Repository	Code Link	Unbiased?
trl (von Werra et al., 2020)	PPO Loss	✗
OpenRLHF (Hu et al., 2024)	PPO Loss	✗
verl (Sheng et al., 2024)	PPO Loss	✗
SimpleRL-Zero (Zeng et al., 2025)	PPO Loss	✗
Open-Reasoner-Zero (Hu et al., 2025)	PPO Loss	✗

Table 2: Many open-sourced PPO implementations contain length bias.

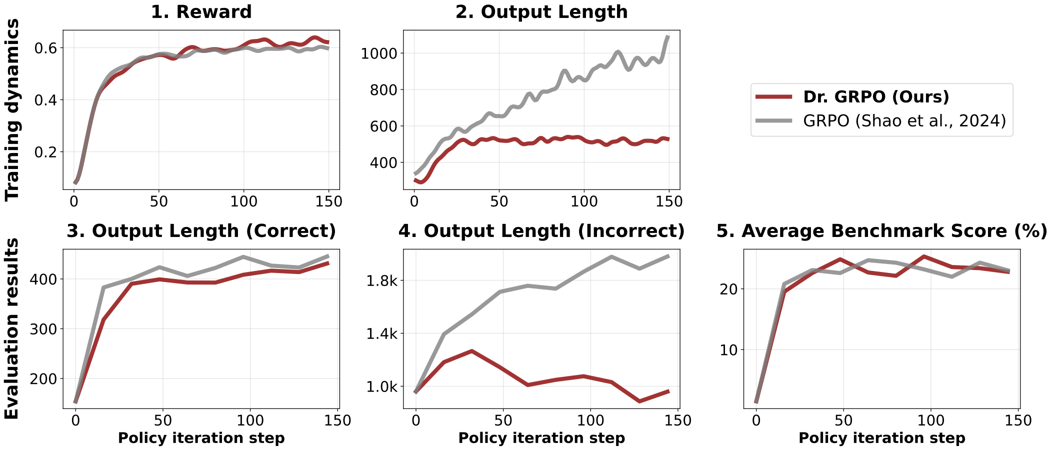

Figure 5: Comparison of Dr. GRPO and GRPO in terms of training dynamics (Top) and evaluation results (Bottom).

implement the unbiased optimization objective, we could replace the `mask.sum(axis=dim)` with a constant value (e.g., generation budget) in the `masked_mean` function in listing 1, as highlighted by the line in green. Notably, these simple modifications recover the PPO objective in Eq. (2), with the advantage estimated by Monte Carlo return with an unbiased baseline (Sutton & Barto, 2018). We give detailed derivations in Sec. A. We refer to our new optimization algorithm as **Dr. GRPO**. We next experimentally validate its effectiveness.

Experimental settings. We implement our algorithm using Oat (Liu et al., 2025a), a modular, research-friendly and efficient LLM RL framework. We adopt the Qwen2.5-1.5B base model and the R1 template (Template 1) for online RL-tuning. We implement the verification-based reward function using Math-Verify[2], with the following minimalistic rule:

$$R(\mathbf{q},\mathbf{o}) = \begin{cases} 1 & \text{if } \mathbf{o} \text{ contains the correct final answer to } \mathbf{q} \\ 0 & \text{otherwise} \end{cases}$$

We run RL on questions sampled from the MATH (Hendrycks et al., 2021) training dataset, and compare the vanilla GRPO with the proposed Dr. GRPO. We evaluate the online model on five benchmarks: AIME2024, AMC, MATH500, Minerva Math and OlympiadBench. More experimental details including hyperparameters can be found in Sec. G.

Results. We report various metrics in Fig. 5 to demonstrate that Dr. GRPO can effectively mitigate the optimization bias and lead to **better token efficiency**. In particular, we first note that both GRPO and Dr. GRPO exhibit similar trend to DeepSeek-R1-Zero (Guo et al., 2025), namely their response length increases along with training reward (Plots 1 & 2). However, we observe that GRPO tends to continually generate longer responses even when the reward improvement slows down (Plot 2). Although such a phenomenon is often referred to as the "emergence" of long-CoT through RL (Zeng et al., 2025; Hu et al., 2025), we argue that

[2] `https://github.com/huggingface/Math-Verify`.

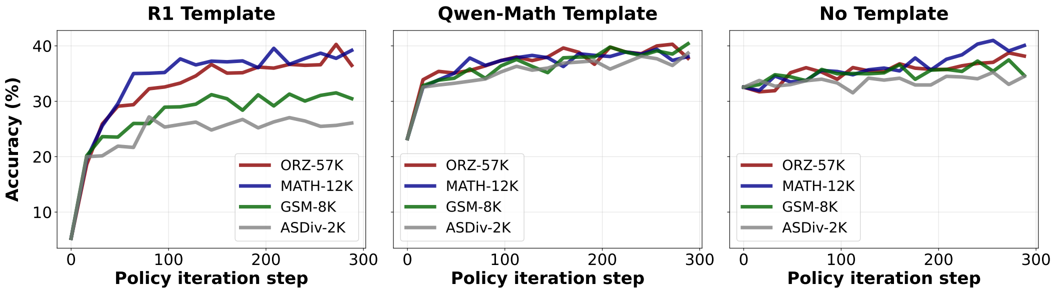

Figure 6: The average benchmark accuracy of different {template, question set} combinations during RL training.

it is also confounded by the response-level length bias (Sec. 3.1) during optimization[3]. In contrast, by computing the unbiased policy gradients, Dr. GRPO prevents the response length from growing wildly during training (Plot 2). Moreover, on evaluation benchmarks, the length of incorrect responses is substantially reduced by Dr. GRPO compared to the baseline (Plot 4), suggesting that an unbiased optimizer also **mitigates overthinking** (Chen et al., 2024).

3.3 A Duet of Template and Question Set Coverage in RL dynamics

Recall that the Qwen2.5-Math base models can readily answer questions with high accuracy without any prompt template (Sec. 2.2). Based on this intriguing observation, we are interested in how different templates affect the RL training. Furthermore, given the general belief that larger question set coverage leads to better performance (Luo et al., 2025; Hu et al., 2025), we also study the interaction between different templates and different levels of question coverage.

Experimental settings. Starting from the Qwen2.5-Math-1.5B base model, we apply R1 template, Qwen-Math template and No template respectively to run RL using Dr. GRPO. All experiments are repeated for different question sets that are detailed in Table 3.

Question set	#	Description
ORZ	57k	Combining AIME, Numina-Math, Tulu3 MATH; diverse and large amount
MATH	12k	High-school math competition questions
GSM	8k	Simpler grade-school math questions
ASDiv	2k	Basic algebra $(+ - \times \div)$ questions

Table 3: Different question sets that have different levels of difficulty and coverage.

Results. Fig. 6 shows the RL curves of different runs, from which we can make several interesting observations: **1)** Templates determine the performance of the initial policies, but RL can improve all policies to a comparable performance of $\sim 40\%$ (given a proper question set); **2)** When using the R1 template, question sets have a significant impact on the dynamics of RL, with too narrow coverage leading to lower plateau performance. However, when using the Qwen-Math template, the best final performance is attained by RL on GSM-8K, demonstrating that training on much simpler (and o.o.d.) questions can largely improve (nearly double) the test accuracy on harder questions. From these observations, we draw the following insights:

- The Qwen2.5-Math-1.5B base model already possesses strong math-solving capabilities (see the starting point in the right plot of Fig. 6). **Applying templates in fact destroys** the capability before RL reconstructs it. This implies that we should be more conservative in claiming the huge gains brought about by pure RL.

[3] We note that both Zeng et al. (2025) and Hu et al. (2025) employ PPO, which is unbiased by formulation. However, their loss implementations still introduce the length bias (see listing 1).

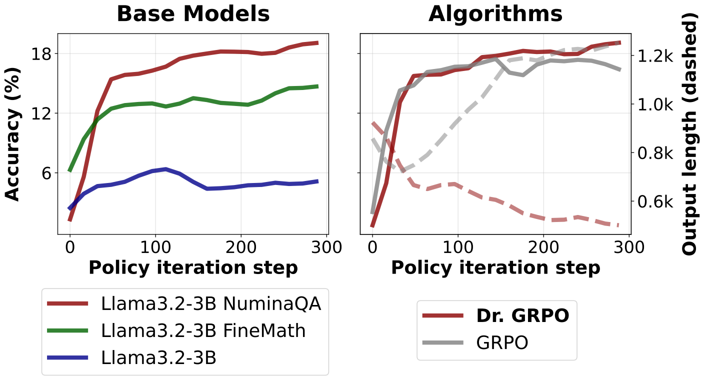

Figure 7: **Left**: The average benchmark performance curves of different base models. **Right**: The comparison between Dr. GRPO and GRPO with respect to reasoning accuracy (solid lines) and model response length (dashed lines).

- When there is a large **mismatch** between base models and templates (e.g., R1 template mismatches Qwen2.5-Math-1.5B), the policy improvement mainly comes from RL-tuning, thus requiring question set to have good coverage (left plot of Fig. 6). **Otherwise**, even a small and completely o.o.d. question set could induce the reasoning ability equally well, by **reinforcing useful reasoning behaviors instead of infusing new knowledge**.

3.4 Domain-Specific Pretraining Improves RL Ceiling

Recent successful R1-Zero-like replications of math reasoners mostly employ Qwen2.5 base models as the initial policies (Zeng et al., 2025; Cui et al., 2025; Hu et al., 2025), which are already strong math solvers and exhibit self-reflection patterns (Sec. 2.2 and 2.3). In this section we hope to explore the other side: **can R1-Zero-like training succeed on originally weak (in terms of math reasoning) base models?** We answer this question affirmatively, with the observation that *math pretraining would improve the ceiling of RL*.

Experimental settings. We adopt the Llama-3.2-3B base model as our starting point, and use the unbiased Dr. GRPO algorithm for RL-tuning with the R1 template. We hypothesize that domain-specific pretraining would help RL, hence we adopt the *Llama-3.2-3B-FineMath*[4], which is continual pretrained on the FineMath dataset (Allal et al., 2025). Moreover, as we hypothesize that Qwen2.5 models are likely to be pretrained on concatenated question-response texts (Sec. 2.2), we similarly prepare a concatenated dataset from NuminaMath-1.5 (Li et al., 2024b), and continual pretrain Llama-3.2-3B-FineMath for 2 epochs with learning rate 1e-5. We refer to the concatanated continual pretrained model as *Llama-3.2-3B-NuminaQA*.

Results. We present the RL curves of different base models in the left plot of Fig. 7. We observe that RL can even improve the vanilla Llama base model, but the gain is minimal. After continual pretraining (and concatenated continual pretraining) to embed math domain knowledge, Llama models can show much stronger RL performance, validating our hypothesis. We also revisit the GRPO's optimization bias with the Llama base model. The right plot of Fig. 7 compares the model performance and response length trained with GRPO and Dr. GRPO. We can clearly see that GRPO can produce the "double-increase" phenomenon, potentially leading to a **misperception** that long-CoT can also emerge on Llama models after math pretraining. Unfortunately, the increase of length might be due to the optimization bias (Sec. 3.1), which can be effectively mitigated by the proposed Dr. GRPO (Sec. 3.2 & right plot of Fig. 7).

4 Closing Remarks

We have taken a critical perspective to examine base models used for R1-Zero-like training, as well as algorithms used for RL. Through the analysis, we demystified how pretraining

[4] https://huggingface.co/HuggingFaceTB/FineMath-Llama-3B.