DEEPSEEK'S
GRPO NOTATION
& WHY IT'S WRONG

by
Jordan Jiosi

Preprint DOI: https://doi.org/10.5281/zenodo.18396691

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 Proximal Policy Optimization (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

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

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:

$$\rho_i(\theta) \approx \frac{\pi_\theta(a_i|s)}{\pi_{\theta_{\text{old}}}(a_i|s)} \tag{1}$$

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

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

importance ratio:

$$\ell_i(\theta) \approx \log \pi_\theta(a_i|s) - \log \pi_{\text{ref}}(a_i|s) \tag{2}$$

1.1 KL Divergence, Trust Regions, and What GRPO Actually Optimizes

Policy-gradient methods are known to be unstable when policy updates are too large; even a single step that substantially alters the action distribution can degrade performance or collapse training. A common stabilization strategy is therefore to constrain successive policies to remain close in distribution, most often through a Kullback–Leibler (KL) divergence penalty or constraint.

The KL divergence between two distributions P and Q is defined as

$$D_{\text{KL}}(P\|Q) \doteq \mathbb{E}_x\left[\log\left(\frac{P(x)}{Q(x)}\right)\right] \quad \text{ɜ} \quad x \sim P \tag{3}$$

Crucially, this quantity is only a KL divergence when the expectation is taken under the same distribution whose deviation is being measured. If the expectation is taken under a different distribution, the expression remains a log-ratio, but it no longer corresponds to a KL divergence in the information-theoretic sense.

For notational clarity, we define the conditional action distribution

$$\pi_\theta^s(a) \doteq \pi_\theta(a|s),$$

so that KL divergences between policies conditioned on the same context may be written without ambiguity.

For policies conditioned on context s, this specialization takes the form

$$D_{\text{KL}}(\pi_\theta^s \,\|\, \pi_{\text{ref}}^s) = \mathbb{E}_{a\sim\pi_\theta^s}\left[\log \frac{\pi_\theta^s(a)}{\pi_{\text{ref}}^s(a)}\right] \tag{4}$$

This distinction is not cosmetic. The KL divergence is an expectation over the *current* policy, and its interpretation depends entirely on that choice of measure.

Trust-Region Methods

In Trust-Region Policy Optimization (TRPO), this idea appears as a constraint or penalty that limits how far an updated policy may drift from a reference distribution. Abstractly, such updates under the rollout assume the form

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

In practice, PPO does not solve this constrained optimization problem directly. Instead, it approximates trust-region behavior through clipping and, optionally, by monitoring or penalizing the empirical KL divergence. The resulting update is heuristic but effective: stability arises from small policy steps rather than from exact enforcement of a KL constraint.

What GRPO Inherits From PPO

GRPO introduces a reference-anchored log-ratio term

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

which is a pointwise random variable. By itself, this quantity is not a KL divergence. It becomes one only after taking an expectation under the same policy that appears in the numerator.

Where GRPO Breaks

GRPO evaluates this log-ratio under samples drawn from the behavior policy $\pi_{\theta_{\text{old}}}$ rather than from the current policy π_θ. When written explicitly as

$$\mathbb{E}_{s,a|\pi_{\theta_{\text{old}}}}\left[\log\left(\frac{\pi_\theta(a|s)}{\pi_{\text{ref}}(a|s)}\right)\right] \quad \text{э} \quad s\sim d^{\pi_\theta},\, a\sim\pi_\theta(a|s)$$

the result is that GRPO's regularization term is not a KL divergence in the strict sense. It is a pointwise log-ratio penalty evaluated under an off-policy distribution.

This mirrors the approximation already employed in PPO, where trust-region behavior arises from clipping or auxiliary KL monitoring rather than from optimizing a true KL-constrained objective. DeepSeek's formulation implicitly assumes this same identification, treating the off-policy expectation as if it were taken under the current policy, even though this assumption is never stated explicitly.

GRPO's Objective Reformulated

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

$$\mathscr{J}_{\text{GRPO}}(\theta) \approx \mathbb{E}_{s,a_{1:G}|\pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\rho_i(\theta) \tilde{A}_i - \beta \, \ell_i(\theta) \right) \right] \tag{5}$$

where θ_{old} denotes the behavior policy used to sample the group, and $\beta > 0$ is a regularization weight [2].

If one introduces PPO-style clipping on the ratio, one obtains the clipped surrogate:

$$\mathscr{J}_{\text{GRPO-clip}}(\theta) \approx \mathbb{E}_{s,a_{1:G}|\pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \min\Big(\rho_i(\theta) \tilde{A}_i, \; \text{clip}\big(\rho_i(\theta), 1-\varepsilon, 1+\varepsilon\big) \tilde{A}_i \Big) - \beta \, \ell_i(\theta) \right] \tag{6}$$

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

GRPO's Objective Reformulated With Gradients

This subsection makes the optimization signal in (Equation 5) explicit by differentiating it with respect to θ. For a single context s and a sampled group $\{a_i\}_{i=1}^{G}$ drawn from the behavior policy $\pi_{\theta_{\text{old}}}$, recall (Equation 1) and (Equation 2).

Recall the pointwise log-ratio in (Equation 2) as a function of the sampled action $\ni \ell_i(\theta) = \ell(a_i, s; \theta)$.

Since the denominator of $\rho_i(\theta)$ is independent of θ,

$$\nabla_\theta \rho_i(\theta) = \rho_i(\theta) \nabla_\theta \log \pi_\theta(a_i|s). \tag{7}$$

Similarly, because π_{ref} is fixed,

$$\nabla_\theta \ell_i(\theta) = \nabla_\theta \log \pi_\theta(a_i|s). \tag{8}$$

The term $\nabla_\theta \log \pi_\theta(a|s)$ is the Jacobian of the scalar log-likelihood with respect to the parameter vector θ. This is the sense in which policy-gradient updates are "first-order": the Jacobian, the gradient, gives the local sensitivity of the log-probability to an infinitesimal parameter change, and therefore determines the immediate direction of improvement.

Consequently, for a small step $\Delta\theta$, a first-order expansion gives

$$\log \pi_\theta(a|s) \approx \log \pi_{\theta_{\text{old}}}(a|s) + \big(\nabla_\theta \log \pi_\theta(a|s)\big)^\top \Delta\theta,$$

making explicit that the update is driven by this first-order sensitivity called the score function.

Plugging (Equation 7) & (Equation 8) into (Equation 5) yields a policy-gradient form:

$$\nabla_\theta \mathscr{J}_{\text{GRPO}}(\theta) \approx \mathbb{E}_{s,a_{1:G}|\pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \Big(\rho_i(\theta) \tilde{A}_i - \beta \Big) \nabla_\theta \log \pi_\theta(a_i|s) \right]. \tag{9}$$

(Equation 9) makes two points immediate.

1. The reference policy does not contribute gradients.

Although $\ell_i(\theta)$ is written as a log-ratio against π_{ref}, the gradient of the regularizer depends only on $\nabla_\theta \log \pi_\theta$. In particular,

$$\nabla_\theta \, \mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(a|s)} \big[\beta \, \ell(a,s;\theta)\big] = \beta \, \mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(a|s)} \big[\nabla_\theta \log \pi_\theta(a|s)\big]. \tag{10}$$

The term $\log \pi_{\text{ref}}(a|s)$ is constant with respect to θ and therefore disappears under differentiation.

2. The regularizer is a forward-KL-style penalty under the behavior distribution.

Define, for any fixed s, the forward KL

$$D_{\mathrm{KL}}(\pi_{\theta_{\text{old}}}(a|s) \,\|\, \pi_\theta(a|s)) \doteq \mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(a|s)}\left[\log \frac{\pi_{\theta_{\text{old}}}(a|s)}{\pi_\theta(a|s)}\right]. \tag{11}$$

Since $\mathbb{E}_{\pi_{\theta_{\text{old}}}}[\log \pi_{\theta_{\text{old}}}]$ does not depend on θ, we have

$$\nabla_\theta D_{\mathrm{KL}}(\pi_{\theta_{\text{old}}}(a|s) \,\|\, \pi_\theta(a|s)) = -\,\mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(a|s)}\left[\nabla_\theta \log \pi_\theta(a|s)\right]. \tag{12}$$

Comparing (Equation 12) with (Equation 10) shows that GRPO's log-ratio regularizer induces the same gradient, up to a sign, as minimizing a forward KL from the behavior policy into the current policy. Equivalently, the regularizer in (Equation 5) acts as a cross-entropy term $-\,\mathbb{E}_{\pi_{\theta_{\text{old}}}}[\log \pi_\theta]$.

Hessians & The Trust-Region Interpretation

Trust-region methods motivate constraining policy changes via a KL divergence because, for small parameter steps $\Delta\theta \doteq \theta - \theta_{\text{old}}$, the KL admits a local quadratic approximation. For each s, (Equation 11) undergoes a second-order Taylor expansion around θ_{old} to yield

$$D_{\mathrm{KL}}(\pi_{\theta_{\text{old}}}(a|s) \,\|\, \pi_\theta(a|s)) \approx \tfrac{1}{2}\,\Delta\theta^\top F_s(\theta_{\text{old}})\,\Delta\theta, \tag{13}$$

where $F_s(\theta_{\text{old}})$ is the conditional Fisher information matrix which is the Hessian of the KL at θ_{old}. This is the sense in which enforcing a KL trust region is "second-order": the Hessian imposes the curvature of the KL to determine the geometry of the update.

By contrast, (Equation 5)'s GRPO regularizer contributes a first-order gradient term proportional to $\mathbb{E}_{\pi_{\theta_{\text{old}}}}[\nabla \log \pi_\theta]$. It does not implement the quadratic trust-region penalty in (Equation 13) directly, and it cannot be interpreted as the gradient of a reverse KL to π_{ref} without changing the sampling distribution.

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.

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.

In short, GRPO inherits PPO's stability heuristics but not its theoretical KL interpretation.

2 DeepSeek's Problematic GRPO Notation: A Novel Analysis

2.1 What Went Wrong

DeepSeek describes GRPO 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.

2.2 Pointwise Log-Ratios $\neq$ KL Divergences

This is the key distinction: a log-ratio evaluated at a single sample is not a divergence, but merely a scalar random variable. Recall the Kullback-Leibler divergence definition between two policies

π_θ and π_{ref} in (Equation 4): without the expectation under π_θ, the quantity $\log\left(\frac{\pi_\theta}{\pi_{\text{ref}}}\right)$ is simply a log-likelihood ratio. DeepSeek's reference 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 $D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ [2].

This expression corresponds to a variational convex-analytic representation of the KL divergence and yields an unbiased estimator *only* when the expectation is taken under the state-action visitation distribution induced by π_θ. That is to say, the KL divergence only holds under expectation with respect to the target distribution.

2.3 Off-Policy Sampling Breaks DeepSeek's Unbiasedness Claim

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

Consequently, three critical issues arise.

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

Second, 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.

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].