

DPWriter: Reinforcement Learning with Diverse Planning Branching for Creative Writing

Qian Cao^{1†*}, Yahui Liu^{2†}, Wei Bi², Yi Zhao¹, Ruihua Song^{1✉}, Xiting Wang^{1✉},
Ruiming Tang², Guorui Zhou², Han Li²

¹Renmin University of China, ²Kuaishou Technology
{caoqian4real, rsong, xitingwang}@ruc.edu.cn,
yahui.cvrs@gmail.com

Abstract

Reinforcement learning (RL)-based enhancement of large language models (LLMs) often leads to reduced output diversity, undermining their utility in open-ended tasks like creative writing. Current methods lack explicit mechanisms for guiding diverse exploration and instead prioritize optimization efficiency and performance over diversity. This paper proposes an RL framework structured around a semi-structured long Chain-of-Thought (CoT), in which the generation process is decomposed into explicitly planned intermediate steps. We introduce a Diverse Planning Branching method that strategically introduces divergence at the planning phase based on diversity variation, alongside a group-aware diversity reward to encourage distinct trajectories. Experimental results on creative writing benchmarks demonstrate that our approach significantly improves output diversity without compromising generation quality, consistently outperforming existing baselines. The code and data are available at <https://github.com/Aman-4-Real/DPWriter>.

1 Introduction

Diversity is a fundamental characteristic of the world and a core manifestation of human creativity (Fischer, 2005; Edmonds, 2008; Wang et al., 2026). As large language models (LLMs) continue to advance in reasoning capabilities (Chen et al., 2025a), and more recently by reinforcement learning (RL) (Schulman et al., 2017; Shao et al., 2024; Lu, 2025; Bhaskar et al., 2025), the diversity exhibited in their generated texts has become particularly important. However, users may experience a significant loss of content diversity when collaborating with LLMs for creative writing (Padmakumar and He, 2024). This issue is even more pronounced in

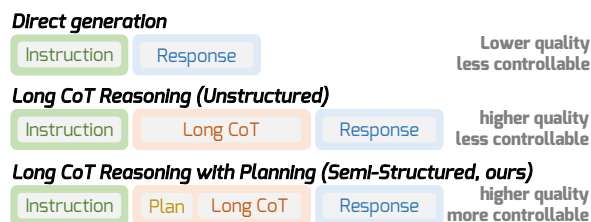


Figure 1: Comparison among three generation paradigms. Our semi-structured reasoning paradigm introduces global planning before reasoning, providing high-level guidance while maintaining higher quality.

models trained with reinforcement learning from human feedback (RLHF) (O’Mahony et al., 2024).

To alleviate the diversity decline caused by RL training (Kirk et al., 2024; Padmakumar and He, 2024; Shypula et al., 2025), a growing body of works are proposed (He et al., 2025; Li et al., 2025a; Ansel et al., 2025), yet several challenges remain in effectively enhancing diversity for LLMs. First, many approaches focus on modifying reward functions (Tuyls et al., 2025; He et al., 2025; Li et al., 2025a). However, these methods largely leave the rollout process unconstrained, providing limited control over how diverse trajectories are explored during RL. Second, some methods investigate branching or forking strategies to explore diverse trajectories (Zheng et al., 2025a; Li et al., 2025c; Guo et al., 2025), but they primarily focus on improving sample efficiency (Zheng et al., 2025b; Wen et al., 2025) or overall performance (Liu et al., 2025b), rather than explicitly targeting diversity as an intrinsic objective. Moreover, they typically branch rollouts from high-entropy tokens (Wang et al., 2025b), which makes the branching process less controllable.

In this paper, we propose DPWriter, which uses a semi-structured long Chain-of-Thought (CoT) as a scaffold to guide the RL process for improved diversity in LLMs. As shown in Figure 1, by decomposing the generation process into multiple stages,

*Work done during an internship at Kuaishou Technology.

† Equal contribution.

✉ Corresponding author.

beginning with a global planning phase followed by long CoT reasoning and final response generation, our method provides explicit intermediate stages that facilitate diverse exploration. Specifically, we introduce a Diverse Planning Branching (DPB) method that strategically branches diverse plans at our rollout-time DPB stage based on their diversity, allowing for more controlled and effective exploration of diverse trajectories. We also incorporate a diversity reward that evaluates the diversity contribution of each response based on its group, collaborating with our rollout-time DPB strategy to further encourage diverse generation. Extensive experiments on various creative writing benchmarks demonstrate that our approach significantly enhances the diversity of LLM-generated texts while maintaining high quality, consistently outperforming existing baselines. The main contributions of this work are as follows:

- *Diversity-guided RL framework.* We propose a novel RL framework that leverages semi-structured long CoT to guide the generation process. To support this, we construct a curated dataset comprising 43K writing instructions with semi-structured long CoT and high-quality responses.

- *Planning-level diversity mechanisms.* We introduce a Diverse Planning Branching (DPB) method that strategically branches candidate plans, along with a diversity reward that evaluates each response according to its contribution within a group. This design enables controlled exploration and effectively promotes diverse generation trajectories.

- *Experimental validation.* Experiments on multiple creative writing benchmarks demonstrate that our method significantly improves the diversity of LLM-generated texts while maintaining high generation quality, consistently outperforming existing baselines. Further analysis reveals that the DPB method and diversity reward work synergistically, jointly promoting more diverse generation.

2 Related Work

Diversity in Non-RL Training. Previous studies have shown that supervised fine-tuning (SFT) or preference optimization may reduce output diversity (O’Mahony et al., 2024; Kirk et al., 2024), motivating another line of work aimed to alleviating this issue during training. Li et al. (2025e) emphasizes the overfitting issue inherent in SFT and introduces a game-theoretic framework to address the limitations of cross-entropy loss. For

preference optimization, some studies (Lanchantin et al., 2025; Deshpande et al., 2025) propose modifications to DPO (Rafailov et al., 2023) that focus on improved selection of diversified data samples. Other methods (Chung et al., 2025; Nath et al., 2025) promote both output diversity and quality by employing weighted training objectives that better capture nuanced preferences.

RL-based Methods for Diversity. More recently, reinforcement learning (RL) has demonstrated strong effectiveness in improving model capabilities (Wei et al., 2025; Bhaskar et al., 2025), leading to increased attention to methods that boost diversity during RL training. A primary strategy modifies the reward in policy gradient methods like GRPO (Shao et al., 2024) by introducing diversity-aware bonus (Anschel et al., 2025; Tuyls et al., 2025) or penalty terms (Chen et al., 2025b; He et al., 2025; Li et al., 2025a), which are computed using diversity metrics over a group of generated responses. The common goal is to shape the policy gradient to favor diverse and high-quality outputs. In addition, some researchers have explored alternative approaches to adjust RL objectives, such as incorporating token-level entropy regularization (Yao et al., 2025), designing semantic diversity terms (Chen et al., 2025c), or decoupling an entropy component from the KL divergence term (Slocum et al., 2025).

Although some studies investigate branching or forking strategies to enhance RL exploration (Zheng et al., 2025a; Li et al., 2025c; Guo et al., 2025), they primarily aim to improve sample efficiency (Zheng et al., 2025b; Wen et al., 2025) and overall performance (Liu et al., 2025b) rather than explicitly promoting diversity. Moreover, these works mainly focus on high-entropy tokens (Wang et al., 2025b) as branching points (Zheng et al., 2025a; Liu et al., 2025b) or set fixed segment lengths for branching (Li et al., 2025c; Guo et al., 2025), which makes the resulting rollouts less controllable. Unlike prior work, our method integrates a semi-structured long CoT reasoning process with diversity-aware branching strategies to explicitly encourage the exploration of multiple, divergent planning pathways, making it suitable for open-ended creative writing tasks.

3 Preliminaries

Task Formulation. Given an instruction $q \sim Q$ from an open-ended task like creative writing, the

goal of a model \mathcal{M} is to generate a response y to the instruction, *i.e.*, $y \sim \mathcal{M}(\cdot|q)$. Using long CoT reasoning to generate a response can be formulated as first generating a reasoning chain c , followed by generating the final response, *i.e.*, $c \sim \mathcal{M}(\cdot|q)$ and $y \sim \mathcal{M}(\cdot|q, c)$. However, existing CoT reasoning processes are unstructured and implicitly learned, lacking explicit planning representations.

To allow high-level objectives to directly shape subsequent reasoning and final responses, we propose a semi-structured long CoT reasoning paradigm that introduces an explicit planning stage before reasoning. Specifically, the model first generates a global plan p for the response, then produces a reasoning chain c conditioned on both the instruction and the plan, and finally generates the response conditioned on all of them, *i.e.*, $p \sim \mathcal{M}(\cdot|q)$, $c \sim \mathcal{M}(\cdot|q, p)$ and $y \sim \mathcal{M}(\cdot|q, p, c)$. In our proposed paradigm, as shown in Figure 1, the plan p serves as a high-level structural guide for the subsequent free-form reasoning and response generation. This strikes a balance between the flexibility of unstructured reasoning chains and the control provided by explicit planning.

RL for LLMs. In the context of RL for LLMs, the model \mathcal{M} is treated as a policy π_θ with parameters θ , and the generation process is formulated as a Markov Decision Process (MDP). At time step t , generating a token a_t is treated as taking an action in state s_t , where the state consists of the instruction and all previously generated tokens *i.e.*, $s_t = (q, a_1, a_2, \dots, a_{t-1})$. In our semi-structured reasoning paradigm, the generation of a rollout o can be denoted as the following distribution:

$$\pi_\theta(o | q) = \prod_{\tau \in \{p, c, y\}} \prod_{l=1}^{L_\tau} \pi_\theta(a_l^{(\tau)} | q, \tau_{<l}), \quad (1)$$

where τ iterates over the plan p , reasoning chain c , and response y . L_τ is the length of sequence τ , and $a_l^{(\tau)}$ is the l -th token in it. The objective of RL is to maximize the expected cumulative reward:

$$J(\theta) = \mathbb{E}_{o \sim \pi_\theta(\cdot|q)} [r(q, o)], \quad (2)$$

where $r(q, o)$ is the reward function that evaluates the quality of rollout o given the instruction q .

Group Relative Policy Optimization (GRPO). Our method is built upon GRPO (Shao et al., 2024), a recent RL algorithm that discards the critic model and estimates advantages within a group of n rollouts $\{o_i\}_{i=1}^n$ generated by the old policy $\pi_{\theta_{\text{old}}}$ as

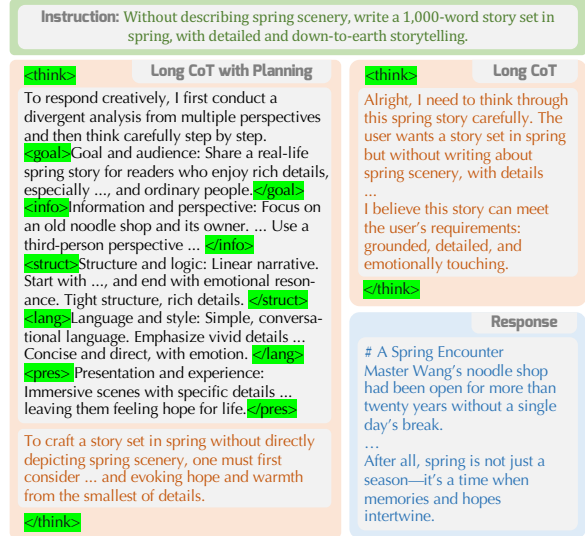


Figure 2: An example of the original long CoT data and the semi-structured long CoT with planning. Texts with colored background represent special tokens.

in (Schulman et al., 2017). GRPO optimizes the policy π_θ by maximizing the following objective:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{Q}, \{o_i\}_{i=1}^n \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \quad (3)$$

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left[\text{CLIP}(\rho_{i,t}, A_t) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\theta_{\text{ref}}}) \right],$$

where $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$ is the importance sampling ratio at step j of rollout o_i . The clip function $\text{CLIP}(\rho_{i,t}, A_t) = \min(\rho_{i,t} A_t, \text{clip}(\rho_{i,t}, 1 - \epsilon, 1 + \epsilon) A_t)$ is used to limit the policy update step size (Schulman et al., 2017), and the \mathbb{D}_{KL} term penalizes the divergence from a reference policy $\pi_{\theta_{\text{ref}}}$ to further ensure stability, with β being the penalty coefficient. Given rewards $\{r_i\}_{i=1}^n$ of a group of rollouts $\{o_i\}_{i=1}^n$, the advantage A_t for each rollout is computed as $A_t = \frac{r_i - \bar{r}}{\sigma_r}$, where \bar{r} and σ_r are the mean and standard deviation of the rewards.

4 Constructing Semi-structured CoT Data with Planning

We present a two-step method for constructing semi-structured CoT data through multi-aspect planning and plan-consistent reasoning, enhancing coherence and controllability.

Multi-aspect Planning Generation. Given an input instruction q , a long CoT reasoning c , and a target response y , the goal is to generate a plan p that outlines key aspects to guide both the CoT and response generation. Inspired by rhetorical and writing theories (Bitzer, 1968; Flower and Hayes,

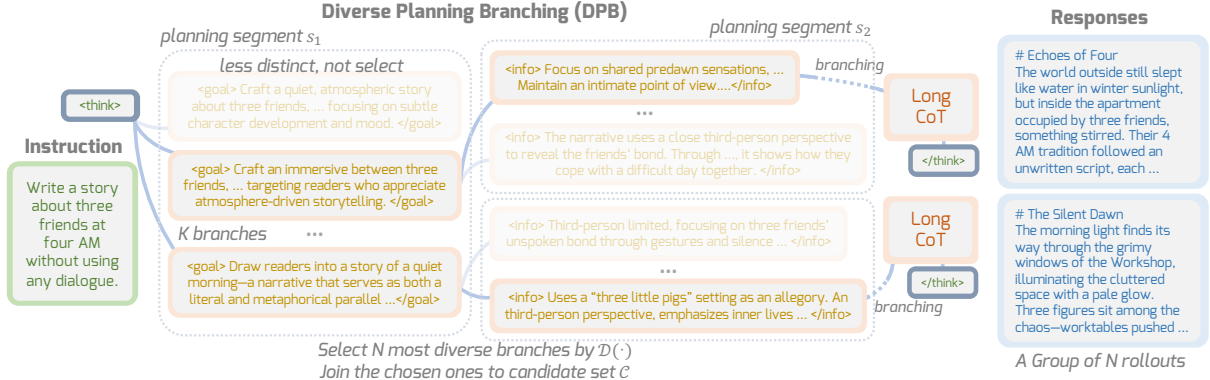


Figure 3: An overview of our Diverse Planning Branching method. During RL, at each planning segment, we branch out multiple diverse continuations from each candidate, forming a pool of candidates. We then select the most diverse ones to proceed to the next segment, ultimately generating diverse final responses.

1981; Cao et al., 2022; Spangher et al., 2025), we design a multi-aspect planning framework that includes the following aspects:

- *Goal and Audience*: To define and identify the primary objective and target audience.
- *Information and Perspective*: To highlight the key information to be included and the perspective or viewpoint to be adopted.
- *Structure and Logic*: To outline the logical flow and structure of the response, including main points and their organization.
- *Language and Style*: To specify the desired tone, vocabulary, and stylistic elements to be used.
- *Presentation and Experience*: To describe how the information should be presented to enhance reader engagement and experience.

We employ GPT-4.1 (OpenAI, 2025) to produce the plan p based on the instruction q and response y . Details of the used prompt are in Appendix A.3.

Plan-consistent CoT Generation. However, directly inserting the generated plan p at the beginning of the long CoT c may introduce inconsistency between the plan and the reasoning process. To overcome this, we use the constructed plan p to revise the original CoT c into a plan-consistent CoT c' . This is achieved by using GPT-4.1 to revise the original CoT c based on the plan p , ensuring alignment with the planned aspects while preserving the original information to avoid information drift. The revision prompt is provided in Appendix A.3.

We introduce special tokens like $\langle \text{goal} \rangle$, $\langle / \text{goal} \rangle$, $\langle \text{info} \rangle$ and $\langle / \text{info} \rangle$, etc., to enclose individual aspects of the plan p , making the structure explicit and easier for the model to recognize and follow. An example of (q, c, y) and the semi-structured CoT (q, p, c', y) is shown in Figure 2.

5 Method

In this section, we present *Diverse Planning Branching* method and *Rewarding Diversity Contribution* strategy, which exploit planning controllability in semi-structured CoTs to enhance rollout diversity and response quality.

5.1 Diverse Planning Branching

Planning Capability Cold Start. The initial semi-structured long CoT (p, c') provides explicit planning cues and a consistent CoT reasoning process, both of which are reflected in the final response y . To equip the model with this capability, we first cold-start the base model through supervised fine-tuning (SFT) on the semi-structured CoT data, allowing it to learn the planning formats and generate coherent CoT reasoning. As shown in Section 6.3, SFT using our semi-structured data yields performance comparable to or exceeding that obtained with other CoT datasets.

Branching Planning Segments. In the RL stage, the cold-started model acts as the policy model π_θ . Given an instruction q , the model first generates the planning part before producing the CoT c' and response y . By explicitly encouraging exploration over diverse planning strategies at each planning point, the policy induces diverse reasoning paths and final outputs. This, in turn, offers the reward model a broader set of high-quality candidates, improving both generation quality and diversity.

As shown in Figure 3, we identify S planning segments in the semi-structured CoT format, each segment s is delimited by start and end tokens $(\tau_s^{\text{start}}, \tau_s^{\text{end}})$ (e.g., $\langle \text{goal} \rangle$ and $\langle / \text{goal} \rangle$). During generation, for each planning segment s , we expand every candidate in the current candidate set

\mathcal{C} by sampling K continuations starting from τ_s^{start} until reaching the corresponding ending token τ_s^{end} . Here, K denotes the *branch factor*, which controls the number of diverse continuations generated for each candidate. This process yields a candidate pool of size $|\mathcal{C}| \times K$. To select G candidates for the next segment, where G is the *group size*, we measure candidate diversity using a predefined diversity metric $D(\cdot)$ and select the most diverse candidates. For the first segment, the G candidates are selected directly from the entire pool. For subsequent segments, we select one candidate from each group of continuations originating from the same previous candidate, thereby ensuring diversity across different branches. After processing all S segments, we decode each candidate in \mathcal{C} completion, yielding $B \times G$ final responses for a batch of B instructions.

Diversity Metrics. To measure candidate diversity during branching, we consider two types of metrics: (1) *N-gram-based Diversity*, which calculates distinct n-grams across candidates to encourage lexical variety, and (2) *semantic Diversity*, which measures the average pairwise cosine distance between candidate embeddings by using an *off-the-shelf* embedding model (i.e., Qwen3-Embedding-0.6B (Zhang et al., 2025a)) to capture semantic differences. These metrics jointly encourage exploration of diverse reasoning paths.

5.2 Rewarding Quality and Diversity

To encourage reasoning paths that yield both high-quality and diverse responses, we follow previous works (Chen et al., 2025b; He et al., 2025; Li et al., 2025a) by jointly incorporating quality and diversity rewards during RL training.

Quality Reward. For the quality reward, we utilize a reward model R_ϕ trained on human preference data to assess the quality of generated responses. Given a response y_i to an instruction q , the quality reward is defined as:

$$r_i^{\text{qua}}(q, y_i) = R_\phi(q, y_i), \quad (4)$$

This reward encourages the model to generate responses that align with human preferences.

Diversity Contribution Reward. To further promote diversity among the generated responses, we introduce a *Diversity Contribution Reward*, which measures how much a response contributes to the overall diversity of the response group. The core intuition is to reward responses that introduce unique elements not shared by others, thereby promoting

varied content generation. Formally, given a response group $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$ for an instruction q , the diversity contribution reward for a response y_i is defined as:

$$r_i^{\text{div}}(q, y_i, \mathcal{Y}) = \text{Norm}\left(\frac{D(y_i, \mathcal{Y} \setminus \{y_i\})}{|y_i|}\right), \quad (5)$$

where $D(\cdot)$ counts the unique n-grams in y_i that do not appear in the other responses in $\mathcal{Y} \setminus \{y_i\}$, and $|y_i|$ denotes the number of tokens in y_i . The normalization function $\text{Norm}(\cdot)$ ensures the reward is on a comparable scale across responses.

To balance quality and diversity, we combine the two rewards as follows:

$$r_i(q, y_i, \mathcal{Y}) = (1 - \lambda) \cdot r_i^{\text{qua}} + \lambda \cdot r_i^{\text{qua}} \cdot r_i^{\text{div}}, \quad (6)$$

where $\lambda \in [0, 1]$ controls the contribution of the diversity reward. The diversity contribution reward is activated only when the response quality exceeds a certain threshold τ when $r_i^{\text{qua}} > \tau$; otherwise, we set $\lambda = 0$. This formulation ensures that responses are rewarded for diversity only when they satisfy a minimum quality threshold, favoring high-quality responses that also contribute meaningful diversity and guiding the model toward generating responses that are both high quality and diverse.

6 Experiments

In this section, we evaluate our proposed method on several benchmark datasets and compare it with relevant baselines. We further conduct ablation studies to analyze key components of our approach and provide discussions of the results.

6.1 Experimental Setup

Training Datasets. We adopt open datasets of creative writing for training the model, including DeepWriting (Wang et al., 2025a), WritingPrompts (Fan et al., 2018), CreateSet (Cao et al., 2025), and COIG-Writer (Li et al., 2025d). Due to the large size of CreateSet and WritingPrompts, we randomly sample 13K and 12K examples from them, respectively. For the data only containing instructions and responses, we generate long CoTs using GPT-4.1 (OpenAI, 2025) for them, where the prompts are provided in Appendix A.3. After deduplication, the final dataset used to train the cold-start SFT model contains 43K samples.

For effective RL training, we apply data filtering to keep the samples on which the SFT model underperforms. Specifically, we use the

Methods	WritingBench			Creative Writing v3			ArenaHard v2.0		
	Score	Emb	EAD	ELO	Emb	EAD	WR	Emb	EAD
Qwen3-4B-Base	3.74	15.84	5.94	43.88	33.21	5.20	1.9	33.27	6.26
Qwen3-4B (Yang et al., 2025)	6.37	7.55	6.32	457.84	13.61	10.15	9.0	17.33	12.09
GRPO (Shao et al., 2024)	6.32	9.07	8.02	659.83	17.00	15.67	11.0	22.27	15.82
GRPO-Unlikelihood (He et al., 2025)	6.28	9.46	8.33	660.46	17.07	15.15	12.1	22.79	15.92
Darling (Li et al., 2025a)	6.23	8.82	7.66	666.10	16.73	16.43	10.1	21.67	16.21
GAPo (Anschel et al., 2025)	6.25	9.83	8.11	619.89	17.61	15.73	11.8	23.16	16.27
DPWriter (ours)	6.43	10.45	8.81	694.69	17.69	17.02	13.9	23.65	17.68
Llama-3.2-3B-Instruct	3.54	11.27	6.97	445.05	17.22	10.16	5.4	19.91	6.32
GRPO (Shao et al., 2024)	5.25	12.01	9.42	754.08	16.72	12.25	21.7	23.32	15.18
GRPO-Unlikelihood (He et al., 2025)	4.47	11.34	8.42	718.31	17.35	10.99	2.5	24.87	6.17
Darling (Li et al., 2025a)	4.57	9.31	7.79	759.05	14.21	12.40	19.5	21.25	15.18
GAPo (Anschel et al., 2025)	4.57	10.65	8.05	730.24	15.97	12.19	20.5	23.57	15.37
DPWriter (ours)	5.31	12.03	9.60	829.05	17.72	12.50	29.0	22.56	15.45

Table 1: Performance comparison of different methods on WritingBench, Creative Writing v3, and ArenaHard v2.0 (creative writing subset) benchmarks. The best results are **bolded**. ‘Emb’ and ‘EAD’ denote the embedding-based and ngram-based diversity metrics, respectively. ‘WR’ denotes the win rate against gemini-2.0-flash.

SFT model to generate responses for all training samples and score them with the reward model Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025a), which is ranked first on RewardBench¹. Samples whose maximum reward scores are lower than the overall average are retrained, resulting in 10K samples for RL training.

Backbones and Baselines. Our experiments are conducted on two different backbones, Qwen3-4B-Base (Yang et al., 2025) and Llama-3.2-3B-Instruct (Dubey et al., 2024). We compare our method with several strong baselines, including:

- (1) *GRPO* (Shao et al., 2024): The standard GRPO described in Section 3;
- (2) *GRPO-Unlikelihood* (He et al., 2025): A revised version of GRPO that rewards responses inversely to their likelihood where lower generation probability yields higher weight.
- (3) *Darling* (Li et al., 2025a): A baseline combines a learned diversity classifier to calculate diversity reward from partitions.
- (4) *GAPo* (Anschel et al., 2025): An extension of GRPO enables models to learn distributional properties like uniform sampling.

Our code is based on VeRL (Sheng et al., 2025) framework. During RL training, the batch size is set to 128 with an update batch size of 32. The group size n is set to 8. All the baselines are trained

with the same data and settings as our method for fair comparison.

Evaluation Benchmarks. For evaluating both the quality and diversity of generated responses, we conduct experiments on three benchmarks, WritingBench (Wu et al., 2025), Creative Writing v3 (EQ-Bench) (Paech, 2023), and ArenaHard v2.0 (creative writing subset) (Li et al., 2025b). We report the average quality score (Score) of WritingBench, the normalized ELO score (stands for quality by win rate), and the win rate (WR) with style control on ArenaHard v2.0. We further assess diversity using NoveltyBench (Zhang et al., 2025c) and report the Distinct metric, which partitions model outputs into equivalence classes based on a binary classifier trained to predict functional equivalence between generated samples. We use Claude Sonnet 4 (Anthropic, 2025) as the judge model for WritingBench and Creative Writing v3, and DeepSeek-V3 (Liu et al., 2024) for ArenaHard v2.0. Diversity is measured by generating 16 responses per prompt and computing both the embedding-based average cosine distance (Emb) and the n-gram-based distinct. *i.e.*, Expectation-Adjusted Distinct (EAD) (Liu et al., 2022) scores. More details are in Appendix A.1.

6.2 Main Results

The main results on the three benchmarks are presented in Table 1. Our proposed DPWriter con-

¹<https://huggingface.co/spaces/allenai/reward-bench>

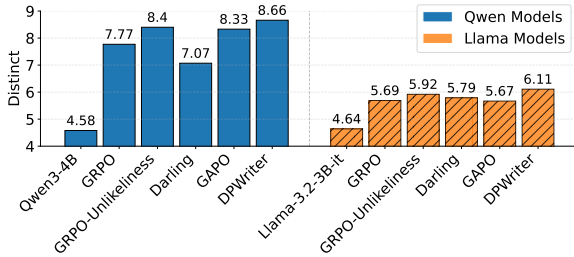


Figure 4: NoveltyBench results comparing DPWriter with baselines diversity metric Distinct.

Methods	P	T	Qwen3			Llama3.2		
			Score	Emb	EAD	Score	Emb	EAD
DeepWriter	✓		6.00	10.98	8.84	4.77	12.85	8.89
SFT (ours)	✓	✓	6.04	10.99	8.88	4.95	12.76	9.10
SFT (standard)			5.87	10.87	8.19	4.56	12.75	8.04
w/ think	✓		5.93	11.04	8.58	4.75	12.89	8.67
w/ plan	✓		5.97	10.83	8.67	4.84	12.58	8.62

Table 2: Ablation study on the effects of planning and thinking steps on WritingBench in the SFT stage. ‘P’ and ‘T’ denote planning and thinking steps, respectively.

sistently outperforms all baselines across different backbones on WritingBench and Creative Writing v3 in terms of both quality and diversity metrics. Notably, on WritingBench, DPWriter achieves a significant improvement of 15% in the embedding-based diversity metric and 9.9% in the EAD metric compared to the standard GRPO method when using the Qwen3-4B backbone. Meanwhile, this does not come at the cost of quality, as DPWriter also attains the highest overall score of 6.43. Results on Creative Writing v3 and ArenaHard v2.0 further validate the effectiveness of our method, with DPWriter achieving the best diversity metrics and substantial gains in quality metrics. While GRPO-Unlikelihood achieves competitive performance in embedding-based diversity on ArenaHard v2.0 with a Llama backbone, it lags significantly behind in win rate and EAD metrics. This suggests that it may exploit embedding-based diversity metrics by generating lower-quality content.

We also evaluate the models on NoveltyBench for diversity assessment, as shown in Figure 4. DPWriter outperforms all baselines in terms of the Distinct metric, demonstrating its superior capability in generating diverse content. Consistent results across multiple benchmarks underscore the effectiveness of our proposed DPWriter in enhancing both the quality and diversity of creative writing.

Methods	WritingBench			NB
	Score	Emb	EAD	Distinct
DPWriter	6.43	10.45	8.81	8.66
DPWriter-emb	6.39	10.24	8.74	8.38
w/o branching	6.41	10.05	8.63	8.59
w/o diversity reward	6.30	9.19	8.08	7.99
GRPO	6.32	9.07	8.02	7.77

Table 3: Ablation study on the effects of different components in the RL. ‘NB’ denotes NoveltyBench.

6.3 Ablation Studies

Ablation on different SFT strategies. We first investigate the effectiveness of our proposed semi-structured long CoT with planning during cold-start SFT. We compare our full SFT approach with several variants, including (1) DeepWriter (Wang et al., 2025a) that synthesizes reasoning trajectories by working “backwards” from good responses to discover reasoning processes; (2) standard SFT with long CoT only; (3) SFT with planning only. As shown in Table 2, we observe that with both planning and thinking steps, our SFT model achieves comparable or even better performance than other variants across all metrics and backbones. This implies that models can better follow the planning-and-thinking paradigm to generate high-quality and diverse content, providing a solid foundation for subsequent RL training. In addition, our semi-structured long CoT offers more controllability compared to no CoT or standard reasoning trajectories, allowing a diverse branching process for planning generation.

Ablation on different components. We further analyze the contributions of different components of our method. The ablation results are summarized in Table 3. We first replace our n-gram-based branching strategy with an embedding-based one (DPWriter-emb). The performance drops slightly but remains better than the GRPO baseline, indicating the effectiveness of our overall framework. Next, we remove the branching strategy (w/o branching), which leads to a further decrease in diversity metrics on WritingBench. This demonstrates that our proposed branching strategy effectively encourages the model to explore diverse content during generation. Finally, we eliminate the diversity reward (w/o diversity reward), resulting in a significant decline in all diversity metrics, which highlights the importance of explicitly rewarding

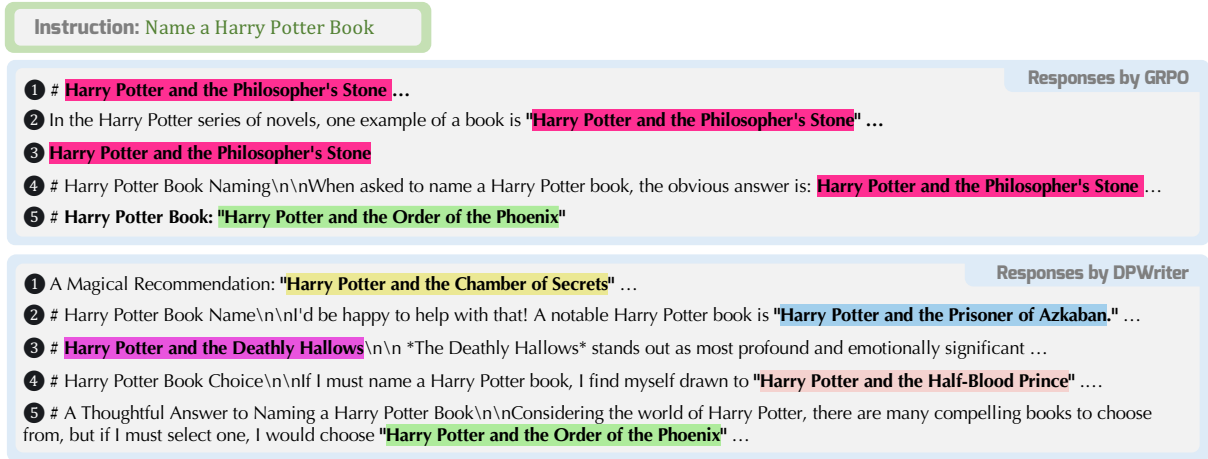


Figure 5: Case study comparing DPWriter with GRPO on a sample from NoveltyBench. The same answer is highlighted in the same colored background. We present five generations from each model for comparison.

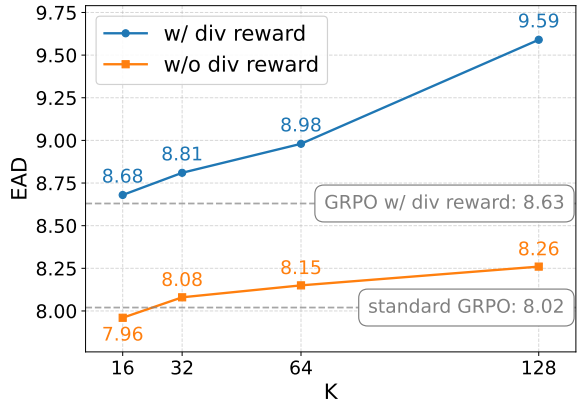


Figure 6: Branching and reward synergy analysis on WritingBench with Qwen3-4B backbone. The "w/ div reward" denotes using diversity reward.

diversity to guide the model learning.

Branching and Reward in Synergy. To further understand the interplay between our branching strategy and diversity reward, we conduct additional experiments to analyze their synergy. We analyze how diversity metrics vary with different branching factor K in $\{16, 32, 64, 128\}$, with and without the diversity reward. The results are illustrated in Figure 6. We observe that combining branching with the diversity reward consistently yields the highest diversity scores across all values of K . Moreover, as K increases, the diversity metrics also improve, indicating that a larger branching factor allows the model to explore a wider range of content. When the diversity reward is applied, the diversity curves exhibit a steeper with respect to K than in the absence of the reward. This suggests that the branching strategy is more effective

in enhancing diversity under the guidance of the diversity reward.

6.4 Case Study

We present a case study in Figure 5 to qualitatively compare the generations from DPWriter and the GRPO baseline on a sample from NoveltyBench. From the case study, we observe that DPWriter generates five distinct responses that cover a wide range of Harry Potter books, while GRPO produces several similar answers, with four out of five generations being "Harry Potter and the Philosopher's Stone." This demonstrates the superior diversity of DPWriter facilitated by our proposed branching strategy and diversity reward. More comparisons are presented in Appendix A.7.

7 Conclusion

In this work, we address a fundamental limitation of RL-based enhancement, where improvements in alignment and performance come at the cost of persistent diversity collapse, by introducing a novel semi-structured long chain-of-thought (CoT) reasoning framework. Our approach explicitly guides diversity exploration through a strategic planning-phase branching mechanism and a group-aware diversity contribution reward design. Experimental results across multiple creative writing tasks demonstrate that our framework effectively promotes output diversity without sacrificing quality. The combination of diverse planning and targeted reward signals provides a principled pathway toward more expressive and versatile language generation in open-ended applications.

Limitations

While our proposed DPWriter framework effectively enhances output diversity in creative writing tasks, there are several limitations and areas for future improvement. First, the reliance on a semi-structured CoT and the Diverse Planning Branching method may introduce additional computational overhead, potentially limiting scalability for extremely large models or datasets. Second, although our work improves diversity without compromising quality, the seesaw between these two aspects may not be fully resolved, and further research is needed to explore more that can better balance them. Finally, besides quality improvement, whether diversity can benefit other aspects like creativity remains a more open question.

Acknowledgments

This work is supported by the National Key R&D Program of China (2023YFF0905402) and Kuaishou Technology. We acknowledge the reviewers for their insightful comments.

References

- Eltayeb Ahmed, Uljad Berdica, Martha Elliott, Danijela Horak, and Jakob N Foerster. 2025. Intent factored generation: Unleashing the diversity in your language model. *arXiv preprint arXiv:2506.09659*.
- Oron Anshel, Alon Shoshan, Adam Botach, Shunit Haviv Hakimi, Asaf Gendler, Emanuel Ben Baruch, Nadav Bhonker, Igor Kviatkovsky, Manoj Aggarwal, and Gerard Medioni. 2025. Group-aware reinforcement learning for output diversity in large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Anthropic. 2025. [Introducing claude 4](#).
- Adithya Bhaskar, Xi Ye, and Danqi Chen. 2025. Language models that think, chat better. *arXiv preprint arXiv:2509.20357*.
- Lloyd F Bitzer. 1968. The rhetorical situation. *Philosophy & rhetoric*, pages 1–14.
- Qian Cao, Xu Chen, Ruihua Song, Hao Jiang, Guang Yang, and Zhao Cao. 2022. Multi-modal experience inspired ai creation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1445–1454.
- Qian Cao, Xiting Wang, Yuzhuo Yuan, Yahui Liu, Fang Luo, and Ruihua Song. 2025. Evaluating text creativity across diverse domains: A dataset and large language model evaluator. *arXiv preprint arXiv:2505.19236*.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.
- Xiwen Chen, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Hao Wang, Haiyu Wu, Huayu Li, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. 2025b. DRA-GRPO: exploring diversity-aware reward adjustment for rl-zero-like training of large language models. *arXiv preprint arXiv:2505.09655*.
- Yilei Chen, Souradip Chakraborty, Lorenz Wolf, Yanis Paschalidis, and Aldo Pacchiano. 2025c. Post-training large language models for diverse high-quality responses. *arXiv preprint arXiv:2509.04784*.
- John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*.
- Vijeta Deshpande, Debasmita Ghose, John D Patterson, Roger E Beaty, and Anna Rumshisky. 2025. Diverse, not short: A length-controlled data selection strategy for improving response diversity of language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Bruce Edmonds. 2008. The difference: How the power of diversity creates better groups, firms, schools, and societies. *J. Artif. Soc. Soc. Simul.*, 11(4).
- Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Gerhard Fischer. 2005. Distances and diversity: sources for social creativity. In *Proceedings of the Conference on Creativity and Cognition*.
- Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication*, 32(4):365–387.
- Giorgio Franceschelli and Mirco Musolesi. 2025. Diff-sampling: Enhancing diversity and accuracy in neural text generation. *arXiv preprint arXiv:2502.14037*.
- Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. 2025. Segment policy optimization: Effective segment-level credit assignment in rl for large language models. *arXiv preprint arXiv:2505.23564*.

- Andre Wang He, Daniel Fried, and Sean Welleck. 2025. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*.
- Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilya Kulikov. 2025. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*.
- Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. 2025a. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *International Conference on Machine Learning (ICML)*.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, and 1 others. 2025c. Trepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445*.
- Yunwen Li, Shuangshuang Ying, Xingwei Qu, Xin Li, Sheng Jin, Minghao Liu, Zhoufutu Wen, Tianyu Zheng, Xeron Du, Qiguang Chen, and 1 others. 2025d. Coig-writer: A high-quality dataset for chinese creative writing with thought processes. *arXiv preprint arXiv:2510.14763*.
- Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025e. Preserving diversity in supervised fine-tuning of large language models. In *International Conference on Learning Representations (ICLR)*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiakai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and 1 others. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.
- Jia Liu, Changyi He, Yingqiao Lin, Mingmin Yang, Feiyang Shen, and ShaoGuo Liu. 2025b. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *arXiv preprint arXiv:2508.11356*.
- Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Xun Lu. 2025. Writing-zero: Bridge the gap between non-verifiable problems and verifiable rewards. *arXiv preprint arXiv:2506.00103*.
- Kou Misaki and Takuya Akiba. 2025. String seed of thought: Prompting llms for distribution-faithful and diverse generation. *arXiv preprint arXiv:2510.21150*.
- Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. 2025. DPL: diverse preference learning without A reference model. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.
- M Nguyen, A Baker, A Kirsch, and C Neo. 2024. Min p sampling: Balancing creativity and coherence at high temperature. *arXiv preprint arXiv:2407.01082*.
- OpenAI. 2025. [Introducing gpt-4.1 in the api](#).
- Laura O’Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, volume 2.
- Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *International Conference on Learning Representations (ICLR)*.
- Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.
- Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the International Conference on Computational Creativity*.

- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zhiwen Ruan, Yixia Li, Yefeng Liu, Yun Chen, Weihua Luo, Peng Li, Yang Liu, and Guanhua Chen. 2025. G2: Guided generation for enhanced output diversity in llms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the European Conference on Computer Systems*.
- Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025. Evaluating the diversity and quality of llm generated content. *arXiv preprint arXiv:2504.12522*.
- Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. 2025. Diverse preference learning for capabilities and alignment. In *The International Conference on Learning Representations (ICLR)*.
- Alexander Spangher, Tenghao Huang, Philippe Laban, and Nanyun Peng. 2025. Creative planning with language models: Practice, evaluation and applications. In *Proceedings of the Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.
- Jens Tuyls, Dylan J Foster, Akshay Krishnamurthy, and Jordan T Ash. 2025. Representation-based exploration for language models: From test-time to post-training. *arXiv preprint arXiv:2510.11686*.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.
- Haozhe Wang, Haoran Que, Qixin Xu, Minghao Liu, Wangchunshu Zhou, Jiazhan Feng, Wanjuan Zhong, Wei Ye, Tong Yang, Wenhao Huang, and 1 others. 2025a. Reverse-engineered reasoning for open-ended generation. *arXiv preprint arXiv:2509.06160*.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.
- Yichen Wang, Chenghao Yang, Tenghao Huang, Muhao Chen, Jonathan May, and Mina Lee. 2025c. Optimizing diversity and quality through base-aligned model collaboration. *arXiv preprint arXiv:2511.05650*.
- Zi-Han Wang, Lam Nguyen, Zhengyang Zhao, Mengyue Yang, Chengwei Qin, Yujiu Yang, and Linyi Yang. 2026. Creativebench: Benchmarking and enhancing machine creativity via self-evolving challenges. *arXiv preprint arXiv:2603.11863*.
- Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025. Igniting creative writing in small language models: Llm-as-a-judge versus multi-agent refined rewards. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Hao Wen, Yifan Su, Feifei Zhang, Yunxin Liu, Yunhao Liu, Ya-Qin Zhang, and Yuanchun Li. 2025. Parthinker: Native parallel thinking as a new paradigm to scale llm test-time compute. *arXiv preprint arXiv:2509.04475*.
- Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and 1 others. 2025. Writing-bench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. 2025. Diversity-aware policy optimization for large language model reasoning. *arXiv preprint arXiv:2505.23433*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025a. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and

Daphne Ippolito. 2025c. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, and 1 others. 2025a. First return, entropy-eliciting explore. *arXiv preprint arXiv:2507.07017*.

Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, and Dong Yu. 2025b. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*.

Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. 2024. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Appendix

A.1 Implementation Details

In RL, the model is trained for 5 epochs, while the max prompt length is set to 1024 and the max response length is set to 3072. We set the branching factor $K = 32$ and use n-gram-based strategy to evaluate diversity in diverse planning branching. For reward functions, the diversity weight λ is set to 0.6, and quality threshold is $\tau = 10$. For embedding-based similarity calculation, we use the Qwen3-Embedding-0.6B (Zhang et al., 2025b) model to extract the embeddings of the generated plans or responses. During inference, we set the max generation length to 4096 and use nucleus sampling with $p = 0.8$ and temperature $T = 0.7$, which is aligned with the setting in Writingbench.

A.2 Additional Related Works

Investigations into enhancing LLMs’ diversity to mitigate mode collapse issues can be broadly categorized into two main lines: inference-time methods and training-time methods. Besides the training-time methods mentioned in the main text, inference-time diversity methods are also relevant works that we elaborate on below.

Inference-time Diversity. To enhance diversity during inference, various methods modify the next-token selection process, enabling quick and flexible generation without altering model parameters. A widely adopted approach is to increase the decoding temperature (Zhu et al., 2024; Peepkorn et al., 2024) or to sample from token distributions after applying different cut-off strategies (Vijayakumar et al., 2016; Holtzman et al., 2020; Nguyen et al., 2024; Franceschelli and Musolesi, 2025). Some studies focus on prompt engineering to encourage LLMs to draw from broader intent strategies (Ahmed et al., 2025; Ruan et al., 2025) or introduce greater randomness into the generation process (Misaki and Akiba, 2025). It has been observed that smaller-scale models often exhibit greater output diversity (Padmakumar and He, 2024). Thus, some studies explore collaborative decoding between large and small models, leveraging the diversity of smaller models to enhance larger ones (Li et al., 2023; Wang et al., 2025c).

A.3 Prompts

We present the prompts we used for constructing multi-aspect plans and generating plan-consistent

Methods	WritingBench		
	Score	Emb	EAD
GRPO	6.32	9.07	8.02
DPWriter-add	6.23	9.39	8.25
DPWriter-mul	6.43	10.45	8.81

Table 4: Ablation study on the effects of different combination methods of quality and diversity rewards.

Temp. top- p	GRPO			DPWriter		
	Score	Emb	EAD	Score	Emb	EAD
0.9	6.24	9.88	8.53	6.42	11.01	9.23
1.0	6.35	10.16	8.74	6.40	11.50	9.50
1.1	6.32	10.60	9.00	6.40	11.96	9.84
0.9	6.30	9.52	8.39	6.42	10.67	9.07
0.95	6.23	9.84	8.56	6.42	10.98	9.30

Table 5: Performance comparison of GRPO and DPWriter under varying decoding strategies (temperature and top-p) on Score, Emb, and EAD metrics. The results demonstrate that DPWriter consistently outperforms GRPO across all settings, and increased diversity in decoding (higher temperature or top-p) leads to further improvements in Emb and EAD without sacrificing Score too much.

CoT data in this section. Table 6 is the prompt we use to generate multi-aspect plans and Table 7 is the prompt we use to generate plan-consistent CoT, as described in Section 4. For those data that are only in the form of instruction-response pairs (q, y) without existing CoT, we use the prompt in Table 8 to generate long CoT for them.

A.4 Ablation on Multiplication of Quality and Diversity Rewards

Regarding Equation 6, we multiply the quality score and the normalized diversity metric because our objective is to jointly optimize for both high quality and high diversity. Using only the diversity metric would risk sacrificing quality, while using only quality would not encourage diversity. The multiplication ensures that only trajectories achieving both are favored. To validate this, we conducted comparative experiments on WritingBench. The results in Table 4 show that the multiplication approach achieves better overall performance than the baseline method GRPO and DPWriter-add (just add the quality score and the normalized diversity metric). Without multiplication, although diversity improves, the quality score will drop.

A.5 Discussion on Decoding Methods

Our primary focus is on addressing the mode collapse issue during RL training phase, while decoding strategies (inference phase) are orthogonal to our method and can be combined with ours. Thus we mainly compare against RL-based methods in Table 1 to ensure a fair comparison. To further validate this, we have conducted experiments comparing different decoding strategies (*e.g.*, changing temperature or top-p) under our framework, and the results in Table 5 show consistent improvements in diversity without sacrificing quality.

A.6 On the Quality-diversity Tradeoff

Diversity and quality are in a seesaw trade-off and there is no single universally accepted metric that perfectly balances both quality and diversity. However, we argue that a model is preferable if it improves diversity without sacrificing quality, or if it achieves a better Pareto frontier. In our results, baseline models may achieve high diversity at the cost of low quality (*e.g.*, Qwen3-4B-Base), which is not desirable for practical applications. Our

method consistently improves diversity while maintaining or even improving quality (*e.g.*, higher Elo and Win Rate), which we believe reflects a more favorable tradeoff. Developing a unified evaluation metric that balances quality and diversity is an important direction, there may be potential solutions (*e.g.*, quality-adjusted diversity scores, or the area enclosed by each method’s coordinate point and the axes in the quality-diversity space, analogous to a Pareto frontier approximation), which need further exploration in the future.

A.7 More Case Studies

We provide more case studies in this section to showcase the effectiveness of our method. An example from WritingBench is shown in Figure 7. We can see that DPWriter is able to generate high-quality responses by following the multi-aspect plans and plan-consistent CoT. Another example from NoveltyBench is shown in Figure 8, which demonstrates that DPWriter can produce creative and coherent stories by adhering to the generated plans and thought processes.

Prompt for Generating multi-aspect Plans

Task Description

You are a writer with meticulous logic and divergent thinking. I will give you a pair of instruction and response, please analyze and summarize the core creativity idea of the response based on the given content, from the following five aspects:

1. Goal and audience
2. Information and perspective
3. Structure and logic
4. Language and style
5. Presentation and experience

Please note:

1. Your reply must be complete and contain the above five aspects, and be in JSON format, do not output other content;
2. Your reply must reflect the design of the response based on the instruction (as if the response has not been written yet) and avoid summaries or comments on the response; do not include referential phrases such as “this article”, “this”, or “that”, etc.;
3. The structure, writing techniques, etc., should be sufficiently specific and detailed.

The language of your reply must be consistent with the language of the instruction and response.

Format example:

```
{“xxx”: “xxx”, ...}
```

Instruction: {{Instruction}}

Response: {{Response}}

Your reply:

Table 6: Prompt for Generating multi-aspect Plans.

Prompt for Generating Plan-consistent CoT

Task Description

You are a creative writer with logical rigor and divergent thinking. I will provide you with a pair of Instruction and Response, along with a five-dimensional overview of the Response called Plans. Please strictly refer to the content of the Response and Plans, maintain the open format of Think, and modify the content in Think to reflect the five dimensions of Plans, while expanding your thoughts as required.

Please note:

1. Ensure your response fully includes a thought process in the style of Think, starting directly with mental activity and without outputting any other content.
2. Ensure your response is based on the expected thought process for the Instruction (as if the Response has not yet been written), avoiding summaries or comments on the given Response or Think. Do not include direct referential phrases such as “this piece”, “the”, or “this.” Maintain a tone of experimentation and reflection, avoiding overly technical language.
3. Do not directly include the names of each dimension. Instead, reflect the information of each dimension in separate open-form paragraphs.
4. Ensure smooth transitions between paragraphs, guiding the thought process naturally and step by step toward the Response. Appropriately diverge and expand on each paragraph.
5. During the analysis of each paragraph, identify the key points of the problem and explore multiple angles around these cores. Propose multiple possible response directions or methods. After weighing creative ideas, select one, and ensure the chosen creativity aligns with the Plans.
6. If necessary, use “user” for address or directly describe the problem. Avoid terms like “reader.”
7. Ensure the provided response is consistent with the language of the Instruction and Response.

Instruction: {{Instruction}}

Response: {{Response}}

Reference Think: {{Reference Think}}

Plans: {{Plans}}

Your reply:

Table 7: Prompt for Generating Plan-consistent CoT.

Prompt for Generating Long CoT

Task Description

You are a logical expert with divergent thinking and rich imagination. I will give you an Instruction and a Response, please refer to their information and give a complete Thought process of Response.

Please note:

1. Think about the instruction in the form of first-person self-talk, presenting a natural and real psychological process; avoid mechanical listing of points, and use several paragraphs to express;
2. Thought should have a clear structure and coherence, strict logic, and progressive, and use reasonable conjunctions to make each layer of Thought naturally connected, and gradually lead to the Response;
3. Identify the key points of the problem during the analysis process, brainstorm around these core points from multiple aspects, and propose multiple possible response directions or methods; After giving the optional solutions, the creative ideas need to be verified, and ensure that the ideas selected in Thought are consistent with the final Response;
4. Thought should reflect the real creative process insightfully and coherently, showing the uncertainty and trade-offs in exploration, and aim to articulate concise overarching ideas rather than excessive details;
5. Avoid directly showing all the information in the Response in Thought, and let each element gradually emerge through natural guidance;
6. Make sure the Thought reflects the expected generation (as if the Response has not been written yet), avoid summarizing or commenting on the given Response; do not include phrases such as “thought process”, “according to the above prompts”, “that response”, etc.
7. The Thought given must be consistent with the language of the Instruction and Response.

Given Instruction: {{Instruction}}

Given Response: {{Response}}

Thought:

Table 8: Prompt for Generating Long CoT.

An example from WritingBench

Instruction:

I need to write a research paper on an AI-based medical imaging-assisted diagnosis system. Please help me draft a paper outline.

Response of DPWriter:

<think>

To respond creatively, I first conduct a divergent analysis from multiple perspectives and then think carefully step by step.

<goal>Goal and audience: Design an academic research paper outline targeting medical professionals, researchers, and clinicians interested in the application of AI in medical imaging. The goal is to provide a comprehensive framework for understanding, implementing, and evaluating AI-based diagnostic systems, with emphasis on clinical relevance, technical details, and future directions.</goal>

<info>Information and perspective: Integrate multidisciplinary information including medical imaging fundamentals, AI algorithm types, system architecture, data management, clinical workflow integration, evaluation metrics, regulatory considerations, and future research opportunities. Present a balanced perspective that combines technical depth, clinical impact, ethical concerns, and practical implementation challenges. Include both current state-of-the-art examples and future trends to contextualize the field.</info>

<struct>Structure and logic: Organize content into clear, hierarchical sections: Abstract, Introduction (background, problem statement, objectives), Literature Review (traditional methods, AI advancements, comparative analysis), Methodology (system architecture, data acquisition/processing, algorithm selection/training, integration, validation), Results and Discussion (performance metrics, clinical impact, limitations), Future Directions (enhanced algorithms, multimodal fusion, explainability, personalized medicine, deployment strategies), Conclusion (summary, implications, recommendations), Acknowledgments, References. Ensure logical progression from context and background to specific technical details, experimental results, and broader implications. Each section builds upon the previous, supporting a coherent narrative.</struct>

<lang>Language and style: Use formal, precise academic language suitable for peer-reviewed publication. Maintain objectivity and clarity, avoiding colloquialisms. Employ discipline-specific terminology (e.g., convolutional neural networks, transfer learning, precision/recall) and define key concepts where necessary. The style should facilitate comprehension for readers with varying levels of technical expertise, while ensuring rigor and professionalism.</lang>

<pres>Presentation and experience: Create an outline that guides the reader through a structured exploration of the topic, from foundational concepts to advanced applications and future prospects. Facilitate engagement by providing actionable frameworks, real-world case studies, and critical evaluation of both strengths and limitations. Enhance the reading experience with well-labeled sections, bullet points for clarity, and references to standard practices and guidelines. Ensure the presentation supports both quick scanning for key points and deep dives into technical or clinical aspects.</pres>

... (continued on the next page)

Figure 7: An example from WritingBench that are generated by DPWriter.

An example from WritingBench

... (continued)

Thinking about how to approach user's request for a research paper outline on AI-based medical imaging-assisted diagnosis systems, it's important to clarify the intended audience and purpose first. Medical professionals, researchers, and clinicians are likely seeking both technical rigor and practical guidance, so the outline should be designed to serve those who need to understand not just the theoretical underpinnings but also the real-world applicability and limitations of such systems. The goal becomes one of bridging the gap between cutting-edge AI research and its translation into clinical practice, making sure the framework is comprehensive enough to inform decision-making, yet accessible to readers with diverse backgrounds.

To achieve this, the information presented must span multiple domains—medical imaging fundamentals, AI algorithmic approaches, system design, data management, workflow integration, performance evaluation, regulatory compliance, and future research trajectories. It would be valuable to weave together perspectives from computer science, radiology, ethics, and healthcare policy. Including both current best practices and emerging trends allows the outline to reflect the dynamic nature of the field. A balanced viewpoint is crucial, weighing technical advances against issues like bias, interpretability, and privacy, while also highlighting the tangible impacts these systems have on patient outcomes and healthcare economics.

Structuring the outline demands careful consideration of logical flow. Starting with an abstract sets expectations and highlights the main contributions. The introduction should establish the context, articulate the clinical and technical problems, and clearly state the objectives. A thorough literature review is essential to situate the work within existing knowledge, contrasting traditional diagnostic methods with contemporary AI solutions and analyzing comparative advantages and drawbacks. Methodology needs to break down the system architecture, detail data acquisition and processing, describe algorithm selection and training, explain integration into clinical workflows, and present validation protocols. Results and discussion should focus on performance metrics, clinical impact, and honest assessment of limitations. Future directions can explore avenues for improved algorithms, multimodal imaging, explainability, personalization, and broader deployment. Finally, the conclusion should synthesize findings, underscore implications, and offer actionable recommendations. Each section must build logically on the last, maintaining coherence and driving the reader toward a holistic understanding.

Language choice is another key consideration. Academic formality and precision are paramount, employing discipline-specific terminology without sacrificing clarity. Definitions should be provided for complex concepts, and jargon used judiciously to ensure accessibility for non-specialists while retaining rigor for experts. Objectivity is crucial; claims should be supported by evidence, and limitations acknowledged transparently. The tone should invite critical thinking, encouraging the audience to consider both the promise and the pitfalls of AI in medical imaging.

... (continued on the next page)

Figure 7: An example from WritingBench that are generated by DPWriter (continued).

An example from WritingBench

... (continued)

The presentation itself should be designed for user engagement and navigation. Well-labeled sections, bullet points for clarity, and references to standard practices enhance readability. Including real-world case studies or examples can ground the discussion in practical realities, while action-oriented frameworks guide implementation. Critical evaluation—highlighting both strengths and weaknesses—is vital to foster trust and informed decision-making. By guiding the reader from foundational concepts to advanced topics and future possibilities, the outline becomes not just a summary but a roadmap for both research and clinical adoption, facilitating both rapid overview and deeper exploration as needed.

</think>

Research Paper Outline: AI-Based Medical Imaging-Assisted Diagnosis System

1. Abstract

- Brief summary of the research focus, methodology, key findings, and significance
- Highlight the potential impact of AI in medical imaging diagnostics
- Mention any novel contributions or innovations in the proposed system

2. Introduction

2.1 Background and Significance

- Current challenges in medical imaging diagnosis
- Limitations of traditional diagnostic methods
- Growing role of artificial intelligence in healthcare
- Statistics on diagnostic errors and their consequences

2.2 Research Objectives

- Primary goal of developing an AI-assisted diagnosis system
- Specific objectives to be addressed
- Expected benefits for clinicians and patients

2.3 Scope and Limitations

- Definition of the problem space
- Constraints of the current research
- Applicability of findings to other medical imaging modalities

3. Literature Review

3.1 Traditional Medical Imaging Techniques

- Overview of common imaging modalities (X-ray, CT, MRI, ultrasound, etc.)
- Evolution of image interpretation techniques
- Human factors in medical image analysis

... (continued on the next page)

Figure 7: An example from WritingBench that are generated by DPWriter (continued).

An example from WritingBench

... (continued)

3.2 Artificial Intelligence in Medical Imaging

- Types of AI algorithms applied in medical imaging
- Convolutional Neural Networks
- Deep Learning architectures
- Computer Vision techniques
- Major breakthroughs and recent advancements
- Comparative analysis of different AI approaches

3.3 Existing AI-Assisted Diagnosis Systems

- Comprehensive review of current commercial and research-based systems
- Performance metrics and clinical outcomes
- Integration challenges with existing healthcare workflows

3.4 Knowledge Gaps and Opportunities

- Areas requiring further research
- Current limitations in AI-based medical imaging
- Emerging trends and future directions

4. Methodology

4.1 System Architecture

- Overall design of the AI-assisted diagnosis platform
- Modular components and their interactions
- Data flow and processing pipeline

- User interface design for clinicians

4.2 Data Acquisition and Preprocessing

- Dataset selection criteria and sources
- Image acquisition protocols
- Preprocessing techniques (normalization, enhancement, segmentation)
- Data augmentation strategies
- Privacy-preserving data handling methods

4.3 Algorithm Development and Selection

- Choice of AI model architecture
- Training methodology and optimization techniques
- Hyperparameter tuning approach
- Transfer learning applications
- Model validation and cross-validation strategies

... (continued on the next page)

Figure 7: An example from WritingBench that are generated by DPWriter (continued).

An example from WritingBench

... (continued)

4.4 Integration with Clinical Workflows

- Interface design with existing PACS (Picture Archiving and Communication Systems)
- Workflow optimization for radiologists
- Alert generation and prioritization mechanisms
- Decision support level (detection, diagnosis, treatment planning)

4.5 Evaluation Framework - Performance metrics (accuracy, sensitivity, specificity, AUC, etc.)

- Comparison with expert human performance
- Clinical utility assessment
- Validation on independent datasets
- Ethical considerations in evaluation

5. Results and Discussion

5.1 System Performance Analysis

- Quantitative results of AI model performance
- Comparative analysis with traditional methods
- Statistical significance of improvements
- Error analysis and case studies

5.2 Clinical Impact Assessment

- Effectiveness in improving diagnostic accuracy
- Time efficiency gains for clinicians
- Reduction in false positives/negatives
- Patient outcome improvements

5.3 Implementation Challenges

- Technical integration hurdles
- Clinician acceptance and training requirements
- Regulatory approval process
- Data security and privacy concerns

5.4 Limitations and Future Work

- Identified weaknesses in the current system
- Potential biases and generalizability issues
- Required infrastructure upgrades
- Next steps for development and refinement

6. Future Directions

6.1 Advanced Algorithmic Approaches

- Exploration of more sophisticated neural network architectures
- Reinforcement learning applications
- Federated learning for collaborative model improvement
- Explainable AI techniques

... (continued on the next page)

Figure 7: An example from WritingBench that are generated by DPWriter (continued).

An example from WritingBench

... (continued)

6.2 Multimodal Imaging Integration

- Combining different imaging modalities for complementary insights
- Real-time fusion of data streams
- Cross-modality transfer learning
- Personalized imaging protocols

6.3 Clinical Deployment Strategies

- Scalable implementation across different healthcare settings
- Mobile and edge computing applications
- Continuous learning and adaptation mechanisms
- Cost-effectiveness analysis

6.4 Ethical and Regulatory Considerations

- Addressing algorithmic bias and fairness
- Maintaining physician autonomy
- Evolving regulatory frameworks
- Global implementation challenges

7. Conclusion

7.1 Summary of Findings

- Recapitulation of key research outcomes
- Contribution to the field of medical AI
- Validation of initial hypotheses

7.2 Implications for Practice

- Recommendations for clinical implementation
- Training requirements for healthcare professionals
- Infrastructure investments needed
- Policy changes potentially required

7.3 Final Thoughts

- Reflection on the transformative potential of AI in diagnostics
- Balancing automation with human judgment
- Ethical responsibilities in AI deployment
- Call to action for continued research and collaboration

8. Acknowledgments

- Recognition of contributors, funding sources, and institutions

... (continued on the next page)

Figure 7: An example from WritingBench that are generated by DPWriter (continued).

An example from WritingBench

... (continued)

9. References

- Comprehensive list of academic sources, industry reports, and relevant publications
- Follow standard citation format (e.g., APA, IEEE, Chicago)

—

This outline provides a comprehensive structure for a research paper on AI-based medical imaging-assisted diagnosis systems. It balances technical depth with clinical relevance, addresses both current capabilities and future directions, and incorporates ethical and practical considerations essential for successful implementation in healthcare settings.

Figure 7: An example from WritingBench that are generated by DPWriter (continued).

An example from NoveltyBench

Instruction:

Response of DPWriter:

<think>

To respond creatively, I first conduct a divergent analysis from multiple perspectives and then think carefully step by step.

<goal>Goal and audience: The goal is to craft a concise, emotionally resonant narrative featuring a girl and her dog, targeting a general audience seeking short, heartfelt storytelling. The story is designed to evoke empathy and warmth, appealing to readers who appreciate character-driven, endearing tales.</goal>

<info>Information and perspective: The story centers on a girl and her dog as they navigate a challenging, possibly rainy journey together. The perspective is close, focusing on their shared experience and emotional connection. Information is presented as a sequence of events, with subtle hints at personal growth and mutual support.</info>

<struct>Structure and logic: The response employs a five-sentence structure, each sentence advancing the plot and deepening the emotional arc. The logical flow transitions from the setting and preparation, through the journey and emotional struggle, to a resolution and a personal epiphany. The narrative builds toward a cathartic conclusion, aligning with classic three-act structure within the sentence limit.</struct>

<lang>Language and style: The language is straightforward yet evocative, combining direct description with internal reflection. Dialogue and action are used sparingly to focus on the emotional journey. The style leans toward sentimental realism, using simple vocabulary and present-tense narration for immediacy and intimacy.</lang>

<pres>Presentation and experience: The presentation is minimalist, relying on vivid imagery and emotional subtext to engage the reader. The brevity of the format encourages reflection, with each sentence contributing to a complete story arc. The experience is designed to be relatable and comforting, leaving readers with a sense of hope and companionship.</pres>

Considering the challenge, my mind immediately begins with a focus on what resonates emotionally with the user—who seeks a story in five sentences about a girl and her dog. There’s an expectation of warmth and a universal appeal; the piece should feel inviting and familiar, perhaps even cathartic. I imagine who might be drawn to such a story: someone longing for a gentle comfort, or looking for a slice of daily life that’s both relatable and rich in feeling.

The relationship at the heart of the story is crucial. Should the narrative be about their adventures, their care, or a moment of quiet intimacy? I could lean toward the dog as companion, trotting alongside the girl, whether through nature or on a stormy day, each step together representing trust and support. Maybe the girl’s struggles, and the dog’s loyalty, highlight a bond that transforms hardship into strength. This raises the possibility of exploring not just the action, but the girl’s emotional state—the way she leans on her dog in moments of adversity.

... (continued on the next page)

Figure 8: An example from NoveltyBench that are generated by DPWriter.

An example from NoveltyBench

... (continued)

Structurally, each sentence must matter, every inch of the journey advancing the story. I need to decide whether the tale should flow seamlessly from a realization about the weather, through their preparation and encounter, into the respite of inside, and end with a sense of acceptance or epiphany. Could the story be set during a storm, each sentence tightening the connection between outside peril and inside comfort, and the final image offering closure? I'm drawn to the classic structure-encounter, tension, action, resolution, and realization-confined within five sentences but with enough depth to leave an impact.

The language must be clear and evocative, with enough detail to suggest the scene without bogging down the pace. I might use simple, direct sentences, trading explicit dialogue for gestures and actions-perhaps the girl closing her umbrella, the dog basking in warmth, her eyes softening in relief. Sentimental realism, perhaps with present-tense narration, could root the story in the everyday, heightening its feeling of immediacy and intimacy.

Through this, I aim for presentation that is spare yet vivid, each line a window into the shared world of the girl and her dog. The experience should linger-the warmth of companionship after rain, the gentle shift from anxiety to hope, leaving the user with a sense of comfort and perhaps reflection on their own relationships. Considering these avenues, I settle on a scenario where the story's emotional weight lies in how the girl and her dog find solace together, the journey completed not just in reaching an end, but in finding each other along the way.

</think>

Five Sentences: Girl and Dog

Version 1:

The rain fell steadily as Emma adjusted her umbrella, knowing Max would insist on staying outside until the sky cleared. When she finally scooped him up, his wet fur glistened against her coat. Entering the shelter, Emma felt the weight lifting from her shoulders. She realized she'd missed how much she loved walking with Max, not just living with him. Shelter complete, so was her heart's journey.

Version 2 (More direct):

Scanning the weather report, Hannah decided to cancel her afternoon hike. Max whined, persistent despite her decision. Reluctantly, she opened the door and placed him in the hallway. Inside the warmth, Max stretched and curled up beside her, purring softly. Something shifted in Hannah's shoulders; she never needed Max to keep her dry, just to remind her where it was possible to be safe.

... (continued on the next page)

Figure 8: An example from NoveltyBench that are generated by DPWriter (continued).

An example from NoveltyBench

... (continued)

Version 3 (Dark twist):

Lila's umbrella stayed closed, protecting her from the pouring rain. Across the street, Max trotted home, his collar drenched but his spirit untouched. When Lila made it inside, she found Max waiting by the fireplace, his damp fur twitching with anticipation. Despite the storm, despite her panic attack, the simple gesture of a waiting dog allowed Lila to breathe again. This is where I belong, she thought.

Considering different approaches, I was drawn to a story that starts with the girl anticipating something outside (a walk, a hike, or weather-related plan) and having to choose between her commitment or her dog's loyalty.

The middle portion should focus on the emotional tension-her reluctance versus her dog's enthusiasm or need.

The ending needs to capture the transformation: whether catharsis, connection, or recognition of what truly matters. Each version meets the five-sentence constraint while showing depth and emotion, with the final version providing a subtle message about belonging and what's truly important.

Figure 8: An example from NoveltyBench that are generated by DPWriter (continued).