

DARL: Encouraging Diverse Answers for General Reasoning without Verifiers

Chongxuan Huang^{1,*}, Lei Lin^{2,†,‡}, Xiaodong Shi^{1,3,‡}, Wenping Hu², Ruiming Tang²

¹ School of Informatics, Xiamen University

² Kuaishou Technology, Beijing, China

³ Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism

huangchongxuan@stu.xmu.edu.cn, mandel@xmu.edu.cn

{linlei, huwenping, tangruiming}@kuaishou.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has demonstrated promising gains in enhancing the reasoning capabilities of large language models. However, its dependence on domain-specific verifiers significantly restricts its applicability to open and general domains. Recent efforts such as RLPR have extended RLVR to general domains, enabling training on broader datasets and achieving improvements over RLVR. However, a notable limitation of these methods is their tendency to overfit to reference answers, which constrains the model’s ability to generate diverse outputs. This limitation is particularly pronounced in open-ended tasks such as writing, where multiple plausible answers exist. To address this, we propose *DARL*, a simple yet effective reinforcement learning framework that encourages the generation of diverse answers within a controlled deviation range from the reference while preserving alignment with it. Our framework is fully compatible with existing general reinforcement learning methods and can be seamlessly integrated without additional verifiers. Extensive experiments on thirteen benchmarks demonstrate overall improvements in reasoning performance. Notably, DARL surpasses RLPR, achieving average gains of 1.3 points on six reasoning benchmarks and 9.5 points on seven general benchmarks, highlighting its effectiveness in improving both reasoning accuracy and output diversity.

1 Introduction

Recent advances in Large Reasoning Models (LRMs), such as DeepSeek-R1 (Guo et al., 2025) and OpenAI’s O1 (Jaech et al., 2024), have demonstrated that reasoning capabilities are crucial for improving model performance across various tasks

* Work done during an internship at Kuaishou Technology.

† Project Lead.

‡ Corresponding author.

(Hu et al., 2025; Liu et al., 2025b; Luo et al., 2025). To this end, Large-Scale Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as an effective training paradigm for LRMs (Xie et al., 2025). Specifically, RLVR assigns rewards solely based on final-answer correctness, encouraging the emergence of intermediate reasoning without explicit supervision. By leveraging clear and verifiable reward signals, RLVR guides models to discover reasoning trajectories that reliably lead to correct solutions.

However, RLVR exhibits notable limitations when compared to LLMs pretrained on general data, which acquire broad foundational abilities (Zhang et al., 2025a). While RLVR performs well in domains with well-defined, objectively correct answers, such as mathematics (Hu et al., 2025; Liu et al., 2025b; Zeng et al., 2025) and programming (Luo et al., 2025), it struggles in areas requiring subjective judgment, including creative writing or open-ended question answering. This limitation arises because RLVR relies on manually designed verifiers, which are difficult to construct for complex, subjective tasks. The need for domain-specific experts to create these verification systems also results in high costs and resource requirements (Yu et al., 2025b). Moreover, the binary reward structure inherent to RLVR, which provides sparse feedback, limits its ability to handle more nuanced or general tasks (Zhang et al., 2025a).

To address these limitations, recent research has explored several directions for extending RLVR to general reasoning tasks. A common strategy leverages generative models to produce fine-grained feedback (Mahan et al., 2024; Li et al., 2023; Zheng et al., 2023), mitigating the limitations of rule-based verifiers, yet it demands large-scale annotations and complex verifier models, which increase computational cost and risk reward hacking, as observed in DeepSeek-V3 and DeepSeek-R1 (Guo et al., 2025). To reduce dependence on labeled

data, unsupervised reward modeling extracts signals directly from model-internal consistency, entropy, or self-generated knowledge (RLIF (Zhao et al., 2025), EM-RL (Agarwal et al., 2025), EMPO (Zhang et al., 2025c), CoVo (Zhang et al., 2025b), RENT (Prabhudesai et al., 2025), TTRL (Zuo et al., 2025), SeRL (Fang et al., 2025)). Although scalable, such closed-loop training often suffers from instability and excessive reliance on pretrained priors. More recent hybrid methods, such as NOVER (Liu et al., 2025a), VeriFree (Zhou et al., 2025), RLPR (Yu et al., 2025b), JEPO (Tang et al., 2025), and LaTRO (Chen et al., 2024), combine self-generated signals with reference answers to build more accurate reward functions. These designs yield more reliable rewards and have achieved leading performance across several reasoning benchmarks. However, generating diverse yet valid outputs is critical for capturing underlying semantic and reasoning structures and for improving generalization, and existing methods tend to overfit to the reference answers, thereby restricting the model’s capacity to produce such diversity.

Motivated by these observations, we argue that an effective general reinforcement learning framework for reasoning should avoid overfitting to specific reference answers or relying on explicit verifiers. Instead, it should encourage controlled exploration beyond the reference, allowing the model to generate diverse yet reference-consistent answers. To this end, we propose *DARL*, a verifier-free, simple yet effective reinforcement learning framework that promotes answer diversity within a bounded deviation range from the reference answer, without sacrificing reference consistency. Our main contributions are summarized as follows:

- First, we propose a general verifier-free reinforcement learning framework that encourages models to generate diverse answers while remaining consistent with the reference. Unlike existing reference-based methods that tend to overfit exact ground-truth expressions, our approach explicitly promotes controlled exploration over alternative answer formulations without compromising alignment with the reference.
- Second, we show that DARL significantly improves answer diversity. At the policy level, our method maintains higher policy entropy during training, indicating enhanced and more stable exploration. At the output level,

likelihood-based analysis shows that DARL assigns higher probability mass to semantically equivalent but lexically diverse answers compared to RLPR. These results demonstrate that our approach promotes controlled answer diversity, which is especially beneficial for open-ended tasks such as writing.

- Third, DARL consistently improves performance across diverse reasoning and general-domain benchmarks. It achieves notable gains on benchmarks with multiple valid reasoning paths (e.g., GPQA and AIME24), while maintaining comparable performance on more constrained tasks such as MMLU-Pro. Moreover, DARL delivers substantial improvements in general-domain settings, with average gains of 11.57 points on logical reasoning, 12.44 points on code generation, and 3.35 points on writing tasks, demonstrating strong generalization in open-ended scenarios.

2 Related Work

Language Model Reasoning. Reinforcement learning has repeatedly shown that narrow and well-defined reward signals can effectively guide agents to achieve superhuman performance on complex tasks (Zhang et al., 2025a). As the primary training signal, rewards determine the optimization trajectory of the model and shape its behavior. In the context of large-scale language model training, rule-based verifiable rewards have been widely adopted to provide scalable and reliable supervision, particularly in domains such as mathematics (Hu et al., 2025; Liu et al., 2025b; Zeng et al., 2025) and code generation (Luo et al., 2025). This approach, formalized as RLVR, replaces learned reward models with programmatic verifiers that emit binary feedback in domains where correctness can be objectively assessed. For example, DeepSeek-V3 (Liu et al., 2024) integrated task-specific rule-based reward systems, while DeepSeek-R1 extended this framework by incorporating both accuracy-based and format-based reward mechanisms. Subsequent studies, including Logic-RL (Xie et al., 2025), DAPO (Yu et al., 2025a), SimpleRL (Zeng et al., 2025), OpenReasoner (Hu et al., 2025), Dr.GRPO (Liu et al., 2025b), and VAPO (Yue et al., 2025), have investigated best practices for balancing exploration and exploitation, the emergence of “aha” moments, and the influence of task difficulty variations on learning dynamics.

Verifier-Free Reasoning. Generative reward modeling and verifier-free approaches have recently emerged as complementary directions for scaling reinforcement learning for LLM reasoning. Early work demonstrated that learned, generation-based verifiers can provide finer-grained, model-aware feedback compared to brittle rule-based checks (e.g., GenRM (Mahan et al., 2024)). However, training standalone reward models typically requires large, complex annotation corpora, increasing system complexity and computational cost (Su et al., 2025; Lu, 2025).

To reduce reliance on expensive labels, a growing body of research investigates unsupervised or self-supervised reward signals that derive intrinsic feedback from model consistency, confidence, or self-generated knowledge (Zhang et al., 2025a). Representative techniques include entropy-minimization and mutual-information objectives, as well as self-scoring and voting schemes. These methods scale without human annotation and can enhance apparent performance, but they are prone to reward-hacking, overconfidence, entropy collapse, and excessive reliance on pretraining priors, potentially limiting true reasoning generalization (Zhang et al., 2025c; Cui et al., 2025b; Fränken et al., 2024; Zuo et al., 2025).

A middle-ground approach leverages ground-truth supervision to enhance self-generated signals, producing more reliable rewards (e.g., JEPO (Tang et al., 2025), LaTRO (Chen et al., 2024), NOVER (Liu et al., 2025a), VeriFree (Zhou et al., 2025), RLPR (Yu et al., 2025b)). While these methods guide models toward reference-aligned answers, they often induce rapid entropy collapse and overfit to specific solution patterns. Our method extends this paradigm by promoting answer diversity: it encourages the model to generate answers that deviate moderately from the ground truth within a controlled range, thereby balancing reliability and diversity in reward optimization.

3 Background

Rule-based Reward. Reinforcement Learning from Verifiable Rewards is a general post-training paradigm that leverages verifiable indicators—such as answer correctness or code execution success—as reward signals. This approach enables models to exhibit capabilities such as planning, reflection, and self-correction in long-chain reasoning tasks (Liu et al., 2025b; Xie et al., 2025).

Given a prompt that requires the model to generate intermediate reasoning steps followed by a final answer (y), the training objective is to optimize the policy model π_θ via reinforcement learning, using a verifiable reward R_{rule} . The reward is computed individually by a rule-based verifier $v : \mathcal{A} \rightarrow \{0, 1\}$, which evaluates the correctness of the final answer:

$$R_{\text{rule}} = v(y) \quad (1)$$

While the outcome reward is inherently sparse, it offers accurate supervision, guiding the model to autonomously produce intermediate reasoning tokens that facilitate the generation of a more reliable final answer.

Verifier-Free Reward. While RLVR has shown strong performance on tasks with explicit verifiers, such as mathematical problem solving and code generation, its applicability to open-domain reasoning is limited, as human-designed verifiers are infeasible for diverse and context-dependent natural language. To address this, recent verifier-free approaches propose generalizable reward functions, leveraging the assumption that high-capacity models inherently produce consistent and self-evaluable outputs, thereby enabling reward estimation without external supervision and enhancing scalability.

Formally, given a reasoning prompt x , the policy model generates a composite output $o = z \oplus y$, where z represents the intermediate reasoning process and y the final answer. The reward r is computed from the model’s intrinsic confidence in the reference answer y^* conditioned on the reasoning chain z , yielding a dense and differentiable training signal:

$$\nabla_\theta J_{\text{VerifierFree}}(\theta) = \mathbb{E}_{o \sim \pi_\theta(\cdot|x)} [r \cdot \nabla_\theta \log \pi_\theta(o|x)] \quad (2)$$

Here, r is derived from model-internal signals rather than external verifiers. For example, RLPR (Yu et al., 2025b) computes r based on the likelihood of generating the ground-truth answer given the intermediate reasoning chain, effectively aligning reward estimation with the model’s probabilistic reasoning process. This verifier-free paradigm broadens the applicability of RLVR to general reasoning tasks while mitigating the reliance on domain-specific verifiers and human annotations.

4 Method

To promote the generation of diverse answers, we introduce a dynamic, diversity-aware reward that

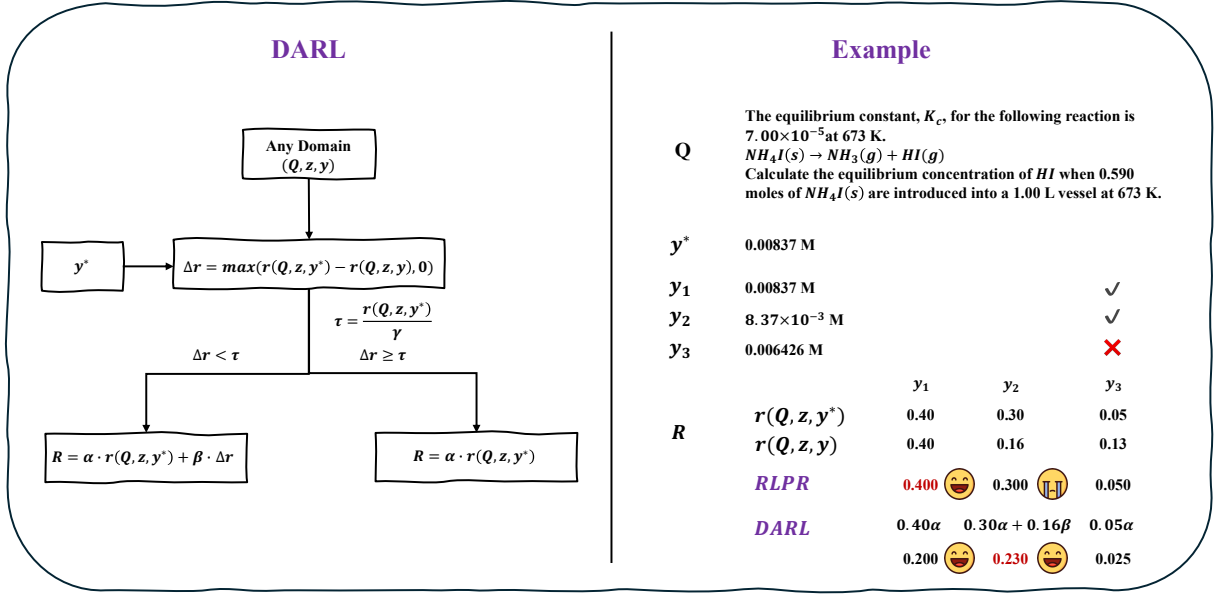


Figure 1: Overview of DARL. Q : input question, z : generated reasoning content before final answer, y : generated final answer, y^* : reference answer.

adaptively adjusts the reward threshold for answer diversity according to the model’s confidence.

Given an input x , the model generates a reasoning trace z followed by an answer y , while y^* denotes the ground-truth answer. The diversity reward is defined as:

$$\Delta r = \max(r(y^*, z, x) - r(y, z, x), 0) \quad (3)$$

where $r(\cdot)$ is a reward function derived from model-internal signals. Here, Δr quantifies the deviation between the model-generated answer and the reference answer. When both $r(y^*, z, x)$ and $r(y, z, x)$ are relatively high and their difference falls within a predefined range, it indicates that, given the current reasoning trace z , the model has sufficient capability to generate y^* . Consequently, the answer y produced under the same reasoning trace is also likely to be correct. In other words, when Δr falls within the acceptable range, y and y^* are highly likely to be consistent answers. Under this condition, the model is encouraged to explore alternative expressions that differ from the reference, and deviations within this range are rewarded, thereby promoting answer diversity while maintaining consistency with the reference answer.

The overall reward combines a reference-aligned term with a bounded answer diversity component:

$$\bar{r} = \alpha \cdot r(y^*, z, x) + \beta \cdot \Delta r \cdot \mathbf{1}[\Delta r \leq \tau] \quad (4)$$

where α and β are trade-off coefficients controlling the balance between exploitation and answer diver-

sity (with $\alpha + \beta = 1$), and τ denotes a fixed diversity threshold. In our implementation, the reward $r(\cdot)$ is computed as the average token probability, following the approach used in RLPR.

$$r(y, z, x) = \frac{1}{|y|} \sum_{i=1}^{|y|} p_{\theta}(y_i | x, z, y_{<i}) \quad (5)$$

To further enhance adaptability, we introduce a dynamic thresholding strategy that scales the diversity boundary according to the model’s confidence in $r(y^*, z, x)$:

$$\bar{\tau} = \alpha \cdot r(y^*, z, x) + \beta \cdot \Delta r \cdot \mathbf{1}\left[\Delta r \leq \frac{r(y^*, z, x)}{\gamma}\right] \quad (6)$$

When the model assigns high confidence to the reference-aligned reward, it indicates sufficient capability to generate answers consistent with the reference. In this regime, we expand the allowable diversity margin to encourage exploration of alternative answers that differ in expression while remaining consistent with the reference. Conversely, when the model’s confidence is low, the diversity range is restricted, biasing learning toward reproducing the reference answer. This prevents premature diversification and stabilizes training when the model’s ability to generate correct answers is limited. Overall, this dynamic answer diversity reward adaptively balances reference alignment and controlled exploration, enabling the generation of multiple reference-consistent solutions while pre-

serving correctness, thereby enhancing both generalization and robustness.

5 Experimental Setup

Dataset. We utilize the dataset collection released with RLPR (Yu et al., 2025b), which was originally curated from the WebInstruct corpus by Ma et al. (2025). The dataset consists of 77.7K training instances of high-quality, non-mathematical reasoning prompts spanning diverse domains, including commonsense reasoning, factual consistency, multi-hop inference, and abstract problem solving.

Models. Our experiments primarily employ Llama3.1 (Dubey et al., 2024) and Qwen2.5 (Yang et al., 2024) as the backbone policy models. Both models have demonstrated strong zero-shot and instruction-following capabilities across a wide range of reasoning benchmarks. Unless otherwise specified, we report results for Llama3.1-8B-Instruct and Qwen2.5-7B-Base variants.

Baselines. For fair and consistent comparison, we adopt the same set of baseline methods as RLPR (Yu et al., 2025b), following its official training configurations. Specifically, our comparisons cover (1) base and instruct models (e.g., Qwen2.5 and Llama3.1), (2) reinforcement learning approaches with rule-based, implicit, or verifier-based rewards (e.g., PRIME (Cui et al., 2025a), SimpleRL-Zoo (Zeng et al., 2025), Oat-Zero (Liu et al., 2025c), and General Reasoner (Ma et al., 2025)), and (3) concurrent verifier-free reinforcement learning methods such as TTRL (Zuo et al., 2025), VeriFree (Zhou et al., 2025) and RLPR (Yu et al., 2025b). All implementations and hyperparameters are consistent with those reported in RLPR, ensuring a fair comparison.

Evaluation. We evaluate all models on a diverse set of benchmarks covering knowledge and factual reasoning, mathematical and symbolic reasoning, logical inference, code generation, and generative writing. The evaluation suite comprises MMLU-Pro (Wang et al., 2024), GPQA (Rein et al., 2023), TheoremQA (Chen et al., 2023), MATH-500 (Cobbe et al., 2021), Minerva (Lewkowycz et al., 2022), AIME24, AutoLogic-cn, AutoLogic-en (Zhu et al., 2025), ZebraLogic (Lin et al., 2025), LiveCodeBench (Jain et al., 2024), HumanEval (Chen, 2021), HumanEval+ (Liu et al., 2023), and WritingBench (Wu et al., 2025). For

WritingBench, model performance is evaluated using the WritingBench-Critic-Model-Qwen-7B, whereas all other benchmarks are assessed with Qwen2.5-72B-Instruct. Together, these benchmarks test a wide spectrum of capabilities, including knowledge-driven reasoning, theorem-based problem solving, logical analysis, code synthesis, and open-ended writing, providing a comprehensive assessment of models’ reasoning abilities and generalization performance.

Answer Diversity Evaluation. Existing training paradigms that maximize the likelihood of ground-truth answers tend to favor reproducing exact reference expressions, which can limit the model’s propensity to generate alternative but consistent answers. To assess the diversity of model-generated answers, we analyze how probability mass is distributed across answer variants that are consistent with the reference.

Specifically, we transform the reference answers in the WebInstruct dataset into lexically distinct but consistent variants using DeepSeek-V3 (Liu et al., 2024). To ensure the reliability of these generated variants, we filter out any duplicate answers, and then manually evaluate a random sample of 2,000 rewritten answers. All sampled variants were confirmed to be valid and consistent with the original reference answers, demonstrating the effectiveness of the DeepSeek-V3 rewriting process. At each training checkpoint, we compute the average log-likelihood assigned by the model to (i) the original ground-truth answers and (ii) the validated rewritten alternatives.

6 Results and Analysis

6.1 Overall Performance

Following RLPR’s benchmark configuration, we evaluate DARTL on a diverse set of reasoning and general-domain tasks. As shown in Table 1, our method consistently improves performance, with gains of 1.79, 1.22, and 7.03 points on GPQA, Minerva, and AIME 24, respectively, while performing comparably on MMLU-Pro and TheoremQA.

The improvements are most pronounced on tasks allowing multiple valid reasoning paths or alternative answers consistent with the reference, such as GPQA and AIME 24. By encouraging exploration of reference-consistent alternatives, DARTL captures underlying reasoning structures without overfitting to a single canonical output.

Model	Base	Verifier	MMLU-Pro	GPQA	TheoremQA	MATH-500	Minerva	AIME 24	General	All
			Avg@2	Avg@4	Avg@2	Avg@2	Avg@2	Avg@16		
Llama Models										
Llama3.1-8B-Inst	Base	–	46.4	31.6	31.3	50.1	32.7	4.2	36.4	32.7
RLVR	Inst	Rule	49.3	36.0	32.0	51.9	35.2	4.6	39.1	34.8
RLPR	Inst	✗	53.6	36.5	36.5	54.1	39.0	8.8	42.2	38.1
DARL	Inst	✗	<u>52.5</u>	36.9	39.4	<u>52.8</u>	<u>37.9</u>	10.0	42.9	38.2
Qwen Models										
Qwen2.5-7B	–	–	45.3	32.4	41.4	63.0	37.6	6.5	39.7	37.7
Qwen2.5-7B-Inst	Base	–	54.5	34.2	47.3	75.4	49.4	9.4	45.3	45.0
Oat-Zero	Math	Rule	45.8	<u>38.8</u>	53.3	<u>80.8</u>	52.1	29.8	46.0	<u>50.1</u>
PRIME	Math	Rule	39.5	32.1	47.7	<u>76.4</u>	45.5	20.4	39.8	43.6
SimpleRL-Zoo	Math	Rule	46.9	38.4	51.1	77.1	51.0	<u>26.5</u>	45.5	48.5
TTRL	Base	Rule	51.1	34.1	48.8	82.1	52.8	15.8	44.7	47.5
SimpleRL-Zoo	Base	Rule	54.1	36.2	49.5	76.3	49.2	14.8	46.6	46.7
RLVR	Base	Rule	55.1	36.2	52.2	76.5	54.9	17.7	47.8	48.8
General Reasoner	Base	Model	55.4	37.4	52.1	77.0	51.7	16.0	48.3	48.3
VeriFree	Base	✗	53.8	36.7	47.6	73.5	49.0	12.5	46.0	45.5
RLPR	Base	✗	56.0	37.6	55.4	78.0	<u>56.5</u>	16.3	49.7	50.0
DARL	Base	✗	<u>55.6</u>	39.4	<u>55.2</u>	76.6	57.7	23.3	50.1	51.3

Table 1: Overall performance on six benchmarks. General: Average of MMLU-Pro, GPQA, TheoremQA. The highest score for each task are highlighted in **bold**, and the second highest are underlined.

A more fine-grained analysis reveals that the effectiveness of diversity is closely tied to the entropy of the answer space. On closed-form discrete tasks such as MMLU-Pro (A/B/C/D), where the answer space is low-entropy and highly constrained, encouraging diversity brings limited benefit, aligning with the intuition that exploration cannot meaningfully expand valid solution coverage.

In contrast, on tasks with high reasoning search complexity, even when the final answer is unique, diversity proves beneficial. For example, on AIME 24, performance improves substantially, suggesting that diverse reasoning trajectories increase the probability of discovering correct solutions. This indicates that diversity primarily operates in the reasoning process space rather than the answer space itself.

6.2 General Performance

To evaluate the generality of our approach, we conduct experiments on a diverse set of benchmarks spanning logical reasoning, code generation, and writing, covering varied input structures, evaluation criteria, and degrees of answer openness. Our method consistently improves performance across all domains, achieving average gains of 11.57 points on AutoLogic, 12.44 points on code-related benchmarks, and 3.35 points on writing

tasks over RLPR.

Notably, the improvements correlate strongly with the openness of the solution space. On multi-solution functional tasks such as HumanEval and HumanEval+, where multiple implementations can satisfy the same specification, diversity leads to substantial gains, consistent with the high entropy of the solution space—encouraging diverse outputs directly increases the likelihood of producing valid solutions.

Similarly, on writing tasks—where diversity is inherently tied to output quality—our method promotes varied yet reference-consistent generations, mitigating overfitting to the reference while preserving correctness.

Overall, these results suggest that DARL adapts effectively to different task regimes: it remains stable in low-entropy settings while delivering significant gains in high-entropy or complex reasoning scenarios. This highlights its ability to improve robustness and generalization across diverse tasks without relying on domain-specific verifiers.

6.3 Training Dynamics of Output Diversity

To further analyze variations in model output diversity, we evaluate it from two complementary perspectives. First, we examine the evolution of policy entropy during training, which serves as an

Model	Base	Verifier	Autologic_cn	Autologic_en	Zebralogic	Livecodebench	Humaneval	Humaneval+	WritingBench	All
Llama Models										
Llama3.1-8B-Inst	Base	–	22.0	25.1	<u>8.9</u>	<u>10.0</u>	61.6	56.7	<u>66.3</u>	35.8
RLVR	Inst	Rule	2.3	20.2	10.1	7.2	38.4	44.5	20.7	20.5
RLPR	Inst	✗	27.8	34.9	8.3	9.0	63.4	62.2	53.5	37.0
DARL	Inst	✗	39.2	42.0	14.2	11.8	64.0	<u>61.6</u>	71.6	43.5
Qwen Models										
Qwen2.5-7B	–	–	5.9	8.4	0.8	2.9	37.8	32.3	38.3	18.1
Qwen2.5-7B-Inst	Base	–	<u>38.3</u>	<u>40.6</u>	10.1	12.9	79.3	<u>70.1</u>	80.7	47.4
Oat-Zero	Math	Rule	13.3	19.1	0.2	6.1	61.0	53.0	53.9	29.5
PRIME	Math	Rule	2.9	1.2	0.0	0.7	3.0	3.7	33.8	6.5
SimpleRL-Zoo	Math	Rule	7.1	14.4	0.7	6.1	61.0	53.0	39.9	26.0
TTRL	Base	Rule	11.1	12.8	2.1	3.6	50.6	44.5	52.0	25.3
SimpleRL-Zoo	Base	Rule	29.0	31.2	6.3	8.2	67.1	62.2	60.9	37.8
RLVR	Base	Rule	25.1	22.4	7.7	9.0	68.9	64.0	54.5	35.9
General Reasoner	Base	Model	29.2	29.0	8.9	10.0	61.6	56.7	66.3	37.4
VeriFree	Base	✗	35.7	39.5	7.6	15.1	72.0	65.9	69.5	43.6
RLPR	Base	✗	29.1	31.9	6.4	8.2	56.7	59.1	67.5	37.0
DARL	Base	✗	40.3	43.9	<u>9.2</u>	<u>13.3</u>	<u>75.0</u>	73.2	<u>70.9</u>	<u>46.5</u>

Table 2: Overall performance on seven general benchmarks. The highest score for each task are highlighted in **bold**, and the second highest are underlined.

Model	Autologic_cn	Autologic_en	zebralogic	Livecodebench	Humaneval	Humaneval+	WritingBench	All
VeriFree	35.7	39.5	7.6	15.1	72.0	65.9	69.5	43.6
DARL	39.3	40.7	9.0	14.3	72.0	68.3	70.0	44.8

Table 3: Overall performance on seven general reasoning benchmarks. The highest score for each task are highlighted in **bold**.

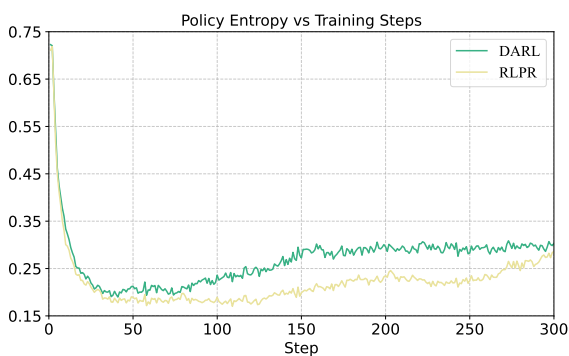


Figure 2: Policy entropy over training steps for DARL and RLPR. Our method consistently maintains higher entropy, indicating increased diversity in generated responses.

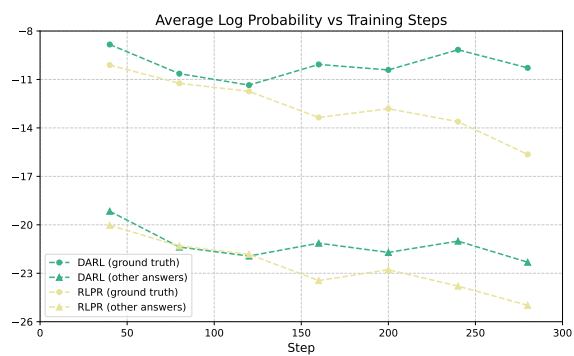


Figure 3: Average log-probability of ground-truth answers and alternative semantically equivalent answers during training. Compared to RLPR, DARL assigns higher likelihood to diverse yet semantically faithful answers.

indicator of the degree of exploration and diversity in the policy over the action space. Second, we assess output-level diversity by computing the average log-likelihood assigned to a set of diverse answer candidates, which captures the model’s tendency to assign probability mass to alternative yet plausible responses.

Our method is associated with higher policy entropy, suggesting improved exploration behavior. Figure 2 illustrates the evolution of policy entropy during training for both RLPR and our proposed method. In the early stage of training, the model’s capability is relatively limited, resulting in a low upper bound on the reward for exploring diverse answers. Consequently, learning is primarily

driven by the RLPR reward signal, and the policy entropy of our method closely follows that of RLPR, exhibiting a rapid decline.

As training progresses into the intermediate stage, the model’s capacity gradually improves, leading to an increased upper bound on the reward associated with exploring diverse answer formulations. During this phase, our method maintains a higher policy entropy than RLPR, indicating a greater degree of exploration and output diversity.

In the later stages of training, the policy entropy of our method stabilizes, whereas that of RLPR continues to increase gradually. This divergence indicates that, although our approach encourages answer diversity, it avoids excessive randomness or output collapse. Instead, it facilitates controlled exploration within a bounded region consistent with the ground-truth answers, suggesting that the proposed diversity exploration range effectively guides the model toward diverse yet reference-consistent outputs.

Our method achieves higher answer diversity and yields improved performance on open-ended tasks. To analyze differences among the diverse answers generated by the model, we quantify output diversity by computing the average log-likelihood assigned to multiple diverse answer candidates. Specifically, for both RLPR and our proposed method, we perform rollouts of eight responses at each training step. For each response, we replace the generated answer with a set of semantically similar but lexically distinct variants and compute the corresponding log-likelihoods under the model. The average log-likelihood across these variants is then used as our answer diversity score.

The results reveal markedly different trends between RLPR and our method. As training progresses, our approach assigns higher probability mass not only to the ground-truth answers but also to alternative, semantically consistent answers, compared to RLPR. This behavior indicates that the proposed method effectively encourages the model to explore and generate diverse answer formulations while remaining aligned with the ground-truth supervision.

6.4 Transferability to Other Verifier-Free RL Frameworks

Beyond RLPR, our method is a general verifier-free paradigm that can be integrated into other reinforcement learning frameworks based on model-internal

Method	Autologic	Humaneval+	WritingBench	Avg.
Qwen2.5-7B				
<i>Base Model</i>	42.35 %	67.68 %	67.54 %	59.19 %
<i>w/ SAD</i>	42.98 %	71.95 %	68.16 %	61.03 %
<i>w/ DAD</i>	43.87 %	73.17 %	70.89 %	62.64 %

Table 4: Ablation study on three representative benchmarks. Base Model is trained without any answer diversity reward. SAD denotes the static answer diversity reward, and DAD denotes the dynamic answer diversity reward. Values in **bold** indicate the best-performing variant.

signals. To evaluate its transferability, we incorporate it into VeriFree and test on a diverse set of general-domain benchmarks.

The results show that adding our approach to VeriFree consistently improves performance, with an average gain of 1.2 points. This indicates that controlled exploration of consistent yet diverse answers serves as a complementary inductive bias, benefiting different verifier-free RL frameworks. By relying solely on model-internal signals and reference supervision, our method preserves the advantages of verifier-free RL while reducing overfitting to exact reference expressions, providing a practical way to enhance answer diversity and generalization across tasks.

6.5 Ablation Study

Table 4 presents an ablation study on three representative benchmarks using Qwen2.5-7B. The Base Model, trained without any answer diversity reward, serves as the reference. Introducing a static answer diversity reward (SAD) consistently improves performance over the Base Model, increasing scores on AutoLogic_en to 42.98, on HumanEval+ to 71.95, and on WritingBench to 68.16, with an overall average improvement of 1.84 points. These results indicate that incorporating answer diversity signals benefits model performance.

Replacing SAD with a dynamic answer diversity reward (DAD) further enhances results across all benchmarks, achieving the highest overall average of 62.64, which corresponds to an improvement of 3.45 points over the Base Model and 1.61 points over SAD. The gains are particularly notable on HumanEval+ with 5.49 points and on WritingBench with 3.35 points, while AutoLogic_en also increases by 1.52 points. These findings demonstrate that dynamically adjusting the diversity reward more effectively balances the exploration of diverse answers while preserving consistency with

the reference answers, yielding robust improvements across heterogeneous benchmarks. Bold numbers indicate the best-performing variant.

7 Conclusion

Our findings emphasize the importance of explicitly encouraging answer diversity in verifier-free reinforcement learning. By introducing a controlled diversity objective under reference supervision, DARL incentivizes models to explore multiple valid outputs rather than overfitting to a single reference answer. This diversity-aware training leads to more flexible and robust behavior, improving generalization across both reasoning and general-domain tasks. We hope this work motivates further research on incorporating explicit diversity-aware objectives into reinforcement learning for language models, particularly in open-ended and multi-step reasoning settings.

Limitations

A primary limitation of DARL is its reliance on direct supervision from ground-truth answers. This dependence restricts its applicability in scenarios where only queries are available without reference answers. In such cases, rule-based verifiers or alternative approaches that do not rely on reference responses may be more suitable. Another limitation lies in its varying effectiveness across tasks with different solution space characteristics. Specifically, DARL tends to provide limited gains on tasks with low-entropy or highly constrained answer spaces, where the set of valid outputs is small and well-defined. In contrast, it is more effective on open-ended or high-entropy tasks, where multiple valid solutions or reasoning paths exist, allowing diversity to better facilitate exploration and performance improvements.

Acknowledgement

We sincerely thank the anonymous reviewers for their insightful and constructive feedback, which has helped improve the quality of this work. We also thank our collaborators for their valuable discussions, support, and contributions throughout the project.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness

of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, Phil Mui, Silvio Savarese, Caiming Xiong, and 1 others. 2024. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. *arXiv preprint arXiv:2411.04282*.

Mark Chen. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. [Theoremqa: A theorem-driven question answering dataset](#). *Preprint*, arXiv:2305.12524.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, and 1 others. 2025a. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025b. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. 2025. Serl: Self-play reinforcement learning for large language models with limited data. *arXiv preprint arXiv:2505.20347*.

Jan-Philipp Fränken, Eric Zelikman, Rafael Rafailov, Kanishk Gandhi, Tobias Gerstenberg, and Noah Goodman. 2024. Self-supervised alignment with mutual information: Learning to follow principles without preference labels. *Advances in Neural Information Processing Systems*, 37:61328–61371.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

- Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Live-codebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving quantitative reasoning problems with language models](#). *Preprint*, arXiv:2206.14858.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.
- Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. ZebraLogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572.
- Wei Liu, Siya Qi, Xinyu Wang, Chen Qian, Yali Du, and Yulan He. 2025a. Nover: Incentive training for language models via verifier-free reinforcement learning. *arXiv preprint arXiv:2505.16022*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. [Understanding r1-zero-like training: A critical perspective](#). *Preprint*, arXiv:2503.20783.
- Xun Lu. 2025. Writing-zero: Bridge the gap between non-verifiable problems and verifiable rewards. *arXiv preprint arXiv:2506.00103*.
- Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, and 1 others. 2025. Deepcoder: A fully open-source 14b coder at o3-mini level. *Notion Blog*.
- Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zhenjun Ma, and Wenhui Chen. 2025. General-reasoner: Advancing llm reasoning across all domains. *arXiv preprint arXiv:2505.14652*.
- Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.
- Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. [Gpqa: A graduate-level google-proof q&a benchmark](#). *Preprint*, arXiv:2311.12022.
- Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*.
- Yunhao Tang, Sid Wang, Lovish Madaan, and Rémi Munos. 2025. Beyond verifiable rewards: Scaling reinforcement learning for language models to unverifiable data. *arXiv preprint arXiv:2503.19618*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *Preprint*, arXiv:2406.01574.
- Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and 1 others. 2025. Writing-bench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.
- Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan

- Li, Dayiheng Liu, Fei Huang, Guanting Dong, Hao-ran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. [Qwen2.5 technical report](#). *ArXiv*, abs/2412.15115.
- Qiyong Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, and 1 others. 2025a. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, and 1 others. 2025b. Rlpr: Extrapolating rlvr to general domains without verifiers. *arXiv preprint arXiv:2506.18254*.
- Yu Yue, Yufeng Yuan, Qiyong Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, and 1 others. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*.
- Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Ke-qing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.
- Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, and 1 others. 2025a. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*.
- Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. 2025b. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.08745*.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025c. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*.
- Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. 2025. Reinforcing general reasoning without verifiers. *arXiv preprint arXiv:2505.21493*.
- Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. 2025. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. *arXiv preprint arXiv:2502.16906*.
- Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, and 1 others. 2025. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.

Method	Autologic	humaneval+	WritingBench	Avg.
DARL-7B (Default: $\beta = 0.01$ and $\gamma = 10$)				
$\beta = 0.00$	42.35 %	67.68 %	67.54 %	59.19 %
$\beta = 0.01$	43.87 %	73.17 %	70.89 %	62.64 %
$\beta = 0.05$	43.24 %	67.68 %	69.69 %	60.20 %
$\beta = 0.10$	46.60 %	68.90 %	68.29 %	61.27 %
$\gamma = 12$	43.30 %	71.95 %	69.60 %	61.62 %
$\gamma = 10$	43.87 %	73.17 %	70.89 %	62.64 %
$\gamma = 8$	45.65 %	73.17 %	71.26 %	63.36 %

Table 5: Parameter sensitivity analysis. Cell shading reflects the relative performance improvement. **Bold** values denote the best-performing configuration.

Training Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think>` `</think>` and `<answer>` `</answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>` `<answer>` answer here `</answer>`.

Table 6: Prompt used for training.

A Training Details

We train our models using the GRPO algorithm, a variant of PPO, without KL control. The training batch size is 768, with an actor mini-batch size of 192, and each step employs 8 rollouts to generate diverse reasoning trajectories. Policy updates are regulated by standard clipping coefficients (0.2 lower bound, 0.27 upper bound) to maintain training stability. Experiments are conducted on 32 GPUs with 80 GB memory each and require approximately 45 hours. Notably, our training setup closely follows that of RLPR, ensuring fair and consistent comparisons.

B Parameter sensitivity analysis

Table 5 presents a sensitivity analysis of the key hyperparameters that regulate answer-diversity control in our framework, namely the diversity reward weight β and the exploration-bound parameter γ . The objective of this analysis is to quantify how these design choices affect performance across diverse evaluation benchmarks (AutoLogic, HumanEval+, and WritingBench), and to validate the

default configuration adopted in our main experiments.

The parameter β determines the relative contribution of the answer-diversity reward in the overall optimization objective. Increasing β amplifies the contribution of the answer-diversity reward in the optimization objective; however, excessively large values may shift the training focus toward promoting diversity rather than improving the model’s ability to generate reference answers. As shown in Table 5, enabling the diversity reward ($\beta > 0$) consistently improves the average performance over the $\beta = 0$ baseline, indicating that explicitly encouraging controlled diversity is beneficial. However, the performance gains are not monotonic with respect to β : while $\beta = 0.01$ achieves the best overall results under the default exploration bound, larger values (e.g., $\beta = 0.05$ and $\beta = 0.10$) do not yield further improvements in the average score. This trend suggests diminishing returns from stronger diversity regularization and potential overemphasis on diversity at the expense of answer quality.

The parameter γ controls the admissible exploration region for answer diversity by specifying an effective upper bound on allowable deviation. Under our formulation, a larger γ enforces a stricter constraint by reducing the exploration upper bound, thereby limiting the extent of diversity the model can explore. Our results indicate that γ plays a crucial role in balancing answer diversity and reference consistency: overly restrictive constraints weaken the effect of the diversity reward, whereas excessively permissive exploration can destabilize training and degrade solution quality. Notably, a moderately relaxed constraint (e.g., $\gamma = 8$) yields higher average performance than the default setting of $\gamma = 10$, suggesting that a slightly broader explo-

Prompt for Rewriting Ground-Truth Answers with DeepSeek-V3

Prompt:

Rephrase the following answer into different wordings while keeping its meaning strictly the same.

- Do NOT add new information.
- Do NOT remove information.
- Keep the logical order consistent.
- Each version must be semantically equivalent but lexically different.
- Generate exactly 3 rewritten versions.
- Output ONLY in the following JSON format:

```
{ "answer1" : "...", "answer2" : "...", "answer3" : "..." }
```

Answer to rewrite:

Figure 4: Prompt used to rewrite ground-truth answers into semantically equivalent but lexically distinct variants with DeepSeek-V3.

ration region allows the model to better exploit the diversity reward.

Taken together, these results indicate that model performance is sensitive to both the magnitude of the diversity reward and the scope of the exploration bounds. Careful calibration of β and γ is therefore crucial to achieving an effective trade-off between reference consistency and controlled diversity. This sensitivity analysis provides empirical support for our default hyperparameter choices and highlights the importance of constraining diversity exploration in verifier-free reinforcement learning.

C Case Study

Figures 6 and 5 present the responses generated by DARL and RLPR, respectively, for the same question. Similarly, Figures 7 and 8 compare the responses of the two methods on another identical input.

An analysis of these responses shows that both methods produce solutions that are fully consistent in terms of logical reasoning and numerical correctness. However, DARL exhibits clear advantages over RLPR in terms of reasoning clarity and presentation quality. In particular, DARL generates responses with more standardized reasoning steps, improved readability, and a more coherent and well-structured format.

Question:

The Great Pyramid of Giza was built as a tomb for Pharaoh Khufu with a square base of 230 meters per side and a height of 147 meters. To calculate the work done in constructing the pyramid, we assume the rocks started at ground level and were lifted into place with no air gaps. The density of the rocks is 2360 kg/m^3 and gravitational acceleration is 9.8 m/s^2 . The problem requires setting up and evaluating an integral using the cross-sectional side length function $s(y)$ at height y to find the total work done against gravity.

Reference:

2.21e12J

RLPR Response:

<think>

To determine the work done to build the Great Pyramid of Giza, we need to calculate the work done to lift each small slice of the pyramid to its final height. First, let's understand the geometry of the pyramid: The base of the pyramid is a square with side length 230 m. The height of the pyramid is 147 m. The side length of the square base at height y can be found using similar triangles. The ratio of the side length of the square cross-section at height y to the base side length (230 m) is equal to the ratio of the height from the top of the pyramid to that height y to the total height of the pyramid (147 m). The height from the top to height y is $147 - y$, and the height of the pyramid is 147 m. Therefore, the side length $s(y)$ of the square cross-section at height y is:

$$s(y) = 230 \cdot \left(\frac{147 - y}{147} \right) = 230 \cdot \left(1 - \frac{y}{147} \right)$$

The area of the cross-section at height y is:

$$A(y) = s(y)^2 = \left(230 \cdot \left(1 - \frac{y}{147} \right) \right)^2 = 230^2 \cdot \left(1 - \frac{y}{147} \right)^2$$

The volume of a thin slice of the pyramid at height y with thickness dy is:

$$dV = A(y) dy = 230^2 \cdot \left(1 - \frac{y}{147} \right)^2 dy$$

The density of the rocks is given as 2360 kg/m^3 , and the acceleration due to gravity is 9.8 m/s^2 . The mass of the thin slice is:

$$dm = (\text{density}) \cdot dV = 2360 \cdot 230^2 \cdot \left(1 - \frac{y}{147} \right)^2 dy$$

The weight of the thin slice is:

$$dF = dm \cdot g = 2360 \cdot 230^2 \cdot \left(1 - \frac{y}{147} \right)^2 \cdot 9.8 dy$$

The work done to lift this slice to its height y is:

$$dW = dF \cdot y = 2360 \cdot 230^2 \cdot 9.8 \cdot \left(1 - \frac{y}{147} \right)^2 \cdot y dy$$

We need to integrate this expression from $y = 0$ to $y = 147$:

$$W = \int_0^{147} 2360 \cdot 230^2 \cdot 9.8 \cdot \left(1 - \frac{y}{147} \right)^2 \cdot y dy$$

So the integral becomes:

$$W = 1227391200 \int_0^{147} \left(1 - \frac{y}{147} \right)^2 \cdot y dy$$

Let us use the substitution $u = 1 - \frac{y}{147}$, then $y = 147(1 - u)$ and $dy = -147 du$. The limits of integration change from $y = 0$ to $y = 147$ to $u = 1$ to $u = 0$.

The integral becomes:

$$W = 1227391200 \int_1^0 u^2 \cdot 147(1 - u) \cdot (-147) du$$

$$\int_0^1 u^2(1 - u) du = \int_0^1 (u^2 - u^3) du$$

$$= \left[\frac{u^3}{3} - \frac{u^4}{4} \right]_0^1 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

$$W \approx 2.652 \times 10^{13} \cdot \frac{1}{12} \approx 2.21 \times 10^{12}$$

So the work done to build the pyramid is approximately 2.21e12 J. </think>

<answer>2.21e12 J</answer>

Figure 5: Case Study of DARL on the WebInstruct Dataset.

Question:

The Great Pyramid of Giza was built as a tomb for Pharaoh Khufu with a square base of 230 meters per side and a height of 147 meters. To calculate the work done in constructing the pyramid, we assume the rocks started at ground level and were lifted into place with no air gaps. The density of the rocks is 2360 kg/m^3 and gravitational acceleration is 9.8 m/s^2 . The problem requires setting up and evaluating an integral using the cross-sectional side length function $s(y)$ at height y to find the total work done against gravity.

Reference:

2.21e12J

DARL Response:

<think>To find the work done to build the pyramid, we consider lifting each infinitesimal horizontal layer to its final height.

At $y = 0$, the side length is 230 m, and at $y = 147$ m, it is 0. Thus,

$$s(y) = 230 - \frac{230y}{147}$$

The volume of a thin slice is

$$dV = s(y)^2 dy = \left(230 - \frac{230y}{147}\right)^2 dy$$

Given density $\rho = 2360 \text{ kg/m}^3$, the mass is

$$dm = \rho dV = \rho \left(230 - \frac{230y}{147}\right)^2 dy$$

The work required to lift this slice to height y is

$$dW = dm \cdot g \cdot y = \rho g y \left(230 - \frac{230y}{147}\right)^2 dy$$

Thus, the total work is

$$W = \int_0^{147} 2360 \times 9.8 \times y \left(230 - \frac{230y}{147}\right)^2 dy$$

Factor:

$$230 - \frac{230y}{147} = 230 \left(1 - \frac{y}{147}\right)$$

$$W = 2360 \times 9.8 \times 230^2 \int_0^{147} y \left(1 - \frac{y}{147}\right)^2 dy$$

Expand:

$$\left(1 - \frac{y}{147}\right)^2 = 1 - \frac{2y}{147} + \frac{y^2}{147^2}$$

$$W = 2360 \times 9.8 \times 230^2 \int_0^{147} \left(y - \frac{2y^2}{147} + \frac{y^3}{147^2}\right) dy$$

Compute:

$$\int_0^{147} y dy = \frac{147^2}{2} = 10804.5$$

$$\int_0^{147} \frac{2y^2}{147} dy = \frac{2}{147} \cdot \frac{147^3}{3} \approx 14260.2$$

$$\int_0^{147} \frac{y^3}{147^2} dy = \frac{147^2}{4} = 5402.25$$

$$10804.5 - 14260.2 + 5402.25 \approx 2000$$

Thus,

$$W = 2360 \times 9.8 \times 230^2 \times 2000$$

$$230^2 = 52900, \quad 2360 \times 9.8 = 23128$$

$$W \approx 23128 \times 52900 \times 2000 \approx 2.45 \times 10^{10} \text{ J}$$

</think>

<answer> 2.45×10^{10} </answer>

Question:

The equilibrium constant, K_c , for the following reaction is 7.00×10^{-5} at 673 K.



Calculate the equilibrium concentration of HI when 0.590 moles of $\text{NH}_4\text{I}(s)$ are introduced into a 1.00 L vessel at 673 K.

Reference:

0.00837 M

DARL Response:

<think>

To determine the equilibrium concentration of HI when 0.590 moles of $\text{NH}_4\text{I}(s)$ are introduced into a 1.00 L vessel at 673 K, we'll follow these steps:

1. **Write the balanced equation and initial conditions:** The reaction is:



Since $\text{NH}_4\text{I}(s)$ is a solid, its concentration does not appear in the equilibrium expression. The equilibrium expression is:

$$K_c = \frac{[\text{NH}_3][\text{HI}]}{1} = [\text{NH}_3][\text{HI}]$$

The value of K_c is given as 7.00×10^{-5} .

2. **Set up the initial concentrations:** Initially, there are 0.590 moles of $\text{NH}_4\text{I}(s)$ in a 1.00 L vessel, which means initially there is no $\text{NH}_3(g)$ and no $\text{HI}(g)$, so the initial concentrations of NH_3 and HI are both 0 M.

When NH_4I starts to decompose, let x be the amount of NH_3 and HI formed at equilibrium. Then:



The changes in concentrations will be:

	$\text{NH}_4\text{I}(s)$	$\text{NH}_3(g)$	$\text{HI}(g)$
Initial (M)	—	0	0
Change (M)	—	+ x	+ x
Equilibrium (M)	—	x	x

The equilibrium concentrations are thus:

$$[\text{NH}_3] = x \quad \text{and} \quad [\text{HI}] = x$$

3. **Substitute into the equilibrium expression:** Using the expression for K_c :

$$K_c = [\text{NH}_3][\text{HI}] = x \cdot x = x^2$$

Given that $K_c = 7.00 \times 10^{-5}$, we have:

$$x^2 = 7.00 \times 10^{-5}$$

4. **Solve for x :** Taking the square root of both sides:

$$x = \sqrt{7.00 \times 10^{-5}}$$

Calculating the square root:

$$x = \sqrt{7.00 \times 10^{-5}} = 8.37 \times 10^{-3}$$

5. **Determine the equilibrium concentration of HI :** The equilibrium concentration of HI is equal to x , which is 8.37×10^{-3} M.

Therefore, the equilibrium concentration of HI is approximately 8.37×10^{-3} M.</think>

<answer> 8.37×10^{-3} M </answer>

Figure 7: Case Study of DARL on the WebInstruct Dataset.

Question:

The equilibrium constant, K_c , for the following reaction is 7.00×10^{-5} at 673 K.



Calculate the equilibrium concentration of HI when 0.590 moles of $\text{NH}_4\text{I}(s)$ are introduced into a 1.00 L vessel at 673 K.

Reference:

0.00837 M

RLPR Response:

<think>

To find the equilibrium concentration of HI when 0.590 moles of $\text{NH}_4\text{I}(s)$ are introduced into a 1.00 L vessel at 673 K, and we know the equilibrium constant $K_c = 7.00 \times 10^{-5}$ for the reaction:



First, note that the reaction involves the dissociation of solid NH_4I into gaseous NH_3 and HI . The concentration of solids does not appear in the equilibrium expression because their concentration is essentially constant and does not change during the reaction.

The balanced chemical reaction is:



Since NH_4I is a solid, we only need to focus on the concentrations of NH_3 and HI at equilibrium. Let's denote the change in the amount of NH_3 and HI at equilibrium as x moles of NH_3 and HI formed at equilibrium.

We start with 0.590 moles of $\text{NH}_4\text{I}(s)$. Since it's a solid and its amount will not change directly in the equilibrium equation, we can focus on the gaseous products.

At equilibrium, the amount of NH_3 and HI will be x moles each.

The initial amount of NH_3 and HI at the start of the reaction is 0 moles because they are formed from the dissociation of NH_4I .

At equilibrium, the amount of NH_3 and HI will both be x moles. Since the reaction starts with 0.590 moles of $\text{NH}_4\text{I}(s)$ and the reaction is 1:1:1, the moles of NH_3 and HI at equilibrium will be x moles each.

Since the volume of the container is 1.00 L, the equilibrium concentrations will be:

$$[\text{NH}_3] = x \text{ M}$$

$$[\text{HI}] = x \text{ M}$$

The equilibrium expression for the reaction will be:

$$K_c = [\text{NH}_3][\text{HI}]$$

We know from the problem that:

$$K_c = 7.00 \times 10^{-5}$$

Substituting the equilibrium concentrations into the K_c expression:

$$7.00 \times 10^{-5} = x \cdot x = x^2$$

To find x , we take the square root of both sides:

$$x = \sqrt{7.00 \times 10^{-5}}$$

Calculating the square root:

$$x = \sqrt{7.00 \times 10^{-5}} \approx 0.00837 \text{ M}$$

Thus, the equilibrium concentration of HI will be x .

So, the equilibrium concentration of HI will be 0.00837 M.

Therefore, the answer is: 0.00837 M</think>

<answer>0.00837 M</answer>

Figure 8: Case Study of RLPR on the WebInstruct Dataset.