Diverse Group Relative Policy Optimization (DGRPO): Upweighting Rare, Accurate Solutions for STEM and Art

Libor Burian
burian.lib@gmail.com

March 24, 2025

Abstract

Language models trained with reinforcement learning often converge to common solution patterns, limiting their creative problem-solving capabilities. This paper introduces Diverse Group Relative Policy Optimization (DGRPO), an extension of Group Relative Policy Optimization (GRPO) that specifically addresses this limitation. While GRPO normalizes rewards within groups of responses to promote accuracy, DGRPO incorporates solution diversity into the advantage calculation through two novel approaches: (1) upweighting less likely but correct tokens to incentivize rare solutions, and (2) quantifying solution uniqueness using cosine similarity of neural embeddings. By introducing a configurable diversity weight parameter, DGRPO allows practitioners to balance accuracy with exploration of diverse solution strategies. My approach encourages language models to discover multiple valid approaches to problems, a critical capability for applications in scientific discovery, mathematical problem-solving, creative coding, and art. DGRPO demonstrates how reinforcement learning can be adapted to reward not just correctness but also the novelty and diversity of solutions in generative AI systems.

1 Background: Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) was introduced as an efficient algorithm for training language models on reasoning tasks, particularly in the DeepSeek-R1 model [Shao et al., 2024]. GRPO builds on the Proximal Policy Optimization (PPO) algorithm but eliminates the need for a separate critic model by estimating baselines from group scores.

For each question q, GRPO samples a group of outputs $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and optimizes the policy model π_θ by maximizing an objective function.

where ϵ and β are hyperparameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})} \tag{2}$$

While GRPO has demonstrated impressive performance on reasoning tasks, its objective function focuses solely on accuracy (as measured by rewards), which can lead to homogenized solution strategies over time.

2 Method 1

Both GRPO and DGRPO are reinforcement learning techniques for training language models, with DGRPO being an extension of GRPO that promotes diversity in correct solutions.

2.1 Key Difference

The core difference is that **GRPO** optimizes for relative performance within a group of responses, while **DGRPO**

$$J_{GRPO}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)]}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i,\right.\right.\right.$$
$$\left.\left.\left.\text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right)\right.$$
$$\left.\left.-\beta D_{KL}\big(\pi_\theta||\pi_{ref}\big)\right)\right]$$

1: The GRPO objective function

additionally rewards less likely (but correct) responses to encourage diversity in solutions.

2.2 Mathematical Explanation

2.2.1 GRPO

In standard GRPO, rewards are normalized within a group of responses to the same prompt:

1. For each prompt x, generate k responses $y_1, y_2, \ldots, y_k$
2. Compute rewards $r_1, r_2, \ldots, r_k$ for each response
3. Calculate the normalized advantage for each response:

$$A_i = r_i - \mu_r$$

Where μ_r is the mean reward across all responses for that prompt.

2.2.2 DGRPO

DGRPO extends this by adding a diversity bonus to correct solutions:

1. Generate k responses and compute base rewards as in GRPO
2. For correct solutions, calculate token probabilities $p(y_i)$
3. Compute the average probability $\bar{p}$ across all correct solutions
4. Calculate a diversity bonus for each correct solution:

$$\text{diversity\_bonus}_i = \max(0, 1 - \frac{p(y_i)}{\bar{p}})$$

5. Final reward becomes:

$$R_i = \begin{cases} r_i + \alpha \cdot \text{diversity\_bonus}_i & \text{if solution is correct} \\ r_i & \text{otherwise} \end{cases}$$

Where α is a scaling factor (set to 0.5 in the implementation).

2.3 Possible benefits of DGRPO over GRPO

1. **Increased Diversity**: DGRPO encourages the model to find multiple correct approaches to the same problem
2. **Better Generalization**: By exploring a wider solution space, the model may generalize better to new problems

DGRPO effectively combines the group-relative optimization of GRPO with a mechanism to promote diversity in model outputs, creating a more versatile and creative reasoning system.

3 Method 2

3.1 Methodology: Diverse Group Relative Policy Optimization (DGRPO)

I propose Diverse Group Relative Policy Optimization (DGRPO), which extends GRPO by incorporating solution diversity into the advantage calculation. My approach quantifies solution rarity using cosine similarity and introduces a configurable parameter to control the trade-off between accuracy and diversity.

3.1.1 Solution Embedding

I convert each solution text s_i to a dense vector embedding $\vec{v_i}$ using a pre-trained neural embedding model:

$$\vec{v_i} = \text{Embed}(s_i) \tag{3}$$

where $\text{Embed}(\cdot)$ represents a neural embedding function that maps text to a high-dimensional vector space. These embeddings capture the semantic content of solutions, allowing for more meaningful comparisons than surface-level lexical matching. The embedding model projects solutions into a space where semantically similar approaches appear closer together, regardless of specific wording or syntactic variations.

3.1.2 Rarity Score Calculation

For each solution, we compute its average similarity to all other solutions in the group:

$$\text{sim}(i,j) = \cos(\vec{v_i}, \vec{v_j}) = \frac{\vec{v_i} \cdot \vec{v_j}}{||\vec{v_i}|| \cdot ||\vec{v_j}||} \tag{4}$$

$$\text{avg\_sim}_i = \frac{1}{G} \sum_{j=1}^{G} \text{sim}(i,j) \tag{5}$$

We then convert this to a rarity score by inverting and normalizing:

$$\text{rarity}_i = 1 - \frac{\text{avg\_sim}_i}{\max_k \text{avg\_sim}_k} \tag{6}$$

This formulation ensures that solutions with lower average similarity (i.e., more unique solutions) receive higher rarity scores. By using neural embeddings instead of lexical features, my approach can identify semantically unique solutions even when they share surface-level similarities with common approaches.

3.1.3 Diversity-Weighted Advantage Calculation

I combine the original reward signal with the rarity score using a diversity weight parameter $\lambda \in [0, 1]$:

$$\text{combined\_score}_i = (1-\lambda) \times \text{normalized\_reward}_i + \lambda \times \text{rarity}_i \tag{7}$$

where $\text{normalized\_reward}_i$ scales the original rewards to the [0,1] range:

$$\text{normalized\_reward}_i = \frac{r_i - r_{\min}}{r_{\max} - r_{\min}} \tag{8}$$

where $r_{\min} = \min(\{r_1, r_2, \ldots, r_G\})$ and $r_{\max} = \max(\{r_1, r_2, \ldots, r_G\})$.

I then compute advantages using the standard GRPO formula but with combined scores:

$$A_i^{DGRPO} = \frac{\text{combined\_score}_i - \text{mean}(\{\text{combined\_score}_1, \ldots, \text{combined\_}}{\text{std}(\{\text{combined\_score}_1, \ldots, \text{combined\_score}_G\})} \tag{9}$$

3.1.4 Objective Function

The DGRPO objective function remains structurally similar to GRPO but uses the diversity-weighted advantages

This objective encourages the model to generate not only correct solutions but also diverse approaches to solving problems.

References

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Minghua Zhang, Yushi Li, Yu Wu, and Daya Guo. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2024.

$$J_{DGRPO}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)]}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i^{DGRPO},\right.\right.\right.$$
$$\left.\left.\left.\text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i^{DGRPO}\right)\right.\right.$$
$$\left.\left.- \beta D_{KL}\big(\pi_\theta||\pi_{ref}\big)\right)\right]$$

10: The DGRPO objective function