



Understanding R1-Zero-Like Training: A Critical Perspective

Dooyoung Kim

Natural Language Processing Lab, SKKU

Contents

1. Introduction

2. Preliminary

3. Method

4. Experiments

5. Conclusion

DeepSeek-R1-Zero

Base model에서 RL을 학습

- 기존의 RL은 Pretraining + SFT + RL
- DeepSeek-R1-Zero는 Pretraining + RL

RL을 진행할수록 Reasoning Capability 증가

- Response length 증가
- Aha moment 등장
 - 응답을 생성하는 과정에서 오류가 있는 경우
 - 스스로 오류를 수정하고 다시 reasoning 하는 현상

Table 2 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

<Source: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning>

DeepSeek-R1-Zero

○ R1-Zero에 적합한 Base 모델 분석

○ 다양한 Base 모델을 기반으로 분석 수행

- Question-Answering Ability: 주어진 질문에 대해 답을 하는가?
- Exploration Ability: 다양한 응답을 생성할 수 있는가?
- Self-Reflection: 스스로 답을 수정하는 패턴이 보이는가?

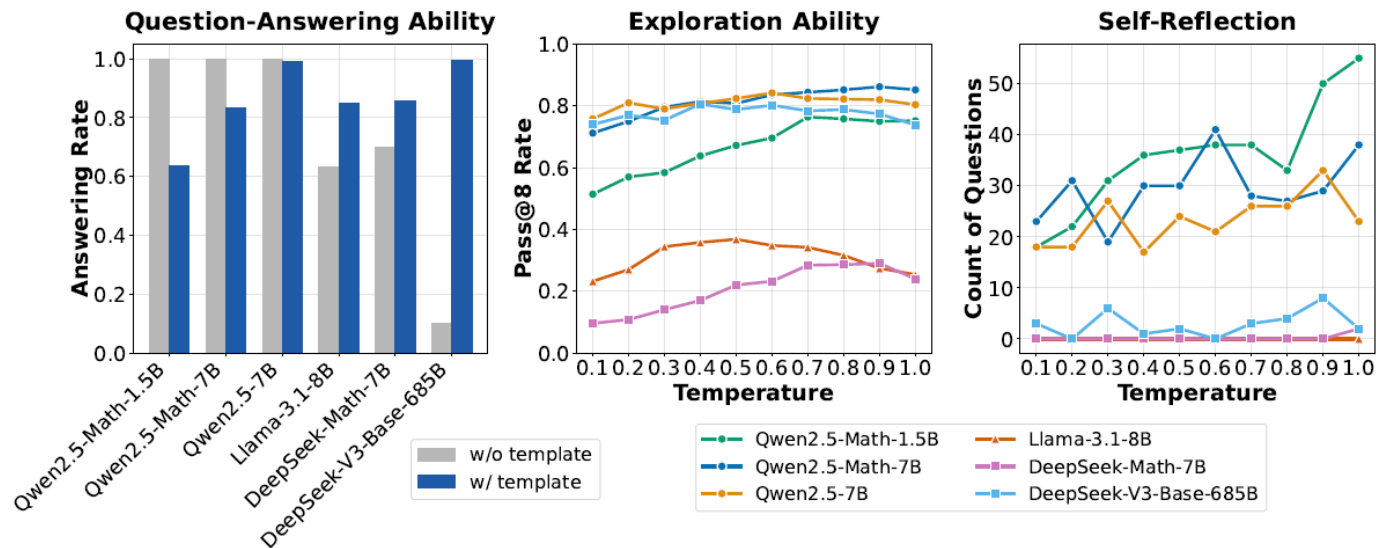


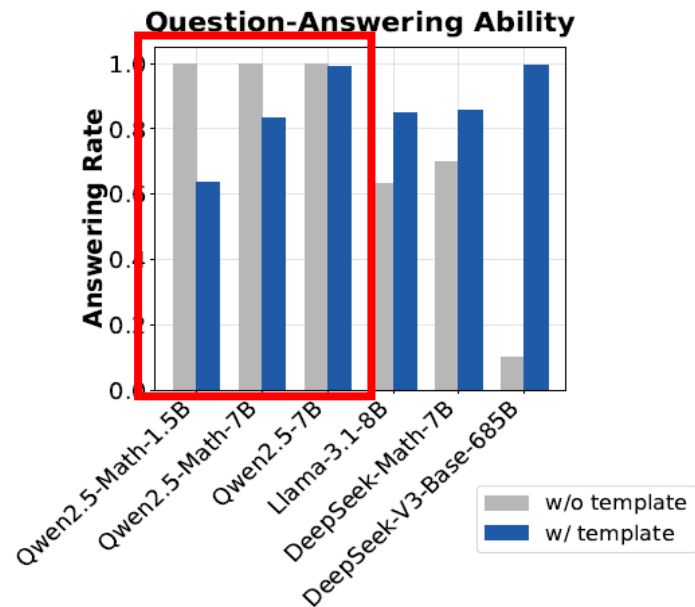
Figure 3: Model attributes across three aspects. **Question-Answering Ability**: the extent to which a pretrained language model provides a direct answer to a question rather than continuing or expanding upon it; **Exploration Ability**: pass@8 measures how well base models explore; **Self-Reflection**: counts are obtained through cross-validation between keyword-based detection and LLM-based detection, as detailed in Appendix D.

DeepSeek-R1-Zero

Qwen2.5는 pseudo SFT 모델일 수 있다

w/o template에서의 Qwen2.5 모델의 성능이 높음

- Qwen2.5는 학습 과정에서 chat model data로 학습함
- 이는 maximize $\log p_{\theta}(q; o)$ 를 사전학습한 것일 수 있음 (SFT)
- R1-Zero를 reproduce 하는데 있어서 Qwen2.5 모델은 적합하지 않을 수 있음



Base model + Template	AIME24	AMC	MATH500	Minerva	OlympiadBench	Avg.
Qwen2.5-Math-1.5B (4-shot prompting)						
R1 template	0.0	20.0	50.4	12.1	15.9	19.7
Qwen template	0.0	9.6	21.2	6.6	2.2	7.9
No template	20.0	32.5	33.0	12.5	22.8	24.2
Qwen2.5-Math-7B (4-shot prompting)						
R1 template	16.7	43.4	61.8	15.1	28.4	33.1
Qwen template	3.3	22.5	61.6	10.7	20.9	23.8
No template	0.0	0.0	0.0	0.0	0.1	0.0
Qwen template	16.7	38.6	50.6	9.9	16.6	26.5
No template	0.2	45.8	69.0	21.3	34.7	38.2

Table 1: Qwen2.5-Math models might be pretrained on concatenated question-answer text, resulting in peak performance when **no template** is applied.

DeepSeek-R1-Zero

- “Aha moment”는 RL 이전에도 발견되던 현상이다
 - Self-reflection이 발생하는 빈도가 RL 이후 증가
 - DeepSeek-R1-Zero에서 self-reflection과 정확도가 큰 관련이 없음
 - Response length가 증가하는 현상도 Reasoning 능력 강화가 원인이 아닐 수 있음

Question(a): An element is randomly chosen from among the first 15 rows ...

Response

In Pascal’s Triangle, every row starts and ends with 1, ...

...

This can be calculated as: *awkward silence* **Wait, I’m overthinking. Let’s try again.** The number of elements in the first n rows of Pascal’s Triangle is $(n + 1)! / [2! * (n - 1)!]$.

...

Question(b): Find the least positive integer n such that $1/\sin 45^\circ \sin 46^\circ + \dots$

Response

First, I notice that the left-hand side of the equation has the form ...

Simplifying, I get ...

Aha! I can use this to get $\sin(a + 2b) = \sin((a + b) + b) = 3 \sin(a + b) \dots$

This looks messy, but I notice that it has the form $\sin a \cos b + \cos a \sin b$ again.

This gives me an idea. Maybe I can ...

...

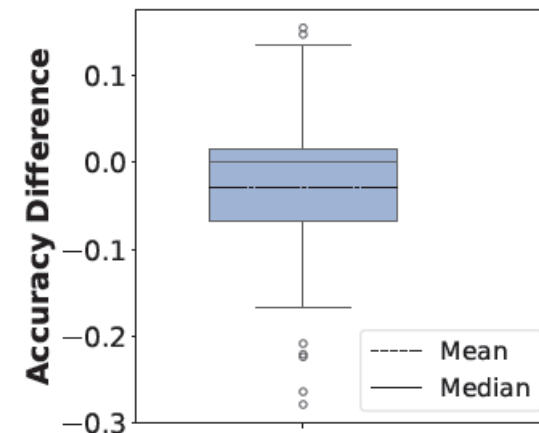


Figure 15: Accuracy difference between responses with and without self-reflection for each question (responses sampled from DeepSeek-R1-Zero).

Figure 13: Cases showing that DeepSeek-V3-Base already exhibits “Aha moment” even before RL tuning.

Return To GRPO

GRPO의 Weakness 1

GRPO의 Objective Function

- Response length bias
- Token별 평균을 취하는 과정에서 길이에 따른 선호도 발생
- 응답 길이로 나누는 것은 Advantage를 응답 길이로 나누는 reweighting과 같다

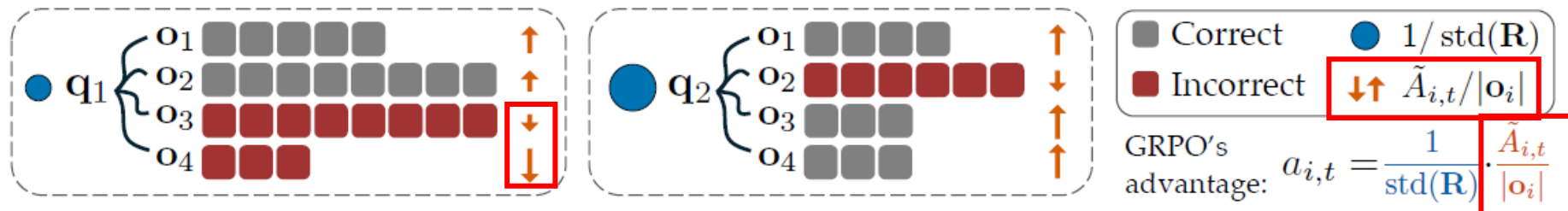


Figure 4: Illustration of the biases in GRPO. Note that the effective advantage of GRPO $a_{i,t}$ is equivalent to a reweighted version of the unbiased advantage $\tilde{A}_{i,t} = R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\mathbf{R})$. The terms $\text{std}(\mathbf{R})$ and $|\mathbf{o}_i|$ could bias the optimization by assigning different weights to different questions and responses, as denoted by the sizes of the blue circles and the lengths of the orange arrows. Upward arrows indicate positive advantages, and vice versa.

Return To GRPO

GRPO의 Weakness 2

Advantage Group 내 리워드의 정규화

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}$$

Group내 reward가 대부분 0 or 1인 경우 $\hat{A}_{i,t}$ 발산

- Difficulty Bias
- 쉽거나 너무 어려운 문제에서 큰 gradient

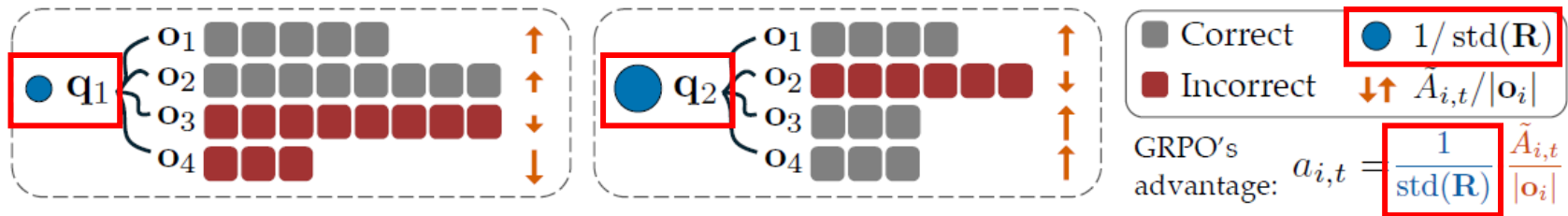
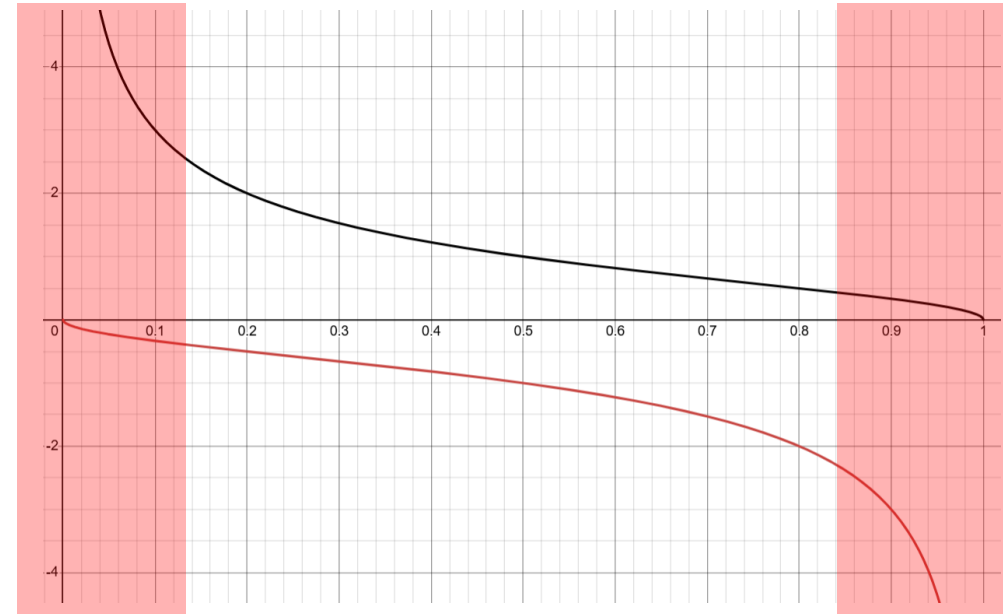


Figure 4: Illustration of the biases in GRPO. Note that the effective advantage of GRPO $a_{i,t}$ is equivalent to a reweighted version of the unbiased advantage $\tilde{A}_{i,t} = R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\mathbf{R})$. The terms $\text{std}(\mathbf{R})$ and $|\mathbf{o}_i|$ could bias the optimization by assigning different weights to different questions and responses, as denoted by the sizes of the blue circles and the lengths of the orange arrows. Upward arrows indicate positive advantages, and vice versa.

Dr.GRPO

- Dr. GRPO: Group Relative Policy Optimization Done Right

- response length bias / difficulty bias 제거
- length, std normalization을 제거해도 되는 수학적 근거 (논문의 Appendix A 참조)

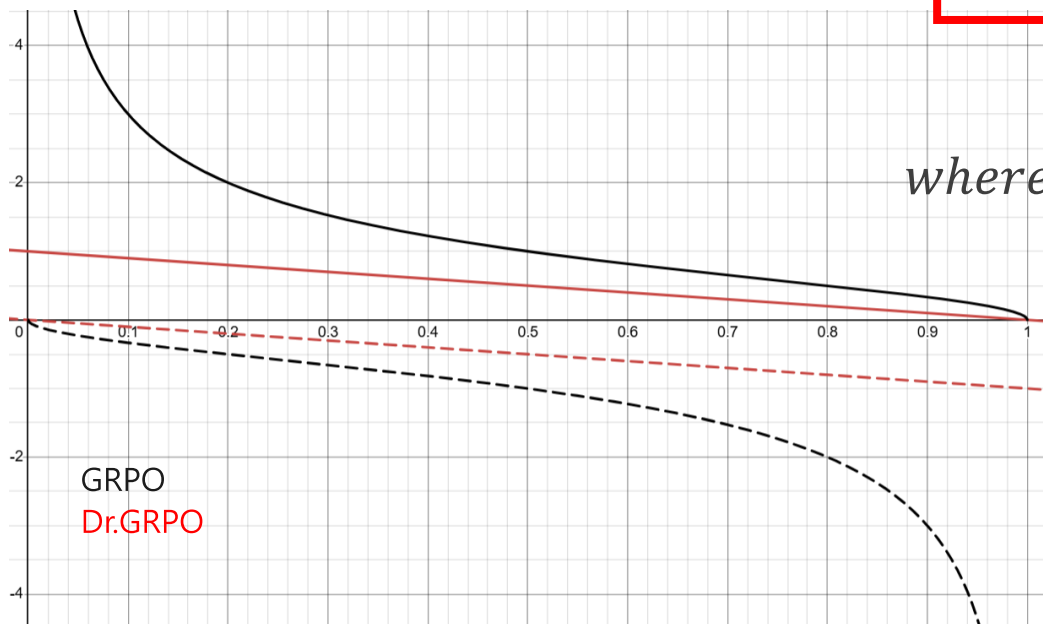
reward 모델이 없으면 KL term을 사용하지 않는 경우가 많음

$$\mathcal{J}_{Dr.GRPO} = \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)} \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_t|} \sum_{i=1}^{|o_t|} \left\{ \min[r_t(\theta) \hat{A}_{i,t}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t}] - \beta \mathbb{D}_{KL}[\pi_{\theta} || \pi_{ref}] \right\}$$

Response length bias의 원인 제거

where $\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}$

Difficulty bias의 원인 제거



Experiments Setting

Dataset

- Train) MATH
- Eval) AIME2024, AMC, MATH500, Minerva Math and OlympiadBench

Model

- Qwen2.5-1.5B + R1 template
- Reward Model -> Math-Verify

Hyperparameters

Parameter	Value
LEARNER	
Optimizer	AdamW
Adam parameters (β_1, β_2)	(0.9, 0.95)
Weight decay	0.0
Gradient norm clipping	1.0
Learning rate scheduler	Constant
Learning rate	1×10^{-6}
Inner proximal update epoch	1
KL loss coefficient	0.0
KL penalty coefficient	0.0
Policy clipping parameter	0.2

Template 1 (R1 template). A conversation between User and Assistant. The User asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the answer. The reasoning process is enclosed within `<think>` `</think>` and answer is enclosed within `<answer>` `</answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>` `<answer>` answer here `</answer>`.
 \nUser: {question}\nAssistant: <think>

Template 2 (Qwen-Math template). `<|im_start|>`system\nPlease reason step by step, and put your final answer within `\\boxed{}`.`<|im_end|>`\n`<|im_start|>`user\n{question}
`<|im_end|>`\n`<|im_start|>`assistant\n

Template 3 (No template). {question}

< Templates (R1, Qwen-Math, No template) >

Parameter	Value
ACTOR	
Maximum response length	3000 tokens
Sampling temperature	1.0
(top P, top k)	(1.0, -1)
Number of responses per question	8

GRPO vs DR.GRPO

Mathematical reasoning under different training settings

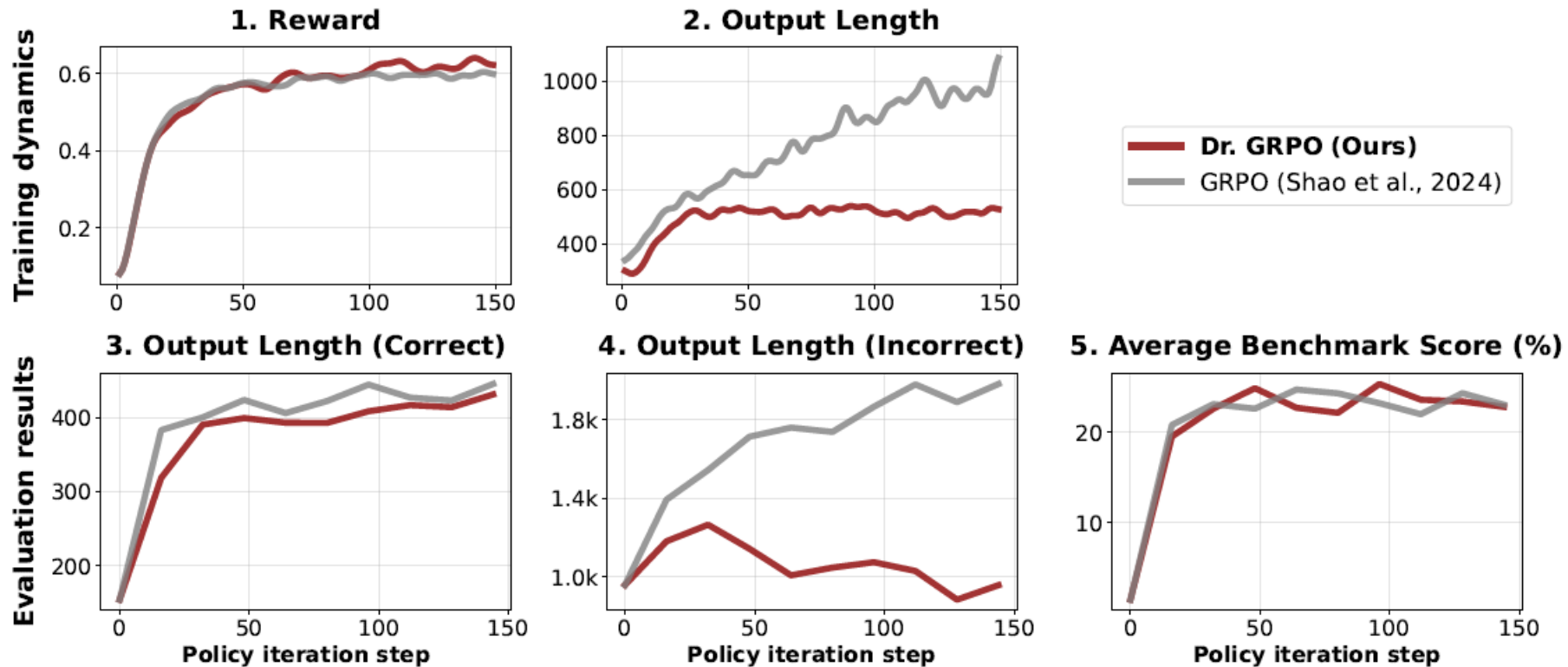


Figure 5: Comparison of Dr. GRPO and GRPO in terms of training dynamics (Top) and evaluation results (Bottom).

Benchmark Scores

Mathematical Benchmark

- Base model에서 성능이 크게 개선
- GRPO와 비교가 없음
- max_length 8k는 회색으로 표기

Base model + Method	AIME24	AMC	MATH500	Minerva	OlympiadBench	Avg.
Qwen2.5-Math-1.5B	20.0	32.5	33.0	12.5	22.8	24.2
Qwen2.5-Math-1.5B*	16.7	43.4	61.8	15.1	28.4	33.1
Oat-Zero-1.5B	20.0	53.0	74.2	25.7	37.6	42.1
R1-Distill-Qwen-1.5B @ 3k	2.5	21.7	52.2	16.3	17.3	22.0
R1-Distill-Qwen-1.5B @ 8k	20.0	49.4	77.4	25.0	35.8	41.5
Qwen2.5-Math-1.5B-Instruct	10.0	48.2	74.2	26.5	40.2	39.8
Llama-3.2-3B	0.0	2.4	6.4	6.3	1.3	3.3
+ RL w. Dr. GRPO	3.3	7.2	10.0	11.0	2.2	6.8
Llama-3.2-3B-FineMath	0.0	3.6	18.4	5.9	2.2	6.0
+ RL w. Dr. GRPO	3.3	10.8	38.0	12.9	9.0	14.8
Llama-3.2-3B-NuminaQA	0.0	0.0	0.6	0.0	0.1	0.14
+ RL w. Dr. GRPO (Oat-Zero-3B)	6.7	18.1	50.0	14.3	14.7	20.7
Llama-3.2-3B-Instruct	6.7	15.7	38.8	11.8	12.6	17.1
Qwen2.5-Math-7B	16.7	38.6	50.6	9.9	16.6	26.5
Qwen2.5-Math-7B*	0.2	45.8	69.0	21.3	34.7	38.2
SimpleRL-Zero-7B	26.7	60.2	78.2	27.6	40.3	46.6
PRIME-Zero-7B	16.7	62.7	83.8	36.0	40.9	48.0
OpenReasoner-Zero-7B @ 3k	13.3	47.0	79.2	31.6	44.0	43.0
OpenReasoner-Zero-7B @ 8k	13.3	54.2	82.4	31.6	47.9	45.9
Oat-Zero-7B	43.3	62.7	80.0	30.1	41.0	51.4
R1-Distill-Qwen-7B @ 3k	10.0	26.2	60.1	23.0	23.1	28.5
R1-Distill-Qwen-7B @ 8k	33.3	68.4	88.1	35.9	47.7	54.7
Qwen2.5-Math-7B-Instruct	16.7	53.0	83.6	29.8	42.7	45.1

Table 4: A comparison on benchmark scores. *Ours* models are RL-tuned by our minimalist recipe (Sec. 1). * means we employ the best template (no template) to generate answers, such that the test scores are highest and can faithfully reflect the capabilities of the base models.

RL with Domain-Specific Pretraining

Weak Base model에서도 RL이 효과적인가?

- NuminaQA: FineMath 모델에 numia-1.5-qa 데이터로 추가적인 R1-zero like 학습
- FineMath: FineMath + FineWeb-Edu 데이터로 사전학습
- Base 모델은 RL로 인한 성능 개선이 적음, 추가적인 학습이 된 모델의 경우 효과적인 성능 개선

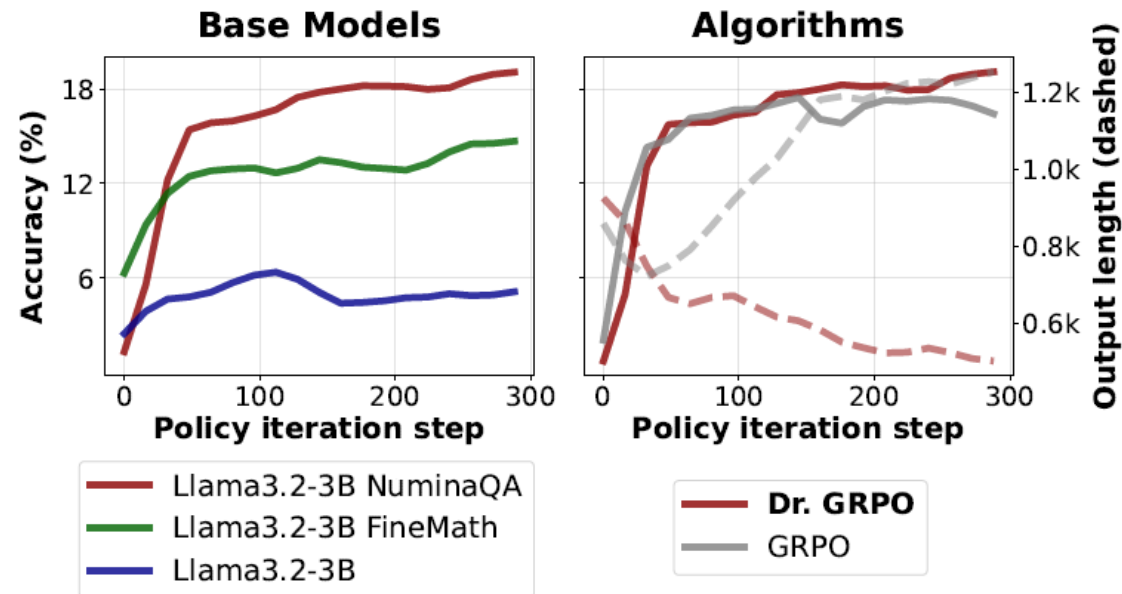


Figure 7: **Left:** The average benchmark performance curves of different base models. **Right:** The comparison between Dr. GRPO and GRPO with respect to reasoning accuracy (solid lines) and model response length (dashed lines).

Benchmark Scores

○ Ablation Study

- Dr.GRPO = GRPO – Len Norm – Std Norm
- Std Norm보다 Len Norm이 효과적
- 생성되는 Token Length가 짧아짐

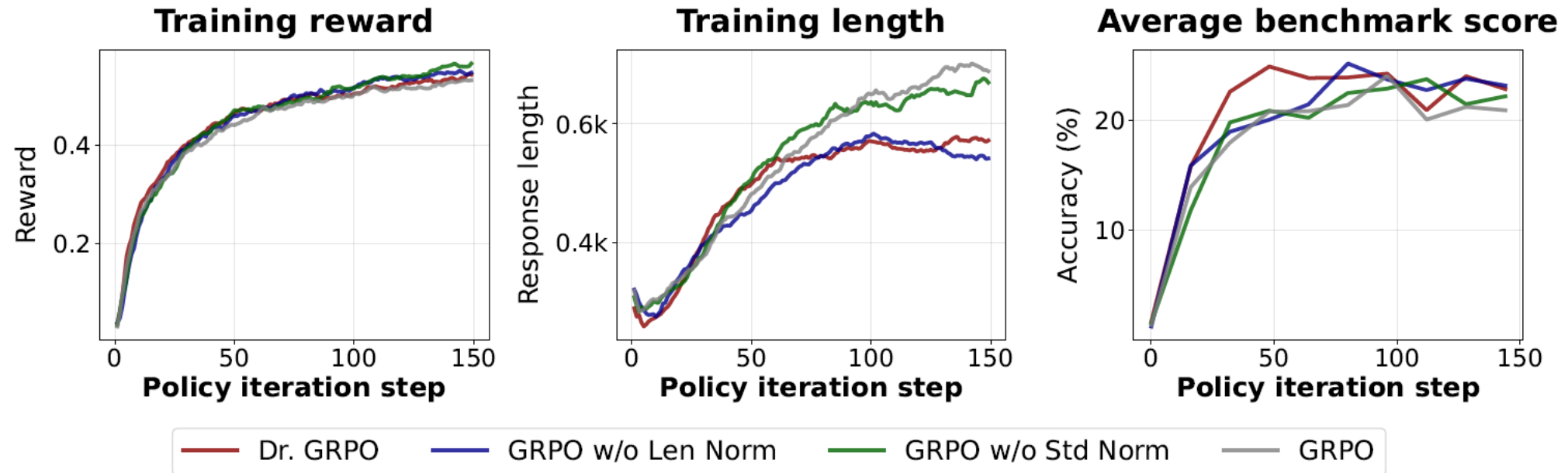


Figure 8: Ablation results on the two bias terms in GRPO.

Conclusion

○ R1-Zero Style RL에 대한 분석

- 일부 Base Model의 경우 이미 SFT 형태의 학습이 되어 있을 수 있음 (Qwen2.5)
- Aha Moment는 RL을 하지 않은 Base Model에서도 발견되는 현상, self-reflection과 성능에는 상관관계 X
- Response Length가 길어지는 현상은 추론 능력 향상이 아니라 bias 문제일 수 있음

○ Dr.GRPO 알고리즘을 제안

- GRPO의 2가지 bias 가능성을 제시: length bias, difficulty bias
- 2가지 bias의 원인인 Length Normalization과 Std Normalization을 제거한 Dr.GRPO 설계
- 성능에서는 GRPO와 큰 차이를 보이지 않지만, 생성된 응답의 길이가 짧은 경향

The Future of Reinforcement Learning in NLP

○ Reward Allocation in RLVR

- 어떻게 Advantage를 설계할 것인가?
- 시퀀스 내 모든 토큰이 동일하게 학습되는 것이 맞는가?

○ Sampling Strategy

- 학습에 사용되는 Sample을 어떻게 추출할 것인가?

○ Entropy & Diversity

- 학습 과정에서 다양성을 어떻게 확보할 것인가?

$$J_{RL} = \mathbb{E}_{\substack{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)}} \frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \hat{A}_{i,t} \times \pi(o_t | [q: o_{<t}])$$

Reward Allocation,
Sampling Strategy



Thank you

Dooyoung Kim

(kdysunleo98@gmail.com)



질문 <https://forms.gle/LtvyMJ7BFwMKpWtz8>

피드백 <https://forms.gle/PAmxLQnRBZVhAMaw8>

응답 <https://docs.google.com/spreadsheets/d/1uWyc0pUfQOwTTZUDyY3gxImU5xcFro90kKJKOcDitnk/edit?usp=sharing>