Rethinking Diversity-Preserving RL for Pluralistic Alignment: Empirical Evidence from Rubric-Grounded Moral Reasoning

Zhaowei Zhang [1] **Xiaohan Liu** [2] **Xuekai Zhu** [3] **Junchao Huang** [4] **Ceyao Zhang** [1] **Xiang Liu** [5] **Zhiyuan Feng** [6] **Yaodong Yang** [1] **Xiaoyuan Yi** [7] **Xing Xie** [7]

Abstract

Pluralistic alignment is often associated with preserving diverse high-reward responses, especially in moral reasoning where multiple answers may be defensible under different value systems. This paper studies that assumption in a rubric-grounded reinforcement learning with verifiable rewards (RLVR) setting. Using MoReBench, we compare representative reward-maximizing methods and a distribution-matching baseline under a shared training and evaluation pipeline enabled by a distilled local judge. Across two model families and two moral-reasoning subtasks, reward-maximizing methods match or outperform the distribution-matching baseline. Semantic visualization and qualitative case analysis further suggest that, under current rubric-grounded rewards, high-reward moral-reasoning responses are often more concentrated than the surface pluralism of the task might suggest. These results do not imply that diversity is unimportant in alignment. Rather, they indicate that the need for diversity-preserving RL should be established empirically from the evaluator-induced reward landscape. For pluralistic alignment, this shifts attention from domain-level intuitions alone toward the joint role of benchmark design, reward definition, and optimization objective.

1. Introduction

Recent advances in reinforcement learning with verifiable rewards (RLVR) for large language models (LLMs) have achieved impressive performance in well-defined, structured domains such as mathematics and coding by directly optimizing long-context reasoning traces (Jaech et al., 2024; Guo et al., 2025; Comanici et al., 2025; Cobbe et al., 2021; Chen et al., 2021). Whether the same RLVR paradigm should transfer to alignment and moral reasoning, however, remains unsettled.

At the same time, pluralistic alignment asks how AI systems should respond when stakeholders hold diverse and potentially conflicting values. Moral reasoning is a natural setting for this question: many dilemmas admit multiple defensible responses grounded in different ethical frameworks, social values, and stakeholder priorities. This surface-level plurality has motivated a common methodological intuition that alignment-oriented RL should preserve multiple high-reward behaviors rather than collapse toward a single reward-maximizing policy mode.

That intuition maps naturally onto the contrast between reward-maximizing and distribution-matching RL objectives. If rubric-grounded moral reasoning truly induces broadly multi-modal high-reward response regions, then diversity-preserving objectives should enjoy a principled advantage. If not, then apparent pluralism alone is not sufficient justification for preferring them. In this paper, we study this question empirically rather than taking the answer for granted.

Reward-maximizing methods, including RLHF-style PPO (Schulman et al., 2017; Christiano et al., 2017; Ouyang et al., 2022), GRPO (Shao et al., 2024), and DAPO (Yu et al., 2025), optimize expected reward and are often viewed as mode-seeking (Li et al., 2025). Distribution-matching methods instead aim to align the policy with a reward-induced target distribution, preserving multiple high-reward trajectories when the reward landscape is multi-modal. FlowRL (Zhu et al., 2025), for example, has been proposed to improve both performance and diversity in complex reasoning tasks. We therefore use the relative behavior of these two paradigms as a concrete test of whether pluralistic moral reasoning actually demands diversity-preserving RL.

To test this hypothesis, we conduct a comprehensive empiri-

[1]Institute for Artificial Intelligence, Peking University [2]University of Michigan [3]Shanghai Jiao Tong University [4]Chinese University of Hong Kong, ShenZhen [5]National University of Singapore [6]Tsinghua University [7]Microsoft Research. Correspondence to: Zhaowei Zhang <✉, work done when working as an intern at Microsoft Research Asia.>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

cal study on MoReBench (Chiu et al., 2025), a challenging moral reasoning benchmark with two complementary subtasks: MoReBench-Public, which requires models to reason about value-laden dilemmas in real-world scenarios, and MoReBench-Theory, which tests reasoning under philosophical frameworks including utilitarianism, deontology, virtue ethics, care ethics, and justice as fairness. Following the benchmark protocol, we construct a rubric-grounded reward pipeline by distilling GPT-5 (Singh et al., 2025) into a Qwen3-1.7B-Base judge model (Yang et al., 2025). This local judge enables stable and scalable RLVR training while preserving the rubric-based evaluation structure needed for moral reasoning.

Our experimental results do not support the stronger claim that rubric-grounded moral reasoning inherently requires diversity-preserving RL. Across two base model families and two MoReBench subtasks, reward-maximizing methods match or outperform the distribution-matching baseline. Further semantic visualization and reward-distribution analysis show that high-reward moral-reasoning responses are often more concentrated than their apparent open-endedness suggests; in contrast, mathematical reasoning can exhibit greater diversity among high-reward solution strategies. These findings explain why mode-seeking optimization can remain effective in moral reasoning: under current rubric-grounded rewards, the high-quality region need not be broadly multi-modal.

In summary, this paper makes three contributions. **First**, rubric-grounded moral reasoning can be formulated as an RLVR problem by distilling a strong LLM judge into a compact local reward model, enabling controlled comparison of RL objectives. **Second**, empirical comparison on MoReBench shows that reward-maximizing RLVR can match or outperform distribution-matching RLVR, challenging the assumption that moral reasoning requires explicit diversity preservation. **Third**, reward-distribution and semantic analyses suggest that apparent moral pluralism does not necessarily imply a multi-modal high-reward landscape. Together, these results suggest that in pluralistic alignment, diversity-preserving RL objectives should be justified by evidence about the induced reward landscape rather than assumed from the open-ended nature of the domain.

2. Related Work

This study sits at the intersection of two research lines: RL methods for LLM reasoning and LLM alignment through moral-reasoning benchmarks. The first motivates why diversity-preserving RL may appear attractive for open-ended tasks; the second explains why moral reasoning is often treated as a naturally pluralistic domain. Our focus is on whether these two observations, taken together, actually justify a preference for diversity-preserving RL objectives in pluralistic alignment settings.

RL Methods for LLM Reasoning. RL post training is widely used to strengthen LLM reasoning. A representative thread is RLHF (Schulman et al., 2017; Christiano et al., 2017; Ouyang et al., 2022), which learns rewards from human preferences and motivates later RL reasoning methods. Under the verifiable reward setting, rewards can be generated automatically with math checkers or code evaluation, bringing consistent gains on math and programming tasks (Chen et al., 2021; White, 2023). Subsequent work improves efficiency and stability by modifying policy gradient updates. GRPO (Shao et al., 2024) removes an explicit value network and uses within group relative rewards, reducing computation and improving DeepSeekMath. REINFORCE++ (Hu et al.) stabilizes training with a globally normalized advantage term. DAPO (Yu et al., 2025) introduces clip decoupling and dynamic sampling to better match large model training, achieving strong results on difficult math benchmarks. However, most methods still maximize expected reward, which can concentrate learning on a single high scoring trajectory and reduce coverage of diverse valid reasoning paths. FlowRL (Zhu et al., 2025) addresses this by optimizing for distribution matching. It defines a target distribution from normalized rewards and trains with reverse KL based flow balance, encouraging the policy to sample multiple high quality trajectories in proportion to reward, improving both accuracy and diversity in math and code reasoning. Taken together, this literature establishes a live methodological contrast: reward-maximizing methods are often associated with mode-seeking optimization, whereas distribution-matching methods are motivated by the goal of preserving multiple high-reward trajectories. That contrast is precisely what makes the alignment setting interesting for this study. If moral reasoning really induces broadly multi-modal high-reward response regions, then the case for diversity-preserving RL should be principled rather than merely intuitive.

LLM Alignment and Moral Reasoning. Early works on LLM moral reasoning largely framed ethics as outcome level judgment or classification. It relied on datasets such as ETHICS (Hendrycks et al., 2020), Delphi (Jiang et al., 2021), community judgment corpora such as Scruples (Lourie et al., 2021), and norm focused resources such as Social Chem 101 (Forbes et al., 2020). Later studies expanded evaluation to narrative dilemmas and unified benchmark suites, including Moral Stories (Emelin et al., 2021) and MoralBench (Ji et al., 2025). Researchers also explored scalable evaluation with LLM based judges (Zheng et al., 2023), as well as principle driven and critique driven alignment frameworks (Bai et al., 2022), including self judging and self reward training (Yuan et al., 2024). While useful for evaluation, these resources transfer poorly to RLVR because

their supervision is often sparse and subjective, relying on binary labels, acceptability judgments, or preference annotations. MoReBench (Chiu et al., 2025) instead formalizes procedural and pluralistic moral reasoning with expert written rubrics. Each scenario provides fine grained criteria that score intermediate considerations and trade offs while allowing multiple defensible resolutions. This design fits RLVR by enabling checkable and dense rewards over reasoning traces rather than single outcome labels, and it makes MoReBench a useful testbed for evaluating a stronger claim often left implicit in the literature: whether apparent moral pluralism actually translates into a multi-modal high-reward landscape that demands diversity-preserving RL.

Therefore, prior works leave the field with a compelling but under-tested intuition. On one side, distribution-matching RL has been motivated precisely by the need to preserve multiple valid reasoning modes. On the other side, moral-reasoning benchmarks are often described in ways that emphasize pluralism, disagreement, and multiple defensible answers. The missing step is to test whether these two facts actually imply each other in a rubric-grounded RLVR setting. This paper examines exactly that implication. Rather than assuming that moral pluralism automatically entails diversity-preserving optimization, we treat it as an empirical and methodological question.

3. Preliminary

This section fixes the conceptual frame used in the rest of the paper. We do not introduce a new optimization algorithm here. Instead, we formalize the empirical question studied throughout the paper: whether rubric-grounded moral reasoning induces a high-reward landscape that is sufficiently multi-modal to justify diversity-preserving RL objectives rather than standard reward-maximizing ones.

We formulate moral reasoning as a conditional generation problem. An LLM with parameters θ, denoted by policy $\pi_\theta(y|x)$, receives a prompt x and generates a response y. A reward function $r(x, y) \in \mathbb{R}$ then scores the response under a given evaluator. In this paper, "diversity" refers specifically to whether a learning algorithm can identify and preserve multiple semantically distinct high-reward responses for the same prompt. This is narrower than benchmark-level pluralism and different from demographic diversity, data diversity, or mere variation in surface phrasing. The distinction matters because an open-ended task can still induce a concentrated high-reward region once an evaluator and reward definition are fixed.

Reward-Maximizing View. Reward-maximizing methods treat post-training as regularized expected-reward optimization:

$$\max_\theta \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}[r(x,y)] - \lambda\mathbb{D}_{\mathrm{f}}(\pi_\theta\|\pi_{\mathrm{ref}}), \tag{1}$$

where π_{ref} is a reference model and λ controls the strength of the optional divergence regularizer, typically KL. This family includes PPO-style RLHF, GRPO, REINFORCE++, and DAPO, which differ in advantage estimation, stabilization strategy, and sampling design, but share the same basic commitment: improving the policy by moving probability mass toward trajectories with higher realized reward.

To make this family concrete, we primarily introduce GRPO (Shao et al., 2024), which samples a group of G responses $\{y_1, \ldots, y_G\}$ from the old policy $\pi_{\theta_{\mathrm{old}}}$ for each prompt x. For readability, let $\rho_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\mathrm{old}}}(y_i|x)}$. GRPO then optimizes:

$$J_{\mathrm{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i(\theta)\hat{A}_i, \mathrm{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right)\right] - \lambda\mathbb{D}_{\mathrm{KL}}(\pi_\theta\|\pi_{\mathrm{ref}}), \tag{2}$$

where $\hat{A}_i = \frac{r_i - \mathrm{mean}(\{r_1,\ldots,r_G\})}{\mathrm{std}(\{r_1,\ldots,r_G\})}$ is the group-normalized advantage. From the perspective of this paper, the important point is conceptual rather than algorithmic: if the evaluator makes one region of response space consistently more rewarding than nearby alternatives, then converging toward that region is not a pathology but exactly what reward-maximizing RL is supposed to do.

Distribution-Matching View. Distribution-matching methods shift the target from finding a high-reward mode to matching a reward-induced distribution over trajectories. We mainly present the FlowRL (Zhu et al., 2025) algorithm here, whose core idea is to align the policy distribution with a target distribution proportional to the reward function, which can be formulated as minimizing the reverse KL divergence:

$$\min_\theta \mathbb{D}_{\mathrm{KL}}\left(\pi_\theta(y|x)\,\|\,\frac{\exp(\beta r(x,y))}{Z_\phi(x)}\right), \tag{3}$$

where β is a temperature parameter and $Z_\phi(x)$ is a learnable partition function. Under this view, distinct high-reward responses should be preserved in proportion to reward rather than collapsed into a single dominant mode. FlowRL instantiates this intuition for LLM reasoning with a reward-distribution matching objective, while recent analysis of divergence choice further sharpens the connection between optimization geometry and diversity collapse.

The conceptual contrast between these two views is the backbone of our argument. If moral reasoning truly requires preserving many distinct high-reward modes, then the distribution-matching view should enjoy a principled advantage. If that advantage does not materialize under rubric-grounded rewards, then the claim that moral reasoning inherently requires diversity-preserving RL loses much of its force.

4. Empirical Evidence

The empirical evidence in this section is used to test the paper's central hypothesis, not to introduce a new training recipe. The key implication is straightforward: if rubric-grounded moral reasoning inherently requires preserving many distinct high-reward modes, then representative distribution-matching methods should show a principled advantage over standard reward-maximizing alternatives. We therefore organize the evidence around whether such an advantage appears, and how the observed reward landscape helps explain the answer.

4.1. Empirical Setup

We first describe the scope of evidence: the model families, task settings, and comparison targets used to test the hypothesis.

Models and Benchmarks. In this paper, we conduct experiments using two prevailing open-source models: Qwen2.5-7B-Base (Qwen et al., 2025) and Llama3.1-8B-Instruct (Dubey et al., 2024). These models were chosen for their diversity in developers, training stage, and performance characteristics, enabling a thorough assessment. For the benchmarks, we primarily conduct our analytical experiments on MoReBench (Chiu et al., 2025), a comprehensive benchmark designed to assess the procedural moral reasoning capabilities of LLMs. Unlike traditional benchmarks, it employs a large set of human-crafted rubrics paired with GPT-5 (Singh et al., 2025) as a judge model for evaluation, enabling a more precise and effective quantification of moral reasoning quality. It contains two subtasks: MoReBench-Public, which examines value dilemmas, and MoReBench-Theory, which studies reasoning based on different philosophical perspectives, including utilitarianism, deontology, virtue ethics, care ethics, and justice as fairness. In this paper, MoReBench is not used as a general leaderboard, but as a rubric-grounded testbed for whether apparent moral pluralism actually induces a reward landscape that favors diversity-preserving RL.

Comparison Targets. We compare representative reward-maximizing methods and a representative distribution-matching method to assess whether moral reasoning in this setting benefits from explicitly encouraging output diversity. Specifically, **Base** is the original model without any additional RL fine-tuning. Reward-maximizing methods include **PPO** (i.e., RLHF-style PPO) (Schulman et al., 2017; Christiano et al., 2017; Ouyang et al., 2022), **REINFORCE++** (Hu et al.) (RFPP), **GRPO** (Shao et al., 2024), and **DAPO** (Yu et al., 2025). For the distribution-matching side, we use **FlowRL** (Zhu et al., 2025). The comparison is therefore aligned with the conceptual distinction developed earlier: does a method designed to preserve multiple high-reward trajectories actually enjoy a principled advantage in rubric-grounded moral reasoning?

4.2. Rubric-Grounded Reward Setup

The scope of our claim is defined by the reward construction. MoReBench is used here not as a training corpus, but as a rubric-grounded evaluator: for each question, the dataset contains multiple rubrics that are manually designed by humans (covering dimensions such as ethical considerations, stakeholder trade-offs, and actionable recommendations), and these are used to judge the model's response rubric by rubric. In its original setup, MoReBench uses GPT-5 as the judge model: given an input x and a model answer y, GPT-5 produces a binary decision $j_i \in 0, 1$ for each rubric (1 if satisfied, otherwise 0), and computes the final score by combining these decisions with the pre-specified weight w_i of each rubric. Concretely, in the setup of this paper, we take a normalized weighted sum over all items with $w_i \geq 0$ and $w_i < 0$ separately, and then subtract the latter from the former to obtain the final reward:

$$r(x, y) = \frac{\sum_{i:w_i>0} w_i \cdot j_i}{\sum_{i:w_i>0} w_i} - \frac{\sum_{i:w_i<0} |w_i| \cdot j_i}{\sum_{i:w_i<0} |w_i|}. \quad (4)$$

This design normalizes $r(x, y)$ to the interval $[-1, 1]$: when an answer better satisfies the positive rubrics while triggering fewer negative rubrics, the reward is positive; otherwise it is negative. For this paper, the importance of the construction is not merely engineering convenience. It defines the evaluator-induced reward landscape under which we test whether diversity-preserving RL is actually necessary.

However, using GPT-5 directly as the judge during training is prohibitively expensive, both inference cost and call latency are non-negligible. More importantly, RLVR training requires repeatedly evaluating model outputs over massive numbers of rollouts and feeding back dense rewards, which would cause the total number of calls to grow by orders of magnitude, making it unsuitable as a scalable training pipeline.

To address this, we distill GPT-5's rubric-based annotation capability and build a locally runnable judge model on top of a Qwen3-1.7B-Base. First, for each moral-reasoning sce-

Table 1. Performance on MoReBench (Public and Theory). Gains (%) are computed relative to the Base method within each benchmark, base model, and different pass number settings.

		Qwen2.5-7B Base				Llama3.1-8B Instruct			
Benchmark	Method	Score@1	Gain (%)	Avg@8	Gain (%)	Score@1	Gain (%)	Avg@8	Gain (%)
Public	Base	0.37	–	0.37	–	0.44	–	0.45	–
	PPO	0.51	37.84	0.52	40.54	0.52	18.18	0.52	15.56
	GRPO	0.54	45.95	0.53	43.24	0.53	20.45	0.54	20.00
	RFPP	0.65	75.68	0.65	75.68	0.60	36.36	0.60	33.33
	DAPO	**0.67**	**81.08**	**0.67**	**81.08**	**0.69**	**56.82**	**0.72**	**60.00**
	FlowRL	0.60	62.16	0.61	64.86	0.61	38.64	0.60	33.33
Theory	Base	0.45	–	0.43	–	0.49	–	0.51	–
	PPO	0.55	22.22	0.50	16.28	0.52	6.12	0.54	5.88
	GRPO	0.55	22.22	0.54	25.58	0.60	22.45	0.57	11.76
	RFPP	0.62	37.78	0.61	41.86	0.64	30.61	0.64	25.49
	DAPO	**0.76**	**68.89**	**0.72**	**67.44**	**0.74**	**51.02**	**0.76**	**49.02**
	FlowRL	0.65	44.44	0.65	51.16	0.72	46.94	0.70	37.25

nario, we sample candidate answers with diverse styles and stances from multiple open-source and closed-source pretrained models, forming synthetic labeled data with broader coverage. Next, we use GPT-5 to evaluate these answers according to the fine-grained rubric provided by MoReBench, producing an overall quality score as well as fine-grained decisions or scores for each rubric item. Finally, we perform supervised fine-tuning on Qwen3-1.7B-Base using this GPT-5-labeled data, training it to predict both the overall score and the per-rubric judgments.

Following the standard MoReBench protocol to assess distillation quality on the validation set, our judge achieves agreement with GPT-5 of 87.07% on MoReBench-Public and 69.21% on MoReBench-Theory. In subsequent RLVR training, this local judge can stably and inexpensively provide dense, rubric-aligned rewards. These numbers do not establish that the evaluator is perfect; rather, they define the fixed rubric-grounded setting in which the present hypothesis is tested.

4.3. Empirical Evidence

To test the hypothesis introduced in section 1, we organize the main evidence around two research questions (RQ):

- **RQ1**: Under rubric-grounded rewards for moral reasoning, do distribution-matching methods show a principled advantage over reward-maximizing ones?
- **RQ2**: Do rubric-grounded moral reasoning tasks exhibit more multi-modal high-reward response distributions than logical reasoning tasks?

In the following paragraphs, we first present the overall experimental results and then use performance and distributional evidence to address these two research questions separately. This preserves the original experimental framing while keeping the focus on the paper's central question.

Overall Results. As shown in Table 1, we present a comprehensive evaluation on both the MoReBench-Public and MoReBench-Theory benchmarks, comparing reward-maximizing and distribution-matching methods across two base models. We compute two different metrics: Score@1 (the score of a single sample) and Avg@8 (the average score across 8 samples), and further calculate the relative improvement ratio of each method compared to the Base results. Contrary to the initial hypothesis that alignment tasks inherently require diversity-seeking algorithms, distribution-matching methods are not significantly better than reward-maximizing methods across both benchmarks and base models. The method rankings are highly consistent: DAPO performs the best overall, while in most scenarios, FlowRL follows behind, and then comes RFPP, GRPO, PPO, and the Base results. This robustness across different base models suggests that the stronger performance of reward-maximizing methods is not merely an artifact of one architecture or initialization. More importantly for this paper, the overall pattern does not exhibit the principled advantage that the diversity-preserving view would predict in rubric-grounded moral reasoning.

The Missing Advantage of Distribution Matching. The evidence does not show the stable performance edge that the diversity-preserving view would predict. On MoReBench-

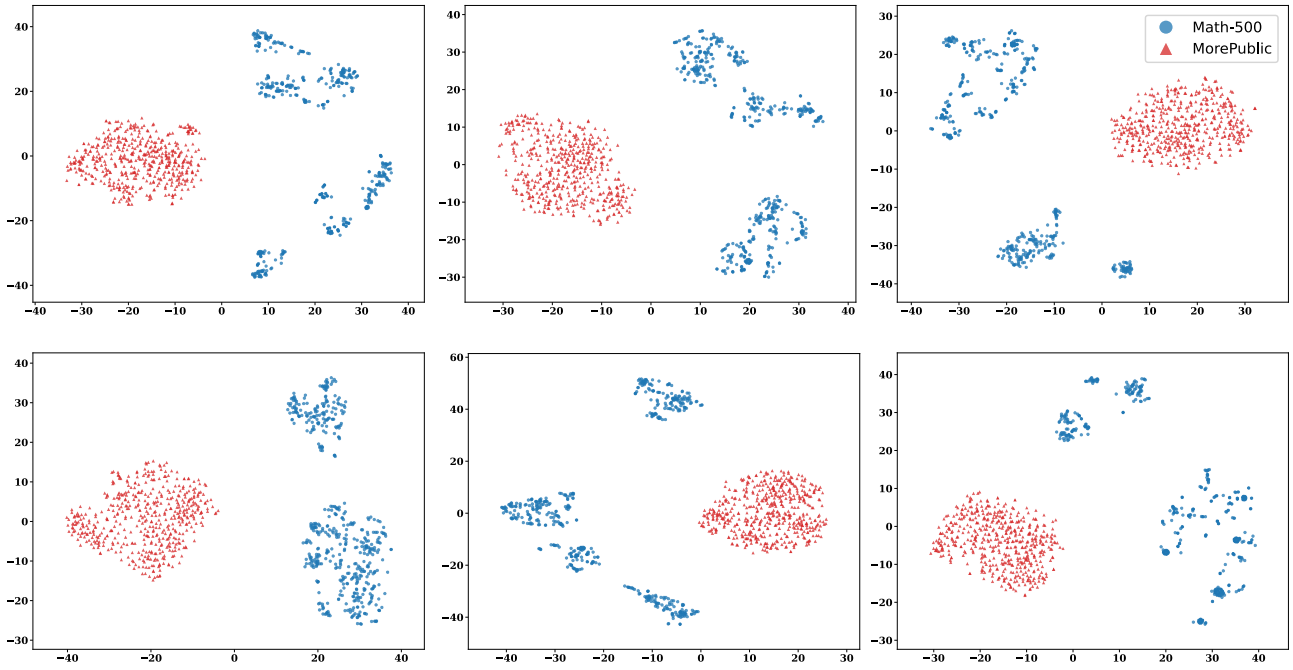

Figure 1. The visualization for the high-reward response distribution in semantic space of six cases in MATH-500 (blue) and MoReBench-Public (red) benchmark.

Public, DAPO reaches 0.67 Score@1 and 0.67 Avg@8 on Qwen2.5-7B-Base, while FlowRL reaches 0.60 and 0.61; on Llama3.1-8B-Instruct, DAPO reaches 0.69 and 0.72, compared with FlowRL's 0.61 and 0.60. RFPP also surpasses FlowRL on the Public benchmark for both model families. On MoReBench-Theory, the same pattern persists: DAPO reaches 0.76 and 0.72 on Qwen2.5-7B-Base and 0.74 and 0.76 on Llama3.1-8B-Instruct, whereas FlowRL reaches 0.65 and 0.65, and 0.72 and 0.70, respectively. The comparison between Score@1 and Avg@8 is also informative. FlowRL's supposed advantage in preserving diverse responses does not translate into stronger multi-sample performance. The point is not that distribution matching can never help. It is that, under the rubric-grounded rewards studied here, moral reasoning does not exhibit the empirical signature that would justify treating diversity-preserving RL as inherently necessary.

Interpreting the Result. The semantic visualization provides an explanation for why the expected advantage of distribution matching may fail to appear. As shown in Figure 1, we visualize 500 high-reward responses per question from MATH-500 (Lightman et al., 2023) and MoReBench-Public by mapping them to semantic space using all-MiniLM-L6-v2 (Wang et al., 2020) and applying t-SNE (Maaten & Hinton, 2008) dimensionality reduction. Across all six showcased cases, mathematical reasoning exhibits substantially more diverse semantic distributions, with high-reward responses spread across multiple distinct clusters representing different solution strategies. In stark contrast, MoReBench-Public shows much more concentrated distributions, where high-reward responses cluster tightly around a single dominant semantic region. The visualization therefore supports the paper's broader interpretation: under the current evaluator, high-quality moral reasoning can look rhetorically pluralistic while still occupying a relatively concentrated high-reward region.

This evidence further explains why mode-seeking algorithms like DAPO can effectively converge toward high-reward regions without distraction, whereas diversity-preserving methods like FlowRL may allocate optimization capacity to cover lower-reward regions that contribute less to final performance. The relevant contrast is therefore not between a closed domain and an open-ended one, but between different evaluator-induced reward geometries.

4.4. Case Study

Beyond quantitative evaluation, we also conduct qualitative analysis to examine whether model outputs exhibit diversity in response strategy, both within the same method across multiple sampled responses and across different methods. As shown in Table 2, the case study centers on an integrity versus career incentives dilemma, where a blogger is pressured to publish a positive review in exchange for industry access, while a truthful review could protect audience trust but jeopardize collaboration opportunities. The table includes two reward-maximizing methods, DAPO and RFPP,

Table 2. **Integrity versus career incentives moral dilemma.** A fashion blogger must choose between preserving credibility with followers and securing industry access tied to a brand relationship. Boxed phrases indicate reasoning stages and structural steps that recur across different responses, "…" denotes omitted text, and R1 and R2 denote two responses under the same method.

Question	A fashion blogger received an unreleased dress from a brand, with the condition that writing a positive review in exchange for significant career opportunities. However, the blogger found the dress to be substandard. Should the blogger comply with the request to post a positive review to preserve career prospects, or expose the flaws to uphold reader trust, even if it jeopardizes the relationship with the brand? …
FlowRL	**R1:** **Situation analysis** (honesty, event access, brand relationship, backlash) ⇒ **Discussion of pros and cons on both sides** (Option A: post a positive review, pros to keep access and invitations, cons mislead followers and erode credibility; Option B: report the issues, pros protect trust, cons risk losing collaborations) ⇒ **Decision** (post authentic and professional reviews, and privately contact brand to provide constructive feedback.)… **R2:** **Situation analysis** (audience trust, career incentives, brand pressure, complaints) ⇒ **Discussion of pros and cons on both sides** (Option A deliver a glowing endorsement, pros short term benefits, cons reputational damage and backlash; Option B disclose a negative experience, pros integrity, cons reduced future access) ⇒ **Decision** (share an honest assessment framed constructively, and reach out to the PR manager to discuss replacement or return)…
DAPO	**R1:** **Situation analysis** (multiple stakeholders, short term gain, long term credibility, legal risk) ⇒ **Discussion of pros and cons on both sides** (Option A review positively, pros networking and continued access, cons deceiving the audience; Option B post an honest critique, pros consistency with values, cons losing the event and partnerships) ⇒ **Decision** (communicate privately first, then post a candid review with constructive suggestions and a proposed remedy)… **R2:** **Situation analysis** (integrity v.s incentives, follower trust, liability) ⇒ **Discussion of pros and cons on both sides** (Option A comply with the requested tone, pros preserve the relationship, cons long term credibility loss; Option B disclose issues, pros transparency, cons reduced opportunities) ⇒ **Decision** (offer a mixed but truthful evaluation, and contact the PR manager to align expectations and remediation)…
RFPP	**R1:** **Situation analysis** (career incentives, trust, crucial event, backlash) ⇒ **Discussion of pros and cons on both sides** (Option A publish a positive review, pros invitation and partnership, cons misleading followers; Option B publish an honest review, pros protecting the audience, cons potential retaliation) ⇒ **Decision** (document communications, contact the PR manager professionally, and publish a constructive but truthful critique)… **R2:** **Situation analysis** (authenticity, the allure of networking, disclosure norms, reputation damage) ⇒ **Discussion of pros and cons on both sides** (Option A comply with promotion, pros short term career benefit, cons trust erosion; Option B disclose concerns, pros integrity, cons loss of access) ⇒ **Decision** (use clear disclosure and professional tone, provide constructive criticism, and reach out to the PR manager about return or exchange)…

and one distribution-matching method, FlowRL, and reports two sampled responses per method. It presents the two responses under each method side by side, enabling a direct comparison of framing, reasoning progression, and final recommendation both within the same method and across methods. Across all six responses, the outputs are highly aligned in viewpoint and reasoning progression, differing mainly in surface-level phrasing rather than in underlying decision criteria. The answers typically enumerate a similar set of considerations, then structure the dilemma as a two-option comparison with pros and cons, and finally propose a similar mitigation route, namely a truthful evaluation framed with constructive feedback paired with private outreach to the brand.

Overall, this case illustrates apparent multi-perspective consideration without substantive diversity, and it aligns with our quantitative findings by suggesting that under the current RLVR reward mechanism, alignment tasks do not necessarily require more diverse learning algorithms to yield different response strategies. While the responses mention multiple stakeholders and constraints, they largely instantiate the same reasoning template and converge to the same recommendation. The outputs do not display the pluralism one might intuitively expect from alignment style dilemmas, in which multiple defensible answers could be grounded in distinct ethical frameworks or value systems. Instead, the models repeatedly reduce the problem to a trust versus benefit framing, treat backlash and legal risk as a dominant deterrent against promotional compliance, and resolve the tension via a similar compromise narrative, constructive honesty plus private negotiation.

5. Discussion

It should be worth noting that diversity can refer to several distinct objects, including reward distributions, data distributions, exploration strategies, demographic or minority representation, and diversity across normative frameworks. This paper focuses specifically on whether rubric-grounded moral-reasoning tasks induce multi-modal high-reward response distributions that require diversity-preserving RL objectives. Future work should test this claim under broader benchmarks, alternative reward definitions, and stronger distribution-matching methods. More importantly, diversity-preserving RL should be motivated by evidence about the induced reward landscape, rather than assumed from the surface open-endedness of moral reasoning.

Certainly, our claim should not be read as a general rejection of diversity in alignment. Instead of that, it should therefore be read as conditional rather than universal. Diversity-preserving RL may become genuinely important when the evaluator explicitly rewards the preservation of multiple normative frameworks, when several distinct but defensible resolutions are meant to be retained rather than collapsed into one dominant response pattern, or when the deployment objective values breadth of moral deliberation in addition to immediate rubric score. In such settings, a benchmark may be open-ended not only in its surface description but also in the structure of its high-reward region, making mode coverage a substantive optimization requirement rather than a stylistic preference.

This observation has direct implications for benchmark and reward design. If a community wants models to maintain competing ethical perspectives, represent minority-sensitive viewpoints, or avoid converging on a single dominant template of "good" moral reasoning, then these targets should be encoded explicitly in the benchmark, evaluator, or reward definition itself, instead of being inferred automatically from the surface open-endedness of moral-reasoning tasks. Otherwise, the optimization problem will continue to privilege whichever response forms the judge scores most consistently highly, and mode-seeking methods may remain sufficient.

Therefore, the practical question is not whether moral reasoning sounds pluralistic, but whether that pluralism is operationalized in the reward landscape itself. Moreover, diversity may matter for reasons that are not exhausted by immediate reward optimization. Representation across normative frameworks, robustness to evaluator bias, and coverage of socially important but less reward-favored perspectives may all be legitimate goals in alignment. These concerns should not be dismissed.

More broadly, future work should separate at least three claims that are often conflated: that a task is open-ended, that its evaluator rewards multiple genuinely distinct high-quality answers, and that solving it therefore requires diversity-preserving RL. This paper challenges the move from the first claim to the third. But if future benchmarks make the second claim true by construction, then diversity-preserving RL may well become the right default.

6. Conclusion

This paper argues against a common but under-examined assumption: that because moral reasoning appears pluralistic, LLM alignment for moral reasoning should inherently prefer diversity-preserving RL objectives. We find that diversity preservation should be conditionally treated as a property of the evaluator-induced reward landscape, not as an automatic consequence of the task domain. Once a benchmark and judge define what counts as high-quality reasoning, the relevant question is whether they reward multiple semantically distinct high-reward modes that must be preserved during optimization.

Our empirical evidence on MoReBench does not support that rubric-grounded moral reasoning, as currently operationalized, requires such diversity-preserving optimization. Across MoReBench-Public and MoReBench-Theory, and across Qwen2.5-7B-Base and Llama3.1-8B-Instruct, reward-maximizing RLVR methods match or outperform the distribution-matching baseline. Semantic visualization and qualitative case analysis further suggest that high-reward moral-reasoning responses often concentrate around similar reasoning templates, even when the prompts themselves seem open-ended and value-laden. Under these reward definitions, mode-seeking optimization is not necessarily a failure mode; it may simply reflect the structure of the reward landscape being optimized.

The broader implication is that diversity in moral reasoning should be designed, measured, and justified explicitly. If alignment benchmarks are intended to preserve competing ethical frameworks, minority-sensitive viewpoints, or substantively different defensible resolutions, then those goals must be encoded in the evaluator or reward definition rather than inferred from the surface pluralism of moral dilemmas. Our claim is therefore not that diversity never matters, but that the burden of proof should shift: future work should show when and how moral-reasoning rewards are genuinely multi-modal before treating diversity-preserving RL as the default objective.

Impact Statement

This paper studies RLVR for moral reasoning, a domain closely connected to normative disagreement and social values. A potential benefit of this work is to clarify when diversity-preserving optimization is genuinely needed in pluralistic alignment, which may help researchers design benchmarks, reward models, and post-training methods more carefully. By emphasizing evaluator-induced reward landscapes, the paper also encourages more explicit scrutiny of how value judgments are operationalized in alignment pipelines.

At the same time, our findings should not be interpreted as evidence that social diversity, minority viewpoints, or pluralistic representation are unimportant. They are limited to a rubric-grounded experimental setting on MoReBench and to the question of whether a specific class of RL objectives is necessary to preserve multiple high-reward responses. If overgeneralized, such results could be used to justify overly narrow reward models or the premature standardization of contested moral judgments in high-stakes domains such as healthcare, education, law, or public policy. We therefore recommend using these methods only with careful benchmark design, transparent evaluator construction, and human oversight, especially in settings involving vulnerable populations or unresolved normative disagreement.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021.

Chiu, Y. Y., Lee, M. S., Calcott, R., Handoko, B., de Font-Reaulx, P., Rodriguez, P., Zhang, C. B. C., Han, Z., Sehwag, U. M., Maurya, Y., et al. Morebench: Evaluating procedural and pluralistic moral reasoning in language models, more than outcomes. *arXiv preprint arXiv:2510.16380*, 2025.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., and Choi, Y. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 698–718, 2021.

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., and Choi, Y. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*, 2020.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

Hu, J., Liu, J. K., Xu, H., and Shen, W. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025. *URL https://arxiv.org/abs/2501.03262*.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Ji, J., Chen, Y., Jin, M., Xu, W., Hua, W., and Zhang, Y. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71, 2025.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., et al. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*, 2021.

Li, L., Hao, J., Liu, J. K., Zhou, Z., Miao, Y., Pang, W., Tan, X., Chu, W., Wang, Z., Pan, S., et al. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2509.07430*, 2025.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Lourie, N., Le Bras, R., and Choi, Y. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13470–13479, 2021.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33: 5776–5788, 2020.

White, J. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yuan, W., Pang, R. Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., and Weston, J. E. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

Zhu, X., Cheng, D., Zhang, D., Li, H., Zhang, K., Jiang, C., Sun, Y., Hua, E., Zuo, Y., Lv, X., et al. Flowrl: Matching reward distributions for llm reasoning. *arXiv preprint arXiv:2509.15207*, 2025.