
The Illusion of Diversity: Aligning LLM Exploration via Effective Entropy

Xiaoliang Fu^{1,*} Jiaye Lin^{2,*,†} Yangyi Fang^{3,*} Cong Qin^{4,*}
Binbin Zheng⁵ Chaowen Hu² Zekai Shao¹
¹Fudan University ²Meituan ³Tsinghua University
⁴Peking University ⁵University of Science and Technology of China
*Equal contribution. †Corresponding author: linjiaye@meituan.com

Abstract

Reinforcement learning with verifiable rewards has shifted the alignment paradigm of large language models from subjective preference to objective correctness, yet a fundamental disconnect persists between exploration diversity and reasoning validity. Existing strategies predominantly rely on “blind” entropy metrics, e.g., token-level or global sequence-level entropy, which structurally fail to distinguish between effective reasoning and potential hallucinations. This misalignment creates an illusion of diversity, where high entropy tends to signal chaotic degeneration rather than valuable exploration. To bridge this gap, we introduce *Effective Entropy*, a validity-aware entropy concept that quantifies diversity exclusively within the valid subspace and systematically filters out the noise of incorrect trajectories. Built on this concept, we instantiate VEPO, a simple Validity-aware Entropy Policy Optimization method that directly incorporates effective entropy into RLVR training. VEPO induces an adaptive attraction dynamic that drives up the probability of under-explored valid trajectories, a mechanism that matches the practical regime of large-scale autoregressive models where sequence probabilities remain extremely small. Experiments on six mathematical reasoning benchmarks with 1.5B and 7B DeepSeek-R1-Distill-Qwen models show consistent aggregate accuracy gains over GRPO and competitive repeated-sampling success. A direct evaluation with matched correct counts finds more distinct strategies on commonly solved MATH500 problems, supported by paired statistical tests and three-seed training comparisons.

1 Introduction

The alignment of Large Language Models (LLMs) has recently witnessed a paradigm shift from preference-based optimization to Reinforcement Learning with Verifiable Rewards (RLVR) [Guo et al., 2025, Lightman et al., 2023, Uesato et al., 2022, Shao et al., 2024]. Unlike subjective human preferences, RLVR anchors model training in objective correctness by utilizing ground-truth signals, such as compiler outputs or mathematical answers, to verify reasoning chains. This shift has driven substantial progress in complex reasoning tasks, where algorithms such as Group Relative Policy Optimization (GRPO) [Shao et al., 2024] show that simple statistical baselines can guide model improvement without the computational overhead of value networks.

However, a disconnect persists between the goal of reasoning and the mechanism used to encourage exploration. Existing RLVR methods often rely on standard entropy to mitigate premature convergence and encourage diversity [Williams and Peng, 1991, Eysenbach and Levine, 2021, Ahmed et al., 2019, He et al., 2025]. We argue that this reliance creates what we call *the illusion of diversity*. As illustrated in Figure 1, standard entropy is agnostic to validity and treats valid and invalid trajec-

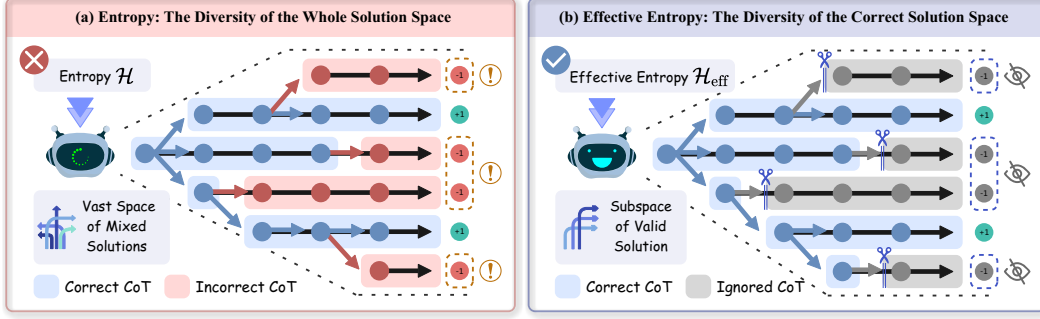


Figure 1: Exploration in RLVR. **(a) Standard entropy:** Conventional entropy operates on the whole solution space. Since the invalid subspace is combinatorially dominant, blind maximization tends to disperse probability toward diverse errors. **(b) Effective entropy:** We measure diversity only within the valid subspace. Filtering out invalid trajectories provides a validity-aware view of exploration and motivates a new optimization direction for reasoning alignment.

tories symmetrically. Because incorrect reasoning paths are combinatorially much more abundant than correct ones, blindly increasing entropy can disperse probability mass into the invalid region. As a result, high entropy may reflect chaotic degeneration rather than useful exploration.

To address this problem, we propose *Effective Entropy* ($\mathcal{H}_{\text{eff}}$), a validity-aware entropy concept that measures diversity only within the valid subspace specified by the reward function. Unlike standard entropy, which quantifies distributional spread over all trajectories, effective entropy focuses on the coverage of correct reasoning paths and ignores the internal distribution of incorrect outputs. This directly reframes exploration in RLVR: the central question is not how to increase uncertainty in general, but how to allocate uncertainty to valid reasoning.

To verify this idea, we instantiate a simple optimization method named **Validity-aware Entropy Policy Optimization (VEPO)**. VEPO directly adds effective entropy as an auxiliary objective to GRPO. Its gradient induces an adaptive attraction dynamic: under-explored valid trajectories receive larger positive learning pressure, encouraging the policy to expand the support of correct solutions. Although the mathematical form of effective entropy contains a theoretical boundary at $\pi(o) > e^{-1}$, this regime is practically unreachable for non-trivial autoregressive LLM trajectories, whose sequence probabilities are products of many token probabilities. Therefore, in realistic LLM training, VEPO behaves as a pure attraction mechanism toward under-explored valid trajectories.

In this paper, our contributions are summarized as follows:

- We identify a critical structural misalignment in existing RLVR exploration, revealing that the reliance on standard entropy fosters “the illusion of diversity” by conflating hallucination with meaningful exploration. To address this, we establish *Effective Entropy* as a validity-aware metric, and demonstrate that it serves as a superior optimization objective for reasoning alignment.
- We instantiate VEPO, a deliberately simple method that directly incorporates effective entropy into RLVR training via an adaptive attraction mechanism, serving as a clean empirical validation of the effective entropy principle rather than a heavily engineered algorithm.
- Experiments on six mathematical reasoning benchmarks across 1.5B and 7B LLMs show consistent aggregate accuracy gains over GRPO. Matched-count strategy measurements and paired training seeds substantiate useful diversity and repeatability within the tested setting.

2 Related Work

2.1 RLVR in LLMs

RLVR has emerged as a distinct paradigm from Reinforcement Learning from Human Feedback (RLHF), anchoring LLM reasoning in objective correctness rather than subjective human ranking. Early approaches often relied on dense process supervision or value networks, such as PPO [Schulman et al., 2017], but their computational burden limits scalability. Outcome-supervised RLVR has

therefore gained increasing attention. GRPO [Shao et al., 2024] is representative of this direction, estimating advantages via group statistics without an explicit value model. However, relying solely on outcome signals can lead to sparse feedback and premature convergence, making effective exploration design critical for further improving reasoning ability.

Rollout allocation provides a complementary route to improving RLVR learning signals. DynaMO [Fang et al., 2026] allocates rollouts using gradient-variance information and modulates token advantages to address gradient attenuation. VEPO retains group-based sampling and instead changes which trajectory uncertainty receives an auxiliary incentive.

2.2 Entropy-Aware Optimization

Policy entropy is a standard metric for characterizing exploration diversity [Williams and Peng, 1991, Eysenbach and Levine, 2021]. In GRPO-style training, entropy often decays as the model converges, which may cause the policy to collapse onto limited reasoning patterns. Early methods mitigate this issue by adding entropy regularization [O’Donoghue et al., 2016] or modifying clipping rules. Recent methods include dynamic clipping strategies such as Clip-Higher [Yu et al., 2025], Clip-Cov [Cui et al., 2025], DCPO [Yang et al., 2025a], and BAPO [Xi et al., 2025], as well as gradient-preserving methods such as CISPO [Chen et al., 2025] and GPPO [Su et al., 2025a,b].

Soft trust-region methods further refine update geometry: DGPO [Fu et al., 2026a] applies bilateral decoupled decay to probability-gradient weights, while MASPO [Fu et al., 2026b] combines soft gating with probability-mass adaptation and asymmetric signal reliability. These mechanisms regulate gradient utilization and stability; effective entropy specifies a validity-aware auxiliary objective.

More recent work shifts from blindly increasing entropy to controlling or utilizing it. AEPO [Wang et al., 2025a] stabilizes the learning process by anchoring policy entropy to a dynamic target. ANCHOR [Qin et al., 2026] likewise studies stable entropy dynamics for efficient LLM reasoning, complementing our focus on the validity of the trajectories contributing to entropy. Entropy Advantage [Cheng et al., 2025] and GTPO [Tan et al., 2025] use entropy to modulate token-level advantages. Advantage Reweighting [Yang et al., 2025b] down-weights overly uncertain tokens to focus learning on more reliable signals. Other works, such as HAPO [Liu et al., 2025] and ASPO [Wang et al., 2025b], further refine heterogeneous optimization behavior. Orthogonal theoretical perspectives also connect SFT and RL [Qin and Springenberg, 2025, Wu et al., 2025] or relate policy optimization to flow matching [Zhu et al., 2025].

Despite these advances, existing entropy-aware methods still treat entropy largely as a global uncertainty quantity. Our work focuses on a different question: whether the entropy being measured actually corresponds to valid reasoning diversity. Effective entropy answers this question by changing the domain of entropy measurement from the entire output space to the reward-verified valid subspace.

Adaptive sample weighting itself has a long history, including fuzzy memberships [Lin and Wang, 2002], manifold regularization [Belkin et al., 2006], and curricula [Bengio et al., 2009]. Our contribution is the verifier-masked policy functional and its derived coefficient, rather than per-sample weighting in general. Effective mutual information [Oh et al., 2025] instead adjusts query–response dependence against a reference joint distribution for distribution-shift analysis; it has a different domain and role from our fixed-query, verifier-gated output-entropy objective.

3 Preliminary

Problem Definition. We formulate the RLVR framework where a policy π_θ generates a reasoning trajectory o conditioned on a query q from dataset Q . The validity of each output is determined by a binary reward function $r(q, o) \in \{0, 1\}$. In practice, the reward can be obtained from deterministic verifiers, such as exact-answer matching, mathematical verification, or compiler execution.

GRPO samples a group of G outputs $\{o_i\}_{i=1}^G$ from the old policy $\pi_{\theta_{\text{old}}}$ for each query and computes their rewards $R_i = r(q, o_i)$. The advantage of the i -th output is estimated by standardizing rewards within the group:

$$\hat{A}_i = \frac{r(q, o_i) - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \quad (1)$$

GRPO then updates the policy by maximizing a clipped surrogate objective. Omitting the KL constraint for brevity, the objective is:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}} \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left[\rho_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(\rho_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right], \quad (2)$$

where

$$\rho_{i,t}(\theta) = \frac{\pi_{\theta}(x_{t,i} \mid x_{<t,i}, q)}{\pi_{\theta_{\text{old}}}(x_{t,i} \mid x_{<t,i}, q)} \quad (3)$$

is the token-level probability ratio.

Existing Use of Entropy. Entropy is commonly introduced to prevent premature convergence. A standard entropy-regularized objective adds an entropy term to the policy loss, while more recent methods use entropy to adapt clipping, rescale advantages, or control learning dynamics. These methods differ in implementation, but they share a common assumption: higher or better-controlled entropy during training implies better and more sustained exploration.

This assumption is incomplete in RLVR. Reasoning tasks care not only about distributional diversity, but also about whether the explored trajectories are valid. Standard entropy can increase even when the model spreads probability among invalid outputs. This motivates the need for a validity-aware entropy concept rather than blind entropy maximization.

4 Effective Entropy

4.1 The Misalignment of Existing Entropy Metrics

To motivate effective entropy, we first examine why common entropy metrics are structurally misaligned with reasoning validity in RLVR reasoning tasks.

Token Entropy Is Locally Blind. Token entropy is typically defined as

$$\mathcal{H}_{\text{tok}} = \frac{1}{T} \sum_t \mathcal{H}(\pi_{\theta}(\cdot \mid x_{<t}, q)). \quad (4)$$

It measures local uncertainty over the next-token distribution. However, in reasoning tasks, correctness is a property of the complete trajectory rather than an isolated token. A model can maintain high token entropy by randomizing locally plausible tokens that ultimately lead to invalid reasoning. Thus, token entropy is not a faithful proxy for useful reasoning exploration.

Sequence Entropy Is Globally Diluted. Sequence entropy is defined over the full trajectory space:

$$\mathcal{H}_{\text{seq}} = - \sum_{o \in \mathcal{O}} \pi_{\theta}(o \mid q) \log \pi_{\theta}(o \mid q). \quad (5)$$

Although it captures global trajectory diversity, it treats all trajectories symmetrically. In reasoning tasks, the valid set is sparse while the invalid set is combinatorially dominant. Therefore, maximizing sequence entropy over $\mathcal{O}$ tends to allocate uncertainty to invalid outputs. A high value of $\mathcal{H}_{\text{seq}}$ only indicates distributional spread, not valid reasoning diversity.

4.2 Definition of Effective Entropy

We introduce *Effective Entropy* to measure the diversity of valid reasoning trajectories. Let $r(o \mid q) \in \{0, 1\}$ be the binary correctness indicator. We define:

$$\mathcal{H}_{\text{eff}}(\pi_{\theta}) = - \sum_{o \in \mathcal{O}} r(o \mid q) \pi_{\theta}(o \mid q) \log \pi_{\theta}(o \mid q). \quad (6)$$

This definition changes the domain of entropy measurement. Standard sequence entropy measures uncertainty over all trajectories, while effective entropy measures uncertainty weighted by validity.

Invalid trajectories contribute zero, and only correct trajectories shape the metric. Effective entropy characterizes validity-weighted spread while also depending on valid probability mass; conditional strategy diversity requires a separate evaluation.

For model training, we express $\mathcal{H}_{\text{eff}}$ in its expectation form:

$$\mathcal{H}_{\text{eff}}(\pi_\theta) = \mathbb{E}_{o \sim \pi_\theta} [-r(o \mid q) \log \pi_\theta(o \mid q)], \quad (7)$$

which enables Monte Carlo estimation from sampled rollouts.

Effective entropy combines valid probability mass with conditional diversity, rather than measuring either alone (Appendix G.5). Terminal-verifier acceptance need not certify every reasoning step.

4.3 Why Effective Entropy Changes the Exploration Paradigm

The key role of effective entropy is conceptual rather than merely algorithmic. It suggests that exploration in RLVR should be evaluated by how much probability spread lies within the verified solution manifold. This view differs from conventional entropy maximization in two aspects.

First, it separates useful diversity from invalid dispersion. Standard entropy cannot distinguish a model that explores many correct strategies from one that produces many different wrong answers. Effective entropy directly encodes this distinction through the reward verifier.

Second, it provides a more aligned optimization signal. If a correct trajectory is under-explored, increasing its probability improves both reward learning and valid diversity. If an invalid trajectory is sampled, it receives no entropy-driven encouragement. This makes effective entropy naturally compatible with sparse verifiable rewards.

5 VEPO: Instantiating Effective Entropy for RLVR

VEPO is a direct instantiation of effective entropy in RLVR training. It is intentionally simple: rather than designing a complex optimizer, we add effective entropy as an auxiliary objective to GRPO to verify whether validity-aware entropy provides a better exploration signal.

5.1 Objective

We define the total optimization objective as:

$$\mathcal{J}_{\text{VEPO}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) + \beta \cdot \mathcal{H}_{\text{eff}}(\pi_\theta), \quad (8)$$

where $\beta > 0$ controls the strength of effective-entropy regularization.

This objective keeps the original RLVR learning signal from GRPO and augments it with a validity-aware diversity signal. Since $\mathcal{H}_{\text{eff}}$ ignores invalid trajectories, VEPO does not reward arbitrary uncertainty. Instead, it encourages the model to allocate additional probability mass to valid trajectories that are currently under-explored during policy optimization.

5.2 Gradient Analysis

Let the effective entropy loss be

$$\mathcal{L}_{\text{eff}}(\theta) = -\mathcal{H}_{\text{eff}}(\pi_\theta). \quad (9)$$

The gradient is:

$$\nabla_\theta \mathcal{L}_{\text{eff}}(\theta) = \mathbb{E}_{o \sim \pi_\theta} [r(o \mid q) (1 + \log \pi_\theta(o \mid q)) \nabla_\theta \log \pi_\theta(o \mid q)]. \quad (10)$$

For valid trajectories, the scalar weight is:

$$w(o) = r(o \mid q) (1 + \log \pi_\theta(o \mid q)). \quad (11)$$

When $r(o \mid q) = 1$ and $\pi_\theta(o \mid q) < e^{-1}$, we have $1 + \log \pi_\theta(o \mid q) < 0$. Minimizing the loss therefore increases the probability of this valid trajectory. In other words, VEPO attracts probability mass toward under-explored valid reasoning paths during RLVR training.

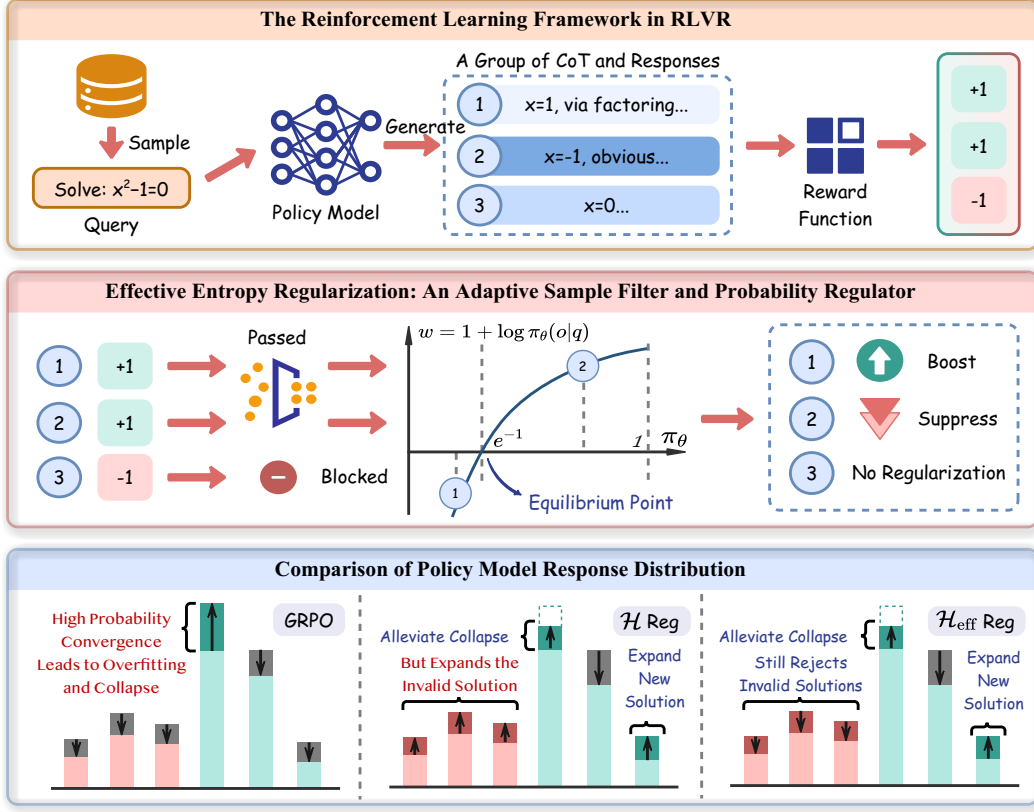


Figure 2: Overview of VEPO. Effective entropy filters exploration through the reward-defined valid subspace. VEPO directly adds this signal to GRPO, inducing an adaptive attraction mechanism that increases the probability of under-explored valid trajectories while ignoring invalid ones.

Theoretical Boundary and Practical LLM Regime. The mathematical form of Eq. 10 has a sign change when $\pi_\theta(o | q) > e^{-1}$. This would correspond to a boundary case where the objective decreases the probability of an overly dominant valid trajectory. However, this regime is practically unreachable for non-trivial autoregressive LLM outputs. A trajectory probability is the product of many conditional token probabilities; even moderately long reasoning chains have extremely small sequence probabilities. Therefore, in realistic LLM RLVR training, valid trajectories almost always satisfy $\pi_\theta(o | q) < e^{-1}$, and VEPO behaves as an adaptive attraction mechanism. We only mention this boundary for theoretical completeness and do not treat it as a practical training behavior.

Invalid Trajectories. For invalid trajectories, $r(o | q) = 0$, so their effective-entropy weight is zero. They receive no entropy-driven attraction. This is the core distinction between effective entropy and standard entropy: VEPO does not spend the exploration budget on invalid outputs.

5.3 Token-Level Implementation

For autoregressive models, we decompose the sequence-level update into token-level gradients. For a sampled trajectory o_i , we compute the sequence-level weight:

$$W_i = r(o_i | q) (1 + \log \pi_\theta(o_i | q)). \quad (12)$$

We detach this weight from the computation graph and apply it to the token log-probabilities:

$$\nabla_\theta \mathcal{L}_{\text{eff}} \approx \frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{T_i} \text{sg}[W_i] \cdot \nabla_\theta \log \pi_\theta(x_{t,i} | x_{<t,i}, q), \quad (13)$$

where $\text{sg}[\cdot]$ denotes stop-gradient.

This implementation is deliberately minimal. It does not introduce additional value networks, process reward models, or complex control mechanisms. Its purpose is to provide a clean test of whether effective entropy is a useful objective for RLVR exploration.

The sequence log-probability and token-gradient aggregation are both raw sums; response-length normalization would define a different functional. Sharing W_i does not identify causally important reasoning tokens. Eq. 10 is exact for on-policy samples; reusing old-policy rollouts produces a local surrogate rather than an unbiased current-policy estimator. Appendix G.6 gives the corresponding detached-weight implementation, and Appendix G.3 isolates correctness gating and length normalization.

6 Experiments

6.1 Experimental Setup

Training Configuration. We evaluate VEPO on the DeepSeek-R1-Distill-Qwen family [Guo et al., 2025], using 1.5B and 7B backbones. All models are fine-tuned on DAPO-Math-17K [Yu et al., 2025], a high-quality dataset for mathematical reasoning. Our training infrastructure follows the GRPO framework. We use a global batch size of 512 and a mini-batch size of 32. The group size is $G = 8$ and the clipping threshold is fixed at $\epsilon = 0.2$. The effective entropy coefficient β is selected through grid search on a held-out validation set.

Verification Protocol. To ensure reliable RL signals, we use a strict dual-oracle verification protocol. During training, we use `math_verify` [Kydliček, 2025] for efficient online reward calculation. During evaluation, we use the more rigorous `prime_math` [Lightman et al., 2023] to reduce reward hacking and improve final metric validity.

Evaluation Benchmarks and Metrics. We evaluate on AIME 2024 [MAA, 2024], AIME 2025 [MAA, 2025], AMC 2023 [MAA, 2023], MATH500 [Hendrycks et al., 2021], Minerva [Lewkowycz et al., 2022], and OlympiadBench [He et al., 2024]. We report two metrics. **Avg@32** measures expected accuracy over 32 sampled trajectories. **Pass@32** measures success under 32 repeated samples. It is affected by per-sample accuracy and does not directly measure semantic strategy diversity; we evaluate that property separately in Section 6.3.

Baselines. We compare VEPO with vanilla GRPO and several entropy-aware baselines: **Entropy-Reg** [O’Donoghue et al., 2016], which adds a standard entropy regularizer; **AEPO** [Wang et al., 2025a], which controls entropy around a dynamic target; **Entropy-Adv** [Cheng et al., 2025] and **GTPO** [Tan et al., 2025], which use entropy to adjust learning signals; and **Adv-Reweight** [Yang et al., 2025b], which down-weights certain uncertain or high-risk updates.

6.2 Main Results

Table 1 summarizes the main results. VEPO improves the six-benchmark aggregate over vanilla GRPO across both 1.5B and 7B backbones, while individual Pass@32 comparisons are mixed. On the 1.5B model, VEPO achieves the best average Avg@32 and Pass@32 among the compared methods. On the 7B model, VEPO also obtains the strongest average Avg@32 and a competitive Pass@32. These results suggest that effective entropy provides a more aligned exploration signal than standard entropy or entropy-derived advantage reshaping methods.

The key point is not that VEPO is a sophisticated optimizer, but that even this simple instantiation of effective entropy yields consistent gains. This supports our central claim: measuring diversity within the valid subspace provides a better foundation for RLVR exploration than measuring entropy over all tokens or all sampled trajectories.

Repeated-Sampling Success and Training Dynamics. Figure 3(a,d) reports repeated-sampling success across k ; these curves do not independently establish semantic strategy diversity. Panels (b,e) show AIME 2025 training dynamics, while (c,f) track token and effective entropy. VEPO sustains learning and higher effective entropy in the tested setting.

Table 1: Main results on mathematical reasoning benchmarks. We report Avg@32 (A@32) and Pass@32 (P@32). Bold indicates the best performance within each backbone.

	<u>AIME24</u>		<u>AIME25</u>		<u>AMC23</u>		<u>MATH500</u>		<u>Minerva</u>		<u>Olympiad</u>		<u>Average</u>	
Method	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32
<i>Backbone: DeepSeek-R1-Distill-Qwen-1.5B</i>														
GRPO	33.2	71.8	27.7	49.9	79.5	94.8	77.6	90.8	26.1	48.8	46.3	64.7	48.4	70.1
Entropy-Reg	32.4	63.3	28.4	56.2	81.3	95.0	78.4	90.5	25.2	47.0	46.3	65.0	48.7	69.5
Entropy-Adv	29.6	67.8	23.9	44.2	77.1	94.1	78.5	89.7	25.2	48.6	46.4	64.3	46.8	68.1
Adv-Reweighting	25.0	59.1	22.3	40.7	70.6	94.6	75.6	90.1	25.1	51.1	42.3	63.2	43.5	66.5
GTPO	23.0	63.2	19.6	45.1	67.4	91.6	74.3	90.5	20.8	48.0	41.5	64.3	41.1	67.1
AEPO	33.9	64.9	28.1	57.1	80.2	97.2	77.4	91.3	24.2	47.1	46.9	66.5	48.5	70.7
VEPO	35.9	72.3	27.8	52.1	85.4	96.5	80.2	91.7	26.5	52.8	47.7	64.0	50.6	71.6
<i>Backbone: DeepSeek-R1-Distill-Qwen-7B</i>														
GRPO	48.2	82.5	37.4	60.5	88.1	96.6	84.8	92.4	37.4	57.2	57.2	73.9	58.9	77.2
Entropy-Reg	49.7	82.7	36.1	60.5	89.4	97.9	84.9	92.3	37.5	54.6	57.9	73.5	59.2	76.9
Entropy-Adv	48.1	83.5	36.8	61.5	88.1	99.8	84.1	91.3	37.4	56.1	56.6	73.5	58.5	77.6
Adv-Reweighting	43.0	76.3	31.9	56.4	83.8	94.4	83.7	93.9	35.4	56.5	55.9	74.7	55.6	75.4
GTPO	43.8	83.4	32.0	63.5	85.1	95.0	83.8	95.3	32.8	57.6	54.5	73.6	55.3	78.1
AEPO	50.0	78.1	37.9	61.9	87.9	97.4	85.4	95.1	36.6	59.4	58.1	76.7	59.3	78.1
VEPO	51.3	86.0	41.1	63.4	91.5	97.5	85.5	92.6	38.3	57.4	59.0	74.2	61.1	78.5

Table 2: Direct strategy diversity on the 441 commonly eligible MATH500-7B problems. Each method contributes exactly eight correct trajectories per problem. Hill-1 is the effective number of strategy modes [Hill, 1973].

Metric	GRPO	VEPO	Difference
Distinct clusters/problem	2.31	2.68	+0.37
Hill-1 effective modes	1.97	2.24	+0.27
Largest-cluster share	0.68	0.61	−0.07

6.3 Direct Strategy Diversity and Statistical Reliability

To separate strategy diversity from accuracy, we compare GRPO and VEPO on the 441 MATH500-7B problems where both produce at least eight correct trajectories in their 32-sample pools. We match exactly eight correct trajectories per method per problem, average over 10,000 rarefactions, and cluster answer-stripped reasoning using frozen gte-Qwen2 embeddings [Alibaba NLP, 2024] and a method-blinded, validated threshold. Appendix G.1 details the protocol and robustness checks.

We strip final-answer spans, pool the remaining chains of thought from both methods, and cluster within each problem using frozen gte-Qwen2-7B-instruct embeddings [Alibaba NLP, 2024], L2 normalization, and average-linkage agglomerative clustering with cosine distance [Müllner, 2011]. A strategy denotes a shared mathematical plan, invariant, decomposition, or substitution, irrespective of wording or notation. Before comparing methods, we fix the pipeline and select one global distance threshold on method-blinded development annotations. Two mathematical annotators label 600 same/different-strategy pairs with Cohen’s $\kappa = 0.82$; the frozen pipeline achieves precision 0.88 and recall 0.84 on a disjoint held-out set.

The paired cluster-count gain is +0.37 (16.0%), with 95% problem-bootstrap CI [0.24, 0.50] and sign-flip permutation $p < 10^{-4}$. Matching correct counts prevents this increase from arising mechanically from higher accuracy. This finite-sample endpoint concerns commonly solved, answer-verified prompts; it does not certify every reasoning step.

Three paired training seeds (0, 1, 2) yield six-benchmark Avg@32 gains of +2.1, +2.4, +2.1 percentage points at 1.5B and +2.3, +2.4, +1.9 at 7B. All paired Avg@32 and Pass@32 gains are positive (Table 7); three seeds support repeatability without strong seed-level significance claims.

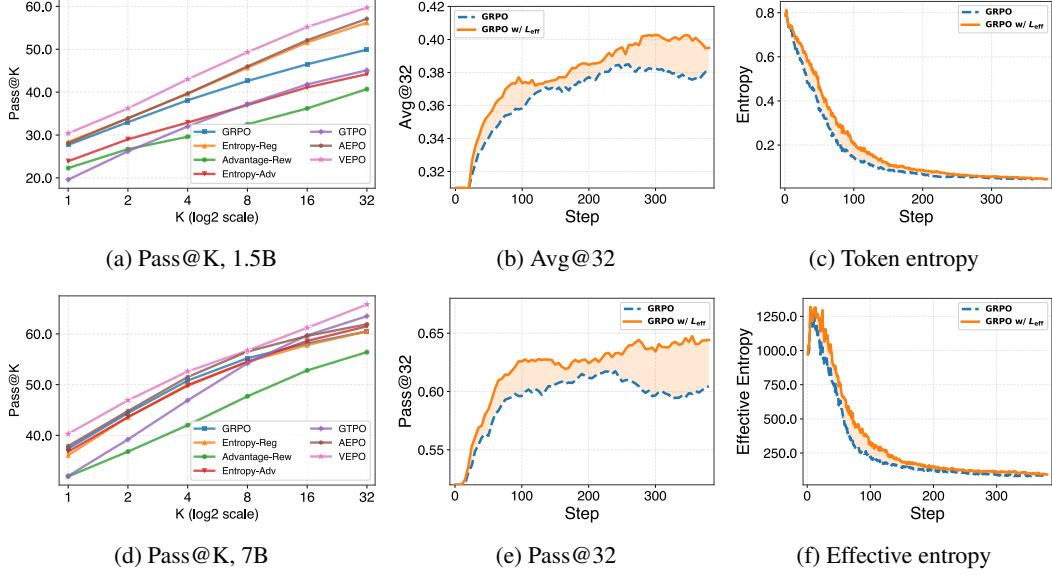


Figure 3: Main empirical results. Subfigures (a) and (d) show Pass@K on AIME 2025 for the 1.5B and 7B models. Subfigures (b) and (e) show Avg@32 and Pass@32 training dynamics on the 7B model. Subfigures (c) and (f) compare token entropy and effective entropy during training. Together, these results show that VEPO improves performance while sustaining stronger validity-aware exploration across different model scales.

6.4 Hyperparameter Analysis

We conduct hyperparameter experiments with $\beta \in \{3e-5, 1e-4, 1e-3, 3e-3, 1e-2\}$. As shown in Figure 4, effective entropy regularization affects the balance between exploration and convergence. Larger β values maintain higher entropy and effective entropy, but may keep the model in an active exploration phase for too long. In contrast, a moderate value such as $\beta = 1e-4$ provides sufficient valid-space exploration while allowing efficient convergence. This is used as our default.

6.5 Attribution Analysis

To understand why effective entropy improves RLVR training, we compare VEPO with GRPO and standard Entropy-Reg. We analyze three aspects: the magnitude of exploration, the structural quality of exploration, and the conversion efficiency from exploration to performance.

Magnitude and Direction. Figure 5a shows that both VEPO and Entropy-Reg can maintain higher token entropy than GRPO. However, similar entropy magnitude does not imply similar exploration quality. To distinguish the direction of exploration more precisely, we define the relative entropy gain of a method M as:

$$G(M) = \frac{\mathcal{H}_M}{\mathcal{H}_{\text{GRPO}}}. \quad (14)$$

As shown in Figure 5b, Entropy-Reg increases total entropy and effective entropy in a coupled manner, suggesting a largely blind expansion of the distribution. In contrast, VEPO yields a stronger gain in effective entropy relative to total entropy. This indicates that VEPO preferentially expands the valid subspace rather than merely spreading probability over all outputs.

Structural Quality. We further define the effective ratio:

$$\rho_{\text{eff}} = \frac{\mathcal{H}_{\text{eff}}}{\mathcal{H}_{\text{seq}}}. \quad (15)$$

This metric measures the fraction of sequence-level uncertainty that is allocated to valid trajectories. As shown in Figure 5c, Entropy-Reg can reduce the effective ratio in later training stages, implying that blind entropy regularization may dilute the policy’s focus on correctness. VEPO maintains

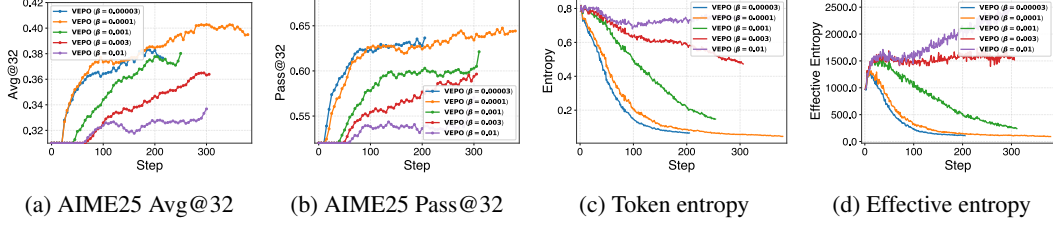


Figure 4: Impact of the effective entropy coefficient β . Larger β maintains stronger exploration but can delay convergence. A moderate coefficient provides the best balance between valid diversity and optimization efficiency across training stages.

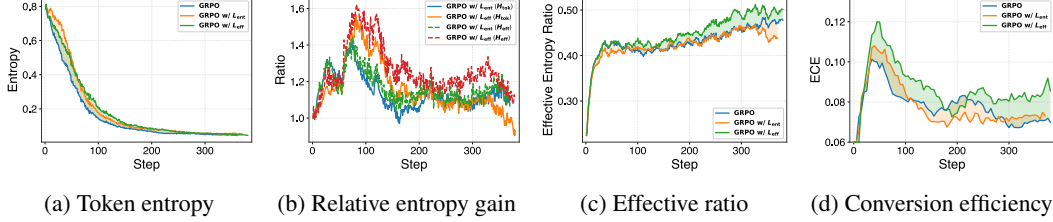


Figure 5: Attribution analysis. Entropy-Reg increases raw entropy but does not reliably improve the structural quality of exploration. VEPO allocates more exploration to the valid subspace, leading to higher effective ratio and stronger conversion efficiency.

a higher effective ratio, showing that its exploration is structurally cleaner and more consistently aligned with verified reasoning.

Conversion Efficiency. Finally, we evaluate whether additional exploration translates into performance gain. Let $\Delta\mathcal{P}$ be the performance gain and $\Delta\mathcal{H}$ be the entropy change relative to initialization. We define Entropy Conversion Efficiency (ECE) as:

$$\text{ECE} = \Delta\mathcal{P} \cdot \exp\left(\text{sgn}(\Delta\mathcal{P}) \cdot \frac{\Delta\mathcal{H}}{\mathcal{H}_{\text{init}}}\right). \quad (16)$$

As shown in Figure 5d, Entropy-Reg provides only transient benefit and later converges toward the GRPO baseline. VEPO maintains a higher ECE trajectory, indicating that validity-aware exploration is more effectively converted into reasoning performance. These diagnostics describe allocation and conversion of entropy during optimization. The matched-correct-count analysis in Section 6.3 provides a complementary, direct assessment of strategy diversity while keeping the number of correct samples fixed within each problem.

7 Conclusion

In this paper, we identify a structural weakness in conventional entropy-based exploration for RLVR. Standard entropy measures distributional spread without considering whether the explored trajectories are valid, thereby conflating useful reasoning diversity with invalid dispersion. To address this issue, we introduce *Effective Entropy*, a validity-aware entropy concept that measures diversity only within the reward-verified valid subspace of reasoning trajectories.

Based on effective entropy, we instantiate VEPO, a simple and direct optimization method that adds effective entropy to GRPO. VEPO induces an adaptive attraction mechanism toward under-explored valid reasoning trajectories and ignores invalid ones. Experiments on six mathematical reasoning benchmarks show consistent aggregate accuracy gains over GRPO and competitive repeated-sampling success against entropy-aware baselines.

Our mathematical RLVR results support validity-aware exploration on the tested backbones, rather than universal transfer. Direct matched-count strategy measurements and paired training seeds strengthen the evidence beyond entropy and pass rates alone. Larger models, interactive agents, process-level verification, and open-ended rewards remain directions for future work.

Acknowledgments and Disclosure of Funding

Funding and Computing Support. Meituan provided the computing resources used in this work. The study was initiated by employees as independent research and was not part of a formally designated Meituan project. The authors received no other funding or support for this work.

Competing Interests. Xiaoliang Fu, Yangyi Fang, Cong Qin, Binbin Zheng, and Zekai Shao were previously employed by Meituan. Jiaye Lin and Chaowen Hu list Meituan as their current affiliation. The authors declare no other competing interests.

References

- Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pages 151–160. PMLR, 2019.
- Alibaba NLP. gte-Qwen2-7B-instruct model card, 2024. URL <https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct>.
- M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006. URL <https://jmlr.org/papers/v7/belkin06a.html>.
- Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009. doi: 10.1145/1553374.1553380.
- A. Chen, A. Li, B. Gong, B. Jiang, B. Fei, B. Yang, B. Shan, C. Yu, C. Wang, C. Zhu, et al. Minimax-ml: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.
- D. Cheng, S. Huang, X. Zhu, B. Dai, W. X. Zhao, Z. Zhang, and F. Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.
- G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.
- B. Eysenbach and S. Levine. Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257*, 2021.
- Y. Fang, J. Lin, X. Fu, C. Qin, H. Shi, C. Hu, L. Pan, K. Zeng, and X. Cai. How to allocate, how to learn? dynamic rollout allocation and advantage modulation for policy optimization. *arXiv preprint arXiv:2602.19208*, 2026. URL <https://arxiv.org/abs/2602.19208>.
- X. Fu, J. Lin, Y. Fang, C. Hu, C. Qin, Z. Shao, B. Zheng, L. Pan, and K. Zeng. From $\log \pi$ to π : Taming divergence in soft clipping via bilateral decoupled decay of probability gradient weight. *arXiv preprint arXiv:2603.14389*, 2026a. URL <https://arxiv.org/abs/2603.14389>.
- X. Fu, J. Lin, Y. Fang, B. Zheng, C. Hu, Z. Shao, C. Qin, L. Pan, K. Zeng, and X. Cai. MASPO: Unifying gradient utilization, probability mass, and signal reliability for robust and sample-efficient LLM reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41348–41365. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1918. URL <https://aclanthology.org/2026.acl-long.1918/>.
- D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, et al. Olympiad-bench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

- J. He, J. Liu, C. Y. Liu, R. Yan, C. Wang, P. Cheng, X. Zhang, F. Zhang, J. Xu, W. Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.
- D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973. doi: 10.2307/1934352.
- H. Kydlíček. Math-Verify: Math Verification Library, 2025. URL <https://github.com/huggingface/math-verify>.
- A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.
- H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL <https://arxiv.org/abs/2305.20050>.
- C.-F. Lin and S.-D. Wang. Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, 13(2):464–471, 2002. doi: 10.1109/72.991432.
- Z. Liu, M. Liu, S. Wen, M. Cai, B. Cui, C. He, and W. Zhang. From uniform to heterogeneous: Tailoring policy optimization to every token’s nature. *arXiv preprint arXiv:2509.16591*, 2025.
- MAA. American mathematics competitions - amc, 2023. URL <https://maa.org/>. Accessed: 2023.
- MAA. American invitational mathematics examination - aime, 2024. URL <https://maa.org/>. Accessed: 2024.
- MAA. American invitational mathematics examination - aime, 2025. URL <https://maa.org/>. Accessed: 2025.
- D. Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011. URL <https://arxiv.org/abs/1109.2378>.
- B. O’Donoghue, R. Munos, K. Kavukcuoglu, and V. Mnih. Combining policy gradient and q-learning. *arXiv preprint arXiv:1611.01626*, 2016.
- C. Oh, Z. Fang, S. Im, X. Du, and Y. Li. Understanding multimodal LLMs under distribution shifts: An information-theoretic approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 46943–46970, 2025. URL <https://proceedings.mlr.press/v267/oh25a.html>.
- C. Qin and J. T. Springenberg. Supervised fine tuning on curated data is reinforcement learning (and can be improved). *arXiv preprint arXiv:2507.12856*, 2025.
- C. Qin, J. Lin, X. Fu, Y. Fang, C. Hu, H. Zhang, Y. Guo, Z. Gu, K. Qin, J. Kang, D. Ding, and P. Zhao. ANCHOR: Taming entropy dynamics for stable and efficient reasoning of large language models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 4046–4057. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3818081. URL <https://doi.org/10.1145/3770855.3818081>.
- J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

- G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.
- Z. Su, L. Pan, X. Bai, D. Liu, G. Dong, J. Huang, W. Hu, F. Zhang, K. Gai, and G. Zhou. Klear-reasoner: Advancing reasoning capability via gradient-preserving clipping policy optimization. *arXiv preprint arXiv:2508.07629*, 2025a.
- Z. Su, L. Pan, M. Lv, Y. Li, W. Hu, F. Zhang, K. Gai, and G. Zhou. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning. *arXiv preprint arXiv:2509.20712*, 2025b.
- H. Tan, J. Pan, J. Lin, T. Chen, Z. Zheng, Z. Tang, and H. Yang. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*, 2025.
- J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.
- C. Wang, Z. Li, J. Bai, Y. Zhang, S. Cui, Z. Zhao, and Y. Wang. Arbitrary entropy policy optimization: Entropy is controllable in reinforcement finetuning. *arXiv e-prints*, pages arXiv–2510, 2025a.
- J. Wang, R. Liu, L. Lin, W. Hu, X. Li, F. Zhang, G. Zhou, and K. Gai. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*, 2025b.
- L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, 2024. doi: 10.18653/v1/2024.acl-long.642. URL <https://aclanthology.org/2024.acl-long.642/>.
- S. Wang, L. Yu, C. Gao, C. Zheng, S. Liu, R. Lu, K. Dang, X. Chen, J. Yang, Z. Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025c.
- R. J. Williams and J. Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.
- Y. Wu, Y. Zhou, Z. Ziheng, Y. Peng, X. Ye, X. Hu, W. Zhu, L. Qi, M.-H. Yang, and X. Yang. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*, 2025.
- Z. Xi, X. Guo, Y. Nan, E. Zhou, J. Shen, W. Chen, J. Liu, J. Huang, Z. Zhang, H. Guo, et al. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025.
- S. Yang, C. Dou, P. Guo, K. Lu, Q. Ju, F. Deng, and R. Xin. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025a.
- Z. Yang, X. Luo, Z. Wang, D. Han, Z. He, D. Li, and Y. Xu. Do not let low-probability tokens over-dominate in rl for llms. *arXiv preprint arXiv:2505.12929*, 2025b.
- Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- X. Zhu, D. Cheng, D. Zhang, H. Li, K. Zhang, C. Jiang, Y. Sun, E. Hua, Y. Zuo, X. Lv, et al. Flowrl: Matching reward distributions for llm reasoning. *arXiv preprint arXiv:2509.15207*, 2025.

A Why Effective Entropy: Domain and Focus of Exploration

In this section, we provide a comparative analysis of why Effective Entropy provides the correct “focus” for reasoning tasks compared to standard formulations. The core distinction lies not in the calculation method, but in the **domain** over which the uncertainty is measured.

Justification of the Inequality $|\mathcal{O}_{\text{invalid}}| \gg |\mathcal{O}_{\text{valid}}|$. The assumption that the invalid subspace combinatorially dominates the valid subspace is fundamental to our critique of standard entropy. We justify this formally through the lens of combinatorial expansion and logical fragility:

1. **Total Space Expansion:** Consider a reasoning chain of length L generated from a vocabulary $\mathcal{V}$. The total output space cardinality is $|\mathcal{O}| = |\mathcal{V}|^L$. For modern LLMs, $|\mathcal{V}| \approx 10^5$, creating a hyper-astronomical search space.
2. **Logical Sparsity:** Validity is a *strict constraint satisfaction problem*. A valid reasoning chain requires a precise sequence of logical transitions. If we model a reasoning path as a chain of K logical steps, and at each step only a small subset of tokens $\mathcal{V}_{\text{correct}} \subset \mathcal{V}$ maintains logical consistency, the size of the valid manifold scales as $|\mathcal{O}_{\text{valid}}| \approx |\mathcal{V}_{\text{correct}}|^L$.
3. **Exponential Divergence:** Since $|\mathcal{V}_{\text{correct}}| \ll |\mathcal{V}|$ (e.g., there are few correct numbers but infinite incorrect ones), the ratio of valid to total sequences is:

$$\frac{|\mathcal{O}_{\text{valid}}|}{|\mathcal{O}|} \approx \left(\frac{|\mathcal{V}_{\text{correct}}|}{|\mathcal{V}|} \right)^L \rightarrow 0 \quad \text{as } L \rightarrow \infty \quad (17)$$

Thus, $\mathcal{O}_{\text{invalid}} = \mathcal{O} \setminus \mathcal{O}_{\text{valid}}$ constitutes the overwhelming majority of the configuration space. Any metric maximizing distribution spread over $\mathcal{O}$ without constraints (like $\mathcal{H}_{\text{seq}}$) will statistically force the probability mass into $\mathcal{O}_{\text{invalid}}$ due to this dimensional curse.

Domain Comparison of Metrics. Let $\mathcal{O}$ be the set of all possible output sequences.

- **Token Entropy** ($\mathcal{H}_{\text{tok}}$) measures uncertainty over the vocabulary V at each step t .
 - *Focus:* Local next-token prediction.
 - *Limitation:* It is agnostic to the set $\mathcal{O}_{\text{valid}}$. A model can have high $\mathcal{H}_{\text{tok}}$ by randomizing tokens that lead to $o \notin \mathcal{O}_{\text{valid}}$.
- **Sequence Entropy** ($\mathcal{H}_{\text{seq}}$) measures uncertainty over the full set $\mathcal{O}$.
 - *Focus:* Global distribution spread.
 - *Limitation:* It treats $\mathcal{O}_{\text{valid}}$ and $\mathcal{O}_{\text{invalid}}$ symmetrically. Given the inequality $|\mathcal{O}_{\text{invalid}}| \gg |\mathcal{O}_{\text{valid}}|$ established above, maximizing this metric predominantly focuses on the invalid subspace.
- **Effective Entropy** ($\mathcal{H}_{\text{eff}}$) measures uncertainty weighted by the validity indicator.
 - *Focus:* Distribution spread strictly within $\mathcal{O}_{\text{valid}}$.
 - *Advantage:* It mathematically ignores the internal dynamics of invalid sequences. It rewards the model only for expanding the support of the distribution covering correct answers.

B Derivations and Theoretical Analysis

B.1 Theoretical Analysis of Global Optima

To rigorously justify why effective entropy provides a more aligned exploration principle for RLVR, we analyze the global optima of three objectives: standard reward maximization, standard entropy regularization, and effective-entropy regularization.

Let $\mathcal{S} = \mathcal{O}_{\text{valid}}$ be the set of correct answers and $\mathcal{N} = \mathcal{O}_{\text{invalid}}$ be the set of incorrect answers, such that the output space $\mathcal{O} = \mathcal{S} \cup \mathcal{N}$. Let the reward function be binary: $r(o) = 1$ if $o \in \mathcal{S}$ and 0 otherwise. We assume $|\mathcal{N}| \gg |\mathcal{S}|$.

Case 1: Standard RL (Max Reward). The objective is $\max_{\pi} \mathbb{E}_{o \sim \pi} [r(o)] = \sum_{o \in \mathcal{S}} \pi(o)$.

- **Solution Space:** The maximum possible reward is 1. Any policy π such that $\sum_{o \in \mathcal{S}} \pi(o) = 1$ is a global optimum.
- **Analysis:** The optimization landscape is degenerate. Without a diversity objective, a **Dirac distribution** on one solution $o^* \in \mathcal{S}$ is permitted, but the reward objective alone does not imply that every optimizer converges to such a distribution.

Case 2: Standard Entropy Regularization (The Illusion of Diversity). We consider the standard regularization objective where diversity is encouraged over the entire output space $\mathcal{O}$:

$$\mathcal{J}_{\text{reg}}(\pi) = \mathbb{E}[r] + \beta \mathcal{H}_{\text{seq}} = \sum_{o \in \mathcal{O}} \pi(o) r(o) - \beta \sum_{o \in \mathcal{O}} \pi(o) \log \pi(o) \quad (18)$$

Using Lagrange multipliers with constraint $\sum \pi(o) = 1$, the Lagrangian is:

$$\mathcal{L} = \sum_{o \in \mathcal{O}} \pi(o) r(o) - \beta \sum_{o \in \mathcal{O}} \pi(o) \log \pi(o) + \gamma \left(\sum_{o \in \mathcal{O}} \pi(o) - 1 \right) \quad (19)$$

Solving $\frac{\partial \mathcal{L}}{\partial \pi(o)} = 0$ yields the standard Boltzmann (Gibbs) distribution:

$$\pi^*(o) = \frac{\exp(r(o)/\beta)}{Z}, \quad \text{where } Z = \sum_{o' \in \mathcal{O}} \exp(r(o')/\beta) \quad (20)$$

- **Leakage into Invalid Subspace:** For any valid solution $o_s \in \mathcal{S}$ and invalid solution $o_n \in \mathcal{N}$:

$$\pi^*(o_s) \propto e^{1/\beta}, \quad \pi^*(o_n) \propto e^{0/\beta} = 1 \quad (21)$$

Crucially, $\pi^*(o_n) \neq 0$. The optimal policy *must* assign non-zero probability to invalid answers to satisfy the entropy constraint.

- **Analysis (The Illusion):** Although $\pi^*(o_s) > \pi^*(o_n)$, the total probability mass assigned to hallucinations is:

$$P(\text{Error}) = \frac{|\mathcal{N}|}{|\mathcal{N}| + |\mathcal{S}| e^{1/\beta}} \quad (22)$$

Since $|\mathcal{N}|$ (combinatorial errors) is typically orders of magnitude larger than $|\mathcal{S}|$, even for moderate β , significant probability mass is forced into the invalid region. This mathematically proves that standard entropy regularization creates an "illusion of diversity" by diluting validity with inevitable hallucinations.

Case 3: VEPO (Max Reward + Effective Entropy). We consider the VEPO objective:

$$\mathcal{J}_{\text{vepo}}(\pi) = \mathbb{E}[r] + \beta \mathcal{H}_{\text{eff}} = \sum_{o \in \mathcal{O}} \pi(o) r(o) - \beta \sum_{o \in \mathcal{O}} r(o) \pi(o) \log \pi(o) \quad (23)$$

Substituting the binary reward structure ($r(o) = 0$ for $o \in \mathcal{N}$ eliminates terms in both sums):

$$\mathcal{J}_{\text{vepo}}(\pi) = \sum_{o \in \mathcal{S}} \pi(o) (1 - \beta \log \pi(o)) \quad (24)$$

Scope and Coefficient Condition. We analyze a finite trajectory simplex with a nonempty valid set and $0 < \beta \leq 1$, which includes all tested coefficients. For a valid coordinate, the derivative of $\pi(o)[1 - \beta \log \pi(o)]$ is $1 - \beta[1 + \log \pi(o)]$, positive whenever $\pi(o) < 1$. Consequently, any positive invalid mass can be shifted to a valid coordinate to improve the objective. This condition is required for the argument below.

Step 1: Probability Assignment to Invalid Sets. By the preceding derivative argument, the optimum under the stated coefficient condition has $\pi^*(o) = 0$ for all $o \in \mathcal{N}$.

Step 2: Distribution within Valid Sets. We solve for π within $\mathcal{S}$ using Lagrange multipliers:

$$\mathcal{L} = \sum_{i \in \mathcal{S}} \pi_i (1 - \beta \log \pi_i) + \gamma \left(\sum_{i \in \mathcal{S}} \pi_i - 1 \right) \quad (25)$$

Setting $\frac{\partial \mathcal{L}}{\partial \pi_k} = 0$:

$$1 - \beta(1 + \log \pi_k) + \gamma = 0 \implies \pi_k = \exp\left(\frac{1 - \beta + \gamma}{\beta}\right) \quad (26)$$

Since the RHS is constant for all $k \in \mathcal{S}$, we have $\pi_i = \pi_j$.

- **Conclusion:** The unique global optimum is the **Uniform Distribution** strictly over the set of correct answers:

$$\pi^*(o) = \begin{cases} \frac{1}{|\mathcal{S}|} & \text{if } o \in \mathcal{S} \\ 0 & \text{if } o \in \mathcal{N} \end{cases} \quad (27)$$

This is an optimum of the idealized expected-reward functional under the stated assumptions; it does not guarantee convergence of clipped GRPO with neural policies.

B.2 Proof of Effective Entropy Gradient

Theorem B.1. *Given the Effective Entropy loss $\mathcal{L}_{\text{eff}}(\theta) = \sum_{o \in \mathcal{O}} r(o|q) \pi_\theta(o|q) \log \pi_\theta(o|q)$, its gradient is given by:*

$$\nabla_\theta \mathcal{L}_{\text{eff}}(\theta) = \mathbb{E}_{o \sim \pi_\theta} [r(o|q)(1 + \log \pi_\theta(o|q)) \nabla_\theta \log \pi_\theta(o|q)] \quad (28)$$

Proof. We start by taking the gradient of the loss function with respect to parameters θ . Note that the reward $r(o|q)$ is independent of θ .

$$\nabla_\theta \mathcal{L}_{\text{eff}} = \nabla_\theta \sum_{o \in \mathcal{O}} r(o|q) \pi_\theta(o|q) \log \pi_\theta(o|q) \quad (29)$$

$$= \sum_{o \in \mathcal{O}} r(o|q) \nabla_\theta (\pi_\theta(o|q) \log \pi_\theta(o|q)) \quad (30)$$

Using the product rule $\nabla(fg) = (\nabla f)g + f(\nabla g)$:

$$\nabla_\theta (\pi_\theta \log \pi_\theta) = (\nabla_\theta \pi_\theta) \log \pi_\theta + \pi_\theta (\nabla_\theta \log \pi_\theta) \quad (31)$$

We apply the log-derivative trick identity: $\nabla_\theta \pi_\theta = \pi_\theta \nabla_\theta \log \pi_\theta$. Substituting this into Eq. (31):

$$\nabla_\theta (\pi_\theta \log \pi_\theta) = (\pi_\theta \nabla_\theta \log \pi_\theta) \log \pi_\theta + \pi_\theta (\nabla_\theta \log \pi_\theta) \quad (32)$$

$$= \pi_\theta (1 + \log \pi_\theta) \nabla_\theta \log \pi_\theta \quad (33)$$

Substituting this back into the summation:

$$\nabla_\theta \mathcal{L}_{\text{eff}} = \sum_{o \in \mathcal{O}} \pi_\theta(o|q) [r(o|q)(1 + \log \pi_\theta(o|q)) \nabla_\theta \log \pi_\theta(o|q)] \quad (34)$$

By definition of expectation $\mathbb{E}_{x \sim p(x)}[f(x)] = \sum p(x)f(x)$, we obtain the final form:

$$\nabla_\theta \mathcal{L}_{\text{eff}} = \mathbb{E}_{o \sim \pi_\theta} [r(o|q)(1 + \log \pi_\theta(o|q)) \nabla_\theta \log \pi_\theta(o|q)] \quad (35)$$

□

B.3 Derivation of Token-Level Decomposition

Here we demonstrate how the sequence-level gradient is decomposed into token-level gradients, justifying the implementation in Eq. (13).

Let $J(o) = r(o|q)(1 + \log \pi_\theta(o|q))$ be the scalar weight for a sampled trajectory o . The gradient estimator for a single sample is:

$$g = J(o) \cdot \nabla_\theta \log \pi_\theta(o|q) \quad (36)$$

For an autoregressive model, the sequence log-probability is the sum of token log-probabilities:

$$\log \pi_\theta(o|q) = \sum_{t=1}^T \log \pi_\theta(x_t | x_{<t}, q) \quad (37)$$

By the **linearity of the gradient operator** ∇_θ , the gradient of a sum is the sum of gradients:

$$\nabla_\theta \log \pi_\theta(o|q) = \nabla_\theta \left(\sum_{t=1}^T \log \pi_\theta(x_t | \dots) \right) = \sum_{t=1}^T \nabla_\theta \log \pi_\theta(x_t | \dots) \quad (38)$$

Substituting this back into the estimator g :

$$g = J(o) \cdot \left(\sum_{t=1}^T \nabla_\theta \log \pi_\theta(x_t | x_{<t}, q) \right) \quad (39)$$

$$= \sum_{t=1}^T [J(o) \cdot \nabla_\theta \log \pi_\theta(x_t | x_{<t}, q)] \quad (40)$$

This proves that applying the sequence-level scalar weight $J(o)$ to the gradient of each token’s log-probability is mathematically equivalent to optimizing the sequence-level Effective Entropy. Note that $J(o)$ must be treated as a constant (detached from the computation graph) during the backward pass of the current step.

C Posterior Evaluation Metric: Entropy Conversion Efficiency

While $\mathcal{H}_{\text{eff}}$ serves as an optimization objective, verifying the efficiency of the learning process requires a posterior metric. In standard RL fine-tuning, we typically observe a pattern where the model gains performance ($\Delta \mathcal{P} > 0$) while becoming more confident, leading to a decrease in entropy ($\Delta \mathcal{H} < 0$).

Design Philosophy. We conceptualize the reduction of entropy as the “cost” or “fuel” consumed to acquire performance. Ideally, we seek a metric that represents the *performance gain per unit of entropy cost*, conceptually similar to $\frac{\Delta \mathcal{P}}{|\Delta \mathcal{H}|}$. However, directly using this ratio introduces significant mathematical issues:

1. **Boundary Instability:** As $\Delta \mathcal{H} \rightarrow 0$ (e.g., in the very early stages or during plateaus), the metric approaches infinity, causing numerical explosions that obscure the trend.
2. **Sign Ambiguity:** A simple ratio fails to naturally distinguish between beneficial convergence and degenerate collapse without complex conditional logic.

To address this, we propose **Entropy Conversion Efficiency (ECE)**, which employs an exponential scaling to ensure continuity and stability while preserving the monotonicity of the return-to-cost ratio. Let $\Delta \mathcal{P}_t = \mathcal{P}_t - \mathcal{P}_{\text{base}}$ and $\Delta \mathcal{H}_t = \mathcal{H}_t - \mathcal{H}_{\text{base}}$. The metric is defined as:

$$\text{ECE}_t = \Delta \mathcal{P}_t \cdot \exp \left(\text{sgn}(\Delta \mathcal{P}_t) \cdot \frac{\Delta \mathcal{H}_t}{\mathcal{H}_{\text{base}}} \right) \quad (41)$$

Primary Analysis Domain ($\Delta \mathcal{P} > 0, \Delta \mathcal{H} < 0$). This is the standard regime for converging models. Here, the term $\frac{\Delta \mathcal{H}}{\mathcal{H}_{\text{base}}}$ is negative.

- The ECE value effectively measures the “exchange rate.” If two models achieve the same performance gain $\Delta \mathcal{P}$, the one that preserves more entropy (smaller $|\Delta \mathcal{H}|$) will have a larger exponential term (closer to 1), resulting in a higher ECE.
- Conversely, a model that collapses (large $|\Delta \mathcal{H}|$) to achieve the same performance is penalized with a smaller exponential factor.

Secondary Domains and Caveats. While ECE is defined globally, interpretations in other quadrants require caution:

- **Exploratory Growth** ($\Delta \mathcal{P} > 0, \Delta \mathcal{H} > 0$): Mathematically, ECE is amplified here ($\exp > 1$), rewarding simultaneous performance and diversity gains. However, this region is susceptible to “hacking,” where a model might artificially inflate entropy (e.g., via irrelevant noise) without genuine reasoning diversity. Therefore, ECE comparisons are most rigorous when $\Delta \mathcal{H} \leq 0$.

Table 3: Classification of entropy-aware algorithms based on their modification of update magnitudes relative to GRPO. The regions are defined by sample type, entropy level, and update action.

Entropy Level	Positive Sample ($r = 1$)		Negative Sample ($r = 0$)	
	Compensate	Penalize	Compensate	Penalize
High Entropy (Low Prob.)	Clip-Higher; Entropy-Reg; Entropy-Adv; GTPO; AEPO; VEPO	Adv-Reweighting	–	Entropy-Reg; Entropy-Adv; Adv-Reweighting; GTPO; Clip-Cov/KL-Cov
Low Entropy (High Prob.)	AEPO	GTPO; 20/80; Clip-Cov/KL-Cov	GTPO	20/80

- **Degeneration** ($\Delta\mathcal{P} < 0$): When performance drops, the $\text{sgn}(\Delta\mathcal{P})$ term inverts the exponent. A collapse in entropy ($\Delta\mathcal{H} < 0$) combined with performance loss results in a heavily penalized (highly negative) score, correctly identifying failure modes.

In our experiments, we focus primarily on the efficiency of converting entropy reduction into performance gain, ensuring that the model does not simply memorize solutions (low cost-efficiency) but learns robustly.

D Qualitative Analysis of Entropy-aware Baselines

By approximating low probability as High Entropy and high probability as Low Entropy, we can categorize the optimization behaviors into an eight-dimensional table (Table 3). We perform a qualitative analysis on the following algorithms, filling them into the table based on their contrast with standard GRPO properties. Note that in this context, "Compensation" refers to boosting the update magnitude, while "Penalty" refers to suppressing it.

- **Clip Higher** [Yu et al., 2025]: Since clipped tokens are predominantly low-probability tokens [Su et al., 2025b], the operation of globally raising the upper bound is equivalent to compensating for the contribution of low-probability (high-entropy) tokens. This belongs to **Region 1** optimization.
- **Entropy Regularization (Entropy-Reg)** [O’Donoghue et al., 2016]: According to its gradient formula, it generally improves high-entropy situations but indiscriminately increases probability. Therefore, it compensates for positive samples and penalizes negative samples (by opposing the probability reduction), belonging to **Region 1 and 4** optimization.
- **Entropy Advantage (Entropy-Adv)** [Cheng et al., 2025]: This method performs compensation on positive samples and suppression on negative samples. Since the degree of interference is positively correlated with entropy, it belongs to **Region 1 and 4** optimization.
- **Advantage Reweighting (Adv-Reweighting)** [Yang et al., 2025b]: Regardless of whether the sample is positive or negative, the smaller the probability (higher the entropy), the larger the penalty on the gradient. Thus, this is **Region 2 and 4** optimization.
- **GTPO** [Tan et al., 2025]: The optimization direction is similar to Entropy-Adv, but due to the reward weighting mechanism, it also affects low-entropy situations. Low entropy results in penalizing positive sample updates and compensating negative sample updates. Therefore, it spans **Region 1, 4, 6, and 7** optimization.
- **AEPO** [Wang et al., 2025a]: This adds a REINFORCE-like regularization term, which can be viewed entirely as treating samples as positive. Thus, it falls into **Region 1 and 5** optimization.
- **20/80 Strategy** [Wang et al., 2025c]: This method stops gradients for low-entropy tokens. This equates to suppressing the update for both positive and negative samples when entropy is low, belonging to **Region 6 and 8** optimization.
- **Clip-Cov & KL-Cov** [Cui et al., 2025]: These methods suppress high-covariance tokens. Based on the analysis, these naturally fall into **Region 4 and 6** optimization.

- **VEPO (Ours):** Our method specifically compensates for high entropy in positive samples, strictly belonging to **Region 1** optimization.

E Implementation Details

E.1 Infrastructure and Environment

Computational Resources. Our experiments were conducted on a Meituan-provided high-performance computing cluster comprising 30 nodes. Each node is equipped with $8 \times$ NVIDIA A100 GPUs (80GB VRAM), interconnected via NVLink for high-bandwidth intra-node communication and InfiniBand for inter-node synchronization. The total compute budget for the experimental sweep, including hyperparameter tuning and baseline reproduction, exceeded 25,000 GPU-hours.

Software Stack. We implemented VEPO within the **VeRL** framework [Sheng et al., 2025], a highly efficient library designed for Post-Training Large Language Models.

- **Training Backend:** We utilized Fully Sharded Data Parallel (FSDP) to manage memory efficiency. For the 7B models, we employed ZeRO-3 offloading to enable full-parameter fine-tuning without requiring parameter-efficient approximations (PEFT).
- **Inference Engine:** To maximize the throughput of the group-based rollout generation, we integrated **vLLM** as the inference backend. We leveraged the PagedAttention mechanism to efficiently manage Key-Value (KV) cache memory, which is critical when generating long chain-of-thought sequences (up to 8192 tokens) to prevent OOM errors during the exploration phase.

E.2 Data and Evaluation Benchmarks

Training Dataset. We utilized the **DAPO-Math-17k** dataset [Yu et al., 2025] for training. To ensure strict alignment between the pre-trained backbone and the RL fine-tuning stage, we used the official tokenizer from the Qwen2.5-Math series. We enforced a maximum sequence length of 8192 tokens, allowing the model sufficient capacity to explore complex, multi-step reasoning paths without artificial truncation.

Reward Function. We implemented a strict dual-oracle verification protocol to prevent reward hacking and ensure metric validity:

- **Training Oracle:** We utilized the lightweight `math_verify` library for efficient online reward calculation. The reward is binary: $r = 1$ if the boxed answer matches the ground truth, and $r = 0$ otherwise.
- **Evaluation Oracle:** For all reported metrics, we employed the more rigorous `prime_math` evaluation suite [Lightman et al., 2023]. This ensures that our reported gains are not artifacts of loose parsing logic.

Benchmark Suite. Table 4 details the characteristics of the evaluation datasets.

E.3 Training Protocols

Optimization Strategy. We employed the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.1. To strictly isolate the algorithmic contribution of VEPO, we adopted a **constant learning rate** strategy across all experiments, avoiding the confounding effects of warm-up or cosine decay schedules.

- **Batch Configuration:** The global batch size is 512 and the mini-batch size is 32.
- **KL Divergence:** We set the KL penalty coefficient $\beta_{KL} = 0$. We rely implicitly on the trust region constraints imposed by the PPO clipping mechanism ($\epsilon = 0.2$) to prevent policy collapse.

Table 4: Detailed profile of the reasoning benchmarks. Datasets are grouped by their strategic role in evaluating the model’s latent reasoning capabilities.

Category	Dataset	Strategic Evaluation Focus
<i>Competition-Level Reasoning (High-Stakes Deduction)</i>		
	AIME 2024/2025	Targets the robustness of long-range reasoning chains; the 2025 version serves as a temporal hold-out to reduce the risk of data leakage.
	AMC 2023	Benchmarks the model against standard competitive mathematics at the entry olympiad level.
<i>Broad-Spectrum Knowledge (Domain Versatility)</i>		
	MATH-500	Provides a curated cross-section of scholastic math to evaluate whether the policy maintains balanced performance across major mathematical sub-fields.
	Minerva	Focuses on the transition from standard mathematical problems to complex, scientific-grade symbolic reasoning.
<i>Stress Testing (Performance Ceiling)</i>		
	OlympiadBench	Challenges the model with rigorous olympiad-level problems to probe the limits of validity-aware exploration.

E.4 Hyperparameter Specifications

We performed a grid search for the Effective Entropy coefficient β on a held-out validation set. Table 5 lists the optimal configurations used for the main results, along with the specific settings for all baseline algorithms.

Table 5: Detailed hyperparameter configurations for VEPO and all baselines. For baselines, we utilized the values recommended in their respective original papers or tuned them for optimal performance on our validation set.

Model Scale	Algorithm	Learning Rate	Key Hyperparameters
<i>Common Settings: Global Batch Size = 512, Mini-batch Size = 32, Max Length = 8192, $\beta_{KL} = 0$</i>			
1.5B	GRPO	1.0×10^{-6}	$\varepsilon = 0.2$
	Entropy-Reg		$\beta_{\text{reg}} = 0.001$
	Entropy-Adv		$\alpha_E = 0.4, \kappa = 2.0$
	Adv-Reweighting		$\alpha_A = 0.1$
	GTPO		$\alpha_1 = 1.0, \alpha_2 = 0.1$
	AEPO		$\mathcal{H}_{\text{target}} = 0.75, T_{\text{low}} = 0.8, T_{\text{high}} = 1.2, \alpha = 0.1$
	VEPO		$\beta = 0.0001$
7B	GRPO	4.63×10^{-7}	$\varepsilon = 0.2$
	Entropy-Reg		$\beta_{\text{reg}} = 0.001$
	Entropy-Adv		$\alpha_E = 0.4, \kappa = 2.0$
	Adv-Reweighting		$\alpha_A = 0.1$
	GTPO		$\alpha_1 = 1.0, \alpha_2 = 0.1$
	AEPO		$\mathcal{H}_{\text{target}} = 0.75, T_{\text{low}} = 0.8, T_{\text{high}} = 1.2, \alpha = 0.1$
	VEPO		$\beta = 0.0001$

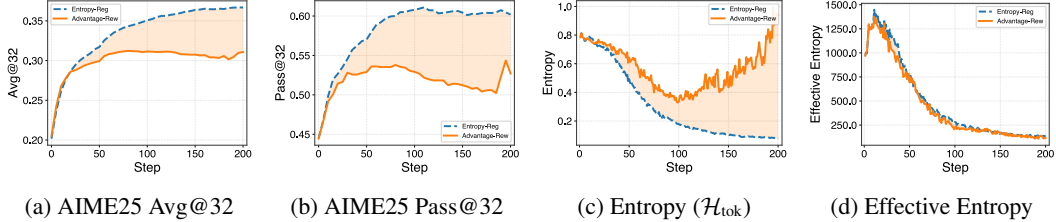


Figure 6: Comparative dynamics between Entropy-Reg and Adv-Reweighting on 7B model. While Adv-Reweighting maintains or even surges in total entropy, its effective entropy remains consistently below Entropy-Reg, resulting in inferior reasoning performance.

F Additional Experimental Results

F.1 The Paradox of Unaligned Diversity

To further illustrate the necessity of aligning exploration with validity, we provide a comparative analysis of the training dynamics between **Entropy-Reg** and **Adv-Reweighting**. As shown in Figure 6, these two baselines exhibit a phenomenon we term the *"Diversity Paradox"*.

High Entropy Without Validity. In the late stages of training (beyond step 150), Adv-Reweighting exhibits a sharp surge in token-level entropy ($\mathcal{H}_{\text{tok}}$). Under traditional metrics, this would be interpreted as a more "diverse" or "exploratory" model. However, this diversity fails to translate into performance; in fact, its Avg@32 and Pass@32 metrics plateau or even decline during this period.

Mechanism Failure. The root cause lies in the *Effective Entropy*. Despite having higher total uncertainty, Adv-Reweighting’s effective entropy is lower than that of the simpler Entropy-Reg. This indicates that the additional entropy in Adv-Reweighting is being injected into the **invalid subspace**. Without VEPO’s validity-aware attraction mechanism, the model generates diverse hallucinations (high total entropy) rather than diverse correct solutions. This confirms that unaligned entropy regularization can be detrimental, acting as a source of chaotic noise that distracts the model from the manifold of correct reasoning.

F.2 Training Efficiency and Computational Overhead

Matched training diagnostics show a 9.8% steady-state overhead for VEPO relative to GRPO. VEPO reaches the paired GRPO run’s final six-benchmark aggregate Avg@32 after 73% of its rollout budget at 1.5B and 78% at 7B. Including the measured overhead gives approximately 80% and 86% of GRPO’s wall-clock time to that target. These are time-to-target measurements for the tested runs, not evidence of zero computational cost.

With $G = 8$, the fractions of rollouts carrying a nonzero auxiliary signal in the early, middle, and late thirds of training are 28.6%, 39.7%, and 45.1% at 1.5B, and 42.1%, 51.5%, and 56.6% at 7B. The corresponding fractions of groups with no auxiliary signal are 7.1%, 2.0%, and 0.9%, versus 1.3%, 0.4%, and 0.1%. The median auxiliary-to-GRPO gradient-norm ratios, including the tuned β , are 0.18 and 0.15. Incorrect trajectories in mixed-reward groups retain the ordinary GRPO update; they receive no additional effective-entropy signal. Terminal verification nevertheless cannot assign credit to useful intermediate steps in an outcome-wrong trajectory.

G Selected Additional Evidence

G.1 Strategy-Diversity Robustness

Matching Correct Samples per Problem. Pass@ k measures repeated-sampling success, which also depends on per-sample accuracy. We therefore measure strategy diversity directly on MATH500 with the 7B backbone. We retain the 441 of 500 problems on which both GRPO and VEPO produce at least eight answer-verified correct trajectories in their fixed pools of 32 samples. On each problem,

we draw exactly eight correct trajectories per method without replacement and average over 10,000 rarefactions with fixed random seeds. This controls the number of correct samples separately for every problem, rather than only matching dataset-level accuracy.

We strip final-answer spans, pool the remaining chains of thought from both methods, and cluster within each problem using frozen gte-Qwen2-7B-instruct embeddings [Alibaba NLP, 2024], L2 normalization, and average-linkage agglomerative clustering with cosine distance [Müllner, 2011]. A strategy denotes a shared mathematical plan, invariant, decomposition, or substitution, irrespective of wording or notation. Before comparing methods, we fix the pipeline and select one global distance threshold on method-blinded development annotations. Two mathematical annotators label 600 same/different-strategy pairs with Cohen’s $\kappa = 0.82$; the frozen pipeline achieves precision 0.88 and recall 0.84 on a disjoint held-out set.

Two internal university faculty members provided unpaid, method-blinded annotations of reasoning-strategy equivalence.

The protocol in Section 6.3 fixes its threshold by development-set pairwise F1 before comparing methods. For cluster proportions s_j , Hill-1 modes are $\exp(-\sum_j s_j \log s_j)$; the largest-cluster share is $\max_j s_j$. Table 6 retains the matched count of eight trajectories per method per problem. Length matching is one-to-one within a problem without replacement.

Table 6: Robustness of the paired VEPO-minus-GRPO cluster-count difference on MATH500-7B. The matched correct count is eight trajectories per method per problem.

Protocol	Difference	95% confidence interval
Primary gte-Qwen2 encoder	+0.37	[0.24, 0.50]
Prespecified e5-Mistral encoder [Wang et al., 2024]	+0.31	[0.18, 0.44]
Within-problem CoT-length matching	+0.34	[0.20, 0.47]

A separate blinded rationale audit finds 92.5% of GRPO and 93.0% of VEPO correct-answer trajectories logically sound: difference +0.5 percentage points, 95% CI $[-3.7, 4.7]$. Answer verification therefore does not certify every intermediate step. The diversity endpoint also conditions on the 441 commonly eligible problems, rather than all MATH500 problems.

G.2 Training-Seed Variability

GRPO and VEPO were trained with the same three paired random seeds, 0, 1, and 2.

Table 7: VEPO minus GRPO across three paired training seeds, in percentage points, averaged over the six mathematical benchmarks.

Backbone	Avg@32 gains	Pass@32 gains
1.5B	+2.1, +2.4, +2.1	+1.3, +1.6, +1.6
7B	+2.3, +2.4, +1.9	+1.3, +1.4, +1.2

Aggregate Avg@32 is 48.4 ± 0.46 versus 50.6 ± 0.60 at 1.5B and 58.9 ± 0.44 versus 61.1 ± 0.40 at 7B (GRPO versus VEPO; $\pm$ denotes seed SD). These three training replicates are distinct from the problem-bootstrap uncertainty in the diversity evaluation. Against AEPO, the 7B aggregate Pass@32 difference is +0.4 percentage points with reported 95% interval $[-0.3, 1.1]$, supporting competitiveness rather than significant superiority.

G.3 Sequence-Weight Controls

The constant-gated control preserves the correctness mask, attraction direction, raw sequence sum, and held-out tuning budget while removing the $\log \pi$ dependence. Its three paired Avg@32 deficits relative to VEPO are 1.2, 1.5, 1.4 percentage points at 1.5B and 1.1, 1.3, 1.2 at 7B. Response-length normalization instead divides each trajectory’s contribution to $\mathcal{H}_{\text{eff}}$ by $T(o)$; it defines a distinct functional and supplies no causal token credit.

Table 8: Sequence-weight controls. Values are unweighted Avg@32 / Pass@32 aggregates over the six mathematical benchmarks in the original-run setting. Length normalization is an additional variant, not the objective used in the main experiments.

Variant	1.5B	7B
Constant-gated sequence sum	49.2 / 70.8	59.9 / 77.7
VEPO, raw effective entropy	50.6 / 71.6	61.1 / 78.5
Response-length-normalized VEPO	–	61.2 / 78.6

G.4 Controlled Verifier Noise

For constant class-conditional false-negative rate α and false-positive rate γ ,

$$\mathbb{E}[\tilde{\mathcal{H}}_{\text{eff}}] = \gamma \mathcal{H}_{\text{seq}} + (1 - \alpha - \gamma) \mathcal{H}_{\text{eff}}. \quad (42)$$

False negatives suppress valid-space exploration, while false positives reintroduce an entropy incentive on incorrect outputs. This simplification does not apply to arbitrary trajectory-dependent verifier errors. Table 9 shows retained VEPO gains within the tested noise range and degradation of both methods as noise increases.

Table 9: Controlled verifier-noise retraining on the 1.5B backbone. Values are three-seed means of six-benchmark Avg@32 / Pass@32. FN and FP denote injected false-negative and false-positive noise.

Verifier setting	GRPO	VEPO
Clean	48.4 / 70.1	50.6 / 71.6
FN 5%	48.2 / 69.8	50.3 / 71.2
FN 10%	47.8 / 69.3	49.9 / 70.8
FP 5%	47.6 / 68.9	49.6 / 70.4
FP 10%	46.8 / 67.7	48.8 / 69.6

Non-binary Rewards. Replacing the binary gate by a score $v(o) \in [0, 1]$ gives reward-weighted entropy, $-\sum_o v(o) \pi(o) \log \pi(o)$, rather than entropy restricted to a binary valid subspace. Partial correctness and process-level verification remain extensions.

G.5 Valid Mass and Conditional Diversity

Effective entropy is an unnormalized partial entropy, not the entropy of the conditional distribution over valid outputs. Writing $Z = \sum_{r(o)=1} \pi_\theta(o | q)$ and $p_{\text{valid}}(o) = \pi_\theta(o | q) / Z$ for $Z > 0$ gives

$$\mathcal{H}_{\text{eff}} = Z \mathcal{H}(p_{\text{valid}}) - Z \log Z. \quad (43)$$

It therefore depends on both valid probability mass and diversity within that mass. Section 6.3 evaluates strategy diversity separately with matched correct counts. Here validity means terminal-verifier acceptance, which need not certify every intermediate reasoning step.

G.6 Detached-Weight Loss Pseudocode

The following pseudocode implements the auxiliary gradient in Eq. 13 on a batch of sampled responses. The mask excludes prompts and padding, and includes the response tokens used in the sequence probability. The raw sequence sum is retained; taking a token mean here would change the stated objective. The loss has the minimization sign, and detaching the weight avoids differentiating it twice.

```
# logp: current-policy response token log probabilities [B, T]
# mask: response-token mask; reward: binary terminal reward [B]
sequence_logp = (logp * mask).sum(dim=-1)
weight = (reward * (1.0 + sequence_logp)).detach()
auxiliary_loss = (weight * sequence_logp).mean()
loss = grpo_loss + beta * auxiliary_loss
loss.backward()
```

This is a computational surrogate whose gradient matches the derived on-policy estimator at the sampled policy. Reusing a batch from an old policy does not by itself make it an unbiased estimator under the updated policy. Exact current-policy expectation would require fresh samples or appropriate sequence-level importance weighting; the original VEPO update is kept as the simple local surrogate.

H Limitations

This work has several limitations. First, effective entropy relies on a reliable validity signal, and is therefore most directly applicable to RLVR settings where correctness can be verified by exact answers, unit tests, program execution, or other deterministic reward functions. Extending the same idea to open-ended tasks with ambiguous or preference-based rewards requires a more careful definition of the valid subspace. Second, VEPO is intentionally designed as a simple instantiation of effective entropy rather than a highly engineered optimizer. Although this simplicity helps isolate the contribution of the concept, more adaptive variants may further improve stability and efficiency. Third, our experiments focus on mathematical reasoning and models up to the tested scales. Whether the same gains hold for larger models, multimodal reasoning, tool-use agents, or long-horizon interactive tasks remains to be further validated. Finally, effective entropy is estimated from sampled rollouts, so its quality depends on sampling coverage: when valid trajectories are extremely rare in early training, additional warm-up, curriculum design, or stronger initial policies may be needed.

The direct strategy-diversity result conditions on commonly solved prompts and on a finite sample budget; embedding clusters are an operational proxy for strategies, not a proof of complete solution-space coverage. Answer acceptance also does not ensure every intermediate step is logically sound. False-positive verifier errors can encourage incorrect outputs, and sequence-level weights do not solve causal token-level credit assignment. The three-seed evidence supports consistency in the tested configuration, with limited precision for seed-level inference.

I Broader Impacts

Validity-aware exploration may improve the reliability of mathematical reasoning systems by favoring answer-verified outputs. A terminal verifier can still accept a correct answer accompanied by an unsound rationale, so downstream educational or scientific uses require independent checks. Poorly specified or malicious verifiers could reward harmful behavior; improving task correctness does not establish general safety. The method also requires repeated rollouts and therefore incurs substantial compute cost. We use existing mathematical datasets and do not collect private user data.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The claims are scoped to mathematical RLVR on the tested 1.5B and 7B backbones. Aggregate accuracy and direct matched-count strategy diversity are reported in Sections 6.2 and 6.3; the complete paired-seed results appear in Appendix G.2.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses limitations in Appendix H. It explicitly notes the dependence on reliable validity signals, the intentionally simple design of VEPO, the current experimental focus on mathematical reasoning and tested model scales, and the sampling-coverage issue when valid trajectories are rare in early training.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the binary-verifier setting and derives the on-policy entropy gradient. Appendix B.1 restricts the finite-simplex optimum to the stated coefficient condition; the local reused-rollout surrogate is distinguished from an exact current-policy estimator.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The paper provides the objective, pseudocode, models, training data, key hyperparameters, and paired training seeds (0, 1, 2). The implementation and executable reproduction scripts are not publicly released.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The existing datasets, models, and encoder are cited. The authors are not publicly releasing the VEPO implementation or reproduction scripts with this version.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [No]

Justification: Section 6.1 and Appendix E describe the models, datasets, optimizer, group size, clipping, and coefficient selection. Appendix G.2 reports paired training seeds 0, 1, 2; Appendix G.1 describes the strategy-diversity evaluation and threshold-selection procedure.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Section 6.3 reports a problem-bootstrap confidence interval and paired sign-flip permutation test for the primary diversity endpoint. Table 7 reports all three paired seed deltas; Appendix G.2 distinguishes seed standard deviations from evaluation uncertainty.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix E describes the Meituan-provided A100 hardware and total experimental compute budget. Appendix F.2 reports matched steady-state overhead and rollout and wall-clock time to a fixed accuracy target.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research uses existing mathematical datasets, cited model backbones, automatic verifiers, and blinded mathematical annotation. It does not collect private user data. Expert annotation was performed by internal university faculty without monetary compensation.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix I discusses reliability benefits, unsound rationales accepted by terminal verifiers, verifier misuse, and computational cost.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new pretrained language model, scraped dataset, image generator, or other high-risk asset. The work proposes an optimization method and evaluates it on mathematical reasoning benchmarks using existing models and datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper properly cites the existing models, datasets, frameworks, and verification tools used in the experiments. However, the current version does not explicitly list the license or terms of use for each asset.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release a new dataset, benchmark, or pre-trained model as a new asset. The main contribution is a training objective and optimization method built on existing models and datasets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: Appendix G.1 summarizes the method-blinded strategy-equivalence annotation and reports validation results. Two internal university faculty members performed the annotation without monetary compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The diversity evaluation uses internal university faculty as expert annotators. This version does not document an institutional determination that this expert annotation is exempt from or outside the scope of human-subjects review, or any applicable approval. No exemption or approval is claimed.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are central to the research, as the paper studies RLVR post-training of DeepSeek-R1-Distill-Qwen backbones and evaluates their mathematical reasoning behavior. The use of LLMs as the experimental models is described in Section 6.1 and Appendix E.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.