Scaf-GRPO: Scaffolded Group Relative Policy Optimization for Enhancing LLM Reasoning

https://arxiv.org/abs/2510.19807

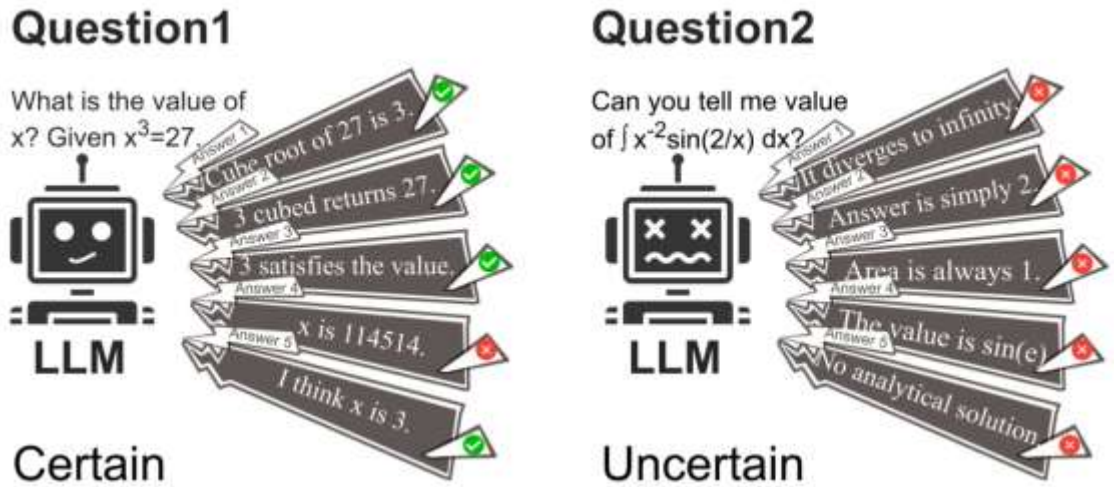
汇报人：陈俞亦　时间：2026-1-06

1. Background

随着强化训练的发展，现有的大模型在复杂推理任务上表现良好，对于复杂的数学问题，编程问题等都能够有一个正确的推理结果。强化学习的发展是其中重要的助推之一，这种训练方式依托于对模型回答的打分，而不是直接计算模型回答与标签之间的距离。这样的训练能够让模型真正学会如何去推理而不是对答案进行模仿。

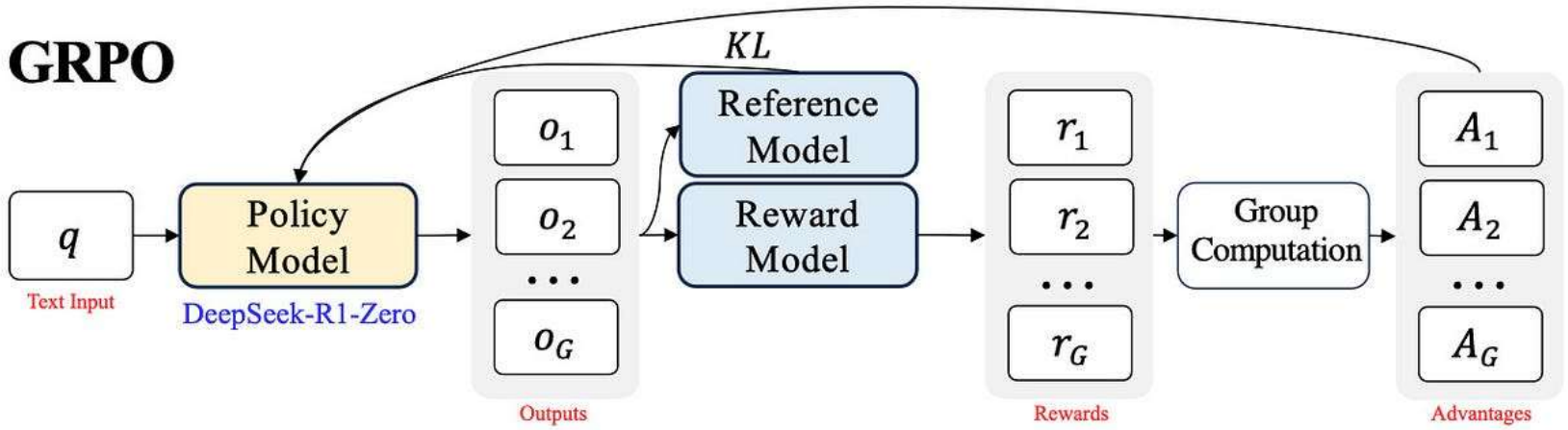
现有问题：Learning Cliff

1. 丢失奖励信号
2. 由于模型训练依托于奖励值的梯度，问题1会导致模型训练梯度消失



2. Introduction

Learning Cliff 现象源于强化学习的本质：基于奖励分数的梯度优化算法。**能否解决一个问题是取决于一个模型当前的能力的**，那么，如果一个模型在当前状态下能力仍不足以应对该训练集子集上,那么在该问题上模型就会获得0分。以GRPO举例：



当所有的 o_i 均为错误答案时，GRPO 在该轮次中的所有得分 r_i 将会是0分，这将导致其相对分数也为 0，那么在该轮次模型将无法产生有效迭代，模型训练将会出现“颠簸”

2. Related works

为了克服”learning cliff“带来的问题，常见的解决方案为：

“Teacher” Policy:

将一个“Off Policy model”生成的“Golden Solution”的前半部分输入当前模型，让模型去推理补充后半部分，这样生成的答案大概率是正确的，有得分的。

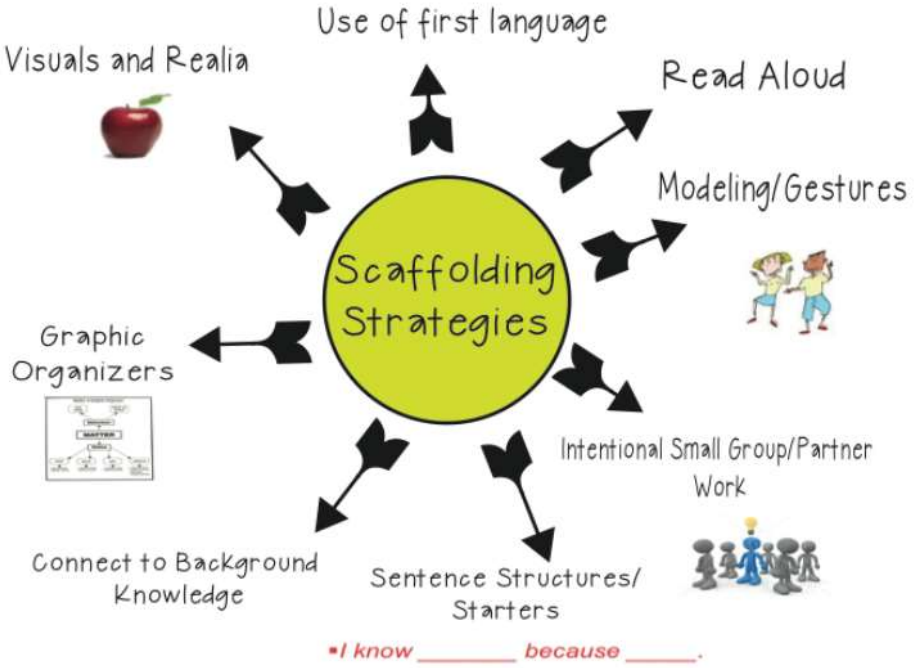
问题在于：

1. 两个模型的策略分布是不同的，这将导致模型分布不匹配从而使得模型训练不稳定，也会引入一些不必要的偏置。
2. 模型会倾向于沿着已给的路线走，失去自主探索能力，也失去了找到更有效率的路径能力。

3. Instructional scaffolding

在学习初期提供**“恰到好处的临时支持”**，随着能力提升**逐步撤除**，让学习者最终能够独立完成任务。

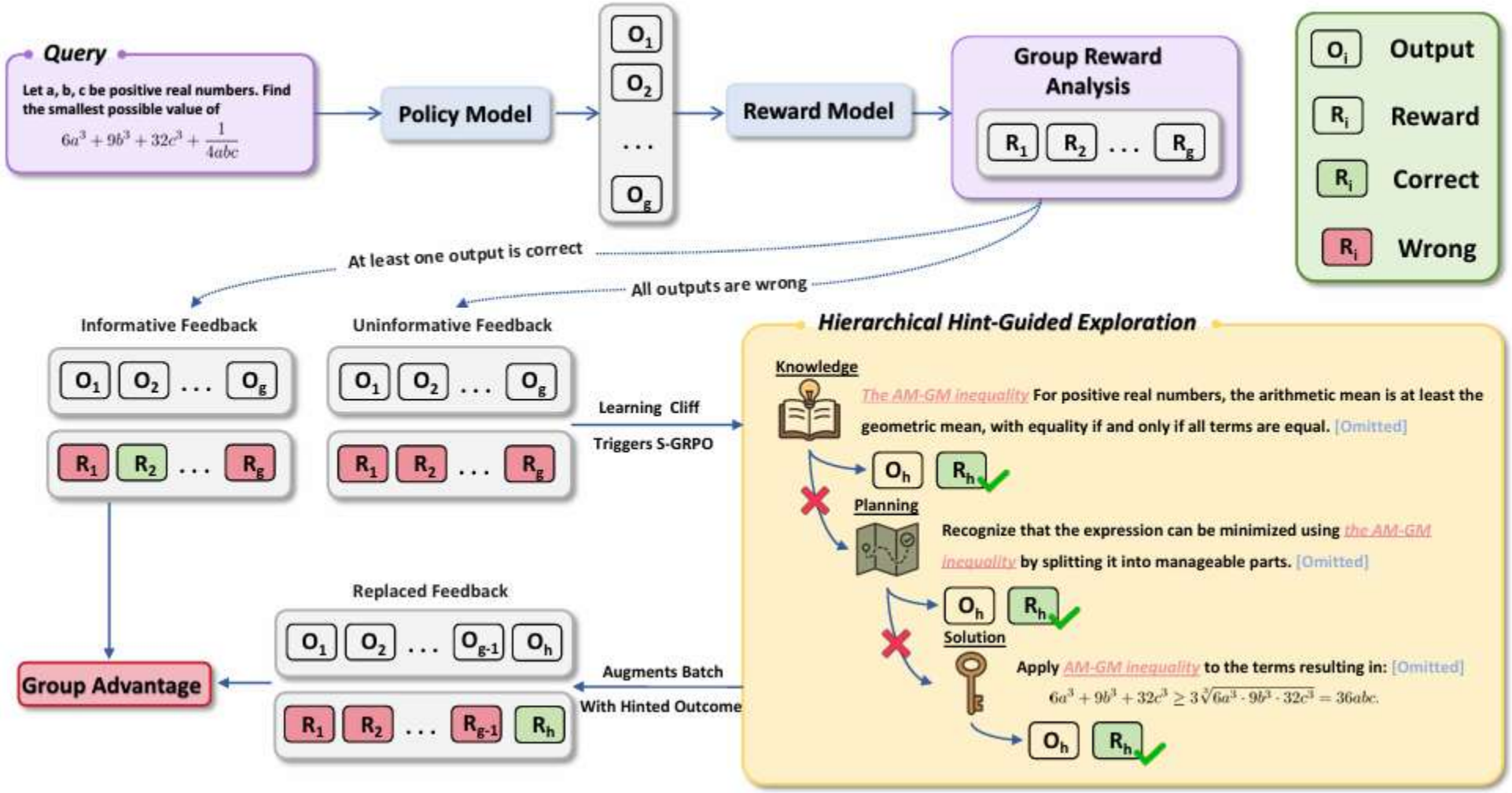
本文模拟了这一过程，在模型**完全无法解决任务**的情况下，对模型进行恰当的提示，让模型”刚好”能够完成任务，而不是直接给答案。



优势在于：

1. 本文所给的提示不是严格的推理步骤，可以最大程度保留模型的探索能力
2. 本文将Hint与问题一并输入模型中，避免了模型策略分布不匹配的问题。

4. Algorithm


Query
Let a, b, c be positive real numbers. Find the smallest possible value of
$6a^3 + 9b^3 + 32c^3 + \frac{1}{4abc}$
Policy Model
O_1 O_2 ... O_g
Reward Model
Group Reward Analysis
R_1 R_2 ... R_g
O_i Output
R_i Reward
R_i Correct
R_i Wrong
At least one output is correct
All outputs are wrong
Informative Feedback
Uninformative Feedback
Learning Cliff
Triggers S-GRPO
Hierarchical Hint-Guided Exploration
Knowledge
The AM-GM inequality For positive real numbers, the arithmetic mean is at least the geometric mean, with equality if and only if all terms are equal. [Omitted]
Planning
Recognize that the expression can be minimized using the AM-GM inequality by splitting it into manageable parts. [Omitted]
Solution
Apply AM-GM inequality to the terms resulting in: [Omitted]
$6a^3 + 9b^3 + 32c^3 \geq 3\sqrt[3]{6a^3 \cdot 9b^3 \cdot 32c^3} = 36abc.$
O_h R_h
Replaced Feedback
O_1 O_2 ... O_{g-1} O_h
R_1 R_2 ... R_{g-1} R_h
Augments Batch
With Hinted Outcome
Group Advantage

4. Structure of K-P-S

Prompt for hint injection

```
### System Prompt:
Please reason step by step, and put your final
answer within \boxed{}.

### User Prompt:
Question: {question}
Knowledge/Planning/Solution Hints: {hints}
```

其中每个阶梯都有四个分级，每次都按照 **Knowledge-Plan-Solution** 来为模型提供指导。此处指导是人工写的标准解法，然后使用 **DeepSeek - R1** 将其拆解为三个部分。

1. 使用原GRPO训练**前15%代**，降低模型因为格式错误等原因带来的0分，让模型拥有最基本的Agent能力（Two Phase）
2. **分辨真正难题和假难题**，假难题直接进行 **GRPO** 迭代，真难题按照 **K – P – S** 步骤进行提示。
3. **S-GRPO** 分段提供 **Hint** 直到找到正确答案，若是实在解不出来就放弃，正确答案随机替换原本的错误答案中的一个，保证一定有真解。

5. GRPO Score

$$\hat{A}'_i = \frac{R(o'_i) - \mu_{\mathcal{G}_{\text{final}}}}{\sigma_{\mathcal{G}_{\text{final}}} + \epsilon_{\text{std}}} \quad \text{for } o'_i \in \mathcal{G}_{\text{final}}.$$

$$J_{\text{Scaf-GRPO}}(\theta) = \hat{\mathbb{E}}_{i,t}\left[\min\left(r'_{i,t}(\theta)\hat{A}'_i, \text{clip}(r'_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}'_i\right)\right]$$

$$r'_{i,t}(\theta) = \begin{cases} \dfrac{\pi_\theta(o'_{i,t}|o'_{i,<t}, q)}{\pi_{\theta_{\text{old}}}(o'_{i,t}|o'_{i,<t}, q)} & \text{if } o'_i \in \mathcal{G}_{\text{final}} \text{ and } o'_i \neq o^*_h \\ \dfrac{\pi_\theta(o'_{i,t}|o'_{i,<t}, q \oplus h^*)}{\pi_{\theta_{\text{old}}}(o'_{i,t}|o'_{i,<t}, q \oplus h^*)} & \text{if } o'_i = o^*_h. \end{cases}$$

基本的GRPO计算与原GRPO一致，当启用了Scaffold 机制时，则原策略分布和现策略分布都要是启用了该机制的分布。

6. Experiment Setting

项目	内容
训练数据集	DeepScaleR-Preview-Dataset（动态难度筛选）
数据筛选策略	Too Easy（丢弃）；Too Hard（保留）；Potentially Solvable（50% 采样）
提示（Hint）生成	DeepSeek-R1 基于 ground-truth solution steps 生成三层提示
实验模型	Qwen2.5-Math-7B /1.5B；Qwen2.5-7B；Llama-3.2-3B-Instruct；DeepSeek-R1-Distill-Qwen-1.5B
对比方法	Vanilla GRPO；Simple-RL；Oat-Zero；LUFFY
评测基准	GaoKao2023en；AIME24；AIME25；AMC；MATH-500；OlympiadBench
OOD 评测	GPQA-Diamond
评测指标	Pass@1 （greedy decoding）
训练设置	10 epochs；best checkpoint；KL = 0
最大输出长度	2048 tokens（Long-CoT：8192）

6. 实验结果

1. **相较于 vanilla GRPO**：Scaf-GRPO 在所有测试模型上取得显著性能提升，有效帮助模型跨越“学习断崖”。
2. **相较于其他方法**：Scaf-GRPO 明显优于 Simple-RL、Oat-Zero 及前缀续写类方法，体现了 in-prompt scaffolding 的优势。
3. **跨架构泛化性**：Scaf-GRPO 在非 Qwen 架构（如 Llama）上依然保持稳定提升，证明其模型无关性。
4. **对 LongCoT 的适用性**：Scaf-GRPO 能进一步增强长链式推理模型的表现，适用于复杂、长推理场景。

Model	AIME 24	AIME 25	AMC	Minerva	MATH-500	Olympiad	Gaokao2023en	Avg.
Qwen2.5-Math-1.5B								
Qwen2.5-Math-1.5B	7.2	3.3	32.5	14.7	32.8	20.6	20.0	18.7
Vanilla GRPO	13.3	10.0	47.5	28.3	72.2	34.8	57.4	37.6
Scaf-GRPO	**20.0**	**13.3**	**60.0**	**29.1**	**73.4**	**36.6**	**57.9**	**41.5**
Qwen2.5-Math-7B								
Qwen2.5-Math-7B	13.3	13.3	42.5	16.5	53.6	18.2	35.1	27.5
Vanilla GRPO	30.0	13.3	60.0	33.4	75.8	41.3	62.6	45.2
SimpleRL-Zero [7]	23.3	13.3	55.0	31.6	76.8	37.2	60.8	42.6
Oat-Zero [20]	30.0	16.7	62.5	34.6	78.0	41.0	62.9	46.5
LUFFY [2]	33.3	16.7	62.5	33.8	75.2	41.7	62.7	46.6
Scaf-GRPO	**43.3**	**20.0**	**70.0**	**36.4**	**80.0**	**43.3**	**63.4**	**50.9**
Qwen2.5-7B								
Qwen2.5-7B	10.0	6.7	37.5	26.4	61.8	34.4	42.6	31.3
Vanilla GRPO	10.0	10.0	50.0	38.5	77.6	40.4	**64.2**	41.5
Scaf-GRPO	**13.3**	**20.0**	**60.0**	**38.6**	**77.8**	**40.8**	63.8	**44.9**
Llama-3.2-3B-Instruct								
Llama-3.2-3B-Instruct	6.7	0.0	20.0	11.8	38.3	12.6	33.5	17.6
Vanilla GRPO	13.3	0.0	35.0	18.7	51.8	18.3	45.7	26.1
Scaf-GRPO	**16.7**	**3.3**	**40.0**	**19.1**	**56.2**	**20.3**	**46.0**	**28.8**
DeepSeek-R1-Distill-Qwen-1.5B								
DeepSeek-R1-Distill-Qwen-1.5B	28.9	20.0	67.5	26.1	83.9	45.8	62.1	47.7
Vanilla GRPO	30.0	21.1	67.5	30.1	83.9	50.2	71.4	50.6
Scaf-GRPO	**33.3**	**23.3**	**77.5**	**32.4**	**85.8**	**50.7**	**72.3**	**53.6**

6. 消融实验：预训练部分对模型效果的影响

预训练部分可以有效帮助模型分辨**什么是伪难题，什么是真难题**。对于伪难题，很多时候，模型的错误并不是模型本身能力的问题，大多是**模型回答格式的错误**。预训练部分可以让模型自己克服这些小错误，让真难题更加突出。这个实验验证了预训练部分对模型的影响。

Model	AIME 24	AIME 25	AMC	Minerva	MATH-500	Olympiad	Gaokao2023en	Avg.
Qwen2.5-Math-1.5B								
Vanilla GRPO	13.3	10.0	47.5	28.3	72.2	34.8	57.4	37.6
Scaf-GRPO (w/o Phase 1)	10.0	10.0	57.5	27.5	71.4	36.3	57.9	38.7
Scaf-GRPO	**20.0**	**13.3**	**60.0**	**29.1**	**73.4**	**36.6**	**57.9**	**41.5**
Qwen2.5-Math-7B								
Vanilla GRPO	30.0	13.3	60.0	33.4	75.8	41.3	62.6	45.2
Scaf-GRPO (w/o Phase 1)	23.3	13.3	70.0	34.2	78.4	41.2	63.1	46.2
Scaf-GRPO	**43.3**	**20.0**	**70.0**	**36.4**	**80.0**	**43.3**	**63.4**	**50.9**

6. 消融实验：Scaffold 式的训练对模型更好

本文在面对真难题时，使用了渐进式的Hint帮助模型生成答案。为了验证模型是否真的需要渐进式的提示，本文分别在渐进式和直接提供答案两种方式下训练。

实验结果显示，在仅提供Solution时其性能相比完整模型下降了 4.9%。这一结果验证了假设：**迫使模型先进行高层次推理，有助于培养更具泛化能力的推理技能。**

Hint Strategy	AIME24	AIME25	AMC23	Minerva	MATH-500	Olympiad	Gaokao2023en	Avg.
Scaf-GRPO (Full K →P →S)	**43.3**	**20.0**	**70.0**	**36.4**	**80.0**	43.3	63.4	**50.9**
w/o Progressive (Solution-Only)	40.0	13.3	65.0	36.2	78.6	**43.7**	62.3	48.4
w/o Knowledge Hint (P →S)	43.3	13.3	70.0	34.2	77.8	42.4	63.1	49.2
w/o Planning Hint (K →S)	43.3	16.7	62.5	35.0	79.4	40.0	63.6	48.6
w/o Solution Hint (K →P)	40.0	10.0	67.5	34.2	78.6	42.2	63.4	48.0
w/o Incremental Chunking	43.3	10.0	62.5	36.0	76.0	41.6	**64.2**	47.7
No Guidance (Vanilla GRPO)	30.0	13.3	60.0	33.4	75.8	41.3	62.6	45.2

6. 消融实验：K – P – S 结构的完整性

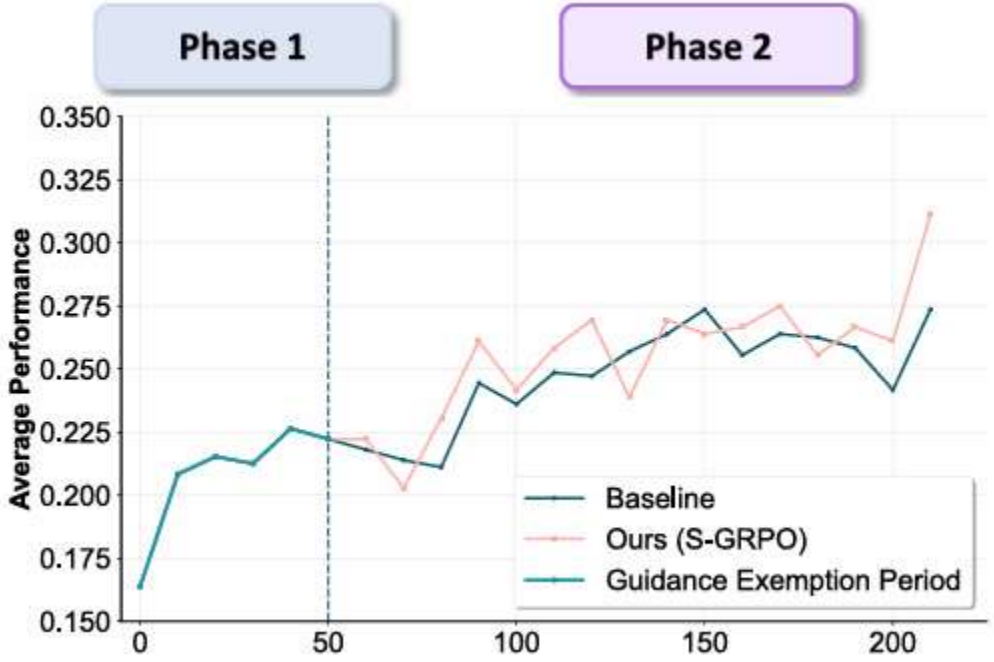
这里本文测试了K-P-S结构的完整性对模型训练的影响，这里采用消融实验的方法，逐层移除其中的一层。无论移除哪一层，模型性能都会下降。其中性能下降最为显著的是移除最终“Solution”提示的 K→P 变体，性能下降了 5.7%。

这一结果表明，该分层结构具有双重作用：**抽象层提示促进高层次推理，而具体层提示则在关键时刻为模型提供必要的兜底支持。**完整的 K→P→S 模型表现最佳，说明各层之间是互补而非冗余的。

Hint Strategy	AIME24	AIME25	AMC23	Minerva	MATH-500	Olympiad	Gaokao2023en	Avg.
Scaf-GRPO (Full K →P →S)	**43.3**	**20.0**	**70.0**	**36.4**	**80.0**	43.3	63.4	**50.9**
w/o Progressive (Solution-Only)	40.0	13.3	65.0	36.2	78.6	**43.7**	62.3	48.4
w/o Knowledge Hint (P →S)	43.3	13.3	70.0	34.2	77.8	42.4	63.1	49.2
w/o Planning Hint (K →S)	43.3	16.7	62.5	35.0	79.4	40.0	63.6	48.6
w/o Solution Hint (K →P)	40.0	10.0	67.5	34.2	78.6	42.2	63.4	48.0
w/o Incremental Chunking	43.3	10.0	62.5	36.0	76.0	41.6	**64.2**	47.7
No Guidance (Vanilla GRPO)	30.0	13.3	60.0	33.4	75.8	41.3	62.6	45.2

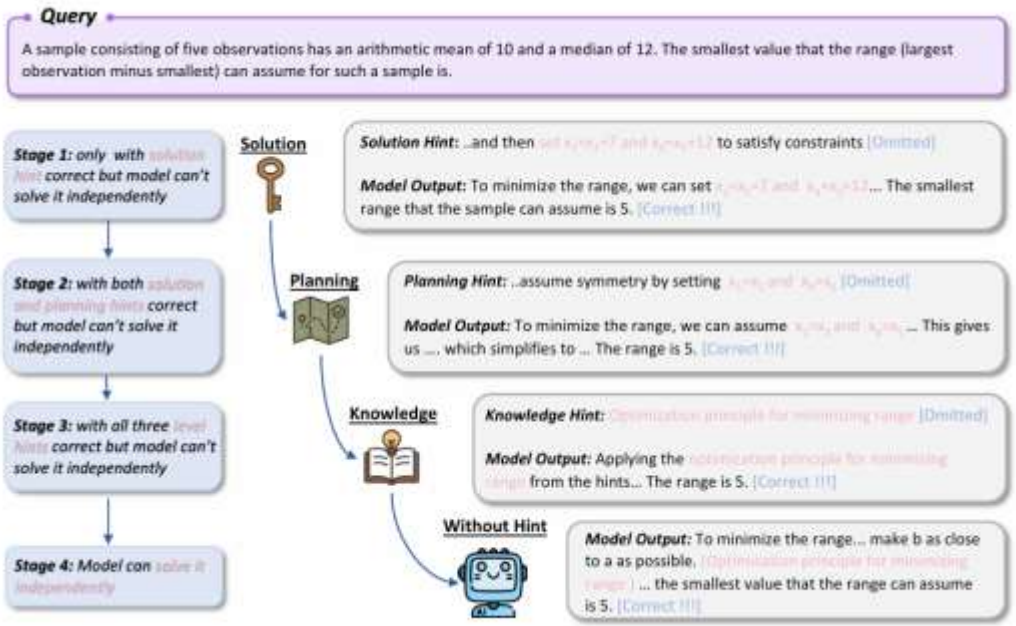
7. Scaffold 有效性分析

vanilla GRPO 的曲线很快趋于平缓，形成所谓的“学习断崖”——即模型在一部分持续存在的“真正困难问题”上无法再获取有效的学习信号。相比之下，Scaf-GRPO 的 scaffold 机制会被激活，使模型能够持续从这些问题中学习，并不断减少零奖励问题的数量。通过将原本不可学习的问题转化为学习机会，Scaf-GRPO 的验证性能持续提升，而基线方法则停滞不前。



7. 模型学习能力分析

Scaf-GRPO 的成功在于**促进能力习得，而不仅仅是模仿答案**。以一个真正困难的问题为例，真正学习的关键证据出现在模型后续再次遇到该问题时——在**完全没有任何提示的情况下**，模型仍能通过整合先前受引导训练中学到的技能成功解题。这表明，提示的作用在于**构建可迁移、持久的推理能力**，而非仅带来一次性的成功，而正是这种在高难问题上的能力积累，使模型得以跨越“学习断崖”。



7. 总结

Scaf-GRPO 为模型提供了支持，使其能够解决原本难以独立完成的问题。这种**策略内引导（on-policy guidance）**既保留了模型的探索自主性，又缓解了前缀续写方法固有的分布一致性问题。该框架使模型能够从先前难以解决的问题中学习，为实现自主推理提供了更有效的路径。

Scaf-GRPO 存在两个主要问题

- 其效果依赖于**高质量的分层提示体系**，而生成这些结构化提示需要相当复杂的数据准备工作。
- 该框架主要适用于**具有可验证解答和结构化推理路径的任务**（如数学），在更开放、主观的领域（如创意写作）中的适用性没有得到验证

Thanks for Your Watching !

汇报人：陈俞亦　时间：2026-1-6