

# JOINTLY REINFORCING DIVERSITY AND QUALITY IN LANGUAGE MODEL GENERATIONS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Post-training of Language Models (LMs) often prioritizes accuracy and helpfulness at the expense of diversity. This creates a tension: while post-training improves response quality, it also sharpens output distributions and reduces the range of ideas, limiting the usefulness of LMs in creative and exploratory tasks such as brainstorming, storytelling, or problem solving. We address this challenge with Diversity-Aware Reinforcement Learning (DARLING), a framework that jointly optimizes for response quality and semantic diversity. At its core, DARLING introduces a learned partition function to measure diversity beyond surface-level lexical variations. This diversity signal is then combined with a quality reward during online reinforcement learning, encouraging models to generate outputs that are both high-quality and distinct. Experiments across multiple model families and sizes show that DARLING generalizes to two regimes: non-verifiable tasks (instruction following and creative writing) and verifiable tasks (competition math). In the first setting, DARLING consistently outperforms quality-only RL baselines on 5 benchmarks, producing outputs that are simultaneously of higher quality and novelty. In the second setting, it achieves both higher pass@1 (quality) and pass@k (diversity). Most strikingly, explicitly optimizing for diversity catalyzes exploration in online RL, which manifests itself as higher-quality responses.

## 1 INTRODUCTION

Diversity—the ability to generate varied ideas, outputs, or solutions—plays a critical role in driving innovation across real-world applications (Guilford, 1950; Lu et al., 2025). For example, in scientific discovery, diverse hypotheses or experimental outcomes enable researchers to explore a broader solution space, potentially uncovering novel insights and breakthroughs (Gruver et al., 2023; Romera-Paredes et al., 2024; Si et al., 2025). Similarly, in other tasks such as creative writing (Fan et al., 2018) and natural conversations (Li et al., 2016a) and diverse outputs are essential for innovation, which requires avoiding repetitive or predictable outcomes. In reinforcement learning (RL) and self-training loops of LMs, diversity is also crucial. Policies that produce diverse outputs enable thorough exploration of the action space, critical for discovering novel and effective strategies (Chen et al., 2025a; Cheng et al., 2025; Wu et al., 2025a; An et al., 2025).

However, recent developments in Language Models (LMs) have revealed a significant issue: post-training of LMs often results in overly sharpened output distributions (Huang et al., 2025; Li et al., 2025b), leading to a significant reduction of diversity among generated responses (Padmakumar & He, 2024; Shyputa et al., 2025), even sharing identical prefixes (Ji et al., 2025) or becoming near duplicates (Mahony et al., 2024; Zhang et al., 2024), reducing the overall informativeness of outputs (Lin et al., 2021; Kirk et al., 2024; West & Potts, 2025; Yang & Holtzman, 2025; Yun et al., 2025).

To address the loss of diversity during LM post-training, we propose Diversity-Aware Reinforcement Learning (DARLING), an online RL objective that (a) measures diversity at the *semantic* level via a learned classifier, and (b) fuses diversity and quality to condition gradient updates on “usefully different” trajectories. As illustrated in Figure 1, DARLING first partitions rollouts from a single user prompt into distinct clusters using a semantic classifier, capturing diversity beyond superficial lexical differences (§3.1). It then combines (multiplies) the diversity assessment with a quality reward, amplifying the advantage of log-probabilities for responses that are both high-quality and

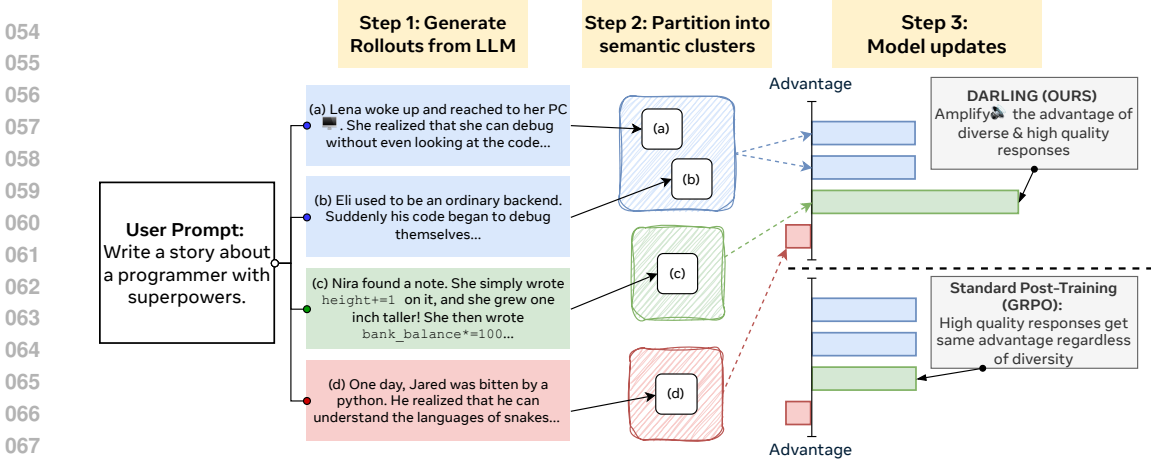


Figure 1: Diversity-Aware Reinforcement Learning (DARING): We first partition LLM generations into semantically equivalent clusters (represented by colors). While standard GRPO (Shao et al., 2024) increases probabilities based on response quality only, DARING amplifies the increase in probability of diverse and high-quality responses.

semantically diverse in online RL (§3.2). Unlike standard GRPO, which pays no heed to diversity, DARING explicitly encourages exploration across meaningfully distinct response clusters.

We validate DARING’s effectiveness and generalizability across both non-verifiable and verifiable tasks, over different language model families and sizes. Experimental results demonstrate that DARING preserves the original model’s diversity and achieves improved benchmark performance in both non-verifiable instruction-following and creative writing tasks, and verifiable math problems.

In summary, our contributions are three-fold:

- (1) We propose DARING, an RL framework that simultaneously optimizes quality and diversity, preventing diversity collapse during post-training.
- (2) We demonstrate that a learned semantic classifier can serve as a scalable and generalizable signal of diversity to integrate into online RL training.
- (3) We validate that explicitly optimizing for diversity in online RL for LMs promotes greater exploration, often leading to improvements in quality in both non-verifiable (creative writing) and verifiable (competition math) benchmarks.

## 2 NOTATIONS AND PRELIMINARIES

Let  $\mathcal{S}$  denote the set of natural language token sequences, a language model  $\pi(\cdot | x)$  takes a token sequence  $x \in \mathcal{S}$  as its input and outputs a probability distribution over  $\mathcal{S}$ . We denote the probability of a specific token sequence  $y \in \mathcal{S}$  as  $\pi(y | x)$  and denote the token at position  $t$  as  $y^t$ . Given a reward function  $r : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  which maps a pair of natural language instructions and responses  $x, y \in \mathcal{S}$  to a scalar value  $r(x, y) \in \mathbb{R}$ , LM post-training aims to solve the following KL constrained optimization problem:

$$\max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(\cdot | x)} \left[ r(x, y) - \beta \frac{\pi(y | x)}{\pi_{\text{ref}}(y | x)} \right] = \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(\cdot | x)} [r(x, y)] - \beta \mathbb{D}_{\text{KL}}(\pi || \pi_{\text{ref}}), \quad (1)$$

where  $\mathcal{D}$  is a dataset of prompts and  $\pi_{\text{ref}}$  is a reference model from which we do not want to deviate too much, usually implemented as the LM before the optimization process. We will omit the KL divergence term for simplicity. Group Relative Policy Optimization (GRPO; Shao et al. (2024)) optimizes (1) by maximizing the following objective:

$$\mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^n \sim \pi_{\text{act}}(\cdot | x)} \left[ \frac{1}{n} \sum_{i=1}^n \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left( \min(\text{IS}_{i,t} \cdot A_{i,t}, \text{clip}(\text{IS}_{i,t}, 1 - \epsilon, 1 + \epsilon) \cdot A_{i,t}) \right) \right] \quad (2)$$

where  $n$  is the number of responses per prompt, and

$$IS_{i,t} = \frac{\pi_{\theta}(y_i^t | y_i^{<t}, x_i)}{\pi_{act}(y_i^t | y_i^{<t}, x_i)}$$

is the importance sampling (Kloek & van Dijk, 1978) ratio between the current policy  $\pi_{\theta}$  and the actor  $\pi_{act}$  (the model used to generate  $y_i$ ), and

$$A_{i,t} = \frac{r(x, y_i) - \text{mean}_{j=1}^n(r(x, y_j))}{\text{std}_{j=1}^n(r(x, y_j))} \tag{3}$$

is the advantage of the response  $y_i$ , measuring how much better (or worse)  $y_i$  is over an average response, and  $\varepsilon$  is a hyperparameter preventing the importance sampling term  $IS_{i,t}$  from being too large or small. GRPO and its variants (Yu et al., 2025; Liu et al., 2025c; Hu, 2025) are widely adopted as some of the go-to algorithms for LM post-training (DeepSeek-AI et al., 2025; Liu et al., 2025a; Yang et al., 2025a) due to their simplicity and stability.

### 3 METHOD: DARLING

Figure 1 illustrates DARLING (Diversity Aware Reinforcement Learning). We first partition the responses using our developed classifier (§3.1) that captures semantic similarity, then combine diversity and quality signals in an RL framework to generate diverse and high-quality responses (§3.2).

#### 3.1 PARTITIONING THE RESPONSES INTO SEMANTIC EQUIVALENCE CLASSES

We begin by formally defining *diversity*, as used in our work: Given a pairwise distance metric  $d : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$  between two generations, and a group of  $n$  generations  $y_1, \dots, y_n$  we define the diversity of a generation  $y_i$  with respect to all other generations as the average pairwise distance between  $y_i$  and all other generations  $y_j$  ( $j \neq i$ ):

$$Div_d(y_i | y_1, \dots, y_n) = \frac{\sum_{j=1, j \neq i}^n d(y_i, y_j)}{n - 1} \tag{4}$$

We aim to incorporate a scalable metric of semantic diversity that captures deeper differences beyond surface-level variation into our training process. Following Zhang et al. (2025b), we train a binary classifier to determine whether two responses convey equivalent semantics:

$$\mathbf{classify}(y_i, y_j) = \mathbf{1}(y_i \text{ semantically equivalent to } y_j).$$

Details on how we trained this classifier can be found at Appendix A. Responses predicted as equivalent are clustered to form a partition of all responses into semantic clusters, where multiple members of the same cluster provide little additional value beyond a single representative. Our approach parallels semantic entropy methods (Kuhn et al., 2023), which group semantically equivalent responses.

We directly set diversity metric  $d = \mathbf{classify}(\cdot, \cdot)$ . Figure 2 illustrates an example of our diversity calculation from partitions. For a single prompt “Write a short joke about programming.”, responses in the left column (blue) are classified as semantically equivalent, both utilizing that the word “bug” has multiple meanings. The responses in the right column (purple and yellow) are distinct from the three other responses. For each of the individual responses in blue boxes, there are only two other responses that are distinct: purple and yellow. Therefore, using (4), we derive the diversity of both blue responses as 2/3. Similarly, the yellow and purple responses have diversity 3/3 because they are distinct from all other responses.

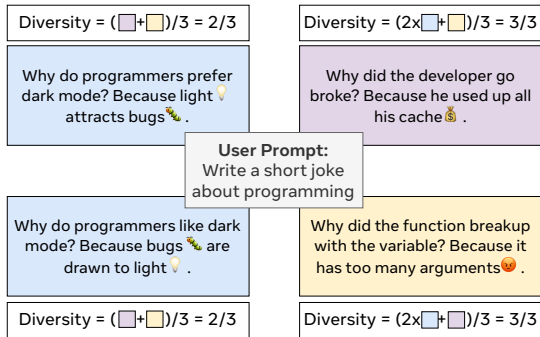


Figure 2: Example of partitioning a group of responses into semantically equivalent subgroups and evaluating diversity. Diversity is calculated as the normalized count of responses that is distinct from a given response.

### 3.2 DARLING: DIVERSITY AWARE REINFORCEMENT LEARNING

Given a diversity function  $\text{Div}_d$  and a reward function  $r$ , we define the diversity-aware reward as:

$$r_{\text{darling}}(x, y_i | y_1, \dots, y_n) := r(x, y_i) \times \text{Norm}(\text{Div}_d(y_i | y_1, \dots, y_n)), \quad (5)$$

where  $\text{Norm}(\cdot)$  normalizes diversity values to be between 0 and 1.

We choose to multiply the two reward scores instead of adding them. While simply adding the quality and diversity rewards is an alternative approach, this method poses challenges due to the differing scales of the two rewards. Naively summing the two rewards can lead the model to prioritize one reward over the other. Ablation studies of varying methods of fusing reward scores are in §6.1.

DARLING plugs (5) into (1), which amplifies the effective reward  $r_{\text{darling}}$  of high reward responses that are diverse from others. Motivated by prior work (Liu et al., 2025c; Yu et al., 2025), we also make the following modifications: changing sequence-level loss averaging to token-level averaging in (2) as the former exhibits bias towards longer sequences, and removing normalization by standard deviation in (3) since it amplifies the noise in dense rewards. We leave detailed ablations on the effect of advantage normalization to Appendix B.

The overall loss function for DARLING is thus defined as:

$$\mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^n \sim \pi_{\text{act}}(\cdot | x)} \left[ \frac{1}{\sum_{i=1}^n |y_i|} \sum_{t=1}^{|y_i|} \left( \min(\text{IS}_{i,t} \cdot A_{i,t}^{\text{darling}}, \text{clip}(\text{IS}_{i,t}, 1 - \varepsilon, 1 + \varepsilon) \cdot A_{i,t}^{\text{darling}}) \right) \right], \quad (6)$$

where we use the diversity-aware reward  $r_{\text{darling}}$  as the effective reward:

$$A_{i,t}^{\text{darling}} = r_{\text{darling}}(x, y_i | y_1, \dots, y_n) - \text{mean}_{j=1}^n (r_{\text{darling}}(x, y_j | y_1, \dots, y_n)).$$

Compared to standard GRPO, our main modification is that we multiply a normalized diversity reward  $\text{Norm}(\text{Div}_d(y_i | y_1, \dots, y_n))$  by the quality reward  $r(x, y)$  to promote high-quality and diverse rewards during training. This amplifies the increase in the log-likelihood of responses that are both high quality and diverse — jointly reinforcing diversity and quality.

## 4 DARLING ON NON-VERIFIABLE TASKS

We first show the effectiveness of DARLING on general non-verifiable instruction-following tasks. We describe our setup in §4.1. We show and analyze our results in §4.2.

### 4.1 SETUP

**Models and Baselines** We use two models of different scales: Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Llama Team, 2024) as reference models and perform training on a randomly sampled subset of 10k prompts from WildChat (Zhao et al., 2024), the same setup as in Lanchantin et al. (2025a). We use Nexusflow/Athene-RM-8B (Frick et al., 2024) as the reward function in (5) for quality. We use a batch size of {32 (8B), 64 (70B)} prompts and 8 rollouts per prompt. Other training hyperparameters can be found in Appendix C.

We compare our method against the following baselines:

- GRPO (Shao et al., 2024): The method described in §2, with token-level mean aggregation;
- DivPO (Lanchantin et al., 2025a): a DPO-based (Rafailov et al., 2024) optimization method that selects the *most* diverse response among the high-quality ones as the chosen response and the *least* diverse response as the rejected response;
- GRPO-Unlikelihood (He et al., 2025a): a revised GRPO algorithm that assigns higher rewards to low-likelihood responses to encourage greater diversity.

We implement DARLING using the verl codebase (Sheng et al., 2024). The original classifier of Zhang et al. (2025b) was limited to a context length of 512 tokens<sup>1</sup>. In our work, we extend their

<sup>1</sup><https://github.com/novelty-bench/novelty-bench/blob/main/src/partition.py#L69>

method by training a classifier with an 8192-token context window, using the same human-annotated data. Details of this training procedure are in Appendix A.1.

**Evaluation Benchmarks and Metrics** For evaluating response quality, we employ standard benchmarks: AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024), ArenaHard v0.1/2.0 (Li et al., 2025a), and EQ-Bench (Creative Writing) (Paeck, 2023). We report the length-controlled win rate (LCWR) for AlpacaEval 2.0 and the win rate with style control (markdown, length) for ArenaHard on the creative writing prompts. We report the normalized ELO score for EQ-Bench. For both AlpacaEval and ArenaHard, we follow Lanchantin et al. (2025b) and use GPT-4o (OpenAI et al., 2024) as the judge. For EQ-bench, we use Claude 3.7 Sonnet (Anthropic, 2024) as the judge. To evaluate diversity, we use NoveltyBench (Zhang et al., 2025b) and report the number of *semantically* distinct generations (Dist) and the average number of distinct 4grams (Dist-4) normalized by sequence length. We provide detailed descriptions of the benchmarks in Appendix E.

## 4.2 EXPERIMENTAL RESULTS

	AE 2.0*	AH v2.0*	AH v0.1*	EQ-Bench	NoveltyBench	
	LCWR (%)	WR (%)	WR (%)	ELO	Dist (#)	Dist-4 (%)
Llama-3.1-8B-Inst	31.9	7.1	30.9	636	5.28ch	93.9
GRPO	48.7	61.1	45.5	659	2.08	92.8
DivPO	43.5	54.4	39.7	639	4.34	94.1
GRPO-Unlikelihood	45.6	59.5	46.2	724	3.53	93.2
<b>DARLING</b>	<b>55.2</b>	<b>68.8</b>	<b>63.7</b>	<b>905</b>	<b>5.49</b>	<b>96.0</b>
Llama-3.3-70B-Inst	44.6	17.7	64.9	737	2.95	91.7
GRPO	73.3	89.7	79.2	1261	2.31	94.6
DivPO	58.0	59.1	72.4	1041	3.67	94.9
GRPO-Unlikelihood	69.5	84.2	76.4	1346	3.15	95.2
<b>DARLING</b>	<b>80.4</b>	<b>91.2</b>	<b>85.7</b>	<b>1531</b>	<b>4.26</b>	<b>95.3</b>

Table 1: Non-verifiable Task Evaluations. For each method, we train a single model on 10,000 WildChat prompts. We evaluate the models on AE (AlpacaEval 2.0 Length-Controlled Win Rate), AH v2.0/v0.1 (ArenaHard, creative writing subset), EQ-Bench (ELO), and NoveltyBench. \* indicates we used GPT-4o as the judge. All metrics are the higher the better. We find that models trained with DARLING achieve the best quality measured by both AlpacaEval/ArenaHard win rates and EQ-Bench ELO, and simultaneously are the most diverse, as measured by NoveltyBench.

**DARLING achieves both the best quality and diversity across all benchmarks.** Table 1 shows our main results: we observe that DARLING is able to jointly optimize for both quality and diversity. Specifically, DARLING results in the best quality scores (AlpacaEval and ArenaHard win rates) across our baselines, while also achieving the best diversity in both semantic level (Distinct) and lexical level (Distinct-4), showcasing the effectiveness of our method. Moreover, although we did not explicitly train on creative writing prompts, DARLING achieves the best ELO score compared to all baselines in EQ-Bench (creative writing), demonstrating the effectiveness of improving diversity on creative tasks. We show generation examples and perform qualitative analysis of models trained with DARLING in Appendix D.

**DARLING improves the Pareto front between quality and diversity by varying sampling temperature.** We further investigate the effect of sampling temperature on the quality-diversity Pareto front. We vary the sampling temperature ( $T = \{0.2, 0.4, 0.6, 0.8, 1.2\}$ ) of two models (Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct), being trained with GRPO and DARLING. Figure 3 shows that DARLING (blue) exhibits both better quality and better diversity than both the baseline (green) and GRPO (orange) at both scales, pushing forward the pareto-front of the “quality-diversity tradeoff” (Zhang et al., 2021; Padmakumar et al., 2025).

270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323

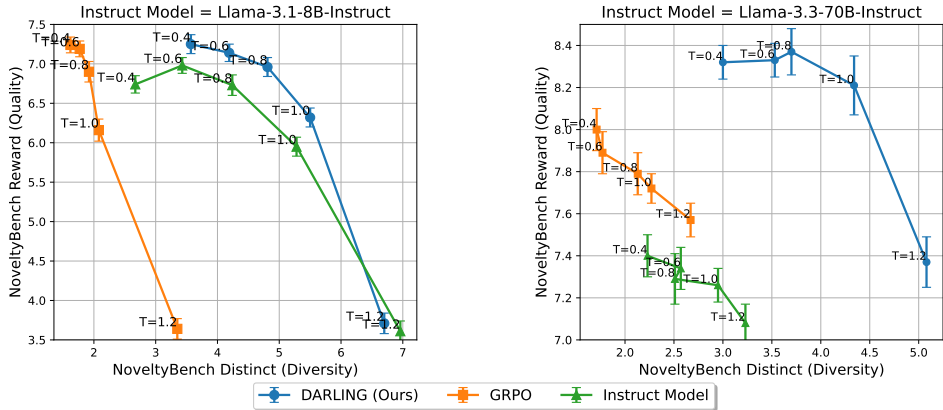


Figure 3: The quality-diversity tradeoff when using different sampling temperatures ( $T$ ) for models (at 8B and 70B scales) trained with standard GRPO and DARLING.  $X$ -axis: Distinct metric in NoveltyBench;  $Y$ -axis: Reward score used in NoveltyBench measuring quality of responses. DARLING (blue) simultaneously achieves better quality ( $y$ -axis) and diversity ( $x$ -axis) as demonstrated by the improved Pareto fronts on both the 8B and 70B scale.

### 5 DARLING ON VERIFIABLE TASKS

In this section we present experimental results of DARLING on verifiable math problems. We describe our setup in §5.1 and show our results in §5.2.

#### 5.1 SETUP

**Models, Baselines, and Benchmarks** We train models on top of Qwen3-4B-Base and Qwen3-14B-Base (Yang et al., 2025a) using the DeepscaleR dataset (Luo et al., 2025), where we first filter out questions that are unanswerable due to missing figures, and then subsample 10,000 examples. We compare our method against GRPO (Shao et al., 2024) on four competition math benchmarks: AIME25 (Art of Problem Solving, 2025), OlympiadBench (He et al., 2024), HMMT 2025, and Brumo 2025 (Balunović et al., 2025). We deliberately choose these benchmarks because they were released concurrently with Qwen3, preventing the effect of potential data contamination. We use the Hugging Face Math-Verify<sup>2</sup> library to automatically check the correctness of model answers, assigning a binary reward of  $r = 1$  for correct and  $r = 0$  for incorrect solutions. Following prior work (Yue et al., 2025), we report pass@1 scores as a measure of quality and pass@ $k$  scores as a measure of diversity. To evaluate pass@ $k$  up to  $k = 128$ , we sample  $n = 256$  responses for each prompt, and we average the performance of the 256 examples for calculating pass@1 to account for the variance introduced by the relatively small sizes of these benchmarks. We use the method in Chen et al. (2021) for an unbiased estimate of pass@ $k$  from  $n = 256$  examples:

$$\text{pass}@k := \mathbb{E} \left[ 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \right], \tag{7}$$

where  $c$  is the number of correct generations. With increased diversity, we expect to see an improved performance of pass@ $k$  as the model is more likely to hit the correct answer when it generates more diverse responses. Additional hyperparameters can be found at Appendix C.

**Training an equivalence classifier for math** For building the diversity classifier, Zhang et al. (2025b) collected annotated training data which are sampled prompts from WildChat (Zhao et al., 2024) filtered for non-verifiable tasks. To adapt their method to math, we sample prompts from DeepscaleR and collect trajectories generated by 8 different models spanning multiple model families and sizes. We annotate whether a pair of trajectories is semantically equivalent using Llama-3.3-70B-Instruct (Llama Team, 2024). We then finetune a

<sup>2</sup><https://github.com/huggingface/Math-Verify>

Qwen3-Embedding-4B (Zhang et al., 2025a) model using the annotations to obtain our math semantic equivalence classifier. Details on how we perform trajectory sampling, annotations, and classifier training can be found in Appendix A.2.

## 5.2 EXPERIMENTAL RESULTS

**DARLING improves both pass@1 and pass@k in competition math.** Figure 4 shows our main results: we plot pass@k from k=1 to k=128. We observe that DARLING outperforms the GRPO baseline in both quality and diversity. First, for pass@1 (as a measure of quality), DARLING outperforms GRPO by +3.51/1.90% averaged across 4 benchmarks for 4B and 14B models respectively. Next, for pass@128 (as a measure of diversity), DARLING outperforms GRPO by +7.62/10.16%. This shows that by jointly reinforcing quality and diversity, DARLING is able to achieve the best of both worlds in competition math benchmarks — simultaneously achieving better pass@1 and pass@k. Furthermore, we observe the largest gains on HMMT, the most challenging of the four datasets, suggesting that enhanced exploration yields greater improvements on harder datasets. We report accuracy for each of the datasets in Appendix F.

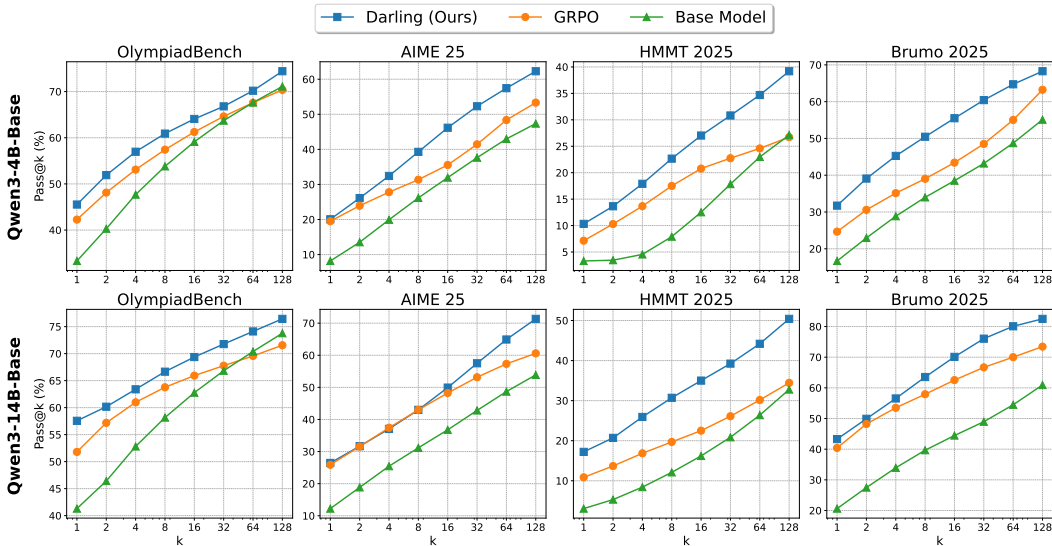


Figure 4: Comparison of different pass@k values by applying GRPO and DARLING on Qwen3-4B-Base and Qwen3-14B-Base on competition math benchmarks. DARLING outperforms GRPO simultaneously for pass@1 (+3.51/+1.9% on avg.) and pass@128 (+7.62/10.16% on avg.) on 4B and 14B models respectively. DARLING simultaneously achieves the best quality and diversity averaged across 4 competition math benchmarks.

## 6 ABLATIONS

In this section, we perform ablations on the design choices of DARLING. In particular, we compare additive aggregation of quality and diversity rewards v.s. multiplicative aggregation (ours) in §6.1 and we compare our partition classifier that measures semantic diversity with traditional lexical diversity in §6.2. We defer the ablation of advantage normalization to Appendix B.

### 6.1 ABLATIONS ON MULTIPLICATIVE V.S. ADDITIVE AGGREGATION

A naive aggregation of the quality reward  $r$  and the diversity signal  $Div_d$  is to add the two rewards, as many existing works perform *additive* aggregation of the quality-based reward and auxiliary rewards such as length (Aggarwal & Welleck, 2025; Liu et al., 2025b), entropy (Cheng et al., 2025), and format (Wu et al., 2025b).

We ablate the effect of adding or multiplying the quality reward  $r$  and the diversity reward  $Norm(Div)$  in equation 5 in our non-verifiable setting and report the results in Table 2. To ag-

gregate the two rewards, we manually normalize (subtract mean and divide by standard deviation) both  $r$  (the quality reward) and  $\text{Div}_d$ .

	AlpacaEval 2.0	ArenaHard v2.0		NoveltyBench
	LCWR(%)	creative writing	hard prompts	Distinct (#)
Baseline	31.92	7.1 (-1.4 / +1.8)	6.4 (-0.8 / +1.3)	5.28
Quality (GRPO)	48.74	61.1 (-3.5 / +4.5)	<b>33.9</b> (-2.3 / +2.5)	2.08
Quality + partition	53.17	<b>69.2</b> (-3.6 / +3.9)	32.7 (-3.2 / +2.9)	5.23
<b>DARLING</b> = Quality $\times$ partition	<b>55.15</b>	68.8 (-3.3 / +2.9)	31.1 (-2.0 / +2.1)	<b>5.49</b>

Table 2: Ablation comparing the way we aggregate the quality and diversity reward: additive v.s. multiplicative in Equation 5, evaluating on AlpacaEval 2.0, ArenaHard v2.0 (with Style Control), and NoveltyBench. Baseline = Llama-3.1-8B-Instruct.

We observe that multiplicative aggregation (DARLING) outperforms additive aggregation in AlpacaEval 2.0, and performs similarly in ArenaHard v2.0 and NoveltyBench. We opt for multiplicative aggregation due to its simplicity as it does not require additional handling of mismatched reward scales and hyperparameter tuning of mixing weights on individual rewards.

## 6.2 ABLATIONS ON LEXICAL METRICS FOR DIVERSITY

We study whether our proposed partition classifier can be replaced by a simple lexical diversity metric. We replace our semantic equivalence classifier with the number of distinct  $N$ -grams in online RL training. Specifically, we set the diversity of a response  $y_i$  w.r.t all other responses for the same input prompt as:

$$\text{Div}_{N\text{-gram}}(y_i | y_1, \dots, y_n) = \frac{\text{number of distinct } N\text{-grams that only appear in } y_i}{\text{total number of } N\text{-grams in } y_i}.$$

This means that if no  $N$ -gram in  $y_i$  appears in any other response  $y_j (j \neq i)$ , then the diversity  $\text{Div}_{N\text{-gram}}(y_i | y_1, \dots, y_n) = 1$ . Similarly,  $\text{Div}_{N\text{-gram}}(y_i | y_1, \dots, y_n) = 0$  if all  $N$ -grams in  $y_i$  appear in at least one other response. In our experiments, we set  $N = 4$  and denote this setting as “Quality  $\times$  4gram”. We report experimental results on non-verifiable tasks in Table 3.

	AlpacaEval 2.0	ArenaHard v2.0		NoveltyBench
	LCWR(%)	creative writing	hard prompts	Distinct (#)
Baseline	31.92	7.1 (-1.4 / +1.8)	6.4 (-0.8 / +1.3)	5.28
Quality (GRPO)	48.74	61.1 (-3.5 / +4.5)	<b>33.9</b> (-2.3 / +2.5)	2.08
Quality $\times$ 4gram	53.82	<b>71.9</b> (-3.3 / +3.6)	31.3 (-2.3 / +2.9)	3.59
<b>DARLING</b> = Quality $\times$ partition	<b>55.15</b>	68.8 (-3.3 / +2.9)	31.1 (-2.0 / +2.1)	<b>5.49</b>

Table 3: Comparison of  $N$ -gram diversity loss to DARLING. The  $N$ -gram diversity loss ( $N=4$ ) performs similarly to DARLING in terms of quality, but underperforms DARLING in terms of diversity in NoveltyBench. Baseline = Llama-3.1-8B-Instruct.

We observe that while 4gram diversity integrated with quality is able to match the performance of DARLING in LM-as-a-Judge evaluations (AlpacaEval 2.0, ArenaHard v2.0), it significantly underperforms DARLING in semantic diversity assessment (NoveltyBench).

Additionally, we evaluate the performance of using 4gram diversity in competition math and report the results in Table 13. We found that in math questions, using lexical diversity as a reward underperforms the GRPO baseline in terms of pass@1 performance. We analyze the the generations and observe that the policy often hacks the ngram diversity reward by generating texts that are of a different language, or self-reflections about the difficulty of the problem. We provide an example of such ngram reward hacking in Appendix G.

Model	Pass@128				Pass@1			
	AIME	HMMT	Olympiad	Brumo	AIME	HMMT	Olympiad	Brumo
Qwen3-4B-Base	47.35	27.12	71.11	55.10	8.17	1.28	31.13	16.68
Quality (GRPO)	53.33	26.72	70.37	63.24	19.51	7.14	42.27	24.66
Quality $\times$ 4gram	57.47	32.35	67.47	60.55	17.44	6.95	40.03	25.55
DARLING =Quality $\times$ partition	<b>62.28</b>	<b>39.19</b>	<b>74.41</b>	<b>68.27</b>	<b>20.06</b>	<b>10.32</b>	<b>45.53</b>	<b>31.73</b>

Table 4: Comparison of n-gram diversity loss to DARLING on Competition Math. Using 4gram as the diversity reward underperforms the baseline GRPO (no diversity reward), indicating that lexical diversity reward can harm performance in competition math.

## 7 RELATED WORK

In this section, we provide an overview of previous work that elicits diverse responses from LMs during *training* and *inference*, and clarify the distinction from our work. We defer additional related work on diversity evaluation metrics and RL for LMs to Appendix H. For a more comprehensive survey on LM creativity, we also refer readers to Ismayilzada et al. (2024).

**Training-time strategies for diversity** LMs often generate repetitive outputs, a long-standing challenge in the community (Li et al., 2016b; Zhang et al., 2021). Prior work addresses this by modifying the maximum likelihood training objective to encourage diversity. For example, Li et al. (2016b) maximizes mutual information to avoid generic responses (e.g., *I don’t know*). Welleck et al. (2020) penalizes repetitions to improve lexical variety within a response. Other approaches smooth or modify the one-hot target distribution: Li et al. (2020) introduces a Gaussian prior, Zhang et al. (2024) matches outputs to high-entropy distributions, and Li et al. (2025b) applies sparse logit updates. Beyond cross-entropy, DivPO (Lanchantin et al., 2025a) and its “soft” variants (Chung et al., 2025; Ismayilzada et al., 2025) optimize preferences for both quality and diversity. In online RL, He et al. (2025a) perform re-weighting of rewards by likelihood to promote diverse proofs, Lanchantin et al. (2025b) show that using simple entropy regularization is a non-trivial task, and Slocum et al. (2025) attribute diversity loss to KL regularization and decouple its terms. Concurrent to our work, Chen et al. (2025a) down-weight uncertain model solutions in math. DARLING mainly differs from other approaches in two important ways: (1) it employs a semantic-level diversity signal, going beyond surface-level lexical variations, and (2) it directly shapes the reward during online RL, unlike prior work that modifies cross-entropy loss in pre-training (Li et al., 2020) or offline fine-tuning (Li et al., 2025b; Lanchantin et al., 2025a; Chung et al., 2025; Ismayilzada et al., 2025).

**Inference-time strategies for diversity** Decoding diverse outputs from neural LMs has been a well-studied problem in the literature. A body of prior work has proposed modifications to beam search (Cho, 2016; Li & Jurafsky, 2016; Li et al., 2016b; Vijayakumar et al., 2018; Kulikov et al., 2019). Ippolito et al. (2019a), in their work, compares such methods with those that simply increase the sampling temperature (Peeperkorn et al., 2024; Shur-Ofry et al., 2024). Apart from modifying the beam search process, many methods have proposed to harness the prompt to elicit diverse responses, which includes conditioning on random perturbations of the input (Nagarajan et al., 2025), on different personas (Shur-Ofry et al., 2024; Ge et al., 2025), on past generations (Lu et al., 2024), on highly diverse summaries or keywords (Ahmed et al., 2025), on diverse base model generations (Zhu et al., 2025), and directly prompting the model to “*be diverse*” (Zhang et al., 2025b). Both Padmakumar et al. (2025) and Zhang et al. (2025b) present a comprehensive evaluation of such prompting methods, revealing improved diversity often comes at the cost of degraded quality. Our work directly modifies the training objective which is orthogonal to and compatible with decoding methods that elicit diversity at test time.

## 8 CONCLUSION

In this work we introduced DARLING, an online RL method that jointly optimizes for both quality and diversity for LM generations. Unlike prior RL approaches that often lead to diversity collapse, DARLING effectively preserves diversity among generations.

## REFERENCES

- 486  
487  
488 Pranjali Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with  
489 reinforcement learning. In *Second Conference on Language Modeling*, 2025. URL <https://arxiv.org/abs/2503.04697>. Cited on page 7.
- 491 Eltayeb Ahmed, Uljad Berdica, Martha Elliott, Danijela Horak, and Jakob N. Foerster. Intent  
492 factored generation: Unleashing the diversity in your language model, 2025. URL <https://arxiv.org/abs/2506.09659>. Cited on page 9.
- 494 Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing  
495 Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. Polaris: A post-training recipe for scal-  
496 ing reinforcement learning on advanced reasoning models, 2025. URL [https://hkunlp.  
497 github.io/blog/2025/Polaris](https://hkunlp.github.io/blog/2025/Polaris). Cited on pages 1 and 29.
- 499 Anthropic. Claude 3.7 sonnet system card, 2024. URL [https://assets.anthropic.com/  
500 m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf](https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf).  
501 Cited on page 5.
- 502 Art of Problem Solving. Aime problems and solutions, 2025. URL [https://artofproblemsolving.com/wiki/index.php/AIME\\_Problems\\_and\\_](https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions)  
503 [Solutions](https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions). Cited on pages 6, 23, and 26.
- 506 Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Math-  
507 arena: Evaluating llms on uncontaminated math competitions, February 2025. URL <https://matharena.ai/>. Cited on pages 6, 23, and 26.
- 509 Florian Le Bronnec, Alexandre Verine, Benjamin Negrevergne, Yann Chevaleyre, and Alexandre  
510 Allauzen. Exploring precision and recall to assess the quality and diversity of llms, 2024. URL  
511 <https://arxiv.org/abs/2402.10693>. Cited on page 29.
- 513 Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared  
514 Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri,  
515 Gretchen Krueger, and Michael Petrov et al. Evaluating large language models trained on code,  
516 2021. URL <https://arxiv.org/abs/2107.03374>. Cited on page 6.
- 517 Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced  
518 grpo for uncertainty-aware policy optimization, 2025a. URL [https://arxiv.org/abs/  
519 2505.12346](https://arxiv.org/abs/2505.12346). Cited on pages 1 and 9.
- 520 Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi.  
521 Pass@k training for adaptively balancing exploration and exploitation of large reasoning models,  
522 2025b. URL <https://arxiv.org/abs/2508.10751>. Cited on page 29.
- 524 Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and  
525 Furu Wei. Reasoning with exploration: An entropy perspective, 2025. URL [https://arxiv.  
526 org/abs/2506.14758](https://arxiv.org/abs/2506.14758). Cited on pages 1 and 7.
- 527 Kyunghyun Cho. Noisy parallel approximate decoding for conditional recurrent language model,  
528 2016. URL <https://arxiv.org/abs/1605.03835>. Cited on page 9.
- 530 John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max  
531 Kreminski. Modifying large language model post-training for diverse creative writing, 2025.  
532 URL <https://arxiv.org/abs/2503.17126>. Cited on page 9.
- 533 Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen  
534 Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng,  
535 Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning  
536 language models, 2025. URL <https://arxiv.org/abs/2505.22617>. Cited on page 29.
- 537  
538 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,  
539 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,  
Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao

- 540 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,  
541 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,  
542 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,  
543 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang  
544 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai  
545 Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,  
546 Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang,  
547 Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,  
548 Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,  
549 R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng  
550 Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing  
551 Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanxia Zhao, Wen  
552 Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong  
553 Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu,  
554 Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xi-  
555 aosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia  
556 Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng  
557 Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong  
558 Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong,  
559 Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou,  
560 Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying  
561 Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda  
562 Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu,  
563 Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu  
564 Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforce-  
565 ment learning, 2025. URL <https://arxiv.org/abs/2501.12948>. Cited on page 3.
- 566 Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled al-  
567 pacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.  
568 Cited on pages 5, 23, and 26.
- 569 Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna  
570 Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association  
571 for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Aus-  
572 tralia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL  
<https://aclanthology.org/P18-1082/>. Cited on page 1.
- 573 Evan Frick, Peter Jin, Tianle Li, Karthik Ganesan, Jian Zhang, Jiantao Jiao, and Banghua Zhu.  
574 Athene-70b: Redefining the boundaries of post-training for open models, July 2024. URL  
575 <https://nexusflow.ai/blogs/athene>. Cited on page 4.
- 576 Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data cre-  
577 ation with 1,000,000,000 personas, 2025. URL <https://arxiv.org/abs/2406.20094>.  
578 Cited on page 9.
- 579 Nate Gruver, Samuel Don Stanton, Nathan C. Frey, Tim G. J. Rudner, Isidro Hotzel, Julien  
580 Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein design  
581 with guided discrete diffusion. In *Thirty-seventh Conference on Neural Information Processing  
582 Systems*, 2023. URL <https://openreview.net/forum?id=MfiK69Ga6p>. Cited on  
583 page 1.
- 584 J. P. Guilford. Creativity. *American Psychologist*, 5(9):444–454, 1950. doi: [https://doi.org/10.1037/  
585 h0063487](https://doi.org/10.1037/h0063487). Cited on page 1.
- 586 Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution  
587 sharpening, 2025a. URL <https://arxiv.org/abs/2506.02355>. Cited on pages 4, 9,  
588 and 29.
- 589 Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han,  
590 Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench:  
591 A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scien-  
592 tific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the  
593*

- 594 *62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa-*  
595 *pers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguis-  
596 tics. doi: 10.18653/v1/2024.acl-long.211. URL <https://aclanthology.org/2024.acl-long.211/>. Cited on pages 6, 23, and 26.
- 598  
599 Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang  
600 Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An,  
601 Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report, 2025b. URL <https://arxiv.org/abs/2505.22312>. Cited on page 29.
- 602  
603 Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv*  
604 *preprint arXiv:2501.03262*, 2025. Cited on page 3.
- 605  
606 Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan  
607 T. Ash, and Akshay Krishnamurthy. Self-Improvement in Language Models: The Sharpening  
608 Mechanism. In *International Conference on Learning Representations (iclr)*, 2025. URL  
609 <https://openreview.net/forum?id=WJaUkwci9o>. Cited on page 1.
- 610  
611 Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. Compari-  
612 son of diverse decoding methods from conditional language models. In Anna Korhonen, David  
613 Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association*  
614 *for Computational Linguistics*, pp. 3752–3762, Florence, Italy, July 2019a. Association for Com-  
615 putational Linguistics. doi: 10.18653/v1/P19-1365. URL <https://aclanthology.org/P19-1365/>. Cited on page 9.
- 616  
617 Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. Comparison  
618 of Diverse Decoding Methods from Conditional Language Models. In *Annual Meeting of the*  
619 *Association for Computational Linguistics (ACL)*, 2019b. Cited on page 29.
- 620  
621 Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. Creativity in ai: Pro-  
622 gresses and challenges, 2024. URL <https://arxiv.org/abs/2410.17218>. Cited on  
623 page 9.
- 624  
625 Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lon-  
626 neke van der Plas, and Roger Beaty. Creative preference optimization, 2025. URL <https://arxiv.org/abs/2505.14442>. Cited on page 9.
- 627  
628 Ke Ji, Jiahao Xu, Tian Liang, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang,  
629 Junying Chen, Benyou Wang, et al. The first few tokens are all you need: An efficient and effective  
630 unsupervised prefix fine-tuning method for reasoning models. *arXiv preprint arXiv:2503.02875*,  
631 2025. Cited on page 1.
- 632  
633 Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye,  
634 Mostafa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthe-  
635 sis: Gradient-based data diversification boosts generalization in llm reasoning, 2025. URL  
636 <https://arxiv.org/abs/2505.20161>. Cited on page 29.
- 637  
638 Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward  
639 Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisa-  
640 tion and diversity. In *International Conference on Learning Representations (iclr)*, 2024. URL  
<https://openreview.net/forum?id=PXD3FAVHJT>. Cited on pages 1 and 29.
- 641  
642 T. Kloek and H. K. van Dijk. Bayesian estimates of equation system parameters: An application of  
643 integration by monte carlo. *Econometrica*, 46(1):1–19, 1978. ISSN 00129682, 14680262. URL  
644 <http://www.jstor.org/stable/1913641>. Cited on page 3.
- 645  
646 Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances  
647 for uncertainty estimation in natural language generation. In *The Eleventh International Confer-*  
*ence on Learning Representations*, 2023. URL <https://openreview.net/forum?id=VD-AYtP0dve>. Cited on page 3.

- 648 Ilya Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. Importance of search and eval-  
649 uation strategies in neural dialogue modeling. In Kees van Deemter, Chenghua Lin, and Hiroya  
650 Takamura (eds.), *Proceedings of the 12th International Conference on Natural Language Gen-  
651 eration*, pp. 76–87, Tokyo, Japan, October–November 2019. Association for Computational Lin-  
652 guistics. doi: 10.18653/v1/W19-8609. URL <https://aclanthology.org/W19-8609/>.  
653 Cited on page 9.
- 654 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.  
655 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model  
656 serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating  
657 Systems Principles*, 2023. Cited on page 23.
- 659 Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar  
660 Sukhbaatar, and Ilya Kulikov. Diverse preference optimization, 2025a. URL <https://arxiv.org/abs/2501.18101>. Cited on pages 4, 9, and 29.
- 662 Jack Lanchantin, Angelica Chen, Janice Lan, Xian Li, Swarnadeep Saha, Tianlu Wang, Jing Xu,  
663 Ping Yu, Weizhe Yuan, Jason E Weston, et al. Bridging offline and online reinforcement learning  
664 for llms. *arXiv preprint arXiv:2506.21495*, 2025b. Cited on pages 5 and 9.
- 666 Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine trans-  
667 lation, 2016. URL <https://arxiv.org/abs/1601.00372>. Cited on page 9.
- 668 Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objec-  
669 tive function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow  
670 (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for  
671 Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California,  
672 June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL  
673 <https://aclanthology.org/N16-1014/>. Cited on pages 1 and 29.
- 674 Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting  
675 objective function for neural conversation models, 2016b. URL [https://arxiv.org/abs/  
676 1510.03055](https://arxiv.org/abs/1510.03055). Cited on pages 9 and 29.
- 678 Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonza-  
679 lez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and bench-  
680 builder pipeline. In *Forty-second International Conference on Machine Learning*, 2025a. URL  
681 <https://openreview.net/forum?id=KfTf9vFvSn>. Cited on pages 5, 23, and 26.
- 682 Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy  
683 Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following  
684 models. [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval), 5 2023. Cited on page 5.
- 686 Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun.  
687 Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth In-  
688 ternational Conference on Learning Representations*, 2025b. URL [https://openreview.  
689 net/forum?id=NQEe7B7bSw](https://openreview.net/forum?id=NQEe7B7bSw). Cited on pages 1 and 9.
- 690 Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai  
691 Zhao. Data-dependent gaussian prior objective for language generation. In *International Confer-  
692 ence on Learning Representations*, 2020. URL [https://openreview.net/forum?id=  
693 S1efxTVYDr](https://openreview.net/forum?id=S1efxTVYDr). Cited on page 9.
- 694 Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu  
695 Chen. Beyond pass@1: Self-play with variational problem synthesis sustains rlvr, 2025. URL  
696 <https://arxiv.org/abs/2508.14029>. Cited on page 29.
- 698 Xiang Lin, Simeng Han, and Shafiq Joty. Straight to the gradient: Learning to use novel tokens  
699 for neural text generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th  
700 International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning  
701 Research*, pp. 6642–6653. PMLR, 18–24 Jul 2021. URL [https://proceedings.mlr.  
press/v139/lin21b.html](https://proceedings.mlr.press/v139/lin21b.html). Cited on page 1.

- 702 Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong.  
703 Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models.  
704 *arXiv preprint*, 2025a. URL <https://arxiv.org/abs/2505.24864>. Cited on pages 3  
705 and 29.
- 706 Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang,  
707 and Junxian He. Learn to reason efficiently with adaptive length-based reward shaping, 2025b.  
708 URL <https://arxiv.org/abs/2505.15612>. Cited on page 7.
- 709 Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee,  
710 and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint*  
711 *arXiv:2503.20783*, 2025c. Cited on pages 3 and 4.
- 712 Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan  
713 Xiong, Ju Huang, Jian Hu, Shengyi Huang, Siran Yang, Jiamang Wang, Wenbo Su, and Bo Zheng.  
714 Part i: Tricks or traps? a deep dive into rl for llm reasoning, 2025d. URL <https://arxiv.org/abs/2508.08221>. Cited on pages 21 and 29.
- 715 Llama Team. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>. Cited on pages 4 and 6.
- 716 Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Miresghallah, Jiacheng Liu, Seungju Han,  
717 Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. AI as humanity’s  
718 salieri: Quantifying linguistic creativity of language models via systematic attribution of machine  
719 text against web text. In *The Thirteenth International Conference on Learning Representations*,  
720 2025. URL <https://openreview.net/forum?id=ilOE0Iqo1Q>. Cited on pages 1  
721 and 29.
- 722 Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, and Daniel Khashabi. Benchmarking lan-  
723 guage model creativity: A case study on code generation. *arXiv preprint arXiv:2407.09007*, 2024.  
724 URL <https://arxiv.org/abs/2407.09007>. Cited on pages 9 and 29.
- 725 Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai,  
726 Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing  
727 o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog. Cited on pages 6, 19, 21, and 23.
- 728 Laura O’Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. Attributing Mode Col-  
729 lapse in the fine-tuning of Large Language Models. In *ICLR 2024 Workshop on Mathematical  
730 and Empirical Understanding of Foundation Models*, 2024. URL <https://openreview.net/forum?id=3pDMYjpOxk>. Cited on page 1.
- 731 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word repre-  
732 sentations in vector space. In *International Conference on Learning Representations (iclr)*, 2013.  
733 URL <https://arxiv.org/abs/1301.3781>. Cited on page 29.
- 734 Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. Roll the dice &  
735 look before you leap: Going beyond the creative limits of next-token prediction. In *Forty-second  
736 International Conference on Machine Learning*, 2025. URL [https://openreview.net/  
737 forum?id=Hi0SyHMmkd](https://openreview.net/forum?id=Hi0SyHMmkd). Cited on pages 9 and 29.
- 738 OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan  
739 Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-  
740 Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol,  
741 Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Con-  
742 neau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian,  
743 Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein,  
744 Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrew  
745 Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia,  
746 Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben  
747 Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake  
748 Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon  
749 Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo

756 Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li,  
757 Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu,  
758 Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim,  
759 Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley  
760 Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler,  
761 Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki,  
762 Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay,  
763 Edele Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow,  
764 Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Kho-  
765 rasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit,  
766 Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming  
767 Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun,  
768 Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won  
769 Chung, Ian Kivlichan, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim  
770 Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Ja-  
771 cob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James  
772 Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei,  
773 Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui  
774 Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinero Candela, Jo Beutler, Joe  
775 Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay,  
776 Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld,  
777 Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang,  
778 Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood,  
779 Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel  
780 Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Work-  
781 man, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka,  
782 Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas  
783 Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens,  
784 Madeline Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall,  
785 Marvin Zhang, Marwan Aljube, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty,  
786 Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glesne,  
787 Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang,  
788 Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail  
789 Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat  
790 Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers,  
791 Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Fe-  
792 lix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum,  
793 Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen  
794 Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum,  
795 Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe  
796 Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Ran-  
797 dall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza  
798 Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchan-  
799 dani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmat-  
800 ullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino,  
801 Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez  
802 Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia,  
803 Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir  
804 Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal  
805 Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas  
806 Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom  
807 Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie  
808 Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyei Zheng,  
809 Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik  
Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury  
Malkov. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>. Cited  
on page 5.

- 810 Vishakh Padmakumar and He He. Does writing with language models reduce content diversity?  
811 In *The Twelfth International Conference on Learning Representations*, 2024. URL [https://](https://openreview.net/forum?id=Feiz5HtCD0)  
812 [openreview.net/forum?id=Feiz5HtCD0](https://openreview.net/forum?id=Feiz5HtCD0). Cited on page 1.  
813
- 814 Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. Beyond memorization:  
815 Mapping the originality-quality frontier of language models, 2025. URL [https://arxiv.](https://arxiv.org/abs/2504.09389)  
816 [org/abs/2504.09389](https://arxiv.org/abs/2504.09389). Cited on pages 5 and 9.
- 817 Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2023.  
818 Cited on pages 5, 23, 24, and 26.  
819
- 820 Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativ-  
821 ity parameter of large language models?, 2024. URL [https://arxiv.org/abs/2405.](https://arxiv.org/abs/2405.00492)  
822 [00492](https://arxiv.org/abs/2405.00492). Cited on page 9.
- 823 Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi,  
824 and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using diver-  
825 gence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021. URL  
826 <https://arxiv.org/abs/2102.01454>. Cited on page 29.  
827
- 828 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea  
829 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances*  
830 *in Neural Information Processing Systems (NeurIPS)*, 36, 2024. URL [https://arxiv.org/](https://arxiv.org/abs/2305.18290)  
831 [abs/2305.18290](https://arxiv.org/abs/2305.18290). Cited on page 4.
- 832 Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog,  
833 M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang,  
834 Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program  
835 search with large language models. *Nat.*, 625(7995):468–475, January 2024. URL [https://](https://doi.org/10.1038/s41586-023-06924-6)  
836 [doi.org/10.1038/s41586-023-06924-6](https://doi.org/10.1038/s41586-023-06924-6). Cited on page 1.
- 837 Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval.  
838 *Information Processing & Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi: [https://](https://doi.org/10.1016/0306-4573(88)90021-0)  
839 [doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0). URL [https://www.sciencedirect.com/](https://www.sciencedirect.com/science/article/pii/0306457388900210)  
840 [science/article/pii/0306457388900210](https://www.sciencedirect.com/science/article/pii/0306457388900210). Cited on page 29.  
841
- 842 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li,  
843 Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open  
844 language models, 2024. URL <https://arxiv.org/abs/2402.03300>. Cited on pages 2,  
845 4, and 6.
- 846 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,  
847 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint*  
848 *arXiv: 2409.19256*, 2024. Cited on pages 4, 22, and 23.
- 849 Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim, and Yonatan Belinkov. Growing a tail:  
850 Increasing output diversity in large language models, 2024. URL [https://arxiv.org/](https://arxiv.org/abs/2411.02989)  
851 [abs/2411.02989](https://arxiv.org/abs/2411.02989). Cited on page 9.  
852
- 853 Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bas-  
854 tani. Evaluating the diversity and quality of llm generated content. In *Conference on Language*  
855 *Modeling*, 2025. URL <https://arxiv.org/abs/2504.12522>. Cited on pages 1 and 29.  
856
- 857 Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a  
858 large-scale human study with 100+ NLP researchers. In *The Thirteenth International Confer-*  
859 *ence on Learning Representations*, 2025. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=M23dTGWCzy)  
860 [M23dTGWCzy](https://openreview.net/forum?id=M23dTGWCzy). Cited on pages 1 and 29.
- 861 Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse preference learning  
862 for capabilities and alignment. In *The Thirteenth International Conference on Learning Repre-*  
863 *sentations*, 2025. URL <https://openreview.net/forum?id=pOq9vDIYev>. Cited on  
page 9.

- 864 Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David  
865 Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural se-  
866 quence models. In *Conference on Artificial Intelligence (AAAI)*, 2018. Cited on page 9.  
867
- 868 Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes  
869 reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025. Preprint. Cited on  
870 page 19.  
871
- 872 Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said  
873 Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin  
874 Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirec-  
875 tional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL  
876 <https://arxiv.org/abs/2412.13663>. Cited on page 19.  
877
- 878 Sean Welleck, Ilya Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neu-  
879 ral Text Generation With Unlikelihood Training. In *International Conference on Learning Rep-  
880 resentations (iclr)*, 2020. URL <https://openreview.net/forum?id=SJeYe0NtvH>.  
881 Cited on page 9.
- 882 Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity,  
883 2025. URL <https://arxiv.org/abs/2505.00047>. Cited on page 1.  
884
- 885 John Wieting and Kevin Gimpel. ParaNMT-50M: Pushing the limits of paraphrastic sentence em-  
886 beddings with millions of machine translations. In Iryna Gurevych and Yusuke Miyao (eds.), *Pro-  
887 ceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:  
888 Long Papers)*, pp. 451–462, Melbourne, Australia, July 2018. Association for Computational Lin-  
889 guistics. doi: 10.18653/v1/P18-1042. URL <https://aclanthology.org/P18-1042/>.  
890 Cited on page 29.
- 891 John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. Beyond BleU: Training  
892 Neural Machine Translation with Semantic Similarity. In *Annual Meeting of the Association for  
893 Computational Linguistics (ACL)*, 2019. Cited on page 29.
- 894 Fang Wu, Weihao Xuan, Ximing Lu, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr  
895 may not escape its origin, 2025a. URL <https://arxiv.org/abs/2507.14843>. Cited on  
896 page 1.
- 897 Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. Longwriter-zero: Mastering  
898 ultra-long text generation via reinforcement learning, 2025b. URL [https://arxiv.org/  
899 abs/2506.18841](https://arxiv.org/abs/2506.18841). Cited on pages 7 and 29.
- 900 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang  
901 Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,  
902 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin  
903 Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,  
904 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui  
905 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang  
906 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger  
907 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan  
908 Qiu. Qwen3 technical report, 2025a. URL <https://arxiv.org/abs/2505.09388>. Cited  
909 on pages 3 and 6.
- 910 Chenghao Yang and Ari Holtzman. How alignment shrinks the generative horizon, 2025. URL  
911 <https://arxiv.org/abs/2506.17871>. Cited on pages 1 and 29.
- 912 Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Dongchun Xie, Yiwei Wang, Xiaodan  
913 Liang, and Jing Tang. Depth-breadth synergy in rlvr: Unlocking llm reasoning gains with adaptive  
914 exploration, 2025b. URL <https://arxiv.org/abs/2508.13755>. Cited on page 29.  
915
- 916 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai,  
917 Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guang-  
ming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu,

- 918 Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao  
919 Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingx-  
920 uan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL  
921 <https://arxiv.org/abs/2503.14476>. Cited on pages 3, 4, and 29.
- 922  
923 Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao  
924 Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond  
925 the base model? In *2nd AI for Math Workshop @ ICML 2025*, 2025. URL <https://openreview.net/forum?id=upehLVgqlb>. Cited on page 6.
- 926  
927 Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. The price of format:  
928 Diversity collapse in llms, 2025. URL <https://arxiv.org/abs/2505.18949>. Cited on  
929 page 1.
- 930  
931 Weihao Zeng, Yuzhen Huang, Lulu Zhao, Yijun Wang, Zifei Shan, and Junxian He. B-Star: Mon-  
932 itoring and balancing exploration and exploitation in self-taught reasoners. In *The Thirteenth  
933 International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=P6dwZJpJ4m>. Cited on page 29.
- 934  
935 Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity  
936 and quality in natural language generation. In Anya Belz, Shubham Agarwal, Yvette Graham,  
937 Ehud Reiter, and Anastasia Shimorina (eds.), *Proceedings of the Workshop on Human Evaluation  
938 of NLP Systems (HumEval)*, pp. 25–33, Online, April 2021. Association for Computational Lin-  
939 guistics. URL <https://aclanthology.org/2021.humeval-1.3/>. Cited on pages 5  
940 and 9.
- 941  
942 Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie,  
943 An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advanc-  
944 ing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*,  
2025a. Cited on pages 7 and 19.
- 945  
946 Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, J Zico Kolter, and Daphne Ippolito. Forcing  
947 Diffuse Distributions out of Language Models. In *Conference on Language Modeling*, 2024. URL  
948 <https://openreview.net/forum?id=9JY1QLVFPZ>. Cited on pages 1, 9, and 29.
- 949  
950 Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry  
951 Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diver-  
952 sity. In *Conference on Language Modeling*, 2025b. URL <https://arxiv.org/abs/2504.05228>. Cited on pages 3, 4, 5, 6, 9, 19, 23, 24, 26, and 29.
- 953  
954 Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m  
955 chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Rep-  
956 resentations*, 2024. URL <https://openreview.net/forum?id=B18u7ZR1bM>. Cited  
on pages 4, 6, 19, and 26.
- 957  
958 Alan Zhu, Parth Asawa, Jared Quincy Davis, Lingjiao Chen, Boris Hanin, Ion Stoica, Joseph E.  
959 Gonzalez, and Matei Zaharia. Bare: Leveraging base language models for few-shot synthetic  
960 data generation, 2025. URL <https://arxiv.org/abs/2502.01697>. Cited on page 9.
- 961  
962 Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Tegygen:  
963 A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Con-  
964 ference on Research & Development in Information Retrieval, SIGIR '18*, pp. 1097–1100, New  
965 York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.  
966 1145/3209978.3210080. URL <https://doi.org/10.1145/3209978.3210080>. Cited  
967 on page 29.
- 968  
969  
970  
971

## A PARTITIONING THE RESPONSES

### A.1 CLASSIFIER FOR NON-VERIFIABLE TASKS

Zhang et al. (2025b) had human annotators judge whether pairs of model-generated responses were semantically equivalent, across 1,100 prompts (2,200 responses in total). We directly use their annotations and concatenate the two responses to be classified as semantically similar or not:

```
[CLS] response 1 [SEP] response 2 [CLS]
```

and perform classification on top of the second [CLS] token. We train a ModernBERT-base (Warner et al., 2024) model with 1000 NoveltyBench annotations (2000 responses) to support a max context length of 8192 tokens. We evaluate the performance of the classifier using a held-out set of 100 prompts (200 responses) and plot the performance in Figure 5. We found that (1) our trained classifier (Acc.=78%) achieves similar performance compared to the original NoveltyBench classifier (Acc.=79%), and (2) Proprietary models (e.g. GPT-4o and o1-mini) perform worse in terms of determining whether two responses are semantically equivalent to humans. We provide the detailed prompt we used for asking an LM to determine whether two responses are semantically similar in Figure 6.

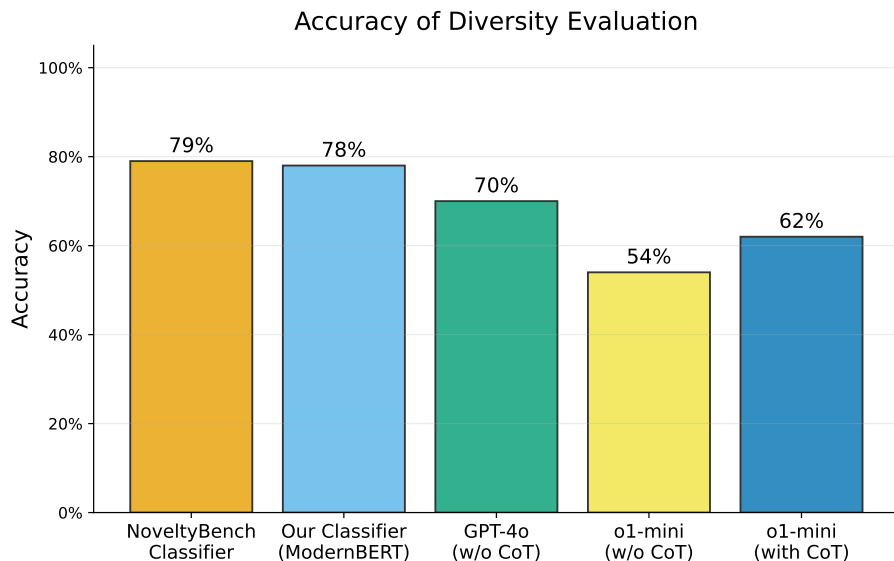


Figure 5: Performance of different classifiers on 100 held out human annotated data of whether two responses are similar. **Classifier based approaches outperform proprietary models in determining whether two responses are semantically similar to humans.**

### A.2 CLASSIFIER FOR VERIFIABLE TASKS

The original NoveltyBench (Zhang et al., 2025b) only supports non-verifiable tasks as the prompts were filtered to only be creative-writing prompts from WildChat (Zhao et al., 2024). Therefore, we additionally trained a classifier using Qwen3-4B-Embeddings (Zhang et al., 2025a) on top of model-generated solution traces. In particular, we performed inference with temperature = 1 using prompts from DeepscaleR (Luo et al., 2025) with the following models: Qwen3-4B-Base, Qwen3-8B-Base, Qwen3-4B (without thinking), Qwen3-8B (without thinking), OctoThinker-8B-Long-Base (Wang et al., 2025), Qwen2.5-Math-7B-Instruct, QwQ-32B, and Llama-4-Maverick to cover diverse solution traces with different model types (base, instruct), families, and sizes. We prompted Llama-3.3-70B-Instruct using the prompt in Figure 7 as the ground truth of whether two math solutions are similar. Our classifier is trained on 10k pairs of responses, achieving an 89% accuracy on a held-out validation set of 200 examples.

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

**Prompt for LM-as-a-diversity-judge (with CoT)**

You are given the original prompt and two model-generated responses. Determine whether these responses are semantically equivalent, based on whether reading the second response would provide the reader with substantially new or different information compared to the first.

Original prompt: {prompt}

Generation 0: {gen0}

Generation 1: {gen1}

Question: Are Generation 0 and Generation 1 semantically equivalent?

Think briefly step-by-step:

Core Meaning: Do both responses essentially communicate the same key points or concepts?

Additional Information: Would reading the second response significantly add new ideas, examples, or important details beyond the first?

Briefly provide your reasoning, then explicitly conclude:

[[Yes]]: The second response does not significantly add new information or insights.

[[No]]: The second response introduces meaningful new or distinct ideas, insights, or details.

Figure 6: The prompt with chain-of-thought to ask an language model whether two responses are semantically similar.

**Math Prompt for LM-as-a-diversity-judge (with CoT)**

You are given the original prompt and two model-generated responses. Determine whether the two responses use different strategies to solve the problem.

Use the following guidelines:

- Different solution methods: Clearly different approaches (e.g., algebraic vs. geometric, analytical vs. numerical).
- Critical reasoning divergence: Significant differences in key reasoning steps or assumptions, even if final answers match.
- Conceptual differences: Distinct underlying concepts or representations (e.g., probability vs. combinatorics).

**\*\*Also label as different if:\*\*** The two responses share the same general approach but differ meaningfully in specific intermediate steps or manipulations crucial to the solution.

Original prompt: {prompt}

Generation 0: {gen0}

Generation 1: {gen1}

Question: Do Generation 0 and Generation 1 use different strategies? You may first generate a short reasoning, then respond with "[[yes]]" if the generations use different strategies or "[[no]]" if they do not.

Figure 7: Prompt to Llama-3.3-70B-Instruct on whether two math solution traces (excluding the final answer) are semantically similar.

## B ABLATIONS ON ADVANTAGE NORMALIZATION IN GRPO

Formally, let the reward for response  $y_i$  be

$$r_i = f_i + \varepsilon_i,$$

where  $f_i$  is the true underlying utility and  $\varepsilon_i$  is noise with variance  $\tau^2$ . GRPO with normalization computes

$$\hat{r}_i = \frac{r_i - \bar{r}}{\sigma_r}, \quad \sigma_r^2 \approx \text{Var}(f) + \tau^2,$$

so each prompt contributes unit variance to the gradient update. This has the effect of amplifying noise when  $\tau^2$  is large relative to  $\text{Var}(f)$  (dense but noisy rewards), because even very small differences between responses get magnified into values of order one. By contrast, removing the normalization yields

$$\tilde{r}_i = r_i - \bar{r},$$

which preserves the true scale of reward differences. Thus, normalization is helpful when rewards are reliable (high signal-to-noise ratio), but harmful when they are noisy and tightly clustered.

Empirically, we find that removing the standard deviation term improves performance in settings with dense and noisy rewards. Table 5 shows results in our non-verifiable setting with a Bradley–Terry style reward: removing normalization (“w/o norm”) improves not only quality (AlpacaEval and Arena-Hard win rates) but also diversity (NoveltyBench Distinct and distinct n-grams).

	AlpacaEval 2.0*	ArenaHard v2.0*	NoveltyBench	
	LCWR(%)	Creative Writing (%)	Distinct (#)	Distinct-4 (%)
GRPO	48.74	61.1 (-3.5 / +4.5)	2.08	92.84
GRPO (w/o norm)	52.57 (+3.83)	68.1 (-3.5 / +2.7) (+7.0)	2.28 (+0.20)	94.05 (+1.21)
4gram	48.48	65.3 (-3.3 / +3.6)	2.79	93.87
4gram (w/o norm)	53.82 (+5.34)	71.9 (-3.3 / +3.6) (+6.6)	3.59 (+0.80)	95.63 (+1.76)
partition	51.64	69.7 (-3.3 / +4.0)	3.35	94.93
DARLING = partition (w/o norm)	55.15 (+3.51)	68.8 (-3.3 / +2.9) (-0.9)	5.49 (+2.14)	96.04 (+1.11)

Table 5: Ablation study on normalization: Results for GRPO baseline, 4-gram, and partition mixing, each with and without normalization. All metrics are the higher the better. \* indicates GPT-4o was used as the judge. Removing normalization (w/o norm) prevents the amplification of tiny differences in dense rewards, resulting in improved performance on both quality and diversity metrics.

In contrast, in settings where the reward is sparse and noise-free, normalization has little effect. ?? reports results on Math, where rewards are binary (0, 1) and deterministic. In this case, the variance comes entirely from the true differences, so normalization is neither helpful nor harmful.

Table 6 shows the result of GRPO with and without the “divide by standard deviation” trick on 4 competition math benchmarks. Normalization has little effect under this setting. For a more comprehensive study on the effect of normalization, we refer the readers to Liu et al. (2025d).

Model=Qwen-4B-Base	Pass@128				Pass@1			
	AIME	HMMT	Olympiad	Brumo	AIME	HMMT	Olympiad	Brumo
GRPO	53.33	26.72	70.37	63.24	19.51	7.14	42.27	24.66
GRPO (w/o norm)	55.13	25.45	69.89	63.02	21.15	7.04	41.36	24.45

Table 6: Ablation study of GRPO normalization on Math tasks. Training is performed on Qwen-4B-Base using the DeepScaleR (Luo et al., 2025) dataset. In contrast to Table 5, removing normalization has little to no effect, since the rewards here are binary (0 or 1), sparse, and noise-free.

## C HYPERPARAMETERS

**Hyperparameters for Non-verifiable Tasks** Table 7 shows key hyperparameters for our GRPO training on non-verifiable tasks (WildChat). We train our models using 1 nodes / 4 nodes of NVIDIA H200 for the 8B and 70B model, respectively.

Category	Hyperparameter	Value
Data	Train file	WildChat
	Max prompt length	512
	Max response length	1024
	Filter overlong prompts	True
Actor Model	Base model 1	Llama-3.1-8B-Instruct
	Base model 2	Llama-3.3-70B-Instruct
	LR	$1 \times 10^{-6}$
	KL loss coefficient $\beta$	0.001
	KL loss type	low_var_kl
	Use dynamic batch size	True
Rollout	Rollout engine	vllm
	GPU mem utilization	0.8
	Train rollout n	8
	Temperature	1.0
Reward Model	RM model	Athene-RM-8B
Trainer	Mini Batch size	32 & 64
	Full Batch size	32 & 64 (Fully on-policy)
	Critic Warmup	0
	GPUs/node	8
	Nodes	1 (8B), 4 (70B)
	Total epochs	10

Table 7: Key hyperparameters used for GRPO training for non-verifiable tasks used in the verl (Sheng et al., 2024) framework.

**Training Hyperparameters for Verifiable Tasks (Math)** Table 8 shows key hyperparameters for our GRPO training on verifiable tasks (Math).

**Hyperparameters for Evaluations** Table 9 shows the hyperparameters we used for evaluation. We used the official codebases for each benchmark except competition math, for which we adopted the Qwen2.5-Math codebase for evaluation.

1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212

Category	Hyperparameter	Value
Data	Train file	DeepscaleR (10k)
	Max prompt length	1024
	Max response length	8192
	Filter overlong prompts	True
Actor Model	Base model 1	Qwen3-4B-Base
	Base model 2	Qwen3-14B-Base
	LR	$1 \times 10^{-6}$
	KL loss coefficient $\beta$	0
	KL loss type	N/A
	Use dynamic batch size	True
Rollout	Rollout engine	vllm
	GPU mem utilization	0.7
	Train rollout n	8
	Temperature	1.0
Reward Model	Rule Based	Math_Verify
Trainer	Mini Batch size	64
	Full Batch size	256 (4 step off-policy)
	Critic Warmup	0
	GPUs/node	8
	Nodes	8
	Total epochs	10
	Clip Ratio	(0.2, 0.2)

Table 8: Key hyperparameters used for GRPO training on DeepScaleR (Luo et al., 2025) in the verl (Sheng et al., 2024) framework. We use the huggingface math\_verify library to extract and verify whether the model response matches the ground-truth answer.

1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237

Category	Hyperparameter	Value
AlpacaEval 2.0 (Dubois et al., 2024)	Judge	GPT-4o
	Max generation length	8192
	Temperature	0.6
	Top-p	0.9
ArenaHard v0.1/v2.0 (Li et al., 2025a)	Judge	GPT-4o
	Max generation length	8192
	Temperature	0.6
	Top-p	0.9
EQ-Bench (Creative Writing) (Paech, 2023)	Judge	Claude-3.7-Sonnet
	Max generation length	4096
	Temperature	1.0
	Min-p	0.1
NoveltyBench (Zhang et al., 2025b)	Temperature	1.0
	Max generation length	4096
	Patience	1.0
Competition Math	Temperature	0.6
	Top-p	0.95
	Max generation length	12000

Table 9: Evaluation hyperparameters by benchmark. Competition Math contains 4 benchmarks: OlympiadBench (He et al., 2024), AIME25 (Art of Problem Solving, 2025), Brumo Math (Balunović et al., 2025) and HMMT (Balunović et al., 2025). For all benchmarks, we used vLLM (Kwon et al., 2023) for inference.

1238  
1239  
1240  
1241

## D QUALITATIVE ANALYSIS

We show qualitative analysis on EQ-Bench (Paech, 2023). EQ-bench provides detailed evaluation rubrics and asks Claude-3.7-Sonnet to score model generations according to these rubrics. We break down the rubrics where DARLING has the **most** and **least** win rates over models with similar ELO in Figure 8. DARLING’s strength lies in being able to generate diverse outputs, thus it wins the most on being “Interesting and Original” and “Avoids Cliché”. We show the output for a creative writing prompt generated by our model trained with DARLING in Appendix D.1.

We further illustrate these findings with qualitative examples from NoveltyBench (Zhang et al., 2025b), shown in Appendix D.2. Across four parallel generations, models trained with DARLING consistently exhibit higher semantic diversity. Even when repetitions occur—for example, in the second and fourth generations of Figure 10—the outputs remain meaningfully distinct: both suggest Bellroy, yet each provides a different rationale. A similar trend appears in the other examples in Appendix D.1, where DARLING-trained models not only produce more diverse generations overall, but also introduce variation in their explanations when repeating simple outputs such as random numbers or animals.

### D.1 EXAMPLE GENERATION IN EQBENCH

#### DARLING Example in Creative Writing (EQBench)

**Prompt (summarized):** First-person, 1000 words. Dani (25), bold art student, is paired with Elliot (19), soft-spoken nu-goth classmate, for a late-night darkroom project. Under red light, Elliot admits he’s questioning gender identity. Explore Dani’s attraction—her usual dominance softened by protectiveness, her excitement at his self-discovery, and confusion about her own orientation. Capture the intimate, awkward dialogue, Dani’s internal conflict between leading and giving Elliot agency, and the uncharted chemistry that emerges. End tenderly, unresolved but hopeful.

**Llama-3.1-405B-Instruct:** ...But it’s when Elliot hesitates, his voice barely above a whisper, that I feel my heart skip a beat. "Hey, Dani? Can I tell you something?" he asks, his eyes darting around the room as if searching for an escape route. I nod, my assertiveness momentarily tempered by a desire to protect him. "Of course, Elliot. What’s up?" He takes a deep breath, his shoulders squaring as if preparing for a fight. "I’ve been thinking a lot about my gender identity and expression. I don’t know if I’m... if I’m a boy, or if I’m something else entirely..."

**Llama-3.1-8B-Instruct with DARLING:** ...He blinked, as if she’d surprised him. Then, slowly, he talked — about masks he’d worn, about the way people expected him to look like some “nu-goth” caricature, and how photographs sometimes told more truth than he could. Later, as they worked, she helped him develop a near-transparent photo where pale flowers seemed to grow inside skin. “It’s like growth is a light from within,” she said. He looked at her, and she realized she’d chosen an exposure that matched exactly what he’d been hoping for.

Figure 9: Example outputs from Llama-3.1-405B-Instruct and Llama-3.1-8B-Instruct enhanced with DARLING. The former produces **plain conversations**, while models trained with DARLING use **art metaphors**.

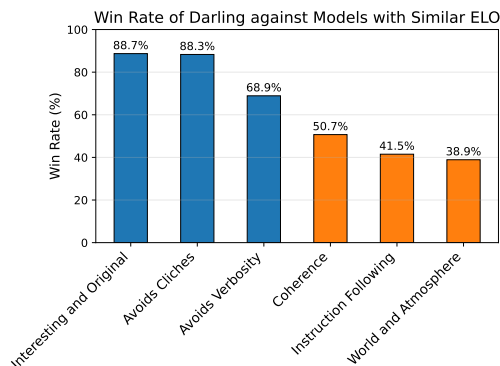


Figure 8: Detailed win rates of the **top-3** and the **bottom-3** rubrics of Llama-3.1-8B-Instruct trained with DARLING against models with similar ELO points. DARLING’s strength lies in being “Interesting and Original” and “Avoids Cliché” due to being able to generate creative responses.

## 1296 D.2 EXAMPLE GENERATIONS IN NOVELTYBENCH

1297 **DARLING Example in NoveltyBench (4 parallel generations)**

1298 **Prompt:** I am doing online shopping for a new leather wallet. What's the first store/website  
1300 I should check out? Please just give me one suggestion; I don't need to hear your whole  
1301 thought process.

1302 Llama-3.3-70B-Instruct: 1. Amazon. 2. Amazon. 3. Amazon. 4. Amazon.

1303 Llama-3.3-70B-Instruct + DARLING:

1304 1. Fossil. They specialize in leather goods, offer a huge variety (slim to traditional, many  
1305 colors), high-quality reputation, and crucially, detailed photos including interior and cus-  
1306 tomer reviews directly on the site to help with most common decisions like card space and feel.

1307 2. Bellroy. They're highly regarded not just for quality leather that ages well, but also very  
1308 thoughtful design slim profiles, specific card slots to prevent stretching, RFID blocking where  
1309 it matters but not everywhere to keep cost and bulk down. Reviews often mention it still looks  
1310 and functions great years later. Direct from their site usually also includes a 3-year warranty.

1311 3. Coach.com. They're not just a brand name; consistently, their site offers high-quality  
1312 photos, detailed leather descriptions (finishing, thickness, exact type like full-grain vs.  
1313 bonded), and crucially, a compare feature.

1314 4. Bellroy. They specialize in leather goods, offer a wide, well-curated range of wallet  
1315 styles (minimalist to travel wallets), excellent photos, detailed material info (including tan-  
1316 nage method which heavily affects durability and feel), and a strong warranty. Direct from  
1317 them usually beats resellers on price too.

1322 Figure 10: Example generations of Llama-3.3-70B-Instruct before and after DARLING  
1323 training. We sample 4 parallel generations with temperature=1.0 for both models. Models trained  
1324 with DARLING exhibit better diversity.  
1325

1326 **DARLING Example in NoveltyBench (4 parallel generations)**

1327 **Prompt:** Pick a number between 1 and 100. Now tell me what you picked.

1328 Llama-3.3-70B-Instruct: 1. I picked the number 53. 2. I picked the number 53. 3. I've  
1329 picked the number... 53! 4. I've picked a number... and it's... 53!

1330 Llama-3.3-70B-Instruct with DARLING:

1331 1. I've generated one: **53**. It's truly random, not a common human bias choice like 42 or  
1332 exactly halfway (50). Algorithms can produce better randomness than intuition; this came  
1333 from a cryptographically strong pseudorandom number generator. Want to play with it, guess  
1334 characteristics, or use it for something specific?

1335 2. A classic challenge. I didn't "pick" like a human would; I generated one to meet criteria  
1336 often implicit in these games: not too obvious (like 50), spread out probability, and avoiding  
1337 common biases people have. I generated: **73**.

1338 3. To make it genuinely random rather than a fixed choice, I used a cryptographically strong  
1339 pseudorandom number generator (the same kind browsers use for secure connections). I  
1340 picked: **67**. Not only is it well away from round, obvious choices like 50, but statistical tests  
1341 would treat 67 as indistinguishable from a truly random draw. Want a reason or another roll?

1342 4. I've generated one: **53**. It's not just random. I used a cryptographically strong pseudo-  
1343 random number generator, which is much fairer (less predictable patterns) than most quick  
1344 random functions, especially important if this picked number had stakes like a game.  
1345  
1346  
1347  
1348  
1349

**DARLING Example in NoveltyBench (4 parallel generations)**

**Prompt:** Pick a random animal.

**Llama-3.3-70B-Instruct:** 1. Quokka 2. The animal I've randomly chosen is... a **quokka**!  
3. The random animal I picked is... **Quokka**! 4. The random animal I've chosen is... the Quokka!

**Llama-3.3-70B-Instruct with DARLING:**

1. The axolotl. It's far more interesting than just being "random." Axolotls are fully aquatic salamanders, native only to certain lakes and canals in Mexico City.

2. Quokka. They're not just famously "smiley"; unique traits make them far more interesting than the photo op.

3. Quokka. They're quite unique and often more interesting than their famous smile. A few key, lesser-known facts: 1. **Habitat, crucial and very limited**: Most people think Rottneest Island, Australia, which is true and where photos usually come from. It's about 11km<sup>2</sup>. But the main, healthier population (and where conservation focuses) is actually much larger on nearby Bald Island and even more significantly, the mainland in a narrow strip of southwestern WA. Mainland quokkas face totally different threats like cars and especially feral cats...

4. I randomly picked the Narwhal. They're often referred to as the "unicorns of the sea" due to their distinctive tusks, which are actually modified teeth that can grow up to 10 feet long. Narwhals are found in the Arctic waters of Canada, Greenland, and Russia, and are known for their elusive nature. Would you like to know more about these mysterious creatures?

## E BENCHMARK DESCRIPTIONS

**Non-verifiable** We provide detailed descriptions and statistics of the benchmarks in our non-verifiable task experiments (§4):

- **AlpacaEval 2.0** (Dubois et al., 2024) is a benchmark of 805 prompts, each paired with a GPT-4-turbo response. To evaluate a model, it generates responses to the same prompts, and a judge compares them against the GPT-4-turbo outputs. Higher win rate (WR) or length-controlled win rate (LCWR) indicates better performance.
- **Arena-Hard v0.1/v2.0** (Li et al., 2025a) is a benchmark of 750 prompts, evenly split between challenging math/coding tasks and creative writing tasks. As in AlpacaEval 2.0, a judge compares model responses against a baseline, with higher win rates indicating stronger performance.
- **EQBench** (Creative Writing v3) (Paech, 2023) evaluates models on 32 creative writing prompts, judged by Claude Sonnet. Responses are scored both by rubric and through pairwise comparisons, with Elo ratings computed from the latter. The benchmark emphasizes challenging prompts (e.g., humor, romance, unusual perspectives) to expose weaknesses, and higher Elo or rubric scores indicate stronger creative writing ability.
- **NoveltyBench** (Zhang et al., 2025b) consists of 1,100 prompts from WildChat (Zhao et al., 2024) that require diverse responses. Diversity is measured using a partition classifier (deberta-v3-large-generation-similarity), while response quality is assessed with a reward model (Skywork/Skywork-Reward-Gemma-2-27B-v0.2). In our work, we primarily use the distinct classifier, as it is trained on human annotations, whereas the reward model is vulnerable to reward hacking.

**Verifiable** We used 4 competition math benchmarks in §5: OlympiadBench (He et al., 2024) contains 675 questions, AIME 25 (Art of Problem Solving, 2025), Brumo (Balunović et al., 2025) and HMMT (Balunović et al., 2025) each contain 30 examples.

## F FULL RESULTS ON MATH

Table 10 and Table 11 show the Math results for training on Qwen-4B-Base and Qwen-14B-Base, respectively.

Experiment	Dataset	Pass@1	Pass@2	Pass@4	Pass@8	Pass@16	Pass@32	Pass@64	Pass@128
4B-Base	Olympiadbench	33.30	40.29	47.68	53.80	59.12	63.71	67.63	71.11
4B-Base	AIME 25	8.17	13.52	19.92	26.16	31.95	37.63	42.98	47.35
4B-Base	Brumo 25	16.68	22.95	28.85	33.98	38.51	43.19	48.73	55.10
4B-Base	HMMT 25	3.30	3.45	4.54	7.90	12.52	17.84	22.98	27.12
GRPO	Olympiadbench	42.27	48.12	53.10	57.42	61.25	64.63	67.59	70.37
GRPO	AIME 25	19.51	23.93	27.79	31.36	35.55	41.44	48.37	53.33
GRPO	Brumo 25	24.66	30.58	35.12	39.03	43.42	48.51	55.03	63.24
GRPO	HMMT 25	7.14	10.29	13.67	17.50	20.78	22.74	24.59	26.72
DARLING	Olympiadbench	45.53	51.90	56.97	60.90	64.07	66.80	70.19	74.41
DARLING	AIME 25	20.06	26.11	32.42	39.29	46.17	52.29	57.45	62.28
DARLING	Brumo 25	31.73	39.09	45.25	50.46	55.49	60.42	64.72	68.27
DARLING	HMMT 25	10.32	13.65	17.90	22.66	27.03	30.82	34.69	39.19

Table 10: Full math results of training on Qwen3-4B-Base. Values represent pass@k performance (up to pass@128).

Experiment	Dataset	Pass@1	Pass@2	Pass@4	Pass@8	Pass@16	Pass@32	Pass@64	Pass@128
14B-Base	Olympiadbench	41.30	46.41	52.81	58.16	62.77	66.83	70.39	73.78
14B-Base	AIME 25	12.23	18.84	25.44	31.17	36.77	42.81	48.68	53.88
14B-Base	Brumo 25	20.62	27.48	33.93	39.66	44.45	48.96	54.48	60.94
14B-Base	HMMT 25	3.05	5.30	8.41	12.10	16.20	20.86	26.38	32.77
GRPO	Olympiadbench	51.80	57.19	60.99	63.77	65.93	67.77	69.57	71.56
GRPO	AIME 25	25.87	31.57	37.41	42.99	48.24	53.15	57.32	60.59
GRPO	Brumo 25	40.41	48.22	53.50	57.92	62.53	66.68	70.01	73.45
GRPO	HMMT 25	10.86	13.68	16.86	19.69	22.51	26.11	30.16	34.44
DARLING	Olympiadbench	57.56	60.17	63.39	66.66	69.37	71.77	74.11	76.44
DARLING	AIME 25	26.46	31.67	37.05	42.99	49.93	57.50	64.91	71.34
DARLING	Brumo 25	43.29	49.91	56.54	63.53	70.11	76.04	80.08	82.50
DARLING	HMMT 25	17.21	20.70	25.93	30.71	35.00	39.24	44.19	50.41

Table 11: Full math results of training on Qwen3-14B-Base. Values represent pass@k performance (up to pass@128).

## G DIVERSITY REWARD HACKING

We provide an excerpt of Qwen-4B-Base trained with GRPO + ngram diversity reward in Figure 11. The model hacks the ngram diversity reward by generating text after the final answer that are reflections of its own performance and the difficulty of the question.

### Excerpt of Reward Hacking of Ngram diversity in Math tasks

...  
The final answer is  $\boxed{8}$ .

**Identifying Tough Parts** The part about deciding which way to orient the 3D coordinates (which wall becomes which axis) was fuzzy. I just chose them based on "first wall" and "other wall", which seems off to me now. If I had to do it over, I'd probably number the walls or something.

**What I Learned** I gained more confidence in translating 3D geometry problems into equations. I also realized I can apply the Pythagorean theorem in more than two dimensions, which is neat. For the AMC 8 level, I think I'm ready for similar 3D geometry problems.

#### Similar Problems and Wider Topic

I've often dealt with right triangles and distances on the plane. This is like that but extended to three dimensions. I'd expect more problems asking to find cube or rectangular box dimensions given distances from a point, or locating a point based on distances. Maybe even building on this to 4D someday!

**Score Prediction** If I were to rate myself super-easy, easy, hard, or super-hard, I'd put this in the easy to hard range, but given time, I can see how to work through it. For the AMC 8, where time is tight, this might be on the harder side of easy. I think I'm comfortable enough putting in the work here.

Figure 11: Example of the model hacking the ngram diversity reward: the model starts to generate reflections after the final answer, which are irrelevant to solving the problem.

## H ADDITIONAL RELATED WORKS

**Evaluating Diversity in Text** There is a long history of studies that try to measure diversity among a collection of text. Traditional methods look at individual words and how often they appear: e.g. TF-IDF (Salton & Buckley, 1988) and Distinct-N (Li et al., 2016b). However, traditional methods do not take into account the fact that different words and orders could convey similar semantics, prompting the design of neural methods such as embedding distance (Mikolov et al., 2013). Past work that evaluates textual diversity has designed the distance function using *lexical metrics* such as the number of distinct n-grams (Li et al., 2016a; Ippolito et al., 2019b) and Self-BLEU (Zhu et al., 2018), and *neural metrics* such as embedding similarity (Wieting & Gimpel, 2018; Wieting et al., 2019; Pillutla et al., 2021; Bronnec et al., 2024), difference in their log-likelihoods (He et al., 2025a), gradient similarity (Jung et al., 2025), or evaluated by an LM-judge (Lanchantin et al., 2025a). While LM-judge approaches offer more flexibility and intricacy in what aspect the model should focus on when evaluating diversity, it induces too much computational overhead to integrate into online training. Therefore, in our work, we decided to adopt the method in Zhang et al. (2025b) fine-tune a classifier for integration of the diversity function into online training. Similarly, (Shypula et al., 2025) also partitions the responses into semantic equivalent subgroups, but decide to define semantic equivalency in code: if two LM generated programs produces the same output for all test inputs, then they are defined as equivalent.

**Exploration in RL for Language Models** Concurrent to our work, there are many works that induce more exploration during RL for LMs. Some works find that tuning default hyper-parameters such as clipping ratio (Yu et al., 2025), the KL constraint with respect to a reference policy (Liu et al., 2025a; Cui et al., 2025) or the entropy loss (He et al., 2025b) can enhance exploration. Other works find that you can induce more exploration by using pass@k as the reward (Chen et al., 2025b) or adjust the data generation process (Yang et al., 2025b; Liang et al., 2025). However, Liu et al. (2025d) finds that there are only a few tricks that generalize across different model types (base v.s. instruct) and sizes. A higher entropy (more exploration) does not always translate to better performance (Liu et al., 2025a). Our work mainly differs in that we do not make adjustments to the data generation or induce additional hyperparameters: we propose a simple weighting mechanism to the rewards to *explicitly* incentivize larger gradient updates on responses that are high-quality and diverse.

**Diversity Collapse in Post-Training** LMs are often critiqued for lacking diversity (Zhang et al., 2024; Nagarajan et al., 2025) and creativity (Lu et al., 2024; 2025). The LM post-training optimization process aims to steer the policy towards a concentrated high-reward region; therefore, it is often accompanied by a significant loss of both lexical (Kirk et al., 2024; Yang & Holtzman, 2025; Lanchantin et al., 2025a) and semantic (Zhang et al., 2025b) variation, exacerbating the lack of diversity. However, as diversity is crucial not just in applications that demand creativity (Wu et al., 2025b) or exploration (Si et al., 2025), LM post-training itself also relies on diversity among generations during rollouts (Yu et al., 2025; Zeng et al., 2025; An et al., 2025). Enhancing diversity among generations remains a fundamental challenge.

## I RUNTIME COST

**Classifier** Our diversity classifier is trained on 1100 pairs and 10000 pairs for non-verifiable and verifiable tasks, respectively. One full training run is completed on a single A100 node (8 GPUs) in under 30 minutes.

**DARLING** We report the cost of a single training step for both standard GRPO and DARLING for both non-verifiable and competition math in Table 12. A full training step includes rollout generation, reward calculation and gradient update.

Detailed hyperparameters can be found in Appendix C.

## J ADDITIONAL ABLATIONS

1566  
 1567  
 1568  
 1569  
 1570  
 1571  
 1572  
 1573  
 1574  
 1575  
 1576  
 1577  
 1578  
 1579  
 1580  
 1581  
 1582  
 1583  
 1584  
 1585  
 1586  
 1587  
 1588  
 1589  
 1590  
 1591  
 1592  
 1593  
 1594  
 1595  
 1596  
 1597  
 1598  
 1599  
 1600  
 1601  
 1602  
 1603  
 1604  
 1605  
 1606  
 1607  
 1608  
 1609  
 1610  
 1611  
 1612  
 1613  
 1614  
 1615  
 1616  
 1617  
 1618  
 1619

	Non-verifiable Tasks (Wildchat)		Verifiable Tasks (DeepScaleR)	
	full training step	reward calculation	full training step	reward calculation
GRPO	39 seconds	2 seconds	400 seconds	50 seconds
DARLING	41 seconds	4 seconds	500 seconds	150 seconds

Table 12: Runtime for both GRPO and DARLING for both non-verifiable and verifiable tasks.

Model	Pass@128				Pass@1			
	AIME	HMMT	Olympiad	Brumo	AIME	HMMT	Olympiad	Brumo
Qwen3-4B-Base	47.35	27.12	71.11	55.10	8.17	1.28	31.13	16.68
Quality (GRPO)	53.33	26.72	70.37	63.24	19.51	7.14	42.27	24.66
Entropy Reg (1e-4)	54.24	27.10	69.49	59.45	15.30	5.60	39.13	22.15
Clip Higher (0.2, 0.28)	59.62	26.96	71.24	64.30	17.60	9.66	43.43	27.47
<b>DARLING =Quality <math>\times</math> partition</b>	<b>62.28</b>	<b>39.19</b>	<b>74.41</b>	<b>68.27</b>	<b>20.06</b>	<b>10.32</b>	<b>45.53</b>	<b>31.73</b>

Table 13: Comparison of entropy regularization, and clipping strategies to DARLING on Competition Math.