

A Step Harder, Not a Step Too Far: The Benefits and Limits of Curriculum Difficulty Progression in GRPO

Ronghan Li¹, Fei Chen^{2*}, Mingyu Lu¹

¹College of Artificial Intelligence, Dalian Maritime University, Dalian, 116026, Liaoning, China.

^{2*}Information Science and Technology College, Dalian Maritime University, Dalian, 116026, Liaoning, China.

*Corresponding author(s). E-mail(s): chenf@dlmu.edu.cn;

Abstract

Curriculum learning typically organizes training samples from easy to hard. In GRPO, medium-difficulty samples are attractive for online selection because they are more likely to yield informative group-relative learning signals, and prioritizing them can improve training efficiency. However, it remains unclear whether persistently selecting such samples also leads to the strongest final model performance, or whether training can benefit from progressing toward harder samples. Meanwhile, sample difficulty changes as the policy evolves, so fixed difficulty estimates can become misaligned with the current policy. We therefore re-estimate sample difficulty under the current policy and, inspired by the easy-to-hard principle of curriculum learning, design and evaluate a staged curriculum that progressively shifts the active difficulty band toward harder samples. Experiments on GSM8K show that persistent medium-difficulty selection does not achieve the strongest observed performance. Across three random seeds, progressing to a moderately harder difficulty band improves test accuracy to 0.6591 ± 0.0077 , compared with 0.6449 ± 0.0024 for the fixed-medium curriculum. However, further progression to a harder difficulty band reduces accuracy to 0.6405 ± 0.0053 . These results show that staged difficulty progression can outperform persistent medium-difficulty selection, but its benefit is non-monotonic. Effective GRPO curricula can therefore benefit from progressing beyond medium difficulty, while avoiding excessive progression toward overly difficult samples.

Keywords: reinforcement learning with verifiable rewards, group relative policy optimization, difficulty dynamics, curriculum learning

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) provides a direct way to improve reasoning models when outputs can be checked by explicit correctness rules. Group Relative Policy Optimization (GRPO) constructs policy updates from multiple responses to the same problem and does not require a separately trained value model [14]. Under binary correctness rewards, problems whose responses are nearly all correct or nearly all incorrect tend to produce low within-group reward variance and thus weak group-relative learning signals [22]. In contrast, prompts whose sampled responses differ in correctness provide nonzero reward variance for group-relative advantage estimation and can therefore yield more informative learning signals [7].

This property raises important questions for curriculum learning in GRPO. Classical curriculum learning typically organizes training samples from easy to hard according to estimated difficulty [2]. In GRPO, however, two practical questions arise: whether difficulty estimates computed before training remain aligned with the evolving policy, and whether progressively increasing curriculum difficulty continues to provide useful learning signals. Because difficulty is relative to the current policy, the same problem may move between easy, medium, and hard difficulty bands as training proceeds. Moreover, a lower success rate does not necessarily correspond to a more informative optimization signal. Consequently, a curriculum constructed from initial difficulty can become temporally misaligned with the current policy. This motivates estimating sample difficulty online under the current policy, so that training decisions reflect the model’s current state rather than stale initial estimates. However, online difficulty estimation alone does not determine which difficulty band should be used as the policy evolves.

Figure 1 summarizes the motivation of this work. In GRPO, intermediate-solvability problems are attractive for online selection because they are more likely to generate mixed correct and incorrect responses and therefore informative group-relative signals. At the same time, difficulty may change as the policy evolves, motivating our investigation of both stale precomputed difficulty estimates and curriculum scheduling across difficulty bands.

Recent work has shown that adaptive curricula and online data selection can improve RLVR data efficiency by using reward statistics, estimated success rates, uncertainty, gradient information, or learnability signals [1, 4, 8, 17, 18, 20, 28]. In GRPO, samples with intermediate success rates are particularly relevant because they are more likely to produce groups containing both correct and incorrect responses, thereby yielding nonzero group-relative learning signals [7, 22]. This raises a further question: is continually selecting such medium-difficulty samples sufficient to maximize final model performance, or can the model benefit from progressing toward harder samples as its policy evolves? If such progression is beneficial, it remains unclear whether the benefit persists as curriculum difficulty continues to increase.

We address these questions through a sequence of curriculum experiments. First, we track the same 1,495 GSM8K training samples throughout a continuous GRPO run to characterize how sample difficulty evolves with the policy. Second, we test whether a static curriculum constructed from initial difficulty becomes misaligned with the current policy. Third, we introduce a fixed-medium curriculum that keeps the active

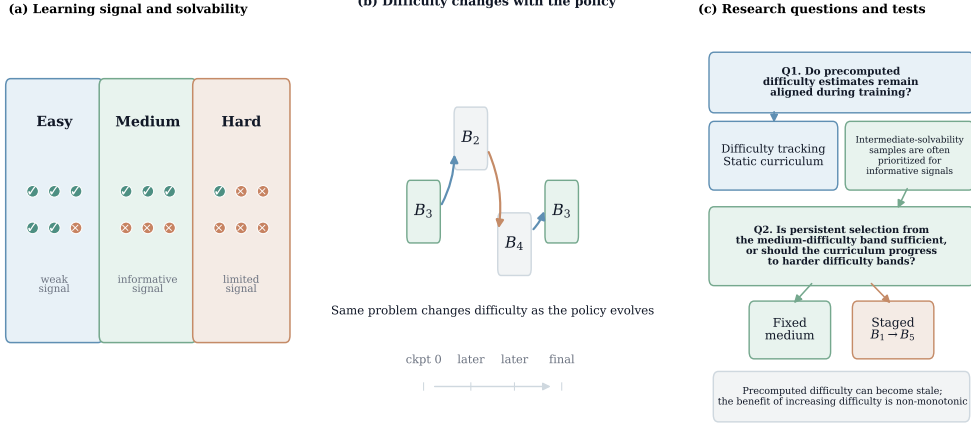


Fig. 1 Motivation and research questions for difficulty-aware training in GRPO. (a) The usefulness of a training problem depends on its current solvability: overly easy or hard problems tend to yield weak group-relative learning signals, whereas intermediate-solvability problems are more likely to produce informative comparisons among sampled responses. (b) Sample difficulty depends on the current policy and can change across difficulty bands as training proceeds, making precomputed difficulty estimates potentially stale. (c) Intermediate-solvability samples are therefore natural candidates for online selection and are emphasized by recent difficulty-aware selection strategies. This motivates two questions examined in this study: whether precomputed difficulty estimates remain aligned with the evolving policy, and whether persistently selecting the medium-difficulty band is sufficient or the curriculum should progress to harder difficulty bands during training.

difficulty band fixed at B_3 while samples are reselected online according to the current policy. This setting tests whether continuously selecting samples from an informative medium-difficulty band is sufficient for maximizing final performance. Finally, we evaluate a staged curriculum in which the active difficulty band progresses from easier to harder bands according to the empirical success rate under the current policy. Across three random seeds, the staged curriculum achieves higher test performance at the moderately harder B_4 band than the fixed-medium curriculum. However, further progression into the harder B_5 band consistently reduces test performance. Together, these experiments suggest that medium-difficulty samples provide informative learning signals but are not sufficient by themselves to maximize final model capability. Progression to a moderately harder difficulty band can improve performance, whereas further increases in curriculum difficulty are not monotonically beneficial. Detailed metric definitions, additional training trajectories, and performance on the fixed diagnostic subsets used for checkpoint selection and stage transitions are provided in the appendices.

Figure 2 provides a visual comparison of the training paradigms considered in this study, highlighting the differences between static difficulty ordering, online sample selection, and staged difficulty progression.

The main contributions of this work are summarized as follows:

- **Policy-dependent difficulty dynamics.** We systematically characterize how sample difficulty changes during GRPO training and show that individual samples exhibit substantial and non-monotonic difficulty changes. We further observe

Comparison of Training Paradigms

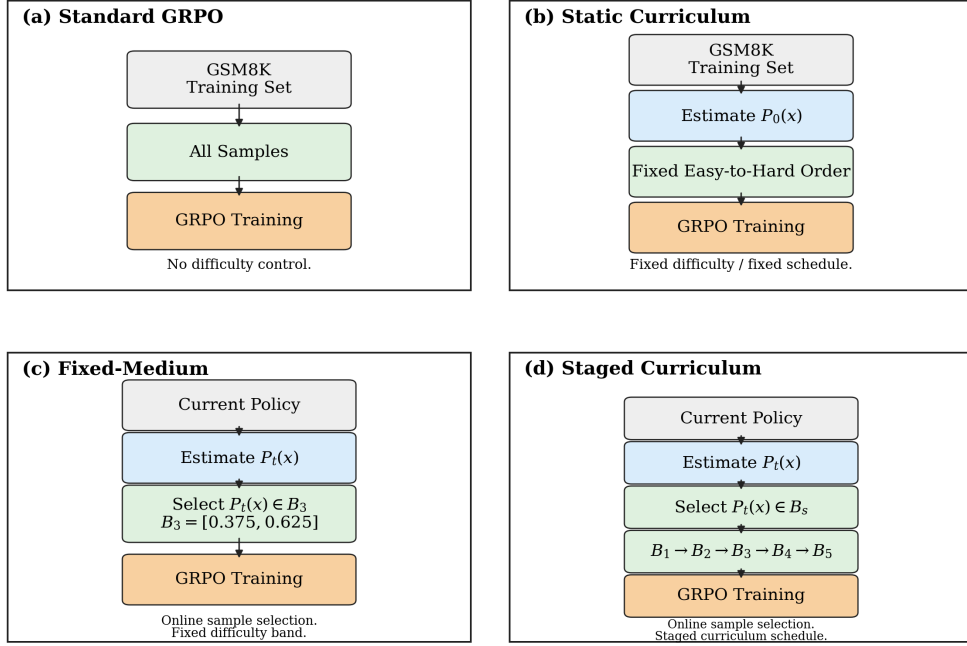


Fig. 2 Comparison of the training paradigms considered in this study. Standard GRPO uses the full training set without difficulty-based selection. The static curriculum uses a fixed easy-to-hard ordering determined by the initial solvability estimate $P_0(x)$. The fixed-medium curriculum reselects samples online according to the current-policy success rate while keeping difficulty band B_3 fixed. The staged curriculum reselects samples online while moving the active difficulty band through $B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5$.

that relatively stable global difficulty rankings can coexist with substantial turnover within specific difficulty bands, highlighting the limitations of relying on fixed difficulty estimates throughout training.

- **Staged difficulty progression.** We design and evaluate a staged curriculum that re-estimates sample difficulty under the current policy and progressively shifts the active difficulty band toward harder samples. Experiments across multiple random seeds show that persistent medium-difficulty selection does not achieve the strongest observed performance, while progressing to a moderately harder difficulty band provides additional gains. However, further progression to harder samples reduces performance, revealing a non-monotonic benefit of increasing curriculum difficulty.

The remainder of this paper is organized as follows. Section 2 reviews related work on GRPO and reinforcement learning with verifiable rewards, dynamic data selection, and curriculum learning with model-dependent difficulty. Section 3 introduces the problem formulation and empirical framework, including policy-dependent sample difficulty and the curriculum settings considered in this work. Section 4 describes the experimental setup for difficulty tracking, static curriculum training, the fixed-medium

curriculum, and the staged curriculum. Section 5 presents the empirical results on difficulty dynamics, the misalignment of precomputed difficulty, and performance under different curriculum conditions. Section 6 discusses the implications of the results, the relationship between online sample selection and difficulty progression, as well as the limitations and future directions of this study. Finally, Section 7 concludes the paper by summarizing the main findings and their implications for difficulty-aware GRPO training.

2 Related Work

2.1 GRPO and RLVR

RLVR has become a major approach for improving reasoning with rule-based or otherwise verifiable rewards. DeepSeekMath introduced GRPO for mathematical reasoning, and subsequent work extended verification-based reinforcement learning to broader reasoning tasks and formal theorem proving [5, 14, 21]. Compared with PPO-based approaches that rely on an additional value model, GRPO constructs relative advantages from multiple sampled responses within the same prompt, thereby reducing reliance on explicit value estimation. Recent studies have further investigated GRPO and related reinforcement learning frameworks from the perspectives of sampling strategy, objective design, and optimization stability [10, 24, 27].

In contrast to these algorithm-level improvements, our work focuses on the interaction between the evolving policy and training samples. Since GRPO derives learning signals from relative rewards within response groups, the usefulness of a problem is closely related to its current solvability under the current policy. However, existing GRPO studies mainly focus on optimization objectives and training stability, while the evolution of individual-sample difficulty with the policy has been less systematically characterized.

2.2 Dynamic Data Selection

Dynamic data selection aims to improve training efficiency by identifying samples that are more valuable given the current model state. Offline selection methods estimate data value using reward trajectories, influence estimation, or multi-factor scoring [9, 18, 19]. In contrast, online selection methods adapt sample choices during optimization according to model-dependent signals, such as reward variance, predicted success probability, Bayesian uncertainty, transition dynamics, or gradient-related information [1, 11, 13, 15, 17, 26]. Other approaches, including PCL and CurES, further exploit intermediate success rates or gradient-efficient sampling strategies to improve training efficiency [4, 25].

These studies motivate treating the utility of training samples as model-dependent rather than fixed. In particular, prioritizing samples with intermediate success rates provides a natural strategy for GRPO, since such samples are more likely to produce mixed correct and incorrect responses and therefore informative group-relative learning signals [7, 22]. Existing online selection methods already adapt sample choices

according to the current model state. However, it remains unclear whether continuously selecting samples with intermediate solvability, even when sample difficulty is re-estimated online, is sufficient to maximize final model performance. Our work complements these studies by systematically characterizing how the difficulty of the same samples evolves with the policy and by examining the sufficiency of persistent selection of intermediate-solvability samples as the policy evolves.

2.3 Curriculum Learning with Model-Dependent Difficulty

Curriculum learning traditionally organizes training data according to estimated difficulty, following the principle of gradually increasing task complexity [2]. Recent reasoning-oriented studies have incorporated curriculum strategies into reinforcement learning. E2H Reasoner explicitly organizes tasks in an easy-to-hard order [12], while Re-Schedule constructs curricula according to reasoning complexity rather than final-answer accuracy [20]. Other approaches, such as AdaRFT and AdaCuRL, further adjust training difficulty according to the evolving model state [8, 16].

These difficulty-aware curriculum approaches suggest that organizing training according to model-dependent difficulty can improve training efficiency by providing samples with appropriate learning challenges. However, traditional easy-to-hard curriculum learning is built on the intuition that models benefit from being exposed to increasingly complex examples as their capabilities improve [2]. This assumption may not always hold in GRPO, where optimization depends on relative rewards among sampled responses. Extremely easy problems may provide limited learning signals due to low reward variance, while extremely difficult problems may also provide limited informative comparisons when most responses are incorrect [7, 22].

Therefore, beyond estimating sample difficulty under the current policy, an important open question in GRPO curriculum learning is whether moving toward harder difficulty bands continues to provide benefits as the policy evolves. In particular, it remains unclear whether the gains from moving to harder difficulty bands persist as curriculum difficulty continues to increase. Our work investigates this issue by characterizing how difficulty evolves with the policy and evaluating a staged easy-to-hard curriculum across increasingly difficult bands. This allows us to examine whether progressively increasing curriculum difficulty remains beneficial, rather than assuming a monotonic benefit from moving to harder difficulty bands.

3 Problem Formulation and Empirical Framework

Tables 1 and 2 summarize the main notation used throughout this paper.

3.1 Difficulty Under the Current Policy

Let π_{θ_t} denote the policy at training state t . For a problem x , we define its solvability under the current policy as

$$p_t(x) = \Pr_{y \sim \pi_{\theta_t}(\cdot|x)} [V(x, y) = 1], \quad (1)$$

Table 1 Main notation for policy-dependent difficulty and curriculum settings.

Symbol	Meaning
x, x_i	A training problem; x_i denotes the i -th problem.
y	A response sampled from the policy for problem x .
t, t_k	Training state or checkpoint; t_k denotes the k -th tracked checkpoint.
θ_t, π_{θ_t}	Model parameters and the corresponding policy at training state t .
$V(x, y)$	Binary correctness verifier for response y to problem x .
$p_t(x)$	True solvability of problem x under the policy at training state t .
$d_t(x)$	True policy-dependent difficulty, defined as $1 - p_t(x)$.
G	Number of sampled responses used to estimate problem solvability.
$r_{t,g}(x)$	Binary correctness reward of the g -th sampled response for problem x at training state t .
$P_t(x)$	Empirical success rate estimated from G responses sampled under the policy at training state t .
$\hat{d}_t(x)$	Estimated policy-dependent difficulty, defined as $1 - P_t(x)$.
$\mathcal{T}$	Sequence of training checkpoints used to track difficulty changes.
$B_s = [l_s, u_s]$	Difficulty band used at curriculum stage s , operationalized by the success-rate interval $[l_s, u_s]$.
$B_1, \dots, B_5$	Predefined difficulty bands used in the staged curriculum.
B_3	Medium-difficulty band kept fixed in the fixed-medium curriculum and used as the third stage of the staged curriculum.
$a_t(x)$	Binary gating variable indicating whether problem x belongs to the active difficulty band at training state t .

where $V(x, y) \in \{0, 1\}$ is the correctness verifier. The corresponding true difficulty is defined as $d_t(x) = 1 - p_t(x)$. With G sampled responses, we estimate solvability as

$$P_t(x) = \frac{1}{G} \sum_{g=1}^G r_{t,g}(x), \quad \hat{d}_t(x) = 1 - P_t(x). \quad (2)$$

Since the true solvability $p_t(x)$ cannot be computed exactly in practice, we use the empirical success rate $P_t(x)$ and the corresponding estimated difficulty $\hat{d}_t(x)$ for all analyses.

3.2 Difficulty Dynamics

Let $\mathcal{T} = \{t_0, t_1, \dots, t_K\}$ denote the ordered sequence of checkpoints used for difficulty tracking.

Table 2 Notation for difficulty-dynamics and difficulty-band membership metrics.

Symbol	Meaning
R_i	Difference between the maximum and minimum empirical solvability of problem x_i over the tracked training trajectory.
$D_{i,\text{init}}$	Maximum absolute deviation of problem x_i from its initial empirical solvability over the tracked trajectory.
$\Delta_{i,k}$	Difference in empirical solvability of problem x_i between consecutive checkpoints t_{k-1} and t_k .
δ	Minimum-change threshold used to identify meaningful direction reversals.
$S_t(B)$	Set of tracked problems whose empirical solvability lies in band B at training state t .
N	Total number of problems in the fixed observation set.
$\text{Retention}_t(B)$, $\text{Outflow}_t(B)$	Fractions of initial band members that remain in or leave band B , respectively.
$\text{Inflow}_t(B)$	Fraction of current members of band B that were outside the band initially.
$\text{Jaccard}_t(B)$	Jaccard similarity between the initial and current membership of band B .
$\text{Mismatch}_t(B)$	Symmetric difference between initial and current band membership, normalized by N .
TV_i	Sum of the absolute empirical-solvability changes of problem x_i across consecutive tracked checkpoints.
E_i	Extra variation caused by back-and-forth changes in the empirical solvability of problem x_i , defined as $TV_i - R_i$.
ρ, τ_b	Spearman and Kendall rank-correlation coefficients, respectively.

We quantify the variation in empirical success rate for each problem over training using its training-period span

$$R_i = \max_{t \in \mathcal{T}} P_t(x_i) - \min_{t \in \mathcal{T}} P_t(x_i). \quad (3)$$

We also measure the maximum absolute deviation from the initial empirical success rate as

$$D_{i,\text{init}} = \max_{t \in \mathcal{T}} |P_t(x_i) - P_0(x_i)|. \quad (4)$$

To identify non-monotonic changes, we examine checkpoint-to-checkpoint differences

$$\Delta_{i,k} = P_{t_k}(x_i) - P_{t_{k-1}}(x_i). \quad (5)$$

After discarding changes with $|\Delta_{i,k}| < \delta$, we count a direction reversal whenever two consecutive retained changes have opposite signs. Here, δ denotes the minimum-change threshold used to suppress small fluctuations caused by finite sampling. Since each empirical success-rate estimate is based on $G = 32$ sampled responses, we set $\delta = 0.125$ as the primary threshold, corresponding to four increments of $1/32$ in the

empirical success rate. Sensitivity analyses with alternative thresholds ($\delta = 0.0625$ and $\delta = 0.25$) are reported in Appendix B. We additionally examine global rank correlation and local membership changes in predefined success-rate intervals. Full definitions of retention, inflow, outflow, Jaccard similarity, mismatch, total variation, extra round-trip variation, and re-entry are given in Appendix A.

3.3 Evaluation of Static Difficulty and Curriculum Scheduling

We examine two related questions that arise because difficulty depends on the current policy. First, we test whether curricula constructed from initial difficulty estimates become misaligned with the current policy as training proceeds. Specifically, we construct a static curriculum by estimating sample difficulty under the initial policy and fixing the resulting difficulty ordering throughout training. This setting allows us to examine whether the initial difficulty estimates remain aligned with difficulty under the current policy.

Second, we study how curriculum difficulty should be scheduled once sample difficulty is re-estimated online under the current policy. In particular, we ask whether persistently using a medium-difficulty band is sufficient to maximize final performance and whether progressing to harder difficulty bands continues to provide benefits as curriculum difficulty increases. We use the term *difficulty band* to denote a predefined range of policy-dependent sample difficulty used for curriculum selection. Each difficulty band B_s is operationalized by a corresponding success-rate interval $[l_s, u_s]$. At curriculum stage s , B_s is referred to as the *active difficulty band*. A sample contributes to the policy loss only when its estimated solvability falls within the success-rate interval associated with the active difficulty band:

$$a_t(x) = \mathbb{I}[P_t(x) \in B_s]. \quad (6)$$

Under the fixed-medium curriculum, the active difficulty band remains fixed at B_3 while eligible samples are reselected online as the policy evolves. Under the staged curriculum, the active difficulty band progresses toward bands associated with lower success rates and therefore greater difficulty. Comparing these curricula allows us to examine whether online sample reselection is sufficient when the difficulty band remains fixed and how performance changes as the curriculum progresses to harder difficulty bands.

This gating mechanism is used only to define the experimental conditions. An effective update is an optimizer step in which at least one sample satisfies the gate and a parameter update actually occurs.

4 Experimental Setup

4.1 Model and Training Configuration

We use GSM8K [3] with Qwen2.5-1.5B-Instruct [23] and LoRA fine-tuning [6]. Due to different computational requirements, we use different numbers of sampled responses for difficulty estimation and online optimization. Specifically, we use $G = 32$ responses

Table 3 Difficulty bands and corresponding success-rate intervals used in the online curriculum conditions.

Difficulty band	Success-rate interval	Correct out of 8	Relative difficulty
B_1	[0.625, 0.875]	5–7	Easier
B_2	[0.500, 0.750]	4–6	Moderately easy
B_3	[0.375, 0.625]	3–5	Medium
B_4	[0.250, 0.500]	2–4	Moderately hard
B_5	[0.125, 0.375]	1–3	Harder

per problem for offline difficulty analysis and $G = 8$ responses per problem during curriculum experiments. The main training runs use a batch size of 8, a gradient accumulation factor of 1, a learning rate of 1×10^{-5} , a temperature of 0.7, top- p of 0.85, and a maximum generation length of 1,024 tokens. Answer correctness is used as the primary verifiable reward. An auxiliary length reward with a fixed weight of 0.2 and a preferred response-length range of 60–650 tokens is also used, but it is not used to compute $P_t(x)$ or make gating decisions. The official GSM8K test set is used only for generalization evaluation.

4.2 Difficulty Tracking and Static Curriculum

Before training, all 7,473 GSM8K training problems are evaluated with 32 generations each to obtain their initial solvability estimates. We first perform a reference GRPO run using the standard GRPO objective without difficulty-based filtering or curriculum scheduling. During this training process, checkpoints are saved at $\{0, 10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000, 4000, 7473\}$.

For performance comparison, we independently run the standard GRPO baseline with the same three random seeds used in the curriculum experiments. Each run is trained for one complete pass over the 7,473 GSM8K training problems. We retrospectively report the best observed test accuracy along each completed training trajectory; these test values are used only for post-hoc performance comparison and do not affect training, early stopping, or parameter updates.

A fixed stratified observation set of 1,495 problems (20% of the training set) is constructed by proportional sampling based on the initial solvability distribution $P_0(x)$, preserving the original difficulty distribution of the full training set. At each checkpoint, the observation set is reevaluated with 32 generations under the corresponding policy to obtain updated solvability estimates $P_t(x)$ and characterize difficulty changes over training.

The static easy-to-hard curriculum first estimates the difficulty of training samples under the initial policy and ranks all training problems in descending order based on their initial solvability $P_0(x)$, where a higher solvability score indicates an easier sample. This ordering is fixed throughout training without further difficulty re-estimation. To assess how the fixed difficulty ordering becomes misaligned with the evolving policy, the next 500 samples in the training order are reevaluated at steps 1000, 2500, 3500, 4500, 5500, and 6500.

The static curriculum is independently repeated with three random seeds. For the multi-seed static curriculum comparison, we retrospectively report the best observed test accuracy along each completed training trajectory; these test values are used only for post-hoc performance comparison and do not affect training, early stopping, or parameter updates. Its best observed test performance is reported in Sec. 5.3, while the complete difficulty-misalignment statistics for the static curriculum and multi-seed results are provided in Appendix D.

4.3 Fixed-Medium and Staged Curricula

The staged curriculum progressively moves the active difficulty band through $B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5$. Based on the initial solvability estimate $P_0(x)$, five fixed diagnostic subsets are constructed sequentially for B_1 through B_5 , with 128 problems assigned to each subset. To ensure that the diagnostic subsets are mutually exclusive despite the overlapping success-rate intervals, any problem already assigned to an earlier diagnostic subset is excluded from subsequent subset construction. These diagnostic subsets are used only to monitor the staged curriculum and are excluded from training, leaving 6,833 training candidates.

Every 50 effective updates, the diagnostic subset corresponding to the current curriculum stage is evaluated, with eight responses generated for each problem. Stage transitions follow a fixed patience-based rule applied to the diagnostic trajectory. A later checkpoint is treated as a new best only when its diagnostic accuracy exceeds the current best by more than the empirical noise tolerance of 0.0107. Once a new best checkpoint is established, training continues for two additional diagnostic evaluations after 50 and 100 additional effective updates, respectively. If neither subsequent evaluation improves upon the current best by more than 0.0107, the stage is considered to have reached a plateau. Neither fluctuations within the noise tolerance nor performance decreases are treated as meaningful improvements.

If a new best is observed during the patience window, the 100-update patience criterion is restarted from that checkpoint. The best checkpoint retained for the current stage is then used to initialize training in the next difficulty band. The GSM8K test set is used only for generalization evaluation and is never involved in stage-transition or checkpoint-selection decisions. The full diagnostic trajectories, noise analysis, and selected checkpoints are reported in Appendix C.

Motivated by online data-selection strategies that prioritize samples with intermediate success rates, we introduce a fixed-medium curriculum for complementary comparison. Such samples are attractive in GRPO because they are more likely to produce mixed correct and incorrect responses and therefore informative group-relative learning signals. However, informative learning signals do not necessarily translate into the strongest final model performance. We therefore ask whether repeatedly selecting the medium-difficulty band is sufficient to maximize final model performance.

In the fixed-medium curriculum, the active difficulty band remains fixed at $B_3 = [0.375, 0.625]$ throughout optimization, while samples are reselected online according to the evolving policy. This curriculum preserves online difficulty estimation and online sample selection while keeping the active difficulty band fixed. Comparing fixed-medium curriculum with the staged curriculum allows us to examine

whether progressing to harder difficulty bands can provide additional benefits beyond repeatedly selecting samples currently in the medium-difficulty band.

Because the fixed-medium curriculum has no stage-transition criterion, we report the best observed test accuracy along each trajectory for comparison; these test values are not used to guide training or parameter updates. This retrospective best-observed value is used only for post-hoc comparison with the staged curriculum and does not affect training or staged checkpoint selection.

All multi-seed experiments use the same three random seeds. For reporting convenience, we refer to them as Seed 1, Seed 2, and Seed 3, corresponding to random seeds 42, 2026, and 100, respectively. For each staged run, stage transitions and checkpoint selection follow the same diagnostic-based criterion described above. We compare performance at the selected B_4 checkpoint with performance after 100 additional effective updates in B_5 . Multi-seed results are reported as mean $\pm$ standard deviation.

5 Results

5.1 State Dependence and Non-Monotonicity of Difficulty

The 1,495 tracked problems become easier on average, with the mean empirical success rate increasing from 0.6554 to 0.7975. At the same time, individual trajectories vary substantially. The mean training-period span is 0.2770, and 54.72% of samples have a training-period span of at least 0.25. At the final checkpoint, the mean absolute change from the initial empirical success rate P_0 is 0.1583, with 50.50% of samples changing by at least 0.125 relative to their initial values.

Non-monotonic changes are common. With $\delta = 0.125$, 53.38% of samples exhibit at least one direction reversal, and 36.79% exhibit at least two. The prevalence of direction reversals is threshold-dependent, rising to 87.02% for $\delta = 0.0625$ and falling to 4.95% for $\delta = 0.25$. Across the five difficulty bands, roughly 81–84% of samples that leave their initial band later re-enter it. Detailed sensitivity and re-entry statistics are reported in Appendix B.

Figure 3 summarizes both the population-level magnitude and non-monotonicity of changes in empirical solvability, while also providing representative problem-level trajectories.

Table 4 Core evidence for state-dependent difficulty dynamics.

Statistic	Result
Mean training-period span R_i	0.2770
Samples with $R_i \geq 0.25$	54.72%
At least one direction reversal, $\delta = 0.125$	53.38%
Final Spearman correlation with P_0	0.8092
Final B_3 retention	20.74%
Final B_3 Jaccard similarity	0.1497

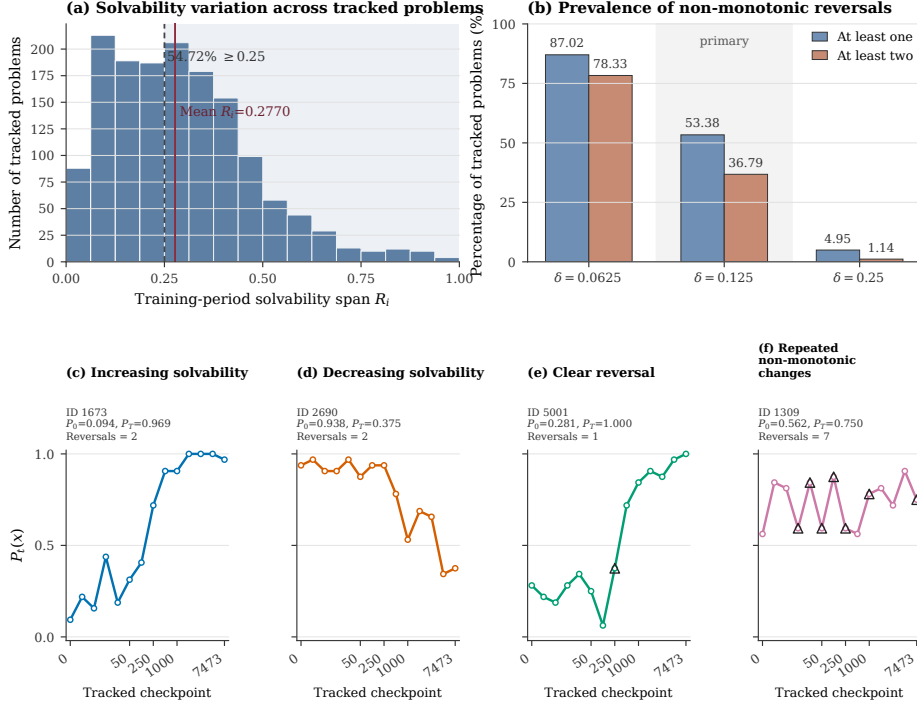


Fig. 3 Magnitude and non-monotonicity of changes in empirical solvability during GRPO training. (a) Distribution of the training-period solvability span R_i across the 1,495 tracked GSM8K problems. The dashed line marks $R_i = 0.25$; 54.72% of samples have a span of at least 0.25, and the mean span is 0.2770. (b) Fractions of tracked problems exhibiting at least one and at least two direction reversals under different minimum-change thresholds δ . At the primary threshold $\delta = 0.125$, 53.38% of samples exhibit at least one reversal, and 36.79% exhibit at least two. (c)–(f) Representative trajectories illustrating increasing solvability, decreasing solvability, a clear direction reversal, and repeated non-monotonic changes, respectively. The label above each panel reports the sample ID, initial solvability P_0 , final solvability P_T , and the number of direction reversals. For the representative trajectories, tracked checkpoints are shown at equal horizontal spacing for readability; tick labels indicate the corresponding training steps.

5.2 Global Ordering Persistence and Local Band-Membership Change

Global rank ordering remains largely preserved even when local band membership changes substantially. The Spearman rank correlation between P_0 and the final P_T remains 0.8092. In contrast, only 20.74% of samples initially in the medium-difficulty band $B_3 = [0.375, 0.625]$ remain there at the final checkpoint, while 65.00% of the samples in B_3 at the final checkpoint entered from outside the initial band. These results show that a relatively stable global difficulty ranking does not imply stable membership within a specific difficulty band. In other words, rank correlation can remain high even when the set of samples occupying a given difficulty band changes substantially.

Full rank and membership tables are given in Appendix B. An additional visualization of the relationship between the initial and final rank percentiles is provided in Appendix B.1, where the concentration near the diagonal offers a more direct view of the preserved global ordering.

Figure 4 visualizes this contrast directly: global rank consistency remains relatively high throughout training, whereas membership in the medium-difficulty band undergoes substantial turnover.

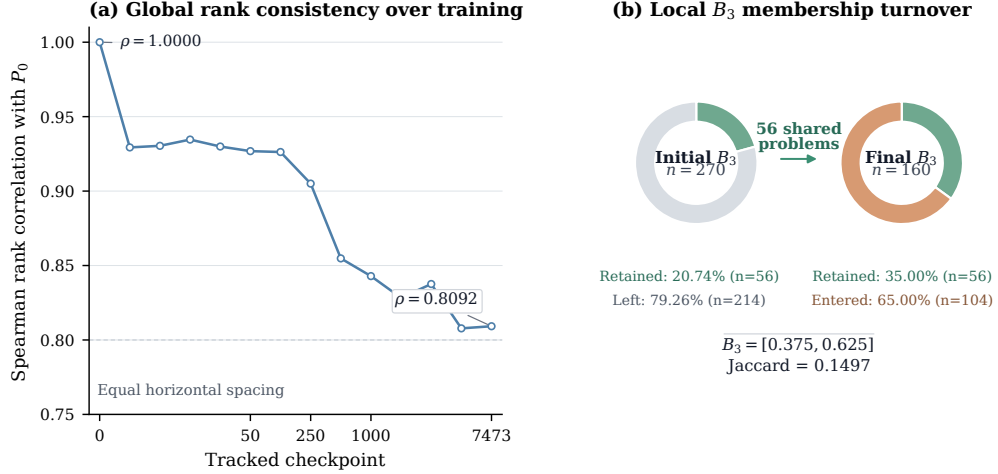


Fig. 4 Global rank consistency and membership turnover in the medium-difficulty band. (a) Spearman rank correlation between the initial solvability ranking and the ranking at each tracked checkpoint for the 1,495 tracked problems. Although rank consistency gradually decreases during training, it remains relatively high at the final checkpoint ($\rho = 0.8092$). Tracked checkpoints are shown at equal horizontal spacing for readability; tick labels indicate the corresponding training steps. (b) Membership turnover in the medium-difficulty band $B_3 = [0.375, 0.625]$. Of the 270 problems initially in B_3 , only 56 remain in the final B_3 , corresponding to 20.74% retention. The final B_3 contains 160 problems, of which 104 entered from outside the initial band, corresponding to 65% inflow. The Jaccard similarity between the initial and final B_3 sets is 0.1497.

5.3 Precomputed Difficulty Misalignment in Static Curriculum Training

The static curriculum provides evidence of difficulty misalignment over training. At step 5500, the next 500 scheduled problems have a planned mean success rate of 0.3986 under the initial policy but a current mean success rate of 0.5573, with a mean absolute deviation of 0.1863. Similar mismatches appear at earlier checkpoints, although their magnitudes are not monotonically related to test accuracy. We therefore interpret this analysis as evidence that precomputed difficulty estimates can become temporally misaligned with the evolving policy, rather than as evidence that such misalignment alone causes performance degradation. The complete table is reported in Appendix D.

Despite this misalignment, static curriculum training remains a competitive baseline. Across Seeds 1, 2, and 3, the best observed test accuracies are 0.6535, 0.6497,

and 0.6535, respectively, corresponding to 0.6522 ± 0.0022 . Therefore, the observed difficulty mismatch should not be interpreted as implying that a precomputed easy-to-hard curriculum becomes ineffective. Rather, these results indicate that the difficulty estimates underlying the fixed schedule can become outdated even when the resulting curriculum still achieves strong performance.

5.4 Performance Across Difficulty Bands

The two online curriculum conditions highlight two complementary effects. First, in one detailed experimental run, persistently selecting the medium-difficulty band does not yield the strongest observed performance. The fixed-medium curriculum reaches a best observed test accuracy of 0.6459 while keeping the difficulty band at $B_3 = [0.375, 0.625]$, whereas the staged curriculum reaches 0.6558 at the selected B_4 checkpoint. This comparison suggests that repeatedly selecting samples currently in the medium-difficulty band, despite their informative group-relative learning signals, is not sufficient to achieve the strongest observed performance; moving to a moderately harder difficulty band can provide additional gains.

Second, the benefit of progression does not persist as curriculum difficulty continues to increase. After the staged run transitions from B_4 to the harder B_5 band, test accuracy decreases from 0.6558 to 0.6391 after 50 additional effective updates and further to 0.6353 after 100 additional effective updates. The corresponding B_5 diagnostic accuracy also decreases from 0.5215 at entry to 0.4971 and 0.4512. The concurrent decline in both diagnostic and test performance suggests that this deterioration is not merely a transient fluctuation in test accuracy.

Taken together, this detailed run indicates that the performance benefit of moving to harder difficulty bands is non-monotonic: remaining in the medium-difficulty band can leave additional gains unrealized, whereas further progression into the harder difficulty band can offset the gains obtained from moderate progression.

Although the staged and fixed-medium curricula follow different training trajectories, the observed performance gap suggests that persistently selecting the medium-difficulty band is insufficient to achieve the strongest observed test performance. Sec. 5.5 examines the consistency of this pattern across random seeds.

Figure 5(a) visualizes the Seed 1 trajectory, including the improvement observed during B_4 and the subsequent decline after progression into B_5 .

Table 5 Single-run outcomes for the main curriculum conditions.

Setting	Evaluation point	Test accuracy
Fixed medium	Best observed checkpoint	0.6459
Staged curriculum	Selected B_4 checkpoint	0.6558
Staged curriculum	$B_5 + 50$ updates	0.6391
Staged curriculum	$B_5 + 100$ updates	0.6353

5.5 Robustness Across Random Seeds

The qualitative pattern described in Sec. 5.4 is observed across all three random seeds. As shown in Table 6, at the selected B_4 checkpoints, the staged curriculum achieves test accuracies of 0.6558, 0.6535, and 0.6679 for Seeds 1, 2, and 3, respectively, corresponding to 0.6591 ± 0.0077 . In comparison, the best observed accuracies along the fixed-medium trajectories are 0.6459, 0.6422, and 0.6467, yielding 0.6449 ± 0.0024 . Performance at the selected B_4 checkpoints exceeds the best observed fixed-medium performance in all three runs by 0.99, 1.13, and 2.12 percentage points, respectively.

For the standard GRPO baseline, the best observed test accuracies are 0.6459, 0.6414, and 0.6459 for Seeds 1, 2, and 3, respectively, corresponding to 0.6444 ± 0.0026 . Performance at the selected B_4 checkpoints of the staged curriculum is higher than the best observed standard GRPO performance in all three runs, by 0.99, 1.21, and 2.20 percentage points, respectively.

Table 6 Multi-seed performance of the curriculum conditions and baselines.

Setting	Seed 1	Seed 2	Seed 3	Mean $\pm$ Std.
Standard GRPO: best observed	0.6459	0.6414	0.6459	0.6444 ± 0.0026
Static curriculum: best observed	0.6535	0.6497	0.6535	0.6522 ± 0.0022
Fixed medium: best observed	0.6459	0.6422	0.6467	0.6449 ± 0.0024
Staged: selected B_4	0.6558	0.6535	0.6679	0.6591 ± 0.0077
Staged: $B_5 + 100$	0.6353	0.6404	0.6458	0.6405 ± 0.0053

Figure 5(b) summarizes the corresponding multi-seed comparison across the main training conditions. The static curriculum provides an additional strong baseline, with best observed test accuracies of 0.6535, 0.6497, and 0.6535 across the three runs, corresponding to 0.6522 ± 0.0022 . Nevertheless, performance at the selected B_4 checkpoints is higher than that of the static curriculum in all three runs by 0.23, 0.38, and 1.44 percentage points, respectively.

Further progression into B_5 shows the opposite pattern across all three seeds. After 100 additional effective updates in B_5 , test accuracies decrease to 0.6353, 0.6404, and 0.6458 for Seeds 1, 2, and 3, respectively, corresponding to 0.6405 ± 0.0053 . Relative to the selected B_4 checkpoints, these results represent decreases of 2.05, 1.31, and 2.21 percentage points, respectively.

Taken together, the multi-seed results support the same conclusion as the detailed analysis in Sec. 5.4: persistent selection of the medium-difficulty band is insufficient to attain the strongest observed test performance, while test performance at the selected B_4 checkpoints is consistently higher than the best observed results from standard GRPO, the fixed-medium curriculum, and the static curriculum across all three seeds. However, further progression into B_5 consistently reduces performance. These results are consistent with a non-monotonic relationship between scheduled difficulty and performance, rather than a monotonic benefit from continually increasing difficulty.

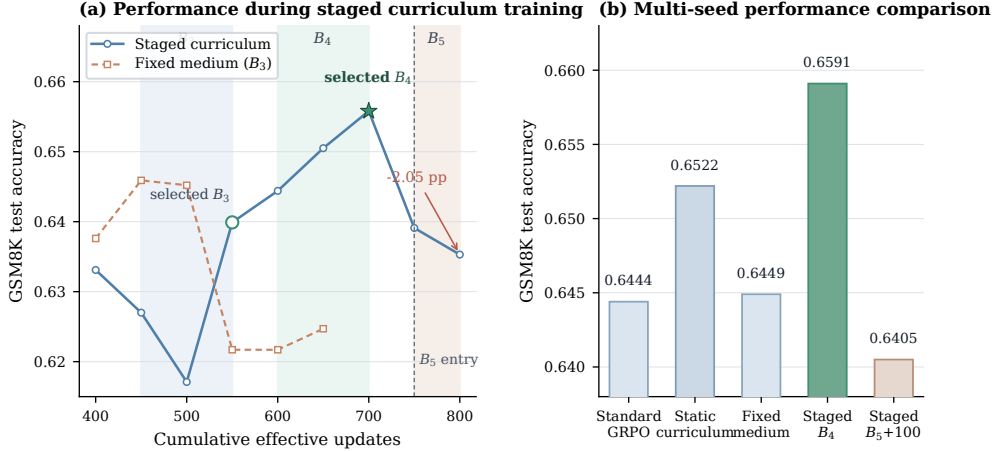


Fig. 5 Performance under the fixed-medium curriculum and the staged curriculum. (a) Seed 1 test-accuracy trajectories for the fixed-medium curriculum and the staged curriculum. The staged curriculum improves after progression into B_4 , reaching the selected B_4 checkpoint at 0.6558, whereas further progression into B_5 reduces accuracy to 0.6353 after 100 additional effective updates. Shaded areas indicate the active difficulty band and are shown only to visualize stage progression. (b) Multi-seed comparison of standard GRPO, static curriculum, fixed-medium training, the selected staged B_4 checkpoint, and performance after 100 additional effective updates in B_5 . The bar chart shows the mean GSM8K test accuracy for each training setting to provide a direct visual comparison of overall performance. Detailed per-seed values are reported in Table 6.

6 Discussion

Our results provide two main insights into difficulty-aware GRPO training. First, sample difficulty is policy-dependent and can change substantially and non-monotonically as training proceeds. Consequently, difficulty estimates obtained under the initial policy may become misaligned with the current policy, even when the global difficulty ranking remains relatively stable. The competitive performance of the static curriculum, however, shows that such misalignment does not by itself make a pre-computed curriculum ineffective. Rather, these observations motivate re-estimating sample difficulty under the current policy when sample selection depends on current solvability.

More importantly, our experiments distinguish informative learning signals from final model performance. Medium-difficulty samples are attractive for online selection because they are more likely to produce mixed correct and incorrect responses and therefore informative group-relative learning signals. However, the fixed-medium curriculum shows that persistently selecting such samples does not yield the strongest observed performance. The staged curriculum instead benefits from progressing to a moderately harder difficulty band, whereas further progression to a harder difficulty band consistently reduces performance across random seeds. This pattern suggests a non-monotonic relationship between curriculum difficulty and model performance: remaining at medium difficulty can leave additional gains unrealized, while increasing difficulty too far can offset the gains obtained from moderate progression.

These findings suggest that online sample selection and difficulty progression should be treated as two related but distinct components of difficulty-aware GRPO training. Online difficulty estimation determines which samples currently belong to a given difficulty band, whereas difficulty progression determines which difficulty band is used at different stages of training. Our results indicate that combining online re-estimation with staged difficulty progression can improve performance beyond persistent medium-difficulty selection, but progression toward harder samples should not be continued indefinitely. Effective curriculum design therefore requires not only identifying informative samples under the current policy, but also controlling how far the curriculum progresses in difficulty.

Several limitations remain. Our experiments focus on a single model family and the GSM8K benchmark, so the generality of the observed difficulty dynamics and curriculum pattern should be evaluated across broader models and reasoning tasks. Online solvability is estimated from only eight sampled responses, which introduces noise into difficulty estimation and sample selection. In addition, the staged curriculum relies on fixed diagnostic subsets and a predefined transition criterion, and it uses a smaller training candidate pool than the other conditions. Comparisons involving the staged curriculum should therefore be interpreted as complete-procedure comparisons rather than strict single-factor evaluations. Future work should examine these findings in broader settings, improve the accuracy and efficiency of online difficulty estimation, and develop more general adaptive mechanisms for determining when curriculum difficulty should increase and when further progression should stop.

7 Conclusion

We investigated how sample difficulty changes during GRPO training and whether persistently selecting medium-difficulty samples is sufficient for achieving strong final performance. The results show that sample difficulty changes with the evolving policy, supporting the use of current-policy difficulty estimates for online sample selection.

More importantly, the curriculum experiments reveal that persistent medium-difficulty selection does not achieve the strongest observed performance. Across three random seeds, progressing to a moderately harder difficulty band provides additional gains, whereas further progression to a harder difficulty band consistently reduces performance. These findings show that the benefit of increasing curriculum difficulty is non-monotonic. Overall, effective GRPO curriculum design should combine online sample selection with staged difficulty progression, moving beyond persistent medium-difficulty selection while avoiding excessive progression toward overly difficult samples.

Acknowledgements. The authors would like to thank ITREC DMU for providing the computational server resources used in this study.

Declarations

Funding

This research was funded by the basic scientific research project of the Department of Education of Liaoning Province, the People’s Republic of China, grant number 984230020.

Competing Interests

The authors declare no competing interests.

Data Availability

The GSM8K dataset used in this study is publicly available.

Code Availability

The code used in this study is available at <https://github.com/NHLLing/GRPO-B>.

Author Contribution

Ronghan Li contributed to the conceptualization, methodology, implementation, experiments, data analysis, and writing of the original manuscript. Fei Chen contributed to research supervision, discussion and refinement of the research ideas and methodology, and review and revision of the manuscript. Mingyu Lu contributed to review and revision of the manuscript. All authors reviewed and approved the final manuscript.

Appendix A Detailed Difficulty-Dynamics Metrics

This appendix gives the detailed metrics used to characterize local difficulty-band changes and non-monotonic trajectories.

For a difficulty band B with associated success-rate interval $[l, u]$, let

$$S_t(B) = \{x_i \mid P_t(x_i) \in [l, u]\}. \quad (\text{A1})$$

We characterize changes in band membership using the following complementary metrics.

$$\text{Retention}_t(B) = \frac{|S_0(B) \cap S_t(B)|}{|S_0(B)|}, \quad (\text{A2})$$

$$\text{Outflow}_t(B) = 1 - \text{Retention}_t(B), \quad (\text{A3})$$

$$\text{Inflow}_t(B) = \frac{|S_t(B) \setminus S_0(B)|}{|S_t(B)|}, \quad (\text{A4})$$

$$\text{Jaccard}_t(B) = \frac{|S_0(B) \cap S_t(B)|}{|S_0(B) \cup S_t(B)|}, \quad (\text{A5})$$

$$\text{Mismatch}_t(B) = \frac{|S_0(B) \triangle S_t(B)|}{N}. \quad (\text{A6})$$

Here N is the number of samples in the fixed observation set and $\triangle$ denotes symmetric difference.

Retention measures the fraction of samples initially belonging to band B that remain in the same band at training state t , while Outflow measures the complementary fraction that has left the band. Inflow measures the fraction of current band members that did not belong to B initially, capturing replacement by newly entering samples. Jaccard similarity quantifies the overlap between the initial and current band memberships. Mismatch measures the size of their symmetric difference relative to the full observation set of N samples, reflecting the overall extent of membership disagreement. Together, these metrics characterize sample persistence, departure, entry, set overlap, and overall membership change.

To distinguish one-directional movement from repeated back-and-forth changes, we use

$$TV_i = \sum_{k=1}^K |P_{t_k}(x_i) - P_{t_{k-1}}(x_i)|, \quad (\text{A7})$$

and define the extra round-trip variation as

$$E_i = TV_i - R_i. \quad (\text{A8})$$

Here, $E_i = 0$ for a monotonic trajectory, whereas $E_i > 0$ indicates variation beyond the observed success-rate span caused by back-and-forth changes. To define re-entry, let

$$I_{i,k}(B) = \mathbf{1}\{P_{t_k}(x_i) \in B\}. \quad (\text{A9})$$

indicate whether sample x_i belongs to band B at checkpoint t_k . A sample is counted as having ever left band B if there exist checkpoints $t_a < t_b$ such that $I_{i,a}(B) = 1$ and $I_{i,b}(B) = 0$. Among these samples, a sample is counted as re-entering band B if there exists a later checkpoint $t_c > t_b$ such that $I_{i,c}(B) = 1$. We define the re-entry rate as

$$\text{ReEntry}(B) = \frac{|\{x_i : \exists a < b < c, I_{i,a}(B) = 1, I_{i,b}(B) = 0, I_{i,c}(B) = 1\}|}{|\{x_i : \exists a < b, I_{i,a}(B) = 1, I_{i,b}(B) = 0\}|}. \quad (\text{A10})$$

Thus, the denominator consists only of samples that have been observed in band B and subsequently left it at least once, rather than all samples that ever belonged to the band or the full observation set.

Appendix B Full Difficulty-Dynamics Statistics

Table B1 Magnitude of sample difficulty changes over the complete GRPO trajectory.

Metric	Mean	Median	≥ 0.125	≥ 0.25	≥ 0.375
Training-period span R_i	0.2770	0.2500	79.87%	54.72%	28.96%
Maximum initial deviation $D_{i,\text{init}}$	0.2207	0.1875	73.65%	37.19%	15.92%

Table B2 Sensitivity of the direction-reversal rate to the minimum-change threshold.

Threshold δ	At least one reversal	At least two reversals
0.0625	87.02%	78.33%
0.125	53.38%	36.79%
0.25	4.95%	1.14%

Table B3 Rank consistency between the initial and current empirical success rates.

Training step	Spearman's ρ	Kendall's τ_b
100	0.9262	0.7988
500	0.8548	0.7050
1000	0.8429	0.6932
7473	0.8092	0.6545

Table B4 Membership changes in the medium-difficulty band $B_3 = [0.375, 0.625]$.

Step	$ S_t $	Retention	Outflow	Inflow	Jaccard	Mismatch
100	239	58.52%	41.48%	33.89%	0.4501	12.91%
500	188	29.63%	70.37%	57.45%	0.2116	19.93%
1000	187	27.04%	72.96%	60.96%	0.1901	20.80%
7473	160	20.74%	79.26%	65.00%	0.1497	21.27%

Table B5 Re-entry rates for the five difficulty bands.

Difficulty band	Ever in band	Ever left	Re-entered	Re-entry rate
B_1	508	498	408	81.93%
B_2	350	344	289	84.01%
B_3	270	267	216	80.90%
B_4	232	230	191	83.04%
B_5	215	211	174	82.46%

B.1 Initial–Final Rank Consistency Visualization

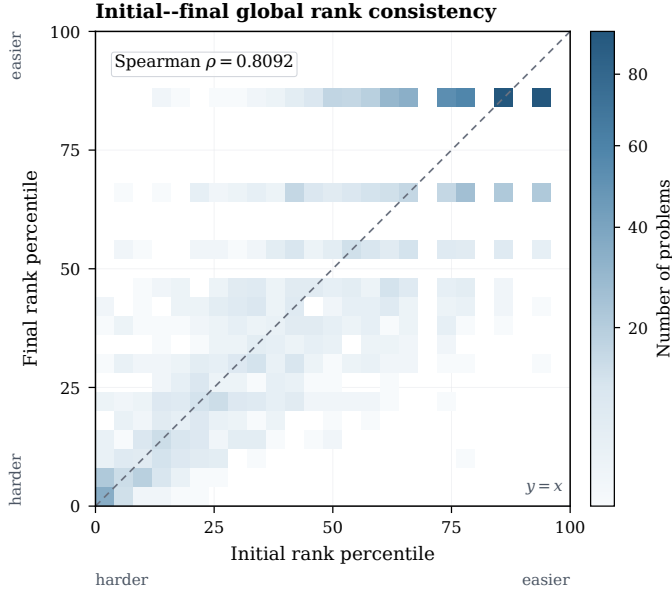


Fig. B1 Initial–final global rank consistency for the 1,495 tracked GSM8K problems. The axes show solvability-rank percentiles at the initial and final checkpoints, with 0 indicating harder problems and 100 indicating easier problems; color denotes the number of problems in each bin. The concentration near the dashed $y = x$ line is consistent with the final Spearman rank correlation of $\rho = 0.8092$.

Appendix C Diagnostic Trajectories and Stage Selection

Five fixed diagnostic subsets of 128 problems are constructed for the staged curriculum, one for each difficulty band B_1 – B_5 . The subsets are constructed sequentially from B_1 through B_5 , and any problem already assigned to an earlier diagnostic subset is excluded from subsequent subset construction. These subsets are used only to monitor training progress and are excluded from the training pool, so they do not contribute to parameter updates. During each curriculum stage, the diagnostic subset corresponding to the active difficulty band is evaluated every 50 effective updates, with eight responses generated for each problem.

To estimate the scale of diagnostic evaluation noise, the same model and diagnostic subset are reevaluated ten times using different sampling seeds. The resulting accuracies have a mean of 0.5226, a standard deviation of 0.0042, a minimum of 0.5156, and a maximum of 0.5283. The empirical 95th percentile of pairwise absolute accuracy differences is approximately 0.0107. We use this value as an empirical noise tolerance when determining whether a later diagnostic evaluation represents a meaningful improvement over the current best checkpoint.

Stage transitions follow a fixed patience-based criterion. Within each stage, a later checkpoint is treated as a new best checkpoint only when its diagnostic accuracy exceeds the current best by more than the empirical noise tolerance of 0.0107. Once a new best checkpoint is established, training continues for two additional diagnostic evaluations after 50 and 100 additional effective updates, respectively. If neither subsequent evaluation improves upon the current best by more than 0.0107, the stage is considered to have reached a plateau. The retained best checkpoint is then used to initialize training in the next difficulty band. If a new best checkpoint is observed during this patience window, the 100-update patience criterion is restarted from that checkpoint. When multiple checkpoints lie within the same high-performing plateau and their diagnostic-accuracy differences remain within the empirical noise tolerance, the earlier checkpoint is retained.

Table C6 summarizes the checkpoints selected for the three random seeds.

Table C6 Selected stage checkpoints for the three random seeds. Diagnostic accuracy is measured on the fixed subset corresponding to the active difficulty band.

Seed	Stage	Success-rate interval	Selected local update	Diagnostic accuracy
1	B_1	[0.625, 0.875]	150	0.8350
1	B_2	[0.500, 0.750]	250	0.8174
1	B_3	[0.375, 0.625]	150	0.7432
1	B_4	[0.250, 0.500]	150	0.6436
2	B_1	[0.625, 0.875]	300	0.9111
2	B_2	[0.500, 0.750]	150	0.8232
2	B_3	[0.375, 0.625]	200	0.7715
2	B_4	[0.250, 0.500]	100	0.6172
3	B_1	[0.625, 0.875]	250	0.9209
3	B_2	[0.500, 0.750]	300	0.8438
3	B_3	[0.375, 0.625]	50	0.7666
3	B_4	[0.250, 0.500]	100	0.6807

Figure C1 provides a visual summary of the diagnostic trajectories underlying the stage-selection decisions reported in Tables C6–C9.

The complete diagnostic trajectories used for stage selection are reported below. Bold values indicate the checkpoints retained for initialization of the subsequent stage.

No post-entry B_5 checkpoint is selected for any of the three seeds. For Seed 1, the B_5 diagnostic accuracy decreases from 0.5215 at entry to 0.4971 after 50 effective updates and 0.4512 after 100 effective updates. For Seed 2, the corresponding diagnostic accuracies are 0.5127, 0.5086, and 0.4521, while for Seed 3 they are 0.5745, 0.5713, and 0.5273. Thus, no run shows a diagnostic improvement after progression into B_5 , and the selected B_4 checkpoint remains the final selected checkpoint for each seed.

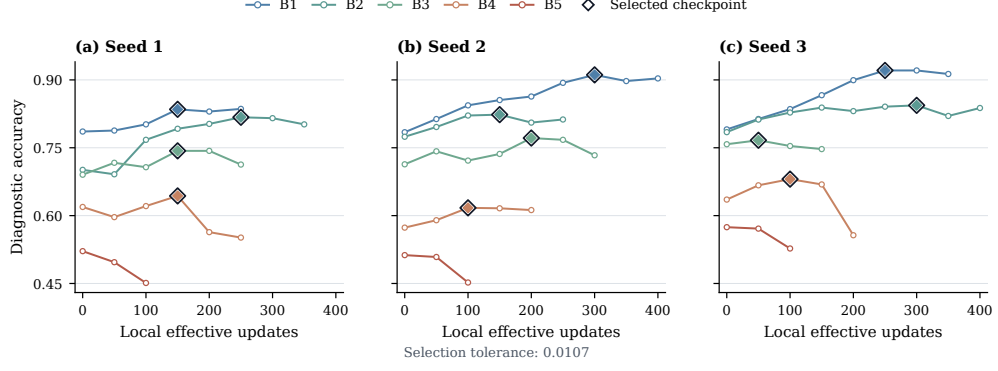


Fig. C1 Diagnostic trajectories underlying stage selection across the three random seeds. Each curve shows diagnostic accuracy as a function of local effective updates within the corresponding difficulty band. Highlighted markers denote checkpoints retained under the predefined patience-based criterion. A later checkpoint is considered a new best only when its diagnostic accuracy exceeds the current best by more than the empirical noise tolerance of 0.0107; after a new best is established, two additional evaluations are performed after 50 and 100 additional effective updates, respectively. No post-entry B_5 checkpoint is selected for any of the three seeds.

Table C7 Diagnostic trajectories for Seed 1. Columns denote local effective updates within each difficulty band.

Stage	0	50	100	150	200	250	300	350	400
B_1	0.7859	0.7881	0.8018	0.8350	0.8301	0.8359	—	—	—
B_2	0.7012	0.6914	0.7676	0.7920	0.8027	0.8174	0.8154	0.8017	—
B_3	0.6904	0.7168	0.7070	0.7432	0.7432	0.7129	—	—	—
B_4	0.6191	0.5967	0.6211	0.6436	0.5635	0.5515	—	—	—
B_5	0.5215	0.4971	0.4512	—	—	—	—	—	—

Table C8 Diagnostic trajectories for Seed 2. Columns denote local effective updates within each difficulty band.

Stage	0	50	100	150	200	250	300	350	400
B_1	0.7844	0.8135	0.8438	0.8555	0.8633	0.8936	0.9111	0.8975	0.9033
B_2	0.7743	0.7959	0.8213	0.8232	0.8057	0.8125	—	—	—
B_3	0.7135	0.7422	0.7217	0.7363	0.7715	0.7676	0.7334	—	—
B_4	0.5735	0.5898	0.6172	0.6162	0.6123	—	—	—	—
B_5	0.5127	0.5086	0.4521	—	—	—	—	—	—

Table C9 Diagnostic trajectories for Seed 3. Columns denote local effective updates within each difficulty band.

Stage	0	50	100	150	200	250	300	350	400
B_1	0.7906	0.8135	0.8354	0.8662	0.8994	0.9209	0.9209	0.9131	–
B_2	0.7845	0.8123	0.8279	0.8388	0.8311	0.8408	0.8438	0.8203	0.8379
B_3	0.7578	0.7666	0.7539	0.7471	–	–	–	–	–
B_4	0.6355	0.6670	0.6807	0.6689	0.5566	–	–	–	–
B_5	0.5745	0.5713	0.5273	–	–	–	–	–	–

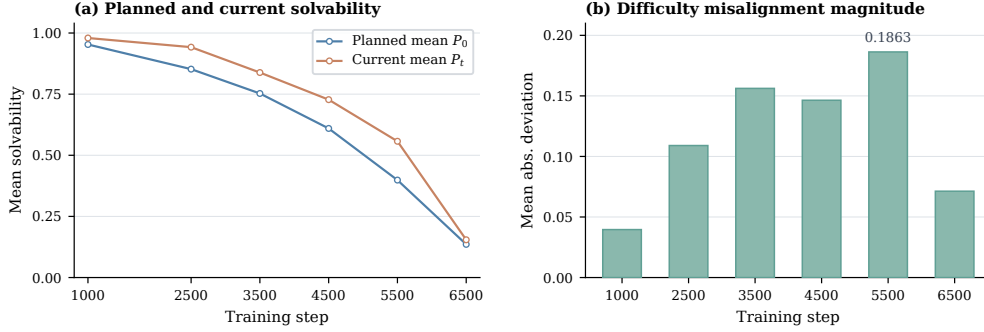


Fig. D1 Difficulty misalignment in the static curriculum. (a) Mean solvability specified by the precomputed curriculum schedule and mean solvability under the current policy at each evaluation point. (b) Mean absolute deviation between the planned and current mean solvability estimates at the evaluated training steps.

Appendix D Additional Training Trajectories

Table D10 reports the mismatch between the initial difficulty estimates used to construct the static curriculum and the corresponding difficulty estimates under the current policy at each evaluation point.

Table D10 Difficulty mismatch between the static curriculum plan and the current policy at each evaluation point.

Training step	Planned mean P_0	Current mean P_t	Mean $ P_t - P_0 $
1000	0.9531	0.9796	0.0396
2500	0.8523	0.9422	0.1090
3500	0.7528	0.8384	0.1562
4500	0.6098	0.7275	0.1464
5500	0.3986	0.5573	0.1863
6500	0.1357	0.1544	0.0713

Figure D1 visualizes the difficulty misalignment summarized in Table D10.

Despite the observed mismatch in precomputed difficulty estimates, the static curriculum remains a competitive baseline across random seeds. Table D11 reports its best observed test performance across random seeds.

Table D11 Best observed test performance of the static curriculum across random seeds.

Seed 1	Seed 2	Seed 3	Mean $\pm$ Std.
0.6535	0.6497	0.6535	0.6522 ± 0.0022

Table D12 shows the test trajectory of the Seed 1 staged curriculum across B_3 , B_4 , and B_5 .

Table D12 Seed 1 test trajectory for the staged curriculum.

Effective updates	Test accuracy	Curriculum state
400	0.6331	Selected B_2 checkpoint
450	0.6270	B_3
500	0.6171	B_3
550	0.6399	Selected B_3 checkpoint
600	0.6444	B_4
650	0.6505	B_4
700	0.6558	Selected B_4 checkpoint
750	0.6391	$B_5 + 50$ updates
800	0.6353	$B_5 + 100$ updates

The complete test trajectories for the fixed-medium curriculum are reported in Table D13. In all three runs, the active difficulty band is fixed at $B_3 = [0.375, 0.625]$, while samples are reselected online according to the evolving policy. Bold values denote the best observed test accuracy along each trajectory. These best observed values are retrospective maxima along the completed trajectories; test accuracy is not used to guide training, early stopping, parameter updates, or sample selection.

The best observed fixed-medium test accuracies are 0.6459 at 450 effective updates for Seed 1, 0.6422 at 700 effective updates for Seed 2, and 0.6467 at 650 effective updates for Seed 3. These values correspond to the fixed-medium results summarized in the multi-seed comparison in the main text.

Table D13 Test trajectories under the fixed-medium curriculum for the three random seeds.

Effective updates	Seed 1	Seed 2	Seed 3
0	0.6141	0.6141	0.6141
50	0.6181	0.6179	0.6133
100	0.6195	0.6194	0.6118
150	0.6223	0.6209	0.6171
200	0.6274	0.6232	0.6224
250	0.6344	0.6270	0.6103
300	0.6406	0.6308	0.6368
350	0.6384	0.6384	0.6384
400	0.6376	0.6338	0.6096
450	0.6459	0.6224	0.6171
500	0.6452	0.6376	0.6247
550	0.6217	0.6331	0.6293
600	0.6217	0.6300	0.6399
650	0.6247	0.6308	0.6467
700	–	0.6422	0.6437
750	–	0.6399	0.6422
800	–	0.6414	0.6141
850	–	0.6399	0.6141
900	–	0.6338	0.6331
950	–	–	0.6262
1000	–	–	0.6328

References

- [1] Bae S, Hong J, Lee MY, et al (2026) Online difficulty filtering for reasoning oriented reinforcement learning. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Rabat, Morocco, pp 700–719, <https://doi.org/10.18653/v1/2026.eacl-long.30>, URL <https://aclanthology.org/2026.eacl-long.30/>
- [2] Bengio Y, Louradour J, Collobert R, et al (2009) Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. Association for Computing Machinery, pp 41–48, <https://doi.org/10.1145/1553374.1553380>
- [3] Cobbe K, Kosaraju V, Bavarian M, et al (2021) Training verifiers to solve math word problems. arXiv preprint URL <https://arxiv.org/abs/2110.14168>, arXiv:2110.14168
- [4] Gao Z, Kim J, Sun W, et al (2026) Prompt curriculum learning for efficient LLM post-training. In: International Conference on Learning Representations, URL <https://arxiv.org/abs/2510.01135>

- [5] Guo D, Yang D, Zhang H, et al (2025) DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645:633–638. <https://doi.org/10.1038/s41586-025-09422-z>, URL <https://doi.org/10.1038/s41586-025-09422-z>
- [6] Hu EJ, Shen Y, Wallis P, et al (2022) LoRA: Low-rank adaptation of large language models. In: *International Conference on Learning Representations*, URL <https://arxiv.org/abs/2106.09685>
- [7] Le TLV, Jeon M, Vu K, et al (2026) No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. In: *International Conference on Learning Representations*, URL <https://arxiv.org/abs/2509.21880>, arXiv:2509.21880
- [8] Li R, Huang H, Wei F, et al (2026) AdaCuRL: Adaptive curriculum reinforcement learning with invalid sample mitigation and historical revisiting. *Proceedings of the AAAI Conference on Artificial Intelligence* 40(27):23123–23131. <https://doi.org/10.1609/aaai.v40i27.39479>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/39479>
- [9] Li X, Zou H, Liu P (2025) LIMR: Less is more for RL scaling. *arXiv preprint* URL <https://arxiv.org/abs/2502.11886>, arXiv:2502.11886
- [10] Liu Z, Chen C, Li W, et al (2025) Understanding R1-Zero-Like training: A critical perspective. *arXiv preprint* URL <https://arxiv.org/abs/2503.20783>, arXiv:2503.20783
- [11] Mao Y, Qu Y, Wang Q, et al (2026) Dynamics-predictive sampling for active RL finetuning of large reasoning models. In: *International Conference on Learning Representations*, URL <https://arxiv.org/abs/2603.10887>
- [12] Parashar S, Gui S, Li X, et al (2026) Curriculum reinforcement learning from easy to hard tasks improves LLM reasoning. In: *International Conference on Learning Representations*, URL <https://arxiv.org/abs/2506.06632>
- [13] Qu Y, Wang Q, Mao Y, et al (2026) Can prompt difficulty be online predicted for accelerating RL finetuning of reasoning models? In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp 1240–1250, <https://doi.org/10.1145/3770854.3780263>, URL <https://arxiv.org/abs/2507.04632>, arXiv:2507.04632
- [14] Shao Z, Wang P, Zhu Q, et al (2024) DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint* URL <https://arxiv.org/abs/2402.03300>, arXiv:2402.03300
- [15] Shen Q, Chen D, Huang Y, et al (2026) BOTS: A unified framework for bayesian online task selection in LLM reinforcement finetuning. In: *International Conference on Learning Representations*, URL <https://arxiv.org/abs/2510.26374>

- [16] Shi T, Wu Y, Song L, et al (2026) Efficient reinforcement finetuning via adaptive curriculum learning. Transactions on Machine Learning Research URL <https://arxiv.org/abs/2504.05520>, arXiv:2504.05520
- [17] Sun Y, Shen J, Wang Y, et al (2025) Improving data efficiency for LLM reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. In: Advances in Neural Information Processing Systems, URL https://papers.nips.cc/paper_files/paper/2025/hash/fd54731cc635e2ed6033e8a0e193eddf-Abstract-Conference.html
- [18] Tan H, Wu X, Wu S, et al (2025) Understanding data influence in reinforcement finetuning. In: Advances in Neural Information Processing Systems, <https://doi.org/10.52202/085713-1294>, URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3755a02b1035fbadd5f93a022170e46f-Abstract-Conference.html
- [19] Tang X, Zhang Z, Liu Y, et al (2026) Towards high data efficiency in reinforcement learning with verifiable reward. In: International Conference on Learning Representations, URL <https://arxiv.org/abs/2509.01321>
- [20] Wang H, Hao Z, Luo J, et al (2026) Scheduling your LLM reinforcement learning with reasoning trees. In: International Conference on Learning Representations, URL <https://arxiv.org/abs/2510.24832>
- [21] Xin H, Ren ZZ, Song J, et al (2025) DeepSeek-Prover-V1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In: International Conference on Learning Representations, URL <https://arxiv.org/abs/2408.08152>
- [22] Xu H, Chen S, Qiu R, et al (2026) Prune as you generate: Online rollout pruning for faster and better RLVR. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pp 13876–13893, <https://doi.org/10.18653/v1/2026.acl-long.632>
- [23] Yang A, Yang B, Zhang B, et al (2024) Qwen2.5 technical report. arXiv preprint URL <https://arxiv.org/abs/2412.15115>, arXiv:2412.15115
- [24] Yu Q, Zhang Z, Zhu R, et al (2025) DAPO: An open-source LLM reinforcement learning system at scale. In: Advances in Neural Information Processing Systems, <https://doi.org/10.52202/085713-3775>, URL <https://arxiv.org/abs/2503.14476>, arXiv:2503.14476
- [25] Zeng Y, Sun Z, Ji B, et al (2026) CurES: From gradient analysis to efficient curriculum learning for reasoning LLMs. In: International Conference on Learning Representations, URL <https://arxiv.org/abs/2510.01037>

- [26] Zhao Y, Xiong K, Ding X, et al (2025) UFO-RL: Uncertainty-focused optimization for efficient reinforcement learning data selection. In: Advances in Neural Information Processing Systems, URL <https://arxiv.org/abs/2505.12457>
- [27] Zheng C, Liu S, Li M, et al (2025) Group sequence policy optimization. arXiv preprint URL <https://arxiv.org/abs/2507.18071>, [arXiv:2507.18071](#)
- [28] Zhou X, Zhu B, Zhang H, et al (2026) Efficient RLVR training via weighted mutual information data selection. arXiv preprint URL <https://arxiv.org/abs/2603.01907>, [arXiv:2603.01907](#)