

**DEEPSEEK'S
GRPO NOTATION
& WHY IT'S WRONG**

by
Jordan Jiosi

Abstract

Group Relative Policy Optimization (GRPO) can be viewed as a PPO-style policy-gradient update in which the advantage signal is constructed from within-group comparisons rather than from a learned value baseline. This paper reformulates DeepSeek’s GRPO objective to make explicit the distinction between

- (i) the PPO/TRPO importance ratio used for off-policy correction
- (ii) the reference-anchored pointwise log-ratio term that is often described as a KL regularizer

This paper’s central claim is not that GRPO is empirically ineffective; rather, under the expectation actually used in the algorithm, the regularization term is a stability proxy and does not literally instantiate KL-regularized optimization. DeepSeek-R1’s presentation therefore conflates a heuristic penalty with a KL-regularized objective as written in their published paper, even though the resulting training procedure can still work well in practice.

Broadly, this paper isolates the mechanics of Group Relative Policy Optimization (GRPO) and reformulates the objective function in a way that makes the role of the reference-policy penalty explicit.

1 Group Relative Policy Optimization (GRPO)

Group Relative Policy Optimization (GRPO) is a group-normalized policy-gradient objective made popular by DeepSeek’s 2024 paper [1]. The core idea is to replace a single-sample advantage estimate with a *relative* advantage computed within a small group of candidate actions, or trajectories, conditioned on the same context, for example the same state or the same initial seeded state. This yields a variance-reduced learning signal while retaining the same fundamental policy-gradient structure used in trust-region style policy optimization, for example PPO-style clipped objectives and KL-anchored updates.

This analysis concerns the interpretation and notation of the GRPO objective rather than its empirical performance. The method itself is not claimed to be incorrect; rather, the focus is on how its regularization term is framed and justified relative to standard policy-gradient and trust-region formulations.

Consider a context variable s denoting a state. Let a policy $\pi_\theta(a|s)$ generate a group of G candidate actions $\{a_i\}_{i=1}^G$, each receiving a scalar return or reward $r_i \approx r(s, a_i)$. Then, we define the group baseline as the mean $\bar{r}(s) \approx \frac{1}{G} \sum_{i=1}^G r(s, a_i)$ and define the group-relative advantage

$$\hat{A}_i \approx r(s, a_i) - \bar{r}(s) \implies \tilde{A}_i \approx \frac{\hat{A}_i}{\sum_r(s) + \epsilon} \implies \sum_r(s) \approx \sqrt{\frac{1}{G} \sum_{i=1}^G (r(s, a_i) - \bar{r}(s))^2}$$

Next, we define a behavior policy $\pi_{\theta_{\text{old}}}$ and a fixed reference policy π_{ref} .

As in standard importance-weighted policy-gradient updates, including TRPO/PPO-style implementations, the importance sampling ratio is taken with respect to the behavior policy:

$$\rho_i(\theta) \approx \frac{\pi_\theta(a_i|s)}{\pi_{\theta_{\text{old}}}(a_i|s)}.$$

This ratio plays the same mechanical role as in PPO; no novel estimator is introduced at this stage.

This is the same importance-weighted policy gradient structure used in PPO and TRPO, where updates are performed under an off-policy sampling distribution but corrected via likelihood ratios.

In this sense, GRPO does not deviate from PPO at the level of gradient estimation; the distinction lies entirely in how the advantage signal is constructed and normalized.

Separately, GRPO introduces an additional reference-anchored log-ratio term, which is *not* the PPO importance ratio:

$$\ell_i(\theta) \approx \log \pi_\theta(a_i|s) - \log \pi_{\text{ref}}(a_i|s).$$

1.1 KL divergence & Trust-Region Updates

This subsection provides the formal grounding required to interpret the GRPO objective where the role of KL-based stability constraints is made explicit.

Policy-gradient methods can become unstable when the policy update is too large; a single bad step can dramatically change the action distribution and degrade performance. A common stabilization approach is to keep the updated policy close to a reference distribution, often the previous policy $\pi_{\theta_{\text{old}}}$ or a fixed reference policy π_{ref} .

The standard measure used to quantify distributional change is the Kullback-Leibler (KL) divergence, a relative-entropy functional:

$$D_{\text{KL}}(P\|Q) \doteq \mathbb{E}_{x \sim P} \left[\log \left(\frac{P(x)}{Q(x)} \right) \right], \quad (1)$$

As formalized in Eq. (Equation 1), this definition is only meaningful as a divergence when the expectation is taken under the same distribution whose deviation is being measured. When this expectation is replaced by off-policy sampling, the expression no longer represents a KL divergence in the information-theoretic sense, even though the pointwise logarithmic form remains unchanged. and for policies conditioned on context s ,

$$D_{\text{KL}}(\pi_\theta(\cdot|s) \parallel \pi_{\text{ref}}(\cdot|s)) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\log \frac{\pi_\theta(a|s)}{\pi_{\text{ref}}(a|s)} \right].$$

A *trust-region style* update can be stated abstractly as maximizing a surrogate improvement term

while constraining KL drift, for example

$$\max_{\theta} \mathbb{E}_{s,a \sim \pi_{\theta_{\text{old}}}} [\rho(\theta) A(s,a)] \quad \text{s.t.} \quad \mathbb{E}_s [D_{\text{KL}}(\pi_{\theta}(\cdot|s) \parallel \pi_{\theta_{\text{old}}}(\cdot|s))] \leq \delta,$$

or equivalently by moving the constraint into the objective as a penalty under a matching expectation. PPO-style algorithms approximate this behavior in practice via clipping and, optionally, by monitoring or penalizing KL.

The GRPO term $\ell_i(\theta)$ is a *pointwise* log-likelihood ratio random variable: it becomes a KL divergence only after taking an expectation under the action distribution of the policy in the numerator. This distinction is exactly what matters when one later claims “KL regularization” while sampling actions under a different policy.

This mirrors the approximation already employed in PPO, where the KL term is used as a trust-region heuristic rather than as an unbiased divergence estimator. In both cases, stability arises from conservative updates rather than from exact optimization of a KL-constrained objective.

Rewriting DeepSeek’s GRPO objective to make its assumptions explicit, and placing it side-by-side with the PPO surrogate objective, one obtains:

$$\mathcal{J}_{\text{GRPO}}(\theta) \approx \mathbb{E}_{s,a_{1:G} | \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G (\rho_i(\theta) \tilde{A}_i - \beta \ell_i(\theta)) \right] \quad (2)$$

where θ_{old} denotes the behavior policy used to sample the group, and $\beta > 0$ is a regularization weight.

While Eq. (Equation 2) is commonly described as KL-regularized, the regularization term as it appears is a pointwise log-ratio surrogate rather than an actual Kullback-Leibler divergence in the strictest sense as composed in DeepSeek’s R1 paper [2].

Two remarks are important for correct interpretation. First, $\ell_i(\theta)$ is a pointwise log-likelihood ratio and *not* itself a KL divergence. Recall from KL divergence that the KL divergence is an expectation

over the policy’s action distribution conditioned on context:

$$D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \approx \mathbb{E}_{s,a|\pi_\theta} \left[\log \left(\frac{\pi_\theta(a|s)}{\pi_{\text{ref}}(a|s)} \right) \right] = \mathbb{E}_{s,a|\pi_\theta} [\ell(\theta)]$$

Thus, adding $-\beta \ell_i(\theta)$ to the per-sample objective can be viewed as a stochastic estimator of a KL-type regularizer only after taking an appropriate expectation but when taken pointwise, it is simply a log-ratio penalty whose interpretation depends entirely on how the expectation is taken, as formalized in Section 2: DeepSeek’s Problematic GRPO Notation: A Novel Analysis.

And this is the critical point: GRPO’s regularization term is mathematically a pointwise log-ratio penalty evaluated under an off-policy distribution. It only becomes a KL divergence after taking an expectation under the current policy, which the algorithm does not perform. The method remains effective in practice, but the interpretation as KL-regularized optimization is formally incorrect.

As detailed in Section section 2, this is better understood as the use of a stability proxy under imperfect modeling assumptions rather than as a literal failure of KL-regularized policy optimization. The method does not fail because it misuses the KL divergence; rather, the misuse reflects a deeper issue: the optimization objective is a proxy for stability, not a literal instantiation of an MDP-constrained problem.

Secondly, GRPO is closely related to PPO-style trust-region optimization. If one introduces PPO clipping on the ratio, one obtains the clipped surrogate:

$$\mathcal{J}_{\text{GRPO-clip}}(\theta) \approx$$

$$\mathbb{E}_{s,a_{1:G}|\pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G \min \left(\rho_i(\theta) \tilde{A}_i, \text{clip}(\rho_i(\theta), 1 - \varepsilon, 1 + \varepsilon) \tilde{A}_i \right) - \beta \ell_i(\theta) \right]$$

recalling $\varepsilon > 0$ is the PPO clipping parameter.

1.2 PPO vs GRPO

GRPO retains the core PPO mechanisms under-the-hood. The update still uses importance sampling ratios, a clipped surrogate objective function, and trust-region like behavior that keeps steps small

in practice.

The substantive difference lies in GRPO removing the learned value function by replacing generalized advantage estimation style advantage estimation with group-normalized rewards, hence the nomenclature paradigm *Group Relative Policy Optimization*.

The trade-off is theoretical clarity for lower variance and reduced resource use, which is a reasonable computational engineering choice, but it weakens the clean interpretation of the regularizer as a KL penalty when combined with off-policy sampling like PPO maintains and enjoys.

The empirical effectiveness of GRPO likely arises from small policy updates, comparative reward modeling, and the surrounding training pipeline rather than from the theoretical properties of the regularization term itself. A more accurate characterization would describe GRPO as PPO with group-based advantage estimation and heuristic entropy-style regularization, rather than as a clean KL-regularized objective.

2 DeepSeek’s Problematic GRPO Notation: A Novel Analysis

2.1 What Went Wrong

DeepSeek’s GRPO’s objective function is described as a KL-regularized variant of PPO. While the algorithm is empirically effective, the formulation conflates several distinct quantities, leading to a misleading interpretation of the regularization term as written in the paper. This section should be read as a concrete instantiation of the modeling/notation lens developed in Warren Powell’s Unified Decision Analysis Framework [3].

2.2 Pointwise Log-Ratios \neq KL Divergences

Recall: the Kullback-Leibler divergence between two policies π_θ and π_{ref}

$$D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) = \mathbb{E}_{s,a \mid \pi_\theta} \left[\log \left(\frac{\pi_\theta(a \mid s)}{\pi_{\text{ref}}(a \mid s)} \right) \right]$$

Without the expectation under π_θ , the quantity $\log\left(\frac{\pi_\theta}{\pi_{\text{ref}}}\right)$ is merely a log-likelihood ratio random variable. Referring to a single-sample realization of this ratio as “the KL” is misleading at best and formally incorrect at worst.

DeepSeek-R1 instead introduces the nonnegative function

$$\frac{\pi_{\text{ref}}(a|s)}{\pi_\theta(a|s)} - \log\left(\frac{\pi_{\text{ref}}(a|s)}{\pi_\theta(a|s)}\right) - 1$$

and labels it an estimator of $D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})$.

This expression corresponds to the convex dual form of the KL divergence and is an unbiased estimator *only* when the expectation is taken under the state-action visitation distribution induced by π_θ .

2.3 Off-Policy Sampling Breaks DeepSeek’s Unbiasedness Claim

In GRPO, the expectation is taken under $\mathbb{E}_{s,a|\pi_{\theta_{\text{old}}}}$ rather than $\mathbb{E}_{s,a|\pi_\theta}$ such that the KL surrogate is inserted into the objective without importance weighting.

Consequently, three critical issues arise.

First off, the surrogate is not, in general, an unbiased estimator of $D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})$ because the expectation is taken under $\pi_{\theta_{\text{old}}}$ rather than π_θ .

Secondly, the usual control-variate interpretation does not strictly apply, because the expectation is taken under a different measure than the one defining the KL; this does not make the method incorrect, but it does invalidate a literal KL interpretation.

Finally, even though the surrogate is nonnegative, that property is orthogonal to whether it estimates the claimed KL divergence.

In contrast, PPO implementations that monitor or penalize KL divergence either compute it exactly from logits or ensure the expectation is taken under the correct policy distribution.

2.4 Practical Implications

Practically, the regularization term behaves more like an entropy-style constraint. The reference policy π_{ref} is treated as fixed and does not contribute gradients directly; instead, it shapes the gradient of $\log \pi_{\theta}$ so as to discourage large deviations from the reference distribution.

This is not inherently incorrect; it is a common and effective heuristic. However, it is not equivalent to optimizing a well-defined KL-regularized reinforcement learning objective under off-policy sampling as defined in the DeepSeek-R1 paper; presenting it as such conflates a heuristic regularization penalty with a KL-regularized objective.

References

- [1] Z. Shao et al., “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024, arXiv:2402.03300v3. arXiv: 2402.03300 [cs.CL].
- [2] DeepSeek-AI, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025, arXiv:2501.12948v1. arXiv: 2501.12948 [cs.CL].
- [3] W. B. Powell, *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*. Wiley, 2022.