MIGE: MULTI-LEVEL INFLUENCE-GUIDED EXPLORATION IN REINFORCEMENT LEARNING FOR LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) substantially improves reasoning in large language models, but policies can collapse to a small set of familiar solution paths. Existing exploration methods mainly preserve token entropy or reward trajectory diversity. Yet high token uncertainty need not mark a consequential reasoning choice, and textually diverse trajectories may follow the same underlying strategy. To address this limitation, we propose Multi-Level Influence-Guided Exploration (MIGE), which guides exploration according to how decisions affect subsequent reasoning. At the strategy level, MIGE computes the log-likelihood of complete executions under alternative strategy contexts to reward strategies with distinct downstream effects. At the step level, it resamples continuations at candidate token positions and rewards forks that change the next intermediate conclusion. Across six mathematical and scientific reasoning benchmarks and two model scales, MIGE achieves the best average performance among standard RLVR and representative exploration baselines while largely preserving successful-solution coverage at Pass@k. Controlled analyses show that influence scores distinguish strategy changes from surface rewrites and locate reasoning forks more reliably than output-distance and uncertainty baselines, supporting downstream influence as an effective signal for meaningful exploration.

1 INTRODUCTION

Reinforcement learning (RL) has become a powerful paradigm for post-training large language models (LLMs), improving their reasoning capabilities by reinforcing desirable outputs while suppressing undesirable ones (Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). A fundamental challenge in RL is balancing exploration and exploitation: while excessive exploitation may accelerate short-term convergence, it can also trap the policy in local optima and hinder further improvement. In LLM reinforcement learning, insufficient exploration often manifests as "entropy collapse", where the distribution over generated trajectories becomes increasingly concentrated, and "reasoning-mode collapse", where model outputs converge to a limited set of homogeneous reasoning patterns. Both behaviors have been observed in recent analyses of RL with verifiable rewards (RLVR) (Cui et al., 2025; Zeng et al., 2025; Wang et al., 2025b). This issue is acute in group-based RL, where homogeneous rollouts weaken the learning signal, especially when all sampled responses receive the same reward (Shao et al., 2024; Yu et al., 2025; Zheng et al., 2025a).

To prevent reasoning-mode collapse and sustain continual improvement in LLM reinforcement learning, we seek to encourage the policy to discover novel reasoning paths that remain effective at solving the task. Achieving this goal presents three challenges. First, meaningful exploration must be distinguished from superficial output variation, since lexically different responses can instantiate the same underlying reasoning strategy (Li et al., 2026b). Second, exploration must be coordinated across complementary levels: the policy should search globally over distinct solution strategies while also exploring local reasoning alternatives that arise during the execution of each strategy, since focusing on either level alone can overlook meaningful reasoning modes at the other. Third, exploration incentives require precise credit assignment so that they promote the decisions responsible for useful novelty without rewarding irrelevant variation or low-quality trajectories and destabilizing learning (Huang et al., 2025; Hu et al., 2026; Li et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches primarily encourage exploration through mechanisms such as entropy regularization and diversity rewards. While effective at mitigating entropy collapse and increasing output diversity, these methods do not necessarily promote meaningful exploration of the reasoning space (Cui et al., 2025; Wang et al., 2025b; Hu et al., 2025; Chen et al., 2025). In particular, token-level uncertainty does not necessarily indicate a genuine reasoning fork, as it may arise from alternative lexical realizations that lead to essentially the same reasoning path (Wang et al., 2025a). Similarly, textual diversity can be driven by variations in wording or linguistic style, without reflecting substantive differences in the underlying reasoning strategies (Kuhn et al., 2023; Li et al., 2025).

To this end, we propose **M**ulti-Level **I**nfluence-**G**uided **E**xploration (MIGE), which evaluates decisions by their influence on subsequent reasoning rather than by token uncertainty or output diversity. Figure 1 contrasts this influence-guided perspective with conventional diversity- and entropy-based exploration. We formulate each reasoning trajectory as comprising a high-level strategy and a sequence of reasoning steps, each producing an intermediate conclusion. This representation enables two complementary mechanisms. At the high level, our strategy influence reward compares the likelihood of sampled continuations across paired strategy contexts and aggregates the resulting distances over nearest neighbors, rewarding novel strategies that have different downstream effects. At the low level, our fork influence reward compares the likelihood of each next conclusion across resampled reasoning-branch contexts to identify local forks that influence the next intermediate conclusion. Finally, we combine task and exploration advantages through multi-level credit assignment to balance solution quality with both global and local exploration.

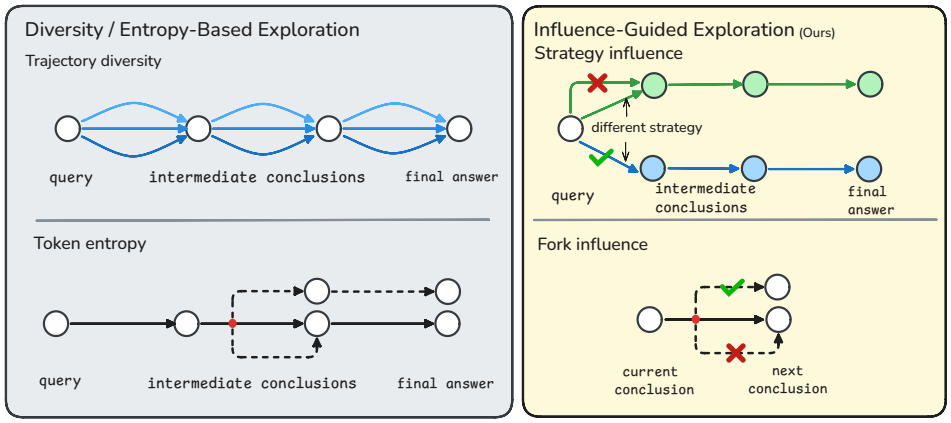


Figure 1: Overview of MIGE versus diversity- and entropy-based exploration. MIGE rewards global strategies that alter downstream trajectories and local branches that change the next conclusion.

Our main contributions are summarized as follows:

- **Influence-based reasoning exploration.** We formulate meaningful exploration by measuring how global strategies and local reasoning choices affect subsequent reasoning, rather than relying on token-level uncertainty or output-level diversity.
- **Multi-level exploration framework.** MIGE jointly rewards novel global strategies and influential local forks while assigning credit to the corresponding decisions.
- **Comprehensive evaluation.** We evaluate MIGE across reasoning benchmarks and analyze strategy novelty, fork localization, component contributions, and robustness.

2 Related Work

Entropy-based exploration. A major line of work promotes exploration in LLM reinforcement learning by preserving policy entropy. Recent studies show that policy entropy often decreases rapidly during RL training, leading to reduced exploration and motivating entropy-preserving optimization (Cui et al., 2025). Further analyses reveal that entropy is highly non-uniform across

reasoning trajectories: a small fraction of high-entropy tokens often coincide with pivotal reasoning decisions or reflective behaviors (Wang et al., 2025b; Cheng et al., 2025). This observation has inspired methods that restrict policy updates to high-entropy positions, augment advantages with entropy, or branch from high-uncertainty decisions (Wang et al., 2025b; Cheng et al., 2025; Zheng et al., 2025b). Other works use entropy-derived weights to shape token- and sequence-level advantages (Tan et al., 2025). ROSE instead combines generation entropy with embedding-based semantic divergence and uses the resulting semantic entropy to select branching positions that promote semantically diverse reasoning paths (Zhao et al., 2026). These methods progressively move from local predictive uncertainty toward sequence- and semantic-level uncertainty. However, entropy remains an indirect signal of reasoning exploration: even when high entropy correlates with critical reasoning forks, alternative tokens or sequences may still follow the same strategy. Thus, maintaining uncertainty does not guarantee distinct reasoning modes.

Diversity-based exploration. Complementary approaches directly encourage diversity among sampled reasoning trajectories. Some methods promote sequence-level exploration through textual or structural diversity rewards, such as n-gram dissimilarity or differences in mathematical expressions (Hu et al., 2025). To reduce sensitivity to superficial wording changes, subsequent methods measure redundancy in representation space: DRA-GRPO reweights trajectories according to embedding-based similarity, while SetPO evaluates the contribution of each trajectory to the diversity of the entire rollout set (Chen et al., 2025; Li et al., 2026a). Other methods explicitly increase trajectory-level exploration through structured rollouts or joint regularization; for example, LATR encourages divergent tree-based rollouts, while DSDR combines global trajectory diversity with local entropy preservation (Xing et al., 2025; Wan et al., 2026). Recent work also considers the quality–diversity trade-off. ExTra rewards novel trajectories only when they are correct and couples this objective with entropy-guided regeneration, while related approaches jointly optimize solution quality and semantic diversity (Hu et al., 2026; Li et al., 2025). Nevertheless, most trajectory-diversity objectives still rely on output-space proxies, including lexical overlap, equation patterns, or whole-response embeddings. Such measures can reduce redundant rollouts, but they do not explicitly determine whether different outputs instantiate different underlying reasoning strategies.

Overall, existing methods largely rely on uncertainty or output diversity as proxies for exploration. Because these signals do not reveal whether a decision changes subsequent reasoning, they provide limited guidance for identifying and crediting meaningful exploration at global and local levels.

3 METHODS

Motivation. Existing exploration signals measure whether next tokens are uncertain or sampled outputs look different, but not whether an alternative decision changes the subsequent reasoning. Two differently worded strategies may induce the same derivation, whereas a small local change may redirect the next conclusion. This motivates using downstream influence as evidence of meaningful novelty. MIGE estimates this influence at two scales: how a global strategy shapes the complete solution and how a local thought branch affects the next intermediate conclusion.

Overview. MIGE has four components. First, a *reasoning trajectory formulation* decomposes a standard chain-of-thought solution into its global strategy, stepwise reasoning, and intermediate conclusions. Second, *high-level strategy influence* measures how alternative strategy contexts change the distribution of complete executions and uses a k-nearest-neighbor score to reward globally novel strategies. Third, *low-level fork influence* uses entropy to locate candidate decisions, then measures how resampled branches change the next-conclusion distribution. Finally, *multi-level influence-guided policy optimization* combines task and the two influence-based exploration advantages, assigning them only to the strategy tokens and selected candidate fork tokens, respectively.

3.1 REASONING TRAJECTORY FORMULATION

We model reasoning as high-level planning followed by stepwise execution, where each step builds on prior conclusions. Given a problem x, a reasoning trajectory is represented as

$$y = s \circ u^{(1)} \circ z^{(1)} \circ \cdots \circ u^{(K)} \circ z^{(K)}, \tag{1}$$

where s is the high-level strategy, $u^{(k)}$ is the reasoning segment at step k, $z^{(k)}$ is its resulting intermediate conclusion, and $\circ$ denotes concatenation. In practice, we use a simple prompt to elicit this

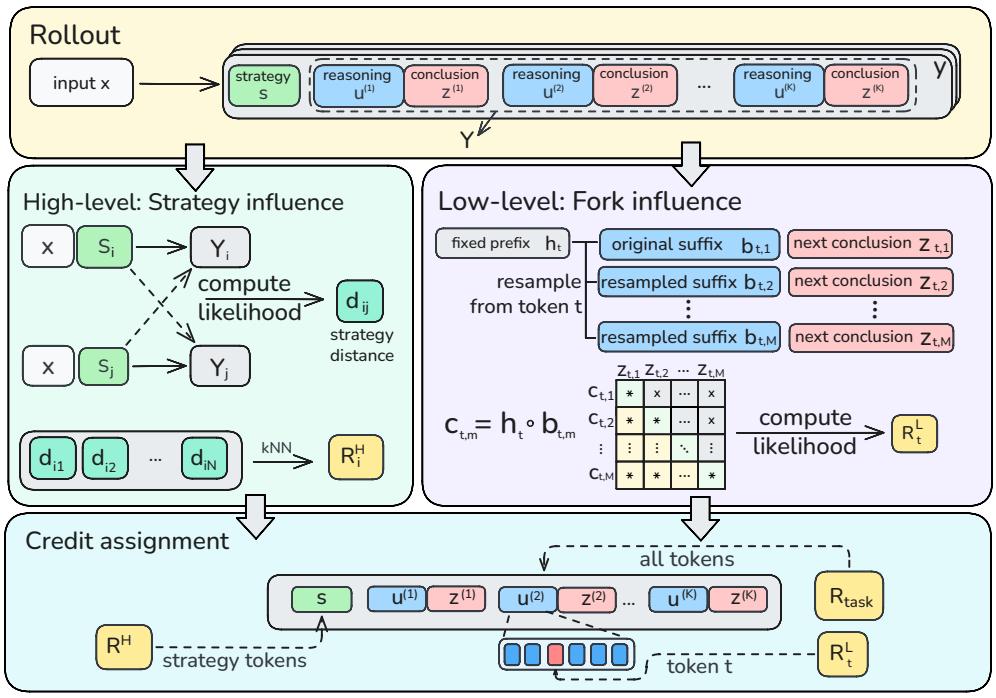


Figure 2: Overview of MIGE. Each rollout is decomposed into a global strategy and stepwise reasoning–conclusion pairs. Strategy influence measures changes in full-execution likelihood across strategy contexts, while fork influence measures changes in next-conclusion likelihood across resampled local contexts. Each reward is assigned to its corresponding decision tokens.

structure and a rule-based parser to extract the strategy and each reasoning–conclusion pair. Both levels then instantiate the same influence principle: a reasoning decision provides a meaningful exploration signal only when replacing it changes the distribution of downstream reasoning. They differ only in the scope of the decision and the horizon over which its influence is measured. At the high level, the decision and outcome are the global strategy and complete execution; at the low level, they are a local reasoning branch and its next intermediate conclusion.

3.2 High-Level Strategy Influence

At the high level, we distinguish strategy changes from surface reformulations by testing whether two strategies favor different complete executions. For each trajectory i, let $Y_i \sim \pi(\cdot \mid x, s_i)$ denote the execution following strategy s_i. For trajectories i and j, we compute the log-likelihood of each execution under both strategy contexts and define their non-negative strategy-influence distance as

$$d^H_{i,j} = \max\left\{0, \frac{1}{2}\left(\frac{1}{|Y_i|}\log\frac{\pi(Y_i \mid x, s_i)}{\pi(Y_i \mid x, s_j)} + \frac{1}{|Y_j|}\log\frac{\pi(Y_j \mid x, s_j)}{\pi(Y_j \mid x, s_i)}\right)\right\}. \tag{2}$$

If exchanging s_i and s_j leaves both executions similarly likely, $d^H_{i,j}$ is near zero; if each execution is supported primarily by its own strategy, the distance is large. For trajectory i, we aggregate these distances into its strategy influence reward:

$$R^H_i = \frac{1}{k}\sum_{j \in \mathcal{N}_k(i)} d^H_{i,j}. \tag{3}$$

Here $\mathcal{N}_k(i)$ contains the indices of the k trajectories with the smallest strategy-influence distance to trajectory i, excluding i itself. Averaging over these nearest neighbors rewards strategy s_i only when it remains distinct from its most similar peers. Compared with a group-wide average, this local comparison is less sensitive to a few outlying trajectories with unusually large distances and therefore provides a more robust measure of strategy novelty.

3.3 LOW-LEVEL FORK INFLUENCE

At the low level, we ask whether resampling at a high-entropy token position yields continuations that induce different distributions over the next intermediate conclusion. For a token position t within the reasoning segment $u_i^{(k)}$ of step k, let h_t denote the prefix before that position and compute

$$H_t^{\text{token}} = -\sum_{v \in \mathcal{V}} \pi(v \mid h_t) \log \pi(v \mid h_t). \tag{4}$$

For simplicity, for each trajectory, we rank its reasoning tokens by H_t^{token} and select the top-q high-entropy positions for resampling (Wang et al., 2025b; Zheng et al., 2025b). At each selected token t, we retain the original continuation and independently resample $M-1$ additional continuations from the same prefix h_t, completing each through the end of the current reasoning step. We index the resulting M continuations by $m \in \{1, \ldots, M\}$, with $m = 1$ denoting the original. Let $b_{t,m}$ denote the reasoning suffix of branch m, starting at position t and ending immediately before its next intermediate conclusion $z_{t,m}$. Thus, $c_{t,m} = h_t \circ b_{t,m}$ is the complete context preceding that conclusion, where $\circ$ denotes concatenation.

Let $\bar{\pi}_t(z) = M^{-1} \sum_{m'=1}^{M} \pi(z \mid c_{t,m'})$ denote the mean likelihood of conclusion z across branch contexts sampled at t. We define the fork influence reward for token t as

$$R_t^L = \max\left\{0, \frac{1}{M} \sum_{m=1}^{M} \log \frac{\pi(z_{t,m} \mid c_{t,m})}{\bar{\pi}_t(z_{t,m})}\right\}. \tag{5}$$

For each branch m, the log-ratio $\log[\pi(z_{t,m} \mid c_{t,m})/\bar{\pi}_t(z_{t,m})]$ compares the likelihood of its conclusion under its own context with the average likelihood of the same conclusion across all branch contexts. It is large when $z_{t,m}$ is specific to $c_{t,m}$ and receives little support from alternative branch contexts. Averaging across branches makes Equation 5 large only when such branch-specific support holds consistently, meaning that each sampled context preferentially supports the conclusion it produces. Thus, a high reward indicates that branches sampled at token t induce different next-conclusion distributions, distinguishing genuine reasoning forks from lexical variation.

3.4 MULTI-LEVEL INFLUENCE-GUIDED POLICY OPTIMIZATION

We augment GRPO with two influence rewards, one for strategies and one for candidate reasoning forks. Within each rollout group, we normalize the task return and strategy influence reward separately. The fork influence reward is not group-normalized; it directly defines a local advantage for selected decision tokens. For token t in rollout i, the advantages are

$$\begin{aligned} A_t = A_i^{\text{task}} &= \frac{r_i^{\text{task}} - \mu_{\text{task}}}{\sigma_{\text{task}} + \epsilon}, & A_i^H &= \frac{R_{i,\text{explore}}^H - \mu_H}{\sigma_H + \epsilon}, \\ A_t^L &= \min\left(\alpha R_{t,\text{explore}}^L, \kappa |A_t|\right), & \widehat{A}_{i,t} &= A_t + \lambda_H \frac{m_{i,t}^H}{|s_i|} A_i^H + m_{i,t}^L A_t^L. \end{aligned} \tag{6}$$

Here, $(\mu_{\text{task}}, \sigma_{\text{task}})$ and (μ_H, σ_H) denote the corresponding group statistics. The coefficients λ_H, α, and κ control the two exploration signals, while the minimum in A_t^L caps the local advantage at a task-dependent scale. The masks $m_{i,t}^H$ and $m_{i,t}^L$ assign high-level credit only to strategy tokens and low-level credit only to selected fork tokens, respectively. Dividing the high-level bonus by $|s_i|$ keeps the total high-level credit independent of strategy length for a fixed A_i^H, reducing the incentive to accumulate exploration credit through longer strategies. Substituting $\widehat{A}_{i,t}$ for the response-level advantage in GRPO gives the final MIGE objective

$$\begin{aligned} J_{\text{MIGE}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\Bigg(&\rho_{i,t}(\theta)\widehat{A}_{i,t}, \\ &\text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\,\widehat{A}_{i,t}\Bigg) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Bigg]. \end{aligned} \tag{7}$$

4 EXPERIMENTS

Our experiments are designed to answer the following research questions.

- **RQ1 (Effectiveness):** Does MIGE improve mathematical reasoning accuracy and coverage of solvable problems over standard RL and representative entropy- and diversity-based exploration methods while preserving effective exploration?
- **RQ2 (Exploration quality):** Does the high-level distance distinguish strategy changes that alter subsequent reasoning from surface variations, and does the low-level score identify observed reasoning forks more reliably than token entropy?
- **RQ3 (Component contributions):** How does each exploration level contribute to reasoning performance, and do the two levels provide complementary gains when combined?

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We train on a subset of DeepScaleR (Luo et al., 2025) and evaluate on AIME 2024–2026, AMC 2023, MATH-500 (Lightman et al., 2023), and GPQA (Rein et al., 2023), with GPQA testing out-of-domain generalization. We report Avg@16 for AIME and AMC, Pass@1 for the remaining benchmarks, and Pass@64 for AIME and AMC.

Models and baselines. We use Qwen3-1.7B and Qwen3-4B (Yang et al., 2025) for the main experiments. The baselines are VANILLA GRPO, DIVER (Hu et al., 2025), and GRPO W/ ENTROPY ADV. (Cheng et al., 2025). The latter two represent output-diversity- and entropy-based exploration. For MIGE, we use $q = 16$ candidate fork positions, sample $M = 4$ branches per position, and set $\lambda_H = 0.3$, $k = 2$, $\alpha = 0.9$, and $\kappa = 0.5$.

4.2 MAIN RESULTS

Table 1: Reasoning performance on Qwen3-1.7B and Qwen3-4B. AIME and AMC use Avg@16, whereas MATH-500 and GPQA use Pass@1. Bold denotes the best result at each model scale.

	In-Domain					Out-of-Domain	
Method	AIME24	AIME25	AIME26	AMC23	MATH-500	GPQA	Avg.
Qwen3-1.7B	13.4	9.8	7.8	42.5	70.2	28.8	28.8
Vanilla GRPO	14.2	10.0	8.1	43.1	71.6	33.8	30.1
DIVER	14.4	11.3	8.5	**45.4**	71.8	33.8	30.9
GRPO+Entropy Adv.	**14.8**	11.7	8.5	43.4	72.4	34.8	30.9
MIGE (ours)	14.0	**12.1**	**9.0**	44.4	**72.4**	**35.9**	**31.3**
Qwen3-4B	23.1	19.4	16.9	66.0	83.6	42.4	41.9
Vanilla GRPO	24.0	21.3	18.3	67.0	83.8	42.9	42.9
DIVER	24.8	20.9	17.5	68.4	83.6	42.9	43.0
GRPO+Entropy Adv.	24.4	21.9	18.4	69.1	83.8	46.0	43.9
MIGE (ours)	**27.4**	**23.2**	**18.7**	**70.9**	**84.8**	**47.5**	**45.4**

RQ1a: Reasoning performance and generalization. Table 1 shows that MIGE achieves the best overall average at both model scales, improving over Vanilla GRPO from 30.1 to 31.3 on Qwen3-1.7B and from 42.9 to 45.4 on Qwen3-4B. More importantly, its margin over the strongest exploration baseline grows from 0.4 to 1.5 points, and at 4B it ranks first on all six benchmarks. This pattern suggests that downstream influence provides a more targeted exploration signal than output diversity or token uncertainty, especially when the underlying policy can generate rich strategy and reasoning-branch candidates; the mixed results on AIME 2024 and AMC 2023 at 1.7B indicate that limited base capability can constrain this benefit. MIGE also exceeds the best baseline on the out-of-domain GPQA benchmark by 1.1 points at 1.7B and 1.5 points at 4B, showing that its gains are not confined to diversity patterns specific to mathematical training tasks.

RQ1b: Reasoning capacity and exploration. Figure 3 shows that MIGE's accuracy gains generally do not come at the cost of successful-solution coverage. At Pass@64, MIGE is best or tied for best in six of the eight model–benchmark settings—three of four at each model scale—indicating that its average-accuracy gains are typically retained under larger sampling budgets. Coverage is not uniformly improved: MIGE trails Vanilla GRPO and GRPO+Entropy Adv. on AIME 2026 at 1.7B, while DIVER performs better on AIME 2025 at 4B. Although the gains are not universal across benchmarks, the overall results indicate that influence-guided exploration improves both average accuracy and successful-solution coverage under larger sampling budgets.

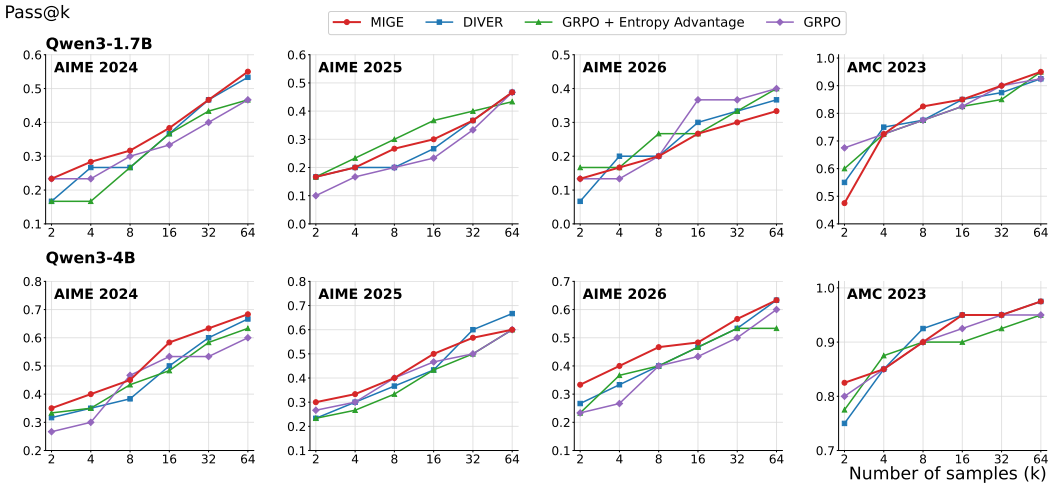


Figure 3: Pass@k on AIME 2024–2026 and AMC 2023 for Qwen3-1.7B (top) and Qwen3-4B (bottom), with $k \in \{2, 4, 8, 16, 32, 64\}$.

4.3 Analysis

4.3.1 High-Level Strategy-Influence Validation

RQ2a: Does the high-level distance capture genuine strategy changes? We select problems with multiple solution methods from AIME, AMC, MATH-500, Minerva Math, and OlympiadBench (He et al., 2024). For each problem, we prompt Qwen3-8B to generate a complete reasoning trajectory in the format described in Section 3.1: the model itself first proposes an original strategy and then executes it step by step. We then use `DeepSeek-V4-Flash`, with the prompt in Appendix C.4, to construct two strategy variants: a surface rewrite that preserves the original solution method and an alternative that uses a genuinely different method. Qwen3-8B then generates the reasoning suffix conditioned on each constructed strategy.

For each metric, we test whether the alternative receives a larger distance from the original than the surface rewrite. We compare strategy influence (Equation 2) with three standard text-similarity metrics: embedding cosine similarity using Qwen3-Embedding-0.6B (Zhang et al., 2025), BLEU (Papineni et al., 2002), and token-level F1. We convert these similarity scores into corresponding distances; the detailed definitions are provided in Appendix A.2. All baselines compare the generated reasoning trajectories, excluding the strategy spans. Table 2 reports the resulting accuracy.

Table 2: High-level strategy validation on Qwen3-8B. Accuracy is the fraction of examples for which a metric assigns a larger distance to the genuine alternative strategy than to the surface rewrite; baseline similarity scores are converted to complementary distances.

Metric	Accuracy
Embedding cosine similarity	70.37%
BLEU	62.96%
Token-level F1	62.96%
Strategy influence (ours)	**88.89%**

On the constructed strategy pairs, the proposed influence score assigns a larger distance to the genuine alternative with 88.89% accuracy, exceeding embedding cosine similarity (70.37%) and both BLEU and token-level F1 (62.96%). The baseline distances compare the generated reasoning trajectories directly and can therefore be affected by incidental wording or shared intermediate calculations. These results suggest that strategy influence better distinguishes genuine strategy changes from surface rewrites than the tested output-distance baselines.

4.3.2 Low-Level Fork-Influence Analysis

RQ2b: Does fork influence better locate reasoning forks? We use the GSM8K test set (Cobbe et al., 2021) because its short arithmetic derivations permit precise inspection of local errors. We sample 30 problems on which Qwen3-8B attains an Avg@8 score below 0.5 and select one incorrect trajectory per problem. We first identify the earliest erroneous reasoning step in each trajectory. For every reasoning token t in that step, we then follow Section 3.3: from prefix h_t, we independently resample $M = 32$ continuations through the next intermediate conclusion. `DeepSeek-V4-Flash` examines the resulting continuations and labels position t as a reasoning fork when resampling produces substantively different intermediate conclusions; differences in wording alone do not qualify. The annotation prompt is provided in Appendix C.5.

After fixing these labels, we score every reasoning token in the erroneous step using token entropy, ROSE semantic entropy (Zhao et al., 2026), and fork influence. We rank tokens under each metric. Top-k hit is the fraction of examples with a labeled fork among the k highest-scoring positions; mean fork rank averages the labeled positions' ranks. Higher hit rates and lower ranks are better.

Table 3: Localization of reasoning-fork tokens annotated from resampled continuations on 30 selected GSM8K problems. Higher Top-k hit and lower mean rank are better.

Metric	Top-1 hit	Top-3 hit	Top-5 hit	Mean rank
Token entropy	23.3%	60.0%	73.3%	4.80
Semantic entropy (ROSE)	20.0%	56.7%	73.3%	5.00
Fork influence (ours)	**70.0%**	**93.3%**	**100.0%**	**1.65**

Fork influence raises Top-1 hit from 23.3% for token entropy and 20.0% for ROSE to 70.0%, and finds a labeled fork among the top five positions in every trajectory. It also reduces mean rank from 4.80 and 5.00, respectively, to 1.65. These results suggest that downstream changes in intermediate conclusions provide a stronger signal for locating reasoning forks than token-level uncertainty alone.

Figure 4 presents a GSM8K problem that asks for the initial amount of money given a final balance of 100 after receiving a total allowance of 40. The erroneous trajectory adds the allowance to the final balance, producing 140 instead of the correct 60. Token entropy is highest at the later token `amount`, after addition has been selected; all 32 continuations resampled there retain 140, yielding zero fork influence. Fork influence is highest at `Add`; resampling there yields 29 addition branches and three subtraction branches with the correct conclusion 60. This contrast separates uncertainty in expressing a fixed decision from the influence of the decision itself.

4.4 Ablation Studies

RQ3: Contributions of high- and low-level exploration. We ablate the two exploration components of MIGE on Qwen3-4B by removing either the strategy influence reward or the fork influence reward. The resulting variants retain only local fork influence or global strategy influence, respectively. We report the same average across six benchmark scores as in Table 1.

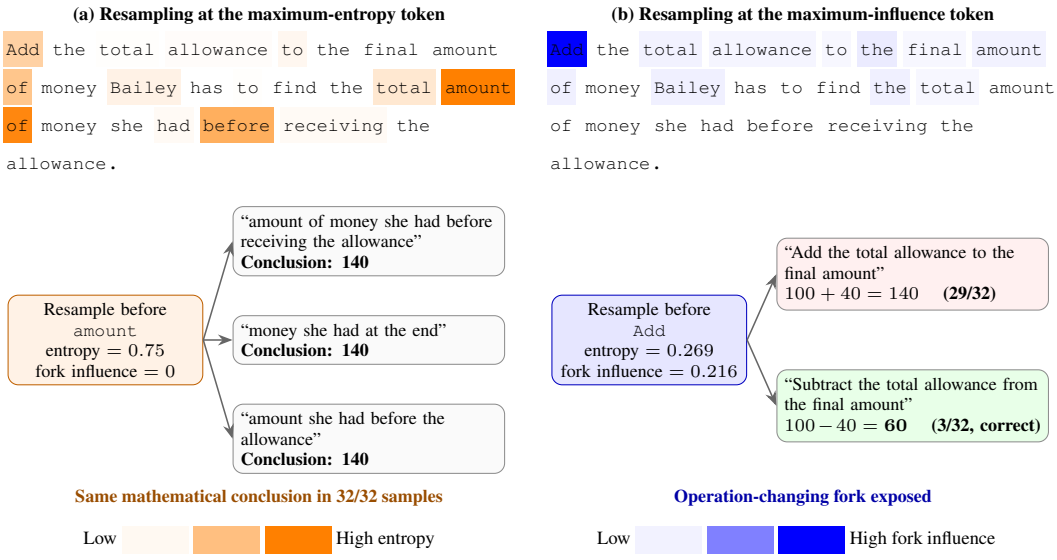


Figure 4: A representative local-fork example. Token entropy peaks after the trajectory has committed to addition, whereas fork influence peaks at the preceding operation token. Token colors show each metric normalized within the erroneous step.

Table 4: Ablation of MIGE on Qwen3-4B; "Avg." follows Table 1.

Variant	Avg.
MIGE	45.4
w/o high-level exploration	44.2
w/o low-level exploration	43.4
w/o exploration	42.9

Effectiveness of individual components. As shown in Table 4, retaining only low-level exploration achieves 44.2, exceeding GRPO w/ Entropy Adv. (43.9) by 0.3 points. Retaining only high-level exploration achieves 43.4, exceeding DIVER (43.0) by 0.4 points. The two single-component variants therefore each improve over the corresponding exploration baseline, showing that both influence signals contribute independently of the other.

Complementarity of the two exploration levels. Full MIGE achieves 45.4. Removing high-level or low-level exploration reduces the average by 1.2 or 2.0 points, respectively, while removing both yields 42.9. The larger drop without low-level exploration indicates that fork influence makes the stronger marginal contribution in this setting. However, adding high-level exploration to the low-level-only variant still improves the average from 44.2 to 45.4. This shows that high- and low-level exploration provide complementary benefits, with their combination achieving the best performance.

5 Conclusion

We introduced Multi-Level Influence-Guided Exploration (MIGE), which guides exploration by measuring how global strategies and local reasoning forks affect subsequent reasoning. Across six benchmarks and two model scales, MIGE improves average performance over standard RLVR and representative exploration baselines while largely preserving Pass@64 coverage. Controlled analyses show that its influence scores distinguish genuine strategy changes from surface rewrites and locate reasoning forks more reliably than the tested alternatives; ablations confirm that the two levels are complementary. Together, these results support influence-guided exploration.

AI USE STATEMENT

We used generative AI tools to brainstorm and refine research ideas, methodology, experimental design, and the analysis workflow. In the experiments, DeepSeek-V4-Flash was used to construct strategy variants and annotate reasoning forks, as described in Section 4.3 and Appendix C. We also used generative AI tools to retrieve and identify relevant literature, draft and revise selected manuscript passages, improve clarity and grammar, and assist with final formatting and reference checks. All AI-assisted outputs were reviewed by the authors: literature and bibliographic metadata were checked against original sources, and experimental artifacts and conclusions were validated by the authors. Generative AI made no final scientific decisions and provided no unverified evidence.

ETHICS STATEMENT

We are not aware of any major ethical concerns arising from this work. Our experiments focus on mathematical and scientific reasoning and use publicly available models and benchmark datasets. No human subjects or private or sensitive data were involved, and the work does not introduce directly harmful applications or insights.

REPRODUCIBILITY STATEMENT

We describe the proposed method in Section 3 and the experimental setup and evaluation in Section 4. Appendix A reports the hardware, training, optimization, and decoding configurations, while Appendix C provides the full prompts and parser implementation. All compared methods use the same base models, training data, and evaluation protocol unless stated otherwise.

REFERENCES

Xiwen Chen, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Hao Wang, Haiyu Wu, Huayu Li, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. DRA-GRPO: Your GRPO needs to know diverse reasoning paths for mathematical reasoning. *arXiv preprint arXiv:2505.09655*, 2025. URL https://arxiv.org/abs/2505.09655.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025. URL https://arxiv.org/abs/2505.22617.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. URL https://arxiv.org/abs/2402.14008.

Wenyang Hu, Junxiang Jia, Zhen Shu, Daniel Dahlmeier, See-Kiong Ng, and Bryan Kian Hsiang Low. ExTra: Exploratory trajectory optimization for language model reinforcement learning. *arXiv preprint arXiv:2606.24994*, 2026. URL https://arxiv.org/abs/2606.24994.

Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. Diversity-incentivized exploration for versatile reasoning. *arXiv preprint arXiv:2509.26209*, 2025. URL https://arxiv.org/abs/2509.26209.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2510.03222*, 2025. doi: 10.48550/arXiv.2510.03222. URL https://arxiv.org/abs/2510.03222.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023. URL https://arxiv.org/abs/2302.09664.

Chenyi Li, Yuan Zhang, Bo Wang, Guoqing Ma, Wei Tang, Haoyang Huang, and Nan Duan. SetPO: Set-level policy optimization for diversity-preserving LLM reasoning. *arXiv preprint arXiv:2602.01062*, 2026a. URL https://arxiv.org/abs/2602.01062.

Gengsheng Li, Jinghan He, Shijie Wang, Dan Zhang, Ruiqi Liu, Renrui Zhang, Zijun Yao, Junfeng Fang, Haiyun Guo, and Jinqiao Wang. R-Diverse: Mitigating diversity illusion in self-play LLM training. *arXiv preprint arXiv:2602.13103*, 2026b. doi: 10.48550/arXiv.2602.13103. URL https://arxiv.org/abs/2602.13103.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025. URL https://arxiv.org/abs/2509.02534.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing O1-Preview with a 1.5b model by scaling RL. https://github.com/agentica-project/deepscaler, 2025. Project repository.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Hongze Tan, Zihan Wang, Jianfei Pan, Jinghao Lin, Hao Wang, Yifan Wu, Tao Chen, Zhihang Zheng, Zhihao Tang, and Haihua Yang. GTPO and GRPO-S: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*, 2025. URL https://arxiv.org/abs/2508.04349.

Zhongwei Wan, Yun Shen, Zhihao Dou, Donghao Zhou, Yu Zhang, Xin Wang, Hui Shen, Jing Xiong, Chaofan Tao, Zixuan Zhong, Peizhou Huang, and Mi Zhang. DSDR: Dual-scale diversity regularization for exploration in LLM reasoning. *arXiv preprint arXiv:2602.19895*, 2026. URL https://arxiv.org/abs/2602.19895.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. Emergent hierarchical reasoning in LLMs through reinforcement learning. *arXiv preprint arXiv:2509.03646*, 2025a. doi: 10.48550/arXiv.2509.03646. URL https://arxiv.org/abs/2509.03646.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025b. URL https://arxiv.org/abs/2506.01939.

Shangyu Xing, Siyuan Wang, Chenyuan Yang, Xinyu Dai, and Xiang Ren. Lookahead tree-based rollouts for enhanced trajectory-level exploration in reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2510.24302*, 2025. URL https://arxiv.org/abs/2510.24302.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025. URL https://arxiv.org/abs/2503.18892.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

Ziqi Zhao, Zhaochun Ren, Jiahong Zou, Liu Yang, Zhiwei Xu, Xuri Ge, Zhumin Chen, Xinyu Ma, Daiting Shi, Shuaiqiang Wang, Dawei Yin, and Xin Xin. Reinforced efficient reasoning via semantically diverse exploration. *arXiv preprint arXiv:2601.05053*, 2026. doi: 10.48550/arXiv.2601.05053. URL https://arxiv.org/abs/2601.05053v1.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a. URL https://arxiv.org/abs/2507.18071.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, Qian Liu, Ge Zhang, and Zejun Ma. First return, entropy-eliciting explore. *arXiv preprint arXiv:2507.07017*, 2025b. URL https://arxiv.org/abs/2507.07017.

A Experimental Details

A.1 Training settings

We implement all methods with `verl` and use vLLM for rollout generation. Experiments are conducted on one NVIDIA RTX PRO 6000 GPU with 96 GB memory. We use the same training setup for Qwen3-1.7B and Qwen3-4B, changing only the base-model checkpoint.

We train on the processed DeepScaleR subset for one epoch with a prompt batch size of 16 and sample eight responses per prompt, yielding 128 trajectories per batch. PPO uses mini-batches of 32 trajectories and one optimization epoch per batch. Prompts exceeding 1,024 tokens are filtered, and responses are limited to 4,096 tokens.

Table 5 summarizes the main optimization and sampling hyperparameters. We use a learning rate of 10^{-6} and weight decay 0.01. The asymmetric PPO clipping interval is $[0.8, 1.28]$, the KL coefficient is 0.001, and the entropy regularization coefficient is zero.

Table 5: Training hyperparameter settings.

Setting	Value	Setting	Value
Maximum prompt length	1,024	Maximum response length	4,096
Prompt batch size	16	Rollouts per prompt	8
PPO mini-batch size	32	Learning rate	1×10^{-6}
Weight decay	0.01	Lower clip ratio	0.20
Upper clip ratio	0.28	KL coefficient	0.001
Entropy coefficient	0	Training temperature	1.0
Evaluation temperature	0.6	Top-p	1.0

A.2 High-Level Analysis Metrics

Let x be the original trajectory and y a surface-rewritten or alternative-strategy trajectory. We express each baseline similarity as a distance. For the embedding baseline, we use

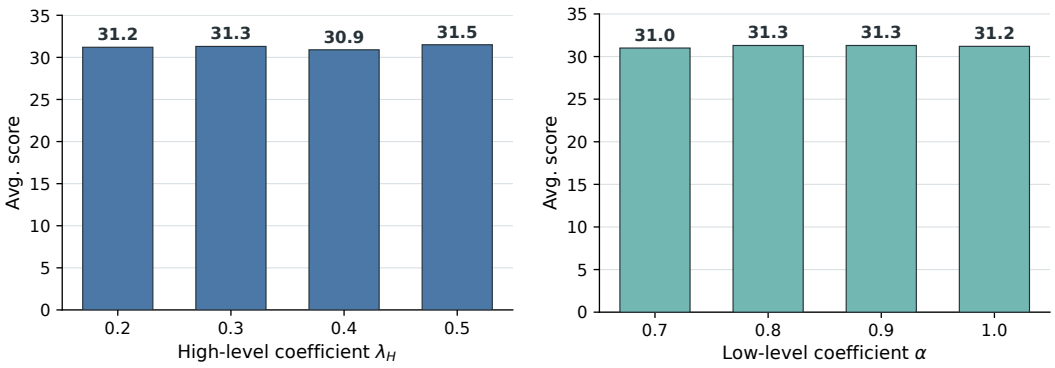
$$d_{\mathrm{emb}}(x, y) = 1 - \frac{e(x)^\top e(y)}{\|e(x)\|_2 \|e(y)\|_2}, \tag{8}$$

where $e(\cdot)$ is the representation produced by Qwen3-Embedding-0.6B. For sentence-level BLEU, with x as the reference and y as the hypothesis, we use $d_{\mathrm{BLEU}}(x, y) = 1 - \mathrm{BLEU}(x, y)$. For token-level F1, let $P = |T_x \cap T_y|/|T_y|$ and $R = |T_x \cap T_y|/|T_x|$, where T_x and T_y are the token multisets of the two trajectories after the same normalization and tokenization. We compute $\mathrm{F1} = 2PR/(P+R)$ and use $d_{\mathrm{F1}}(x, y) = 1 - \mathrm{F1}$. Thus, larger values indicate greater trajectory differences.

B Hyperparameter Sensitivity

We examine the sensitivity of MIGE to the high-level exploration coefficient λ_H and the low-level coefficient α on Qwen3-1.7B. We vary one coefficient at a time while holding the remaining training configuration fixed. Figure 5 reports the average evaluation score. Both panels use the same non-truncated vertical scale so that the small absolute differences are not visually exaggerated.

Performance is stable throughout both ranges. Varying λ_H from 0.2 to 0.5 yields average scores between 30.9 and 31.5, a maximum spread of 0.6 points. The result is even less sensitive to α: values from 0.7 to 1.0 produce scores between 31.0 and 31.3, with $\alpha = 0.8$ and $\alpha = 0.9$ tied at 31.3. Performance is therefore stable across the evaluated ranges.



(a) High-level coefficient λ_H **(b) Low-level coefficient α**

Figure 5: Hyperparameter sensitivity on Qwen3-1.7B. The vertical axis spans 0–35 in both panels, and exact scores are printed above the bars.

C PROMPTS

C.1 STRUCTURED REASONING PROMPT

For MIGE, we append the following strategy-first instruction to every problem:

> **[Problem]**
>
> Output your overall strategy first. Then think step by step and output the final answer within `\boxed{}`.

C.2 BASELINE PROMPT

For the other baselines, we append the following instruction to every problem:

> **[Problem]**
>
> Think step by step and output the final answer within `\boxed{}`.

C.3 RULE-BASED STRUCTURE PARSER

The parser uses the same rules for Qwen3-1.7B and Qwen3-4B and retains exact token offsets. It identifies execution steps from explicit Step headings, numbered Markdown headings, or numbered lists. An unnumbered step-by-step section is treated as one reasoning step.

A leading section headed by "strategy", "approach", or "plan" is used as the strategy span; otherwise, the prefix before Step 1 is used. Within each execution step, the parser extracts the final qualifying mathematical conclusion, and the preceding text forms the reasoning span.

Listing 1 shows the core parser implementation.

Listing 1: Core rule-based extraction of the strategy, reasoning steps, and intermediate conclusions.

```
import re

STEP_HEADING = re.compile(
    r"^(?:(?:#{1,6})[ \t]+)?"
    r"(?:[^\w\s#$\\]{0,4}[ \t]*)?"
    r"(?:\*\*|__)?[ \t]*step[ \t]*"
    r"(?P<number>\d+(?:\.\d+)*)"
    r"(?:\*\*|__)?[ \t]*(?::|[.)-](?:[ \t]|$)|$)", re.I)
NUMBERED_HEADING = re.compile(
    r"^(?:#{2,6})[ \t]+(?:\*\*|__)?"
    r"(?P<number>\d+)[.)][ \t]+\S", re.I)
NUMBERED_LIST = re.compile(
```

```
    r"^(?P<number>\d+)[.)][ \t]+(?:\*\*|__).+(?:\*\*|__)[ \t]*:?[ \t]*$")
STRATEGY_HEADING = re.compile(
    r"\b(?:overall[ \t]+)?(?:strategy|approach|plan)\b", re.I)
EXECUTION_HEADING = re.compile(
    r"\b(?:step[- ]by[- ]step|detailed|full)[ \t]+"
    r"(?:solution|reasoning|analysis|calculation|computation|explanation|"
    r"breakdown|execution|optimization)", re.I)
RESULT_CUE = re.compile(
    r"\b(?:therefore|thus|hence|consequently|"
    r"we (?:get|find|obtain|conclude)|this (?:gives|means|implies)|"
    r"the result is)\b", re.I)
MATH_RELATION = re.compile(
    r"(?:=|\\(?:leq|geq|equiv|approx|in)\b|[<>\u2264\u2265\u2248\u2208])")
VALUE_SIGNAL = re.compile(
    r"(?:\\boxed\s*\{|\$[^$\n]*\d[^$\n]*\$|\b\d+(?:\.\d+)?\b)")

def find_step_headings(text):
    explicit = [line for line in lines(text)
                if STEP_HEADING.match(line.text.strip())]
    if explicit:
        execution = execution_sections(text)
        if execution:
            after = [h for h in explicit if h.start >= execution[-1].end]
            if after:
                return after       # discard numbered strategy outline
        return explicit
    numbered = [line for line in lines(text)
                if NUMBERED_HEADING.match(line.text.strip())]
    if is_consecutive_from_one(numbered):
        return numbered
    execution = execution_sections(text)
    if not execution:
        return []
    listed = numbered_list_titles_after(text, execution[-1], NUMBERED_LIST)
    if is_consecutive_from_one(listed):
        return listed
    return [as_coarse_step(execution[-1])]

def find_strategy_span(text, headings):
    # Prefer an explicit strategy/approach/plan section before execution.
    candidates = [line for line in lines(text)[:16]
                  if STRATEGY_HEADING.search(strip_markup(line.text))]
    start = len(text) - len(text.lstrip())
    if candidates:
        following_steps = [h.start for h in headings
                           if h.start > candidates[0].start]
        if following_steps:
            end = min(following_steps)
            if text[candidates[0].end:end].strip():
                return start, end

    # Structural fallback: an unheaded, nontrivial prefix before Step 1.
    if headings and headings[0].number.split(".")[0] == "1":
        end = headings[0].start
        if len(text[start:end].strip()) >= 24:
            return start, end
    return None

def conclusion_strength(paragraph):
    if (not balanced_math(paragraph)
            or paragraph.rstrip("*_ ").endswith("?")
            or is_next_action_list(paragraph)
            or is_weak_or_transition(paragraph)):
        return -1
    if has_complete_box(paragraph):
        return 10
    has_cue = RESULT_CUE.search(paragraph) is not None
    has_relation = MATH_RELATION.search(paragraph) is not None
    has_value = VALUE_SIGNAL.search(paragraph) is not None
    if has_cue and (has_relation or has_value):
        return 7
    if has_relation and ("$" in paragraph or "\\" in paragraph):
        return 4
    if has_value and ("$" in paragraph or "**" in paragraph):
        return 3
    return -1

def extract_step_conclusion(text, body_start, body_end):
    # The last sufficiently strong paragraph is the local conclusion.
    for start, end in reversed(paragraph_spans(text, body_start, body_end)):
        paragraph = text[start:end]
```

```
        strength = conclusion_strength(paragraph)
        if strength >= 3:
            if strength >= 7:
                start = last_conclusion_sentence_start(text, start, end)
            return start, end
    return None

def parse_structured_cot(text):
    headings = find_step_headings(text)
    if not headings or not has_complete_final_box(text, headings[-1].end):
        return None
    strategy = find_strategy_span(text, headings)  # optional for low level

    steps = []
    for i, heading in enumerate(headings):
        if strategy is not None and heading.start < strategy[1]:
            continue                    # strategy heading, not execution
        section_end = (headings[i + 1].start
                       if i + 1 < len(headings) else len(text))
        conclusion = extract_step_conclusion(text, heading.end, section_end)
        if conclusion is None:
            return None                 # reject partially parsed responses
        result_start, result_end = conclusion
        if not text[heading.end:result_start].strip():
            return None
        steps.append({
            "heading": text[heading.start:heading.end],
            "reasoning": text[heading.end:result_start],
            "conclusion": text[result_start:result_end],
        })
    strategy_text = text[slice(*strategy)] if strategy is not None else None
    return {"strategy": strategy_text, "steps": steps}
```

Handling unsuccessful parses. A trajectory without an extracted strategy is excluded from high-level influence computation and receives no high-level bonus. An invalid step parse receives no low-level bonus, while its task reward and standard GRPO advantage are retained. Invalid or duplicate fork continuations are resampled within the attempt budget; positions with fewer than M valid branches receive no low-level bonus.

C.3.1 Robustness of the Rule-Based Structured-Reasoning Parser

We evaluate the parser on 400 held-out Qwen3-1.7B and Qwen3-4B rollouts with a 4,096-token response limit. After excluding 14 length-truncated responses, Table 6 reports coverage on the remaining 386. "Both valid" means that both the strategy and step structure are extracted.

Table 6: Rule-based structured-reasoning parser coverage.

Model	Strategy valid	Step valid	Both valid
Qwen3-1.7B	93.8%	80.9%	78.9%
Qwen3-4B	87.0%	78.6%	73.4%

Across the two models, coverage is 90.4% for strategies, 79.8% for steps, and 76.2% for both.

For local-conclusion extraction, we sample four continuations from each of 64 fork prefixes. The parser extracts the next local conclusion from 212 of these 256 counterfactual branches (82.8%).

These results support the parser as a simple, computationally inexpensive way to obtain structured spans for MIGE in most evaluated cases without additional model inference, although extraction remains imperfect. Potential ways to improve extraction include format-oriented supervised fine-tuning that teaches the model to emit explicit `<strategy>`, `<reasoning>`, and `<conclusion>` tags with matching closing tags, or an LLM-as-a-judge approach that identifies and validates spans when rule-based extraction is ambiguous. These alternatives would require additional training or inference, and we leave their evaluation to future work.

C.4 Strategy-Variant Construction Prompt

We use this prompt with `DeepSeek-V4-Flash` for the strategy variants in Section 4.3.1:

Given a problem and an original strategy, produce two variants.
Surface rewrite: Restate the original strategy using different wording and organization while preserving exactly the same underlying solution method.
Alternative strategy: Propose a genuinely different method for solving the same problem. The alternative must differ in its core mathematical approach, not merely in wording, notation, or the order of presentation.
Do not solve the problem or provide step-by-step reasoning. Return only the two strategies, clearly labeled "Surface rewrite" and "Alternative strategy."
Problem: [Problem]
Original strategy: [Original strategy]

C.5 Reasoning-Fork Annotation Prompt

We use this prompt with `DeepSeek-V4-Flash` for the fork annotations in Section 4.3.2:

You are given one token position in an erroneous reasoning step, the original intermediate conclusion, and one continuation sampled from the prefix ending immediately before that token. The sampled continuation extends through its next intermediate conclusion. Determine whether the sampled continuation produces an intermediate conclusion that is substantively different from the original. A difference must change the mathematical result, inference, operation, or downstream reasoning direction. Stylistic or lexical changes do not count.
Return "Changed conclusion" if the two conclusions are substantively different and "Same conclusion" otherwise. Quote both conclusions and briefly justify the label.
Problem: [Problem]
Original erroneous step: [Erroneous step]
Token position: [Token position]
Original intermediate conclusion: [Original conclusion]
Resampled continuation: [Continuation]

D Case Studies

We sample two correct responses to the same problem from the final Qwen3-4B MIGE checkpoint. Colors distinguish the response components: **strategy**, **reasoning**, and **conclusion**.

Problem. A rectangular box P has distinct edge lengths a, b, and c. The sum of the lengths of all 12 edges is 13, the sum of the areas of all six faces is $11/2$, and the volume is $1/2$. The longest interior diagonal is m/n, where m and n are relatively prime positive integers. Find $m+n$.

Response 1: Direct Symmetric Identity

We are given a rectangular box (a rectangular prism) with edge lengths a, b, and c, all distinct. The following are known:

1. **Sum of all 12 edges.** Each of the 12 edges consists of four of each length a, b, c, so
$$4(a+b+c) = 13 \quad \Rightarrow \quad a+b+c = \frac{13}{4}.$$
2. **Sum of the areas of all 6 faces.** The six faces consist of:
   - two faces with area ab,
   - two faces with area bc,
   - two faces with area ca.

   So the total area is
$$2(ab+bc+ca) = \frac{11}{2} \quad \Rightarrow \quad ab+bc+ca = \frac{11}{4}.$$

3. **Volume of the box.**

$$abc = \frac{1}{2}.$$

We are to find the **length of the longest interior diagonal**, which is the space diagonal of the rectangular box. This is given by

$$d = \sqrt{a^2 + b^2 + c^2}.$$

So, our goal is to compute $\sqrt{a^2 + b^2 + c^2}$ using the equations above.
Step 1: Use identities to find $a^2 + b^2 + c^2$.
Recall that

$$(a + b + c)^2 = a^2 + b^2 + c^2 + 2(ab + bc + ca).$$

We know that $a + b + c = 13/4$ and $ab + bc + ca = 11/4$.
So,

$$\left(\frac{13}{4}\right)^2 = a^2 + b^2 + c^2 + 2 \cdot \frac{11}{4},$$

$$\frac{169}{16} = a^2 + b^2 + c^2 + \frac{22}{4} = a^2 + b^2 + c^2 + \frac{11}{2},$$

$$\Rightarrow \quad a^2 + b^2 + c^2 = \frac{169}{16} - \frac{11}{2} = \frac{169}{16} - \frac{88}{16} = \frac{81}{16}.$$

Step 2: Compute the length of the longest interior diagonal.

$$d = \sqrt{a^2 + b^2 + c^2} = \sqrt{\frac{81}{16}} = \frac{9}{4}.$$

Step 3: Final Answer.
We are told to write the answer in the form m/n, where m and n are relatively prime. In this case, $9/4$ is already in simplest form, so

$$m = 9, \qquad n = 4 \quad \Rightarrow \quad m + n = 13.$$

Final Answer:

$$\boxed{13}$$

RESPONSE 2: CUBIC-ROOT RECONSTRUCTION

We are given a rectangular box with **distinct** edge lengths a, b, c, and we are given the following:

- The **sum of all 12 edges** of the box is 13.
- The **sum of the areas of all 6 faces** is $11/2$.
- The **volume** of the box is $1/2$.

We are to find the **length of the longest interior diagonal** (which is the space diagonal), and express it as a reduced fraction m/n, then give $m + n$.
Step 1: Define the variables and set up equations.
Let a, b, c be the **distinct** edge lengths (so each is different). Then:

- Each edge is repeated four times in the box. So the sum of all edges is $4(a + b + c)$. In the problem, this is given to be 13:

$$4(a + b + c) = 13 \quad \Rightarrow \quad a + b + c = \frac{13}{4}.$$

- The areas of the six faces of a rectangular box are

$$2(ab + bc + ca).$$

So,

$$2(ab + bc + ca) = \frac{11}{2} \quad \Rightarrow \quad ab + bc + ca = \frac{11}{4}.$$

- The volume is

$$abc = \frac{1}{2}.$$

So now we have the following three equations:

$$(1) \quad a+b+c = \frac{13}{4}, \qquad (2) \quad ab+bc+ca = \frac{11}{4}, \qquad (3) \quad abc = \frac{1}{2}.$$

These are the **symmetric sums** of the roots of a cubic polynomial:

$$x^3 - (a+b+c)x^2 + (ab+bc+ca)x - abc = 0.$$

Substitute the known values:

$$x^3 - \frac{13}{4}x^2 + \frac{11}{4}x - \frac{1}{2} = 0.$$

Multiply through by 4 to eliminate denominators:

$$4x^3 - 13x^2 + 11x - 2 = 0.$$

Step 2: Solve the cubic equation.
We need to find the roots of

$$4x^3 - 13x^2 + 11x - 2 = 0.$$

Try possible rational roots using the Rational Root Theorem: possible roots are $\pm 1, \pm 2, \pm 1/2, 1/4$. Try $x = 1$:

$$4(1)^3 - 13(1)^2 + 11(1) - 2 = 4 - 13 + 11 - 2 = 0.$$

So $x = 1$ **is a root**. We can factor this out. Let us perform polynomial division (or use synthetic division) to divide the cubic by $(x-1)$:

1	4	−13	11	−2
		4	−9	2
	4	−9	2	0

So the cubic factors as

$$(x-1)(4x^2 - 9x + 2) = 0.$$

Now solve

$$4x^2 - 9x + 2 = 0.$$

Using the quadratic formula,

$$x = \frac{9 \pm \sqrt{81-32}}{8} = \frac{9 \pm \sqrt{49}}{8} = \frac{9 \pm 7}{8}.$$

So

$$x = \frac{16}{8} = 2 \quad \text{or} \quad x = \frac{2}{8} = \frac{1}{4}.$$

The roots of the cubic are

$$x = 1, \qquad x = 2, \qquad x = \frac{1}{4}.$$

So the edge lengths of the box are

$$a = \frac{1}{4}, \qquad b = 1, \qquad c = 2.$$

Or any permutation, but since they are distinct, we can use them in any order for computing the space diagonal.

Step 3: Find the length of the longest interior diagonal.
The **space diagonal** of a rectangular box is given by

$$d = \sqrt{a^2 + b^2 + c^2}.$$

So compute

$$a = \frac{1}{4}, \qquad b = 1, \qquad c = 2,$$

$$a^2 = \left(\frac{1}{4}\right)^2 = \frac{1}{16}, \qquad b^2 = 1^2 = 1, \qquad c^2 = 2^2 = 4.$$

So,

$$d = \sqrt{\frac{1}{16} + 1 + 4} = \sqrt{\frac{1}{16} + 5} = \sqrt{\frac{1}{16} + \frac{80}{16}} = \sqrt{\frac{81}{16}} = \frac{9}{4}.$$

Step 4: Final Answer.
The length of the **longest interior diagonal** is

$$\frac{9}{4}.$$

So, $m = 9$, $n = 4$, and $m + n = \boxed{13}$.
Final Answer: $\boxed{13}$

CONTROLLED PROMPT–REWARD COMPARISON

The two responses instantiate distinct solution routes for the same problem. We separate the effects of output format and training signal through a controlled comparison with the standard GRPO checkpoint. Comparing GRPO under plain chain-of-thought (CoT) and *strategy-first CoT*, which asks the model to state its overall strategy before step-by-step reasoning, measures the effect of the output format. Comparing GRPO and MIGE under strategy-first CoT holds the output format fixed and measures the effect of the training signal.

For each condition, we sample 128 responses to the rectangular-box problem with temperature 0.6, top-$p = 1.0$, at most 2,048 generated tokens, and matched random seeds. We classify a response as *cubic-root reconstruction* only when it forms the cubic and recovers the edge lengths through a correct residual quadratic factor. Responses that mention the cubic but ultimately use $a^2+b^2+c^2 = (a+b+c)^2 - 2(ab+bc+ca)$ are assigned to the *direct symmetric identity* route.

Table 7: Solution-route frequencies for the rectangular-box case study.

Checkpoint	Output format	Direct identity	Cubic reconstruction
MIGE	Strategy-first CoT	103 (80.5%)	25 (19.5%)
GRPO	CoT	108 (84.4%)	20 (15.6%)
GRPO	Strategy-first CoT	110 (85.9%)	18 (14.1%)

Within GRPO, strategy-first CoT changes the cubic-reconstruction frequency from 15.6% to 14.1%, suggesting that the prompt alone does not increase route diversity. Holding the format fixed, MIGE raises this frequency from 14.1% to 19.5%, shifting probability toward the less frequent valid route. This matched-format result supports broader strategy-level exploration.