REASONING ARENA: TRACE TOURNAMENTS WHEN VERIFIABLE REWARDS FALL SHORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has become a leading paradigm for improving the reasoning ability of large language models through outcome-based supervision. However, verifiable rewards frequently become uninformative at the group level: when all sampled traces of a given prompt receive identical rewards, group-relative advantage estimation provides no gradient signal, even though the traces may differ substantially in reasoning quality. We propose REASONING ARENA, an adaptive training framework that routes such non-diverse reward groups to a judge system instead of discarding them. Beyond examining the final answer, REASONING ARENA constructs *trace* tournaments, where reasoning traces are compared head-to-head to expose finer-grained preferences within the group, converting reasoning quality into rich relative reward signals. To make reward estimation efficient, rather than exhaustively comparing every pair, each new trace is evaluated against a small, dynamically updated pool of previously generated traces as *anchors* to efficiently establish a relative ranking. We then fit a Bradley-Terry model on the incomplete comparison graph, enabling scalable RL integration without quadratic pairwise comparisons. Empirical results demonstrate that REASONING ARENA consistently outperforms the RLVR baseline by 7.6% on average in competition mathematics and coding benchmarks. By converting otherwise wasted zero-advantage samples into useful gradient updates, our method accelerates training by 27% to 41%, saving nearly 50% of generation compute, and substantially improves overall reasoning performance.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable progress in following instructions, generating coherent long-form responses, and conducting multi-step reasoning (Ouyang et al., 2022; Kojima et al., 2022). Recently, the reasoning capabilities of LLMs have been further advanced by the paradigm of reinforcement learning with verifiable rewards (RLVR), which trains LLMs directly on outcome-level signals provided by rule-based verifiers. Group-relative policy optimization methods, such as GRPO (Shao et al., 2024), drive this advancement by sampling a group of N reasoning traces for a given prompt, scoring each trace with the verifier, and updating the policy using advantages computed relative to the group mean. This group-level contrast incentivizes the model to explore more effective thinking trajectories.

This paradigm is bottlenecked, however, by the lack of variance of rewards within each sampled group. Since GRPO relies on advantages computed relative to the group average, the estimator carries a gradient signal only when traces within a group receive different rewards. When the data difficulty is misaligned with the model's current capabilities, either too difficult or too easy, a substantial fraction of groups may end up with all traces marked incorrect or all correct. This results in a degenerate group where every advantage is exactly zero. We refer to such cases as *non-diverse reward groups*. In these instances, the rewards collapse to a constant outcome across the entire group. Even though the underlying traces may differ substantially in reasoning quality, soundness, or conciseness, the verifier falls short by providing no signal to distinguish them. As a result, the group contributes zero gradient to policy updates, meaning the expensive compute used to generate these traces is completely wasted.

The existing literature has primarily approached the zero-variance problem through data curation, using prefiltering or difficulty prediction to avoid trivially easy or impossibly hard prompts (Yu et al.,

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.

2025a; Zheng et al., 2026). However, discarding these non-diverse groups forfeits the fine-grained learning signals latent in their reasoning trajectories. A complementary line of work attempts to recover signals directly from these groups via entropy-guided advantage shaping (Le et al., 2026; Wu et al., 2025). However, the reward reshaping relies entirely on the model's intrinsic token distribution and cannot distinguish between a rigorous proof and a confident hallucination. Consequently, the non-diverse reward group remains the fundamental barrier to learning meaningful reasoning behaviors on these prompts while making RL training inefficient.

To unlock the fine-grained learning signals hidden within non-diverse groups, we propose REASONING ARENA, an adaptive training framework that establishes a hybrid-reward learning paradigm. REASONING ARENA detects non-diverse groups online and dynamically routes them to a tournament-based reward system, while preserving the original verifiable reward for diverse groups. This ensures that verifiable rewards remain the gold standard whenever they produce informative advantages; the judge is only invoked precisely where the verifier falls short. For non-diverse reward groups, REASONING ARENA constructs *trace tournaments*: an LLM judge compares the reasoning *traces* head-to-head within the group. Comparing intermediate traces rather than final answers is critical: it allows the judge to distinguish solutions that share identical verifier outcomes but differ in logical soundness and reasoning quality, eliciting fine-grained learning signals that have not been effectively utilized. To make trace tournaments scalable in asynchronous RL, we introduce a live tournament strategy that pairs traces with live opponents, where we maintain a running pool of current best, worst, and median traces as anchors to efficiently establish a relative ranking. We then fit a Bradley-Terry (BT) model over the resulting incomplete comparison graph, enabling robust and efficient advantage estimation without incurring the high inference cost of full quadratic pairwise comparisons.

We evaluate REASONING ARENA extensively across reasoning domains, including competition mathematics and code generation. Our empirical results demonstrate that by effectively extracting gradient signals from non-diverse reward groups, REASONING ARENA consistently improves overall model performance over standard RLVR-only and pure LLM-as-a-judge baselines, preserving out-of-domain generalization. The gains are accompanied by faster training (27% to 41%) and a substantial reduction in generation compute of nearly 50%, since non-diverse reward groups that would otherwise be discarded now provide meaningful gradients. Our cost analysis further shows that the live opponent selection strategy performs competitively with full round-robin tournament evaluations while effectively reducing the quadratic comparison cost to a linear scale.

Our contributions are summarized as follows: (1) We identify the inefficiency and wasted learning signals caused by discarded non-diverse groups in RLVR, and propose adaptive routing between verifiable and judge-based rewards as a principled remedy that preserves verifiable supervision wherever it is informative; (2) We propose REASONING ARENA that introduces *trace* tournaments as the reward mechanism for non-diverse reward groups, seamlessly scaled to RL training via a live tournament with Bradley-Terry estimation on incomplete graphs; (3) We demonstrate empirical gains on math and code reasoning tasks, along with improvements in training efficiency, OOD generalization, and the effective utilization of samples that would otherwise yield zero advantage.

2 RELATED WORK

Reinforcement Learning from Verifiable Rewards. RLVR has emerged as the dominant paradigm for training reasoning models, moving from preference-based alignment (Ouyang et al., 2022; Rafailov et al., 2023) to outcome-based supervision driven by rule-based verification (Guo et al., 2025). Group-relative policy optimization methods (Shao et al., 2024; Zheng et al., 2025; Chen et al., 2025) rely heavily on the group-mean advantage estimator, consequently inheriting a critical zero-variance failure mode on non-diverse reward groups. DAPO (Yu et al., 2025a) introduces dynamic sampling that filters out fully correct or fully incorrect groups during training, but this avoids rather than solves the problem: the generation compute is nevertheless wasted, and it forfeits the potential of learning from these samples. DEPO (Tang et al., 2025) and GRESO (Zheng et al., 2026) similarly skip uninformative prompts at the data-curation or rollout-selection stage. A closer line of work derived signals from non-diverse groups themselves: RL-ZVP (Le et al., 2026) and ZAPO (Wu et al., 2025) repurpose zero-variance prompts via entropy-guided advantage shaping, and RLPR (Yu et al., 2025b) replaces the verifier with a probabilistic surrogate. Since the recovered signal in all these methods is fundamentally intrinsic to the policy itself, it cannot distinguish a rigorous

proof from a hallucinated derivation. REASONING ARENA diverges by introducing an *external* judge reward applied *adaptively*. By composing verifiable and judge rewards at the per-group level within a single RLVR gradient step, non-diverse groups now gain rich gradient signals while the exact verifier continues to drive learning wherever it remains informative.

LLM-as-a-Judge for Rich Rewards. Beyond deterministic verifiers, LLM-as-a-judge has emerged as a flexible alternative for reward formulation. This approach was popularized by RLAIF (Lee et al., 2024) and subsequently refined through pointwise rubric-based scoring (Viswanathan et al., 2026; Gunjal et al., 2026; Shao et al., 2025a). As pointwise judging is sensitive to prompt design and prone to surface-level bias (Liu et al., 2026b; Zhou et al., 2024a), recent work shifts toward pairwise elicitation (Liu et al., 2024; Rezaei et al., 2025) and tournament-based ranking. Precisely, tournament rewards have been applied to non-verifiable domains (Zhang et al., 2026; Feng et al., 2026), including rubric-driven generation (Jia et al., 2026), text-to-image (Wang et al., 2025; Wan et al., 2025), and even text-to-video generations (Long et al., 2025). They all target open-ended domains where no verifier exists, and use tournaments as the sole reward source applied uniformly to every group. REASONING ARENA operates in a fundamentally different setting: reasoning tasks where a verifier exists but falls short of differentiating the intermediate reasoning trajectories of identical solution correctness. Rather than replacing the verifier, we embed tournaments *within* RLVR and route rollouts to them adaptively. By invoking the judge strictly on non-diverse reward groups, REASONING ARENA brings tournament-based judging into the verifiable-reward regime as a targeted complement, recovering the rich relative-ranking signal while preserving the exact verifier elsewhere.

3 PRELIMINARY

Reinforcement Learning from Verifiable Rewards. We consider reinforcement learning for reasoning tasks with verifiable outcomes. Let $x \sim \mathcal{D}$ denote a prompt and π_θ a policy that generates a reasoning trace $y = (a_1, \ldots, a_T)$. During training, a rollout engine samples a group of N traces $\mathcal{G}(x) = \{y_i\}_{i=1}^N$ from the behavior policy π_{old}. With a given binary verifiable reward function $R_v(x, y_i) = V(x, y_i) \in \{0, 1\}$, group-relative methods (Shao et al., 2024) compute an advantage for each trace relative to the other samples within the same prompt group. We adopt CISPO (Chen et al., 2025) as our underlying RLVR algorithm in this work, which clips the importance-sampling weight rather than the policy-gradient term to retain learning signals from rare tokens. The RLVR objective is:

$$\mathcal{J}_{\text{RLVR}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, \mathcal{G}\sim\pi_{\text{old}}}\left[\frac{1}{\sum_{i=1}^N |y_i|}\sum_{i=1}^N\sum_{t=1}^{|y_i|} \text{sg}\big(\hat{\rho}_{i,t}(\theta)\big)\cdot A_i \cdot \log \pi_\theta(a_{i,t} \mid x, y_{i,<t})\right], \quad (1)$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The term $\hat{\rho}_{i,t}(\theta)$ is the clipped token-level importance-sampling weight, and A_i is the group-relative trace advantage that scales every token-level update for trace i:

$$\hat{\rho}_{i,t}(\theta) = \text{clip}\left(\frac{\pi_\theta(a_{i,t} \mid x, y_{i,<t})}{\pi_{\text{old}}(a_{i,t} \mid x, y_{i,<t})},\, 1-\epsilon_{\text{lo}}^{\text{IS}},\, 1+\epsilon_{\text{hi}}^{\text{IS}}\right), \qquad A_i = \frac{R_v(x, y_i) - \mu_{\mathcal{G}}}{\sigma_{\mathcal{G}} + \epsilon}, \quad (2)$$

with $\mu_{\mathcal{G}} = \frac{1}{N}\sum_{j=1}^N R_v(x, y_j)$ and $\sigma_{\mathcal{G}}^2 = \frac{1}{N}\sum_{j=1}^N (R_v(x, y_j) - \mu_{\mathcal{G}})^2$ denoting the within-group reward mean and variance, respectively. The defining characteristic of Eq. equation 2 is that there must be a non-zero advantage for at least one sample in a group to produce a learning signal. Without a non-zero advantage function, the group cannot produce any meaningful gradient contribution.

The Non-Diverse Reward Group Problem. We formalize the reward diversity of a rollout group directly through its within-group reward variance:

$$D(\mathcal{G}) = \mathbb{I}[\,\sigma_{\mathcal{G}} > 0\,]\,. \quad (3)$$

A group is *reward-diverse* when $D(\mathcal{G}) = 1$ and *non-diverse* when $D(\mathcal{G}) = 0$. In the context of binary verifiable rewards, non-diverse groups manifest in exactly two forms: *all-correct* groups, where $R_v(x, y_i) = 1$ for every trace, and *all-incorrect* groups, where $R_v(x, y_i) = 0$ for every trace. By construction, this lack of reward diversity enforces a strict mathematical collapse:

$$D(\mathcal{G}) = 0 \quad \Longleftrightarrow \quad \sigma_{\mathcal{G}} = 0 \quad \Longleftrightarrow \quad A_i = 0 \;\; \forall\, i \in \{1, \ldots, N\}. \quad (4)$$

Consequently, a non-diverse group contributes absolutely no reward-driven policy gradient under Eq. equation 1, regardless of how much the underlying trajectories differ in reasoning steps. This

failure mode dominates two distinct training regimes: *all-incorrect* groups are highly prevalent early in training when the policy remains weak, and *all-correct* groups, which grow as the policy improves (Figure 1). In modern asynchronous RL systems, the batch must still be filled to the target size before a training step can proceed: once a non-diverse group is derived, all of its N already-generated traces are discarded. The engine must then sample additional rollouts to fill the batch, which is both costly and inefficient, as the per-trace generation cost dictates the overall pipeline latency. This structural inefficiency motivates our approach to extract and utilize the reasoning variation within these non-diverse groups, converting otherwise wasted compute into constructive gradient updates.

4 REASONING ARENA

To address the non-diverse reward group problem, REASONING ARENA adaptively routes each rollout group to distinct reward sources based on the variance of the received rewards. The primary motivation is to retain the exactness of the verifiable reward where it provides meaningful gradient signals, while dynamically falling back to a judge system precisely when the verifier falls short. For reward-diverse groups, we maintain the standard RLVR update; conversely, non-diverse groups are routed to a *trace tournament*. Rather than evaluating final answers, this tournament conducts trace-by-trace comparisons to elicit granular signals regarding the reasoning quality of each trajectory. To scale this mechanism for asynchronous RL, we further introduce a live opponent strategy combined with Bradley-Terry fitting, which significantly reduces the computational overhead.



Figure 1: The presence of non-diverse reward groups is prevalent in RLVR. The fraction of non-diverse groups is decomposed into all-incorrect and all-correct groups for training with `Ministral-3-8B-Instruct`. All-wrong groups dominate early training, while all-correct groups increase as the policy improves; both regimes receive zero group-relative advantage under standard RLVR despite having already consumed rollout compute.

Adaptive group routing. The rule-based verifier R_v from Eq. equation 2 is highly efficient and, whenever a group exhibits reward diversity, $D(\mathcal{G}) = 1$, it is capable of separating the generated traces to provide a robust gradient signal. By comparison, LLM judges are computationally expensive and inherently noisier. However, LLM-as-a-judge offers a distinct advantage: it can evaluate intermediate reasoning steps to rank trajectories that share the exact same final answer. This makes the judge perfectly suited for non-diverse reward groups, which is the precise scenario where the verifier falls short and yields zero gradient. To leverage the strengths of both approaches, REASONING ARENA dynamically integrates these two reward mechanisms at the per-group level. It routes to a more informative reward source conditioned on whether the verifier successfully produces within-group variance:

$$R^\star(x, y_i) = \begin{cases} R_v(x, y_i), & \text{if } D(\mathcal{G}) = 1, \\ R_j(x, y_i), & \text{if } D(\mathcal{G}) = 0, \end{cases} \tag{5}$$

where R_j denotes a judge-derived reward (e.g., pointwise scoring or trace tournaments) defined in Section 4. Substituting this into the group statistics of Eq. equation 2, $R^\star$ yields the adaptive objective:

$$\mathcal{J}_{\text{RA}}(\theta) = \sum_{\mathcal{G}: D(\mathcal{G})=1} \mathcal{J}_{\text{RLVR}}(\theta; \mathcal{G}, R_v) + \sum_{\mathcal{G}: D(\mathcal{G})=0} \mathcal{J}_{\text{Arena}}(\theta; \mathcal{G}, R_j). \tag{6}$$

On reward-diverse groups, Eq. equation 6 simply reduces to standard RLVR. The router isolates the judge's intervention strictly to groups that would otherwise contribute zero gradient. Therefore, the judge reward acts as a targeted regularization term that extracts hidden reasoning variations.

From pointwise scoring to trace tournaments. Since every trace in a non-diverse reward group shares an identical verifier outcome, the final answer provides no distinguishing information, forcing the judge to evaluate the underlying reasoning trajectory itself. A straightforward realization of R_j is a pointwise judge evaluating each trace in isolation based on a predefined rubric, which we later evaluate as the `Adaptive Pointwise` baseline. However, absolute pointwise scores are notoriously difficult to calibrate consistently across different traces (Li et al., 2025; Zhang et al.,

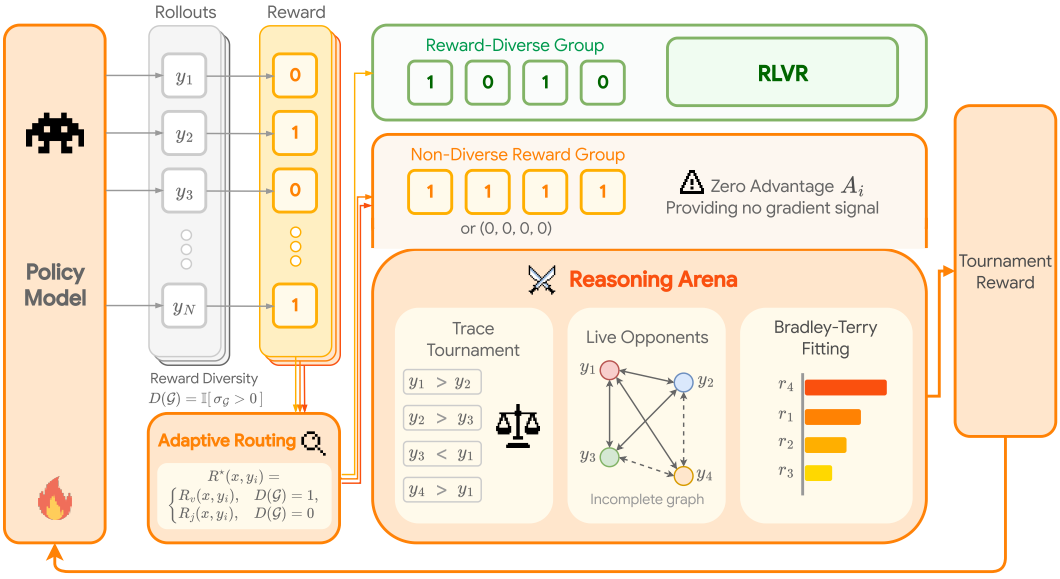

Figure 2: Overview of REASONING ARENA. Rollout groups are adaptively routed based on their outcome diversity. Reward-diverse groups receive standard verifiable rewards. Conversely, non-diverse reward groups with zero advantage are dynamically routed to a trace tournament. By pairing traces with live opponents and fitting a Bradley-Terry model on the incomplete comparison graph, REASONING ARENA efficiently extracts rich relative reward signals from reasoning traces.

2026), and they are susceptible to superficial heuristics, such as length, formatting, and hedging (Zhou et al., 2024a; Liu et al., 2026b), which inadvertently incentivize reward hacking. Motivated by findings that pairwise preferences align more robustly with human judgment (Liu et al., 2024; Zhou et al., 2024a), REASONING ARENA employs pairwise comparisons between traces drawn from the same prompt and policy model.

For any given pair of reasoning trace (y_i, y_j), the judge processes the context (x, y_i, y_j) and outputs a categorical verdict $v \in \{\mathrm{A}, \mathrm{B}, \mathrm{Tie}\}$. We map this verdict to a soft outcome for the first trace:

$$o_{ij} = \begin{cases} \gamma, & v = \mathrm{A}, \\ 1/2, & v = \mathrm{Tie}, \\ 1 - \gamma, & v = \mathrm{B}, \end{cases} \qquad o_{ji} = 1 - o_{ij}, \tag{7}$$

where $\gamma \in (1/2, 1]$ is a soft-margin hyperparameter that interpolates between an uninformative tie ($\gamma = 1/2$) and a deterministic win ($\gamma = 1$). The tie value is strictly constrained to $1/2$ by the antisymmetry condition $o_{ij} + o_{ji} = 1$, which is essential for ensuring unbiased win-rate aggregation and maintaining a valid probability distribution when fitting the Bradley-Terry model.

Order debiasing. Pairwise LLM judges often exhibit strong, prompt-dependent position bias (Zhou et al., 2024b). To mitigate this, we adopt a simple permutation calibration (Wang et al., 2024; Zhou et al., 2024a). During sampling, the presentation order of every queried pair is randomized independently, converting systematic positional offsets into zero-mean noise. During aggregation, each observed match (i, j, o_{ij}) is symmetrically augmented with its mirror $(j, i, 1 - o_{ij})$, explicitly enforcing the antisymmetry $o_{ij} + o_{ji} = 1$ that a biased judge might otherwise violate.

Deriving the tournament reward. A naive tournament approach on a non-diverse group executes a full round-robin format, judging every pair (y_i, y_j) to produce a complete match set of $\binom{N}{2}$ comparisons. The per-trace tournament reward then becomes its win-rate:

$$R_j^{\mathrm{wr}}(y_i) = \frac{1}{N-1} \sum_{j \neq i} o_{ij}. \tag{8}$$

While mathematically unbiased on a complete graph, a round-robin incurs a prohibitive $O(N^2)$ inference cost per group, increasing the staleness in asynchronous RL pipelines under the same judge compute budget as the long-tail trace completes generation and gets paired to comparison late.

To resolve this, REASONING ARENA executes a *live* tournament that is fully parallelized with asynchronous RL. As soon as a new trace y_k finishes generation, the live tournament dynamically selects

three opponents as anchors for comparison from the traces that have already arrived. Specifically, the current best, worst, and median traces are chosen as the live opponents based on a live leaderboard ranked by $R_j^{\mathrm{wr}}(y_i)$ progressively. This setup includes extreme traces that always bracket the newly generated trace, while the median opponent acts as a middle anchor to rank the new generation. This strategy yields informative comparisons at a constant rate, reducing the per-group judging complexity from $O(N^2)$ to $O(N)$.

Bradley-Terry Reward Aggregation. Since the live opponent rule generates an incomplete, non-uniformly sampled comparison graph, utilizing the raw win-rate R_j^{wr} would confound the trace's true reasoning quality with the opponent-selection policy. To resolve this, we define the latent reasoning log-strength β_i of each trace y_i and estimate it using the Bradley-Terry (BT) model (Bradley & Terry, 1952). The probability that trace y_i is preferred over y_j by the judge is defined via the logistic sigmoid function:

$$P(y_i \succ y_j) = \sigma(\beta_i - \beta_j) = \frac{1}{1 + \exp\big(-(\beta_i - \beta_j)\big)}. \tag{9}$$

To robustly fit these strengths, we first construct a symmetrized match set $\widetilde{M}$ by mirroring every observed match: each outcome (i, j, o_{ij}) is explicitly duplicated as $(j, i, 1 - o_{ij})$. We then estimate the strength vector $\boldsymbol{\beta} \in \mathbb{R}^N$ by minimizing the L2-regularized soft cross-entropy loss:

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{(i,j,o_{ij}) \in \widetilde{M}} [o_{ij} \log P(y_i \succ y_j) + (1 - o_{ij}) \log P(y_j \succ y_i)] + \frac{1}{2}\|\boldsymbol{\beta}\|_2^2. \tag{10}$$

This formulation natively supports the continuous soft outcomes $o_{ij} \in [0, 1]$ from Eq. equation 7, preserving the judge's relative rating granularity. The objective is also strictly convex, and we optimize it efficiently using the L-BFGS-B algorithm (Liu & Nocedal, 1989). The L2 penalty ($\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$) shrinks the ratings toward zero, preventing extreme values when the comparison graph is sparse and anchoring the global optimum even on disconnected subgraphs. Finally, the fitted latent log-strengths $\boldsymbol{\beta}^\star$ are min-max normalized to produce the final scalar rewards $R_j \in [0, 1]$.

5 Experiments

Models and datasets. We use `Ministral-3-8B-Instruct-2512` (Liu et al., 2026a) as the policy model for follow-up RL training. For the primary LLM judge, we employ `DeepSeekMath-V2` (Shao et al., 2025b) to evaluate reasoning trajectories in the mathematical domain. To investigate the scalability of the reward source, we conduct additional ablations using `Qwen3-235B-A22B` (Yang et al., 2025) and `Qwen3.5-122B-A10B` as the judge. Our RL training uses the STEM RL data mixture in (Liu et al., 2026a). We explicitly filter out coding and visual-reasoning data during training, ensuring that the code task serves for out-of-distribution (OOD) evaluation. Our evaluation tasks include competition mathematics: AIME 2024/2025/2026, Beyond AIME (ByteDance-Seed, 2025), graduate-level domain reasoning: GPQA-Diamond (Rein et al., 2024), and code reasoning: LiveCodeBench v6 (Jain et al., 2025). We report the average pass ratio@16 for the math and GPQA-Diamond benchmarks, and pass@5 for LiveCodeBench to reduce variance.

Baselines. We compare Reasoning Arena against representative baselines:

- **RLVR** (Chen et al., 2025): We adopt CISPO as the underlying RLVR algorithm, which resamples for non-diverse reward groups, and we directly conduct RL training on the policy model with verifiable rewards only, which serves as the primary baseline.
- **RLAIF** (Lee et al., 2024): We implement a pointwise LLM-as-a-judge baseline with the same policy optimization algorithm. The reward is generated via a DeepSeekMath-style proof-quality rubric, scoring each trace discretely in $\{0, 0.5, 1\}$ based on logical soundness.
- **RL-ZVP** (Le et al., 2026): We implement RL-ZVP with CISPO, which substitutes an intrinsic entropy proxy as the rewards for non-diverse groups. It penalizes low-entropy tokens in all incorrect responses, whereas high-entropy tokens incur smaller penalties.
- **ArenaRL** (Zhang et al., 2026): Learning from a tournament-only reward baseline. This is structurally equivalent to the round-robin tournament in Reasoning Arena applied to all groups without adaptive routing, allowing us to ablate the impact of the routing mechanism.

Table 1: Main results on competition mathematics, scientific reasoning, and code benchmarks. Adaptive routing consistently improves over verifier-only RLVR and judge-only baselines. REASONING ARENA-Live achieves the best average score, improving over RLVR by $+7.6\%$.

Method	AIME 24	AIME 25	AIME 26	Beyond AIME	GPQA-D	LCB v6	Avg.
RLVR (Chen et al., 2025)	58.5	43.8	46.0	28.2	54.8	46.7	46.3
RLAIF (Lee et al., 2024)	53.1	46.9	48.3	27.6	56.9	50.7	47.3
RL-ZVP (Le et al., 2026)	53.5	48.3	51.3	29.5	58.0	50.9	48.6
ArenaRL (Zhang et al., 2026)	60.4	50.2	56.9	31.9	59.6	50.4	51.6
Adaptive Pointwise	62.9	51.5	54.4	31.9	58.5	48.5	51.3
REASONING ARENA	**66.7**	**54.8**	54.6	33.6	59.5	51.6	53.5
REASONING ARENA-Live	63.5	51.7	**59.0**	**36.4**	**60.5**	**52.2**	**53.9**
Δ(w.r.t. RLVR)	**+5.0**	**+7.9**	**+12.9**	**+8.3**	**+5.7**	**+5.6**	**+7.6**



Figure 3: Performance over RL training steps. REASONING ARENA and REASONING ARENA-Live improve steadily on in-domain math benchmarks and preserve gains on GPQA-Diamond and OOD LiveCodeBench, whereas verifier-only RLVR often plateaus or regresses after early training.

In addition to the above baselines, we include `Adaptive Pointwise`, a REASONING ARENA variant, where we replace the tournament reward with a pointwise rubric judge similar to RLAIF, while retaining the adaptive routing mechanism. In terms of our methods, we report the REASONING ARENA, which routes non-diverse reward groups to a round-robin tournament using win-rate reward of Eq. equation 8, and REASONING ARENA-Live, which implements the live opponent selection rule in §4 and Bradley-Terry reward aggregation with $\gamma = 0.75$

Main results. Table 1 summarizes the final evaluation performance. Standard RLVR reaches an average score of 46.3, indicating that verifiable rewards alone are effective but leave substantial learning signal unused. RLAIF improves the average only mildly, suggesting that replacing the verifier everywhere is not optimal in verifiable domains. In contrast, our adaptive routing approaches all give a consistent gain: REASONING ARENA-Live reaches an average performance of 53.9, outperforming RLVR by 7.6 on average. Notably, it improves AIME 2026 by 12.9 points, with consistent improvements on the rest of the tasks, which also substantially surpasses the RL-ZVP baseline that leverages an intrinsic entropy reward for non-diverse reward group learning. This demonstrates that converting hidden signals in non-diverse reward groups from an external source is critical for performance improvements, effectively expanding the usable learning signals while preserving the original verifiable supervision.

Figure 3 compares performance over RL steps. The margin between RLVR and REASONING ARENA widens steadily during training, indicating that non-diverse reward groups provide usable gradients throughout optimization. This divergence is particularly pronounced in later training stages. As the policy improves, all-correct groups become more prevalent. While standard group-relative RLVR assigns zero advantage to these groups, REASONING ARENA successfully extracts fine-grained,

Table 2: Training efficiency and the judge reward costs with a group $N = 8$ and the live tournament with three opponents. α is the fraction of groups that were routed to the trace tournament, so *effective calls* are $\alpha \cdot c$. All metrics are averaged over the entire training. *We note that the step time is strongly correlated with the response length, and the growing response length in RL is a typical pattern that reflects the thinking ability is learning. However, longer responses also nevertheless increase wall-clock step time independently of generation efficiency.

Method	Judge calls c	Routed groups α	Effective calls	Step time	#Gens / step	Perf.
RLVR (Chen et al., 2025)	0	0%	0.0	17.7	908	46.3
RLAIF (Lee et al., 2024)	$N = 8$	100%	8.0	4.2*	712	47.3
RL-ZVP (Le et al., 2026)	0	49%	0	7.5	480	48.6
ArenaRL (Zhang et al., 2026)	$\binom{N}{2} = 28$	100%	28.0	10.5	465	51.6
Adaptive Pointwise	$N = 8$	47%	3.8	10.6	660	51.3
REASONING ARENA	$\binom{N}{2} = 28$	45%	12.5	10.4	459	53.5
REASONING ARENA-Live	18	43%	7.7	13.0 ↓ 4.7	455 ↓ 453	53.9



Figure 4: Training efficiency and generation behavior. Though REASONING ARENA-Live produces long reasoning traces, adaptive routing avoids repeatedly resampling zero-advantage groups, reducing *half* of the number of generations needed per optimizer step. The live tournament keeps wall-clock step time below verifier-only RLVR while retaining the performance gains via tournament rewards.

trace-level preferences. Furthermore, the sustained progress on LiveCodeBench and GPQA-Diamond demonstrates that our adaptive tournament preserves OOD generalization by learning from an external distribution, mitigating the domain overfitting frequently observed in pure verifier-driven training.

Efficiency. Table 2 details the judge-call complexity for the evaluated reward configurations. With a group size of $N = 8$, a full trace tournament dictates $\binom{8}{2} = 28$ judge calls per routed group. The live opponent strategy reduces this to a maximum of 18 $(1+2+3\times5)$ calls while also improving the overall performance average. Because the adaptive routing exclusively targets non-diverse groups, the empirical judge cost is heavily discounted by the fraction of reward-diverse groups present in the batch. Crucially, this adaptive mechanism significantly improves overall training efficiency. In standard RLVR, a non-diverse group yields zero variance and therefore zero advantage, meaning the computational cost of generating those N trajectories is entirely wasted. REASONING ARENA instead reduces the wall-clock step time by 27% to 41% while saving nearly 50% of the generations per step, massively reducing the generation overhead. By converting these previously discarded rollouts into rich tournament signals, REASONING ARENA maximizes sample efficiency and training efficiency with a practically feasible judge costs for scalable RL training.

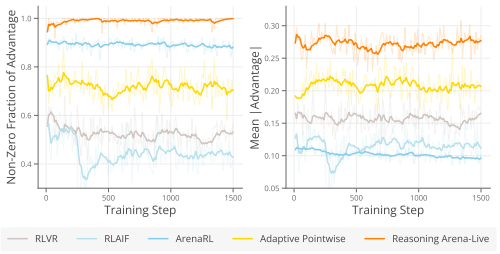

Figure 5: Tournament rewards restore usable advantages on non-diverse reward groups. Left: the fraction of traces with non-zero advantage remains highest for REASONING ARENA-Live across training. Right: REASONING ARENA-Live also yields the largest mean absolute advantage, indicating that trace tournaments provide denser and stronger credit assignment than verifier-only or pointwise judge rewards.

Ablating the adaptive routing mechanism. The comparison between ArenaRL and REASONING ARENA highlights the effect of adaptive routing. ArenaRL applies tournament rewards uniformly and

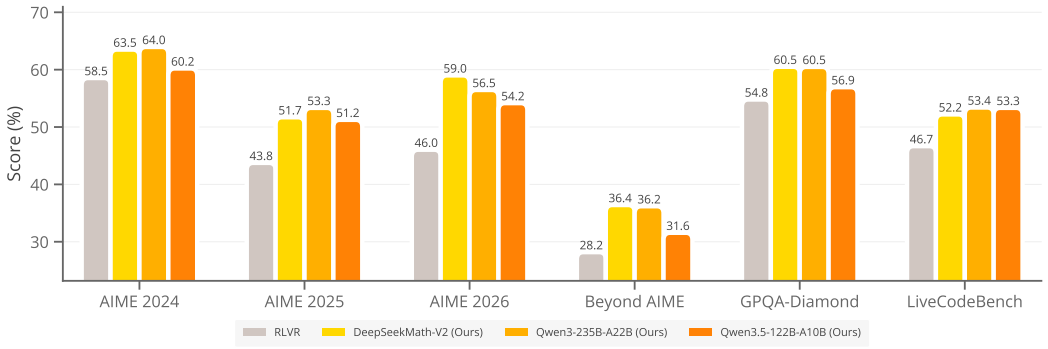

Figure 6: Extending the judge-model in REASONING ARENA-Live to more model families. All judge choices substantially improve over RLVR, showing the robustness of REASONING ARENA.

improves over RLVR, but it discards the special reliability of exact verifiers on reward-diverse groups. REASONING ARENA instead uses tournaments only when the verifier provides no within-group contrast. This strategic routing yields higher average performance than ArenaRL while concentrating computational resources on a targeted subset of the data. Figure 5 reflects the sample efficiency in RL by different learning strategies. It is notable that though ArenaRL comes with high non-zero fraction of advantage samples, the L1-norm is relatively small, whereas REASONING ARENA effectively improves the sample utilization to nearly 100% while providing denser reward signals. This observation further consolidates the adaptive routing mechanism in both improving training and sample efficiency.

Tournaments vs. pointwise scoring. The comparison of REASONING ARENA against the Adaptive Pointwise ablates the form of the routed reward source. Both methods employ the same adaptive routing, differing only in their scoring of non-diverse reward groups. Trace tournaments improve the average from 51.3 to 53.5 with full tournaments. This gap suggests that merely introducing an LLM judge is insufficient, but the form of reward evaluation is more critical. Pointwise scoring forces the judge to assign an absolute numerical value in isolation. This makes it notoriously difficult to calibrate across traces without exhaustive rubrics, leaving it vulnerable to superficial heuristics. Pairwise comparison, however, inherently grounds the evaluation by presenting the judge with two trajectories sharing identical verifiable outcomes. This naturally anchors the baseline, forcing the reward distribution to reflect true relative reasoning quality rather than formatting artifacts.

Effectiveness of different judge models. We further investigate the robustness of our approach with respect to various judge models. Replacing `DeepSeekMath-V2 (685B)` in REASONING ARENA with broader models, such as `Qwen3-235B-A22B` and `Qwen3.5-122B-A10B`, reveals a scaling behavior. Higher-capacity models exhibit superior consistency in their pairwise preferences, particularly when judging complex, multi-step logical derivations. While `DeepSeekMath` is highly effective for mathematical domains, the `Qwen` variants also provide robust reward signals for non-diverse reward groups, showing that REASONING ARENA is robust to judge models.

6 CONCLUSION

We introduced REASONING ARENA, an adaptive training framework designed to resolve the non-diverse reward group problem in RLVR. Standard group-relative methods fail to provide gradient signals when a deterministic verifier assigns identical outcomes to all traces in a sampled group. Rather than discarding these non-diverse reasoning traces, REASONING ARENA adaptively routes them to a head-to-head trace tournament, leveraging an LLM judge to extract fine-grained relative preferences from the intermediate reasoning steps. To scale this mechanism for asynchronous RL training, we developed a live opponent selection strategy coupled with Bradley-Terry reward aggregation. This allows the framework to infer consistent rewards from sparse comparison graphs without the prohibitive cost of full quadratic evaluations. Empirical results across mathematical and code reasoning tasks demonstrate that REASONING ARENA consistently outperforms standard RLVR and pointwise judge baselines. By converting otherwise wasted zero-advantage samples into informative gradient updates, REASONING ARENA provides a principled, compute-efficient, and sample-efficient method for composing exact verification with LLM judges.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing and correcting grammar errors. We have not used generative AI tools for all other purposes.

REPRODUCIBILITY STATEMENT

We fully disclose all the prompt templates in Appendix. E. All models used in this work are accessible as open-source models, and we include all discussions about the implementation details in both Sec. 5 and Appendix. B

REFERENCES

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

ByteDance-Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads. *Hugging Face repository*, 2025.

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

Dylan Feng, Bhavesh Kumar, and Leonard Tang. Tourno: Tournament optimization for reinforcement learning in non-verifiable domains, 2026. URL https://www.haizelabs.com/blog/tourno.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=c1bTcrDmt4.

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

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Ruipeng Jia, Yunyi Yang, Yuxin Wu, Yongbo Gai, Siyuan Tao, Mengyu Zhou, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Open rubric system: Scaling reinforcement learning with pairwise adaptive rubric. *arXiv preprint arXiv:2602.14069*, 2026.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Thanh-Long V. Le, Myeongho Jeon, Kim Vu, Viet Dac Lai, and Eunho Yang. No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kiXFIESZKv.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, pp. 26874–26901, 2024. URL https://proceedings.mlr.press/v235/lee24t.html.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. Large language models are miscalibrated in-context learners. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11575–11596, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.603. URL https://aclanthology.org/2025.findings-acl.603/.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026a.

Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989. doi: 10.1007/BF01589116. URL https://doi.org/10.1007/BF01589116.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=9gdZI7c6yr.

Yixin Liu, Yue Yu, DiJia Su, Sid Wang, Xuewei Wang, Song Jiang, Bo Liu, Arman Cohan, Yuandong Tian, and Zhengxing Chen. Examining reasoning llms-as-judges in non-verifiable llm post-training. *arXiv preprint arXiv:2603.12246*, 2026b.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö Arık. Vista: A test-time self-improving video generation agent. *arXiv preprint arXiv:2510.15831*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*, 2025.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025a.

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

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

Xinyu Tang, Zhenduo Zhang, Yurou Liu, Wayne Xin Zhao, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Towards high data efficiency in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2509.01321*, 2025.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=RPRqKhjrr6.

Xingchen Wan, Han Zhou, Ruoxi Sun, Hootan Nakhost, Ke Jiang, Rajarishi Sinha, and Sercan Ö Arık. Maestro: Self-improving text-to-image generation via agent orchestration. *arXiv preprint arXiv:2509.10704*, 2025.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025.

Guanzhong Wu, Weiduo Liao, Caigao JIANG, Xingyu Lu, Lintao Ma, and Ying Wei. Useless, or untapped? unlocking the full value of zero-advantage samples for better policy optimization, 2025. URL https://openreview.net/forum?id=FPVB4qSXCZ.

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

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

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

Qiang Zhang, Boli Chen, Fanrui Zhang, Ruixue Ding, Shihang Wang, Qiuchen Wang, Yinfeng Huang, Haonan Zhang, Rongxiang Zhu, Pengyong Wang, et al. Arenarl: Scaling rl for open-ended agents via tournament-based relative ranking. *arXiv preprint arXiv:2601.06487*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for LLM reasoning via selective rollouts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=x5lITYXmW2.

Han Zhou, Xingchen Wan, Yinhong Liu, Nigel Collier, Ivan Vulić, and Anna Korhonen. Fairer preferences elicit improved human-aligned large language model judgments. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1241–1252, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.72. URL https://aclanthology.org/2024.emnlp-main.72/.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=L3FHMoKZcS.

A Limitations and Future Work

Reasoning Arena is an adaptive reward framework orthogonal to specific model architectures and reinforcement learning algorithms. By extracting fine-grained signals from non-diverse reward groups, it yields substantial improvements over purely verifiable rewards. While our proposed adaptive routing effectively recovers these otherwise wasted signals, we note that the inherent prevalence of non-diverse reward groups remains dependent on the underlying data mixture and the base capabilities of the policy model. Nevertheless, Reasoning Arena mitigates the computational burden of exhaustively prefiltering or re-grading the training corpus to construct a perfectly difficulty-matched dataset, which is itself costly and difficult to process per policy model to train.

Due to the strict latency and off-policy constraints inherent to asynchronous RL, passing complete reasoning trajectories can easily exceed 160,000 input tokens per pair. Consequently, Reasoning Arena currently relies on the concise reasoning traces natively generated within the policy model's response trunk. In future work, intelligently incorporating truncated reasoning tokens could further enhance the judge's evaluation accuracy. Furthermore, the applicability of Reasoning Arena could be naturally expanded to tool-use agents, where tens to hundreds of intermediate tool calls and high-level planning steps can serve as the reasoning trace for the judge to evaluate.

Although our approach significantly accelerates overall sample and training efficiency, it operates at the cost of additional GPU resources or API compute required to serve the judge models. This establishes a practical tradeoff among runtime acceleration, generation efficiency, and total resource allocation. Overall, Reasoning Arena serves as a highly generalizable approach for overcoming the non-diverse reward group problem of verifiable rewards, offering a systematic framework for efficient RLVR optimization.

B Implementation Details

Models and Datasets. In this work, we use the following models and datasets, all of which are released for research use under their respective licenses.

We use 1) `Ministral-3-8B-Instruct-2512` Liu et al. (2026a) (Mistral Research License: `https://mistral.ai/licenses/MRL-0.1.md`) as the policy model trained with online RL; 2) `DeepSeekMath-V2` Shao et al. (2025b) (Apache 2.0 License: `https://huggingface.co/deepseek-ai/DeepSeek-Math-V2/blob/main/LICENSE`) as the LLM judge for pairwise tournament and pointwise scoring; 3) `Qwen3-235B-A22B` Yang et al. (2025) (Apache 2.0 License: `https://huggingface.co/Qwen/Qwen3-235B-A22B/blob/main/LICENSE`) and `Qwen3.5-122B-A10B` (Apache 2.0 License: `https://huggingface.co/Qwen/Qwen3.5-122B-A10B/blob/main/LICENSE`) as alternative judge backbones for ablation studies on judge families.

We evaluate on 1) AIME 2024, AIME 2025, and AIME 2026 (problems copyrighted by the Mathematical Association of America) as the competition-level mathematics benchmarks; 2) BeyondAIME ByteDance-Seed (2025) (CC0 1.0 Universal Public Domain Dedication: `https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME`); 3) GPQA-Diamond Rein et al. (2024) (CC-BY-4.0 License: `https://github.com/idavidrein/gpqa/blob/main/LICENSE`) as a graduate-level Google-proof Q&A benchmark; and 4) LiveCodeBench Jain et al. (2025) (MIT License: `https://github.com/LiveCodeBench/LiveCodeBench/blob/main/LICENSE`) as the coding benchmark.

Setups. We train Ministral 3 8B Instruct using FP8 quantization with a maximum sequence length of $81,920$, a constant learning rate at 4e-7 for 1500 optimization steps with 50 warm-up steps; each minibatch contains 480 RL sampled rollouts, and the group size is $N = 8$, so 60 groups per batch. We set $\epsilon_{\text{hi}}^{\text{IS}} = 2.0$ and $\epsilon_{\text{lo}}^{\text{IS}} = 1.0$ for one-sided clipping. A full round-robin tournament requires $\binom{8}{2} = 28$ judge inferences per non-diverse group, while Reasoning Arena-Live with three opponents necessitates at most 18 $(1 + 2 + 3 * 5)$ inferences, such that early traces will only compare to the available opponents, and later traces compare against three anchors. The judge operates with a $32,768$-token context window and receives the ground-truth answer for reference as part of its prompt context. The judge is accessed by API calls. We refer to live algorithm in 1 and the prompt templates in App. E.

Table 3: Derived resources and modeled costs for 1500 steps at \$4 per H200 GPU-hour, with judge costs estimated at 0.5\$/1.6\$ per 1M input/output tokens. Costs are rounded to \$10; green arrows show total-cost reductions relative to RLVR.

Method	Score	Gens. (10^3)	Days	GPU hours	Judge calls	GPU cost(\$)	Judge est.(\$)	Total est.(\$)	
RLVR	46.3	1362.0	18.44	10620	0.0	42,480	0	42,480	
RLAIF	47.3	1068.0	4.38	2520	8.0	10,080	11,620	21,700	
RL-ZVP	48.6	720.0	7.81	4500	0.0	18,000	0	18,000	
ArenaRL	51.6	697.5	10.94	6300	28.0	25,200	26,560	51,760	
Adaptive Pointwise	51.3	990.0	11.04	6360	3.8	25,440	5,120	30,560	
REASONING ARENA	53.5	688.5	10.83	6240	12.5	24,960	11,700	36,660	↓ 13.7%
REASONING ARENA-Live	53.9	682.5	13.54	7800	7.7	31,200	7,150	38,350	↓ 9.7%

Algorithm 1 REASONING ARENA-Live for one prompt x.

Require: Group $\mathcal{G}$ of N concurrently generated traces; verifier R_v and judge.
Completed traces $\mathcal{S} \leftarrow \emptyset$; comparison results $\mathcal{M} \leftarrow \emptyset$.
for each trace y_k as it completes generation and verification **do**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{y_k\}$.
  if $D(\mathcal{S}) = 0$ **then**
    Select up to three anchors from $\mathcal{S} \setminus \{y_k\}$: current best, worst, and median.
    Compare y_k with the anchors asynchronously; update $\mathcal{M}$ and the leaderboard as results arrive.
  end if
end for
if $D(\mathcal{G}) = 1$ **then**
  Discard judge results, cancel pending judge tasks, and set $R^\star = R_v$.
else
  Await pending comparisons; set $R^\star = R_j$ using Bradley–Terry aggregation on $\mathcal{M}$.
end if
Compute group-relative advantages from $R^\star$ and pass the group to the trainer.

Computation costs. 1) The policy model is served on a single node of 8×H200 GPUs as a generator and is bottlenecked by the number of generations that are wasted due to the discarded non-diverse groups, which is reflected in Figure 4 (c). Reasoning Arena saved 50% of this generator compute. 2) In asynchronous RL, the trainer is bottlenecked by the step time, which defines the GPU hours requested for running a complete run, and we serve the trainer on two nodes of 8×H200 GPUs, and REASONING ARENA accelerates training by 34% on average. 3) The judge is served independently as API requests, and Reasoning Arena is trading concurrent API requests (calls × groups × steps, as in Table 2) for saving the computation resources in the trainer and generator, which acts as a trade-off. In total, for each REASONING ARENA-live run, we spent about 4.95B estimated API tokens, 2,820 fewer local H200 GPU-hours (26.6%), and 9.7% lower estimated total cost relative to RLVR, and most importantly, together with 7.6% performance enhancement in average.

C ADDITIONAL EXPERIMENTS

Analysis of the trace tournament mechanism. We inspected three logged comparisons with identical final answers but checkable differences in intermediate reasoning (Appendix D). In the both-correct pair (Appendix D.1), the judge identifies an invalid algebraic substitution and an unsupported step, preferring the response that correctly derives the same solution set. In the triangle-counting pair (Appendix D.2), both responses incorrectly answer 1 rather than 8. The judge prefers the response that derives $(s-a)(s-b)(s-c) = 16s$ from Heron's formula, while explicitly rejecting its false uniqueness claim. Our independent finite enumeration uses this valid reduction to recover all eight triangles. In the tetrahedron pair (Appendix D.3), both responses give the same wrong volume, but the judge recognizes a correct difference-body reduction and volume scaling while identifying the remaining erroneous factor. These preferences align with verified differences in intermediate reasoning even when final-answer correctness and any answer-only distance to the

reference are identical within each pair. They provide qualitative evidence that the tournament can recover reasoning-sensitive signals; they do not estimate how frequently its preferences are correct across training.

Ablating all-correct and all-incorrect non-diverse reward groups. Non-diverse groups come in two forms: all-correct and all-incorrect. We rerun both forms of reward source, and our experiments show that the gain from REASONING ARENA is dominated by the reward signals learned from all-incorrect groups. This reflects the practice in preparing the training mixture for a target model, such that the desired data distribution ranges from medium- to hard-level questions, whereas easy questions that are commonly solvable have been mostly pre-filtered via a grading pipeline.

Precisely, the average ratio between all-incorrect groups vs. all-correct groups in our runs is $35\% : 9\%$, which represents almost four times the ratio of groups that constitute a gradient step. Meanwhile, all-correct groups only represent 5% (see Fig. 1) in the initial range of training steps. Therefore, relying only on the all-correct groups will still keep most generations discarded, and consequently, we observe that it only has a marginal impact on the overall training, while not reaching the same efficiency improvement as the combined REASONING ARENA achieved. Overall, a more controlled experiment will require a careful balancing of all-easy and all-hard samples in the pre-defined training mixture, which we will endeavor to include in the future work.

D A DEMO OF TRACE TOURNAMENT

We present three logged comparisons in which the responses give identical final answers but differ in substantive reasoning steps. The first pair gives the same correct solution set; in each of the other two pairs, both responses give the same incorrect numerical answer. Thus, within each pair, neither final-answer correctness nor any answer-only distance to the reference can explain the preference. The two both-incorrect cases distinguish valid intermediate reductions from unsupported conclusions, while the judge explicitly identifies the preferred response's remaining error. These selected cases illustrate reasoning-sensitive feedback; they do not estimate its prevalence or the judge's overall reliability. The listings reproduce the original questions, reference answers, complete responses, and recorded judge rationales and verdicts without abridgment. Mathematical checks outside the listings provide our analysis of each case.

D.1 BOTH ANSWERS CORRECT: DETECTING AN INVALID ALGEBRAIC DERIVATION

Question.

```
[user]: Solve the system of equations in $\mathbb{R}$: $x+y=z, x^{2}+y^{2}=z, x^{3}+y^{3}=z$.
Put the final answer in \boxed{}.
```

Reference answer.

```
(0,0,0),(0,1,1),(1,0,1),(1,1,2)
```

Response A.

```
To solve the system of equations:
1. \(x + y = z\),
2. \(x^2 + y^2 = z\),
3. \(x^3 + y^3 = z\),

we start by expressing \(z\) in terms of \(x\) and \(y\) from the first equation:
\[ z = x + y. \]

Substituting \(z\) into the second equation, we get:
\[ x^2 + y^2 = x + y. \]

Using the identity \(x^2 + y^2 = (x + y)^2 - 2xy\), we substitute and simplify:
\[ (x + y)^2 - 2xy = x + y, \]
\[ z^2 - 2xy = z, \]
\[ 2xy = z^2 - z, \]
\[ xy = \frac{z^2 - z}{2}. \]

Next, we substitute \(z = x + y\) into the third equation:
\[ x^3 + y^3 = x + y. \]

Using the identity \(x^3 + y^3 = (x + y)(x^2 - xy + y^2)\), we substitute and simplify:
\[ (x + y)(x^2 - xy + y^2) = x + y, \]
\[ x^2 - xy + y^2 = 1 \quad \text{(assuming \(x + y \neq 0\))}. \]
```

```
We now have a quadratic equation in terms of \(z\):
\[ xy = \frac{z^2 - z}{2}, \]
\[ x^2 - xy + y^2 = 1. \]

Substituting \(xy\) from the expression for \(xy\) into the quadratic equation:
\[ z^2 - \frac{z^2 - z}{2} = 1, \]
\[ 2z^2 - z^2 + z = 2, \]
\[ z^2 + z - 2 = 0. \]

Solving the quadratic equation:
\[ z^2 + z - 2 = 0, \]
\[ z = \frac{-1 \pm \sqrt{1 + 8}}{2} = \frac{-1 \pm 3}{2}, \]
\[ z = 1 \quad \text{or} \quad z = -2. \]

Since \(z\) represents a real number, we discard \(z = -2\) and consider \(z = 1\) and \(z =
    0\).

For \(z = 0\):
\[ x + y = 0, \]
\[ xy = 0. \]
This gives the solution \((0, 0, 0)\).

For \(z = 1\):
\[ x + y = 1, \]
\[ xy = 0. \]
This gives the solutions \((0, 1, 1)\) and \((1, 0, 1)\).

For \(z = 2\):
\[ x + y = 2, \]
\[ xy = 1. \]
This gives the solution \((1, 1, 2)\).

Thus, the real solutions to the system of equations are:
$\boxed{(0, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 2)}$.
```

Response B.

```
To solve the system of equations in \(\mathbb{R}\):
\[ x + y = z, \]
\[ x^2 + y^2 = z, \]
\[ x^3 + y^3 = z, \]

we start by expressing \(z\) in terms of \(x\) and \(y\) from the first equation:
\[ z = x + y. \]

Next, we use the identity for the sum of squares:
\[ x^2 + y^2 = (x + y)^2 - 2xy. \]
Substituting \(z = x + y\) into \(x^2 + y^2 = z\), we get:
\[ z^2 - 2xy = z, \]
which simplifies to:
\[ 2xy = z^2 - z, \]
\[ xy = \frac{z^2 - z}{2}. \]

For the third equation, we use the identity for the sum of cubes:
\[ x^3 + y^3 = (x + y)^3 - 3xy(x + y). \]
Substituting \(z = x + y\) and \(xy = \frac{z^2 - z}{2}\) into \(x^3 + y^3 = z\), we get:
\[ z = z^3 - 3 \left( \frac{z^2 - z}{2} \right) z, \]
\[ z = z^3 - \frac{3z^3 - 3z^2}{2}, \]
\[ z = \frac{2z^3 - 3z^3 + 3z^2}{2}, \]
\[ z = \frac{-z^3 + 3z^2}{2}, \]
\[ 2z = -z^3 + 3z^2, \]
\[ z^3 - 3z^2 + 2z = 0, \]
\[ z(z^2 - 3z + 2) = 0, \]
\[ z(z - 1)(z - 2) = 0. \]

This gives us three possible values for \(z\): \(z = 0\), \(z = 1\), and \(z = 2\).

For each value of \(z\), we find the corresponding \(x\) and \(y\):

1. If \(z = 0\):
   \[ x + y = 0, \]
   \[ xy = \frac{0^2 - 0}{2} = 0. \]
   This implies \(x = 0\) and \(y = 0\), so the solution is \((0, 0, 0)\).

2. If \(z = 1\):
   \[ x + y = 1, \]
   \[ xy = \frac{1^2 - 1}{2} = 0. \]
   This implies either \(x = 0\) and \(y = 1\), or \(x = 1\) and \(y = 0\). So, the solutions
     are \((0, 1, 1)\) and \((1, 0, 1)\).
```

```
3. If \(z = 2\):
   \[ x + y = 2, \]
   \[ xy = \frac{2^2 - 2}{2} = 1. \]
   This implies \(x = 1\) and \(y = 1\), so the solution is \((1, 1, 2)\).

Thus, the complete set of real solutions to the system is:
\[
\boxed{(0, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 2)}
\]
```

Judge rationale and verdict.

```
**Analysis**

Both responses aim to solve the system

\[
x+y=z,\quad x^{2}+y^{2}=z,\quad x^{3}+y^{3}=z
\]

and both list the same four solutions \((0,0,0),\;(0,1,1),\;(1,0,1),\;(1,1,2)\).
However, the reasoning quality differs significantly.

**Response A** makes a critical algebraic mistake.
After obtaining \(xy=\frac{z^{2}-z}{2}\) and \(x^{2}-xy+y^{2}=1\) (under the assumption \(x+y\
    neq0\)), it attempts to substitute \(xy\) into the second equation but writes

\[
z^{2}-\frac{z^{2}-z}{2}=1,
\]

which is incorrect.  (The correct expression should involve either \(z-xy\) (using \(x^{2}+y
    ^{2}=z\)) or \(z^{2}-3xy\) (using \((x+y)^{2}=z^{2}\)).)
From that erroneous equation it deduces \(z^{2}+z-2=0\), giving \(z=1\) or \(z=-2\).
It then separately considers the case \(x+y=0\) (\(z=0\)) - which is necessary because
    division by \(x+y\) was performed - and obtains \((0,0,0)\).
Surprisingly, after solving only for \(z=0\) and \(z=1\), it adds the solution for \(z=2\)
    without any derivation, simply stating "For \(z=2\) ...".  This addition is unjustified
    and makes the solution incomplete and logically flawed.

**Response B** proceeds cleanly.  It substitutes \(z=x+y\) and uses the identities

\[
x^{2}+y^{2}=(x+y)^{2}-2xy,\qquad x^{3}+y^{3}=(x+y)^{3}-3xy(x+y)
\]

to obtain \(xy=\frac{z^{2}-z}{2}\) and then

\[
z = z^{3}-3\!\left(\frac{z^{2}-z}{2}\right)\!z .
\]

Simplifying yields \(z(z-1)(z-2)=0\), correctly giving the three possibilities \(z=0,1,2\).
    For each value the corresponding \(x,y\) are found from \(x+y=z\) and \(xy=\frac{z^{2}-z
    }{2}\), producing all four solutions in a systematic, error-free manner.

**Conclusion**

Response B is mathematically correct, logically coherent, and fully justified.  Response A
    contains algebraic errors and an unjustified step, making it clearly inferior.  Hence the
     better response is **B**.

\[
\boxed{B}
\]
```

Mathematical check

Using $xy = (z^2 - z)/2$ in the cubic identity yields $z(z-1)(z-2) = 0$, from which the four listed solutions follow. A incorrectly replaces $x^2 - xy + y^2$ by $z^2 - xy$, whereas the correct expression is $z^2 - 3xy$. At the valid solution $(1, 1, 2)$, A's substituted expression $z^2 - (z^2 - z)/2$ equals 3, not 1. The judge identifies this error and the unsupported introduction of $z = 2$. Its preference therefore tracks a checkable derivation error despite identical correct final answers.

D.2 Both answers incorrect: a valid reduction toward an exact count

Question.

```
[user]: How many different triangles with integer side lengths are there such that the sum of
     the lengths of any two sides exceeds the length of the third side by at least 5 units,
     and that the area is numerically twice the perimeter? (Two triangles are regarded to be
     the same if they are congruent.)
Put the final answer in \boxed{}.
```

Reference answer.

```
8
```

Response A.

```
To solve the problem, we need to find all integer-sided triangles that satisfy two conditions:
1. The sum of any two sides exceeds the length of the third side by at least 5 units.
2. The area of the triangle is numerically twice the perimeter.

Firstly, we identify the triangle inequalities with the given stricter condition. For a
     triangle with sides \(a\), \(b\), and \(c\), the conditions are:
\[ a + b \geq c + 5 \]
\[ a + c \geq b + 5 \]
\[ b + c \geq a + 5 \]

We then consider the area condition:
\[ \text{Area} = 2 \times \text{Perimeter} \]

Using Heron's formula for the area of a triangle, we have:
\[ \text{Area} = \sqrt{s(s-a)(s-b)(s-c)} \]
where \(s = \frac{a+b+c}{2}\) is the semi-perimeter. The condition for the area can be written
      as:
\[ \sqrt{s(s-a)(s-b)(s-c)} = 2 \times (a+b+c) \]
which simplifies to:
\[ s(s-a)(s-b)(s-c) = 16s^2 \]
\[ (s-a)(s-b)(s-c) = 16s \]

We then systematically check possible values of \(s\) and the corresponding values of \(a\),
     \(b\), and \(c\) to satisfy both conditions. After extensive checking, we find that the
     only set of sides that satisfy both the stricter triangle inequality and the area being
     twice the perimeter is \(13\), \(14\), and \(15\).

We verify this triangle:
- The perimeter is \(13 + 14 + 15 = 42\).
- The semi-perimeter is \(s = 21\).
- Using Heron's formula, the area is:
\[ \text{Area} = \sqrt{21(21-13)(21-14)(21-15)} = \sqrt{21 \times 8 \times 7 \times 6} = \sqrt
     {7056} = 84 \]
- Since \(2 \times \text{Perimeter} = 84\), the area condition is satisfied.

Thus, the only integer-sided triangle that satisfies both conditions is the triangle with
     sides \(13\), \(14\), and \(15\).

$\boxed{1}$
```

Response B.

```
To solve the problem, we need to find triangles with integer side lengths that satisfy two
     conditions:
1. The sum of the lengths of any two sides exceeds the length of the third side by at least 5
     units.
2. The area of the triangle is numerically equal to twice the perimeter.

We start by identifying triangles that meet the first condition and then check if they also
     satisfy the second condition. The triangle (13, 14, 15) is a candidate because it is
     known to satisfy the geometric constraints involved.

First, we verify that the triangle (13, 14, 15) meets the first condition:
- \(13 + 14 = 27 \geq 15 + 5 = 20\)
- \(13 + 15 = 28 \geq 14 + 5 = 19\)
- \(14 + 15 = 29 \geq 13 + 5 = 18\)

Next, we check the area and perimeter to see if the second condition is satisfied:
- The semi-perimeter \( s = \frac{13 + 14 + 15}{2} = 21 \)
- The area \( A = \sqrt{s(s-a)(s-b)(s-c)} = \sqrt{21 \cdot 8 \cdot 7 \cdot 6} = \sqrt{7056} =
     84 \)
- The perimeter \( P = 13 + 14 + 15 = 42 \)
- Twice the perimeter is \( 2P = 84 \)

Since the area is equal to twice the perimeter, the triangle (13, 14, 15) satisfies both
     conditions.

After thorough searching and verification, we conclude that there is only one such triangle up
      to congruence.
```

```
$\boxed{1}$
```

Judge rationale and verdict.

```
**Analysis**

Both solution attempts give the answer \(\boxed{1}\), which contradicts the known correct
    answer (8).  Hence both are incorrect.  However, we can compare their reasoning quality.

- **Response A** explicitly states the stricter triangle inequalities, correctly applies Heron
    's formula, and derives the algebraic condition \((s-a)(s-b)(s-c)=16s\).  This derivation
     is mathematically sound and reduces the problem to a Diophantine equation that could be
    systematically explored.  The response then *claims* to have performed a systematic check
     and found only \((13,14,15)\).  Although the claim of uniqueness is false and no
    checking details are shown, the presence of the derived equation demonstrates a more
    structured approach and a deeper engagement with the problem.

- **Response B** merely verifies that the triangle \((13,14,15)\) satisfies the two given
    conditions and asserts, without any supporting mathematics, that it is the only one.  No
    derivation or reduction is offered, and there is no indication of how one would search
    for other possibilities.  The reasoning is essentially "pick a known triangle and claim
    it's unique".

While both answers are wrong, **Response A** is superior because it contains a correct and
    useful algebraic simplification that moves toward a complete solution, whereas Response B
     is little more than a guess.  Therefore, Response A is the better attempt.

**Verdict**
\(\boxed{A}\)
```

Mathematical check

With semiperimeter $s = (a+b+c)/2$, the area condition is $\Delta = 4s$. Heron's formula therefore gives A's valid reduction $(s-a)(s-b)(s-c) = 16s$. Set $u = b+c-a$, $v = a+c-b$, and $w = a+b-c$. These are integers at least 5 and have the same parity; the reduction becomes

$$uvw = 64(u+v+w).$$

Relabeling the sides so that $u \le v \le w$ gives $64 < uv \le 192$, and hence the finite search

$$5 \le u \le 13, \qquad u \le v \le \left\lfloor \frac{192}{u} \right\rfloor, \qquad w = \frac{64(u+v)}{uv-64}.$$

Keeping integer $w \ge v$ of the same parity and recovering the sides as $(v+w)/2$, $(u+w)/2$, and $(u+v)/2$ yields exactly eight triangles:

$$(9,40,41), \quad (9,75,78), \quad (10,24,26), \quad (10,35,39),$$
$$(11,25,30), \quad (12,16,20), \quad (13,14,15), \quad (15,15,24).$$

All satisfy the stated inequalities and area condition. For example, $(12,16,20)$ has area 96 and perimeter 48, immediately disproving both responses' claim that $(13,14,15)$ is unique. A does not carry out the complete search, but its reduction provides a correct route to it. The judge selects A for that valid intermediate step while explicitly identifying the false uniqueness claim. This distinguishes mathematical progress from verification of a single candidate despite identical wrong final counts.

D.3 Both answers incorrect: a valid reduction with a mistaken factor

Question.

```
[user]: Two regular tetrahedra with edge lengths of \(\sqrt{2}\) are given, which are
    transformed into each other by central symmetry. Let \(\phi\) be the set of midpoints of
    segments whose endpoints belong to different tetrahedra. Find the volume of the figure
    \(\phi\).
Put the final answer in \boxed{}.
```

Reference answer.

```
\frac{5}{6}
```

Response A.

```
To find the volume of the set of midpoints \(\phi\) of segments joining points from one
     regular tetrahedron to another, where the second tetrahedron is the central reflection of
      the first, we need to consider the geometric properties of the tetrahedra and their
     midpoints.

A regular tetrahedron with edge length \(\sqrt{2}\) has a volume given by:
\[ V = \frac{a^3 \sqrt{2}}{12} \]
where \(a = \sqrt{2}\). Substituting \(a = \sqrt{2}\) into the formula, we get:
\[ V = \frac{(\sqrt{2})^3 \sqrt{2}}{12} = \frac{2\sqrt{2} \cdot \sqrt{2}}{12} = \frac{2 \cdot
     2}{12} = \frac{4}{12} = \frac{1}{3} \]

The set of midpoints \(\phi\) of segments joining points from the two tetrahedra forms a
     convex hull. Given the symmetry and the central reflection nature of the transformation,
     the volume of \(\phi\) is equal to the volume of the original tetrahedron, which is \(\
     frac{1}{3}\).

Therefore, the volume of the set of midpoints \(\phi\) is:
\[
\boxed{\dfrac{1}{3}}
\]
```

Response B.

```
To solve the problem, we start by noting that the volume of a regular tetrahedron with edge
     length \(a\) is given by the formula \(V = \frac{a^3}{6\sqrt{2}}\). For \(a = \sqrt{2}\),
      the volume \(V\) of the tetrahedron is:

\[
V = \frac{(\sqrt{2})^3}{6\sqrt{2}} = \frac{2\sqrt{2}}{6\sqrt{2}} = \frac{2}{6} = \frac{1}{3}.
\]

The problem involves two regular tetrahedra, one of which is a central symmetric image of the
     other. The set \(\phi\) is the set of all midpoints of segments connecting a point in the
      first tetrahedron to a point in the second tetrahedron. This set \(\phi\) can be
     represented as \((\text{T1} - \text{T1})/2\), where \(\text{T1}\) is the first
     tetrahedron and \(\text{T2}\) is its central symmetric image.

For a convex body \(K\), the volume of \(K - K\) is known to be \(8 \times \text{vol}(K)\) in
     three dimensions. Therefore, the volume of \(\text{T1} - \text{T1}\) is:

\[
\text{vol}(\text{T1} - \text{T1}) = 8 \times \text{vol}(\text{T1}) = 8 \times \frac{1}{3} = \
     frac{8}{3}.
\]

When we scale this by \(1/2\), the volume of \(\phi\) is:

\[
\text{vol}(\phi) = \frac{1}{8} \times \text{vol}(\text{T1} - \text{T1}) = \frac{1}{8} \times \
     frac{8}{3} = \frac{1}{3}.
\]

Thus, the volume of \(\phi\) is:

\[
\boxed{\dfrac{1}{3}}.
\]
```

Judge rationale and verdict.

```
**Analysis**

The problem asks for the volume of the set \(\phi\) of midpoints of segments joining points
     from two regular tetrahedra (edge length \(\sqrt{2}\)) that are central reflections of
     each other.
Both responses compute the volume of a single tetrahedron correctly as \(\frac13\).

**Response A**
After the volume calculation, it merely states:
> “Given the symmetry and the central reflection nature of the transformation, the volume of
     \(\phi\) is equal to the volume of the original tetrahedron, which is \(\frac13\).”

No justification or derivation is given; this is essentially an unsupported guess.  The
     assertion that \(\phi\) has the same volume as the original tetrahedron is not explained
     and is in fact false.

**Response B**
- It correctly places the origin at the centre of symmetry so that the second tetrahedron is
     \(-T_1\); then \(\phi = \frac{T_1 + (-T_1)}{2} = \frac{T_1 - T_1}{2}\) (up to translation
     , which does not affect volume).
```

```
- It recognises that scaling by \(\frac12\) multiplies volume by \(\frac1{2^3}= \frac18\).
- It attempts to find \(\operatorname{vol}(T_1 - T_1)\) by claiming: “For a convex body \(K\),
     the volume of \(K - K\) is known to be \(8 \times \operatorname{vol}(K)\) in three
     dimensions.”  This statement is **incorrect** (it holds only when \(K\) is centrally
     symmetric; a tetrahedron is not).  Using this erroneous factor gives \(\operatorname{vol
     }(T_1 - T_1) = 8\cdot\frac13 = \frac83\) and finally \(\operatorname{vol}(\phi)=\frac18\
     cdot\frac83=\frac13\).

Although the factor \(8\) is wrong, Response B exhibits a structured approach: it reduces the
     problem to the volume of a difference body and correctly handles scaling.  With the
     proper factor (for a simplex, \(\operatorname{vol}(K-K)=20\operatorname{vol}(K)\) by
     Rogers–Shephard) the same reasoning would yield the correct answer \(\frac56\).  Response
      A provides no such reasoning; it is merely a guess.

**Conclusion**
Both answers are numerically wrong, but Response B contains a logical derivation (with one
     mistaken fact) while Response A offers no derivation at all.  Hence **Response B is the
     better solution attempt**.

\[
\boxed{B}
\]
```

Mathematical check

After translating the center of symmetry to the origin, the midpoint body is $(K - K)/2$. Here $\mathrm{vol}(K) = 1/3$ and $\mathrm{vol}(K - K) = 20\,\mathrm{vol}(K)$, hence

$$\mathrm{vol}(\phi) = \frac{1}{8} \cdot 20 \cdot \frac{1}{3} = \frac{5}{6}.$$

The volume can also be checked directly: the midpoint body is congruent to a unit cube with eight corner tetrahedra, each of volume $1/48$, removed, giving $1 - 8/48 = 5/6$. Both responses have identical final answers and identical numerical error. The judge explicitly rejects B's factor 8 while recognizing its correct geometric reduction and scaling, distinguishing a localized mistaken fact from A's unsupported volume assertion.

E PROMPT TEMPLATE

Tournament prompt for pairwise preference (REASONING ARENA)

Compare the two solution attempts below and decide which is better.

[PROBLEM]
{problem}

[REFERENCE ANSWER]
The correct answer is: {final_answer}

[RESPONSE A]
{response_a}

[RESPONSE B]
{response_b}

Which response is better? Compare the final answers and reasoning quality. Your response MUST end with exactly one of: \boxed{A}, \boxed{B}, or \boxed{Tie}.

Keep your analysis under 500 words to ensure you have space for the final verdict.

Pointwise prompt (Adaptive Pointwise)

Evaluate the model response to the following math question. Score it as 1 (completely correct), 0.5 (minor errors), or 0 (incorrect/fatal errors). Your response MUST end with exactly one of: `\boxed{1}`, `\boxed{0.5}`, or `\boxed{0}`.

[PROBLEM]
{problem}

[REFERENCE ANSWER]
The correct answer is: {final_answer}

[RESPONSE]
{response}

Pointwise and DeepSeekMath-style rubric (RLAIF)

Please evaluate the quality of the model response to a math problem. The response may require a rigorous proof or a final numerical/symbolic answer. If the response provides an answer, it should also include a valid proof or derivation demonstrating why the answer is correct.

Scoring criteria:
- Score 1: The response is completely correct, with all steps executed properly and clearly demonstrated.
- Score 0.5: The response is generally correct, but with some details omitted or minor errors that do not invalidate the overall approach.
- Score 0: The response does not actually address the required problem, contains fatal errors, or has severe omissions that undermine its validity.

[PROBLEM]
{problem}

[REFERENCE ANSWER]
The correct answer is: {final_answer}

[RESPONSE]
{response}

Begin with "Here is my evaluation of the response:" and provide a detailed analysis of the key steps. Then conclude with "Based on my evaluation, the final overall score should be:" followed by exactly one of: `\boxed{1}`, `\boxed{0.5}`, or `\boxed{0}`.