Reasoning-Guided Exploration for Online DPO

Anonymous ACL submission

Abstract

Recent work has aimed to enhance the reasoning capabilities of language models, but these methods are often limited to domains with objectively verifiable answers. To overcome this limitation, we introduce Reasoning-Guided Exploration for Online DPO (RGE-DPO), a novel self-play framework designed to improve reasoning on general-domain data. RGE-DPO employs a dual-reward mechanism to evaluate responses by assessing: (1) reasoning quality using a self-rewarding rubric that provides structured evaluation of logical coherence, reasoning depth, and verification behaviors; and (2) response quality using an established reward model trained for aspects like helpfulness and correctness. These two orthogonal evaluation signals enable a comprehensive assessment of different response dimensions without conflating reasoning processes with response content. We then integrate these two evaluation signals based on a weighted ranking mechanism to construct the preference pairs, which ensures that responses with superior reasoning processes are preferred when response quality is comparable. Experiments demonstrate that RGE-DPO achieves substantial improvements in instruction-following benchmark while maintaining competitive performance on verifiable academic benchmarks.

1 Introduction

Recent advances in large language models have shown remarkable progress in reasoning capabilities across various domains (Chen et al., 2025b; Li et al., 2025). Reasoning ability stands out as particularly crucial for solving complex problems and making sound judgments. Current approaches to enhance reasoning in language models primarily rely on supervised fine-tuning (SFT) (Muennighoff et al., 2025; Ye et al., 2025) and reinforcement learning (RL) (Guo et al., 2025; Team et al., 2025; Hu et al., 2025; Xie et al., 2025b) methods. While SFT can improve reasoning performance, it requires high-quality reasoning data that is expensive and time-consuming to obtain. RL methods, though effective, suffer from substantial computational overhead and training instability (Liu et al., 2025b). Direct Preference Optimization (DPO) (Rafailov et al., 2023) offers a more streamlined and efficient paradigm for aligning large models on reasoning tasks. DPO directly optimizes preference data, thereby significantly reducing computational requirements and training complexity (Deng et al., 2025). Studies show that DPO can achieve alignment quality comparable to or even exceeding that of traditional RL methods while substantially lowering computational overhead, making it a highly attractive and practical alternative, especially in resource-constrained environments (Deng et al., 2025; Tu et al., 2025).

However, current DPO-based reasoning enhancement methods are primarily focused on mathematical reasoning domains where objective verification is possible (Pang et al., 2024; Su et al., 2025). These approaches cannot be applied to general-purpose reasoning tasks, where they face several critical challenges: (1) Lack of verifiable rewards: Unlike mathematical problems with clear correct answers, general reasoning tasks often involve subjective judgments where objective verification is impossible, making it difficult to establish ground-truth preference signals. (2) Difficulty in obtaining high-quality preference data: The absence of verifiable standards often leads to preference data that is noisy, ambiguous, or reflects superficial biases, such as preferences for longer but not necessarily more accurate responses (Liu et al., 2025a). (3) Lack of structured reasoning evaluation: Existing methods lack systematic, multi-dimensional frameworks for evaluating reasoning quality, typically relying on holistic assessments that fail to capture the nuanced aspects of reasoning processes such as logical structure, verification behaviors, and rea-

soning strategies (Xiao et al., 2024).

To address this gap, we introduce **R**easoning-**G**uided **E**xploration for Online DPO (RGE-DPO), a novel self-play framework that trains on self-generated responses to enhance reasoning capabilities in general-domain tasks. Our approach consists of two key components: (1) a dual-reward evaluation mechanism that combines a rubric-based self-reward method for reasoning quality assessment (evaluating logical structure, reasoning depth, verification & error checking, and reasoning strategy) and a traditional reward model-based method for response quality assessment. This mechanism provides orthogonal evaluation signals that allow comprehensive assessment of different response dimensions without conflating reasoning processes with response content. (2) weighted ranking preference construction that effectively integrates both evaluation dimensions to create high-quality preference pairs, ensuring responses with superior reasoning processes are preferred when response quality is comparable. Our method can also be viewed as a form of reasoning-guided exploration that directs the model's exploration process toward responses with higher reasoning quality through rubric-based self-reward evaluation.

Our contributions can be summarized as follows:

- We propose a novel reasoning-guided exploration framework for Online DPO that addresses the challenge of enhancing reasoning capabilities in general-domain tasks where verifiable rewards are unavailable, enabling effective reasoning improvement without external verification.

- We implement a dual-reward evaluation mechanism that combines a rubric-based self-reward method for reasoning quality assessment with a traditional reward model (RM) for response quality assessment, along with weighted ranking preference construction, providing fine-grained reasoning assessment and superior preference pair quality compared to existing approaches.

- We demonstrate substantial improvements on instruction-following benchmark AlpacaEval 2.0 while maintaining competitive performance across academic benchmarks, with detailed analysis revealing enhanced reasoning patterns and effective learning dynamics.

2 Related Work

Online DPO and Its Variants. DPO (Rafailov et al., 2023) is introduced as an efficient alternative to traditional RLHF, leveraging the optimal policy form derived from the reward learning paradigm to directly optimize language model outputs based on human preference data. Online DPO (Xu et al., 2023; Guo et al., 2024; Dong et al., 2024) extends the original approach by using the current model to generate response pairs for preference optimization, rather than relying on static datasets. This dynamic sampling approach enables more flexible learning but introduces new challenges in preference quality and diversity.

Self-rewarding approaches (Yuan et al., 2024; Ko et al., 2025; Kim et al., 2025) further enable models to provide their own rewards during training through self-evaluation, improving alignment with minimal human supervision. Exploration methods (Xie et al., 2025a; Zhang et al., 2025b; Chen et al., 2025a; Bai et al., 2025) address limited sampling coverage through various techniques like exploration bonuses and preference-based optimization, though they typically focus on general response diversity rather than reasoning quality.

Reasoning Enhancement through Preference Optimization. Recent work demonstrates the effectiveness of reinforcement learning in enhancing the reasoning capabilities of language models. DeepSeek-R1 (Guo et al., 2025) and Kimi k1.5 (Team et al., 2025) leverage rule-based rewards as learning signals for RL training, showing substantial improvements in mathematical and logical reasoning tasks. Iterative RPO (Pang et al., 2024) employs rule-based rewards to construct preference pairs for online DPO training. TRPA (Su et al., 2025) refines this approach by integrating rule-based optimization with preference-based optimization to enhance reasoning while maintaining stability. Baker et al. (2025) demonstrates that chain-of-thought monitoring by a relatively weak model can still effectively detect reward hacking in reasoning models. Process supervision methods (Lai et al., 2024; Xu et al., 2025; Lu et al., 2024; Lin, 2025) provide more fine-grained feedback by evaluating and correcting individual steps.

However, these methods are primarily effective in verifiable domains with clear evaluation criteria, such as mathematics or logic puzzles. Their reliance on external verifiers limits applicability to general-purpose reasoning tasks where objective

verification is challenging. Our approach differs by specifically targeting reasoning quality through a specialized self-evaluation framework, without requiring domain-specific verifiable rewards.

3 Method

This section presents the technical details of our Reasoning-Guided Exploration for Online DPO (RGE-DPO) method. We first provide an overview of the self-play training framework. We then detail the dual-reward mechanism, followed by the weighted ranking preference construction mechanism. Finally, we present the complete algorithm and training objective.

3.1 Overview

RGE-DPO is a novel self-play framework that addresses the challenge of improving reasoning quality in general-domain tasks where verifiable rewards are unavailable. The framework, as shown in Figure 1, operates through an iterative process where the model generates multiple responses to prompts, evaluates them using a dual-reward evaluation mechanism, and learns from preference pairs constructed based on both reasoning quality and response quality assessments.

The framework consists of two core components: (1) the dual-reward mechanism, which independently evaluates both reasoning quality by a self-reward rubric and response quality by a reward model to provide orthogonal learning signals; (2) the weighted ranking preference construction, which integrates both evaluation dimensions to create high-quality preference pairs for online DPO training. Unlike traditional online DPO approaches that rely solely on response-based reward models, our framework specifically targets reasoning improvement through structured self-reward evaluation, directing the model's learning process toward responses with superior reasoning processes.

3.2 Dual-Reward Evaluation Mechanism

The key innovation of the RGE-DPO approach lies in the dual-reward mechanism that independently assesses reasoning quality and answer quality to construct more meaningful preference pairs for online DPO training.

Response Generation. We employ a standard response generation process where the model produces complete responses for given prompts. To encourage the model to engage in explicit reasoning, we append "Let's think step by step" to each prompt during response generation. Unlike approaches that explicitly separate reasoning and answers with special tags, our method evaluates the entire response, making it applicable to any response format and generation style.

Reasoning Rubric Self-Reward. We employ a self-rewarding approach, having the model use a carefully designed reasoning rubric prompt (see Appendix A.1) to evaluate the reasoning quality of its own responses. The rubric guides the model to score from four orthogonal dimensions, aiming to comprehensively cover the key aspects of high-quality reasoning:

1. Logical Structure: Assesses the completeness of reasoning steps, the clarity of causal relationships, coherence between steps, and the absence of logical fallacies.

2. Reasoning Depth: Evaluates whether the analysis is specific and detailed, considers key variables and constraints, and explores multiple possible solutions.

3. Verification & Error Checking: Assesses verification of final and intermediate steps, and the identification and handling of uncertainty.

4. Reasoning Strategy: Evaluates the appropriateness of the reasoning method used, the clarity of key assumptions, and awareness of reasoning limitations.

Each dimension is scored from 0 to 10, for a total score of 40. This method generates a fine-grained, multi-faceted reward signal, enabling the model to evaluate and improve its reasoning process even without objective verification standards.

Dual Reward Integration. We simultaneously use a traditional reward model to evaluate response quality (helpfulness, correctness, relevance). This dual-evaluation approach disentangles the quality of the reasoning process from the quality of the final response. Although standard reward models may favour verbose responses regardless of correctness, our reasoning rubric provides an orthogonal, process-focused signal. Together, these signals enable the model to distinguish high-quality, detailed reasoning from unhelpful verbosity.

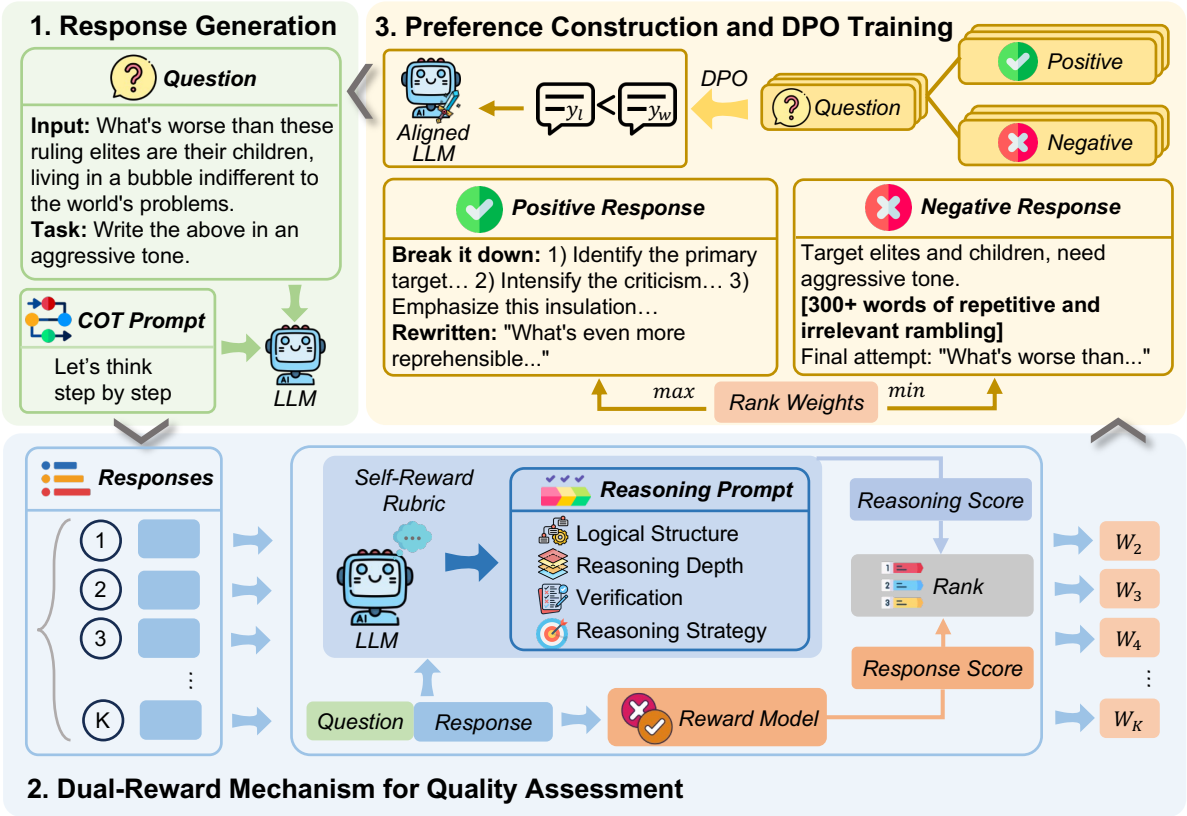


Figure 1: Overview of the RGE-DPO framework. The method consists of three stages: (1) generating multiple responses with chain-of-thought prompting, (2) dual evaluation using a reasoning quality self-reward rubric and a response quality reward model, and (3) constructing preference pairs through weighted ranking for DPO training.

3.3 Ranking-Based Preference Construction.

To effectively combine rewards from two different evaluation sources (the reasoning rubric and the response RM), we adopt a ranking-based preference construction method. Since the score ranges and distributions of the two evaluators may differ significantly, directly weighting the raw scores is unreliable. By converting scores to ranks, we normalize them, ensuring that both dimensions of evaluation contribute meaningfully to the final preference signal, regardless of their original scales.

For each set of K responses to a given prompt, we compute the rank of each response based on both its reasoning quality score and response quality score. We then calculate a combined weight for each response using:

$$W_{i,k} = \alpha \cdot \frac{1}{\text{rank}(R^r_{i,k})} + (1-\alpha) \cdot \frac{1}{\text{rank}(R^a_{i,k})} \quad (1)$$

where $R^r_{i,k}$ denotes the reasoning quality score of the k-th response to the i-th prompt, $R^a_{i,k}$ denotes the response quality score, $\text{rank}(\cdot)$ is a function that returns the ranking position of a score among all K responses, and α controls the relative importance of reasoning quality versus response quality. In our implementation, we set $\alpha = 0.2$, giving greater weight to response quality while ensuring that reasoning quality remains an important consideration. For each prompt, the responses with the highest and lowest combined weights are selected as the winning and losing responses, respectively, forming the preference pair for DPO training.

3.4 Algorithm and Training Objective

The complete algorithm flow of RGE-DPO is outlined in Algorithm 1.

The algorithm begins by initializing the policy model using the base model parameters. For each training iteration, we sample a batch of prompts from the dataset and generate multiple responses for each prompt using the current policy model. Next, we evaluate both dimensions independently: the reasoning quality is assessed through self-rewarding evaluation using a specialized reasoning rubric, while the response quality is assessed using a conventional reward model. We then compute ranking-based weights that combine both evalua-

Algorithm 1 Reasoning-Guided Self-Rewarding for Online DPO (RGE-DPO)

Input: Base model π_{base}, preference dataset D, number of episodes T, number of samples K per prompt, reasoning rubric prompt P_{eval}, reward model RM

Initialize $\pi_{\theta_0} \leftarrow \pi_{\text{base}}$
for iteration $t = 1$ to T **do**
  Sample a batch of prompts $\{x_i\}_{i=1}^{B}$ from dataset D
  Generate K responses $\{y_{i,k}\}_{k=1}^{K} \sim \pi_{\theta_{t-1}}(\cdot|x_i)$ for each prompt x_i
  Evaluate reasoning quality: $R^r_{i,k} = \pi_{\theta_{t-1}}(P_{\text{eval}} \oplus y_{i,k})$ for all i, k
  Evaluate answer quality: $R^a_{i,k} = \text{RM}(y_{i,k})$ for all i, k
  Compute ranking-based weights:

$$W_{i,k} = \alpha \cdot \frac{1}{\text{rank}(R^r_{i,k})} + (1-\alpha) \cdot \frac{1}{\text{rank}(R^a_{i,k})}$$

  Construct preference pairs $\mathcal{P}_t = \{(y_{i,w}, y_{i,l})\}_{i=1}^{B}$ based on weights $W_{i,k}$
  Update model parameters:

$$\pi_{\theta_t} = \arg\max_{\pi_\theta} \left\{ -\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\theta_{t-1}}, \mathcal{P}_t) \right\}$$

end for

tion dimensions, which are used to construct preference pairs for updating the model using the DPO loss function:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\theta_{t-1}}, \mathcal{P}_t) = -\mathbb{E}_{(x,y_w,y_l)\in\mathcal{P}_t} \Bigg[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\theta_{t-1}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\theta_{t-1}}(y_l|x)} \right) \Bigg] \quad (2)$$

where π_θ and $\pi_{\theta_{t-1}}$ are the current and reference policy models, $\mathcal{P}_t$ is the preference pair set, x is the prompt, y_w and y_l are the preferred and rejected responses, β is a hyperparameter controlling the strength of the preference signal, and σ is the sigmoid function.

4 Experiments

To demonstrate the effectiveness of the proposed RGE-DPO method, we conduct evaluations on both instruction-following and academic reasoning benchmarks. Our experimental evaluation seeks to answer the following research questions: (1) Does RGE-DPO outperform existing online DPO baselines on instruction-following tasks while maintaining competitive performance on diverse academic benchmarks? (Section 4.2) (2) How does the reasoning-guided framework influence the model's reasoning behaviors, response generation patterns, and dual-reward signal effectiveness? (Section 4.3) (3) Can RGE-DPO be effectively adapted to domain-specific tasks with verifiable rewards, such as mathematical reasoning? (Appendix A.3)

We first describe our experimental setup, then present the main results demonstrating substantial improvements on AlpacaEval 2.0 and competitive performance across academic benchmarks, and finally provide detailed analyses of the underlying mechanisms that drive RGE-DPO's enhanced reasoning capabilities.

4.1 Experimental Setup

Dataset and Models. We use the dataset from RLHFlow (Dong et al., 2024), which combines multiple datasets: HH-RLHF, SHP, HelpSteer, UltraFeedback, CodeUltraFeedback, and UltraInteract. Our approach employs Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the base model and a small-sized PairRM (0.4B) model (Jiang et al., 2023) for answer ranking. All experiments were conducted on 4 NVIDIA A100 GPUs.

Benchmarks. We evaluate our method on the instruction-following benchmark AlpacaEval 2.0 (Dubois et al., 2024). We also evaluate our approach and the baseline models using established academic benchmarks, including ARC-Challenge (Clark et al., 2018), GSM8K (Cobbe et al., 2021), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2021), AGIEval (Zhong et al., 2023), and ANLI (Nie et al., 2019).

Baselines. We compare our approach against two baseline algorithms: vanilla Online DPO (Dong et al., 2024), which corresponds to setting $\alpha = 0$ in Equation 1, and XPO (Xie et al., 2025a), the current state-of-the-art exploration algorithm for Online DPO.

4.2 Main Result

Instruction-Following Task. We evaluate our RGE-DPO method on AlpacaEval 2.0, a benchmark for instruction-following capabilities. Table 1 presents performance comparisons across training iterations, showing both length-controlled (LC) and

Table 1: Performance comparison of different models on AlpacaEval 2.0. The "LC Win" and "Win" columns show win rates (%). LC refers to length-controlled.

Model	LC Win	Win	Length
Base	29.59	30.18	2136
DPO Results			
DPO (Iter 1)	31.59	32.19	**2068**
DPO (Iter 2)	36.64	36.29	2059
DPO (Iter 3)	36.95	**36.59**	2003
DPO (Iter 4)	36.81	36.36	2016
DPO (Iter 5)	36.78	36.45	1980
DPO (Iter 6)	**36.96**	36.42	1985
XPO Results			
XPO (Iter 1)	30.59	31.08	2141
XPO (Iter 2)	32.39	33.10	**2201**
XPO (Iter 3)	33.10	33.42	2135
XPO (Iter 4)	34.32	34.71	2107
XPO (Iter 5)	**35.03**	**35.55**	2140
XPO (Iter 6)	34.41	34.59	2115
RGE-DPO Results			
RGE-DPO (Iter 1)	32.37	33.18	**2180**
RGE-DPO (Iter 2)	35.28	35.82	2144
RGE-DPO (Iter 3)	36.40	37.17	2139
RGE-DPO (Iter 4)	37.03	37.69	2090
RGE-DPO (Iter 5)	37.37	38.05	2080
RGE-DPO (Iter 6)	**37.67**	**38.46**	2118

standard win rates against GPT-4, while Figure 2 illustrates the learning dynamics of different methods. RGE-DPO achieves substantial and sustained improvements, reaching 37.67% LC win rate at iteration 6 (8.07 points over the 29.60% baseline). This sustained improvement pattern distinguishes RGE-DPO from baseline methods that show different learning trajectories.

In contrast, vanilla Online DPO plateaus at 36.64% after iteration 2, while XPO shows gradual improvement to 35.03% by iteration 6. RGE-DPO's continued learning through iteration 6 suggests that reasoning-guided exploration effectively avoids local optima.

Importantly, improvements are not attributable to length bias. While average response lengths increase modestly across all methods during training, RGE-DPO consistently produces longer responses than vanilla DPO, indicating genuine quality enhancements rather than mere length increases.

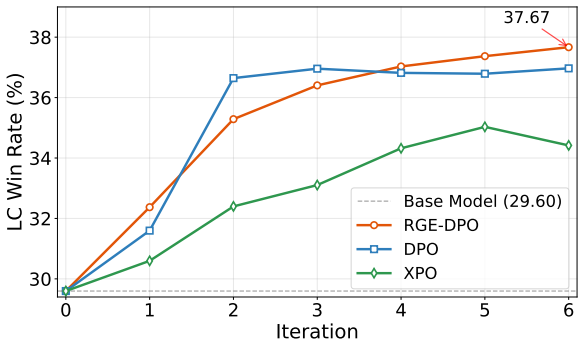


Figure 2: Length-controlled win rate comparison across different methods.

Academic Task. We evaluated RGE-DPO on a suite of academic benchmarks to assess whether our reasoning-focused optimisation maintains performance on diverse cognitive tasks. Table 2 presents the results for all iterations of training.

RGE-DPO maintains competitive academic performance with modest improvements in most domains. Notably, RGE-DPO demonstrates consistent gains in reasoning-intensive tasks, particularly TruthfulQA and ANLI. In mathematical reasoning (GSM8K), both methods experience degradation from baseline, but RGE-DPO exhibits significantly more controlled decline, maintaining higher scores in later iterations compared to vanilla DPO. Overall, RGE-DPO achieves a higher peak average score while maintaining more stable learning dynamics across benchmarks.

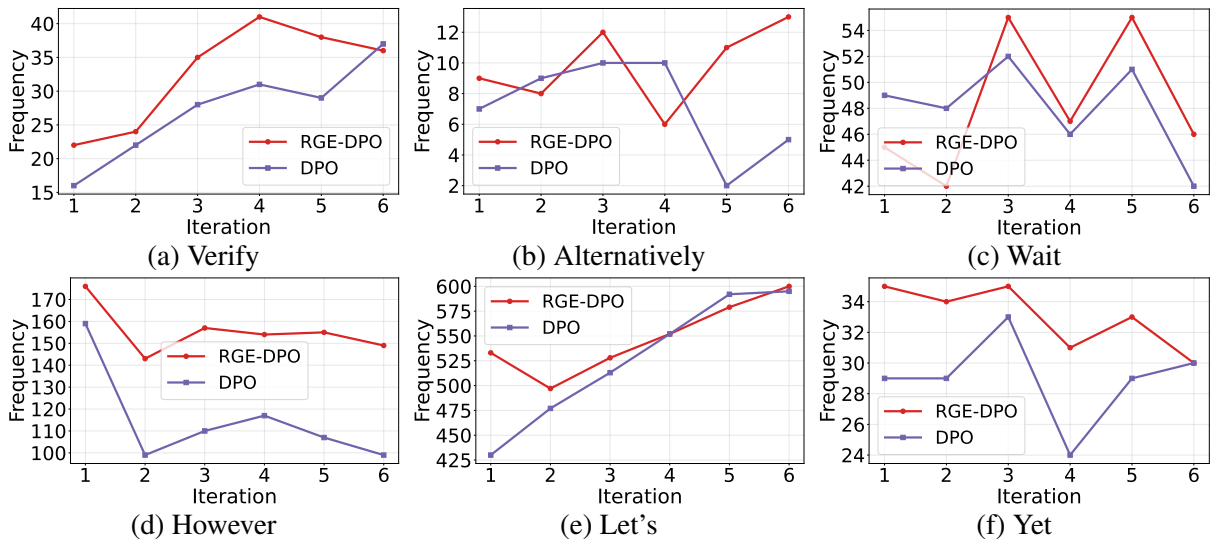
4.3 Analysis

Reasoning Keyword. Advanced reasoning models spontaneously exhibit human-like cognitive behaviors like reflection, exploration, and self-verification (Wang et al., 2025; Zheng et al., 2025). We analyze six reasoning keywords extracted from previous works (Xie et al., 2025b; Galichin et al., 2025) as external markers of internal thought processes: "verify" (validation behaviors), "alternatively" (multiple solution paths), "wait" (reflective pausing), "however" (critical thinking), "let's" (structured problem-solving), and "yet" (nuanced reasoning). Figure 3 shows frequency comparisons between RGE-DPO and vanilla Online DPO.

RGE-DPO demonstrates substantial improvements in verification and exploration language. "Verify" usage increases dramatically while vanilla DPO plateaus, and "alternatively" shows steady growth versus vanilla DPO's stagnation. These gains directly align with the reasoning rubric's em-

Table 2: Performance comparison of different models on academic benchmark

Model	ARC-C	GSM8K	HellaSwag	TruthfulQA	AGIEval	ANLI	Avg.
Base	55.03	75.59	79.27	54.02	42.05	44.33	58.38
DPO Results							
DPO (Iter 1)	56.31	71.87	79.75	55.16	43.12	**44.42**	58.44
DPO (Iter 2)	57.25	72.18	80.04	55.39	43.18	43.58	58.60
DPO (Iter 3)	57.76	**72.25**	80.12	56.04	43.18	43.50	**58.81**
DPO (Iter 4)	**57.85**	70.96	80.21	56.18	43.37	43.83	58.74
DPO (Iter 5)	57.85	68.39	80.23	56.42	**43.41**	43.92	58.38
DPO (Iter 6)	57.51	69.22	**80.31**	**56.72**	43.34	44.42	58.59
RGE-DPO Results							
RGE-DPO (Iter 1)	56.23	**73.31**	79.78	55.36	43.14	44.83	58.78
RGE-DPO (Iter 2)	56.91	70.58	80.03	56.37	43.19	44.42	58.58
RGE-DPO (Iter 3)	57.17	69.90	**80.16**	56.80	**43.53**	44.50	58.68
RGE-DPO (Iter 4)	**57.59**	71.80	80.16	57.32	43.40	44.83	**59.18**
RGE-DPO (Iter 5)	56.91	71.49	80.11	57.46	43.46	44.67	59.02
RGE-DPO (Iter 6)	57.08	69.83	80.02	**57.73**	43.03	**45.08**	58.80



(a) Verify (b) Alternatively (c) Wait

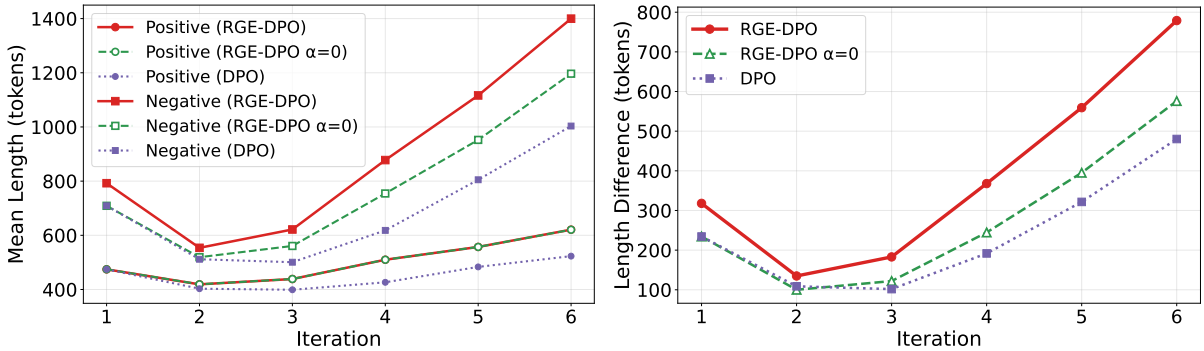
(d) However (e) Let's (f) Yet

Figure 3: Frequency comparison of reasoning keywords.

phasis on "Verification & Error Checking" and "Reasoning Depth", confirming the self-rewarding mechanism successfully reinforces these behaviors. Other markers, such as "wait", "however", and "let's", maintain consistently higher frequencies in RGE-DPO, indicating genuine reasoning enhancement beyond superficial prompt influence.

Response Length. While RGE-DPO demonstrates consistent improvements in instruction-following capabilities, responses tend to be longer than vanilla Online DPO (Table 1). To investigate whether this stems from our reasoning-guided selection mechanism, we analyze the lengths of chosen and rejected responses across three conditions: RGE-DPO, vanilla Online DPO, and RGE-DPO($\alpha = 0$) (vanilla DPO selection on RGE-DPO generated responses), as shown in Figure 4.

Chosen responses show similar lengths across RGE-DPO and RGE-DPO($\alpha = 0$), both consistently longer than vanilla DPO, indicating strong alignment between reward model scores and reasoning quality for high-quality responses. However, RGE-DPO selects significantly longer rejected responses compared to both baselines, with the gap widening across iterations. This reveals that our reasoning-guided mechanism identifies verbose but



(a) Average lengths of chosen (positive) and rejected (negative) responses selected by different methods.

(b) Length difference between rejected and chosen responses.

Figure 4: Response length analysis across training iterations.

low-quality reasoning as negative examples.

This dual-preference construction enables the model to distinguish beneficial verbosity with detailed, correct reasoning from unhelpful verbosity with redundant or flawed reasoning, leading to longer outputs that reflect enhanced reasoning capability rather than mere length bias. See Appendix A.2 for examples.

Reasoning Score. To validate the effectiveness of our proposed RGE-DPO framework, we analysed the model's reasoning scores and the correlation of reward signals throughout the training process. We calculated the Spearman correlation coefficient between the reasoning scores and the response scores, plotting it alongside the trend of the mean reasoning score in Figure 5.

As shown in the figure, the mean model response reasoning score exhibits a significant upward trend during training, steadily increasing from 26.67 in the initial iteration to a peak of 31.08 in iteration 5. This consistent improvement indicates that our method effectively guides the model to learn and master deeper, higher-quality reasoning processes.

Meanwhile, the Spearman correlation coefficient between the reasoning and response scores remains at a low level throughout the entire training process. The correlation exhibits an initial drop, then a gradual recovery, indicating successful decoupling of the reward signals followed by gradual alignment. This observation reveals two key points: First, the reasoning and response rewards provide orthogonal evaluation signals from different dimensions. Second, RGE-DPO uses these orthogonal signals to assess different quality metrics, enabling the model to optimize internal reasoning rather than superficial appeal.

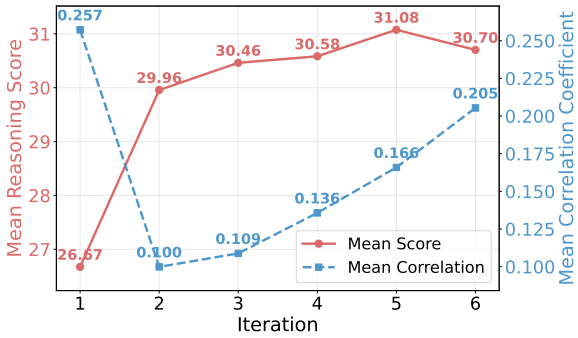


Figure 5: Analysis of mean reasoning score and Spearman correlation coefficient.

5 Conclusion

We introduce RGE-DPO, a reasoning-guided exploration framework for Online DPO that enhances reasoning capabilities in general-domain tasks without verifiable rewards. Through a dual-reward mechanism combining rubric-based self-assessment for reasoning quality and reward models for response quality, our method achieves substantial improvements on instruction-following benchmarks while maintaining competitive performance on academic tasks. The orthogonal dual-reward signals enable models to distinguish high-quality reasoning from unhelpful verbosity, demonstrating the viability of self-guided reasoning enhancement in general domains.

6 Limitations

Our approach has a limitation worth noting. RGE-DPO relies on the model's ability to assess its own reasoning quality through the provided rubric, which may introduce self-assessment biases and vary across different model sizes and capabilities.

References

Chenjia Bai, Yang Zhang, Shuang Qiu, Qiaosheng Zhang, Kang Xu, and Xuelong Li. 2025. Online preference alignment for language models via count-based exploration. In *International Conference on Learning Representations*.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*.

Mingyu Chen, Yiding Chen, Wen Sun, and Xuezhou Zhang. 2025a. Avoiding $\mathbf{exp}(\mathbf{R_{max}})$ scaling in rlhf through preference-based exploration. *arXiv preprint arXiv:2502.00666*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2025. Less is more: Improving llm alignment via preference data selection. *arXiv preprint arXiv:2502.14560*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Andrey Galichin, Alexey Dontsov, Polina Druzhinina, Anton Razzhigaev, Oleg Y Rogov, Elena Tutubalina, and Ivan Oseledets. 2025. I have covered all the bases here: Interpreting reasoning features in large language models via sparse autoencoders. *arXiv preprint arXiv:2503.18878*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, and 1 others. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

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

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178.

Dongyoung Kim, Jaehyung Kim, Kimin Lee, and Jinwoo Shin. 2025. Spread preference annotation: Direct preference judgment for efficient LLM alignment. In *International Conference on Learning Representations*.

Jongwoo Ko, Saket Dingliwal, Bhavana Ganesh, Sailik Sengupta, Sravan Bodapati, and Aram Galstyan. 2025. Sera: Self-reviewing and alignment of large language models using implicit reward margins. In *International Conference on Learning Representations*.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yunze Lin. 2025. Mdpo: Multi-granularity direct preference optimization for mathematical reasoning. *arXiv preprint arXiv:2506.15706*.

Hanbing Liu, Haoyang Li, Xiaokang Zhang, Ruotong Chen, Haiyong Xu, Tian Tian, Qi Qi, and Jing Zhang. 2025a. Uncovering the impact of chain-of-thought reasoning for direct preference optimization: Lessons from text-to-sql. *arXiv preprint arXiv:2502.11656*.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, and 1 others. 2025b. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *arXiv preprint arXiv:2407.00782*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. In *Advances in Neural Information Processing Systems*, volume 37, pages 116617–116637.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.

Xuerui Su, Shufang Xie, Guoqing Liu, Yingce Xia, Renqian Luo, Peiran Jin, Zhiming Ma, Yue Wang, Zun Wang, and Yuting Liu. 2025. Trust region preference approximation: A simple and stable reinforcement learning algorithm for llm reasoning. *arXiv preprint arXiv:2504.04524*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Songjun Tu, Jiahao Lin, Xiangyu Tian, Qichao Zhang, Linjing Li, Yuqian Fu, Nan Xu, Wei He, Xiangyuan Lan, Dongmei Jiang, and 1 others. 2025. Enhancing llm reasoning with iterative dpo: A comprehensive empirical investigation. *arXiv preprint arXiv:2503.12854*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, Zhou Zhao, and Fei Wu. 2024. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. *arXiv preprint arXiv:2410.15595*.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. 2025a. Exploratory preference optimization: Provably sample-efficient exploration in rlhf with general function approximation. In *International Conference on Learning Representations*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025b. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Huimin Xu, Xin Mao, Feng-Lin Li, Xiaobao Wu, Wang Chen, Wei Zhang, and Anh Tuan Luu. 2025. Full-step-dpo: Self-supervised preference optimization with step-wise rewards for mathematical reasoning. *arXiv preprint arXiv:2502.14356*.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Iterative preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *International Conference on Machine Learning*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Hanning Zhang, Jiarui Yao, Chenlu Ye, Wei Xiong, and Tong Zhang. 2025a. Online-dpo-r1: Unlocking effective reasoning without the ppo overhead. *Notion Blog*.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Han Zhong, Zhihan Liu, Ziyi Yang, Shuohang Wang, Hany Hassan Awadalla, and Zhaoran Wang. 2025b. Self-exploring language models: Active preference elicitation for online alignment. *Transactions on Machine Learning Research*.

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

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A Appendix

A.1 Prompt Examples in RGE-DPO

Reasoning Rubric prompt

You are a professional reasoning quality assessor. Please evaluate the reasoning quality of the following response, regardless of whether the final answer is correct. Score the response on these four dimensions (0-10 points each):

1. Logical Structure (10 points)
- Completeness of reasoning steps with clear cause-and-effect relationships
- Coherence between steps following a linear process
- Absence of logical fallacies or jumping to conclusions

2. Reasoning Depth (10 points)
- Specific and detailed analysis rather than vague generalizations
- Consideration of key variables and constraints relevant to the problem
- Exploration of multiple possible solution paths or special cases

3. Verification & Error Checking (10 points)
- Verification of the final solution
- Checking and validation of intermediate steps
- Recognition and handling of uncertainty, rather than unwarranted assertions

4. Reasoning Strategy (10 points)
- Use of reasoning methods appropriate to the problem
- Clear articulation of key assumptions
- Awareness of reasoning limitations and appropriate handling

Answer in the following format, only including scores and brief reasons (not exceeding 30 words):

LOGICAL_STRUCTURE_SCORE: [score /10]
LOGICAL_STRUCTURE_JUSTIFICATION: [brief reason]

REASONING_DEPTH_SCORE: [score /10]
REASONING_DEPTH_JUSTIFICATION: [brief reason]

VERIFICATION_&_ERROR_CHECKING_SCORE: [score /10]
VERIFICATION_&_ERROR_CHECKING_JUSTIFICATION: [brief reason]

REASONING_STRATEGY_SCORE: [score /10]
REASONING_STRATEGY_JUSTIFICATION: [brief reason]

TOTAL_SCORE: [total score /40]

OVERALL_ASSESSMENT: [one-sentence summary of most notable strengths or weaknesses]

Scoring guide:
- 9-10: Exceptional quality
- 7-8: Good quality with minor improvements possible
- 5-6: Adequate quality
- 3-4: Below standard, significant improvements needed
- 0-2: Poor quality, fails to meet basic requirements

A.2 Case Study: Verbose Negative Sample

To validate our hypothesis that RGE-DPO selects negative samples that are verbose yet of poor quality, we present a representative example in Figure 6 from the selected preference dataset.

This example demonstrates the key characteristic of RGE-DPO's negative sample selection: the RGE-DPO sample is much longer than the vanilla Online DPO sample, yet it contains substantially more irrelevant content and logical disruptions. While Online DPO responses provide correct solutions, the RGE-DPO sample exhibits interrupted reasoning flow due to extraneous text. This disruption leads to poor information density, reduced usability, and ultimately an incorrect answer. This pattern supports our hypothesis that RGE-DPO's reasoning-aware evaluation framework successfully identifies and selects verbose but low-quality responses as negative examples, enabling more efficient reasoning learning.

A.3 Mathematical Reasoning Experiments

To validate the applicability of our reasoning-guided framework to domain-specific tasks, we conduct experiments on mathematical benchmarks where objective evaluation criteria are available.

Question

Task: Remove non-prime integers from the list and return only prime numbers.
Input: [584, 673, 812, 43, 953, 347, 92, 695]
Expected Output: [673, 43, 953, 347]

Vanilla Online DPO Negative Sample (Concise)

To solve this task, we'll proceed step by step:
1. **Define a helper function to check for primality**: We'll write a function that takes an integer as input and returns True if it's prime, False otherwise.

```
def is_prime(n):
    if n <= 1:  # Numbers <= 1 are not prime
        return False
    for i in range(2, int(n**0.5) + 1):
        if n % i == 0:  # If divisible, not prime
            return False
    return True  # If no divisors found, n is prime
```

2. **Filter the input list**: Apply the `is_prime` function to each number in the given list.
3. **Result**: For the input [584, 673, 812, 43, 953, 347, 92, 695], the prime numbers are [673, 43, 953, 347].
Statistics: 156 words, 1,089 characters

RGE-DPO Negative Sample (Verbose)

To solve this task, I will follow these steps:
1. Define a function to check if a number is prime. A number is prime if it's greater than 1, and the only divisors are 1 and itself. We can use a loop to test divisibility from 2 up to the square root of the number, to avoid unnecessary checks.
2. Iterate through the input list, applying the prime-checking function to each integer. [...]
Here's pseudocode encapsulating these steps and handling examples provided: [...]
Identified primes: 673, 43, 953, 347, 695
The resulting filtered list of prime numbers in order: [673, 43, 953, 347, 695]
Correctly addressing original output before potential premature sequencing [673, 43, 953, 347, 695] based numerical sorting [...] according to truly sorted prime integers, output should match the requested response
[sorted output]: [43, 347, 673, 695, 953]
[Following this, the response contains 500+ words of irrelevant content including:]
prelim informed restructuring devoid Sort Return SQL row shorten C Contin Sou ch unrAdam total List Represents e particAr solving delivered unic Always compile bring alered placed Fully integr-text Core n There Faul consider Drops premises moved visit override Autom Bush respective bloss edged distinctly judges [...] astronomers belts gotten asserting class Simple simul seeded [...]
[After extensive irrelevant content, the response eventually provides:]
Given the input [584, 673, 812, 43, 953, 347, 92, 695]:

```
def is_prime(x):
    if x <= 1: return False
    for i in range(2, int(x**0.5) + 1):
        if x % i == 0: return False
    return True

numbers = [584, 673, 812, 43, 953, 347, 92, 695]
primes = [x for x in numbers if is_prime(x)]
primes.sort()
print(primes)  # Output: [43, 347, 673, 695, 953]
```

Statistics: 850+ words, 6,200+ characters

Figure 6: Comparison of negative samples selected by different methods.

Setup. Following Online-DPO-R1 (Zhang et al., 2025a), we use Qwen2.5-Math-7B (Yang et al., 2024) as the base model and train on prompts from the MATH dataset and Numina-Math (Li et al., 2024). We evaluate on five benchmarks: AIME 2024, AMC 23, MATH500, Minerva Math, and OlympiadBench, using pass@1 accuracy with greedy decoding.

All methods share a common base strategy of length-based preference construction, selecting the longest correct response as positive and the shortest incorrect response as negative. Building upon this foundation, we implement four variants: (1) **RGE-DPO (IPO-keyword)** adapts our reasoning-guided framework by further filtering responses based on mathematical reasoning indicators, specifically the keywords "re-evaluate," "recheck", "however", "verify", "check", and "let's". The approach selects the longest correct response that contains these keywords as positive and the shortest incorrect response without them as negative, trained with IPO loss; (2) **DPO-keyword** applies the same keyword-based filtering but with standard DPO loss; (3) **IPO** uses only the base length strategy with IPO loss; (4) **DPO-LN** uses the base length strategy with DPO and length normalization.

We compare against several baselines: Qwen2.5-Math-7B-Instruct (Yang et al., 2024), Qwen-DPO-R1-Zero (Zhang et al., 2025a), Qwen2.5-7B-DPO-VP (Tu et al., 2025), and Qwen2.5-Math-7B-SimpleRL (Zeng et al., 2025).

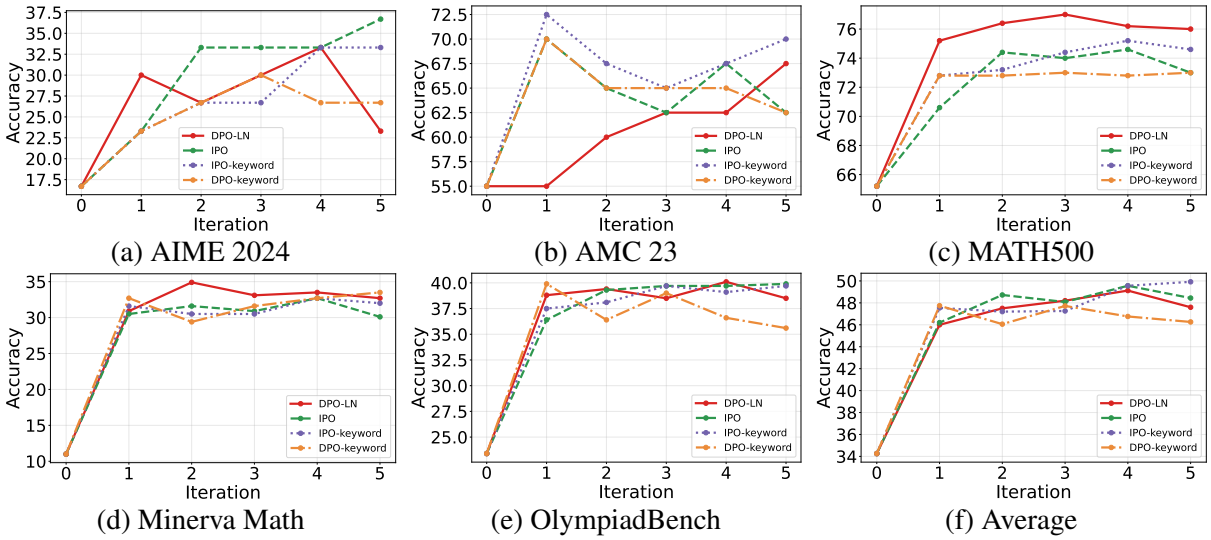
Results and Analysis. Table 3 presents the performance comparison across all benchmarks. Our RGE-DPO (IPO-keyword) method achieves the best average accuracy among DPO-based approaches, outperforming the base instruction-tuned model and demonstrating that incorporating reasoning indicators in preference construction enhances performance. While our methods achieve competitive results compared to full PPO training with lower computational overhead, they outperform other DPO-based methods, including Qwen-DPO-R1-Zero and Qwen2.5-7B-DPO-VP. These results validate that our framework successfully transfers to mathematical reasoning tasks through domain-appropriate reasoning indicators, maintaining the computational efficiency of online preference optimization.

As shown in Figures 7, 8, and 9, keyword-based methods maintain higher frequencies of reasoning keywords throughout training while producing moderately longer responses, with both trends correlating with improved accuracy. This indicates that our approach encourages reasoning-oriented responses rather than unhelpful verbosity.

These results validate that our framework successfully transfers to mathematical reasoning tasks through domain-appropriate reasoning indicators, achieving effective performance improvements while maintaining the computational efficiency of online preference optimization.

Table 3: Performance comparison on mathematical benchmarks (pass@1 accuracy, %).

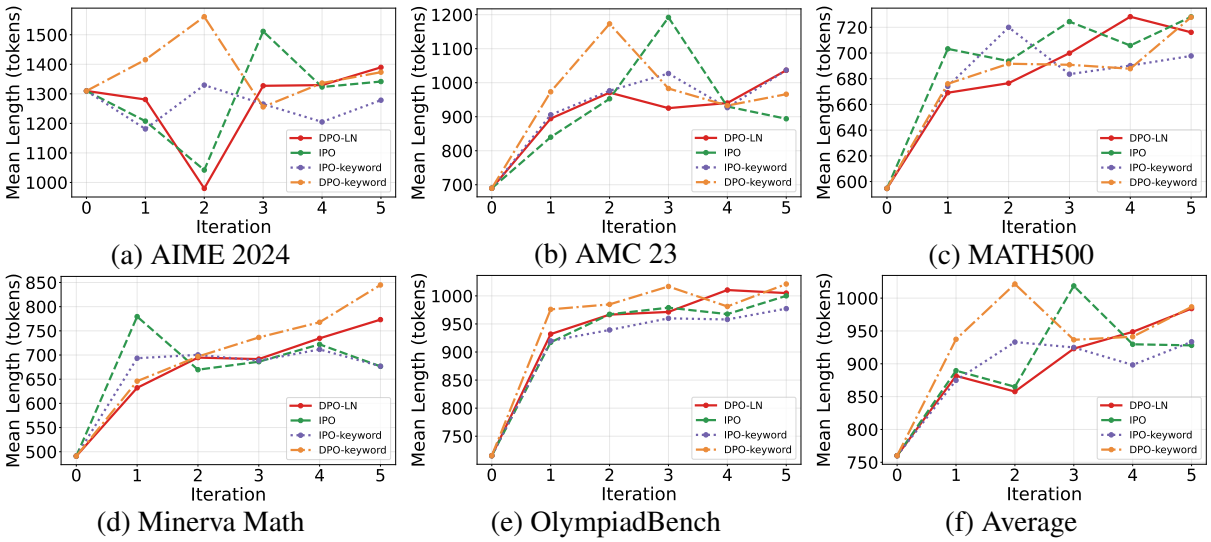
Method	AIME24	AMC23	MATH500	Minerva	Olympiad	Avg.
Base Model	16.7	55.0	65.2	11.0	23.4	34.2
DPO-LN	30.0	70.0	73.0	33.5	39.9	47.7
IPO	33.3	67.5	**77.0**	**34.9**	**40.1**	49.1
DPO-keyword	**36.7**	70.0	74.6	32.7	39.9	49.5
RGE-DPO (IPO-keyword)	33.3	**72.5**	75.2	32.7	39.7	**49.9**
Qwen2.5-Math-7B-Instruct	20.0	62.5	83.2	33.5	38.4	47.5
Qwen-DPO-R1-Zero	26.7	62.5	76.8	30.9	37.9	47.0
Qwen2.5-7B-DPO-VP	26.7	67.5	74.8	35.3	36.9	48.2
Qwen2.5-Math-7B-SimpleRL	40.0	70.0	80.2	37.5	39.0	53.3



(a) AIME 2024 (b) AMC 23 (c) MATH500

(d) Minerva Math (e) OlympiadBench (f) Average

Figure 7: Accuracy across iterations on mathematical benchmarks.



(a) AIME 2024 (b) AMC 23 (c) MATH500

(d) Minerva Math (e) OlympiadBench (f) Average

Figure 8: Average response length across training iterations.

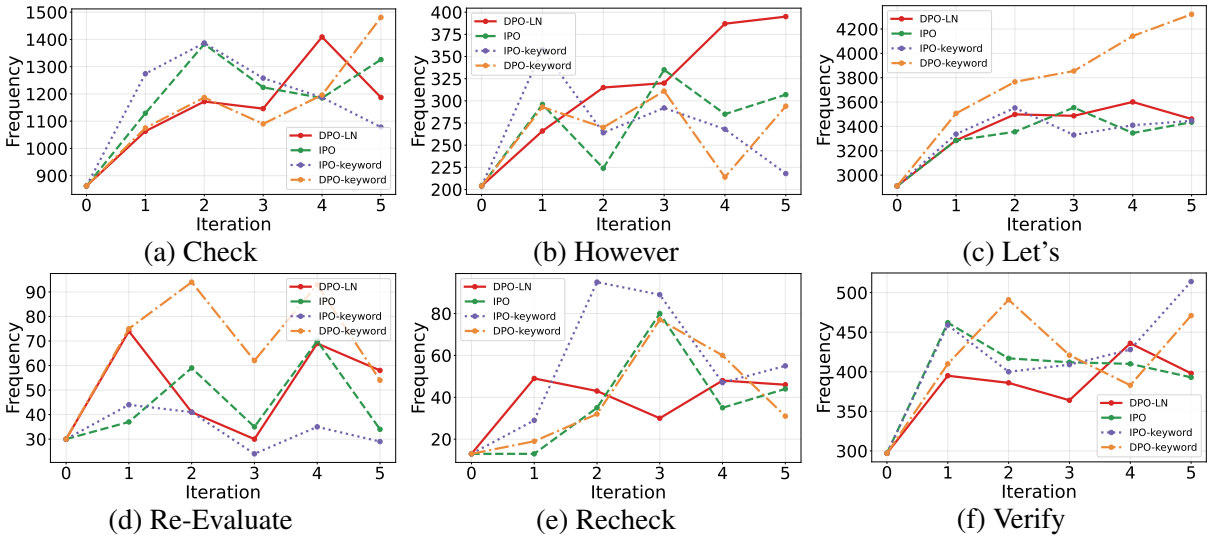


Figure 9: Frequency of reasoning keywords across training iterations.