

000 IMPROVING POLICY OPTIMIZATION VIA ENHANCED 001 EXPLORATION 002 003 004

005 **Anonymous authors**

006 Paper under double-blind review

007 008 ABSTRACT

011 Reinforcement learning has become the standard approach for aligning large lan-
012 guage models to complex reasoning tasks. However, these methods often overlook
013 rare valuable responses, as learning signals are dominated by high-probability,
014 frequently sampled outputs. To address this, we propose EXploration-Enhanced
015 Policy Optimization (EXPO), a novel approach that dynamically reweights the
016 advantage of each response based on its generation probability. EXPO amplifies
017 gradients from rare valuable samples, ensuring they contribute meaningfully to
018 policy updates and guide the model toward underexplored, high-value solutions.
019 We evaluate EXPO on multiple mathematical reasoning benchmarks. It consis-
020 tently outperforms strong baselines across model scales: on Qwen2.5-Math-1.5B,
021 EXPO surpasses DAPO by +3.0%; on Llama-3.2-3B-Instruct, by +3.6%; and on
022 the larger Qwen2.5-Math-7B, it outperforms the DAPO by +4.6%, Dr.GRPO by
023 +5.3% and instruction-tuned baseline by +9.1%. These gains demonstrate EXPO’s
024 effectiveness in leveraging valuable but underrepresented responses for better pol-
025 icy learning.

026 1 INTRODUCTION

027 The development of reasoning-centric LLMs,
028 including OpenAI-o3 (OpenAI, 2025),
029 DeepSeek-R1 (DeepSeek-AI et al., 2025), and
030 Kimi-K2 (Bai et al., 2025), has significantly
031 advanced the frontier of LLM capabilities,
032 particularly in tackling complex reasoning
033 tasks in mathematics and programming (Yue
034 et al., 2025). This progress is primarily driven
035 by Reinforcement Learning with Verifiable
036 Rewards (RLVR). Most implementations rely
037 on policy gradient algorithms, with Proximal
038 Policy Optimization (PPO) (Schulman et al.,
039 2017) and its recent variants (Fan et al., 2025;
040 Ren et al., 2025; Cozma et al., 2025) being
041 widely adopted for their stability and empirical
042 effectiveness. These methods iteratively refine
043 the policy by estimating the advantage of
044 sampled responses and reinforcing high-reward
045 behaviors.

046 Despite their promising results, these methods
047 remain fundamentally constrained by the base
048 LLM’s initial capabilities (Yan et al., 2025;
049 Zhao et al., 2025). Reinforcement learning amplifies existing behaviors by biasing the policy to-
050 ward high-reward paths, which improves sampling efficiency. However, this gain comes at a cost:
051 it narrows the model’s reasoning scope (Yue et al., 2025). We observe that this limitation stems from
052 the suppression of low-probability trajectories, even when they produce exceptionally high rewards.
053 As illustrated in Figure 2, this suppression causes a collapse in response diversity and weakens the
model’s ability to solve problems creatively. We analyze the root cause lies in the expectation-based

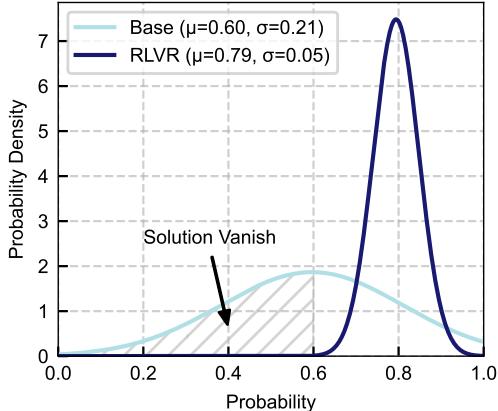


Figure 1: **Probability Density of Response Before and After RLVR Training.** RLVR sharpens the distribution and eliminates low-probability responses, reducing diversity and suppressing rare valuable reasoning paths.

054 nature of the policy gradient objective, which weights gradient updates by the likelihood of sampled
 055 trajectories. As a result, high-probability actions dominate learning, while rare valuable responses
 056 contribute negligibly to the gradient.

057 To overcome this limitation, we propose that the policy should be explicitly incentivized to learn
 058 from its valuable and often ignored discoveries. Amplifying the learning signal from these rare
 059 valuable samples is critical not only for escaping local optima but also for unlocking the model’s
 060 full creative and problem-solving capacity. Motivated by this insight, we introduce **Exploration-Enhanced Policy Optimization (EXPO)**, a novel algorithm that dynamically reshapes the training
 061 objective to prioritize high-reward, low-probability responses.

062 EXPO achieves this by introducing a dynamic weighting mechanism that modulates the advantage
 063 of each sampled response based on its likelihood under the current policy. Specifically, for desirable
 064 (high-reward) responses, the weight is inversely proportional to their generation probability, thereby
 065 amplifying gradients for rare valuable outputs. Conversely, for undesirable responses, the weight
 066 places stronger emphasis on penalizing frequent mistakes, encouraging the policy to avoid harmful
 067 or suboptimal behaviors. We summarize our contributions as follows:

- 070 • We identify and analyze a fundamental limitation of standard policy gradient methods such
 071 as GRPO and its variants in aligning large language models: their tendency to overlook
 072 low-probability, high-reward responses, which hinders effective exploration.
- 073 • We propose EXPO, a novel and lightweight algorithm that mitigates this bias by dynam-
 074 ically reweighting advantages based on response probability, thereby focusing learning on
 075 the most informative samples.
- 076 • We demonstrate through extensive experiments on multiple mathematical tasks that EXPO
 077 consistently outperforms strong baselines, yielding models that better explore the reward
 078 landscape and generate higher-quality, more diverse outputs.

079 2 PRELIMINARIES

080 We build on recent advances in policy gradient methods for LLM post-training to improve perfor-
 081 mance. Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) replaces PPO’s value
 082 network with group-based advantage normalization. For a prompt q , it samples G responses
 083 $\{o_1, \dots, o_G\}$, computes rewards $\{R_1, \dots, R_G\}$, and normalizes advantages within the group:

$$084 A_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G) + \epsilon}. \quad (1)$$

085 where ϵ ensures numerical stability. The GRPO objective includes KL regularization:

$$086 \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} (\min(r_{i,t} A_{i,t}, \text{clip}(r_{i,t}, 1 - \varepsilon, 1 + \varepsilon) A_{i,t}) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})) \right]. \quad (2)$$

087 with $r_{i,t} = \pi_\theta(o_{i,t} \mid q, o_{i < t}) / \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i < t})$ and β controlling KL penalty strength. DAPO
 088 (Yu et al., 2025) extends GRPO with four key improvements: (1) asymmetric clipping bounds
 089 ($\varepsilon_{\text{low}}, \varepsilon_{\text{high}}$) for more flexible updates, (2) dynamic sampling to adjust group composition, (3) token-
 090 level loss for finer control, and (4) reward shaping for long responses. Its objective removes the
 091 per-response normalization and KL term, instead applying a diversity constraint:

$$092 \mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E} \left[\frac{1}{\sum_i |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min(r_{i,t} A_{i,t}, \text{clip}(r_{i,t}, 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) A_{i,t}) \right], \quad (3)$$

093 s.t. $0 < |\{o_i \mid \text{is_equivalent}(o_i, a)\}| < G$.

094 which ensures sampled responses are not all identical to the reference answer a , preserving output
 095 diversity. These methods form the foundation for our approach, which further addresses their shared
 096 limitation: the suppression of rare valuable responses due to expectation-based gradient weighting.

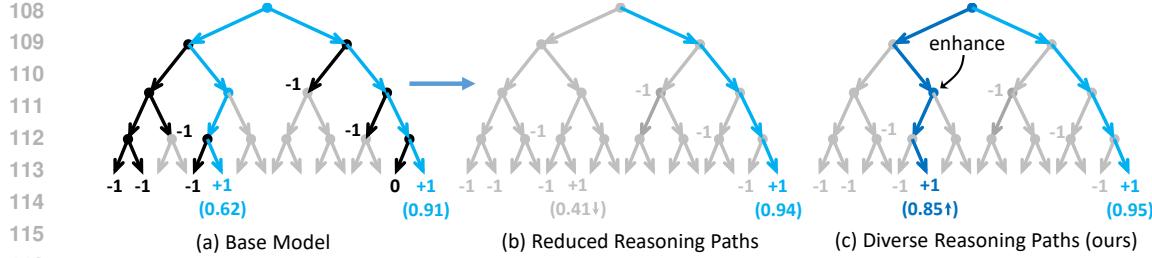


Figure 2: **Illustration of Reasoning Path Dynamics During Policy Optimization.** Solutions are scored +1 if correct and -1 otherwise; numbers in () indicate their respective probabilities. We observe: (a) Before optimization, the base model explores a diverse set of reasoning paths. (b) After standard reinforcement learning optimization, the reasoning space collapses: only high-reward, high-probability paths remain dominant, at the expense of diversity. (c) Our method preserves and enhances exploration by dynamically amplifying learning signals from rare but high-reward paths, achieving both improved performance and sustained reasoning diversity.

3 MOTIVATION

RLVR for large language models typically begins by sampling a large batch of responses, which is then partitioned into smaller mini-batches for iterative training. Responses that align with desired behavior receive higher rewards, encouraging the model to reinforce them; conversely, poorly aligned responses are penalized, prompting the model to suppress them. In essence, RLVR amplifies existing positive behaviors, but this comes at a cost.

As illustrated in Figure 1, we observe that after RLVR training, the model’s output distribution becomes more concentrated: response probabilities increase on average, and their variance decreases. Many low-probability responses that existed in the base model, including numerous correct or high-quality ones, vanish entirely. This phenomenon is often attributed to *entropy collapse*, where the policy distribution narrows over time, leading to a significant loss in solution diversity.

We offer a complementary, yet often overlooked, perspective. Since RLVR training relies on sampling responses from the current policy before updating, high-probability responses are sampled more frequently and thus dominate gradient updates. In contrast, low-probability responses, even when they yield high rewards, are rarely sampled and exert minimal influence on learning. This imbalance intensifies over time: as the model becomes more confident in its frequent outputs, those outputs increasingly steer future updates, creating a self-reinforcing loop that suppresses diversity and encourages homogeneity.

This intuition is formally validated by examining the policy gradient objective. The core issue lies in its expectation-based formulation:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} [G(\tau) \nabla_{\theta} \log \pi_{\theta}(\tau)] = \sum_{\tau} \pi_{\theta_{\text{old}}}(\tau) \cdot G(\tau) \nabla_{\theta} \log \pi_{\theta}(\tau). \quad (4)$$

where $G(\tau)$ denotes the advantage of trajectory τ . Because each term is weighted by the sampling probability $\pi_{\theta_{\text{old}}}(\tau)$, even exceptionally rewarding responses contribute negligibly to the gradient if they are unlikely under the current policy.

This leads to what we term **statistical short-sightedness**: rare but excellent responses, often happy accidents, are drowned out by the statistical mass of common, mediocre ones. The model thus becomes proficient at refining what it already knows, while systematically ignoring its flashes of brilliance. The result is premature convergence: models that generate safe, predictable outputs, but fail to discover superior, creative solutions lurking in the long tail of the response space. This motivates our proposal: to move beyond the standard expectation and introduce a mechanism that **dynamically amplifies the learning signal from high-reward, low-probability discoveries**, compelling the model to attend to and learn from its promising and low-probability outputs.

162 4 METHOD
163164 4.1 EXPLORATION-ENHANCED POLICY OPTIMIZATION
165

166 To address the suppression of rare valuable responses in policy gradient methods, we introduce
167 Exploration-Enhanced Policy Optimization, a novel algorithm that dynamically reweights the learning
168 signal based on a response’s probability. EXPO amplifies gradients from outputs that are high-
169 reward but low-probability under the current policy, ensuring they contribute meaningfully to policy
170 optimization. The core mechanism modifies the advantage A_i of each response y_i using a sequence-
171 level **dynamic weight** α_i :

$$\hat{A}_i = (1 + \alpha_i)A_i. \quad (5)$$

172 where A_i is the standard group-normalized advantage. The weight α_i is large for rare valuable
173 responses and near zero otherwise. We define α_i as:

$$\alpha_i = \text{clip}((1 - \text{clip}(\tilde{p}_i, \delta, 1))^\gamma, 0, \alpha_{\max}). \quad (6)$$

174 with the effective probability \tilde{p}_i set to: $p_i = \pi_{\theta_{\text{old}}}(y_i | x)$ if $A_i > 0$ (amplify rare valuable responses),
175 $1 - p_i$ if $A_i < 0$ (penalize frequent mistakes). Here, $\gamma \geq 0$ controls the focus on rare responses. The
176 value of the advantage A_i is very sensitive to model training, to ensure stable training, we introduce
177 two mechanisms:

- 180 • **Progressive Adjust** (δ): We set $\delta = t/T$, as training progresses ($\delta \rightarrow 1$), amplification of
181 rare responses fades smoothly, preventing late-stage instability.
- 182 • **Weight Clamping** (α_{\max}): We set an upper limit on $\alpha_i \leq \alpha_{\max}$ (e.g., 0.5), bounding the
183 scaling factor $(1 + \alpha_i) \in [1.0, 1.5]$ to avoid extreme updates.

184 Substituting the reweighted advantage \hat{A}_i into the DAPO objective yields the EXPO objective.

$$\begin{aligned} 185 \mathcal{J}_{\text{EXPO}}(\theta) &= \mathbb{E}_{(q, a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \\ 186 &\left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}} \right) \hat{A}_{i,t} \right) \right]. \quad (7) \end{aligned}$$

187 This formulation ensures that rare valuable responses have a much stronger effect on the policy up-
188 date, enabling EXPO to learn from exceptional but infrequent outputs that standard methods over-
189 look.

190 4.2 GRADIENT ANALYSIS

191 To better understand how EXPO reshapes the learning signal, we analyze its objective gradient. For
192 clarity, we omit the PPO-style ratio clipping (i.e., assume clipping bounds are not active), and recall
193 that both the advantage A_i and the dynamic weight α_i are computed using the frozen old policy $\pi_{\theta_{\text{old}}}$
194 and are thus treated as constants during gradient computation.

195 Under these conditions, the gradient of the EXPO objective (Eq. 7) with respect to θ is:

$$\nabla_{\theta} \mathcal{J}_{\text{EXPO}}(\theta) = \mathbb{E}_{(q, a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G (1 + \alpha_i) A_i \sum_{t=1}^{|o_i|} \nabla_{\theta} \log \pi_{\theta}(o_{i,t} | q, o_{i < t}) \right]. \quad (8)$$

196 where we use the identity $\nabla_{\theta} r_{i,t}(\theta) = r_{i,t}(\theta) \nabla_{\theta} \log \pi_{\theta}(o_{i,t} | q, o_{i < t})$. In contrast, the gradient of
197 the standard DAPO objective (Eq. 3) is:

$$\nabla_{\theta} \mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q, a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G A_i \sum_{t=1}^{|o_i|} \nabla_{\theta} \log \pi_{\theta}(o_{i,t} | q, o_{i < t}) \right]. \quad (9)$$

198 The only difference between the two gradients is the multiplicative factor $(1 + \alpha_i)$ in EXPO. For
199 high-reward responses ($A_i > 0$), α_i decreases with probability, it amplifies the influence of rare
200 valuable outputs while reducing the impact of common ones. For low-reward responses ($A_i < 0$), α_i
201 increases with probability, which discourages the model from assigning high likelihood to frequent
202 mistakes. **This ensures that exploration is guided by quality, not just novelty.**

216

5 EXPERIMENTS

217

5.1 SETUP

220 **Datasets.** For training, we use the MATH
 221 dataset (Hendrycks et al., 2021), focusing on
 222 problems from difficulty levels 3 to 5. As
 223 shown in Figure 3, we further group these prob-
 224 lems into five tiers based on the base model’s
 225 performance, with the *Hard* tier exhibiting the
 226 most effective training dynamics, our exper-
 227 iments use **1,000** samples for training from this
 228 tier, see Appendix A for details. For eval-
 229 uation, we benchmark our method across five
 230 mathematical reasoning datasets: (1) **AIME24**:
 231 A collection of 30 high-school olympiad-level
 232 problems from the 2024 American Invitational
 233 Mathematics Examination (Li et al., 2024). (2)
 234 **AMC**: A set of 83 intermediate-difficulty prob-
 235 lems from the American Mathematics Com-
 236 petitions, primarily in multiple-choice format (Li
 237 et al., 2024). (3) **MATH500**: A randomly sam-
 238 pled subset of 500 problems from the MATH
 239 dataset, spanning algebra, geometry, and num-
 240 ber theory (Hendrycks et al., 2021). (4) **Min-
 241 ervaMath**: A benchmark of 272 multi-step reason-
 242 ing problems (Lewkowycz et al., 2022). (5)
 243 **OlympiadBench**: A challenging suite of 675 high-difficulty mathematics problems (He et al.,
 244 2024).

245 **Models and Baselines.** Our experiments employ several models from the Llama and Qwen fam-
 246 ilies, including Llama-3.2-3B-Instruct (Dubey et al., 2024), Qwen2.5-Math-1.5B and
 247 Qwen2.5-Math-7B (Yang et al., 2024). We compare EXPO against several strong baselines: (1)
 248 **SimpleRL-Zero**: a replicate of the DeepSeek-R1 training on small models with limited data (Zeng
 249 et al., 2025b). (2) **OpenReasoner-Zero**: an open source implementation of large-scale reasoning-
 250 oriented RL training (Hu et al., 2025). (3) **PRIME-Zero**: process reinforcement through implicit
 251 rewards (Cui et al., 2025). (4) **Oat-Zero**: an unbiased optimization method that improves token
 252 efficiency while maintaining reasoning performance (Dr.GRPO; Liu et al., 2025b). (5) **DAPO**: a
 253 decoupled clip and dynamic sampling policy optimization algorithm (Yu et al., 2025).

254 **Evaluation Metrics.** Following established practice (Liu et al., 2025b; Zeng et al., 2025a), our
 255 primary metric is **Pass@1** (Chen et al., 2021). Pass@ k measures whether at least one of k inde-
 256 pendently generated solutions is correct. We focus on the more strict Pass@1 setting, which evaluates
 257 the accuracy of a single generated response and serves as a robust indicator of model reliability.

258 **Implementation Details.** We conduct reinforcement learning training using the `ver1` framework
 259 (Sheng et al., 2024). We set the clipping threshold to $\epsilon = 0.2$, and during training, we sample 16
 260 rollouts per prompt at a temperature of 1.0, with a maximum response length of 2048 tokens. The
 261 global batch size is 16, with a per-GPU mini-batch size of 4 and a learning rate of 1×10^{-6} . For
 262 inference, we use the vLLM library (Kwon et al., 2023), setting temperature to 0.0 and top-p to
 263 1.0. To ensure rigorous evaluation on mathematical problems, we incorporate verification functions
 264 from Math-Verify. All experiments are conducted on a cluster of 1 compute node, equipped with 4
 265 NVIDIA A40 40GB GPUs.

266

5.2 MAIN RESULTS

267 We present the main results of EXPO versus strong baselines on five challenging math reasoning
 268 benchmarks in Table 1. EXPO achieves the best overall performance, reaching 52.1% average ac-
 269 curacy using only 1,000 training samples, demonstrating its efficiency and strong ability to improve

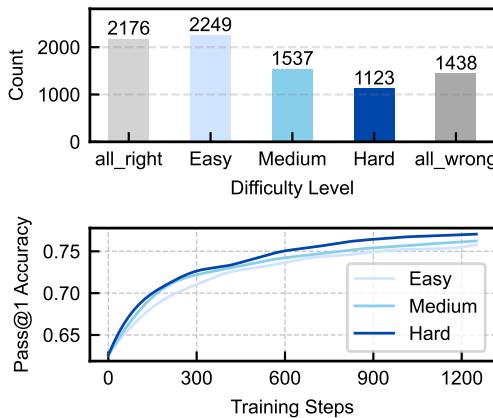


Figure 3: **Analysis of Problem Difficulty Distribution and Training Dynamics.** (a) Distribution of problems across difficulty levels. (b) Training curves of Pass@1 accuracy.

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

complex reasoning. This is a +17.1% gain over the base model and +9.1% over its instruction-tuned version, showing the clear benefit of our method. Compared to existing RLVR methods, EXPO consistently performs better. It outperforms the strongest prior baseline, Oat-Zero, by 2.0% on average. While some methods excel on specific tasks, for example, Oat-Zero on AIME24 and OpenReasoner-Zero on Math500 and OlympiadBench, EXPO delivers more balanced results: it ranks first on MinervaMath and AMC23, and second on AIME24 and Math500. The most convincing evidence of EXPO’s effectiveness comes from comparisons with our own reimplementations. EXPO beats standard GRPO by +5.4% on average and also surpasses recent variants like Dr.GRPO by +5.3% and DAPO by +4.6%. These consistent gains confirm that EXPO’s design enables more effective learning and delivers better final performance.

Table 1: **Performance Comparison of Various Baselines on Multiple Benchmarks.** Previous RLVR methods and our implementation are based on Qwen2.5-Math-7B. **Avg.** indicates mean accuracy across all test datasets. Top results are in **bold**, and runner-up results are underlined. Performance improvements (Δ) are relative to each baseline method.

| Algorithm | AIME24 | Math500 | OlympiadBench | MinervaMath | AMC23 | Avg. | Δ |
|------------------------------|-------------|-------------|---------------|-------------|-------------|-------------|---------------|
| Qwen2.5-Math-7B | 14.7 | 64.0 | 30.7 | 27.2 | 38.6 | 35.0 | + 17.1 |
| Qwen2.5-Math-7B-Instruct | 12.5 | 80.4 | 41.0 | 32.7 | 48.5 | 43.0 | + 9.1 |
| Previous RLVR methods | | | | | | | |
| SimpleRL-Zero | 27.0 | 76.0 | 34.7 | 25.0 | 54.9 | 43.5 | + 8.6 |
| OpenReasoner-Zero | 16.5 | 82.4 | 47.1 | 33.1 | 52.1 | 46.2 | + 5.9 |
| PRIME-Zero | 17.0 | 81.4 | 40.3 | 39.0 | 54.0 | 46.3 | + 5.8 |
| Oat-Zero | 33.4 | 78.0 | <u>43.4</u> | 34.6 | 61.2 | <u>50.1</u> | + 2.0 |
| Our Implementation | | | | | | | |
| GRPO | 14.8 | 80.0 | 42.1 | 41.2 | 55.4 | 46.7 | + 5.4 |
| Dr.GRPO | 16.6 | 81.2 | 43.4 | <u>44.5</u> | 48.2 | 46.8 | + 5.3 |
| DAPO | 13.3 | 79.6 | 39.3 | 43 | 62.5 | 47.5 | + 4.6 |
| EXPO (ours) | 30.0 | <u>81.8</u> | 41.3 | 44.9 | 62.5 | 52.1 | |

5.3 ANALYSIS

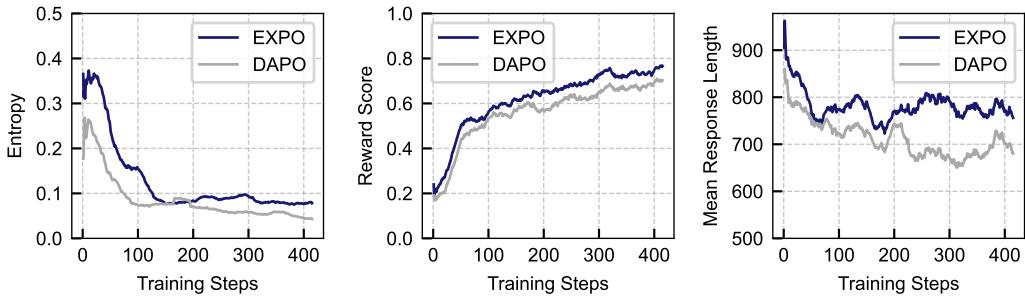
5.3.1 TRAINING DYNAMICS

We compare the training dynamics of EXPO and DAPO across three key metrics: policy entropy, reward score, and response length. As shown in the left panel of Figure 4, EXPO maintains higher entropy throughout training compared to DAPO, indicating a more exploratory behavior and sustained diversity in generated responses. This aligns with EXPO’s design goal of amplifying rare, high-reward trajectories, preventing premature convergence to narrow policy regions. The middle panel reveals that EXPO achieves a consistently higher reward score, demonstrating its effectiveness in optimizing for quality while preserving exploration. Notably, the reward improvement is not accompanied by a reduction in response length (right panel), where both methods exhibit similar trends, initially decreasing before stabilizing around 700-800 tokens. However, EXPO maintains slightly longer responses on average, suggesting it preserves expressiveness without sacrificing coherence or reward. Together, these dynamics confirm that EXPO strikes a better balance between exploration and exploitation, enabling robust learning from rare but valuable samples.

5.3.2 IMPACT OF LLM BACKBONE

To verify that the advantages of EXPO are not limited to a single model architecture, we extended our evaluation to different LLM backbones. As shown in Table 2, we conducted experiments on smaller Qwen2.5-Math-1.5B and Llama-3.2-3B-Instruct models. On the Qwen2.5-Math-1.5B, EXPO once again achieves the highest average accuracy of 43.9%, delivering a substantial improvement of +3.5% over the standard GRPO baseline and +1.8% over the strong competitor, Dr. GRPO. More importantly, we observe a similar trend on the Llama-3.2 model. Despite this model having a different architecture and a lower initial performance, EXPO still emerges as the most effective algorithm with an average accuracy of 24.8%. This represents a clear gain of +1.6% over GRPO and even larger gains over Dr. GRPO by +4.0% and DAPO +3.6 %. These consistent results across two

324 distinct model families robustly demonstrate that the performance improvements offered by EXPO
 325 are a general property of our algorithm and not dependent on a specific model architecture.
 326



338 **Figure 4: Training Dynamics of EXPO.** Left: Policy entropy, showing EXPO maintains higher ex-
 339 ploration throughout training. Middle: Reward score, indicating EXPO achieves consistently higher
 340 performance. Right: Mean response length, revealing both methods stabilize at similar lengths, with
 341 EXPO slightly preserving longer outputs.

342 **Table 2: Performance Comparison under Different LLM Backbones.** Qwen2.5-Math-1.5B and
 343 Llama-3.2-3B-Instruct are evaluated. The results highlight the impact of both model architecture and
 344 optimization strategy, with EXPO achieving the highest average scores across both LLM backbones,
 345 demonstrating its consistent effectiveness in enhancing reasoning performance.

| Algorithm | AIME24 | Math500 | OlympiadBench | MinervaMath | AMC23 | Avg. | Δ |
|-----------------------|--------|---------|---------------|-------------|-------|------|----------|
| Qwen2.5-Math-1.5B | 10.0 | 62.2 | 29.2 | 16.2 | 42.5 | 32.0 | +11.9 |
| + GRPO | 13.3 | 73.2 | 32.7 | 30.1 | 52.5 | 40.4 | +3.5 |
| + Dr. GRPO | 20.0 | 74.2 | 37.6 | 25.7 | 53.0 | 42.1 | +1.8 |
| + DAPO | 13.3 | 71.8 | 32.3 | 29.4 | 57.5 | 40.9 | +3.0 |
| + EXPO | 23.3 | 71.6 | 34.1 | 30.5 | 60.0 | 43.9 | |
| Llama-3.2-3B-Instruct | 6.7 | 41.0 | 12.1 | 17.3 | 15.0 | 18.4 | +6.4 |
| + GRPO | 6.7 | 44.8 | 17.2 | 22.1 | 25.0 | 23.2 | +1.6 |
| + Dr. GRPO | 6.7 | 50.0 | 14.7 | 14.3 | 18.1 | 20.8 | +4.0 |
| + DAPO | 6.7 | 45.2 | 14.2 | 19.9 | 20.0 | 21.2 | +3.6 |
| + EXPO | 6.7 | 50.6 | 16.3 | 22.8 | 27.5 | 24.8 | |

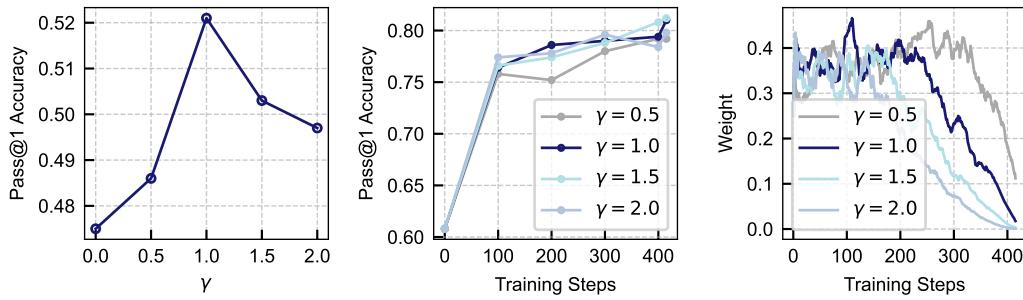
5.3.3 IMPACT OF γ COEFFICIENT

360 The hyperparameter γ controls the sensitivity
 361 of the dynamic reweighting factor α_i to re-
 362 response rarity. We conduct an ablation study to
 363 evaluate its impact on performance and train-
 364 ing dynamics. As shown in left panel of Fig-
 365 ure 5, Pass@1 accuracy peaks at $\gamma = 1.0$, with
 366 a drop for both lower and higher values. This
 367 indicates that moderate amplification is opti-
 368 mal: too weak ($\gamma < 1.0$) fails to boost rare
 369 high-reward responses, while excessive ampli-
 370 fication ($\gamma > 1.0$) may overemphasize noise or
 371 unstable signals. The middle panel shows that
 372 larger γ values accelerate early convergence, as
 373 seen in steeper initial curves for $\gamma = 1.5$ and
 374 2.0. However, these models tend to plateau
 375 earlier, suggesting that aggressive amplification
 376 may lead to premature stabilization. Finally, the right panel reveals how γ shapes the evolution
 377 of dynamic weights. Higher γ leads to stronger initial amplification and faster decay of weight
 378 magnitude, indicating more aggressive focus on rare events early in training, followed by rapid de-
 379 emphasis. Together, these results confirm that $\gamma = 1.0$ strikes the best balance between exploration

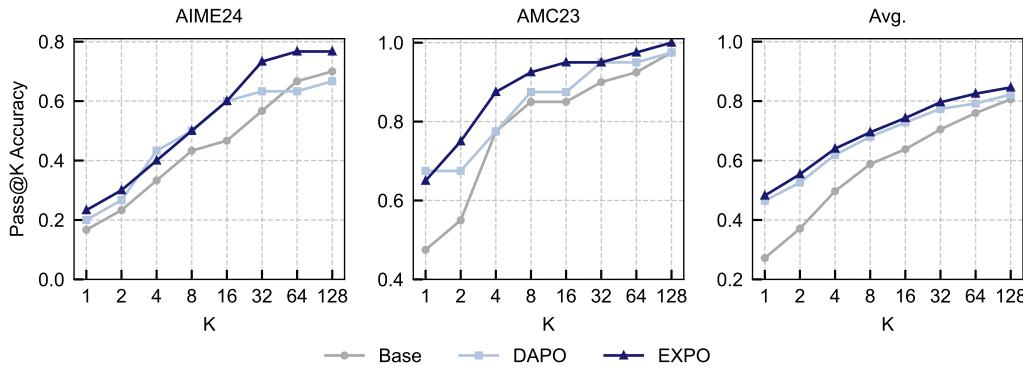
378 **Table 3: Average Number of Correct Solutions**
 379 **across Mathematical Benchmarks.** Higher val-
 380 ues indicate greater diversity in valid reasoning
 381 solutions. All results are based on the Qwen2.5-
 382 Math-7B.

| Algorithm | DAPO | EXPO |
|---------------|------|-------|
| AIME24 | 25.8 | 27.2 |
| Math500 | 98.5 | 100.8 |
| OlympiadBench | 78.7 | 79.9 |
| MinervaMath | 41.4 | 42.3 |
| AMC23 | 48.4 | 50.0 |

378 and stability, enabling EXPO to effectively learn from rare valuable responses without sacrificing
 379 training robustness.
 380



392 **Figure 5: Ablation Study of Reweighting Factor γ in EXPO.** (a) Pass@1 accuracy vs. γ : peak
 393 performance at $\gamma = 1.0$. (b) Training curves for different γ : higher values improve early con-
 394 vergence but may plateau earlier. (c) Evolution of dynamic weights: larger γ leads to stronger
 395 amplification and faster decay of rare-response weights.



411 **Figure 6: Pass@K Accuracy across Mathematical Reasoning Benchmarks.** EXPO consistently
 412 outperforms both the base model and DAPO at all values of K. More results on Math500, Olympiad-
 413 Bench and MinervaMath can be found at Appendix C

415 5.3.4 DIVERSITY OF REASONING SOLUTIONS

417 To quantify the diversity of correct reasoning solutions generated by the model, we sample 128 solu-
 418 tions per problem using a high temperature $T = 1.0$ to encourage stochasticity and avoid repetitive
 419 outputs. We then count the average number of correct solutions across five mathematical bench-
 420 marks. This metric reflects the breadth of the model’s reasoning, an indicator of creative and robust
 421 problem-solving. As shown in the table 3, EXPO consistently generates more diverse correct solu-
 422 tions than DAPO across all benchmarks, with the largest improvements on Math500 (+2.3) and
 423 AMC23 (+1.6). This is a direct consequence of EXPO’s core design. By dynamically amplifying
 424 gradients from rare valuable trajectories, EXPO actively resists premature convergence to dominant
 425 or repetitive solution patterns. As a result, the model preserves and refines a wider variety of valid
 426 reasoning paths throughout training, enabling it to internalize and deploy a richer, more diverse set
 427 of problem-solving strategies at inference time.

428 5.3.5 EFFECT ON REASONING CAPACITY BOUNDARY

430 The core function of EXPO is to encourage the generation of diverse reasoning solutions. A natural
 431 question is: does this diversity help improve reasoning capability boundary? To investigate this, we
 432 adopt the pass@k metric following [Yue et al. \(2025\)](#), which measures whether a model can generate

432 at least one correct solution in k independent attempts. This provides an upper-bound estimate
 433 of the model’s latent reasoning capacity. As shown in Figure 6, we observe a clear divergence in
 434 trends between small and large k . At small k , RLVR-trained models outperform the base model,
 435 consistent with the well-known finding that RLVR improves single-sample correctness. However,
 436 as k increases, the base model steadily closes the gap. For example, on AIME24, DAPO leads the
 437 base model by +3.3% at $k = 1$, but falls behind by -3.3% at $k = 128$, suggesting that DAPO’s
 438 gains come from concentrating probability mass on known correct paths, rather than discovering
 439 new ones. In contrast, EXPO maintains a consistent performance advantage over the base model
 440 even at $k = 128$, achieving a final average accuracy of 84.6% versus 80.6%, a +4.0% absolute
 441 gain. This demonstrates that EXPO not only improves single-attempt accuracy but also genuinely
 442 expands the model’s reasoning boundary by preserving and reinforcing diverse, high-reward solution
 443 trajectories.

444 6 RELATED WORK

445 **Mathematical Reasoning with LLMs** Mathematical reasoning is a gold standard for evaluating
 446 LLMs (Zhang et al., 2025), requiring symbolic abstraction, logical consistency, and multi-step
 447 deduction, cognitive traits central to science and engineering. Research mainly splits into two
 448 paradigms. *Formal reasoning*, based on systems like Lean or Coq (Zheng et al., 2022; Azerbayev
 449 et al., 2023; Xin et al., 2025), ensures correctness via machine-checkable proofs, ideal for theo-
 450 rem proving. *Informal reasoning*, using natural language or code without formal guarantees, better
 451 mirrors human problem solving: flexible, heuristic, and often tool-free (Sun & Zhang, 2025; Singh
 452 et al., 2025). It excels in tasks like word problems and symbolic computation, where plausible,
 453 high-quality outputs matter more than formal proof. We adopt informal reasoning, as it matches
 454 real-world settings where formal systems are unavailable. In such settings, models must learn from
 455 sparse rewards and discover rare, high-value reasoning paths. This is the core challenge EXPO
 456 addresses by amplifying signals from low-probability, high-reward responses.

457 **Policy Optimization for LLMs** Reinforcement learning has significantly improved LLM reason-
 458 ing, as shown in models like OpenAI-o3, DeepSeek-R1 and Kimi-K2 (OpenAI, 2025; DeepSeek-AI
 459 et al., 2025; Bai et al., 2025). Progress largely builds on verifiable rewards (Zeng et al., 2025b; Hu
 460 et al., 2025; Cui et al., 2025), which offer reliable training signals. Follow-up work uses test-time
 461 adaptation (Muennighoff et al., 2025; Zuo et al., 2025) and structured prompting (Wang et al., 2023;
 462 Sun & Zhang, 2025) to boost performance within existing limits, while newer RL methods (Shao
 463 et al., 2024a; Liu et al., 2025a; Yu et al., 2025) refine objectives for reasoning, yet remain mostly
 464 on-policy, amplifying known behaviors instead of discovering new ones. Recent work (Zhao et al.,
 465 2025; Yue et al., 2025) identifies a key issue: on-policy learning rarely explores beyond current
 466 behavior, favoring safe, frequent outputs and optimize within model boundaries rather than expanding
 467 reasoning horizons. EXPO tackles this by dynamically reweighting gradients to amplify signals
 468 from low-probability, high-reward responses. This enables learning from the model’s most val-
 469 uable, previously ignored outputs, preserving reasoning diversity and preventing model collapse into
 470 narrow solution modes, thereby improving overall performance.

471 7 CONCLUSION

472 We identify a critical limitation in standard policy gradient methods for LLM alignment: their ten-
 473 dency to suppress low-probability, high-reward responses, a bias that narrows reasoning scope, col-
 474 lapses diversity, and stifles creative problem-solving. To address this, we propose Exploration-
 475 Enhanced Policy Optimization, a lightweight algorithm that dynamically reweights policy gradients
 476 to amplify learning signals from rare but valuable outputs while penalizing frequent mistakes. By
 477 reshaping the optimization objective around response rarity and reward, EXPO enables models to
 478 escape local optima, sustain exploration, and internalize diverse reasoning strategies. Extensive ex-
 479 periments across multiple mathematical benchmarks confirm that EXPO consistently outperforms
 480 strong baselines like DAPO and GRPO, achieving higher accuracy, greater solution diversity, and
 481 more stable training, demonstrating that explicitly valuing happy accidents is not just beneficial, but
 482 essential for unlocking the full reasoning potential of large language models.

486 REFERENCES
487

488 Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev,
489 and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level math-
490 ematics. *CoRR*, abs/2302.12433, 2023. doi: 10.48550/ARXIV.2302.12433. URL <https://doi.org/10.48550/arXiv.2302.12433>.

491 Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen,
492 Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du,
493 Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng
494 Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao
495 Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing
496 Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng
497 Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei
498 Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu
499 Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu,
500 Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu,
501 Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan
502 Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song,
503 Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi
504 Wang, Dinglu Wang, Feng Wang, and Haiming Wang. Kimi K2: open agentic intelligence.
505 *CoRR*, abs/2507.20534, 2025. doi: 10.48550/ARXIV.2507.20534. URL <https://doi.org/10.48550/arXiv.2507.20534>.

506
507 Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared
508 Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large
509 language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

510
511 Andrei Cozma, Landon Harris, and Hairong Qi. KIPPO: koopman-inspired proximal policy op-
512 timization. *CoRR*, abs/2505.14566, 2025. doi: 10.48550/ARXIV.2505.14566. URL <https://doi.org/10.48550/arXiv.2505.14566>.

513
514 Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu
515 Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint*
516 *arXiv:2502.01456*, 2025.

517
518 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,
519 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,
520 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao
521 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,
522 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,
523 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,
524 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang
525 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai
526 Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,
527 Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang,
528 Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,
529 Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,
530 R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye,
531 Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing
532 reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.
533 48550/ARXIV.2501.12948. URL <https://doi.org/10.48550/arXiv.2501.12948>.

534
535 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
536 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony
537 Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark,
538 Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière,
539 Bethany Biron, Bin Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris
Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong,
Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny
Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino,

540 Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael
 541 Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson,
 542 Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar,
 543 Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra,
 544 Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar,
 545 Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng
 546 Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park,
 547 Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasudevan Alwala, Kartikeya
 548 Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of
 549 models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783.

550 Tiantian Fan, Lingjun Liu, Yu Yue, Jiaze Chen, Chengyi Wang, Qiying Yu, Chi Zhang, Zhiqi Lin,
 551 Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Bole Ma, Mofan Zhang, Gaohong Liu, Ru Zhang,
 552 Haotian Zhou, Cong Xie, Ruidong Zhu, Zhi Zhang, Xin Liu, Mingxuan Wang, Lin Yan, and
 553 Yonghui Wu. Truncated proximal policy optimization. *CoRR*, abs/2506.15050, 2025. doi: 10.
 554 48550/ARXIV.2506.15050. URL <https://doi.org/10.48550/arXiv.2506.15050>.

555 Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu,
 556 Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for
 557 promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint*
 558 *arXiv:2402.14008*, 2024.

559 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,
 560 and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In
 561 Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing
 562 Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, Decem-
 563 ber 2021, virtual*, 2021.

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

567 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.
 568 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
 569 serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating
 570 Systems Principles*, 2023.

571 Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ra-
 572 masesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative
 573 reasoning problems with language models. *Advances in Neural Information Processing Systems*,
 574 35:3843–3857, 2022.

575 Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif
 576 Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in
 577 ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*,
 578 13:9, 2024.

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

582 Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min
 583 Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025b.
 584 doi: 10.48550/ARXIV.2503.20783.

585 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke
 586 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time
 587 scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.

588 OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, April 2025. Available:
 589 <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.

594 Allen Z. Ren, Justin Lidard, Lars Lien Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Ma-
 595 jumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy opti-
 596 mization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025,*
 597 *Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL [https://openreview.net/](https://openreview.net/forum?id=mEpqHvbD2h)
 598 [forum?id=mEpqHvbD2h](https://openreview.net/forum?id=mEpqHvbD2h).

599 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 600 optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL [http://arxiv.org/abs/](http://arxiv.org/abs/1707.06347)
 601 [1707.06347](http://arxiv.org/abs/1707.06347).

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

605 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li,
 606 Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open
 607 language models. *CoRR*, abs/2402.03300, 2024b. doi: 10.48550/ARXIV.2402.03300. URL
 608 <https://doi.org/10.48550/arXiv.2402.03300>.

609 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,
 610 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint*
 611 *arXiv: 2409.19256*, 2024.

612 Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool
 613 integration for llms via reinforcement learning. *CoRR*, abs/2505.01441, 2025. doi: 10.48550/
 614 ARXIV.2505.01441. URL <https://doi.org/10.48550/arXiv.2505.01441>.

615 Lin Sun and Can Zhang. Exchange of perspective prompting enhances reasoning in large language
 616 models. *CoRR*, abs/2506.03573, 2025. doi: 10.48550/ARXIV.2506.03573. URL <https://doi.org/10.48550/arXiv.2506.03573>.

617 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha
 618 Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language
 619 models. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.

620 Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanja Zhao, Haocheng Wang, Bo Liu, Liyue
 621 Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Haowei Zhang, Qihao Zhu, Dejian Yang, Zhibin Gou,
 622 Z. F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: Harnessing proof assistant feedback
 623 for reinforcement learning and monte-carlo tree search. In *The Thirteenth International Confer-
 624 ence on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net,
 625 2025. URL <https://openreview.net/forum?id=I4YAIwrsXa>.

626 Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang.
 627 Learning to reason under off-policy guidance, 2025. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2504.14945)
 628 [2504.14945](https://arxiv.org/abs/2504.14945).

629 An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu,
 630 Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu,
 631 Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert
 632 model via self-improvement. *CoRR*, abs/2409.12122, 2024. doi: 10.48550/ARXIV.2409.12122.

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

636 Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao
 637 Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the
 638 base model? *CoRR*, abs/2504.13837, 2025. doi: 10.48550/ARXIV.2504.13837. URL <https://doi.org/10.48550/arXiv.2504.13837>.

639 Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-
 640 zoo: Investigating and taming zero reinforcement learning for open base models in the wild.
 641 *CoRR*, abs/2503.18892, 2025a. doi: 10.48550/ARXIV.2503.18892.

648 Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model
 649 and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient.
 650 <https://hkust-nlp.notion.site/simplerl-reason>, 2025b. Notion Blog.
 651

652 Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou,
 653 Zhongzhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu
 654 Wang, Songtao Huang, Yue Liao, Hongru Wang, Mengyue Yang, Heng Ji, Michael Littman, Jun
 655 Wang, Shuicheng Yan, Philip Torr, and Lei Bai. The landscape of agentic reinforcement learning
 656 for llms: A survey, 2025. URL <https://arxiv.org/abs/2509.02547>.
 657

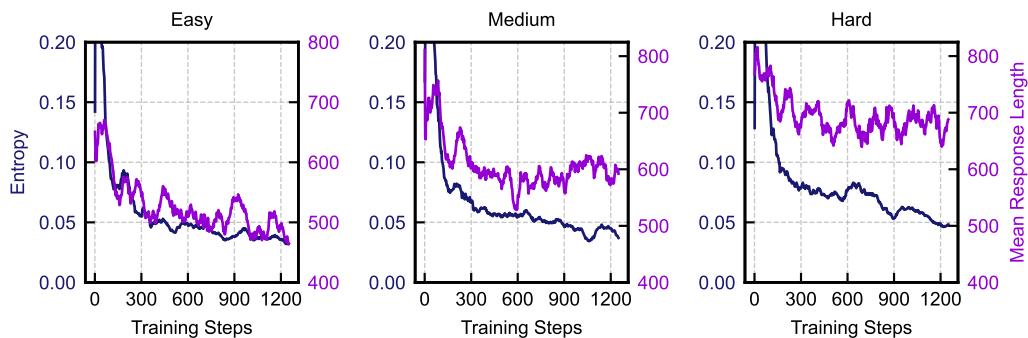
658 Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach.
 659 Echo chamber: RL post-training amplifies behaviors learned in pretraining, 2025. URL <https://arxiv.org/abs/2504.07912>.
 660

661 Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: a cross-system benchmark
 662 for formal olympiad-level mathematics. In *The Tenth International Conference on Learning
 663 Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL
 664 <https://openreview.net/forum?id=9ZPegFuFTFv>.
 665

666 Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen
 667 Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding,
 668 and Bowen Zhou. Ttrl: Test-time reinforcement learning, 2025. URL <https://arxiv.org/abs/2504.16084>.
 669

671 A TRAINING DATA

673 We build our training set from MATH dataset problems at difficulty levels 3–5. For each problem,
 674 we generate four responses using Qwen2.5-Math-7B and classify it into one of five difficulty tiers
 675 based on correctness: all-right: 4 correct, Easy: 3 correct, Medium: 2 correct, Hard: 1 correct, all-
 676 wrong: 0 correct. We exclude all-right and all-wrong problems, as they offer little learning signal,
 677 being either too easy or too hard. This leaves Easy, Medium, and Hard problems to study how data
 678 difficulty affects RL fine-tuning. As shown in Figure 7, training on Hard problems yields faster
 679 convergence and sustained gains. From the 1,123 problems in the Hard tier, we randomly sample
 680 1,000 for our final training set, balancing challenge and tractability to maximize learning signal.
 681



693
 694 **Figure 7: Training Dynamics by Difficulty Level.** Entropy (blue) and mean response length (purple)
 695 across training steps for Easy, Medium, and Hard problems. Hard problems show sustained
 696 variability.
 697

698 B CODE IMPLEMENTATION

700 It is easy to implement EXPO based on open-source RL framework. For example, we show the
 701 minimum viable implementation of EXPO that only modifies a few line of DAPo loss in verl.

```

702
703 1 def compute_policy_loss(advantages, sentence_logps, step_ratio, gamma
704 2     =1.0):
705 3     mask_adv = advantages.mean(dim=1) < 0
706 4     sentence_logps_detach = torch.clamp(torch.exp(sentence_logps.detach()
707 5         ), 0, 1)
708 6     sentence_logps_detach[mask_adv] = 1 - sentence_logps_detach[mask_adv]
709 7     sentence_logps_detach = torch.clamp(sentence_logps_detach, step_ratio
710 8         , 1)
711 9     alpha = 1 - sentence_logps_detach
712 10    alpha = alpha.pow(gamma)
713 11    alpha = torch.clamp(alpha, 0, 0.5)
714 12    return alpha
715
716

```

Listing 1: Function `compute_policy_loss` for adaptive response weighting.

C MORE EVALUATION RESULTS

We evaluate Qwen2.5-Math-7B under three settings: (1) Base, (2) DAPO, and (3) EXPO. Results are reported across five mathematical benchmarks using pass@K with $K \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, sampling temperature $T=1.0$, and correctness judged by any correct sample. As shown in below tables, EXPO consistently outperforms both Base and DAPO at all K values. With Qwen2.5-math-7B, The gap is largest at $K=1$, highlighting EXPO’s strength in single-sample accuracy, crucial for real-world use. At $K=128$, EXPO achieves 84.6% average, surpassing DAPO and Base. These results confirm that EXPO improves not only accuracy but also the diversity and reasoning capability boundary.

Table 4: Pass@K performance of Llama-3.2-3B-Instruct without finetuning across multiple mathematical reasoning benchmarks. K denotes the number of sampled responses per problem, with success measured if any response is correct. All evaluations use identical prompting, decoding, and temperature settings for fair comparison.

| K | AIME24 | Math500 | OlympiadBench | Minerva | AMC23 | Avg. |
|-----|--------|---------|---------------|---------|-------|------|
| 1 | 0.0 | 25.0 | 5.5 | 7.7 | 2.5 | 8.1 |
| 2 | 3.3 | 40.2 | 9.3 | 11.0 | 22.5 | 17.3 |
| 4 | 3.3 | 50.6 | 15.6 | 17.3 | 25.0 | 22.4 |
| 8 | 6.7 | 60.6 | 21.2 | 24.6 | 35.0 | 29.6 |
| 16 | 6.7 | 71.2 | 29.3 | 30.9 | 60.0 | 39.6 |
| 32 | 6.7 | 77.8 | 38.7 | 37.5 | 70.0 | 46.1 |
| 64 | 16.7 | 86.0 | 46.4 | 41.9 | 85.0 | 55.2 |
| 128 | 26.7 | 88.4 | 53.5 | 48.2 | 92.5 | 61.9 |

Table 5: Pass@K performance of Llama-3.2-3B-Instruct finetuned with DAPO across multiple mathematical reasoning benchmarks.

| K | AIME24 | Math500 | OlympiadBench | Minerva | AMC23 | Avg. |
|-----|--------|---------|---------------|---------|-------|------|
| 1 | 3.3 | 44.6 | 13.8 | 16.2 | 30.0 | 21.6 |
| 2 | 6.7 | 53.4 | 18.7 | 23.2 | 32.5 | 26.9 |
| 4 | 6.7 | 61.0 | 24.4 | 27.9 | 40.0 | 32.0 |
| 8 | 16.7 | 66.6 | 28.7 | 32.7 | 42.5 | 37.4 |
| 16 | 20.0 | 72.2 | 33.2 | 39.3 | 55.0 | 43.9 |
| 32 | 20.0 | 76.0 | 37.9 | 43.4 | 62.5 | 48.0 |
| 64 | 26.7 | 79.4 | 42.1 | 49.3 | 67.5 | 53.0 |
| 128 | 30.0 | 82.6 | 45.9 | 53.7 | 72.5 | 56.9 |

756
 757 Table 6: Pass@K performance of Llama-3.2-3B-Instruct finetuned with EXPO across multiple
 758 mathematical reasoning benchmarks.

| K | AIME24 | Math500 | OlympiadBench | Minerva | AMC23 | Avg. |
|-----|--------|---------|---------------|---------|-------|------|
| 1 | 6.7 | 44.4 | 14.5 | 16.5 | 32.5 | 22.9 |
| 2 | 10.0 | 55.4 | 19.7 | 23.5 | 35.0 | 28.7 |
| 4 | 10.0 | 63.6 | 24.7 | 28.7 | 35.0 | 32.4 |
| 8 | 13.3 | 70.0 | 29.9 | 32.4 | 50.0 | 39.1 |
| 16 | 13.3 | 74.4 | 36.0 | 38.2 | 57.5 | 43.9 |
| 32 | 16.7 | 81.2 | 40.3 | 44.5 | 65.0 | 49.5 |
| 64 | 26.7 | 85.4 | 47.0 | 49.6 | 72.5 | 56.2 |
| 128 | 33.3 | 88.4 | 51.0 | 57.0 | 85.0 | 62.9 |

769
 770 Table 7: Pass@K performance of Qwen2.5-math-7B without finetuning DAPO across multiple
 771 mathematical reasoning benchmarks.

| K | AIME24 | Math500 | OlympiadBench | Minerva | AMC23 | Avg. |
|-----|--------|---------|---------------|---------|-------|------|
| 1 | 16.7 | 42.4 | 18.2 | 11.0 | 47.5 | 27.2 |
| 2 | 23.3 | 58.4 | 30.2 | 18.4 | 55.0 | 37.1 |
| 4 | 33.3 | 73.2 | 39.3 | 25.0 | 77.5 | 49.7 |
| 8 | 43.3 | 84.8 | 47.9 | 33.1 | 85.0 | 58.8 |
| 16 | 46.7 | 90.0 | 55.6 | 41.5 | 85.0 | 63.8 |
| 32 | 56.7 | 93.6 | 63.6 | 48.5 | 90.0 | 70.5 |
| 64 | 66.7 | 95.0 | 69.3 | 56.6 | 92.5 | 76.0 |
| 128 | 70.0 | 96.6 | 75.6 | 63.2 | 97.5 | 80.6 |

783
 784 Table 8: Pass@K performance of Qwen2.5-math-7B finetuned with DAPO across multiple mathe-
 785 matical reasoning benchmarks.

| K | AIME24 | Math500 | OlympiadBench | Minerva | AMC23 | Avg. |
|-----|--------|---------|---------------|---------|-------|------|
| 1 | 20.0 | 76.2 | 37.8 | 30.9 | 67.5 | 46.5 |
| 2 | 26.7 | 83.4 | 45.2 | 40.4 | 67.5 | 52.6 |
| 4 | 43.3 | 88.4 | 51.9 | 48.2 | 77.5 | 61.9 |
| 8 | 50.0 | 90.6 | 58.8 | 52.9 | 87.5 | 68.0 |
| 16 | 60.0 | 93.0 | 64.7 | 58.8 | 87.5 | 72.8 |
| 32 | 63.3 | 94.8 | 69.2 | 64.7 | 95.0 | 77.4 |
| 64 | 63.3 | 96.2 | 72.6 | 68.8 | 95.0 | 79.2 |
| 128 | 66.7 | 98.0 | 76.3 | 72.4 | 97.5 | 82.2 |

796
 797 Table 9: Pass@K performance of Qwen2.5-math-7B finetuned with EXPO across multiple mathe-
 799 matical reasoning benchmarks.

| K | AIME24 | Math500 | OlympiadBench | Minerva | AMC23 | Avg. |
|-----|--------|---------|---------------|---------|-------|------|
| 1 | 23.3 | 79.2 | 39.6 | 33.8 | 65.0 | 48.2 |
| 2 | 30.0 | 85.6 | 47.4 | 39.0 | 75.0 | 55.4 |
| 4 | 40.0 | 89.6 | 54.7 | 48.2 | 87.5 | 64.0 |
| 8 | 50.0 | 92.4 | 58.8 | 54.0 | 92.5 | 69.5 |
| 16 | 60.0 | 93.8 | 64.3 | 58.5 | 95.0 | 74.3 |
| 32 | 73.3 | 95.0 | 69.5 | 65.1 | 95.0 | 79.6 |
| 64 | 76.7 | 96.6 | 72.9 | 68.8 | 97.5 | 82.5 |
| 128 | 76.7 | 97.6 | 76.4 | 72.1 | 100.0 | 84.6 |

810 D CASE STUDY
811812 D.1 COMPARISON BETWEEN EXPO AND BASELINES
813814 **Question:** Find the greatest common divisor of 3339, 2961, and 1491. (Correct answer is
815 21).
816817 **GRPO:**818 To find the greatest common divisor (GCD) of the numbers 3339, 2961, and 1491, we can
819 use the Euclidean algorithm. The Euclidean algorithm is an efficient method for computing
820 the greatest common divisor of two numbers. We can extend this method to more than two
821 numbers by iteratively applying it to pairs of numbers. The steps are as follows: 1. First, find
822 the GCD of the first two numbers, 3339 and 2961. 2. Then, find the GCD of the result from
823 step 1 with the third number, 1491. Let's start with the first step: finding the GCD of 3339
824 and 2961 using the Euclidean algorithm. The Euclidean algorithm states that $\text{GCD}(a, b) =$
825 $\text{GCD}(b, a \bmod b)$ and continues until the remainder is 0. The last non-zero remainder is the
826 GCD of the two numbers. We will apply this iteratively, the GCD of 3339 and 2961 is 126.
827 Now, we need to find the GCD of 126 and 1491. The GCD of 126 and 1491 is 9. Therefore,
828 the GCD of 3339, 2961, and 1491 is 9. The final answer is 9.
829830 **DR.GRPO:**831 To find the greatest common divisor (GCD) of the numbers 3339, 2961, and 1491, we can
832 use the Euclidean algorithm. The Euclidean algorithm finds the GCD of two numbers by
833 repeatedly applying the division algorithm: the GCD of a and b is the same as the GCD of
834 b and the remainder of a divided by b . First, we can find the GCD of 3339 and 2961 using
835 the Euclidean algorithm. Then, we can use that GCD to find the GCD with the third number,
836 1491. Here's the step-by-step approach: 1. Calculate the GCD of 3339 and 2961. 2. Use the
837 result to calculate the GCD with 1491. Let's start by calculating the GCD of 3339 and 2961.
838 We use the Euclidean algorithm, the GCD of 3339 and 2961 is 1. Since the GCD of 3339
839 and 2961 is 1, the GCD of 1 and any other number is 1. So, the GCD of 1 and 1491 is 1. The
840 final answer is 1.841 **DAPO:**842 To find the greatest common divisor (GCD) of 3339, 2961, and 1491, we will first find the
843 GCD of 3339 and 2961 using the Euclidean algorithm, and then we will find the GCD of
844 the result with 1491. The Euclidean algorithm for finding the GCD of two numbers a and b
845 (where $a > b$) is: 1. Divide a by b to get the remainder r . 2. Replace a with b and b with
846 r . 3. Repeat the process until the remainder is 0. The last non-zero remainder is the GCD of
847 a and b . Let's first find the GCD of 3339 and 2961. We will then find the GCD of the result
848 with 1491. The greatest common divisor (GCD) of 3339, 2961, and 1491 is 213. The final
849 answer is 3.850 **EXPO:**851 To find the greatest common divisor (GCD) of 3339, 2961, and 1491, we can use the Eu-
852 clidean algorithm. The Euclidean algorithm is an efficient method for computing the greatest
853 common divisor (GCD) of two numbers. We can extend this method to more than two num-
854 bers by iteratively applying the algorithm. The steps to find the GCD of multiple numbers are
855 as follows: 1. Find the GCD of the first two numbers. 2. Use the result to find the GCD with
856 the next number. 3. Repeat until you have the GCD of all the numbers. We start by finding
857 the GCD of 3339 and 2961, and then we find the GCD of that result with 1491. Let's do this
858 step-by-step. First, we find the GCD of 3339 and 2961, the GCD of 3339 and 2961 is the last
859 non-zero remainder, which is 3. Next, we find the GCD of 3 and 1491: $1491 = 3 * 497 + 0$
860 The GCD of 3 and 1491 is 3. The final answer is 21.861 D.2 DIFFERENT SOLUTIONS FOR THE SAME PROBLEM
862863 **Question:**

864 You have 5 shirts, 6 pairs of pants, and 8 hats. How many outfits can you make consisting of
 865 one shirt, one pair of pants, and one hat?

866 (Correct answer is 240).

867

EXPO (Solution 1):

To determine the total number of different outfits that can be made, we need to calculate the product of the number of choices for each item of clothing. Specifically, there are 5 choices for a shirt, 6 choices for a pair of pants, and 8 choices for a hat. The total number of different outfits can be calculated as follows: Total outfits = Number of shirts \times Number of pairs of pants \times Number of hats = $5 \times 6 \times 8$ We can calculate this product step by step. First, we calculate the product of the number of shirts and pairs of pants: $5 \times 6 = 30$ Then, we multiply the result by the number of hats: $30 \times 8 = 240$ So, the total number of different outfits that can be made is 240. The final answer is 240.

868

EXPO (Solution 2):

We have 5 shirts, 6 pairs of pants, and 8 hats. To find the total number of possible outfits consisting of one shirt, one pair of pants, and one hat, we can use the principle of counting combinations. The total number of possible outfits is the product of the number of choices for each item. So, we need to calculate: Total outfits = (Number of shirts) \times (Number of pairs of pants) \times (Number of hats) Given: Number of shirts = 5, Number of pairs of pants = 6, Number of hats = 8. So, the total number of outfits is: Total outfits = $5 \times 6 \times 8$. The total number of possible outfits consisting of one shirt, one pair of pants, and one hat is 240. The final answer is 240.

869

EXPO (Solution 3):

You have 5 shirts, 6 pairs of pants, and 8 hats. To determine how many different outfits can be made consisting of one shirt, one pair of pants, and one hat, we can use the basic principle of counting. According to the principle of counting, if there are m ways to do one thing and n ways to do another, then there are $m \times n$ ways to do both. Extending this to three choices, if there are a ways to do the first thing, b ways to do the second, and c ways to do the third, then there are $a \times b \times c$ ways to do all three. In this problem: There are 5 ways to choose a shirt. There are 6 ways to choose a pair of pants. There are 8 ways to choose a hat. Therefore, the total number of different outfits is $5 \times 6 \times 8$. The total number of different outfits that can be made consisting of one shirt, one pair of pants, and one hat is 240. The final answer is 240.

870

871

E LLM USAGE STATEMENT

872

873 LLMs were used in the preparation of this paper solely for grammatical correction and language
 874 polishing, and they help improve the clarity and fluency of the writing. All intellectual contributions,
 875 including concept development, algorithm design, experimentation, and interpretation of results, are
 876 entirely the work of the authors.

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900