse Rewards View paper