On the Impossibility of Unbiased and Length-Invariant Policy Optimization with Outcome Rewards

Fei Ding*
Alibaba Group

Yongkang Zhang
Alibaba Group

Yuhao Liao
Tsinghua University

Zijian Zeng
Tsinghua University

Huiming Yang
Tsinghua University

Abstract

Group Relative Policy Optimization (GRPO) is the dominant reinforcement learning algorithm for training reasoning capabilities in large language models, notably adopted by DeepSeek-R1. The recent improvement Dr. GRPO (COLM 2025) identifies the response-level length bias caused by per-trajectory length normalization in GRPO and proposes removing this normalization, claiming the resulting optimizer is "unbiased." We show that this claim is incomplete. Specifically, we establish an *impossibility theorem*: under the standard outcome reward + GRPO setting, no length-based weighting scheme can simultaneously achieve the following two properties. (P1) *Gradient unbiasedness*: the gradient estimator is an unbiased estimate of the true policy gradient. (P2) *Length invariance*: each trajectory's effective contribution to the gradient is independent of its token length. GRPO approximately satisfies P2 but violates P1; Dr. GRPO satisfies P1 but violates P2. We characterize the complete tradeoff spectrum via the parametric family $f_\alpha(L) = L^{\alpha-1}$, where $\alpha = 0$ recovers GRPO, $\alpha = 1$ recovers Dr. GRPO, and provide quantitative analysis showing that Dr. GRPO's length bias can cause longer trajectories to dominate gradient updates by a factor proportional to the length ratio. Our results reveal that neither algorithm is universally "done right"; they occupy opposite ends of a fundamental and unavoidable tradeoff.

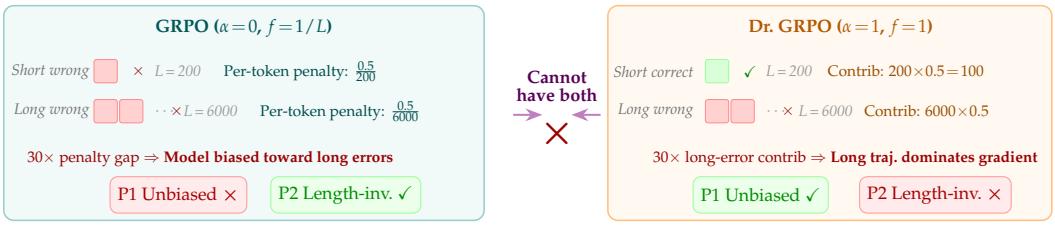

Figure 1: Complete unbiasedness is impossible.

1 Introduction

Reinforcement learning (RL) has become a core technique for improving reasoning capabilities of large language models (LLMs). DeepSeek-R1-Zero (DeepSeek-AI et al., 2026) demonstrated an important finding: without supervised fine-tuning, directly applying RL to a base LLM can elicit complex reasoning behaviors, including chain-of-thought and self-reflection. The core algorithm of this training paradigm is Group Relative Policy Optimization (GRPO) (Shao et al., 2024). GRPO is a critic-free RL algorithm that estimates advantages by comparing multiple responses sampled for the same prompt.

*Corresponding author: dignfei@gmail.com

A salient empirical observation during GRPO training is the persistent growth of response length (DeepSeek-AI et al., 2026; Zeng et al., 2025; Hu et al., 2025). Liu et al. (2025) critically examined this phenomenon and identified two sources of optimization bias in GRPO. The first is *response-level length bias* caused by per-trajectory length normalization $\frac{1}{|\mathbf{o}_i|}$, and the second is *question-level difficulty bias* caused by standard deviation normalization. They proposed Dr. GRPO, removing both normalization terms, claiming to restore an "unbiased" optimization objective. Dr. GRPO has been widely adopted by the community and achieved state-of-the-art results on mathematical reasoning benchmarks at the time.

In this paper, we challenge the completeness of this claim. We confirm that Dr. GRPO's gradient estimator is indeed an unbiased estimate of the policy gradient (as they rigorously proved in their Appendix A). However, we show that removing the length normalization term $\frac{1}{|\mathbf{o}_i|}$ introduces another form of bias: *length bias in the optimization dynamics*. This bias causes longer trajectories to contribute disproportionately more to gradient updates. More fundamentally, we establish the following impossibility result:

> **Main Result (Informal).** Under the outcome reward + GRPO setting, no length-based weighting scheme can simultaneously achieve gradient unbiasedness and length invariance. GRPO and Dr. GRPO represent the two extremes of this unavoidable tradeoff.

Our contributions are as follows:

- We formalize two desirable properties of group-based RL optimizers, namely *gradient unbiasedness* (P1) and *length invariance* (P2), and prove they are mutually exclusive under outcome rewards (theorem 6).
- We characterize the tradeoff spectrum via the parametric family $f_\alpha(L) = L^{\alpha-1}$ ($\alpha \in [0, 1]$), where $\alpha = 0$ corresponds to GRPO and $\alpha = 1$ corresponds to Dr. GRPO (theorem 8).
- We provide quantitative analysis showing that Dr. GRPO's length bias can be severe: at length ratio r, the longer trajectory captures $\frac{r}{1+r}$ of the gradient signal (theorem 9).

2 Preliminaries

Token-level MDP. Language model generation is modeled as a token-level Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, p_\mathcal{Q})$. At step t, the state $s_t = [\mathbf{q}, o_1, \ldots, o_{t-1}]$ is the concatenation of the prompt and previously generated tokens. The policy $\pi_\theta(\cdot|s_t)$ selects the next token o_t from the vocabulary $\mathcal{A}$. Generation terminates upon producing an end-of-sequence token or exhausting the token budget. The objective is to maximize the expected return:

$$J(\pi_\theta) = \mathbb{E}_{\mathbf{q}\sim p_\mathcal{Q}}\left[\mathbb{E}_{\mathbf{o}\sim\pi_\theta(\cdot|\mathbf{q})}\left[R(\mathbf{q}, \mathbf{o})\right]\right], \tag{1}$$

where $R(\mathbf{q}, \mathbf{o}) = \sum_{t=1}^{|\mathbf{o}|} r(s_t, o_t)$ is the trajectory return. Under the standard *outcome reward* setting for reasoning tasks (DeepSeek-AI et al., 2026), a scalar reward is assigned at the end of generation: $R(\mathbf{q}, \mathbf{o}) = 1$ if $\mathbf{o}$ contains the correct answer, and 0 otherwise.

Policy gradient. The Monte Carlo policy gradient (Williams, 1992; Sutton & Barto, 2018) of Eq. (1) is:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\mathbf{q},\mathbf{o}\sim\pi_\theta}\left[\sum_{t=1}^{|\mathbf{o}|} \nabla_\theta \log \pi_\theta(o_t|\mathbf{q}, \mathbf{o}_{<t}) \cdot A(o_t|\mathbf{q}, \mathbf{o}_{<t})\right], \tag{2}$$

where $A(o_t|\mathbf{q}, \mathbf{o}_{<t}) = R(\mathbf{q}, \mathbf{o}) - B(\mathbf{q}, \mathbf{o}_{<t})$ is the advantage and B is any baseline independent of o_t (Sutton & Barto, 2018). Under outcome rewards, the advantage is identical for all tokens in a trajectory since the return does not depend on t.

Group-relative baseline. Both GRPO and Dr. GRPO sample G responses $\{\mathbf{o}_1, \ldots, \mathbf{o}_G\}$ for each prompt and use the group mean as the baseline: $B = \text{mean}(\mathbf{R})$, where $\mathbf{R} = \{R(\mathbf{q}, \mathbf{o}_1), \ldots, R(\mathbf{q}, \mathbf{o}_G)\}$. The advantage for all tokens in trajectory $\mathbf{o}_i$ is:

$$\tilde{A}_i = R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\mathbf{R}). \tag{3}$$

GRPO (Shao et al., 2024). GRPO maximizes the following surrogate objective (omitting the clipping mechanism as it does not affect our analysis):

$$J_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{o}_i|}\sum_{t=1}^{|\mathbf{o}_i|}\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}\cdot\frac{\tilde{A}_i}{\text{std}(\mathbf{R})}. \tag{4}$$

Dr. GRPO (Liu et al., 2025). Dr. GRPO removes the per-trajectory length normalization $\frac{1}{|\mathbf{o}_i|}$ and the standard deviation normalization $\text{std}(\mathbf{R})$:

$$J_{\text{Dr.GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|\mathbf{o}_i|}\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}\cdot\tilde{A}_i. \tag{5}$$

Liu et al. (2025) proved in their Appendix A that the gradient of Eq. (5) recovers the unbiased Monte Carlo policy gradient with a group-relative baseline. Furthermore, the advantage $\tilde{A}_i$ is equivalent to REINFORCE Leave-One-Out (RLOO) (Kool et al., 2019; Ahmadian et al., 2024) up to a constant factor.

Remark 1. Dr. GRPO's advantage estimator is equivalent to RLOO (Ahmadian et al., 2024) up to a constant factor; see Liu et al. (2025) Appendix A. Therefore our analysis also applies to RLOO, but we focus on Dr. GRPO since it explicitly claims to resolve the length bias issue.

Unified framework. To unify the analysis of both methods, we introduce a *weighted gradient estimator* parameterized by a weighting function $f : \mathbb{N} \to \mathbb{R}_+$:

$$\hat{g}_f = \frac{1}{G}\sum_{i=1}^{G} f(|\mathbf{o}_i|)\cdot\tilde{A}_i\cdot\sum_{t=1}^{|\mathbf{o}_i|}\nabla_\theta \log \pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t}). \tag{6}$$

GRPO corresponds to $f(L) = 1/L$ and Dr. GRPO corresponds to $f(L) = 1$ (both omitting the $\text{std}(\mathbf{R})$ factor since it is a question-level scalar orthogonal to length bias analysis).

3 Main Result: Impossibility Theorem

Notation and setup. Consider the length-weighted gradient estimator

$$\hat{g}_f = \frac{1}{G}\sum_{i=1}^{G} f(L_i)\,\tilde{A}_i\,S_i, \tag{7}$$

where

$$L_i := |\mathbf{o}_i|, \qquad S_i := \sum_{t=1}^{L_i}\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, \mathbf{o}_{i,<t}), \tag{8}$$

with the group mean baseline

$$\tilde{A}_i := R_i - \frac{1}{G}\sum_{j=1}^{G} R_j. \tag{9}$$

In what follows, Π denotes the policy class under consideration. We assume all expectations below exist and that within-group trajectories are conditionally i.i.d. given the prompt and the current policy.

Assumption 2 (Fixed-length realizability)**.** There exists a set of lengths $\mathcal{L} \subseteq \mathbb{N}$ such that for every $L \in \mathcal{L}$, the policy class Π contains a policy $\pi^{(L)}$ under which, given the prompt, the trajectory length equals L almost surely, while the token content retains non-degenerate randomness.

Assumption 3 (Update scale functional)**.** Fix an update scale functional

$$\rho : \mathbb{R}^d \to \mathbb{R}_+,$$

used to measure the magnitude of a single-trajectory score sum. We only require ρ to be positively homogeneous of degree one for non-negative scalars, i.e., for all $\alpha \geq 0$ and all $v \in \mathbb{R}^d$,

$$\rho(\alpha v) = \alpha \rho(v). \tag{10}$$

Typical examples include vector norms or the non-negative projection magnitude along a fixed direction.

Definition 4 (Trajectory-level correctness P1)**.** The estimator $\hat{g}_f$ satisfies *trajectory-level correctness* over the policy class Π if there exists a constant $c > 0$, independent of the trajectory length distribution, such that for every policy $\pi \in \Pi$,

$$\mathbb{E}_\pi[\hat{g}_f] = c\, \nabla_\theta J(\pi). \tag{11}$$

Definition 5 (Length neutrality P2)**.** Let

$$\Gamma_{\pi,\rho}(L; a) := \mathbb{E}_\pi\left[\rho(S) \,\middle|\, L(\tau) = L,\ \tilde{A}(\tau) = a\right], \tag{12}$$

where a denotes a fixed effective training signal, i.e., a realized value of the group-relative advantage.

The estimator $\hat{g}_f$ satisfies *length neutrality* under the scale functional ρ if for every policy $\pi \in \Pi$, every realizable length L, and every fixed a,

$$f(L)\, \Gamma_{\pi,\rho}(L; a) \tag{13}$$

is independent of L.

Theorem 6 (Structural conflict at the policy-class level)**.** *Under the outcome-level reward and group mean baseline setting, consider a weight function depending only on length,*

$$f : \mathbb{N} \to \mathbb{R}_+.$$

Suppose Assumptions 2 and 3 hold.

If there exist a policy $\pi^\star \in \Pi$, an effective training signal value $a^\star$, and two distinct lengths $L_1, L_2 \in \mathcal{L}$ such that

$$\Gamma_{\pi^\star,\rho}(L_1; a^\star) \neq \Gamma_{\pi^\star,\rho}(L_2; a^\star), \tag{14}$$

then no such f can simultaneously satisfy P1 (trajectory-level correctness) and P2 (length neutrality) over the policy class Π.

Proof. We show that P1 and P2 impose mutually contradictory constraints on f.

Step 1: If P1 holds over the policy class Π, then $f(L)$ must be a constant function.

Pick any $L_0 \in \mathcal{L}$. By Assumption 2, there exists a policy $\pi^{(L_0)} \in \Pi$ under which the trajectory length equals L_0 almost surely given the prompt, while the token content remains random. Under this policy, for all i,

$$L_i = L_0, \tag{15}$$

so the estimator can be written as

$$\hat{g}_f = f(L_0) \cdot \frac{1}{G} \sum_{i=1}^{G} \tilde{A}_i S_i. \tag{16}$$

Expanding the baseline,

$$\tilde{A}_i = R_i - \frac{1}{G} \sum_{j=1}^{G} R_j = \left(1 - \frac{1}{G}\right) R_i - \frac{1}{G} \sum_{j \neq i} R_j. \tag{17}$$

Therefore,

$$\mathbb{E}_{\pi^{(L_0)}}[\tilde{A}_i S_i] = \left(1 - \frac{1}{G}\right) \mathbb{E}_{\pi^{(L_0)}}[R_i S_i] - \frac{1}{G} \sum_{j \neq i} \mathbb{E}_{\pi^{(L_0)}}[R_j S_i]. \tag{18}$$

For $j \neq i$, since within-group trajectories are conditionally i.i.d., R_j and S_i are independent; moreover, by the score function identity,

$$\mathbb{E}_{\pi^{(L_0)}}[S_i] = 0. \tag{19}$$

Hence,

$$\mathbb{E}_{\pi^{(L_0)}}[R_j S_i] = \mathbb{E}_{\pi^{(L_0)}}[R_j]\,\mathbb{E}_{\pi^{(L_0)}}[S_i] = 0. \tag{20}$$

Thus,

$$\mathbb{E}_{\pi^{(L_0)}}[\tilde{A}_i S_i] = \left(1 - \frac{1}{G}\right)\mathbb{E}_{\pi^{(L_0)}}[R_i S_i]. \tag{21}$$

By the REINFORCE identity,

$$\mathbb{E}_{\pi^{(L_0)}}[R_i S_i] = \nabla_\theta J(\pi^{(L_0)}), \tag{22}$$

yielding

$$\mathbb{E}_{\pi^{(L_0)}}[\hat{g}_f] = f(L_0)\,\frac{G-1}{G}\,\nabla_\theta J(\pi^{(L_0)}). \tag{23}$$

If P1 holds over the policy class Π, there exists a length-independent constant $c > 0$ such that

$$\mathbb{E}_{\pi^{(L_0)}}[\hat{g}_f] = c\,\nabla_\theta J(\pi^{(L_0)}). \tag{24}$$

Therefore,

$$f(L_0)\,\frac{G-1}{G} = c. \tag{25}$$

Since L_0 is arbitrary in $\mathcal{L}$, $f(L)$ must be the same for all $L \in \mathcal{L}$. That is, there exists a constant $c_0 > 0$ such that

$$f(L) \equiv c_0, \qquad \forall L \in \mathcal{L}. \tag{26}$$

Step 2: If P2 holds, then under the theorem's assumptions f cannot be a constant function.

By Definition 5, if the estimator $\hat{g}_f$ satisfies length neutrality P2 under the scale functional ρ, then for every policy $\pi \in \Pi$ and every effective training signal value a for which the conditional expectation is defined, there exists a constant $C_{\pi,a}$ depending only on (π, a) and not on the length L, such that for all realizable lengths $L \in \mathcal{L}$,

$$f(L)\,\Gamma_{\pi,\rho}(L; a) = C_{\pi,a}. \tag{27}$$

Now fix the policy $\pi^\star \in \Pi$, the effective training signal value $a^\star$, and the two distinct lengths $L_1, L_2 \in \mathcal{L}$ from the theorem's assumptions, satisfying

$$\Gamma_{\pi^\star,\rho}(L_1; a^\star) \neq \Gamma_{\pi^\star,\rho}(L_2; a^\star). \tag{28}$$

We show that f cannot be a constant function.

Suppose for contradiction that f is constant, i.e., there exists a constant $c_0 > 0$ such that

$$f(L) \equiv c_0, \qquad \forall L \in \mathcal{L}. \tag{29}$$

Substituting (29) into (27) with $\pi = \pi^\star$ and $a = a^\star$, we obtain for all realizable lengths $L \in \mathcal{L}$,

$$c_0\,\Gamma_{\pi^\star,\rho}(L; a^\star) = C_{\pi^\star,a^\star}. \tag{30}$$

In particular, for L_1 and L_2,

$$c_0\,\Gamma_{\pi^\star,\rho}(L_1; a^\star) = C_{\pi^\star,a^\star}, \tag{31}$$

and

$$c_0\,\Gamma_{\pi^\star,\rho}(L_2; a^\star) = C_{\pi^\star,a^\star}. \tag{32}$$

Since $c_0 > 0$, these two equations imply

$$\Gamma_{\pi^\star,\rho}(L_1; a^\star) = \Gamma_{\pi^\star,\rho}(L_2; a^\star), \tag{33}$$

contradicting (28).

Therefore, under the theorem's assumptions, any weight function f satisfying P2 cannot be a constant function.

Step 3: Contradiction.

Step 1 shows: if P1 holds over the policy class Π, then $f(L)$ must be a constant function. Step 2 shows: if P2 holds and there exists a policy for which $\Gamma_{\pi,\rho}(L;a)$ varies non-trivially with length, then $f(L)$ cannot be a constant function.

These are contradictory. Therefore, under the theorem's assumptions, no weight function f depending only on length can simultaneously satisfy P1 and P2 over the policy class Π. □

Illustrative example. Consider two trajectories for the same prompt with lengths $L_s \ll L_\ell$, compared under the same effective training signal. If under some pre-specified scale functional ρ, the longer trajectory has a larger typical score-sum magnitude, i.e.,

$$\Gamma_{\pi,\rho}(L_\ell;a) > \Gamma_{\pi,\rho}(L_s;a),$$

then constant weights preserve this length-induced scale disparity, while any length compensation attempting to eliminate this disparity must deviate from constant weights. This example serves only to illustrate the structural conflict in the theorem and does not form part of the proof.

Scope of the theorem. Theorem 6 does not claim that a specific functional form (e.g., $1/L$) is necessarily optimal; it merely states: when the typical score-sum magnitude under fixed effective training signal varies non-trivially with length, no unified weight function depending only on length can simultaneously satisfy P1 and P2.

Furthermore, the theorem only excludes weight functions that depend *solely on length*; more general estimator designs, such as weighting schemes that depend on token position, context, score geometry, or finer-grained credit assignment, are not within the scope of this exclusion.

Remark 7 (Essence of the conflict)**.** P1 requires that a uniform length weight does not alter the original trajectory-level policy gradient objective; P2 requires that this weight compensates for the non-trivial variation of score-sum magnitude with length. When P1 constrains $f(L)$ to be a constant function while P2 demands it to vary with length, the two become structurally irreconcilable.

3.1 Examples

For ease of understanding, we provide toy cases in sections A and B, analyzing how GRPO's length bias manifests as correct responses tending to be shorter and incorrect responses tending to be longer, while Dr. GRPO's length bias manifests as both correct and incorrect responses tending to be longer.

4 Corollaries and Analysis

4.1 Tradeoff Spectrum

Corollary 8 (Parametric Tradeoff Family)**.** *Consider the parametric family* $f_\alpha(L) = L^{\alpha-1}$, $\alpha \in [0,1]$*:*

- $\alpha = 0$*:* $f_0(L) = 1/L$ *— GRPO. Approximately satisfies P2 (length invariant) but violates P1 (biased gradient).*
- $\alpha = 1$*:* $f_1(L) = 1$ *— Dr. GRPO. Satisfies P1 (unbiased gradient) but violates P2 (length biased).*
- $\alpha \in (0,1)$*: intermediate tradeoff. Partially biased gradient, partially length-dependent.*

Gradient estimation bias is proportional to $|\alpha - 1|$ and length bias is proportional to α, establishing an inverse relationship.

fig. 2 visualizes this tradeoff.

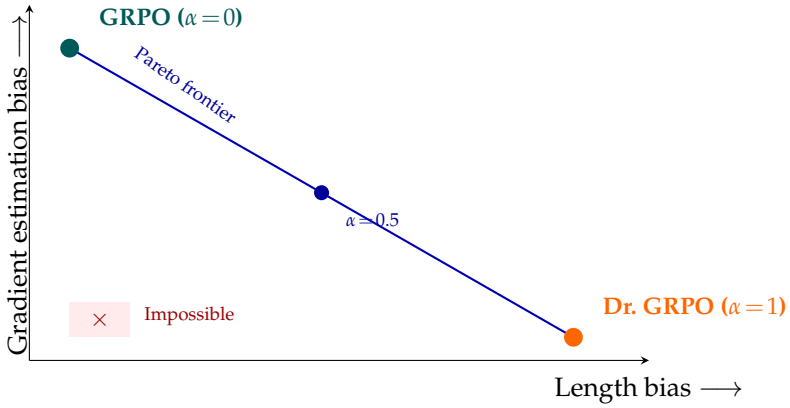

Figure 2: The impossibility tradeoff. The origin (zero bias on both axes) is unreachable. GRPO ($\alpha = 0$) and Dr. GRPO ($\alpha = 1$) occupy opposite ends of the Pareto frontier parameterized by $f_\alpha(L) = L^{\alpha-1}$.

4.2 Quantifying Dr. GRPO's Length Bias

Corollary 9 (Dr. GRPO's length bias)**.** *Under Dr. GRPO ($f(L) = 1$) with $G = 2$ and binary outcome reward, let $\mathbf{o}_1$ and $\mathbf{o}_2$ be two trajectories with lengths L_1 and L_2. Their advantages satisfy $|\tilde{A}_1| = |\tilde{A}_2| = 0.5$. The effective gradient weight of trajectory $\mathbf{o}_i$ is:*

$$w_i = \frac{L_i}{L_1 + L_2}. \tag{34}$$

For length ratio $r = L_{\max}/L_{\min}$, the longer trajectory captures:

$$w_{\text{long}} = \frac{r}{1+r} \tag{35}$$

of the total gradient magnitude, approaching 100% as $r \to \infty$. Under GRPO ($f(L) = 1/L$), $w_1 = w_2 = 0.5$, independent of length.

Proof. With $G = 2$ and binary reward, exactly one trajectory is correct ($R = 1$) and one incorrect ($R = 0$), giving $\text{mean}(\mathbf{R}) = 0.5$ and $|\tilde{A}_1| = |\tilde{A}_2| = 0.5$. Under Dr. GRPO, the gradient contribution magnitude of $\mathbf{o}_i$ is proportional to $f(|\mathbf{o}_i|) \cdot |\tilde{A}_i| \cdot |\mathbf{o}_i| = 1 \cdot 0.5 \cdot L_i$. The share is $w_i = L_i/(L_1 + L_2)$. Under GRPO, the contribution is $(1/L_i) \cdot 0.5 \cdot L_i = 0.5$, independent of length. □

Example 10 (Extreme case)**.** Let $G = 2$, $\mathbf{o}_1$ correct ($R = 1$, length 10 tokens), $\mathbf{o}_2$ incorrect ($R = 0$, length 10,000 tokens). The advantages are $\tilde{A}_1 = +0.5$, $\tilde{A}_2 = -0.5$. Under Dr. GRPO, $\mathbf{o}_2$'s gradient contribution is $10{,}000 \times 0.5 = 5{,}000$, while $\mathbf{o}_1$'s is only $10 \times 0.5 = 5$. The longer trajectory captures $\frac{5000}{5005} = 99.9\%$ of the gradient signal, nearly completely drowning out the reinforcement of the correct answer. Under GRPO, both contribute 50%. A step-by-step derivation of this example (including gradient decomposition and its effect on parameter updates) is given in section B.

table 1 shows the severity of this effect at various length ratios.

Example 11 (Practical relevance)**.** Liu et al. (2025) reported in their Table 5 that DeepSeek-R1-Zero produces correct answers averaging 4,965 tokens and incorrect answers averaging 8,206 tokens (a ratio of approximately 1:1.65). Under Dr. GRPO with $G = 2$, the incorrect (longer)

Table 1: Gradient weight shares of Dr. GRPO vs. GRPO at different length ratios ($G = 2$, binary reward). Under GRPO, both trajectories always receive equal weight.

Length ratio r	w_{long} (Dr. GRPO)	w_{short} (Dr. GRPO)	w_{long} (GRPO)	w_{short} (GRPO)
1:1	50.0%	50.0%	50.0%	50.0%
2:1	66.7%	33.3%	50.0%	50.0%
5:1	83.3%	16.7%	50.0%	50.0%
10:1	90.9%	9.1%	50.0%	50.0%
50:1	98.0%	2.0%	50.0%	50.0%
100:1	99.0%	1.0%	50.0%	50.0%

trajectory would capture approximately $\frac{8206}{4965+8206} \approx 62.3\%$ of the gradient, deviating 24.6 percentage points from the balanced 50%. While this proportion may appear moderate for a single update, the bias accumulates over hundreds of training iterations, systematically favoring longer responses.

4.3 Quantifying GRPO's Gradient Bias

For completeness, we also characterize the gradient bias introduced by GRPO.

Corollary 12 (GRPO's gradient bias)**.** *Under GRPO ($f(L) = 1/L$), the gradient estimator satisfies:*

$$\mathbb{E}[\hat{g}_{1/L}] - \nabla_\theta J = \mathbb{E}\left[\frac{1}{G}\sum_i \tilde{A}_i \left(\frac{1}{|\mathbf{o}_i|} - 1\right) \nabla_\theta \log \pi_\theta(\mathbf{o}_i|\mathbf{q})\right]. \tag{36}$$

This bias is non-zero when the trajectory length $|\mathbf{o}_i|$ is correlated with the score function $\nabla_\theta \log \pi_\theta(\mathbf{o}_i|\mathbf{q})$. This is generally always the case since the policy determines when the EOS token is generated.

Proof. By direct computation: $\mathbb{E}[\hat{g}_{1/L}] = \mathbb{E}\left[\frac{1}{G}\sum_i \tilde{A}_i \frac{1}{|\mathbf{o}_i|} \nabla_\theta \log \pi_\theta(\mathbf{o}_i|\mathbf{q})\right]$ and $\nabla_\theta J = \mathbb{E}\left[\frac{1}{G}\sum_i \tilde{A}_i \nabla_\theta \log \pi_\theta(\mathbf{o}_i|\mathbf{q})\right]$. The difference follows directly by linearity. Since $|\mathbf{o}_i|$ is determined by when π_θ generates the EOS token, $|\mathbf{o}_i|$ and $\nabla_\theta \log \pi_\theta(\mathbf{o}_i|\mathbf{q})$ are dependent, making the bias generally non-zero. □

4.4 Extension to General Group Size

Corollary 13 (General G + binary reward)**.** *For group size G with binary reward, if K out of G responses are correct, the advantages are $\tilde{A}_{\text{correct}} = 1 - K/G$ and $\tilde{A}_{\text{incorrect}} = -K/G$. Under Dr. GRPO, the effective weight of trajectory $\mathbf{o}_i$ is still proportional to $|\mathbf{o}_i| \cdot |\tilde{A}_i|$. The length bias exists for all G: longer trajectories always contribute more to the gradient, regardless of their correctness:*

$$w_i = \frac{|\mathbf{o}_i| \cdot |\tilde{A}_i|}{\sum_{j=1}^{G} |\mathbf{o}_j| \cdot |\tilde{A}_j|}. \tag{37}$$

5 Discussion

"Done Right" is a misnomer. Dr. GRPO (Liu et al., 2025), titled "Understanding R1-Zero-Like Training: A Critical Perspective," positions its contribution as fixing GRPO's optimization biases. The phrase "GRPO Done Right" implies a single correct formulation. Our impossibility theorem (theorem 6) shows this is not the case: GRPO and Dr. GRPO navigate different points on the inherent tradeoff between gradient unbiasedness and length invariance. Calling one of them "done right" obscures the fact that both make legitimate but different tradeoff choices.

When does the tradeoff matter? The practical importance of the tradeoff depends on the variance of response lengths. When all responses to a given prompt have similar lengths (e.g., simple arithmetic), the difference between $\alpha = 0$ and $\alpha = 1$ is negligible. When response lengths vary substantially, the choice of α materially affects training dynamics. This situation is typical in reasoning tasks: correct solutions may be concise while incorrect attempts tend to be verbose (DeepSeek-AI et al., 2026).

Practical guidance. While we do not propose a specific algorithm, our analysis suggests: (i) When response length variance is high, a smaller α (closer to GRPO) may be preferable to prevent longer trajectories from dominating the gradient. (ii) When gradient bias is the primary concern (e.g., early in training when the policy changes rapidly), a larger α (closer to Dr. GRPO) provides more accurate gradient estimates. (iii) The optimal α may vary across training phases, suggesting that a curriculum approach could be beneficial.

Implications for training dynamics. A practical implication of theorem 9 deserves attention: when long correct responses receive L times more reinforcement signal than short correct responses, the policy may gradually shift toward generating longer outputs. The complete causal chain from gradient dominance to behavioral change also involves clipping, learning rate, and multi-step optimization, which lie beyond the scope of our single-step analysis. However, the systematic asymmetry in gradient signals provides a necessary condition for this trend. Conversely, under GRPO ($\alpha = 0$), a 10-token short correct response and a 10,000-token long correct response receive the same total reinforcement signal. This provides no incentive at the gradient level to favor longer or shorter outputs.

Relationship to other biases. Our analysis complements Yang et al. (2026). The latter studies a different bias in GRPO: *difficulty bias*. This bias refers to the group-relative advantage estimator systematically underestimating advantages for difficult prompts and overestimating them for easy prompts. The length bias we identify is orthogonal, arising from within-group length variation rather than between-group difficulty variation. The standard deviation normalization in GRPO contributes to difficulty bias (Liu et al., 2025); our impossibility result is independent of whether $\text{std}(\mathbf{R})$ normalization is used.

Limitations. Our impossibility result is specific to the *outcome reward* setting, where each trajectory is assigned a scalar reward broadcast to all tokens. Under *process reward* (Schulman et al., 2018), different tokens receive different advantage estimates and the problem structure changes. The advantage is no longer constant across tokens, and the $\sum_t$ aggregation is no longer simply $|\mathbf{o}_i| \cdot \tilde{A}_i$. Extending the impossibility analysis to process rewards is an interesting future direction. Furthermore, our analysis focuses on single-step gradient estimators. The interaction between length bias and multi-step optimization dynamics (e.g., through PPO-style clipping) warrants further investigation.

6 Conclusion

We have established a fundamental impossibility result for group-based policy optimization under outcome rewards: gradient unbiasedness and length invariance cannot coexist. This reveals that GRPO and Dr. GRPO are not in a "biased" vs. "correct" relationship, but instead represent two principled tradeoff choices on the Pareto frontier. We hope this clarification helps the community make more informed algorithmic decisions, recognizing that the appropriate operating point depends on the specific characteristics of the training setting, especially the distribution of response lengths.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, 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 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2026. URL https://arxiv.org/abs/2501.12948.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL https://arxiv.org/abs/2503.24290.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free!, 2019. URL https://openreview.net/forum?id=r1lgTGL5DE.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018. URL https://arxiv.org/abs/1506.02438.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2 edition, 2018.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/bf00992696. URL http://dx.doi.org/10.1007/BF00992696.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your group-relative advantage is biased, 2026. URL https://arxiv.org/abs/2601.08521.

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. https://hkust-nlp.notion.site/simplerl-reason, 2025. Notion Blog.

A Intuitive Examples of Asymmetric Length Behavior from GRPO's Length Normalization

This section illustrates through simple examples that in GRPO, dividing by the length $|o_i|$ for each response leads to the following asymmetric phenomenon:

- Incorrect answers are more likely to grow longer (penalties are diluted);
- Correct answers do not obviously grow longer and may even be slightly suppressed.

A.1 Why Incorrect Answers Tend to Grow Longer (Penalty Dilution)

Consider two incorrect answers for the same question:

- Short incorrect answer: length 100;
- Long incorrect answer: length 1000;

Assume both have the same advantage (both negative, e.g., -0.5). In GRPO, the per-token weight is "advantage divided by length," so:

- Short incorrect answer: approximately -0.005 per token;
- Long incorrect answer: approximately -0.0005 per token.

We observe:

The per-token penalty for the long incorrect answer is only one-tenth that of the short one.

This means:

- If the model generates a shorter incorrect answer, it receives stronger per-token suppression;
- If the model generates a longer incorrect answer, the error is "spread out" and becomes harder to suppress.

Therefore, during training, the model is more likely to retain "verbose but incorrect" answers rather than "brief but incorrect" ones, manifesting as:

Incorrect answers gradually grow longer.

A.2 Why Correct Answers Do Not Obviously Grow Longer

Similarly, consider two correct answers:

- Short correct answer: length 100;
- Long correct answer: length 1000;

Assume both have the same advantage (both positive, e.g., $+0.5$). In GRPO:

- Short correct answer: approximately $+0.005$ per token;
- Long correct answer: approximately $+0.0005$ per token.

We observe:

The per-token reward for the long correct answer is smaller.

This means:

- Each token in the short correct answer is strongly reinforced;
- The long correct answer, while correct overall, receives weaker per-token reinforcement.

Therefore, increasing length does not bring stronger learning signals but instead "dilutes" existing rewards. The result is:

Correct answers are not noticeably pushed to grow longer; the advantage of long answers is even weakened.

A.3 Summary

Through the above examples, we can see that length normalization produces the following asymmetric effects:

- For incorrect answers: the longer the response, the weaker the per-token penalty $\Rightarrow$ more likely to grow longer;
- For correct answers: the longer the response, the weaker the per-token reward $\Rightarrow$ no obvious tendency to grow longer.

Therefore, length normalization in GRPO primarily manifests as a "lengthening effect" on incorrect answers, rather than a symmetric lengthening of all answers.

B Detailed Derivation of the Dr. GRPO Extreme Example

We provide a step-by-step derivation for theorem 10 ($G = 2$, $|\mathbf{o}_1| = 10$, $|\mathbf{o}_2| = 10{,}000$) to illustrate how Dr. GRPO's length bias manifests in gradient updates.

Setup. Let $\mathbf{o}_1$ be the correct response ($R(\mathbf{q}, \mathbf{o}_1) = 1$, length $|\mathbf{o}_1| = 10$ tokens) and $\mathbf{o}_2$ be the incorrect response ($R(\mathbf{q}, \mathbf{o}_2) = 0$, length $|\mathbf{o}_2| = 10{,}000$ tokens).

Step 1: Compute advantages.

$$\text{mean}(\mathbf{R}) = \frac{1+0}{2} = 0.5, \tag{38}$$

$$\tilde{A}_1 = 1 - 0.5 = +0.5 \quad \text{(correct trajectory: reinforce)}, \tag{39}$$

$$\tilde{A}_2 = 0 - 0.5 = -0.5 \quad \text{(incorrect trajectory: penalize)}. \tag{40}$$

Step 2: Dr. GRPO objective. Early in training, $\pi_\theta \approx \pi_{\theta_{\text{old}}}$, the importance ratio $\rho_{i,t} \approx 1$, and clipping does not activate:

$$J_{\text{Dr.GRPO}} = \frac{1}{2}\left[\sum_{t=1}^{10} \tilde{A}_1 + \sum_{t=1}^{10000} \tilde{A}_2\right] \tag{41}$$

$$= \frac{1}{2}\left[10 \times (+0.5) + 10{,}000 \times (-0.5)\right] \tag{42}$$

$$= \frac{1}{2}\left[5 - 5{,}000\right] = -2{,}497.5. \tag{43}$$

Step 3: Gradient decomposition. The gradient decomposes into contributions from both trajectories:

$$\nabla_\theta J_{\text{Dr.GRPO}} = \frac{1}{2}\left[\underbrace{(+0.5)\sum_{t=1}^{10}\nabla_\theta \log \pi_\theta(o_{1,t}|\cdot)}_{\text{from } \mathbf{o}_1\text{: 10 terms}} + \underbrace{(-0.5)\sum_{t=1}^{10000}\nabla_\theta \log \pi_\theta(o_{2,t}|\cdot)}_{\text{from } \mathbf{o}_2\text{: 10,000 terms}}\right]. \tag{44}$$

Step 4: Gradient magnitude analysis. Let g denote the typical per-token gradient norm $\mathbb{E}[\|\nabla_\theta \log \pi_\theta(o_t|\cdot)\|]$. The expected gradient magnitude from each trajectory is:

$$\|\text{gradient from } \mathbf{o}_1\| \approx 0.5 \times 10 \times g = 5g, \tag{45}$$

$$\|\text{gradient from } \mathbf{o}_2\| \approx 0.5 \times 10{,}000 \times g = 5{,}000g. \tag{46}$$

The ratio is $5g : 5{,}000g = 1 : 1{,}000$. The incorrect long trajectory dominates $\frac{5000}{5005} = 99.9\%$ of the gradient.

Step 5: Asymmetric effect on learning. This gradient dominance creates a systematic asymmetry:

Signal type	Source	# Tokens	Gradient magnitude
Reinforce correct answer	$\mathbf{o}_1$	10	$5g$
Penalize this specific incorrect answer	$\mathbf{o}_2$	10,000	$5{,}000g$

The parameter update is almost entirely determined by $\mathbf{o}_2$. The signal reinforcing the correct short answer ($\mathbf{o}_1$) accounts for only 0.1% of the total gradient.

Step 6: Comparison with GRPO. Under GRPO ($f(L) = 1/L$), the same example gives:

$$\|\text{gradient from } \mathbf{o}_1\| \approx \frac{1}{10} \times 0.5 \times 10 \times g = 0.5g, \tag{47}$$

$$\|\text{gradient from } \mathbf{o}_2\| \approx \frac{1}{10000} \times 0.5 \times 10{,}000 \times g = 0.5g. \tag{48}$$

The ratio is $0.5g : 0.5g = 1 : 1$. Both trajectories contribute equally regardless of the length difference.

Step 7: Reversed scenario. Now consider the opposite case: $\mathbf{o}_1$ is a long correct response ($|\mathbf{o}_1| = 10{,}000$) and $\mathbf{o}_2$ is a short incorrect response ($|\mathbf{o}_2| = 10$). Under Dr. GRPO, the correct trajectory captures $\frac{10000}{10010} = 99.9\%$ of the gradient, providing $10{,}000 \times 0.5 \times g = 5{,}000g$ of reinforcement signal to the long correct pattern, while the penalty signal for the short incorrect answer is only $5g$, which is negligible. This strongly reinforces the long response pattern, creating an incentive at the gradient level to favor longer outputs.

Summary. In both scenarios (long correct/short incorrect and short correct/long incorrect), the longer trajectory dominates the gradient under Dr. GRPO. When the long trajectory is correct, its reasoning pattern is strongly reinforced; when the long trajectory is incorrect, the reinforcement of the short correct answer is drowned out. Both effects create a systematic gradient-level preference for longer responses.

C Analysis at Non-Extreme Length Ratios

A natural concern is that the extreme example in section B (length ratio 1,000:1) may exaggerate the practical importance of length bias. In this appendix, we systematically analyze how the bias behaves across the full range of length ratios, including the moderate ratios commonly encountered in practice.

C.1 Gradient Weight as a Function of Length Ratio

By theorem 9, under Dr. GRPO ($G = 2$, binary reward), the longer trajectory's gradient weight is $w_{\text{long}} = r/(1+r)$, where $r = L_{\text{long}}/L_{\text{short}}$. table 2 provides a fine-grained breakdown from 1:1 to 100:1 ratios.

Table 2: Gradient weight shares under Dr. GRPO across the full range of length ratios ($G = 2$, binary reward). Correct trajectory length fixed at $L_{\text{correct}} = 100$ tokens; incorrect trajectory length varies.

$L_{\text{incorrect}}$	Ratio r	w_{correct} (Dr. GRPO)	$w_{\text{incorrect}}$ (Dr. GRPO)	Deviation from 50%	GRPO
50	0.5	66.7%	33.3%	16.7 pp	50/50
100	1.0	50.0%	50.0%	0 pp	50/50
150	1.5	40.0%	60.0%	10.0 pp	50/50
200	2.0	33.3%	66.7%	16.7 pp	50/50
300	3.0	25.0%	75.0%	25.0 pp	50/50
500	5.0	16.7%	83.3%	33.3 pp	50/50
800	8.0	11.1%	88.9%	38.9 pp	50/50
1,000	10.0	9.1%	90.9%	40.9 pp	50/50
2,000	20.0	4.8%	95.2%	45.2 pp	50/50

Even at a moderate 2:1 ratio (e.g., 100 vs. 200 tokens), the longer trajectory already captures 66.7% of the gradient, creating a 2× imbalance. At 5:1 (100 vs. 500 tokens, well within the range observed in practice), the longer trajectory captures 83.3%.

fig. 3 visualizes how the gradient weight share varies continuously with the length ratio, and fig. 4 shows the weight breakdown when pairing a fixed 100-token correct trajectory with incorrect trajectories of varying lengths.

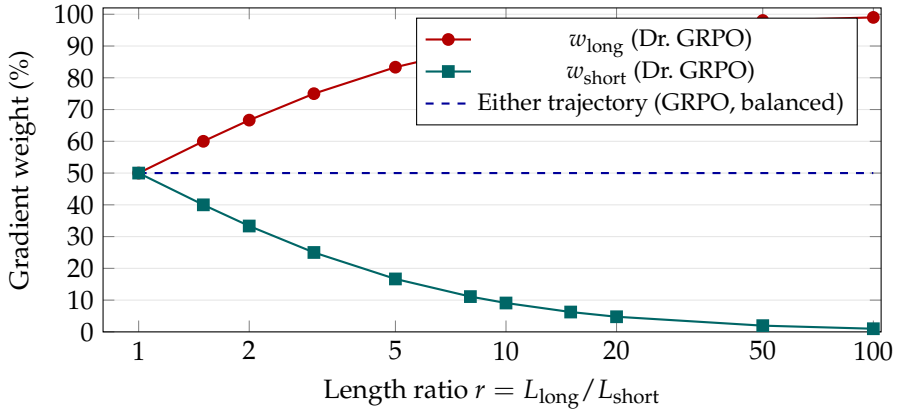

Figure 3: Gradient weight share vs. length ratio ($G = 2$, binary reward). Under Dr. GRPO, the long trajectory's weight increases monotonically with r, approaching 100% as $r \to \infty$. Under GRPO, both trajectories always receive exactly 50%, regardless of r.

C.2 Concrete Scenario Analysis

We analyze five representative scenarios to illustrate how the bias manifests in different training contexts:

Key observation. Even the "slightly longer incorrect" case (100 vs. 150 tokens, ratio 1.5:1) produces a 10 percentage point deviation from the balanced 50/50 allocation. The bias is exactly zero *only* when lengths are equal, consistent with our impossibility theorem.

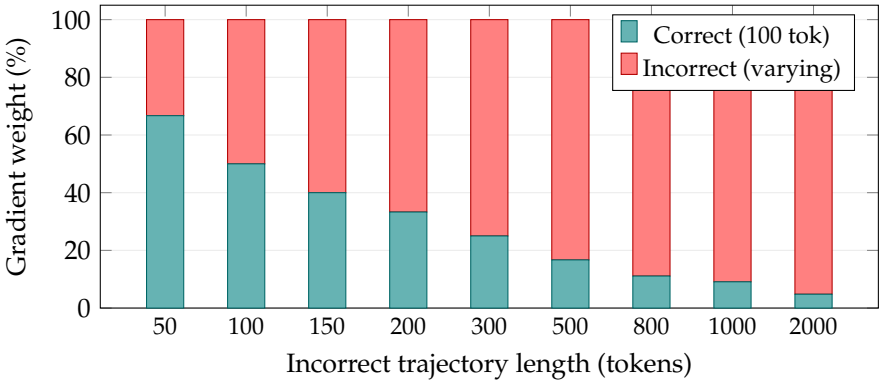

Figure 4: Dr. GRPO gradient weight breakdown: correct trajectory fixed at 100 tokens vs. incorrect trajectories of varying lengths ($G = 2$, binary reward). As the incorrect trajectory grows longer, it increasingly dominates the gradient. At 500 tokens ($5\times$ ratio), the incorrect trajectory captures 83.3% of the gradient.

Table 3: Gradient weights under Dr. GRPO for representative training scenarios ($G = 2$, binary reward). Bias > 0 indicates the correct trajectory dominates; bias < 0 indicates the incorrect trajectory dominates.

Scenario	L_{correct}	$L_{\text{incorrect}}$	w_{correct}	$w_{\text{incorrect}}$	Net bias
Long correct + short incorrect	1,000	100	90.9%	9.1%	$+0.818$
Short correct + long incorrect	100	1,000	9.1%	90.9%	-0.818
Equal length	500	500	50.0%	50.0%	0
Slightly longer correct	150	100	60.0%	40.0%	$+0.200$
Slightly longer incorrect	100	150	40.0%	60.0%	-0.200

C.3 Why Moderate Ratios Still Matter

One might argue that a 60/40 allocation (ratio 1.5:1) is tolerable for a single gradient step. We identify three reasons why even moderate ratios are practically significant:

(1) Cumulative effect. Each training step has its own length ratio drawn from the policy's output distribution. Even if the per-step bias is moderate (e.g., 60/40 allocation), the bias is *directionally consistent*. In LLM reasoning training, incorrect responses tend to be longer than correct ones (as documented in Liu et al. (2025) Table 5: correct 4,965 tokens vs. incorrect 8,206 tokens). Over hundreds of training steps, this unidirectional bias accumulates, systematically underweighting the reinforcement signal for correct answers.

(2) Positive feedback loop. When the policy shifts toward generating longer responses (due to length bias), the length ratios in subsequent rollouts increase, further amplifying the bias. For example, the policy may start generating 200-token incorrect responses (ratio 2:1, weight 66.7%), and once the bias pushes it toward longer outputs, future incorrect responses may grow to 500 tokens (ratio 5:1, weight 83.3%). This forms a self-reinforcing cycle.

(3) Interaction with binary reward. Under binary outcome rewards, training batches typically contain a mix of prompts with different correct/incorrect compositions. For difficult prompts, only one out of G responses may be correct. That sole correct response tends to be short (quickly finding the answer), while the $G - 1$ incorrect responses are long (trying multiple approaches, hitting the maximum length). The length bias systematically underweights the correct signal precisely when it is most valuable.

C.4 Length Bias Ratio: A Formal Metric

We define the *Length Bias Ratio* (LBR) to quantify how much each trajectory deviates from the ideal uniform weight:

Definition 14 (Length Bias Ratio)**.** For trajectory $\mathbf{o}_i$ among G trajectories under Dr. GRPO with equal advantage magnitudes:

$$\text{LBR}(\mathbf{o}_i) = \frac{|\mathbf{o}_i|}{\frac{1}{G}\sum_{j=1}^{G}|\mathbf{o}_j|} = \frac{G \cdot |\mathbf{o}_i|}{\sum_{j=1}^{G}|\mathbf{o}_j|}. \tag{49}$$

$\text{LBR} = 1$ indicates no length bias (the trajectory receives its "fair share" of gradient weight). $\text{LBR} > 1$ indicates over-representation; $\text{LBR} < 1$ indicates under-representation. Under GRPO ($f(L) = 1/L$), $\text{LBR} = 1$ for all trajectories.

Table 4: Length Bias Ratio for selected length pairs under Dr. GRPO ($G = 2$).

$\lvert\mathbf{o}_1\rvert$	$\lvert\mathbf{o}_2\rvert$	$\text{LBR}(\mathbf{o}_2)/\text{LBR}(\mathbf{o}_1)$	Interpretation
100	100	1.0×	Balanced
100	200	2.0×	$\mathbf{o}_2$ receives 2× more weight
100	500	5.0×	$\mathbf{o}_2$ receives 5× more weight
100	1,000	10.0×	$\mathbf{o}_2$ receives 10× more weight
50	500	10.0×	$\mathbf{o}_2$ receives 10× more weight

LBR grows *linearly* with the length ratio, with no diminishing returns or saturation. This is a direct consequence of Dr. GRPO's $f(L) = 1$ weighting: removing the $1/|\mathbf{o}_i|$ normalization makes gradient contributions strictly proportional to trajectory length.