Scaffolded Group Relative Policy Optimization (Scaf-GRPO): A Framework for Robust LLM Reasoning Enhancement

AI Research Assistant

Abstract

Large Language Models (LLMs) have shown remarkable capabilities in complex reasoning tasks, often enhanced through Reinforcement Learning from Verifiable Rewards (RLVR). However, these methods are frequently hampered by the "learning cliff" phenomenon, where models fail on problems significantly beyond their current capabilities, leading to persistent zero-reward signals and stalled learning. This paper introduces Scaffolded Group Relative Policy Optimization (Scaf-GRPO), a novel progressive training framework designed to overcome this challenge. Inspired by pedagogical scaffolding, Scaf-GRPO strategically provides hierarchical, minimal in-prompt guidance only when a model's independent learning plateaus. The framework employs a two-phase approach: an initial guidance exemption period to diagnose "true-hard" problems, followed by hierarchical hint-guided exploration. By augmenting the training batch with minimally guided successful trajectories, Scaf-GRPO restores the learning gradient, enabling models to learn from previously intractable problems. Experimental results on challenging mathematics benchmarks demonstrate Scaf-GRPO's significant superiority over vanilla GRPO and existing prefix-based guidance methods, fostering genuine skill acquisition and improved generalization. This work provides a robust methodology for extending the frontier of autonomous reasoning in LLMs.

1 Introduction

The advent of Large Language Models (LLMs) has revolutionized capabilities across diverse domains, from natural language understanding to complex reasoning tasks in mathematics, programming, and logic [**?**]. A pivotal technique for enhancing these complex reasoning abilities is Reinforcement Learning from Verifiable Rewards (RLVR), where models learn by exploring various strategies and receiving feedback on their final outcomes, circumventing the need for expensive, step-by-step human annotations [**?**]. This paradigm allows models to autonomously discover effective problem-solving procedures.

Despite its promise, RLVR is fundamentally constrained by a phenomenon we term the "learning cliff." This occurs when an LLM encounters problems

that are far beyond its current capabilities. In such scenarios, all exploratory attempts consistently fail, leading to a persistent zero-reward signal. For policy optimization algorithms like Group Relative Policy Optimization (GRPO) [**?**], this collapses the advantage calculation to zero, effectively rendering these difficult problems "invisible" to the learning gradient and stalling progress. Consequently, these problems form a persistent "long tail" of challenges that the model cannot conquer autonomously, preventing it from leveraging the most difficult examples to achieve a higher level of competence.

To address this critical bottleneck, existing strategies often incorporate off-policy guidance from a more capable "teacher" policy, typically by providing a prefix of a correct "golden" solution [**?**]. While this ensures a positive reward signal, it introduces significant issues such as distributional mismatches between teacher-generated prefixes and student-generated suffixes, necessitating complex algorithmic corrections. More importantly, this "on-rails" guidance stifles the model's ability to explore alternative, potentially more novel or efficient, reasoning strategies.

This paper introduces **Scaf-GRPO (Scaffolded Group Relative Policy Optimization)**, a novel progressive training framework inspired by pedagogical scaffolding [**?**]. Scaf-GRPO provides hierarchical, minimal, and progressive assistance to help LLMs bridge their capability gaps without enforcing rigid solution prefixes. Our in-prompt scaffolding approach is guided by two primary objectives: first, to maintain policy consistency by having the model process both the problem and the hint under a single, unified policy, thereby avoiding the distributional mismatches of prefix-based methods. Second, to preserve exploration flexibility, as our hints act as "signposts" rather than "railroads," guiding the model without fixing its path and allowing it to discover its own unique solution strategies.

2 Background: GRPO and the Learning Cliff

Reinforcement Learning from Verifier Reward (RLVR) has become a cornerstone for enhancing LLM reasoning. In this paradigm, models generate solutions, and an external verifier provides an outcome-based reward (e.g., correct/incorrect). DeepSeek-R1 [**?**] demonstrated that even with sparse, binary rewards, models can learn complex reasoning strategies.

Group Relative Policy Optimization (GRPO) [**?**] is an on-policy RL algorithm used for training LLMs that eliminates the need for a trainable value function. For a given prompt q, the policy π_θ generates a group of N trajectories, $G = \{o_1, \ldots, o_N\}$. After obtaining a terminal reward $R(o_i)$ for each trajectory from a verifier, GRPO computes a normalized advantage $\hat{A}_i$ as:

$$\hat{A}_i = \frac{R(o_i) - \mu_G}{\sigma_G + \epsilon_{std}} \tag{1}$$

where μ_G and σ_G are the mean and standard deviation of rewards in the group G, and ϵ_{std} is a small constant for numerical stability. The policy is then updated

by maximizing a clipped surrogate objective:

$$J_{GRPO}(\theta) = \hat{\mathbb{E}}_{i,t}\left[\min\left(r_{i,t}(\theta)\hat{A}_i, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right)\right] \tag{2}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|o_{i,<t},q)}{\pi_{\theta_{old}}(o_{i,t}|o_{i,<t},q)}$ is the probability ratio between the current and old policies, and ϵ is the clipping hyperparameter.

The key limitation, the *learning cliff*, arises when all trajectories in G receive a zero reward. In this scenario, μ_G and σ_G become zero, causing $\hat{A}_i$ to collapse to zero for the entire group and stalling the learning process. This lack of a learning signal for difficult problems prevents models from improving on the most challenging examples.

3 The Scaf-GRPO Framework

Scaf-GRPO modifies the GRPO training process by strategically augmenting the trajectory group G when a learning cliff is detected. The framework operates in two carefully designed phases: an initial guidance exemption period and a subsequent cyclical phase of hierarchical hint-guided exploration.

3.1 Phase 1: Diagnosing True-Hard Problems (Guidance Exemption)

A crucial principle in effective teaching is to avoid providing unnecessary help. Not all initial failures indicate a fundamental capability gap; many are "pseudo-hard" samples stemming from issues like unfamiliarity with output formats or nascent reasoning skills. Scaf-GRPO incorporates a guidance exemption period, typically the initial 15% of training steps. During this phase, the model attempts solutions purely through on-policy exploration. This period allows the model to overcome basic execution issues independently. Guidance is only activated when the rate of solving zero-reward queries stagnates, identifying problems as "true-hard" and thus candidates for intervention.

3.2 Phase 2: Hierarchical Hint-Guided Exploration

Once a problem is identified as "true-hard," Scaf-GRPO activates its guidance mechanism using a pre-defined, three-tiered hint hierarchy, $H = \{H_{knowledge}, H_{planning}, H_{solution}\}$. These tiers offer distinct levels of guidance:

1. $H_{knowledge}$ **(Knowledge Hint):** Points to the key concept or formula required.

2. $H_{planning}$ **(Planning Hint):** Outlines a high-level strategic framework for the solution.

3. $H_{solution}$ **(Solution Hint):** Provides a concrete calculation step.

To provide the minimal necessary guidance, the framework executes a deterministic search through this hierarchy, proceeding from the most abstract to the most concrete hint ($H_{knowledge} \rightarrow H_{planning} \rightarrow H_{solution}$). Within each tier, guidance is offered incrementally. The search terminates as soon as the model generates a correct solution, thereby identifying the minimal effective guidance required. This approach encourages the internalization of reasoning skills rather than the memorization of solutions.

3.3 On-Policy Batch Augmentation and Unified Loss

The core of Scaf-GRPO is its on-policy intervention, designed to reactivate the learning signal during a learning cliff. When all initial trajectories $G = \{o_1, \ldots, o_N\}$ from $\pi_\theta(\cdot|q)$ yield zero reward, the advantage $\hat{A}_i$ collapses, halting the gradient update. Scaf-GRPO intervenes by finding a minimal hint h^* that enables policy π_θ to generate a successful trajectory $o_h^* \sim \pi_\theta(\cdot|q \oplus h^*)$, where $\oplus$ denotes the concatenation of the hint into the prompt. This successful trajectory replaces a random failed trajectory $o_j \in G$ to form an augmented group, $G_{final} = (G \setminus \{o_j\}) \cup \{o_h^*\}$.

Crucially, Scaf-GRPO does not alter the mathematical form of the GRPO loss function. Instead, it modifies the data used for the loss computation. The advantage calculation is performed on this conditionally augmented batch:

$$\hat{A}'_i = \frac{R(o'_i) - \mu_{G_{final}}}{\sigma_{G_{final}} + \epsilon_{std}} \quad \text{for } o'_i \in G_{final} \tag{3}$$

The learning objective remains the clipped surrogate objective, but it is now applied to the trajectories in G_{final}. The probability ratio for a given trajectory $o'_i \in G_{final}$ at timestep t is critically computed with respect to the trajectory's specific originating prompt:

$$r'_{i,t}(\theta) = \begin{cases} \frac{\pi_\theta(o'_{i,t}|o'_{i,<t},q)}{\pi_{\theta_{old}}(o'_{i,t}|o'_{i,<t},q)} & \text{if } o'_i \in G_{final} \text{ and } o'_i \neq o_h^* \\ \frac{\pi_\theta(o'_{i,t}|o'_{i,<t},q\oplus h^*)}{\pi_{\theta_{old}}(o'_{i,t}|o'_{i,<t},q\oplus h^*)} & \text{if } o'_i = o_h^* \end{cases} \tag{4}$$

This on-policy augmentation ensures the batch contains non-zero reward variance, restoring a meaningful advantage signal and allowing learning to resume on previously intractable problems. This preserves the on-policy principle, avoiding the high variance and instability associated with off-policy corrections.

4 Experimental Setup and Results

The effectiveness of Scaf-GRPO was demonstrated through extensive experiments on several challenging mathematics benchmarks, including AIME24, AIME25, AMC, Minerva, MATH-500, Olympiad, and GaoKao2023en. Experiments were conducted on diverse LLM architectures, such as the Qwen2.5 series, Llama-3.2-3B-Instruct, and the Long Chain-of-Thought model DeepSeek-R1-Distill-Qwen-1.5B.

4.1 Key Findings

- **Significant Performance Gains:** Scaf-GRPO consistently achieved substantial performance improvements across all tested models and benchmarks. For instance, on the AIME24 benchmark, the Qwen2.5-Math-7B model saw a relative `pass@1` score boost of 44.3% over the vanilla GRPO baseline.
- **Outperformance of Baselines:** Scaf-GRPO demonstrated clear superiority over vanilla GRPO and other leading methods, including prefix-based guidance approaches like LUFFY [**?**]. This highlights the efficacy of the in-prompt scaffolding strategy compared to altering trajectories.
- **Generalization Across Models and Tasks:** The framework proved to be model-agnostic, showing consistent gains on different architectures (Qwen, Llama) and specializations (math-tuned, instruction-tuned, Long-CoT models). Furthermore, Scaf-GRPO fostered robust reasoning skills that generalized to out-of-distribution tasks, such as the expert-level scientific questions in the GPQA-Diamond benchmark [**?**].
- **Ablation Studies Validation:** Ablation studies confirmed the critical role of each component: the guidance exemption period prevents over-reliance on hints, and the progressive, hierarchical nature of hints fosters more generalizable reasoning skills over direct solution provision.

5 Discussion and Future Work

Scaf-GRPO presents a significant step forward in enhancing LLM reasoning by effectively addressing the learning cliff. Its pedagogical approach, which provides minimal, hierarchical guidance only when needed, allows models to internalize reasoning skills rather than merely imitating solutions. This preserves the exploratory autonomy of LLMs and avoids the distributional consistency issues inherent in prefix-continuation methods.

5.1 Limitations

The practical deployment of Scaf-GRPO currently relies on the availability of a high-quality, tiered hint hierarchy. Generating these structured hints requires a non-trivial data preparation effort. Additionally, the framework is primarily designed for tasks with verifiable solutions and structured reasoning paths, such as mathematics. Its applicability to more open-ended, subjective domains like creative writing is less direct.

5.2 Future Work

Future research should focus on several promising directions:

1. **Automated Hint Generation:** Developing methods to automatically generate high-quality, tiered hint hierarchies would significantly enhance the framework's scalability and reduce manual data preparation. This could involve using more powerful LLMs to self-generate hints or leveraging symbolic reasoning systems.

2. **Adaptive Scaffolding Mechanisms:** Exploring adaptive guidance mechanisms where the level and type of assistance dynamically adjust to the model's improving proficiency. This personalized learning process could optimize the balance between guidance and autonomous exploration.

3. **Extending to Other Domains:** Investigating the applicability of Scaf-GRPO to other complex reasoning domains, such as scientific discovery, logical inference, or code generation, where structured problem-solving can benefit from targeted guidance.

4. **Theoretical Analysis of Convergence:** A deeper theoretical analysis of Scaf-GRPO's convergence properties and its robustness against various types of learning cliffs would provide further insights into its mechanisms.

6 Conclusion

In this work, we introduced Scaf-GRPO, a novel training framework that effectively overcomes the "learning cliff" phenomenon in reinforcement learning for large language models. By providing hierarchical, minimal in-prompt guidance, Scaf-GRPO enables models to solve problems previously beyond their reach. This on-policy guidance preserves exploratory autonomy and mitigates the distributional consistency issues inherent in prefix-continuation methods. Our extensive experiments demonstrate that Scaf-GRPO significantly outperforms vanilla GRPO and strong prefix-based baselines across challenging mathematics benchmarks, establishing a more effective path toward robust and autonomous reasoning in LLMs.

References

[1] Xichen Zhang, Sitong Wu, Yinghao Zhu, Haoru Tan, Shaozuo Yu, Ziyi He, Jiaya Jia. *Scaf-GRPO: Scaffolded Group Relative Policy Optimization for Enhancing LLM Reasoning*. arXiv preprint arXiv:2510.19807, 2025.

[2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv preprint arXiv:2501.12948, 2025.

[3] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. *DAPO: An*

open-source LLM reinforcement learning system at scale. arXiv preprint arXiv:2503.14476, 2025.

[4] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. *Learning to reason under off-policy guidance*. arXiv preprint arXiv:2504.14945, 2025.

[5] Laura E. Berk and Adam Winsler. *Scaffolding Childrens Learning: Vygotsky and Early Childhood Education*. National Association for the Education of Young Children, Washington, DC, 1995.

[6] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. *OlympiadBench: A challenging benchmark for promoting AGI with Olympiad-level bilingual multimodal scientific problems*. arXiv preprint arXiv:2402.14008, 2024.