Distributional Clarity: The Hidden Driver of RL-Friendliness in Large Language Models

Anonymous ACL submission

Abstract

Language model families exhibit striking disparity in their capacity to benefit from reinforcement learning: under identical training, models like Qwen achieve substantial gains, while others like Llama yield limited improvements. Complementing data-centric approaches, we reveal that this disparity reflects a hidden structural property: **distributional clarity** in probability space. Through a three-stage analysis—from phenomenon to mechanism to interpretation—we uncover that RL-friendly models exhibit intra-class compactness and inter-class separation in their probability assignments to correct vs. incorrect responses. We quantify this clarity using the **Silhouette Coefficient** (S) and demonstrate that (1) high S correlates strongly with RL performance; (2) low S is associated with severe logic errors and reasoning instability. To confirm this property, we introduce a Silhouette-Aware Reweighting strategy that prioritizes low-S samples during training. Experiments across six mathematical benchmarks show consistent improvements across all model families, with gains up to 5.9 points on AIME24. Our work establishes distributional clarity as a fundamental, trainable property underlying RL-Friendliness.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become the dominant approach for enhancing LLM reasoning (Guo et al., 2025; Hu et al., 2025; Yu et al., 2025b; Wang et al., 2025b; Zhu et al., 2025). However, beneath this success lies a puzzling asymmetry: when trained with identical RLVR pipelines such as GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025a), Qwen models (Team, 2024) consistently achieve substantial gains in mathematical reasoning, while Llama (Dubey et al., 2024) yields only limited improvements (Liu et al., 2025c; Gandhi et al., 2025; Zeng et al., 2025). This disparity reflects differences in *RL-Friendliness*—the capacity of a model to benefit from reinforcement learning.

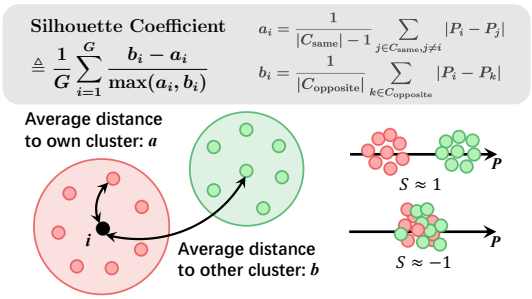


Figure 1: Schematic illustration of the Silhouette Coefficient (S). We adapt this metric to quantify distributional clarity. High S values represent ideal landscapes with compact and separated clusters, while low values indicate overlapping distributions.

Why do foundation models differ in their capacity to benefit from RL training? Prior work has approached this question from a data-centric perspective. Gandhi et al. (2025) identified specific reasoning patterns as critical differentiators, demonstrating that fine-tuning with data containing these patterns can partially bridge the gap during subsequent RL training. Similarly, OctoThinker (Wang et al., 2025c) showed that exposing Llama to high-quality mathematical corpora during mid-training better prepares it for reinforcement learning. In parallel to these data-centric approaches, we examine an orthogonal aspect: the intrinsic generation properties of different model families, aiming to uncover the key drivers of RL-Friendliness.

We introduce a three-stage analysis framework that systematically dissects RL-Friendliness *from phenomenon to mechanism to interpretation*. At the phenomenological level, we find that while RL-friendly and less RL-friendly models solve largely overlapping problem sets, RL-friendly models achieve consistently higher pass rates on these shared problems, implying superior probability assignment to correct solutions. At the mechanistic stage, we identify a key distributional differ-

ence on probability: RL-friendly models exhibit **intra-class compactness and inter-class separation**-probability scores for correct and incorrect responses cluster densely within their groups yet remain significantly separated. We introduce the **Silhouette Coefficient (S)** (Murphy, 2022) to quantify this distributional clarity and observe a strong positive correlation between S and pass rates.

At the interpretative level, we show that distributional clarity governs reasoning quality at the semantic level by examining error attribution and solution stability. RL-friendly models produce predominantly low-severity errors (calculation issues) and exhibit consistent reasoning paths, whereas less RL-friendly models generate high-severity errors (logic flaws) and show unstable reasoning. Crucially, samples with low S are disproportionately associated with severe errors and instability, while high-S samples show stable reasoning and minor errors. This reveals that distributional clarity constitutes a key factor in effective RL optimization.

To validate the critical role of distributional clarity in RL-Friendliness, we introduce a *Silhouette-Aware Reweighting* strategy that prioritizes low-S samples during training. By targeting samples with poor distributional clarity, we force the model to improve its worst-performing areas while maintaining its strengths. Experiments on six mathematical reasoning benchmarks demonstrate consistent improvements for all model families, particularly on challenging datasets like AIME24, where gains range from 1.8 to 5.9 points across model families. These gains across diverse models validate our analysis and demonstrate that addressing distributional clarity is key to enhance RL-Friendliness.

In summary, our contributions are as follows:

1. **A three-stage diagnostic framework.** We demonstrate that RL-Friendliness reflects not merely *what* models can solve, but *how reliably* they solve it. Through systematic analysis from phenomenon to mechanism to interpretation, we identify distributional clarity—specifically intra-class compactness and inter-class separation in the probability assignments to correct vs. incorrect responses—as the fundamental structural property governing RL-Friendliness.
2. **A mechanistic understanding.** We introduce the Silhouette Coefficient (S) to quantify distributional clarity and show that it serves as a unified metric bridging structure and behavior: it simultaneously reflects performance, correlates with error severity, and tracks reasoning stability. This establishes *how* distributional clarity governs RL-Friendliness: high S ensures stable training dynamics aligned with optimization goals, while low S manifests as severe logical errors and unstable reasoning—impeding the reinforcement of reliable behaviors.
3. **A practical intervention.** We introduce a Silhouette-Aware Reweighting strategy to validate our analysis, which prioritizes low-S samples during training. Experiments across six benchmarks yield consistent gains for all model families, confirming that distributional clarity is not merely explanatory but trainable—transforming structural diagnosis into practical enhancement.

2 Anatomy of RL-Friendliness: A Three-Stage Analysis

In this section, we conduct a systematic analysis to uncover the underlying factors contributing to the disparity in RL-friendliness across different model families. We apply our three-stage framework, progressing from phenomenon to mechanism to interpretation. At the *phenomenological level*, we examine performance to quantify outcome disparities on shared problem sets. At the *mechanistic level*, we analyze probability distributions to identify the structural characteristics that drive these performance differences. At the *interpretative level*, we assess reasoning behaviors to provide semantic interpretations of these distributional patterns in terms of error severity and solution stability.

2.1 Phenomenon: The Performance Disparity

To investigate RL-Friendliness disparities, we begin by quantifying model performance through *pass rates*—the empirical probability of generating correct responses. For each query q, we sample K independent responses and compute the pass rate $\rho(q)$—the fraction verified as correct. Higher $\rho(q)$ indicates greater probability mass assigned to correct reasoning paths.

We conducted a comparative analysis using Qwen2.5-7B (Team, 2024) and OctoThinker-8B-Hybrid-Base (Wang et al., 2025c), both trained using the DAPO algorithm under identical settings. Figure 2 visualizes the per-query pass rates on AIME 2024 ($K = 256$). We focus on the intersection of solvable problems—queries where both models achieve $\rho > 0$.

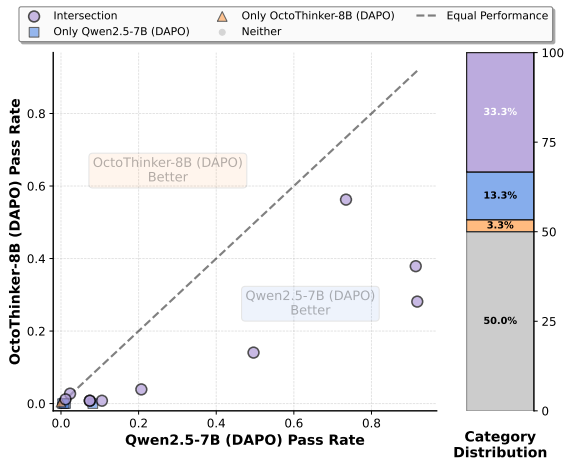


Figure 2: Per-problem pass rate comparison between Qwen2.5-7B (DAPO) and OctoThinker-8B (DAPO) on AIME 2024. Each point represents a query. Points below the diagonal indicate Qwen achieves higher pass rates. A similar distribution pattern is observed on MATH-500 (see Figure 9).

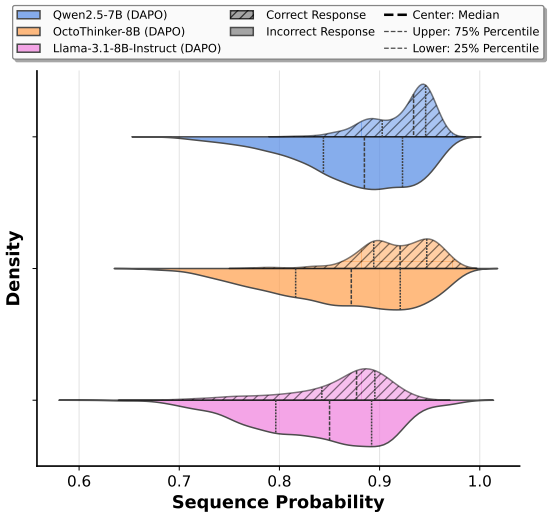


Figure 3: Probability Distributions: Kernel density estimates of sequence probabilities for correct (top) and incorrect (bottom) responses. Qwen exhibits clear separation, whereas Llama and OctoThinker show significant overlap.

Figure 2 shows substantial overlap in the fundamental capabilities of the two models. Specifically, the intersection set accounts for approximately 71.4% of the total solvable problems for Qwen and 90.9% for OctoThinker. However, within this intersection, the behavior diverges markedly. Most data points fall below the diagonal ($y = x$), indicating that Qwen consistently achieves higher pass rates than OctoThinker on the same problems.

This asymmetry reveals a critical insight: **RL-Friendliness is not merely about what a model can solve, but also how reliably it solves it.** Less RL-friendly models fail to assign sufficient probability mass to the correct solutions, while RL-friendly models maintain high reliability in generating correct responses for shared solvable instances.

2.2 Mechanism: Compactness and Separation

The performance disparity observed in Section 2.1, we hypothesize, arises from fundamental differences in how models distribute confidence over correct and incorrect responses. To investigate the mechanism driving this behavior, we quantify response confidence as the geometric mean of token-level probabilities, computing a length-normalized sequence score $P(o|q)$. Given a query q and a generated response $o = (y_1, y_2, \ldots, y_L)$, we calculate:

$$P(o|q) = \left(\prod_{i=1}^{L} P(y_i|q, y_{<i}) \right)^{\frac{1}{L}} \quad (1)$$

Figure 3 visualizes the distribution of sequence-level probabilities for correct versus incorrect responses for Qwen2.5-7B, Llama-3.1-8B, and OctoThinker-8B (all DAPO-trained) on AIME 2024. These distributions reveal striking structural differences across models. Most notably, Qwen's distribution exhibits two distinct characteristics:

- **intra-class compactness**—probability scores cluster tightly within correct and incorrect groups
- **inter-class separation**—a substantial margin exists between the correct and incorrect clusters

In contrast, Llama and OctoThinker display significant overlap: probabilities of incorrect responses frequently match or exceed those of correct ones, creating ambiguous decision boundaries.

To rigorously quantify this structural property, we introduce the **Silhouette Coefficient** (S), adapted from cluster analysis to our one-dimensional probability distributions. For a given query, let $\mathcal{O}$ denote the set of all generated response probabilities, which is partitioned into two clusters: correct responses and incorrect responses. For each response score $P_i \in \mathcal{O}$, let C_{same} be the cluster containing P_i and C_{opposite} be the complementary cluster. We compute the average intra-cluster distance a_i and the average inter-cluster distance b_i as follows:

$$a_i = \frac{1}{|C_{\text{same}}| - 1} \sum_{P_j \in C_{\text{same}}, j \neq i} |P_i - P_j| \quad (2)$$

$$b_i = \frac{1}{|C_{\text{opposite}}|} \sum_{P_k \in C_{\text{opposite}}} |P_i - P_k| \quad (3)$$

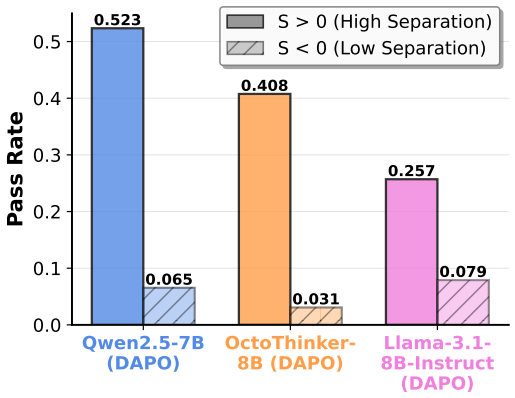


Figure 4: Impact of distributional structure on pass rates. Queries with positive S values achieve significantly higher performance across all models.

Based on these distances, the Silhouette value s_i for the individual sample is defined as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{4}$$

The final query-level coefficient $S \in [-1, 1]$ is obtained by averaging s_i across all samples (i.e., $S = \frac{1}{|\mathcal{O}|} \sum_{P_i \in \mathcal{O}} s_i$). As illustrated in Figure 1, this metric captures total distributional quality: values approaching 1 indicate ideal structure—compact clusters with clear separation—while negative values reveal overlap between correct and incorrect groups.

These two properties—compactness and separation—govern RL-Friendliness through different mechanisms. We provide theoretical grounding in Appendix B, showing that gradient variance in GRPO-style algorithms is predominantly determined by the probability distributions of correct and incorrect responses. When these distributions exhibit high compactness, the gradient signal remains stable, ensuring consistent parameter updates. Moreover, clear separation between correct and incorrect clusters directly aligns with the RL training objective, which is to increase the probability of correct responses while decreasing that of incorrect ones. **High S thus indicates both stable training dynamics and alignment with the optimization goal.**

Empirically, **this distributional clarify positively correlates with performance**. Figure 4 compares pass rates across queries partitioned by S into high-clarity ($S > 0$) and low-clarity ($S < 0$) groups.[1] The gaps are substantial: for Qwen, high-S queries achieve 52.3% pass rates versus 6.5% for low-S queries—an 8× difference. OctoThinker exhibits an even more dramatic pattern: 40.8% versus 3.1%, a 13× gap. This validates S as a reliable indicator of model performance and RL training potential.

2.3 Interpretation: Error Severity and Solution Stability

To understand *what* distributional clarity corresponds to semantically, we conduct a fine-grained behavioral analysis on MATH-500 along two dimensions: (i) the nature of generated errors and (ii) the stability of the reasoning strategies employed.

Error Severity Attribution. We first investigate whether the incorrect responses from different models stem from similar types of failures. Using an LLM-as-a-judge approach[2], we construct a hierarchical error taxonomy and classify failures into three severity levels: *High Severity (Fundamental)* involving core logic or knowledge flaws; *Mid Severity (Execution)* involving calculation errors; and *Low Severity (Presentation)* involving formatting issues. Figure 5a shows that Qwen exhibits a significantly healthier error profile, with a lower proportion of high-severity errors compared to OctoThinker and Llama.

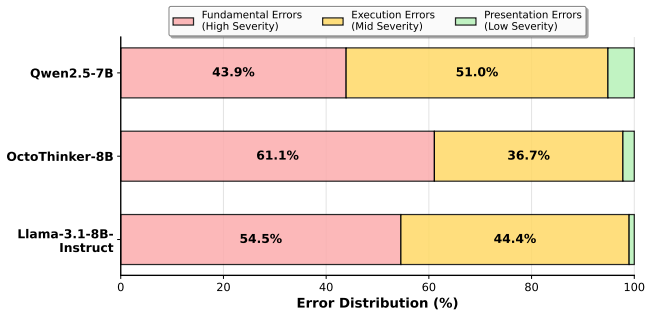
Crucially, we find a clear association: **high-severity errors are disproportionately concentrated in responses with poor distributional clarity**. Figure 5b reports, for each error category, the fraction of responses with negative Silhouette Coefficients ($S < 0$). For Llama, 62.2% of fundamental errors have $S < 0$, indicating weak separation between correct and incorrect responses when the model makes logic-breaking mistakes. In contrast, low-severity errors are more likely to retain $S > 0$, suggesting that the probability distribution correctly identifies promising approaches.

Solution Stability Analysis. Beyond error analysis, we measure the stability of correct reasoning by computing the proportion of distinct solution methods among all correct responses for each query. Two solutions are deemed to share the same method if they utilize identical key theorems, formulas, or logical strategies, ignoring variations in variable naming or verbosity. We cluster correct responses using an automated judge. Clustering details and prompts are in Appendix A.3.

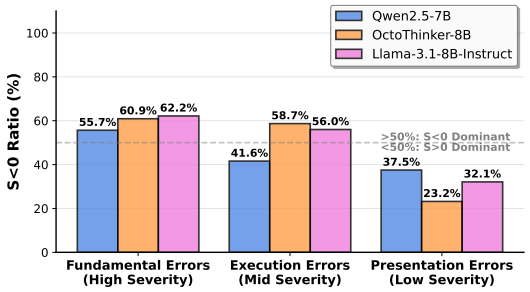
Distinct solutions ratio differs systematically between RL-friendly and less RL-friendly models.

[1] We exclude queries where all responses are correct or all incorrect, as S is undefined in these cases.

[2] Detailed error taxonomy definitions and evaluation prompts are provided in Appendix A.2.

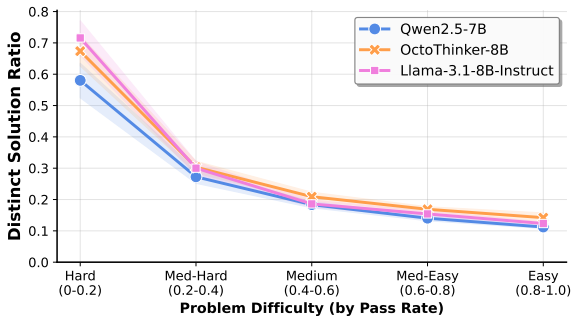


(a) Distribution of Error Severity

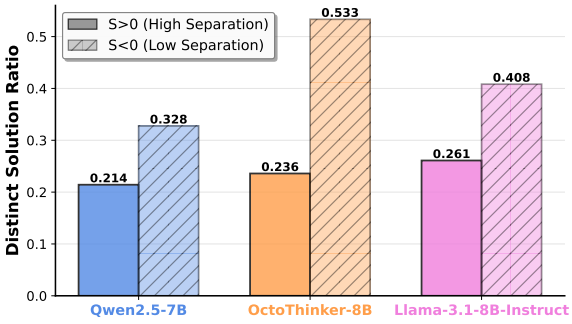


(b) Error Severity and Silhouette Coefficient Relationship

Figure 5: Error attribution analysis on MATH-500. (a) Proportion of High (Fundamental), Mid (Execution), and Low (Presentation) severity errors across models. (b) Percentage of responses with $S < 0$ (poor distributional clarity) within each error category. High-severity errors are strongly correlated with low distributional clarity.



(a) Distinct Solution Ratio vs. Problem Difficulty



(b) Distinct Solution Ratio by Distributional Clarity

Figure 6: Solution stability analysis. (a) Average distinct solution ratio across problem difficulty tiers. Lower ratio implies greater consistency. (b) Solution ratio for queries grouped by Silhouette Coefficient. High clarity ($S \geq 0$) correlates with more stable, consistent reasoning patterns.

As Figure 6a shows, Qwen exhibits consistently low ratio, particularly on hard problems, indicating stable convergence to effective strategies. In contrast, OctoThinker and Llama show significantly more distinct solutions. Closer inspection reveals that high ratio in less RL-friendly models reflects reasoning instability rather than genuine methodological breadth: these models frequently reach correct answers through spurious correctness and hallucinated steps that register as distinct methods, inflating the distinct ratio metric.

We demonstrate that this instability is directly tied to distributional clarity. Figure 6b compares distinct solution ratio for queries with $S \geq 0$ versus $S < 0$. Across all models, high-clarity queries exhibit significantly lower ratio, meaning the model consistently reproduces the similar valid reasoning logic. Conversely, low-clarity queries produce fragmented reasoning paths. This establishes the causal chain: **poor distributional clarity** ($S < 0$) **generates reasoning instability (high ratio), which fundamentally undermines RL training by preventing reliable identification of behaviors to reinforce**. Conversely, high S enables stable reasoning policies that RL can effectively reinforce.

3 Empirical Validation via Silhouette-Aware Reweighting

To validate that distributional clarity drives RL-Friendliness, we propose Silhouette-Aware Reweighting—a strategy that modulates the training signal to prioritize queries with poor distributional clarity. We adopt DAPO (Yu et al., 2025a) as our training backbone.

Standard Formulation. Formally, given a query q and a group of G outputs $\{o_i\}_{i=1}^{G}$ with rewards R_i, the standard advantage $\hat{A}_{i,t}$ in DAPO (and GRPO) is computed via group normalization:

$$\hat{A}_{i,t} = \frac{R_i - \mu(\{R_j\}_{j=1}^{G})}{\sigma(\{R_j\}_{j=1}^{G})} \tag{5}$$

where μ and σ denote the mean and standard deviation of the group rewards.

Silhouette-Aware Reweighting. To explicitly force the model to focus on samples exhibiting distributional ambiguity, we introduce a weighted advantage $\tilde{A}_{i,t}$ defined as:

$$\tilde{A}_{i,t} = \hat{A}_{i,t} \cdot w(q) \tag{6}$$

where $w(q)$ is a query-specific weight derived from the distributional properties of the model outputs.

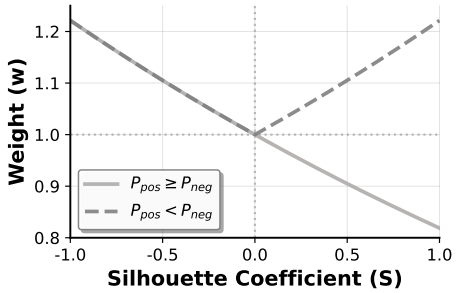


Figure 7: Silhouette-Aware Reweighting function $w(S) = \exp(-\beta \cdot S)$ with $\beta = 0.2$.

To handle cases where the model assigns higher probabilities to incorrect responses, we introduce a rectified Silhouette metric. Let P_{pos} and P_{neg} denote the average sequence-level probabilities for correct and incorrect response groups. When $P_{pos} < P_{neg}$, simply optimizing for high S values might inadvertently push the distributions further apart in the wrong direction. To mitigate this, we define:

$$S' = \begin{cases} S & \text{if } P_{pos} \geq P_{neg} \\ -|S| & \text{if } P_{pos} < P_{neg} \end{cases} \tag{7}$$

This rectified metric becomes strictly negative when the distribution is inverted, ensuring the reweighting emphasizes reversing the probability mass back to the correct direction.

Based on this rectified metric, we define the reweighting factor $w(q)$ as follows:

$$w(q) = \exp(-\beta \cdot S') \tag{8}$$

where $\beta > 0$ controls the sensitivity of the reweighting. Figure 7 illustrates this function assigns larger weights ($w > 1$) to queries with negative S', amplifying the gradient signal for samples with poor or inverted distributional clarity. Conversely, for queries with high positive S' where the distribution is already well-structured, the weight decays ($w < 1$), preventing the model from overfitting to easy samples.

4 Experiments

4.1 Experimental Setup

Models and Datasets. We validate our approach on three backbone models: Qwen2.5-7B (Team, 2024), OctoThinker-8B-Hybrid-Base (Wang et al., 2025c), and Llama-3.1-8B-Instruct (Dubey et al., 2024). OctoThinker-8B-Hybrid-Base is a model derived from Llama-3.1-8B via mid-training on reasoning corpora. For the Llama family, we select the Instruct version because the Base model exhibits insufficient instruction-following capabilities for direct RLVR. Regarding training data, we utilize the DAPO-Math-17k dataset (Yu et al., 2025a) for both Qwen and OctoThinker. For Llama, we utilize the MATH dataset (Hendrycks et al., 2021) following prior studies (Zeng et al., 2025; Yue et al., 2025) due to its initial reasoning performance.

Evaluation Benchmarks. We select six widely recognized benchmarks in the mathematical reasoning domain for testing: AIME 2024, AIME 2025, MATH-500 (Hendrycks et al., 2021), AMC, Minerva (Lewkowycz et al., 2022), and OlympiadBench (Huang et al., 2024).

Implementation Details. The reweighting hyperparameter β is set to 0.2. Additional details regarding the training configurations and evaluation setting are provided in Appendix C.

4.2 Main Results

Table 1 presents the performance comparison between DAPO and our Silhouette-aware variant, providing strong empirical validation for our analysis.

Unlocking Potential in Less RL-Friendly Models. Our method outperforms the baseline across all models, with the most significant average gains observed in less RL-friendly families. For OctoThinker, the approach nearly doubles the pass rates on AIME 2024 (4.9% $\rightarrow$ 8.2%) and AIME 2025 (2.1% $\rightarrow$ 5.0%). Llama also achieves notable improvements on AIME 2024 and Math500. These results confirm that explicitly targeting samples with poor distributional clarity effectively alleviates the optimization barrier for these models.

Enhancing Strong Models. Even for the already RL-friendly Qwen, our strategy yields further gains (12.2% $\rightarrow$ 18.1% on AIME 24). This indicates that optimizing for distributional clarity is a universally beneficial objective, helping even capable models to further refine their decision boundaries.

Distributional Clarity Drives Performance. Figure 8 tracks Silhouette Coefficient (S) and pass rate evolution across training, revealing strong correlation ($r = 0.815$) that confirms the causal relationship between distributional clarity and performance. Model families exhibit distinct trajectories: less RL-friendly models (OctoThinker, Llama) start with extremely low S, where standard DAPO yields marginal clarity improvements while our strategy induces substantial S increases corresponding to significant performance gains. Qwen starts with higher S but still advances toward the high-performance region, demonstrating that dis-

Model	AIME24 *avg@256*	AIME25 *avg@256*	MATH500 *avg@32*	AMC23 *avg@16*	Minerva *avg@16*	Olympiad *avg@16*	Average
Qwen Family							
Qwen2.5-7B	5.5	2.6	50.7	30.5	19.4	23.1	22.0
+ DAPO	12.2	11.8	79.7	**70.2**	**36.4**	42.4	42.1
+ DAPO-Silhouette	**18.1**	**12.0**	**80.4**	**70.2**	34.3	**43.2**	**43.0**
OctoThinker Family							
OctoThinker-8B	1.5	0.7	39.0	18.0	13.3	13.2	14.3
+ DAPO	4.9	2.1	59.4	46.9	27.6	25.8	27.8
+ DAPO-Silhouette	**8.2**	**5.0**	**62.6**	**47.2**	**29.6**	**27.8**	**30.1**
Llama Family							
Llama-3.1-8B-Instruct	3.7	0.4	46.8	24.7	21.7	15.4	18.8
+ DAPO	6.1	**0.5**	51.4	25.2	25.1	**19.4**	21.3
+ DAPO-Silhouette	**7.9**	0.4	**53.1**	**27.7**	**25.4**	**19.4**	**22.3**

Table 1: Main results across six mathematical reasoning benchmarks. Our Silhouette-aware strategy consistently outperforms the standard DAPO baseline, with particularly significant gains for less RL-friendly models (OctoThinker and Llama) on challenging benchmarks like AIME24.

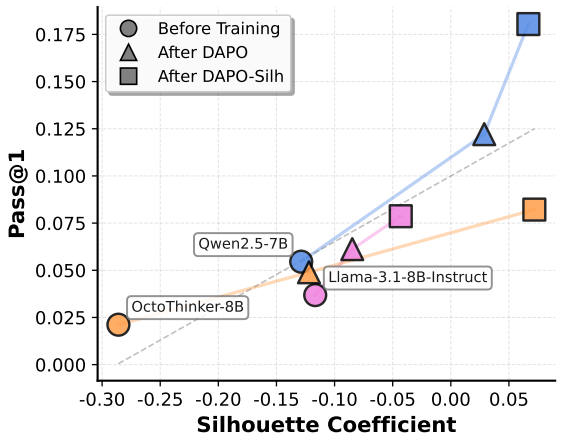


Figure 8: Evolution of Silhouette Coefficient and Pass@1 across training stages on AIME24. The strong positive correlation ($r = 0.815$) confirms that performance gains are driven by improved distributional clarity.

tributional clarity remains a bottleneck even for capable models.

4.3 Robustness and Ablation Analysis

Our analysis framework posits that the core driver of RL-Friendliness is distributional clarity—specifically, intra-class compactness and inter-class separation—rather than any particular choice of measurement metric or algorithm. To verify the generality of our approach, we conduct comprehensive ablation studies using the OctoThinker-8B backbone, as summarized in Table 2.

Metric Generalization. We first investigate whether the performance gain is an artifact of the Silhouette Coefficient. We implement an alternative variant using the Fisher Ratio (Bishop and Nasrabadi, 2006), a classical statistic that explicitly measures the ratio of inter-class variance to intra-class variance. By integrating this metric into DAPO via a similar reweighting strategy (details in Appendix C.3), we observe performance improvements comparable to the Silhouette Coefficient. This confirms that the benefit stems from emphasizing distributional clarity itself, independent of the specific mathematical formulation used to quantify it.

Mechanism Validation. To verify that distributional clarity specifically drives the gains, we evaluate three alternative strategies (Table 2). First, we test an Inter-class Separation Only variant, which modifies the weight to solely maximize the margin between correct and incorrect probabilities, ignoring intra-class compactness. While this yields improvements over the baseline (28.8% vs. 27.8%), it falls short of the full Silhouette strategy (30.1%). This underscores that intra-class compactness is necessary for optimal RL-Friendliness. We further compare against Pass-Rate Reweighting and Random Reweighting. Pass-Rate Reweighting, which prioritizes low-success samples, achieves only 28.7%, indicating that the benefits of our Silhouette strategy arise from distributional clarity rather than hard sample mining based on coarse outcome statistics. Random Reweighting yields 27.5%, below the baseline, confirming that improvements require informative signals, not arbitrary reweighting. Collectively, these comparisons establish the necessity of Silhouette-Aware Reweighting for capturing signals that partial metrics and heuristic baselines fail to exploit.

Hyperparameter Sensitivity. We analyze the impact of reweighting intensity β by evaluating $\beta \in \{0.1, 0.2, 0.5\}$. Table 2 show that our method

Method Setting	AIME24	AIME25	MATH500	AMC23	Minerva	Olympiad	Average
Default Setting							
DAPO	4.9	2.1	59.4	46.9	27.6	25.8	27.8
DAPO-Silhouette ($\beta = 0.2$)	**8.2**	5.0	62.6	47.2	29.6	27.8	30.1
Metric Generalization							
DAPO-Fisher Ratio	7.5	4.5	63.6	44.4	29.0	**29.9**	29.8
Mechanism Validation							
Inter-class Separation Only	5.8	2.2	62.4	48.0	26.7	27.6	28.8
Pass-Rate Reweighting	5.0	2.6	61.1	**49.7**	27.2	26.8	28.7
Random Reweighting	5.0	1.8	60.5	42.8	28.8	26.3	27.5
Hyperparameter Sensitivity							
DAPO-Silhouette ($\beta = 0.1$)	6.9	3.4	62.8	44.7	28.0	26.9	28.8
DAPO-Silhouette ($\beta = 0.5$)	6.6	**6.3**	**65.5**	46.6	**30.4**	29.6	**30.8**

Table 2: Robustness and ablation analysis on OctoThinker-8B. We confirm the effectiveness of our strategy across different metrics (Fisher Ratio) and varying hyperparameter settings (β). Mechanism Validation compares our full Silhouette approach against a variant optimizing only separation, outcome-based reweighting (Pass-Rate), and random noise.

proves highly robust, with all β values consistently outperforming the DAPO baseline. The aggressive setting ($\beta = 0.5$) achieves the highest overall average (30.8%), particularly excelling on MATH-500 and Minerva. However, the moderate setting ($\beta = 0.2$) yields the best performance on the challenging AIME 2024 benchmark (8.2% vs. 6.6%). Consequently, we adopted $\beta = 0.2$ as the default for our main experiments to balance performance across diverse difficulty benchmarks, while noting that higher β values may offer further potential.

5 Related Work

RLVR. RLVR has emerged as a dominant paradigm for enhancing reasoning capabilities (Guo et al., 2025; Shao et al., 2025b; Team et al., 2025; Chen et al., 2025a; Liu et al., 2025a). Recent studies have explored reshaping reward or advantage by leveraging the model's intrinsic generation metrics to improve exploration efficiency. For instance, Cheng et al. (2025) incorporates token-level entropy into the advantage function, while others utilize model uncertainty—derived from perplexity or confidence scores—to refine credit assignment (Dai et al., 2025; Xie et al., 2025; Wang et al., 2025a; Chen et al., 2025b; Cui et al., 2025). Unlike these methods that focus on single-sample uncertainty, we reshape the advantage function using the group-level probability landscape to enforce distributional clarity.

Data-Centric Approaches to Reasoning. Prior research largely attributes RL efficacy to data composition (Shao et al., 2025a; Mo et al., 2025; Tu et al., 2025), advocating for specific reasoning patterns like self-verification (Gandhi et al., 2025) or high-quality mathematical mid-training (Wang et al., 2025c; Tian et al., 2025). Similarly, systematic studies highlight the benefits of front-loading reasoning data (Akter et al., 2025) and bridging syntactic gaps (Liu et al., 2025b; Zhang et al., 2025) to establish a robust foundation. Unlike these approaches that focus on *what* models learn, we propose an orthogonal perspective: analyzing *how* the intrinsic probability landscape affects trainability. We identify distributional clarity—specifically intra-class compactness and inter-class separation—as a prerequisite for effective RL optimization.

6 Conclusion

In this work, we investigated the disparity in RL-Friendliness across foundation models, proposing a shift from purely data-centric views to an intrinsic distributional perspective. Our three-stage analysis identified distributional clarity as a critical determinant of RL-Friendliness. we demonstrated that poor distributional clarity (quantified by the Silhouette Coefficient) is strongly associated with high-severity logic errors and reasoning instability. Validating this insight, our Silhouette-Aware Reweighting strategy significantly enhanced the performance of less RL-friendly models by prioritizing samples with ambiguous distributions. Our findings highlight that beyond data composition, the intrinsic distributional properties of foundation models serve as a fundamental prerequisite for effective reinforcement learning.

Limitations

Our Silhouette-Aware Reweighting strategy relies on group-relative statistics derived from multiple rollouts (set to $G = 16$) to estimate distributional clarity. While this statistical approach effectively stabilizes the training signal, the precision of the Silhouette Coefficient inherently benefits from a sufficient sample size. In scenarios with extremely restrictive sampling budgets or where generation diversity is severely collapsed, the distributional estimation may become less robust. However, it is worth noting that modern RLVR algorithms, such as GRPO and DAPO, already necessitate group sampling for baseline advantage estimation; thus, our approach leverages existing computational structures rather than introducing additional sampling overhead.

Additionally, our study focuses on models generating standard Chain-of-Thought reasoning paths, rather than the extremely long reasoning trajectories (e.g., exceeding 10k tokens) recently observed in some specialized reasoning models. We adopted this setting to facilitate a controlled comparison across diverse open-weight model families, such as Llama-3.1, which may not inherently support such extended contexts without specific adaptation. Furthermore, limiting the response length allows for extensive ablation studies within accessible computational resources, though investigating the distributional properties of these ultra-long reasoning processes remains a promising direction.

References

Syeda Nahida Akter, Shrimai Prabhumoye, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, Yejin Choi, and Bryan Catanzaro. 2025. Front-loading reasoning: The synergy between pretraining and post-training data. *arXiv preprint arXiv:2510.03264*.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, and 1 others. 2025a. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. 2025b. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

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

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, and 1 others. 2025. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, and 1 others. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Emmy Liu, Graham Neubig, and Chenyan Xiong. 2025b. Midtraining bridges pretraining and posttraining distributions. *arXiv preprint arXiv:2510.14865*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. 2025. Mid-training of large language models: A survey. *arXiv preprint arXiv:2510.06826*.

Kevin P Murphy. 2022. *Probabilistic machine learning: an introduction*. MIT press.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, and 1 others. 2025a. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*.

Zhihong Shao, Yuxiang Luo, Chengda Lu, ZZ Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025b. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *arXiv preprint arXiv:2511.22570*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yijun Tian, Shaoyu Chen, Zhichao Xu, Yawei Wang, Jinhe Bi, Peng Han, and Wei Wang. 2025. Reinforcement mid-training. *arXiv preprint arXiv:2509.24375*.

Chengying Tu, Xuemiao Zhang, Rongxiang Weng, Rumei Li, Chen Zhang, Yang Bai, Hongfei Yan, Jingang Wang, and Xunliang Cai. 2025. A survey on llm mid-training. *arXiv preprint arXiv:2510.23081*.

Jinpeng Wang, Chao Li, Ting Ye, Mengyuan Zhang, Wei Liu, and Jian Luan. 2025a. Icpo: Intrinsic confidence-driven group relative preference optimization for efficient reinforcement learning. *arXiv preprint arXiv:2511.21005*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025c. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*.

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, Mark Coates, and Jian-Yun Nie. 2025. It takes two: Your GRPO is secretly DPO. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. 2025. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025a. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, and 1 others. 2025b. Rlpr: Extrapolating rlvr to general domains without verifiers. *arXiv preprint arXiv:2506.18254*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Charlie Zhang, Graham Neubig, and Xiang Yue. 2025. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*.

A Details of Three-Stage Analysis

In this section, we provide supporting materials and detailed implementation settings for the three-stage analysis framework presented in Section 2. This includes additional performance visualizations (Phenomenon Level) and the fine-grained taxonomies and prompts used for behavioral interpretation (Interpretation Level).

A.1 Phenomenon Level: Additional Pass Rate Analysis

In Section 2.1, we presented the per-problem pass rate comparison on the AIME 2024 benchmark. To further validate that the observed performance disparity is a consistent phenomenon across different mathematical reasoning tasks, we conduct the corresponding analysis on the MATH-500 benchmark.

As shown in Figure 9, the results on MATH-500 exhibit a pattern highly consistent with AIME 2024. We observe a substantial intersection in the problem sets solvable by both models, indicating that they share comparable latent capabilities. However, within this large shared domain, Qwen2.5-7B achieves higher pass rates. This further confirms that RL-friendly models possess a superior capacity to assign high probability mass to correct solutions across diverse problem sets.

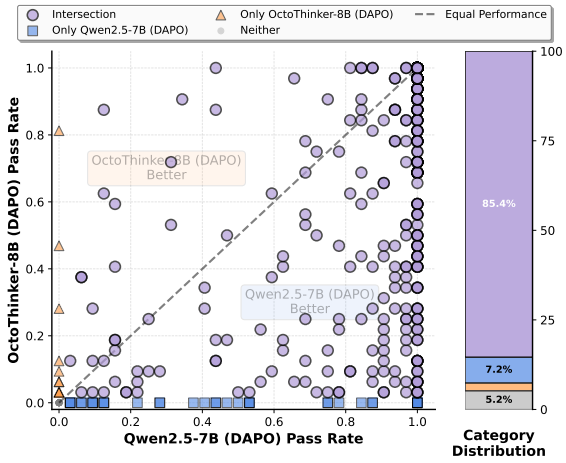


Figure 9: Per-problem pass rate comparison between Qwen2.5-7B (DAPO) and OctoThinker-8B (DAPO) on the MATH-500 benchmark. Despite a large intersection of mutually solvable problems, Qwen demonstrates higher pass rates on these shared queries compared to OctoThinker, mirroring the trend observed in AIME 2024.

A.2 Interpretation Level: Error Attribution Details

Error Taxonomy. We construct a comprehensive taxonomy to categorize incorrect responses based on the root cause of the failure. As shown in Table 3, these categories are grouped into three severity levels: High (Fundamental), Mid (Execution), and Low (Presentation). This hierarchy allows us to distinguish between deep cognitive failures and superficial implementation mistakes.

Evaluation Prompt. We employ Qwen3-32B (Yang et al., 2025) as the judge model to classify errors. The prompt includes the full taxonomy definitions to ensure consistent attribution. The exact prompt template is provided in Figure 10.

A.3 Interpretation Level: Solution Stability Details

In this subsection, we describe the methodology used to quantify the stability of reasoning paths generated by the models.

Clustering Algorithm. To group correct solutions based on their semantic method, we employ an incremental clustering approach driven by an LLM judge. Unlike traditional embedding-based clustering which may fail to capture subtle logic differences, our method utilizes pairwise comparisons to determine if two solutions rely on the same core mathematical strategy. The detailed procedure is outlined in Algorithm 1. For each problem, we maintain a list of distinct solution clusters. For every new correct response, we compare it against the representative solution of existing clusters. If a match is found, the response is assigned to that cluster; otherwise, it forms a new cluster.

Evaluation Prompt. The LLM judge determines whether two solutions share the same fundamental approach. The specific criteria for same method (e.g., identical theorems or logical strategies) versus different method (e.g., geometric vs. algebraic approaches) are explicitly defined in the system prompt to ensuring consistency. The complete prompt used for this pairwise comparison is presented in Figure 11.

Category	Code	Description
Fundamental Errors (High)	E1.1	**Misunderstanding Question:** Misread variables, function definitions, or the goal.
	E1.2	**Constraint Violation:** Ignored constraints (e.g., "positive integers", "distinct").
	E2.1	**Knowledge Error:** Used wrong mathematical formulas, theorems, or facts.
	E2.2	**Planning/Method Error:** The chosen approach or model was fundamentally flawed.
	E5.1	**Repetition Loop:** Entered a degenerative loop repeating the same sequence.
	E5.2	**Irrelevant/Incoherent:** Failed to address the specific question or generated noise.
Execution Errors (Mid)	E3.1	**Calculation/Execution Error:** Correct formula but failed arithmetic or algebraic manipulation.
	E3.2	**Step Hallucination:** Invented intermediate values or steps without basis.
Presentation Errors (Low)	E4.1	**Format Error:** Correct answer but failed to format (e.g., missing `\boxed{}`).
	E4.2	**Premature Stop:** Generation cut off before completion.

Table 3: Fine-grained taxonomy for error attribution. Errors are categorized by severity: High (Fundamental), Mid (Execution), and Low (Presentation).

Prompt for Error Attribution Analysis

You are a mathematical reasoning analyst. Your task is to analyze a MODEL RESPONSE to a math question that resulted in an INCORRECT answer. Identify the primary cause of the error based on the following taxonomy.

Taxonomy for Incorrect Outcomes:
1. **[E1.1] Misunderstanding Question:** Misread variables, function definitions, or the goal.
2. **[E1.2] Constraint Violation:** Ignored constraints (e.g., "positive integers", "distinct", "minimum").
3. **[E2.1] Knowledge Error:** Used wrong mathematical formulas, theorems, or facts.
4. **[E2.2] Planning/Method Error:** The chosen approach/model was fundamentally flawed.
5. **[E3.1] Calculation/Execution Error:** Formula was correct, but arithmetic or algebraic manipulation failed.
6. **[E3.2] Step Hallucination:** Invented intermediate values or steps without basis.
7. **[E4.1] Format Error:** Answer calculated correctly but failed to format (e.g., missing `\boxed{}`).
8. **[E4.2] Premature Stop:** Generation cut off before completion.
9. **[E5.1] Repetition Loop:** The model entered a degenerative loop, repeating the same phrase, number, or sequence endlessly.
10. **[E5.2] Irrelevant/Incoherent:** The response fails to address the specific question. It may be unrelated text, pure text completion, gibberish, or nonsensical noise.
11. **[E6.1] Other:** The response fits none of the above categories.

Output Format (JSON):
{
"`category_code`": "E3.1",
"`reason`": "Brief explanation of why this category was chosen."
}

[Question]
{question}

[Ground Truth]
{ground_truth}

[Model Response]
{model_response}

Analyze the response based on the Ground Truth and provide the attribution category code and reason in JSON.

Figure 10: The LLM-as-a-judge prompt used for error attribution analysis.

Algorithm 1 Incremental Solution Method Clustering

Require: Set of correct solutions $\mathcal{R} = \{o_1, o_2, \ldots, o_N\}$ for a query q
Require: LLM Judge $\mathcal{J}(o_a, o_b) \rightarrow \{0, 1\}$
Ensure: Set of method clusters $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$

Initialize clusters $\mathcal{C} \leftarrow \emptyset$
for each solution o_i in $\mathcal{R}$ **do**
    $assigned \leftarrow$ FALSE
    for each cluster c_k in $\mathcal{C}$ **do**
        $o_{rep} \leftarrow$ Representative(c_k) ▷ First element of cluster
        if $\mathcal{J}(o_i, o_{rep}) == 1$ **then**
            Add o_i to c_k
            $assigned \leftarrow$ TRUE
            break
        end if
    end for
    if $assigned ==$ FALSE **then**
        Create new cluster $c_{new} = \{o_i\}$
        Add c_{new} to $\mathcal{C}$
    end if
end for
return $\mathcal{C}$

Prompt for Solution Method Comparison

You are a mathematics expert and logic analyst. Your task is to compare two CORRECT solutions to the same math problem and determine if they use the **same core mathematical method/approach**.

Criteria for same Method:

- They use the same key theorems, formulas, or logical strategy (e.g., both use "Coordinate Geometry" or both use "Mass Point Geometry").
- Ignore differences in variable names, calculation order, or verbosity.
- Ignore differences in how the answer is formatted if the derivation logic is identical.

Criteria for Different Method:

- One uses a geometric approach while the other uses an algebraic/trigonometric approach.
- One uses a brute-force enumeration while the other uses a combinatorial formula.
- One uses a specific theorem (e.g., Fermat's Little Theorem) while the other uses pattern finding.

Output JSON Format:
{
"`is_same_method`": *boolean*,
"`reason`": "Brief explanation of the similarity or difference."
}

[Question]
{question}

[Solution A]
{solution_a}

[Solution B]
{solution_b}

Do Solution A and Solution B use the fundamentally same mathematical method?

Figure 11: The LLM-as-a-judge prompt used to determine if two solutions utilize the same reasoning method.

B Gradient Variance Analysis

For a query q with correct responses sampled from $\pi^+(\cdot|q)$ and incorrect responses from $\pi^-(\cdot|q)$, the GRPO policy gradient can be written as (Wu et al., 2025):

$$\nabla_\theta J \propto \mathbb{E}_{o_j\sim\pi^+}[\nabla_\theta \pi_\theta(o_j|q)] - \mathbb{E}_{o_k\sim\pi^-}[\nabla_\theta \pi_\theta(o_k|q)] \tag{9}$$

Define $g_+ = \mathbb{E}_{o_j\sim\pi^+}[\nabla_\theta \pi_\theta(o_j|q)]$ and $g_- = \mathbb{E}_{o_k\sim\pi^-}[\nabla_\theta \pi_\theta(o_k|q)]$ as the expected gradients over correct and incorrect responses. Assuming that correct and incorrect responses are sampled independently, the variance of the policy gradient is:

$$\mathrm{Var}(\nabla_\theta J) = \mathrm{Var}(g_+) + \mathrm{Var}(g_-) \tag{10}$$

For a language model with softmax output layer, the derivative of the probability π with respect to the logit z is $\frac{\partial \pi}{\partial z} = \pi(1-\pi)$. By the chain rule, the gradient of the probability with respect to model parameters θ can be written as:

$$\nabla_\theta \pi_\theta(y|x) = \pi(y|q)(1-\pi(y|q)) \cdot \nabla_\theta z \tag{11}$$

Since hidden representations in transformers are normalized and exhibit stable variance, the stochasticity of gradient updates is predominantly governed by the $\pi(1-\pi)$ term. We obtain:

$$\mathrm{Var}(\nabla_\theta J) \propto \mathrm{Var}_{\pi^+}[\pi(1-\pi)] + \mathrm{Var}_{\pi^-}[\pi(1-\pi)] \tag{12}$$

where we use the shorthand $\mathrm{Var}_{\pi^+}[\pi(1-\pi)]$ to denote $\mathrm{Var}_{o_j\sim\pi^+}[\pi(o_j|q)(1-\pi(o_j|q))]$. When probabilities of correct responses are tightly clustered around some value π_+, the variance $\mathrm{Var}_{o_j\sim\pi^+}[\pi(1-\pi)]$ is small. Similarly, when incorrect responses cluster tightly around π_-, the variance $\mathrm{Var}_{o_k\sim\pi^-}[\pi(1-\pi)]$ is small.

C Experimental Setup Details

In this section, we provide the detailed configurations for both the training and evaluation phases.

C.1 Training Configuration

We implement our training pipeline using the verl framework (Sheng et al., 2025). All models are trained on a node of 8 NVIDIA H800 GPUs. The training utilizes the DAPO algorithm with the hyperparameters detailed in Table 4.

Hyperparameter	Value
Data Configuration	
Max prompt length	2048
Max response length	8192
DAPO Algorithm Configuration	
Advantage estimator	GRPO
Clip ratio (low)	0.2
Clip ratio (high)	0.28
Responses per prompt	16
Sampling temperature	1.0
Sampling top-p	1.0
KL in reward	False
KL loss	False
Optimization Configuration	
Optimizer	AdamW
Learning rate	1e-6
Learning rate warmup steps	10
Weight decay	0.1
Gradient clipping	1.0
Batch size	512
Mini-batch size	32
Total training steps	200

Table 4: Training hyperparameters.

C.2 Evaluation Configuration

During evaluation, we generate multiple reasoning paths for each query. The decoding parameters are set to `temperature=0.6` and `top_p=0.95`, with a maximum token limit of 8192.

The number of sampled rollouts (K) varies by benchmark:

- **AIME 2024, AIME 2025:** $K = 256$.
- **MATH-500:** $K = 32$.
- **AMC 23, Minerva, OlympiadBench:** $K = 16$.

The final performance is reported as the average value calculated across these K samples.

C.3 Fisher Ratio Implementation

To verify the robustness of our approach across different distributional metrics, we implement a variant based on the Fisher Ratio (F). While the Silhouette Coefficient focuses on geometric cluster separation, the Fisher Ratio statistically quantifies the separation by comparing the squared difference of means to the sum of variances.

For a given query, let $\mu_{pos}, \sigma^2_{pos}$ and $\mu_{neg}, \sigma^2_{neg}$ denote the mean and variance of the sequence-level probabilities for correct and incorrect responses, respectively. The standard Fisher Ratio F is calcu-

lated as:

$$F = \frac{(\mu_{pos} - \mu_{neg})^2}{\sigma_{pos}^2 + \sigma_{neg}^2} \tag{13}$$

Similar to our Silhouette strategy, we must handle cases where the distribution is inverted (i.e., incorrect responses have higher probabilities than correct ones). We define a rectified Fisher score F':

$$F' = \begin{cases} F & \text{if } \mu_{pos} \geq \mu_{neg} \\ -F & \text{if } \mu_{pos} < \mu_{neg} \end{cases} \tag{14}$$

The reweighting factor $w(q)$ is then computed using an exponential decay function. Unlike the Silhouette Coefficient which is bounded in $[-1, 1]$, the Fisher Ratio can vary widely in magnitude. Therefore, we apply a clamping mechanism to ensure training stability:

$$w(q) = \text{clip}\left(\exp(-\beta \cdot F'), 0.95, 1.05\right) \tag{15}$$

where $\text{clip}(x, \text{min}, \text{max})$ restricts the weight within the specified range. For our experiments, we set the sensitivity hyperparameter $\beta = 0.01$. Finally, the advantage is modulated as $\hat{A}_{new} = \hat{A}_{old} \cdot w(q)$.