A Policy Gradient Derivations

In the context of RL for LLM post-training, we typically maximize the value of

$$\mathcal{J}(\pi_\theta) = \mathop{\mathbb{E}}_{\mathbf{q}\sim p_\mathcal{Q}} \left[\mathop{\mathbb{E}}_{\mathbf{o}\sim\pi_\theta(\cdot|\mathbf{q})} [R(\mathbf{q},\mathbf{o})] \right], \tag{4}$$

where $R(\mathbf{q},\mathbf{o}) = \sum_{t=1}^{|\mathbf{o}|} r(\mathbf{q},\mathbf{o}_{\le t})$ is the return (Sutton & Barto, 2018) of the trajectory $\mathbf{q};\mathbf{o}$, and $r(\mathbf{q},\mathbf{o}_{\le t})$ represents the token-level reward for t-th token in response $\mathbf{o}$.

The Monte Carlo policy gradient (Sutton & Barto, 2018) of Eq. (4) is

$$\begin{aligned}
\nabla_\theta \mathcal{J}(\pi_\theta) &= \mathop{\mathbb{E}}_{\mathbf{q}\sim p_\mathcal{Q}} \left[\mathop{\mathbb{E}}_{\mathbf{o}\sim\pi_\theta(\cdot|\mathbf{q})} [\nabla_\theta \log \pi_\theta(\mathbf{o}|\mathbf{q}) R(\mathbf{q},\mathbf{o})] \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{q}\sim p_\mathcal{Q}} \left[\mathop{\mathbb{E}}_{\mathbf{o}\sim\pi_\theta(\cdot|\mathbf{q})} [\nabla_\theta \sum_{t=1}^{|\mathbf{o}|} \log \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t}) R(\mathbf{q},\mathbf{o})] \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{q}\sim p_\mathcal{Q}} \left[\mathop{\mathbb{E}}_{\mathbf{o}\sim\pi_\theta(\cdot|\mathbf{q})} [\sum_{t=1}^{|\mathbf{o}|} \nabla_\theta \log \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t}) \sum_{t'=t}^{|\mathbf{o}|} r(\mathbf{q},\mathbf{o}_{\le t'})] \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{q}\sim p_\mathcal{Q}} \left[\mathop{\mathbb{E}}_{\mathbf{o}\sim\pi_\theta(\cdot|\mathbf{q})} \left[\sum_{t=1}^{|\mathbf{o}|} \nabla_\theta \log \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t}) \left(\sum_{t'=t}^{|\mathbf{o}|} r(\mathbf{q},\mathbf{o}_{\le t'}) - B(\mathbf{q},\mathbf{o}_{<t}) \right) \right] \right],
\end{aligned} \tag{5}$$

where $B(\mathbf{q},\mathbf{o}_{<t})$ is a variance reduction term, which is invariant with respect to o_t so that

$$\begin{aligned}
\mathop{\mathbb{E}}_{o_t\sim\pi_\theta(\cdot|\mathbf{q},\mathbf{o}_{<t})} [\nabla_\theta \log \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t}) B(\mathbf{q},\mathbf{o}_{<t})] &= \mathop{\mathbb{E}}_{o_t\sim\pi_\theta(\cdot|\mathbf{q},\mathbf{o}_{<t})} [\nabla_\theta \log \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t})] B(\mathbf{q},\mathbf{o}_{<t}) \\
&= [\sum_{o_t} \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t}) \nabla_\theta \log \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t})] B(\mathbf{q},\mathbf{o}_{<t}) \\
&= [\sum_{o_t} \nabla_\theta \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t})] B(\mathbf{q},\mathbf{o}_{<t}) \\
&= [\nabla_\theta \sum_{o_t} \pi_\theta(o_t|\mathbf{q},\mathbf{o}_{<t})] B(\mathbf{q},\mathbf{o}_{<t}) \\
&= [\nabla_\theta 1] B(\mathbf{q},\mathbf{o}_{z<t}) = 0.
\end{aligned}$$

Typically, we set $B(\mathbf{q},\mathbf{o}_{<t}) = \mathop{\mathbb{E}}_{\mathbf{o}_{\ge t}\sim\pi_\theta(\cdot|\mathbf{q},\mathbf{o}_{<t})} [\sum_{t'=t}^{|\mathbf{o}|} r(\mathbf{q},\mathbf{o}_{\le t'})]$, which is the expected cumulative reward in the future (also known as the value of the current state), and denote $A(o_t|\mathbf{q},\mathbf{o}_{<t}) = \sum_{t'=t}^{|\mathbf{o}|} r(\mathbf{q},\mathbf{o}_{\le t'}) - B(\mathbf{q},\mathbf{o}_{<t})$ as the advantage. In the case of outcome reward, $\sum_{t'=t}^{|\mathbf{o}|} r(\mathbf{q},\mathbf{o}_{\le t'}) = \sum_{t=1}^{|\mathbf{o}|} r(\mathbf{q},\mathbf{o}_{\le t}) = R(\mathbf{q},\mathbf{o})$.

By setting $B(\mathbf{q},\mathbf{o}_{<t}) = \text{mean}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\})$, the policy gradient of Eq. (5) becomes

$$\nabla_\theta \mathcal{J}(\pi_\theta) = \mathop{\mathbb{E}}_{\mathbf{q}\sim p_\mathcal{Q}} \left[\mathop{\mathbb{E}}_{\{\mathbf{o}_i\}_{i=1}^G\sim\pi_\theta(\cdot|\mathbf{q})} [\frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{|\mathbf{o}|} \nabla_\theta \log \pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t}) \tilde{A}_{i,t}] \right], \tag{6}$$

where

$$\tilde{A}_{i,t} = \frac{R(\mathbf{q},\mathbf{o}_i) - \text{mean}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\})}{\cancel{\text{std}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\})}}.$$

We adopt the PPO (Schulman et al., 2017b) objective to compute Eq. (6):

$$\begin{aligned}
&\mathcal{J}(\pi_\theta) = \mathbb{E}[\mathbf{q}\sim p_\mathcal{Q}, \{\mathbf{o}_i\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot|\mathbf{q})] \\
&\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})} \tilde{A}_i, \text{clip} \left(\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}, 1-\epsilon, 1+\epsilon \right) \tilde{A}_i \right] \right\},
\end{aligned}$$

from which we conclude that both std and $|\mathbf{o}|$ should not appear in the RL objective.

Unbiasedness of $\tilde{A}_{i,t}$. We note that $\tilde{A}_{i,t}$ computed above is equivalent to that of REINFORCE Leave-One-Out (RLOO) (Ahmadian et al., 2024; Kool et al., 2019) up to a scaling factor, which can be subsumed into the learning rate without affecting the RL dynamics. Specifically,

$$\begin{aligned}\frac{G}{G-1}\cdot\tilde{A}_{i,t} &= \frac{G}{G-1}R(\mathbf{q},\mathbf{o}_i) - \frac{G}{G-1}\frac{1}{G}\sum_{j=1}^{G}R(\mathbf{q},\mathbf{o}_j)\\ &= \frac{G}{G-1}R(\mathbf{q},\mathbf{o}_i) - \frac{1}{G-1}\sum_{j=1,j\neq i}^{G}R(\mathbf{q},\mathbf{o}_j) - \frac{1}{G-1}R(\mathbf{q},\mathbf{o}_i)\\ &= \hat{A}_{i,t}^{\text{RLOO}}.\end{aligned}$$

B Detailed Benchmark Results

We show the detailed benchmark results for three scales (1.5B, 3B and 7B) in Table 4. We also include the instruct models at the same scale and R1-Distill models for comparison. Note that since we employ the Qwen2.5-Math base models, which have a context length of 4k, we thus limit the generation budget at 3k for all baselines compared. For models that are trained for a longer context (OpenReasoner-Zero end R1-Distill-Qwen), we also report their performance at 8k generation budget.

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.

C Extended Empirical Results

In this section we present two extended empirical results for (1) the ablation of different bias terms in GRPO and (2) statistical significance of Dr. GRPO's results. We RL-tune the

Qwen2.5-1.5B base model on a mixture of 3K diverse math questions drawn from ASDiv, MATH, and AIME (pre-2023).

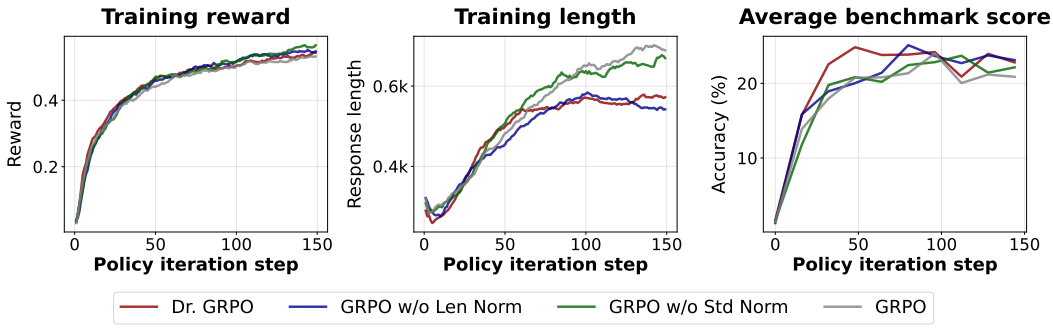

Figure 8: Ablation results on the two bias terms in GRPO.

Fig. 8 shows the training and evaluation curves for the following variants: Dr. GRPO, GRPO w/o length normalization, GRPO w/o standard deviation (std) normalization and Vanilla GRPO. From the middle subplot, we observe that both Dr. GRPO and the variant without length normalization generate shorter responses compared to the other two. This confirms that the length bias term has a more significant influence on response length–consistent with our expectations.

In terms of performance, Dr. GRPO and the other ablated variants consistently outperform vanilla GRPO in both training rewards and evaluation accuracy. This indicates that removing bias terms (either length or std) improves policy learning, validating our motivation for Dr. GRPO.

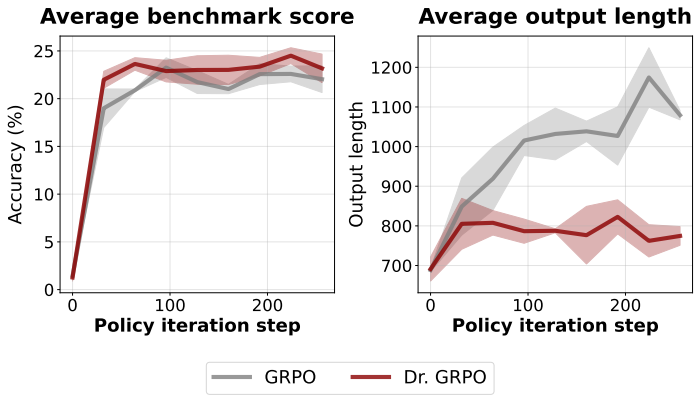

Figure 9: Evaluation results of 3 independent RL runs. The mean curves are drawn in solid lines and the standard deviation is plotted in the shaded areas.

Fig. 9 compares GRPO and Dr. GRPO across three independent runs. We observe that Dr. GRPO consistently demonstrates statistically significant improvements–both in token efficiency and final accuracy–across different random seeds.

D Keyword-based Detection and LLM-Based Identification of Self-Reflection Behaviors

We construct a pool of carefully selected keywords and phrases that signal self-reflection behaviors in the LLM's responses. However, LLM-generated responses often contain hallucinations and off-topic content, leading to the presence of simple, ambiguous keywords that do not necessarily indicate genuine self-reflection. For instance, terms like "wait"