RLMR: Reinforcement Learning with Mixed Rewards for Creative Writing

JianXing Liao[1*], Tian Zhang[1], Xiao Feng[1†], Yusong Zhang[1], Haorui Wang[1*], Bosi Wen[2*], Ziying Wang[3*], Runzhi Shi[3*]

[1]Tencent Hunyuan Team
[2]Tsinghua University
[3]Peking University
jasonliao@std.uestc.edu.cn, alicexfeng@tencent.com[†]

Abstract

Large language models are extensively utilized in creative writing applications. Creative writing requires a balance between subjective writing quality (e.g., literariness and emotional expression) and objective constraint following (e.g., format requirements and word limits). Existing reinforcement learning methods struggle to balance these two aspects: single reward strategies fail to improve both abilities simultaneously, while fixed-weight mixed-reward methods lack the ability to adapt to different writing scenarios. To address this problem, we propose Reinforcement Learning with Mixed Rewards (RLMR), utilizing a dynamically mixed reward system from a writing reward model evaluating subjective writing quality and a constraint verification model assessing objective constraint following. The constraint following reward weight is adjusted dynamically according to the writing quality within sampled groups, ensuring that samples violating constraints get negative advantage in GRPO and thus penalized during training, which is the key innovation of this proposed method. We conduct automated and manual evaluations across diverse model families from 8B to 72B parameters. Additionally, we construct a real-world writing benchmark named WriteEval for comprehensive evaluation. Results illustrate that our method achieves consistent improvements in both instruction following (IFEval from 83.36% to 86.65%) and writing quality (72.75% win rate in manual expert pairwise evaluations on WriteEval). To the best of our knowledge, RLMR is the first work to combine subjective preferences with objective verification in online RL training, providing an effective solution for multi-dimensional creative writing optimization.

Introduction

Large language models (LLMs) are widely applied to creative writing tasks, from traditional poetry composition to modern fiction generation, and from literary scriptwriting to commercial copywriting, fulfilling diverse writing demands across domains and genres. To further enhance LLM performance in creative writing tasks, reinforcement learning techniques have been widely applied during the post-training phase. Through methods such as Group Relative Policy Optimization (GRPO), researchers aim to guide models toward generating higher-quality creative content through reward signals.

However, existing reinforcement learning reward strategies suffer from fundamental limitations. The evaluation criteria for creative writing are inherently dual in nature: on one hand, they require assessing subjective writing qualities such as literariness, emotional expression, and originality; on the other hand, they necessitate verifying objective constraint following, including length constraints, format requirements, and specific writing styles. Different creative writing scenarios exhibit significant variations in their emphasis on subjective versus objective evaluation.

Current reward strategies face two major challenges. First, single reward strategies struggle to simultaneously optimize both subjective and objective dimensions. As illustrated in Figure 1, under single-signal strategies, reward models only score writing quality without reflecting constraint following. Second, existing multi-reward signal fusion strategies typically employ fixed-weight summation. Such fixed-weight mechanisms fail to dynamically adjust weights based on actual sample performance within groups, making them unsuitable for different writing scenarios.

To address these issues, we propose Reinforcement Learning with Mixed Rewards (RLMR), a dynamic mixed-reward framework for creative writing. By coupling a writing reward model for evaluating subjective writing quality with a constraint verification model for assessing objective constraint following, we implement an adaptive mechanism that dynamically allocates reward weights based on constraint following within sampled group responses. Unlike existing methods that use fixed-weight fusion, our core innovation lies in dynamically adjusting the constraint following reward weight according to writing quality within sampled groups. This ensures that samples violating constraints receive negative advantage values in GRPO calculations, thereby being systematically penalized during policy gradient updates.

To validate our method's effectiveness, we conducted training on various scales of Qwen and DeepSeek model families and performed both automated and manual evaluations on multiple creative writing and instruction-following benchmarks. RLMR shows substantial gains in both writing quality and constraint following compared to single-reward and linear weighting baseline methods. Manual evaluation

[*]Work done when these authors interned at Tencent.
[†]Corresponding author.

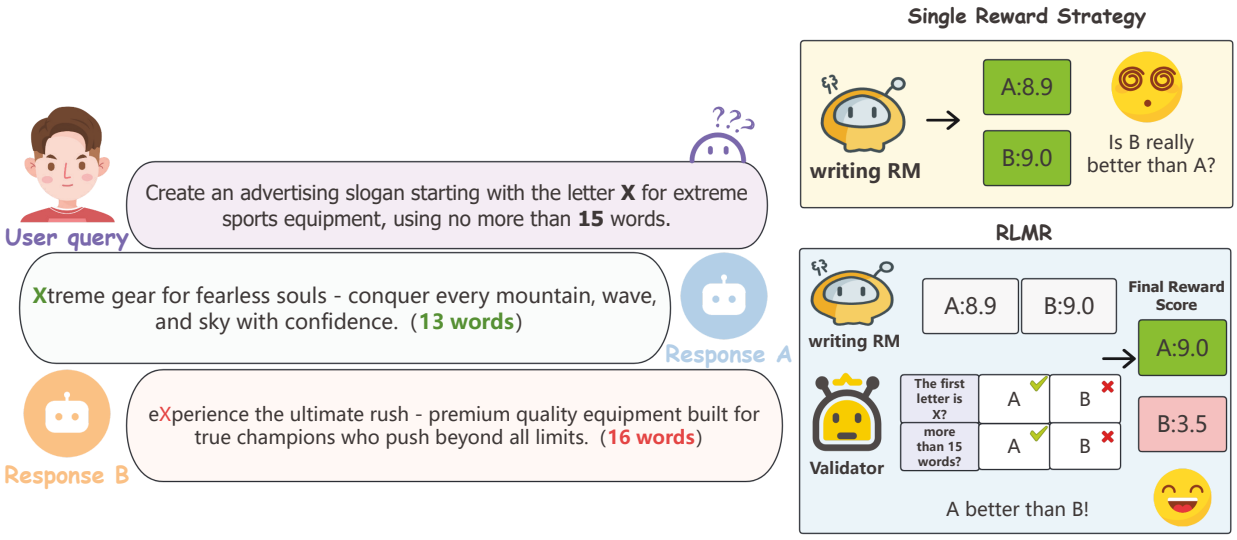


Figure 1: Comparison of single reward strategy versus our mixed RLMR approach. Given a task requiring an advertising slogan starting with "X" using no more than 15 words, Response A follows constraints but scores lower (8.9), while Response B violates constraints but scores higher (9.0). Single reward strategies incorrectly prefer Response B, while our RLMR combines writing quality and instruction following signals to correctly identify Response A as superior through dynamic penalty adjustments.

confirms significant preference for our approach over traditional strategies. These results effectively validate that our method resolves the trade-off between subjective and objective evaluation criteria in creative writing optimization.

Our key contributions include:

1. Identifying the inherent limitations of single reward signals and fixed-weight mixing strategies in creative writing tasks.
2. Proposing RLMR and developing a dynamic reward adjustment mechanism that ensures constraint-violating samples receive negative advantages during training, enabling better balance between writing quality and constraint following among multiple reward signals.
3. Demonstrating consistent improvements across diverse model families and scales through comprehensive automated and manual evaluations, proving the effectiveness of our method.

Related Work

To further improve LLM performance and align it with human preferences, reinforcement learning, especially RLHF, has become a mainstream optimization approach. Algorithms such as Proximal Policy Optimization (PPO) (Ouyang et al. 2022) and Group Relative Policy Optimization (GRPO) (Shao et al. 2024) are widely used to align LLM behavior with human preferences. PPO ensures training stability by limiting the extent of policy updates through clipped probability ratios, but requires separate value function training which increases computational overhead. GRPO optimizes policy gradients by estimating baselines from sampled groups, avoiding the need for separate value function training while maintaining competitive performance. Given GRPO's computational efficiency and effectiveness in creative writing scenarios, we choose it as our reinforcement learning framework.

Mixed reward strategies have become increasingly important in reinforcement learning, integrating multi-dimensional reward signals to guide model training more comprehensively. Peng et al. (Peng et al. 2025a) proposed the Agentic Reward Modeling framework, which combines human preference rewards with verifiable correctness signals (factuality and instruction following) to provide more reliable rewards for large language models. Jia et al. (Jia et al. 2025) introduced Writing-Zero, proposing a writing-principle-based pairwise Generative Reward Model (GRM) that leverages self-principled critique to transform subjective assessments into reliable, verifiable rewards for creative writing tasks. Wu et al. (Wu et al. 2025) developed LongWriter-Zero framework for ultra-long text generation, employing specialized reward models targeting length control, writing quality, and structural formatting with a composite reward function that averages individual advantages to balance multiple reward dimensions.

However, these existing mixed reward approaches all rely on fixed-weight fusion mechanisms, which suffer from fundamental limitations. First, fixed weights cannot adapt to varying constraint compliance patterns within different sample groups. When most responses in a group violate constraints, fixed-weight strategies still assign positive gradients

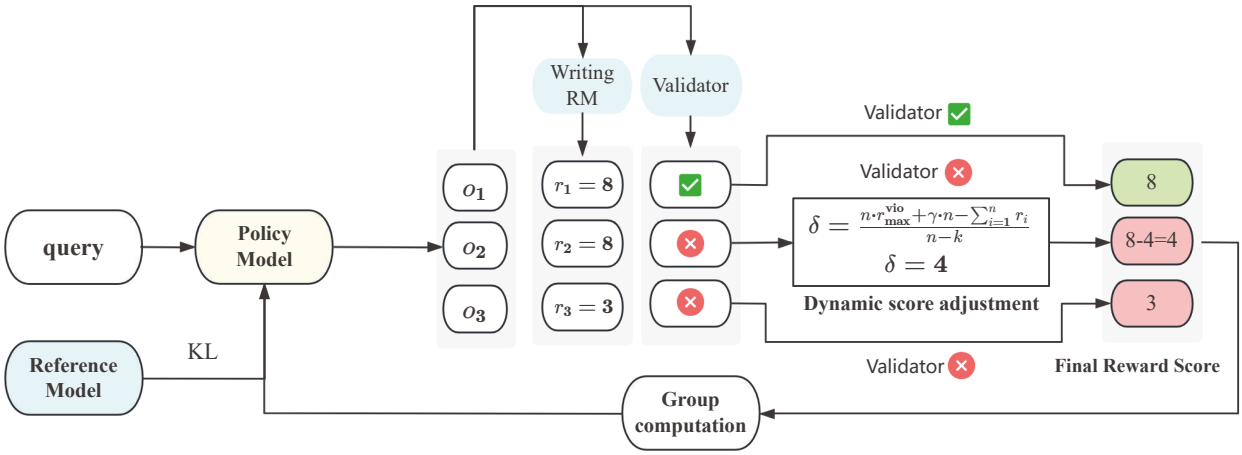


Figure 2: Overview of our Dynamic Mixed-Reward GRPO Framework. The policy model generates responses (o_1, o_2, o_3) evaluated by both writing quality (Writing RM) and constraint compliance (Validator). In this example: $n = 3$ (total samples), $r_{\max}^{\text{vio}} = 8$ (highest reward among violating samples), $\gamma = 1$ (minimum gap below the mean), $k = 1$ (number of violating samples), $\sum_{i=1}^{n} r_i = 19$ (sum of original rewards). The framework calculates penalty $\delta = 4$ and deducts it from violating samples ($o_2 : 8 \rightarrow 4$). After adjustment around mean=5, only high-quality compliant samples (o_1) receive positive gradients (green), while both low-quality samples (o_3) and constraint-violating samples (o_2) receive negative gradients (red).

to high-quality but constraint-violating samples, contradicting creative writing requirements. Second, the relative importance between subjective quality assessment and objective constraint following cannot be accurately determined, making weight assignment difficult. To address these issues, we propose a dynamic mixed-reward GRPO framework that adaptively adjusts penalty weights based on actual constraint compliance performance within each sampled group, ensuring constraint-violating samples consistently receive negative advantages during training. This dynamic adjustment approach is better suited for creative writing tasks.

RLMR Framework for Creative Writing

To effectively combine subjective and objective reward signals, we propose a mixed-reward GRPO framework. This framework integrates a writing reward model for evaluating writing quality with a verification model for assessing instruction compliance. By adjusting reward scores based on verification results, we achieve improved instruction-following capability while maintaining writing quality.

Reward Models

Our RLMR framework employs two reward models: a writing reward model that evaluates subjective writing quality and a constraint verification model that assesses objective compliance with task requirements.

Writing Reward Model. The writing reward model r_{write} evaluates the overall quality of creative writing outputs. We train this model on a large language model using human-annotated preference pairs (y_w, y_l) for creative writing prompts x. Following the Bradley-Terry preference model, we optimize:

$$\mathcal{L}_{\text{write}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma(r_{\text{write}}(x, y_w) - r_{\text{write}}(x, y_l))\right] \tag{1}$$

where y_w and y_l denote preferred and non-preferred responses, and σ is the sigmoid function. Unlike general reward models, our writing reward model captures creative writing features including literary expression, emotional depth, originality, narrative coherence, and stylistic maturity.

Constraint Verification Model The verification model identifies constraint violations in creative writing tasks, including word limits, formatting requirements, and content restrictions. For query q and response o, the model outputs:

$$V(o, q) = \bigwedge_{i=1}^{n} \text{verify}(o, c_i) \tag{2}$$

where $C = \{c_1, c_2, ..., c_n\}$ represents n identified constraints, and $\bigwedge$ denotes logical conjunction. A response is compliant only if all constraints are satisfied.

Dynamic Reward Adjustment Strategy

Fixed-weight reward fusion inadequately balances writing quality and constraint compliance. We introduce a dynamic adjustment mechanism that modifies original rewards before computing GRPO advantages. This ensures constraint-violating samples receive systematic penalties while preserving GRPO's comparative structure.

In standard GRPO, policy $\pi_{\theta_{\text{old}}}$ generates G responses $\{o_1, ..., o_G\}$ for query q with rewards $\{r_1, ..., r_G\}$. Advantages are computed as:

$$\hat{A}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \tag{3}$$

Our strategy ensures constraint-violating samples obtain negative advantages after normalization, acting as negative examples during optimization. Compliant samples receive positive advantages and are prioritized for learning.

For each query, we sample n responses $\mathcal{S} = \{s_1, ..., s_n\}$ with original rewards $\{r_1, ..., r_n\}$. We first identify constraint-violating samples through the verification model and adjust their rewards accordingly:

$$r_i' = \begin{cases} r_i & \text{if } V(s_i, q) = \text{True} \\ r_i - \delta & \text{if } V(s_i, q) = \text{False} \end{cases} \tag{4}$$

where $\delta > 0$ is the penalty term to be determined. Let k denote the number of constraint-violating samples in the group. The adjusted mean becomes:

$$\bar{r}' = \frac{1}{n}\sum_{i=1}^{n} r_i' = \frac{1}{n}\left(\sum_{i=1}^{n} r_i - k\delta\right) \tag{5}$$

To guarantee that all constraint-violating samples receive negative advantages after normalization, we require that for any violating sample j where $V(s_j, q) = \text{False}$:

$$r_j' < \bar{r}' - \gamma \tag{6}$$

where $\gamma > 0$ controls the minimum gap below the adjusted mean. This ensures violating samples will have sufficiently negative advantages to be suppressed during training. γ is generally set to 0.5 in this paper.

To determine the appropriate penalty δ, let $r_{\text{max}}^{\text{vio}}$ be the highest original reward among all constraint-violating samples. Substituting Equations (4) and (5) into inequality (6), we derive the penalty bound:

$$\delta \geq \frac{n \cdot r_{\text{max}}^{\text{vio}} + n \cdot \gamma - \sum_{i=1}^{n} r_i}{n - k} \tag{7}$$

Setting δ above this bound ensures all violating samples produce negative advantages, systematically suppressing them during gradient updates while preserving the relative ordering among compliant samples. This dynamic adjustment mechanism allows the model to learn from high-quality compliant responses while avoiding the reinforcement of constraint violations.

Dynamic Sampling Strategy Inspired by DAPO (Yu et al. 2025), we address gradient vanishing in creative writing RL training. When all sampled responses receive identical scores, zero advantages yield zero gradients. In creative tasks, this occurs with over-optimized samples, under-optimized samples, and samples where all responses violate constraints.

We implement a composite filtering strategy that removes three types of ineffective samples: (1) groups where all rewards exceed a high threshold, (2) groups where all rewards fall below a low threshold, and (3) groups where all responses fail verification. When filtered samples are insufficient, we dynamically resample new prompts to maintain adequate contrastive signals for effective training.

Experiments and Results

In this section, we show experiments to test our dynamic mixed-reward GRPO framework for creative writing. We describe the setup, share results, and give analysis.

Experimental Setup

Training Query Construction We construct our GRPO training queries from real-world seed data, we apply the self-instruct (Wang et al. 2023) methodology to expand the dataset diversity while maintaining realistic writing scenarios. To ensure balanced genre representation, we employ DeepSeek-V3 to classify generated queries by writing genre and adjust the sampling distribution to match real-world proportions observed in our seed data. This process yields a final training set of 8,739 queries.

Evaluation Benchmarks We test model performance on writing quality and instruction following using four benchmarks:

WritingBench (Yao et al. 2025) covers 6 main categories and 100 subdomains like academic, finance, politics, literature, education, and marketing. It has 1,239 real-world prompts, each with 5 custom criteria. We use Claude-4-Sonnet to score outputs.

WriteEval is our custom dataset containing 890 samples collected from real-world scenarios and augmented with LLM-generated instructions to match authentic writing styles. The dataset uniformly covers 30 primary writing genres and 377 secondary categories, including Chinese-specific genres such as folk texts, classical Chinese, and composition writing. For each instruction, we solicited responses from six competitive Chinese writing models: Claude-4-Sonnet, Gemini-2.5-Pro, DeepSeek-R1, DeepSeek-V3, Doubao-1.5-Thinking, and Hunyuan-TurboS. Human experts conducted blind evaluation to select the best response from each set as reference answers. For automated evaluation, Claude-4-Opus compares model outputs against reference answers to determine win rates: Win Rate $= \frac{\text{Number of wins}}{\text{Total comparisons}} \times 100\%$ where a "win" indicates the model output is judged superior to the reference answer. Detailed prompt templates are provided in the appendix.

ComplexBench (Wen et al. 2024) checks complex instruction following with combined constraints. It builds hard prompts that need to meet multiple rules. Scoring uses questions to check each part.

IFEval (Zhou et al. 2023) is Google's benchmark for verifiable instructions like word count or keywords. It has 25 types across 500 prompts. We use prompt-level strict-accuracy for evaluation.

Baseline Methods To evaluate our dynamic mixed-reward strategy, we compare against three baseline methods that represent the spectrum of existing reward strategies in creative writing optimization:

(1) Writing Reward Only GRPO: This baseline trains using only writing quality rewards without any constraint verification signals. This method represents the traditional approach in RLHF where models are optimized solely based on human preference signals for output quality (Ouyang

Model	Method	Writing Quality		Instruction Following	
		WritingBench	WriteEval	ComplexBench	IFEval
Qwen2.5-32B	Original Model	6.14	3.93%	74.78%	83.36%
	GRPO Baseline (Writing RM only)	7.05	7.95%	68.42%	80.41%
	GRPO Baseline (Verification Model only)	5.73	1.24%	**83.94%**	82.77%
	Linear Weighting	7.13	6.40%	73.91%	84.04%
	RLMR(w/o DAPO)	7.34	9.31%	77.83%	**87.14%**
	RLMR(Ours)	**7.93**	**11.56%**	79.04%	86.65%
Qwen2.5-72B	Linear Weighting	6.43	10.22%	74.78%	85.58%
	RLMR(Ours)	**7.81**	**17.18%**	**80.21%**	**87.79%**
Qwen3-8B	Linear Weighting	7.61	26.64%	77.16%	83.14%
	RLMR(Ours)	**8.13**	**31.69%**	**82.01%**	**86.43%**
DeepSeek-R1-Distill-Llama-8B	Linear Weighting	5.68	1.46%	**53.91%**	56.38%
	RLMR(Ours)	**7.41**	**3.57%**	52.35%	**60.94%**

Table 1: Performance comparison across different models and methods on writing quality and instruction-following benchmarks. Our dynamic mixed-reward approach consistently outperforms baseline methods across all model scales.

et al. 2022; Stiennon et al. 2020). Following established RLHF practices, this baseline uses a reward model trained on human-annotated preference pairs to score creative writing outputs (Dong et al. 2024).

(2) Verification Signal Only GRPO: This baseline uses only binary constraint verification signals (pass/fail) without considering writing quality. This approach aligns with recent work on Reinforcement Learning with Verifiable Rewards (RLVR), where models are trained using deterministic verification functions for tasks with clear correctness criteria (Cobbe et al. 2021; Mroueh 2025). TheBy comparing against these methods, we demonstrate that our dynamic mixed-reward strategy addresses the limitations of both single-reward and fixed-weight approaches, providing a more effective solution for creative writing optimization.

(3) Linear Weighting Strategy: Following the approach proposed by Peng et al. (2025b), this baseline combines writing rewards with verification signals through fixed-weight linear combination. Specifically, we normalize both writing rewards and verification scores to the [0,1] range and compute their arithmetic mean: $(s_{\text{normalized\_writing}} + s_{\text{normalized\_verification}})/2$. This method represents the current state-of-the-art in mixed-reward strategies, as demonstrated in the Agentic Reward Modeling framework (Peng et al. 2025b), which successfully integrates human preference rewards with verifiable correctness signals including factuality and instruction following.

Reward Model and Training Setup

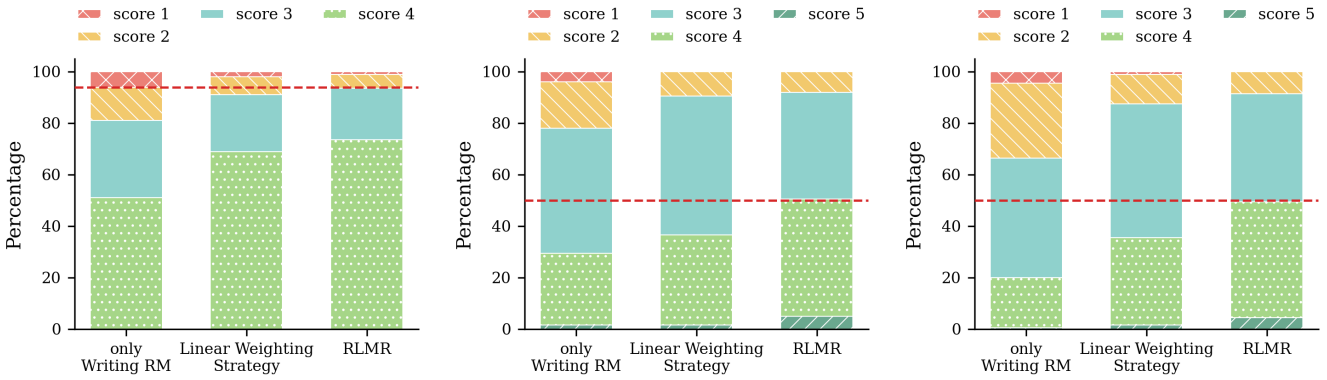
Writing Reward Model. We use a Pointwise Bradley-Terry Reward Model (Bradley and Terry 1952; Ouyang et al. 2022) for continuous feedback. It trains on Tencent-Hunyuan-Large (Sun et al. 2024) with 200,000 labeled samples. Each sample has a prompt and two responses; humans pick the better one based on quality, adherence, style, and experience. We use this model for rewards in RLHF to match human preferences.

Constraint Verification Model. We use Qwen2.5-72B-Instruct with prompts to check constraints. It makes checklists and verifies each one. We employ binary verification (all constraints satisfied or not) rather than proportion-based scoring because partial constraint satisfaction is functionally equivalent to complete failure in creative writing tasks. This binary approach ensures the model learns to generate responses that satisfy all constraints simultaneously, rather than trading off between different constraint types. See appendix for prompt details.

Experimental Results

Automated Evaluation Results We test our framework on four models: Qwen2.5-32B and Qwen2.5-72B (Team 2024; Yang et al. 2024), Qwen3-8B (Yang et al. 2025), and DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI 2025). Table 1 shows results across methods and benchmarks.

The automated evaluation results reveal compelling evidence for the effectiveness of our dynamic mixed-reward approach. Results from Qwen2.5-32B clearly expose the inherent problems with single reward signals. When training with writing RM alone, writing quality improves modestly from 6.14 to 7.05, yet instruction following suffers substantial degradation: ComplexBench performance drops from 74.78% to 68.42%, while IFEval accuracy falls from 83.36% to 80.41%. The reverse pattern emerges when using only the constraint verification model—instruction following on ComplexBench rises from 74.78% to 83.94%, but writing quality plummets from 6.14 to 5.73, with WriteEval performance collapsing from 3.93% to a mere 1.24%. This stark trade-off demonstrates that single signals cannot balance subjective creative quality with objective constraint adherence.



(a) Instruction Following (b) Content Quality (c) Overall Performance

Figure 3: Human evaluation score distributions across three dimensions. The red dashed line indicates the satisfactory threshold (score ≥ 3 for Content Quality and Overall Performance, score = 4 for Instruction Following). Our RLMR method consistently shows higher proportions of satisfactory scores compared to baseline methods.

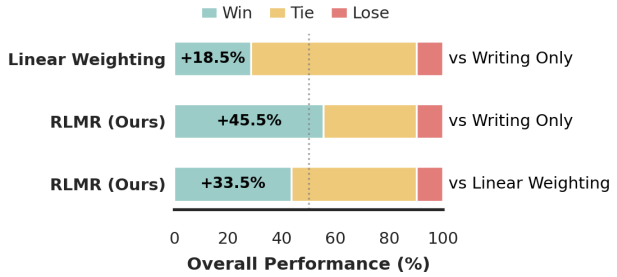


Figure 4: Pairwise comparison results for Overall Performance. "Win" indicates the left method outperforms the right method; "tie" indicates comparable performance; "lose" indicates the left method underperforms. The red dashed line represents equal performance (50%). RLMR demonstrates significant advantages over both baseline methods.

Given these limitations, mixed-reward strategies emerge as a natural solution by combining writing RM with constraint verification signals. The most classical approach is linear weighting, which averages the two reward types with fixed coefficients. On Qwen2.5-32B, this approach elevates writing quality to 7.13 while preserving reasonable instruction following capabilities, successfully avoiding the severe bias problems observed with single-signal methods. These results underscore the critical importance of integrating both subjective and objective evaluation dimensions in creative writing optimization.

However, our RLMR method delivers even greater improvements, consistently outperforming linear weighting across all tested models. On Qwen2.5-32B, RLMR pushes writing quality further to 7.93 and achieves an 11.56% WriteEval win rate, substantially surpassing linear weighting's 7.13 and 6.40% respectively. This pattern of superior performance extends to other architectures: Qwen3-8B sees writing quality rise from 7.61 to 8.13, with WriteEval win rates jumping from 26.64% to 31.69%. Similarly, Qwen2.5-72B confirms this trend, with WriteEval performance climbing from 10.22% to 17.18%.

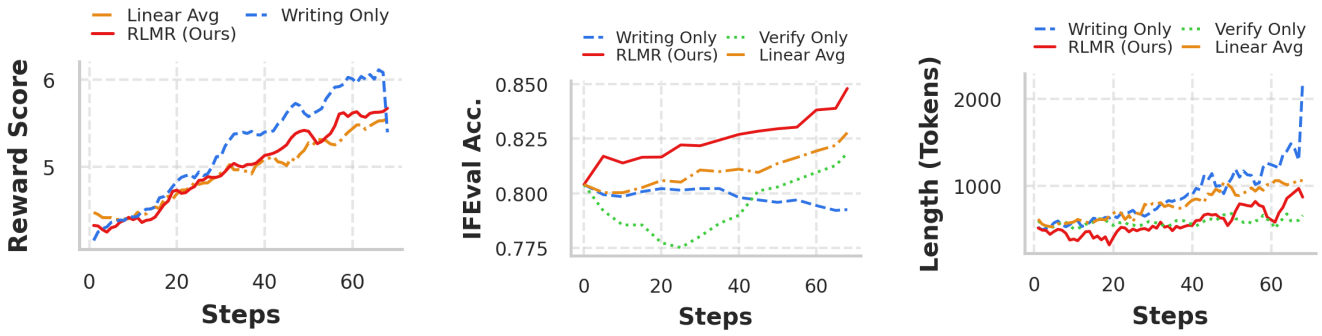
The robustness of these improvements becomes evident when examining results across diverse model scales. Our experiments span architectures ranging from 8B to 72B parameters, including both Qwen and DeepSeek families, with all models demonstrating consistent advantages under RLMR.

Manual Evaluation Results We conducted human evaluation on 200 randomly sampled instances from the WriteEval dataset to assess model performance across three dimensions: Instruction Following, Content Quality, and Overall Performance. Detailed scoring criteria and guidelines are provided in the appendix. For Instruction Following, we consider a score of 4 as complete instruction adherence. For Content Quality and Overall Performance, scores of 3 or above are considered satisfactory.

Figure 3 presents the score distribution across all three evaluation dimensions. The results clearly demonstrate the limitations of single-reward strategies. The writing-only baseline shows inferior performance across multiple dimensions compared to mixed-reward approaches, with notably lower satisfactory rates in instruction following and content quality. Among mixed-reward strategies, our RLMR method achieves higher satisfactory rates across all dimensions.

Specifically, for Instruction Following, RLMR shows the highest proportion of perfect scores (score 4), indicating superior constraint adherence. In Content Quality, RLMR demonstrates a more favorable distribution with increased proportions in higher score ranges (scores 4-5), suggesting better content generation capabilities. The Overall Performance dimension reveals similar trends, with RLMR achieving the most balanced distribution toward higher satisfaction levels.

Figure 4 shows the results of direct pairwise comparisons for Overall Performance. RLMR achieves substantial win



(a) Writing Reward Scores (b) IFEval Performance (c) Response Length

Figure 5: Training dynamics across different metrics. (a) Writing reward model scores during training. (b) IFEval performance during training. (c) Generated response length during training.

rates against both baseline methods: 45.5% win rate versus writing-only baseline and 33.5% win rate versus linear weighting strategy. These results demonstrate that our RLMR strategy achieves higher usability and satisfaction rates in creative writing tasks, confirming the practical effectiveness of our approach.

Experimental Analysis

The experimental results demonstrate that single reward signals fail to balance writing quality and instruction following effectively. Using only writing rewards improves creative quality but reduces constraint adherence. Using only verification signals severely harms writing quality while providing limited gains in instruction following. These findings confirm that creative writing optimization requires careful integration of both subjective and objective evaluation criteria.

Our dynamic mixed-reward strategy significantly outperforms linear weighting approaches across all tested models and benchmarks. This superiority stems from fundamental limitations of fixed-weight methods. Writing quality scores and constraint verification signals operate on different scales and distributions. Writing rewards typically follow continuous distributions, while constraint verification produces binary outcomes. The scalar inconsistency between these two signals makes it difficult to determine appropriate weighting coefficients. Moreover, optimal weighting coefficients need adjustment for different reward models, making fixed-weight approaches impractical across diverse model configurations.

Our dynamic adjustment mechanism addresses these limitations by calculating penalty terms based on actual constraint compliance patterns within each sample group. Rather than applying uniform weights, the approach modulates penalties according to the theoretical bounds derived in Equation (7). This ensures constraint-violating samples consistently receive negative advantages and are suppressed during training.

Figure 5 shows training dynamics across key metrics. The writing RM only baseline achieves the highest writing reward scores during training (Figure 5a), but this improvement reveals classic reward hacking behavior. Despite high reward scores, its IFEval performance deteriorates significantly (Figure 5b), dropping below both the original model and other baselines. This divergence between reward scores and actual instruction-following capability demonstrates that the model learns to exploit the reward model rather than genuinely improving writing quality.

The reward hacking behavior is further evidenced by the dramatic increase in response length (Figure 5c). The writing RM only baseline shows uncontrolled length growth, reaching over 1400 tokens on average, which explains its poor instruction-following performance. When models generate excessively long outputs, they cannot properly adhere to specific constraints like word count limits, format requirements, or conciseness instructions.

In contrast, our RLMR method maintains balanced optimization across all metrics. It achieves steady improvement in writing reward scores while preserving strong IFEval performance, demonstrating that our dynamic reward adjustment successfully prevents the model from exploiting either reward signal. The controlled response length further confirms that RLMR learns to generate high-quality content without resorting to length inflation. This balanced training dynamic validates the effectiveness of our dynamic penalty mechanism in creating models that excel at both creative quality and constraint adherence.

Conclusion

We propose RLMR, a dynamic mixed-reward GRPO framework balancing creative quality and constraint adherence. By dynamically penalizing violations to ensure negative advantages, RLMR overcomes single-reward and fixed-weight limitations. Experiments across diverse architectures demonstrate substantial improvements in quality and compliance, confirmed by human evaluation. Our method prevents reward hacking while maintaining stable optimization. Future work includes extensions to dialogue and code generation.

References

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

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

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Dong, H.; Xiong, W.; Pang, B.; Wang, H.; Zhao, H.; Zhou, Y.; Jiang, N.; Sahoo, D.; Xiong, C.; and Zhang, T. 2024. RLHF Workflow: From Reward Modeling to Online RLHF. *arXiv preprint arXiv:2405.07863*.

Jia, R.; Yang, Y.; Gai, Y.; Luo, K.; Huang, S.; Lin, J.; Jiang, X.; and Jiang, G. 2025. Writing-Zero: Bridge the Gap Between Non-verifiable Tasks and Verifiable Rewards. *arXiv preprint arXiv:2506.00103*.

Mroueh, Y. 2025. Reinforcement Learning with Verifiable Rewards: GRPO's Effective Loss, Dynamics, and Success Amplification. *arXiv preprint arXiv:2503.06639*.

Ouyang, L.; Wu, J.; Jiang, X.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Peng, H.; Qi, Y.; Wang, X.; Yao, Z.; Xu, B.; Hou, L.; and Li, J. 2025a. Agentic Reward Modeling: Integrating Human Preferences with Verifiable Correctness Signals for Reliable Reward Systems. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15934–15949. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Peng, H.; Qi, Y.; Wang, X.; Yao, Z.; Xu, B.; Hou, L.; and Li, J. 2025b. Agentic Reward Modeling: Integrating Human Preferences with Verifiable Correctness Signals for Reliable Reward Systems. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15934–15949. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Sun, X.; Chen, Y.; Huang, Y.; Xie, R.; Zhu, J.; Zhang, K.; Li, S.; Yang, Z.; Han, J.; Shu, X.; Bu, J.; Chen, Z.; Huang, X.; Lian, F.; Yang, S.; Yan, J.; Zeng, Y.; Ren, X.; Yu, C.; Wu, L.; Mao, Y.; Xia, J.; Yang, T.; Zheng, S.; Wu, K.; Jiao, D.; Xue, J.; Zhang, X.; Wu, D.; Liu, K.; Wu, D.; Xu, G.; Chen, S.; Chen, S.; Feng, X.; Hong, Y.; Zheng, J.; Xu, C.; Li, Z.; Kuang, X.; Hu, J.; Chen, Y.; Deng, Y.; Li, G.; Liu, A.; Zhang, C.; Hu, S.; Zhao, Z.; Wu, Z.; Ding, Y.; Wang, W.; Liu, H.; Wang, R.; Fei, H.; Yu, P.; Zhao, Z.; Cao, X.; Wang, H.; Xiang, F.; Huang, M.; Xiong, Z.; Hu, B.; Hou, X.; Jiang, L.; Ma, J.; Wu, J.; Deng, Y.; Shen, Y.; Wang, Q.; Liu, W.; Liu, J.; Chen, M.; Dong, L.; Jia, W.; Chen, H.; Liu, F.; Yuan, R.; Xu, H.; Yan, Z.; Cao, T.; Hu, Z.; Feng, X.; Du, D.; Yu, T.; Tao, Y.; Zhang, F.; Zhu, J.; Xu, C.; Li, X.; Zha, C.; Ouyang, W.; Xia, Y.; Li, X.; He, Z.; Chen, R.; Song, J.; Chen, R.; Jiang, F.; Zhao, C.; Wang, B.; Gong, H.; Gan, R.; Hu, W.; Kang, Z.; Yang, Y.; Liu, Y.; Wang, D.; and Jiang, J. 2024. Hunyuan-Large: An Open-Source MoE Model with 52 Billion Activated Parameters by Tencent. arXiv:2411.02265.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560.

Wen, B.; Ke, P.; Gu, X.; Wu, L.; Huang, H.; Zhou, J.; Li, W.; Hu, B.; Gao, W.; Xu, J.; et al. 2024. Benchmarking Complex Instruction-Following with Multiple Constraints Composition. *arXiv preprint arXiv:2407.03978*.

Wu, Y.; Bai, Y.; Hu, Z.; Lee, R. K.-W.; and Li, J. 2025. LongWriter-Zero: Mastering Ultra-Long Text Generation via Reinforcement Learning. *arXiv preprint arXiv:2506.18841*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.21783*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yao, L.; et al. 2025. WritingBench: A Comprehensive Benchmark for Generative Writing. *arXiv preprint arXiv:2503.05244*.

Yu, Y.; Liu, Y.; Chen, H.; et al. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint arXiv:2503.14476*.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*.