arXiv:2602.03143v1 [cs.LG] 3 Feb 2026

Self-提示语言模型增强强化学习

Baohao Liao [*†1,2] **Hanze Dong** [*1] **Xinxing Xu** [1] **Christof Monz** [2] **Jiang Bian** [1]

摘要

组相对策略优化（GRPO）最近已成为一种实用方案，可用于使大语言模型与可验证目标保持一致。然而，在稀疏终端奖励下，GRPO 往往会陷入停滞，因为同一组中的推演经常获得相同奖励，导致相对优势值坍缩，更新消失。我们提出带特权监督的自提示对齐 *GRPO* (SAGE)，这是一种同策略强化学习框架，在训练期间注入特权提示，以在相同的终端验证器奖励下重塑展开分布。对于每个提示 x，模型会采样一个简洁的提示 h（例如，计划或分解），然后生成一个 τ 以 (x, h)为条件的解答。关键在于，任务奖励 $R(x, \tau)$ 保持不变；提示只会在有限采样下提升组内结果的多样性，从而防止 GRPO 优势值在稀疏奖励下坍缩。在测试时，我们设置 $h = \varnothing$，并部署无提示策略，不使用任何特权信息。此外，采样多样化的自提示可作为一种自适应课程，比来自初始策略或更强外部模型的固定提示更有效地追踪学习者的瓶颈。在 6 个基准测试和 3 个大语言模型上的实验表明，SAGE 的表现始终优于 GRPO，平均而言，在 Llama-3.2-3B-Instruct 上提升 +2.0，在 Qwen2.5-7B-Instruct 上提升 +1.2，在 Qwen3-4B-Instruct 上提升 +1.3。代码见https://github.com/BaohaoLiao/SAGE。

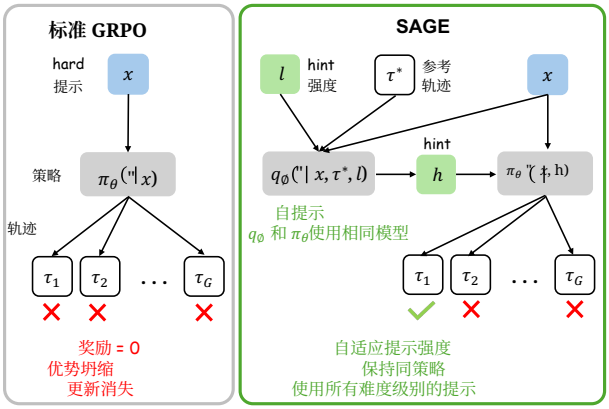

图 *1.* 我们提出的方法 SAGE 概览。当大语言模型无法为某个提示采样到任何正确轨迹时，大语言模型会从该提示的参考解答中**自行生成**提示。随后，将该提示与困难提示一同作为大语言模型的输入，从而避免优势坍缩，并确保采样到正确轨迹以更新策略模型。

1. 引言

强化学习（RL）已成为训练和对齐大语言模型的核心工具，尤其是在监督信号最自然地通过可验证目标来表示时，例如精确匹配正确性、单元测试或自动检查器（Ouyang等人，2022年；Schulman等人，2017；DeepSeek-AI等人，2025年）。在这种情形下，目标很直接：最大化提示上的期望奖励。然而，优化过程可能很脆弱：在有限采样下，策略梯度估计器可能表现出高方差，甚至在困难提示上退化。

一个显著的例子是稀疏终端奖励下的组相对策略优化（GRPO）（Shao 等人，2024年）。GRPO 会对每个 rollout 组内的奖励进行中心化（通常还会标准化），依赖组内结果差异产生非零更新。对于带有 $0/1$ 验证器的情况，困难提示经常会产生这样的组：其中所有推演获得相同奖励（通常全为零）。此时，组内中心化后的优势值会坍缩，小批量策略梯度估计则恒为零。重要的是，这是一种**有限样本病态现象**：底层期望目标未必是平坦的，但对于许多提示，该估计量无法提供学习信号。

现有解决方案大多着眼于改进数据收集。一种常见的基线做法是跳过信息量不足的更新（例如，退化组），并重新采样提示。这虽然能提升性能，却会隐式地使训练偏向更容易的提示（Yu 等人，2025年；Xiong 等人，2025a）。更系统的

[*]共同贡献 [†]本工作是在微软实习期间完成的。 [1]微软研究院 [2]阿姆斯特丹大学语言技术实验室。联系邮箱：Hanze Dong <hanze-dong@microsoft.com>。

预印本。*2026*年*2*月*4*日。

方法包括采用自适应采样或课程式调度，为难度较高的提示分配更多推演（Yao 等人，2025年；Xiong 等人，2025年；Li等人，2025年；Zhang等人，2025c），还包括利用离线数据或外部生成的候选内容（例如，来自更强模型的内容）来引导学习（Zhang等人，2025a；Yan等人，2025年；Zhang等人，2025年）。尽管这些策略有效，但它们可能会扭曲训练分布，或引入必须谨慎处理的上下文／分布失配。

我们提出 SAGE（采用特权监督的自提示对齐 GRPO），这是一种基于特权提示的互补方法。训练期间，我们提供额外提示 h，即参考解答 $\tau^\star$ 的有损压缩，并从提示条件化策略 $\pi_\theta(\cdot \mid x,\ h)$ 进行 rollout。提示只会重塑 rollout 分布，以提高在有限组中观察到混合结果的概率。测试时，我们部署无提示策略。我们将策略自身生成的提示称为自提示，将生成此类提示的过程称为自提示。

这种退化现象可以明确地表示出来。设 $p_\theta(x)$ 为无提示成功概率，G 为组大小。一个 rollout 组包含混合结果的概率为 $1-(1-p_\theta(x))^G - p_\theta(x)^G \approx G p_\theta(x)$，因此，只要 $G p_\theta(x) \ll 1$，更新就会消失。只有当提示提高有效成功概率，使得在相同的 G下混合结果组变得常见时，提示才有用。

贡献。 (i) 如图 1 所示，我们提出了 SAGE，这是一种同策略强化学习框架，在训练期间利用特权提示对推演进行条件化，同时保持任务奖励不变，并在测试时移除提示。(ii) 我们开发了一种依赖策略的提示强度调度器，仅在组内奖励崩溃时激活提示，从而实现自动课程。(iii) 我们提出了一种在线自提示方案，在训练期间定期刷新提示分布，使其与学习者保持校准，避免固定提示过弱或过强。(iv) 我们通过分析将 GRPO 崩溃刻画为伯努利奖励下的门开启概率，并通过实验证明，SAGE 能够提升具有挑战性的推理基准上的样本效率和最终准确率。

2. 特权提示强化学习

当提示偏短时，原始 GRPO 在困难情形下，组往往获得相同的奖励，导致组内优势坍缩并使学习停滞。我们通过在训练期间注入特权提示 来应对这一失效模式：在保持奖励不变的同时重塑展开分布，以便在有限采样下呈现信息丰富的轨迹。

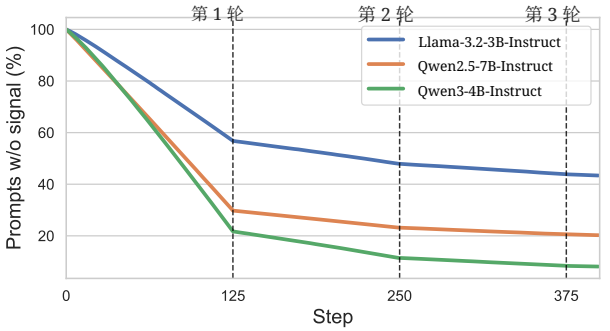

图2。 截至训练步骤，从未采样到正确轨迹的提示所占的百分比。这里我们使用 64k 个提示进行训练，并在每个步骤中为每个提示采样 8 条轨迹。大量提示在强化学习过程中被浪费了，尤其是对于较弱的大语言模型而言，因为它们无法为训练提供任何信号。

2.1. 设置与 GRPO 停滞

设 $x \sim \mathcal{D}$ 为一个提示。策略 π_θ 生成轨迹$\tau = (y_1, \ldots\ldots, y_T)$，记为 $\tau \sim \pi_\theta(\cdot \mid x)$。我们使用二元奖励 $R(x,\ \tau) \in \{0,1\}$. 定义成功概率

$$p_\theta(x) = \Pr_{\tau \sim \pi_\theta(\cdot|x)}[R(x,\tau)=1]. \tag{1}$$

GRPO 实现通常使用组内标准差对组内优势进行标准化。对于一组 Grollout $\{\tau_i\}_{i=1}^G$，令

$$R_i = R(x,\tau_i),\ \bar{R} = \tfrac{1}{G}\sum_{i=1}^{G} R_i,\ s^2 = \tfrac{1}{G}\sum_{i=1}^{G}(R_i-\bar{R})^2,$$

并定义标准化优势值

$$A_i = \frac{R_i - \bar{R}}{s+\epsilon}, \tag{2}$$

其中 $\epsilon \geq 0$ 是数值稳定项，且 $R_i \in \{0,1\}$。当 $p_\theta(x)$极小时，一个组往往全为零，并且对所有 i都有 $A_i = 0$。非退化组出现的概率为

$$1-(1-p_\theta(x))^G - p_\theta(x)^G \approx 1-(1-p_\theta(x))^G \approx G p_\theta(x),$$

因此，每当 $G p_\theta(x) \ll 1$时，训练就会在大多数提示上停滞。图 2 在实践中说明了这一现象：对于许多困难提示，在很长一段训练过程中从未采样到正确轨迹，因而无法产生学习信号。

2.2. 特权提示作为采样

当训练期间可获得参考轨迹 $\tau^\star$ 时，我们将提示 h 生成为 $\tau^\star$的有损压缩。该提示作为附加上下文追加到提示中。

提示强度和无提示情形。我们通过离散强度等级 $\ell \in \{0,1,\ldots,L\}$ 控制提示信息量，

其中，较大的 ℓ 表示包含更多关于参考轨迹 $\tau^\star$ 的信息。我们采样

$$\ell \sim p(\ell), \qquad h \sim q(h \mid x, \tau^\star, \ell), \tag{3}$$

其中，$\ell = 0$ 对应无提示设定，而 $q(h \mid x, \tau^\star, 0) = \delta_\varnothing(h)$（即，$h = \varnothing$ 确定性地）。

依赖策略的调度方式 ℓ。只有当提示未能提供学习信号时，才应使用提示。我们通过一个简单的统计量（例如坍塌指示器 $c(x) = \mathbb{I}\big[\mathrm{Var}\big(\{R(x, \tau_i)\}_{i=1}^G\big) = 0\big]$，该统计量由策略模型 π_θ下的一个小型探测组计算得出）来决定提示 ℓ 的采样是否依赖策略。一个最简调度器是

$$p(\ell \mid x) = \begin{cases} \delta_0, & c(x) = 0, \\ p(\ell), & c(x) = 1, \end{cases} \tag{4}$$

因此 $\ell > 0$ 仅在组坍缩时才会被激活。

有提示时，我们从 $\pi_\theta(\cdot \mid x, h)$ 中采样。成功率会提高：$p_\theta^{(\ell)}(x) = \Pr_{\tau \sim \pi_\theta(\cdot \mid x, h)}[R(x, \tau) = 1]$，$h \sim q(\cdot \mid x, \tau^\star, \ell)$。如果提示能充分提高 $p_\theta^{(\ell)}(x)$，使得 $G p_\theta^{(\ell)}(x)$ 不再微乎其微，那么提示就有用；这样，非退化组便会更常见，强化学习也能获得更新。

2.3. 使用提示的 GRPO 与最终损失

给定 x，采样 ℓ 和 h，然后抽取一组推演

$$\tau_i \sim \pi_\theta(\cdot \mid x, h), \qquad R_i = R(x, \tau_i), \quad i = 1, \ldots, G.$$

使用式（A_i 2）计算。以提示为条件的损失为

$$\mathcal{L}(\theta) = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} A_i \sum_{t=1}^{T_i} \log \pi_\theta(y_{i,t} \mid x, h, y_{i,<t})\right] + \beta\, \mathbb{E}\big[\mathrm{KL}\big(\pi_\theta(\cdot \mid x, h) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x, h)\big)\big], \tag{5}$$

其中，期望是针对 $x \sim \mathcal{D}$、根据式（3）进行的提示采样，以及来自 $\pi_\theta(\cdot \mid x, h)$ 的推演。在测试时，我们设定$\ell = 0$，$h = \varnothing$ 并运行策略 $\pi_\theta(\cdot \mid x, \varnothing) \equiv \pi_\theta(\cdot \mid x)$。

总结。稀疏奖励可能导致 GRPO 停滞，因为对于许多提示 x，有限组中没有正样本，优势值也会随之归零。特权提示通过改变这些提示对应的展开分布，同时保持奖励不变，从而解决这一问题。依赖策略的调度器仅在组内优势值归零时启用提示，由此形成自动课程。训练仍然是同策略的，因为推演取自 $\pi_\theta(\cdot \mid x, h)$。部署时使用$\ell = 0$，且不需要提示或特权信息。

3. 分析

3.1. 标准化 GRPO 作为门控更新目标

固定一个上下文（x，h），并进行 G 次 rollout，获得奖励 $R_i \in \{0,1\}$。令 $\bar{R} = \frac{1}{G}\sum_i R_i$ 且 $s^2 = \frac{1}{G}\sum_i (R_i - \bar{R})^2$。标准化 GRPO 使用 $\overline{A_i} = \frac{R_i - \bar{R}}{s + \epsilon}$.

推论 3.1（信号能量等于门概率）。定义优势能量 $E := \frac{1}{G}\sum_{i=1}^G A_i^2$. 若 $\epsilon > 0$，

$$E = \frac{s^2}{(s+\epsilon)^2} \in [0, 1], \tag{6}$$

它关于 s 单调递增，并且在 $s = 0$时仍会收敛到 0 。

对于 GRPO，提示级更新幅度主要取决于该组是否非退化 ($s > 0$)。换言之，训练的行为类似于一种门控过程：只有当 rollout 组包含混合结果时才会进行更新。

命题 3.2（伯努利奖励下的门控开启概率）。设 $p_\theta(x, h) = \Pr[R(x, \tau) = 1 \mid x, h]$。则

$$\Pr[s > 0 \mid x, h] = 1 - (1 - p_\theta(x, h))^G - p_\theta(x, h)^G, \tag{7}$$

$\Pr[s > 0 \mid x, h]$ 在 $p_\theta = \frac{1}{2}$ 处达到最大值。在稀疏区间 $p_\theta(x, h) \ll 1$ 中，$\Pr[s > 0 \mid x, h] \approx G\, p_\theta(x, h)$。

因此，SAGE 应选择提示强度，使困难提示词脱离 $G p_\theta \ll 1$ 所处的区间，并避免提示过强，以免将 $p_\theta \approx 1$ 推回到重新关闭门控的状态。

命题 3.3（最优提示分布取决于策略）。固定一个提示 x 和组大小 $G \geq 2$。令$p_\theta(x, h) = \Pr_{\tau \sim \pi_\theta(\cdot \mid x, h)}[R(x, \tau) = 1]$ 并定义

$$u(p) := 1 - (1 - p)^G - p^G. \tag{8}$$

在伯努利奖励下，$u(p_\theta(x, h)) = \Pr[s > 0 \mid x, h]$。

对于任意分布 $q(\cdot \mid x)$，定义期望概率

$$J_x(\theta, q) := \mathbb{E}_{h \sim q(\cdot \mid x)}[u(p_\theta(x, h))]. \tag{9}$$

那么，u 在 $[0, 1]$，上是对称且严格凹的，并在 $p = \frac{1}{2}$ 处取得最大值。因此，任意最大化点

$$q_\theta^\star(\cdot \mid x) \in \arg\max_q J_x(\theta, q) \tag{10}$$

必须将其概率质量置于校准提示 上，使得 $p_\theta \approx \frac{1}{2}$。–

一般而言，校准提示集合取决于 θ。除非$p_\theta(x, h)$ 在 θ 中对几乎所有 q-意义下的 h均保持不变，否则固定的 q 无法在整个训练过程中始终对 $J_x(\theta, q)$ 保持（近似）最优。因此，在线更新 q 可减少这种门控失配，并提高非退化 *GRPO* 更新的频率。

注 3.4（为何不对每个提示采样多个提示。）。令$u(p) = 1 - (1-p)^G - p^G$ 表示式（7）中的门概率。对于 $G \geq 2$，u 在 $[0, 1]$ 上是凹函数，因为 $u''(p) = -G(G-1)\big(p^{G-2} + (1-p)^{G-2}\big) \leq 0$. 因此，对于任意提示分布 $h \sim q$，

$$\mathbb{E}_{h \sim q}\big[u(p_\theta(x, h))\big] \leq u(\mathbb{E}_{h \sim q}[p_\theta(x, h)]). \tag{11}$$

在平均成功率固定时，提示之间的额外随机性只会降低非退化组的预期出现频率。受（11）启发，我们每轮对每个提示只采样一次提示实现，并将计算资源用于 G或更佳的强度调度。

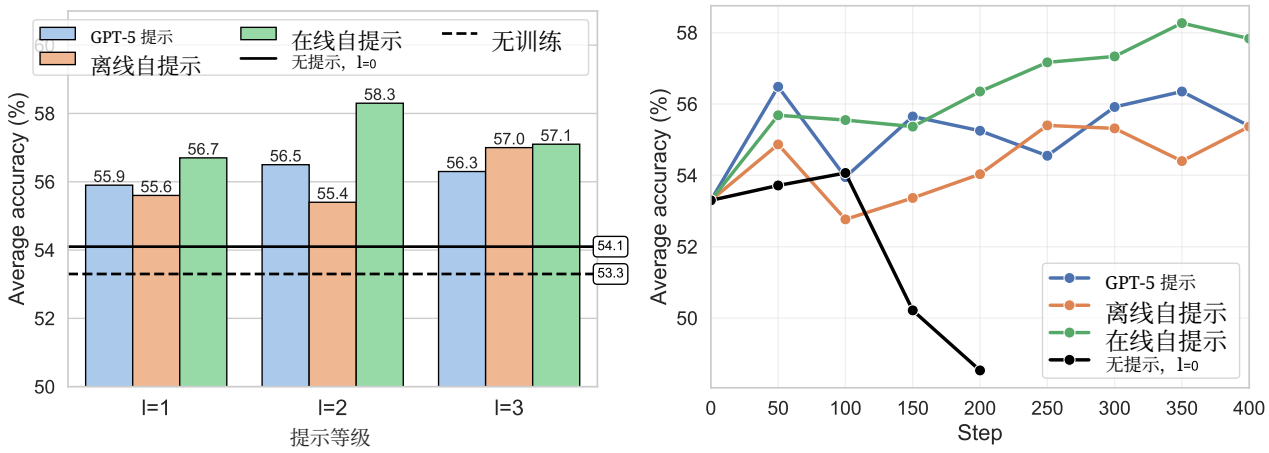

图3. Qwen3-4B-Instruct 在 6 个基准测试上的平均准确率。此处的 4.5k 个训练提示极其困难，其正确轨迹在训练过程中从未被采样过，如图 2所示。此处每个提示在每个步骤的 rollout 数量设为 32，以促进探索。 **左侧**：不同提示下的表现。无提示训练仅略微提升表现，因为困难提示的奖励信号十分稀疏。然而，使用任何提示进行训练都能提升表现。在所有方法中，在线自提示在不同提示等级下始终表现最佳。 **右侧**：提示等级为 $l = 2$时，平均准确率随训练步数的变化情况。随着训练推进，无提示训练甚至会导致表现下降，因为奖励信号过于稀疏，使模型过拟合于少数可解提示。然而，在线自提示能稳步提升表现。详细数据请参见表 C.1 。

总的来说，标准化 GRPO 将稀疏奖励学习转化为最大化门概率。所提出的方法应通过以下方式将这一点付诸实践：(i) 确保基于 h 的同策略条件化；(ii) 在门关闭时调度 ℓ；(iii) 在线更新自提示生成器，以保持 p_θ 校准。

4. SAGE 的设计

4.1. 同策略训练

为什么提示必须放在条件上下文中。SAGE 将提示 h附加到提示中，并从提示条件化策略 $\pi_\theta(\cdot \mid x, h)$ 中采样推演。这不仅仅是一种建模选择：这是在增广上下文上保持训练遵循同策略的关键。损失为 $\sum_t \log \pi_\theta(y_t \mid x, h, y_{<t})$。如果改为采样 $\tau \sim \pi_\theta(\cdot \mid x, h)$，却计算 $\log \pi_\theta(\tau \mid x)$（即在对数概率中删除 h），那么该更新便不再对应于采样过程中任何同策略目标的梯度。实践中，这种不匹配的表现类似于离策略更新，并且在稀疏奖励下稳定性明显更差。我们还进行了一项受控消融实验（第 5.2 节）：保持采样过程完全一致（使用提示进行推演），但改变条件设置。

4.2. 在线自提示

在该算法中，我们可以通过两种方式生成特权提示：**(1) 离线提示（固定）**。固定的提示生成器（例如，从 $\tau^\star$中一次性提取）较为简单，但它不会适应学习者，并且可能在训练过程中逐渐失准。**(2)在线提示**。我们会定期使用当前策略的副本来刷新提示生成器 π_θ。

在SAGE中，提示生成是在线的。我们通过提示策略 π_θ 生成与参考轨迹对齐的仅过程计划，来实现 $q_\phi(h \mid x, \tau^\star, \ell)$。

我们评估了三种变体：**(1) 固定特权提示：** q_ϕ由 π_{θ_0} 推导得出，并在初始化后冻结。**(2)在线特权提示（SAGE）：** q_ϕ 由 π_θ 推导得出，并在训练期间更新。**(3)外部教师提示：** q_ϕ 由更强的冻结模型生成（如有）。

如图 3所示，与无提示相比，添加提示在各提示等级下都能提升性能，这与提示能在奖励稀疏的情况下提高采样到信息丰富轨迹的概率相一致。此外，在线自提示的表现始终最佳，表明持续更新的自提示比固定提示更适合学习者。

4.3. 依赖策略的调度

我们通过离散强度变量 $\ell \in \{0, 1, \ldots, L\}$控制提示信息量，其中 $\ell = 0$ 对应可部署的无提示设定。调度器具有依赖策略的特性，即利用策略 π_θ 下近期推演收集的统计数据来调整 ℓ（停止梯度）。直观而言，只有当当前策略针对某个提示提供的学习信号不足时，我们才会提高 ℓ。

方案 1（SAGE-LIGHT）：轮次级准确率阈值。 令 $\hat{p}_{t-1}(x)$ 表示在上一轮次中，使用当前提示等级下策略模型的 rollout 测得的提示x 的经验成功率。给定目标阈值$\alpha \in (0, 1)$，当提示过于

算法 1 SAGE / SAGE-LIGHT：带特权监督的自提示对齐GRPO

要求： 训练集 $\mathcal{D} = \{(x,\ \tau^\star)\}$，策略模型 π_θ，组大小 G，KL权重 β，稳定器 $\epsilon > 0$，最大提示级别 L，参考策略 π_{ref}，基于 π_θ 的提示生成器 $q_\phi(h \mid x,\ \tau^\star,\ \ell)$，阈值 α（仅限 SAGE-LIGHT）。

初始化策略参数 θ；初始化逐提示级别映射 $\ell(x) \leftarrow 0$，适用于所有 $x \in \mathcal{D}$。
重复执行轮次：
  对于每个小批量 $\{(x_b,\ \tau_b^\star)\}_{b=1}^{B}$：
    对于 $b = 1, \ldots, B$ **do**
      如果 SAGE-LIGHT**则**
        如果 轮次 > 1 **且** $\bar{R}_b < \alpha$ **则** $\ell(x_b) \leftarrow$ 最小$\{\ell(x_b) + 1,\ L\}$
        采样 $h_b \sim q_\phi(h \mid x_b,\ \tau_b^\star,\ \ell(x_b))$
      else
        针对 $\ell = 0, \cdots\cdots, L$ **do**
          采样 $\tilde{h}_b \sim q_\phi(h \mid x_b, \tau_b^\star, \ell)$
          采样 $\tilde{\tau}_{b,\ i} \sim \pi_\theta(\cdot \mid x_b,\ \tilde{h}_b)$ 针对 $i = 1,\ \ldots\ldots,\ G$ 并使用计算资源 $\tilde{R}_{b,\ i} \leftarrow R(x_b,\ \tilde{\tau}_{b,\ i})$
          如果 $\sum_{i=1}^{G} \tilde{R}_{b,\ i} > 0$**或** $\ell = L$**则**
            $h_b \leftarrow \tilde{h}_b$，$\tau_{b,i} \leftarrow \tilde{\tau}_{b,i}$，$R_{b,i} \leftarrow \tilde{R}_{b,i}$；**中断**
      （如果使用 SAGE-LIGHT）采样 $\tau_{b,\ i} \sim \pi_\theta\ (\cdot \mid x_b,\ h_b)$，针对 $i = 1$、……、G，并计算 $R_{b,\ i} \leftarrow R\ (x_b,\ \tau_{b,\ i})$。
      计算 $\bar{R}_b \leftarrow \frac{1}{G}\sum_{i=1}^{G} R_{b,\ i}$，$s_b \leftarrow \sqrt{\frac{1}{G}\sum_{i=1}^{G}(R_{b,\ i} - \bar{R}_b)^2}$，$A_{b,\ i} \leftarrow \frac{R_{b\ ,\ i} - \bar{R}_b}{s_b + \epsilon}$。
    $\mathcal{L}_{\text{pg}} \leftarrow -\frac{1}{BG}\sum_{b=1}^{B}\sum_{i=1}^{G} A_{b,\ i} \sum_{t=1}^{T_b\ ,\ i}$ 日志 $\pi_\theta(y_{b,\ i,\ t} \mid x_b,\ h_b,\ y_{b,\ i,\ <t})$
    $\mathcal{L}_{\text{kl}} \leftarrow \frac{1}{B}\sum_{b=1}^{B} \text{KL}(\pi_\theta(\cdot \mid x_b, h_b) \,\|\, \pi_{\text{ref}}(\cdot \mid x_b, h_b))$
    更新 $\theta \leftarrow \theta - \eta\, \nabla_\theta \big(\mathcal{L}_{\text{pg}} + \beta\, \mathcal{L}_{\text{kl}}\big)$
部署： 设置 $\ell = 0$so $h = \varnothing$，然后运行 $\pi_\theta(\cdot \mid x)$。

(SAGE)

难：

$$\ell_t(x) = \min\big\{\, \ell_{t-1}(x) + 1,\ L \,\big\} \quad \text{if } \hat{p}_{t-1}(x) < \alpha, \quad (12)$$

否则保持 $\ell_t(x) = \ell_{t-1}(x)$ 不变。因此，只有当成功持续罕见时，才会启用提示。

方案 2（SAGE）：组退化触发机制。GRPO 需要组内结果存在差异，才能产生非零更新。因此，我们采用一种更局部的触发机制，根据探测组是否包含任何正向样本来判断。对于在当前强度 $\ell_{t-1}(x)$ 下从 $\pi_\theta(\cdot \mid x,\ h)$ 推演得到的小型探测组 $\{\tau_i\}_{i=1}^{G}$，定义 $z(x) = \mathbb{I}\Big[\sum_{i=1}^{G} R(x,\ \tau_i) = 0\Big]$，即当组内没有正向 rollout 时，$z(x) = 1$。然后，我们仅对这类退化提示提高提示强度：

$$\ell_t(x) = \min\big\{\, \ell_{t-1}(x) + 1,\ L \,\big\} \quad \text{if } z(x) = 1, \quad (13)$$

否则保持 $\ell_t(x) = \ell_{t-1}(x)$ 不变。此规则针对稀疏奖励中特定的有限样本病态现象：当组内没有正向 rollout 时，标准化 GRPO 优势会坍缩，策略梯度估计器也会随之消失。

讨论。 SAGE-LIGHT（方案 1）的计算资源效率较高，因为它只在轮次级别更新提示强度（不会为 rollout 增加额外计算），但对奖励突然崩塌的反应可能较慢。SAGE（方案 2）的反应更灵敏，并通过无正样本触发机制直接针对 GRPO 的失效模式，代价是需要额外的探测 rollout。我们报告了两种方案的结果，并发现无正样本触发机制通常能更快地从困难提示上的训练停滞中恢复，并取得更好的表现。

总体而言，算法 1概述了 SAGE 的实现。每个提示在每个轮次抽取 一个 提示（给定 ℓ），该提示的所有 Grollout 共享相同的上下文。此设计减少了提示带来的不必要方差。

5. 实证结果

模型。我们使用数学专业化程度各异的大语言模型：Llama-3.2-3B-Instruct（Meta，2024年）、Qwen2.5-7B-Instruct（Yang等人，2024年）和 Qwen3-4B-Instruct-2507（Yang等人，2025年），分别代表低、中、高水平的数学导向优化；其中最后一个模型通过强化学习进行了大规模训练。

训练集。 我们的训练数据来自 OpenR1- Math-220k（Hugging Face，2025），使用 Nu- minaMath 1.5（Li等人，2024）中的提示，以及由 DeepSeek-R1（DeepSeek-AI等人，2025）生成的推理轨迹。初始数据集包含 94k 条提示。为确保答案可验证，我们使用 *Math-Verify* 工具（Kydl´ı˘cek）移除 DeepSeek-R1 推理轨迹未能正确通过验证的提示，最终得到 64k 条提示。这些提示用于图2。由于资源有限，我们从该集合中进一步子采样出 15k 条提示，并将相应的 DeepSeek-R1 推理轨迹限制在 8,192 个词元以内。这一限制是必要的，因为我们的一个基线模型 LUFFY（Yan等人，2025）依赖这些推理轨迹，而过长的轨迹会显著增加强化学习训练时间。由于我们没有根据通过率筛选提示，最终得到的 15k 条提示涵盖了广泛的难度等级，类似于实际的强化学习训练数据集。

表1。三个大语言模型在分布内和分布外任务上的准确率。**最佳**和次佳结果分别以粗体和下划线标示。与基线方法相比，SAGE 和 SAGE-LIGHT 在不同大语言模型上的平均表现始终更优。

方法	分布内							分布外			
	AIME24 / 25	AMC23	MATH-500	Minerva	奥林匹克竞赛	Avg.	Δ	GPQA	MMLU-Pro	Avg.	Δ
Llama-3.2-3B-Instruct	6.5 / 0.6	22.8	44.7	17.8	14.2	17.8	0	17.9	27.0	22.5	0
SFT	0.4 / 0.6	9.5	26.9	5.1	6.5	8.2	−9.6	11.6	18.8	15.2	−7.3
GRPO	6.7 / 0.8	29.5	52.1	20.5	21.8	21.9	+4.1	26.3	39.8	33.1	+10.6
LUFFY	4.4 / 0.4	18.6	38.9	14.3	11.9	14.7	−3.1	16.0	26.7	21.4	−1.1
Scaf-GRPO	7.7 /**2.3**	28.8	51.7	19.4	19.5	21.5	+3.7	24.1	38.0	31.0	+8.5
SAGE-LIGHT	8.8 / 1.9	32.2	54.1	**20.8**	20.1	23.0	+5.2	26.8	39.6	33.2	+10.7
SAGE	**9.2** /0.8	**34.7**	**56.3**	20.1	**22.0**	**23.9**	**+6.1**	**27.3**	**40.7**	**34.0**	**+11.5**
Qwen2.5-7B-Instruct	13.8 / 6.7	53.4	75.7	38.1	39.2	37.8	0	37.1	56.4	46.7	0
SFT	3.5 / 7.1	30.9	56.2	20.0	21.7	23.2	−14.6	9.5	35.6	22.5	−24.2
GRPO	15.0 / **13.5**	55.5	79.2	39.1	44.5	41.1	+3.3	37.2	57.6	47.4	+0.7
LUFFY	**17.1/ 13.5**	55.2	**81.3**	39.0	44.2	41.7	+3.9	**38.1**	59.1	**48.6**	**+1.9**
Scaf-GRPO	14.6 / 12.7	58.8	78.0	**39.8**	42.0	41.0	+2.2	36.6	58.4	47.5	+0.8
SAGE-LIGHT	**17.1** / 11.7	58.1	79.9	38.6	**46.1**	41.9	+4.1	36.6	58.8	47.7	+1.0
SAGE	16.0 / 12.5	**60.3**	80.0	39.3	45.9	**42.3**	**+4.5**	38.0	**59.3**	**48.6**	**+1.9**
Qwen3-4B-Instruct	52.1 / 43.1	92.2	93.6	46.1	67.7	65.8	0	57.6	70.9	64.3	0
SFT	14.4 / 22.1	55.2	78.9	38.1	40.2	41.5	−24.3	29.4	52.6	41.0	−23.3
GRPO	55.8 / 45.0	**95.0**	96.0	**50.1**	70.4	68.7	+2.9	57.0	72.0	64.5	+0.2
LUFFY	42.3 / 36.0	86.4	91.1	48.2	59.4	60.6	−5.2	31.9	46.6	39.3	−25.0
Scaf-GRPO	**59.8**/ 45.2	92.2	95.1	48.9	69.8	68.5	+2.7	54.3	72.1	63.2	−1.1
SAGE-LIGHT	59.2 / 47.1	92.2	95.1	49.5	70.5	68.9	+3.1	57.1	72.0	64.5	+0.2
SAGE	58.1 / **52.1**	94.2	**95.4**	49.2	**71.2**	**70.0**	**+4.2**	**57.8**	**72.5**	**65.2**	**+0.9**

评测集。我们主要使用六个广泛采用的数学基准测试评估模型：AIME24（MAA 委员会，2024）、AIME25（MAA 委员会，2025）、AMC23（Li等人，2024）、MATH-500（Hendrycks等人，2021）、Minerva数学（Lewkowycz等人，2022）和奥林匹克竞赛基准（He等人，2024）。此外，我们还纳入两个非数学基准测试：GPQA-diamond（Rein等人，2024）和 MMLU-Pro（Wang等人，2024），以评估训练后模型的泛化能力。**值得注意的是，我们只在训练期间使用提示。评测时，输入大语言模型的只有提示本身。**

基线方法。我们将 SAGE 与以下基线方法进行比较：(1) 监督微调（SFT），即使用来自 DeepSeek-R1 的推理轨迹对模型进行微调；(2) GRPO（Shao 等人，2024年），该方法不使用任何提示进行学习；(3) LUFFY（Yan等人，2025年），该方法将一条同策略轨迹替换为 DeepSeek-R1 中对应的正确轨迹；(4) Scaf-GRPO（Zhang等人，2025年），该方法融入了 GPT-5.2 在低推理力度设置下生成的提示。值得注意的是，SFT、LUFFY 和 Scaf-GRPO 都依赖更强大的外部大语言模型，而 SAGE 仅通过自生成提示进行学习。为确保公平比较，我们使用相同的 1.5 万个采样提示，通过其开源实现复现 LUFFY 和 Scaf-GRPO，并仅对齐批次大小和训练步数。

实现细节。所有实验均在 8 张 A100 GPU 上运行，训练使用 verl（Sheng等人，2025年），采样使用 vLLM（Kwon等人，2023年）。遵循 DAPO（Yu 等人，2025年）的做法，我们通过设置 $\beta = 0$禁用 KL 项，并采用非对称裁剪，其中 $\epsilon_{\text{low}} = 0.2$ 和 $\epsilon_{\text{high}} =$ 0。28。除非另有说明，训练和评测的最大响应长度均设为 8096，[1] 批次大小为 128，每个提示生成 8 条轨迹，[2] 总训练步数为 500。每 50 步进行一次评测，并报告所有检查点中的最佳平均准确率。我们将 L 设为 3，并将 $\alpha = 0.35$ 用于 SAGE-LIGHT。用于提示生成与注入的提示详见附录 B。所有方法完整的训练与评测设置见附录C。

5.1. 主要结果

我们在表 1 中报告了所有方法和大语言模型在八个基准测试上的结果，相应的训练动态见图4。在三种基础模型上，SAGE 的平均表现始终优于所有基线方法，在各项基准测试上的平均成绩分别提升了 +6.1（Llama-3.2）、+4.5（Qwen2.5）和 +4.2（Qwen3）。尽管各大语言模型针对数学任务的优化程度各不相同，我们仍对它们使用固定的训练集。因此，对 Qwen3 而言，该训练集相对容易，而对 Llama 而言则更具挑战性；这一差异也反映在图 2 中。尽管如此，SAGE 始终能提升所有大语言模型的表现，展现出稳健且有效的泛化能力。

SAGE 与监督微调的比较。由于容易过拟合，监督微调的表现最差，甚至不如基础大语言模型

[1]主要结果采用 8096，这是 LUFFY 的要求；其余实验则受资源限制，采用 2048。[2]由于响应长度较长导致训练速度较慢，Qwen3-4B-Instruct 使用 4 条轨迹。

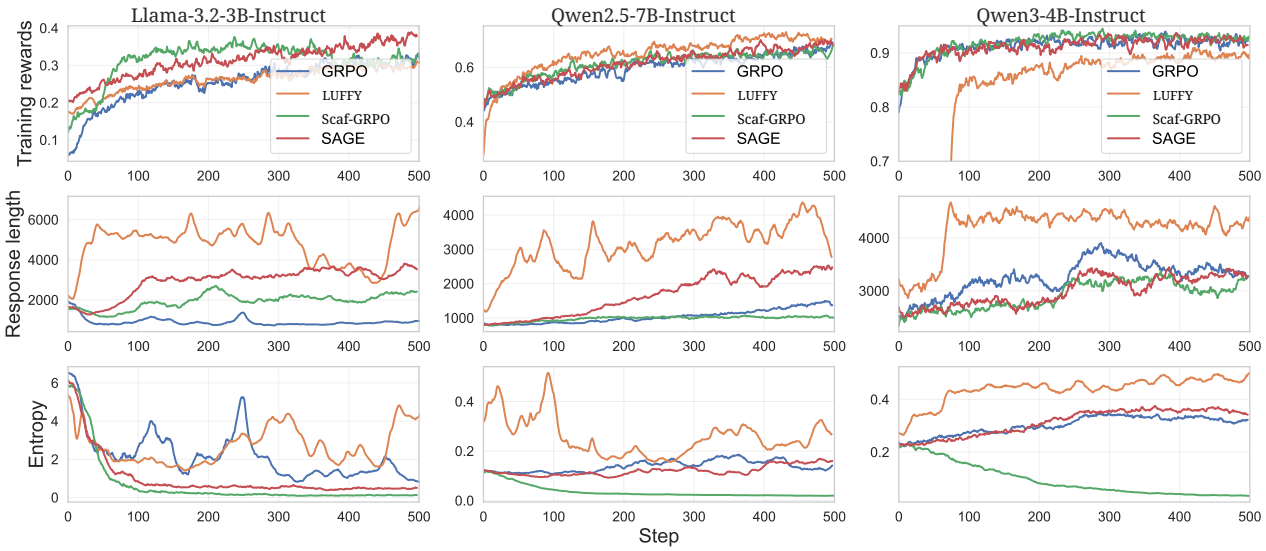

图 *4*。 不同方法的训练动态。对于训练奖励，应关注其趋势而非数值，因为添加提示（SAGE 和 Scaf-GRPO）会改变提示的难度，而使用正确的离策略轨迹（LUFFY）会提高奖励。(1) LUFFY 表现出最明显的不稳定性：对于 Llama，其熵非常高；对于 Qwen3，在训练开始时奖励非常低。这是因为它模仿了离策略轨迹，而该轨迹的分布可能与策略模型不一致。(2) Scaf-GRPO 的熵最低，意味着探索较少。(3) SAGE 保留了在策略特性，熵适中，且响应长度稳定增长，这通常意味着推理模式更佳。

表 *2*。 没有任何训练信号的提示所占百分比，即在整个训练过程中都未采样到这些提示的任何正确轨迹。

方法	Llama-3.2-3B	Qwen2.5-7B	Qwen3-4B
Base	56.9%	29.8%	21.8%
GRPO	40.2%	10.3%	1.3%
SAGE	30.0%	8.2%	1.0%

训练数据。相比之下，SAGE 保留了强化学习的特性，有选择地使模型的分布向正确轨迹集中。

SAGE 与 GRPO 的比较。表2 报告了从未提供训练信号的提示所占比例。与 GRPO 相比，SAGE 对提示词集的利用效率显著更高。这一效果在较弱的大语言模型 Llama-3.2 上尤为明显：借助自生成提示，SAGE 成功利用了多出 10% 的提示，因此相较 GRPO 实现了最大的性能提升（+2.0）。对于较强的模型 Qwen3，SAGE 的表现则与 GRPO 更为相似，提示利用率几乎相同。尽管如此，尽管只多利用了 0.3% 的提示，SAGE 相较 GRPO 仍实现了 +1.3 的准确率提升。这些困难提示在强化学习中发挥着关键作用，这与先前的研究（Xiong 等人，2025年；Yu 等人，2025年）一致；这些研究倾向于在通过率较低的提示词集上进行强化学习。此外，如图4所示，对于 Llama-3.2 和 Qwen2.5，SAGE 的响应长度增长都快于 GRPO，这是因为 SAGE 能从在 GRPO 下无法提供任何信号的困难提示中学习。

SAGE 与 LUFFY 的比较。继监督微调之后，LUFFY 的离策略行为程度位居第二，因为其一条轨迹由不同的模型生成。如图4所示，在训练早期，响应长度急剧增加，反映了大语言模型模仿更强模型的倾向。然而，这种离策略设置会因策略模型与更强模型之间的不匹配而导致训练不稳定性。具体而言，使用 LUFFY 训练的 Llama-3.2 熵过高，且响应长度大幅波动；而使用 LUFFY 训练的 Qwen3 在训练开始时奖励极低。如表1所示，LUFFY 仅在 Qwen2.5 上优于 GRPO（但仍不如 SAGE），而在 Llama-3.2 和 Qwen3 上的表现均不如基础模型。

SAGE 与 Scaf-GRPO 的对比。Scaf-GRPO 依赖于更强模型生成的提示（例如，在我们的设置中使用 GPT-5.2）。如图 4 所示，在所有方法中，其熵最低，表明探索程度有限。这种行为可能源于提示透露了过多信息。相比之下，SAGE 的熵水平与 GRPO 相当，[3] 同时在表 1 中始终优于 Scaf-GRPO。此外，从自生成提示中学习能够实现更加端到端的训练过程，并简化实现。

SAGEvs, SAGE-LIGHT. SAGE-LIGHT的准确率略低，但提升了效率，仅需53%的

[3]GRPO 在 Llama-3.2 上的熵异常高。

表 3. Qwen2.5-7B-Instruct 上的训练时间。

GRPO	LUFFY	Scaf-GRPO	SAGE	SAGE-LIGHT
1.0× (25.3小时)	1.2×	1.5×	2.3×	1.2×

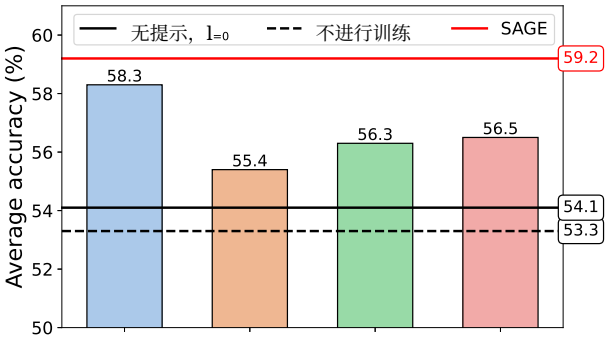

图 5. 在与图 3 使用相同提示词集进行训练的 Qwen3-4B-Instruct 上开展的消融实验。

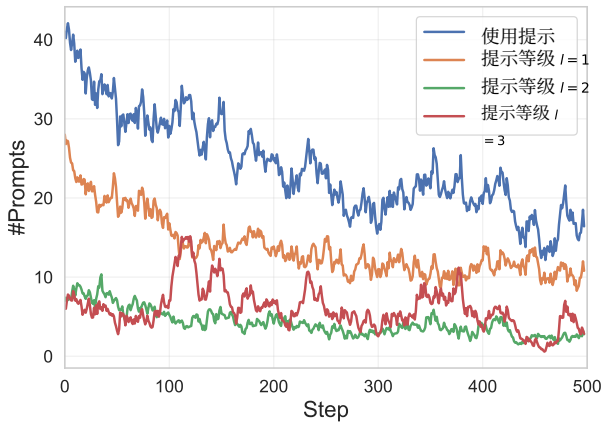

图 6. 在 Llama-3.2-3B-Instruct 上训练期间使用提示的提示数量。批次大小为 128。模型随训练步使用更少的提示，表明模型变得更强大。

SAGE 的训练时间。

分布外性能。 在表1中，分布外基准上的性能与分布内基准上的性能呈现出相似的模式。SAGE和SAGE-LIGHT始终分别取得最佳和次佳准确率，表明其具有更强的泛化能力。

5.2. 讨论

延迟。 SAGE 的一个潜在局限是延迟较高，因为当无法为提示采样到正确轨迹时，它必须即时生成并使用提示。表3 报告了不同强化学习方法的训练时间。其中，SAGE 的训练成本最高，而 SAGE-LIGHT 所需时间仅略多于 GRPO。对于高度复杂的提示，SAGE 可能会跨多个等级（从 $l = 0$ 到 $l = 3$）采样提示，这会增加计算开销。相比之下，SAGE-LIGHT 利用上一轮次的提示词准确率来选择合适的提示等级，因此只从单个等级进行采样。这两种 SAGE 变体为具有不同效率要求的实践者提供了灵活的权衡方案，并且都始终优于基线方法。

采用更多提示的离线方法。 在图 3中，在线自提示会在每个训练步骤生成一个新提示，而离线自提示则依赖于训练前生成的固定提示。在线自提示表现更优的一种可能解释是提示多样性更高。为检验这一点，我们进行了额外的消融实验（图 5）：使用多个提示的离线自提示。具体而言，在训练前，我们使用基础大语言模型生成 10 个提示，温度设为温度=1.0。训练期间，每一步都会为同一提示使用不同的提示，确保相同的提示在不同时间与多样化的提示配对。我们观察到，提示多样性的提高确实改善了性能，相较于标准离线自提示提升了 +0.9。不过，在线自提示仍以 2.0 的优势胜过这一变体。我们认为，以在线方式生成的提示比单纯的多样性带来更多益处。

离策略。 我们消融策略梯度是否与采样上下文相匹配：SAGE 优化 $\log \pi_\theta(\cdot \mid x,\ h)$，而离策略变体则使用 h 进行采样，但优化$\log \pi_\theta(\cdot \mid x)$。图 5 显示，离策略变体（56.5）相比同策略 SAGE（59.2），甚至相比单一提示等级基线（58.3），均出现明显下降。

相同提示等级与 SAGE 的对比。 SAGE 并不依赖固定的提示等级。相反，仅当当前等级无法产生正确响应时，它才会自适应地提高提示等级。在图 5 中，该策略相比使用常数提示等级（例如在线 $l = 2$）带来了明显的性能提升，改进幅度为 +0.9。该设计能够更有效地利用困难提示词，使大语言模型能够从更弱的提示逐步学习，直至 $l = 0$。

相对于步骤，提示更少。 在图6中，我们可以观察到大语言模型在训练过程中使用的提示越来越少。这表明自提示确实增强了强化学习。大语言模型变得越来越强大，并能够逐渐在没有提示帮助的情况下解决困难问题。

案例研究。 一个关于提示如何提供帮助的示例（附录A）。

6. 相关工作

数据重采样与外部引导。数据选择和过滤广泛应用于大语言模型的在线强化学习中（Zhang等人；Dong等人，2023年；Xiong 等人，2023年；Dong等人，2024年；Shi 等人，2024年；Liao 等人，2025年；Feng 等人，2025年），

对于 GRPO 风格的方法而言，这些做法尤为重要，因为在奖励稀疏时，组内优势值可能会坍缩。以往研究主要通过重塑训练分布或注入外部引导来缓解这一问题。一种常见的权宜之计是跳过退化组，并对提示进行重采样或提高其采样权重（Yu 等人，2025年；Xiong 等人，2025a；Yao 等人，2025年；Li等人，2025年），这能提升效率，但会使训练偏向成功概率不为零的提示。另一种思路是通过加入来自更强教师模型、参考模型或离线缓冲区的正向轨迹来引导学习（Zhang等人，2025a；Yan等人，2025年），但将其与同策略展开混合时，可能会引入上下文或分布失配。相比之下，SAGE 通过特权提示改变展开分布，同时保留纯粹的同策略目标，既不丢弃困难提示词，也不依赖静态外部缓冲区。

特权提示与 SAGE。尽管在强化学习领域，利用计划或标准解答等中间指导已有悠久历史（Ng等人，1999；Szepesv´ari，2022年；Ouyang等人，2022年），近期针对大语言模型的改编通常通过启发式的“批次手术”来实现提示。例如，Zhang等人（2025年）通过在检测到失败后，向 rollout 批次中添加带提示轨迹来缓解崩溃。这种方法会在同一组内混合不同上下文（例如，x 和 $x,\ h$），模糊了组内基线和优势归一化的含义。此外，外部提示生成器可能无法根据学习者当前的能力进行校准。SAGE通过三个关键设计选择脱颖而出：（1）将提示形式化为一种明确增强的同策略采样过程，从而确保GRPO损失定义良好；（2）采用依赖于策略的强度调度器，仅在“学习门控”关闭时启用提示；（3）通过滞后策略进行在线自提示，确保提示分布跟随学习者不断演变的支持范围，而非依赖可能不匹配的外部教师。

7. 结论

我们发现，在稀疏的 0/1 奖励下，GRPO 存在有限样本退化现象。当组奖励相同时，优势标准化会发生坍缩，小批量梯度在困难提示上消失。我们提出 SAGE，这是一种特权过程提示方法，在训练期间注入源自参考解答的提示，以改变展开分布，同时保留原始奖励定义。该方法根据检测到的组坍缩情况，采用依赖策略的调度机制控制提示强度；推理时则使用无提示策略。大量实验验证了该方法在各项任务中的改进效果。

致谢

本研究部分得到了荷兰科学研究组织（NWO）项目 VI.C.192.080 的支持。

影响声明

这项工作可以降低大语言模型在可验证任务上的强化学习训练成本，并提升其稳定性。风险包括被滥用来优化有害的可验证目标。

参考文献

DeepSeek-AI、Guo, D.、Yang, D.、Zhang, H.、Song, J.、Zhang, R.、Xu, R.、Zhu, Q.、Ma, S.、Wang, P.、Bi, X.、Zhang, X.、Yu, X.、Wu, Y.、Wu, Z. F.、Gou, Z.、Shao, Z.、Li, Z.、Gao, Z.、Liu, A.、Xue, B.、Wang, B.、Wu, B.、Feng, B.、Lu, C.、Zhao, C.、Deng, C.、Zhang, C.、Ruan, C.、Dai, D.、Chen, D.、Ji, D.、Li, E.、Lin, F.、Dai, F.、Luo, F.、Hao, G.、Chen, G.、Li, G.、Zhang, H.、Bao, H.、Xu, H.、Wang, H.、Ding, H.、Xin, H.、Gao, H.、Qu, H.、Li, H.、Guo, J.、Li, J.、Wang, J.、Chen, J.、Yuan, J.、Qiu, J.、Li, J.、Cai, J. L.、Ni, J.、Liang, J.、Chen, J.、Dong, K.、Hu, K.、Gao, K.、Guan, K.、Huang, K.、Yu, K.、Wang, L.、Zhang, L.、Zhao, L.、Wang, L.、Zhang, L.、Xu, L.、Xia, L.、Zhang, M.、Zhang, M.、Tang, M.、Li, M.、Wang, M.、Li, M.、Tian, N.、Huang, P.、Zhang, P.、Wang, Q.、Chen, Q.、Du, Q.、Ge, R.、Zhang, R.、Pan, R.、Wang, R.、Chen, R. J.、Jin, R. L.、Chen, R.、Lu, S.、Zhou, S.、Chen, S.、Ye, S.、Wang, S.、Yu, S.、Zhou, S.、Pan, S.、Li, S. S.、Zhou, S.、Wu, S.、Ye, S.、Yun, T.、Pei, T.、Sun, T.、Wang, T.、Zeng, W.、Zhao, W.、Liu, W.、Liang, W.、Gao, W.、Yu, W.、Zhang, W.、Xiao, W. L.、An, W.、Liu, X.、Wang, X.、Chen, X.、Nie, X.、Cheng, X.、Liu, X.、Xie, X.、Liu, X.、Yang, X.、Li, X.、Su, X.、Lin, X.、Li, X. Q.、Jin, X.、Shen, X.、Chen, X.、Sun, X.、Wang, X.、Song, X.、Zhou, X.、Wang, X.、Shan, X.、Li, Y. K.、Wang, Y. Q.、Wei, Y. X.、Zhang, Y.、Xu, Y.、Li, Y.、Zhao, Y.、Sun, Y.、Wang, Y.、Yu, Y.、Zhang, Y.、Shi, Y.、Xiong, Y.、He, Y.、Piao, Y.、Wang, Y.、Tan, Y.、Ma, Y.、Liu, Y.、Guo, Y.、Ou, Y.、Wang, Y.、Gong, Y.、Zou, Y.、He, Y.、Xiong, Y.、Luo, Y.、You, Y.、Liu, Y.、Zhou, Y.、Zhu, Y. X.、Xu, Y.、Huang, Y.、Li, Y.、Zheng, Y.、Zhu, Y.、Ma, Y.、Tang, Y.、Zha, Y.、Yan, Y.、Ren, Z. Z.、Ren, Z.、Sha, Z.、Fu, Z.、Xu, Z.、Xie, Z.、Zhang, Z.、Hao, Z.、Ma, Z.、Yan, Z.、Wu, Z.、Gu, Z.、Zhu, Z.、Liu, Z.、Li, Z.、Xie, Z.、Song, Z.、Pan, Z.、Huang, Z.、Xu, Z.、Zhang, Z. 和 Zhang, Z. Deepseek-r1：通过强化学习激励大语言模型的推理能力，2025年。网址https://arxiv.org/abs/2501.12948。

Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., SHUM, K., 和 Zhang, T. RAFT：面向生成式基础模型对齐的奖励排序微调。机器学习研究汇刊

，2023年。ISSN 2835-8856。网址：https://openreview.net/forum?id=m7p5O7zblY。

Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., Jiang, N., Sahoo, D., Xiong, C., and Zhang, T. RLHF 工作流程：从奖励建模到在线 RLHF。*arXiv* 预印本 *arXiv:2405.07863*，2024年。

Feng, Y., Kwiatkowski, A., Zheng, K., Kempe, J., and Duan, Y. Pilaf：用于奖励建模的最优人类偏好采样。*arXiv* 预印本 *arXiv:2502.04270*，2025年。He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., 等。奥林匹克竞赛基准：一项旨在推动通用人工智能发展的挑战性基准，涵盖奥林匹克竞赛级别的双语多模态科学问题。*arXiv* 预印本 *arXiv:2402.14008*，2024年。Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. 使用 MATH 数据集衡量数学问题求解能力。*arXiv* 预印本 *arXiv:2103.03874*，2021年。Hu, J., Wu, X., Zhu, Z., Xianyu, Wang, W., Zhang, D., and Cao, Y. OpenRLHF：易于使用、可扩展且高性能的 RLHF 框架。*arXiv* 预印本 *arXiv:2405.11143*，2024 年。Hugging Face。Open R1：对 DeepSeek-R1 的完全开放复现，2025年1月。网址 https://github.com/huggingface/open-r1。Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. 使用 PagedAttention 高效管理大语言模型服务的内存。载于*ACM SIGOPS* 操作系统原理第 *29* 届研讨会论文集，2023年。Kydlıˇcek, H. Math-Verify：数学验证库。网址 https://github.com/huggingface/math-verify。Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., 等。使用语言模型解决定量推理问题。神经信息处理系统进展，35:3843–3857，2022年。Li, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S., Rasul, K., Yu, L., Jiang, A. Q., Shen, Z., 等。NuminaMath：AI4Maths 中规模最大的公开数据集，包含 86 万组竞赛数学问题及其解答。*Hugging Face* 仓库，13(9):9，2024年。Li, Z., Chen, C., Yang, T., Ding, T., Sun, R., Zhang, G., Huang, W., and Luo, Z.-Q. 背包强化学习：通过优化预算分配释放大语言模型的探索能力。*arXiv* 预印本 *arXiv:2509.25849*，2025年。

Liao, B., Xu, Y., Dong, H., Li, J., Monz, C., Savarese, S., Sahoo, D. 和 Xiong, C. 面向高效大语言模型推理的奖励引导推测解码。*arXiv* 预印本 *arXiv:2501.19324*，2025年。Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S. 和 Lin, M. 理解类 r1-zero 训练：批判性视角。*arXiv* 预印本 *arXiv:2503.20783*，2025年。MAA 委员会。AIME 问题与解答。 https://artofproblemsolving.com/wiki/index.php/AIME_问题_与_解答，2024年。MAA 委员会。AIME 问题与解答。 https://artofproblemsolving.com/wiki/index.php/AIME_问题_与_解答，2025年。Meta。介绍 Meta Llama 3：迄今能力最强的开放可用大语言模型。*Meta AI* 博客，2024年。https://ai.meta.com/blog/meta-llama-3/。Ng, A. Y., Harada, D. 和 Russell, S. 奖励变换下的策略不变性：奖励塑形的理论与应用。载于国际机器学习会议，第 99 卷，第 278–287 页。Citeseer，1999 年。Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. 等。训练语言模型以遵循人类反馈下的指令。神经信息处理系统进展，35:27730–27744，2022年。Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J. 和 Bowman, S. R. GPQA：研究生水平的谷歌不可解问答基准。载于首届语言建模会议，2024年。Schulman, J., Wolski, F., Dhariwal, P., Radford, A. 和 Klimov, O. 近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017年。Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y., Wu, Y. 和 Guo, D. DeepSeekMath：突破开放语言模型数学推理的极限。*arXiv* 预印本 *arXiv:2402.03300*，2024年。Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H. 和 Wu, C. HybridFlow：灵活高效的 RLHF 框架。载于第二十届欧洲计算机系统会议论文集，第 1279–1297 页，2025年。Shi, R., Zhou, R. 和 Du, S. S. 在线直接偏好优化中采样器的关键作用。*arXiv* 预印本 *arXiv:2409.19605*，2024年。

Szepesv´ari, C. 强化学习算法。Springer Nature, 2022年。Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z. 等。MMLU-Pro：更稳健、更具挑战性的多任务语言理解基准。神经信息处理系统进展, 37:95266–95290，2024年。Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N. 和 Zhang, T. 基于人类反馈的迭代偏好学习：在 KL 约束下连接 RLHF 的理论与实践。2023年。Xiong, W., Yao, J., Xu, Y., Pang, B., Wang, L., Sahoo, D., Li, J., Jiang, N., Zhang, T., Xiong, C. 等。从拒绝采样到 REINFORCE：一种极简的大语言模型推理方法。*arXiv* 预印本 *arXiv:2504.11343*，2025a。Xiong, W., Ye, C., Liao, B., Dong, H., Xu, X., Monz, C., Bian, J., Jiang, N. 和 Zhang, T. REINFORCE-ADA：非线性强化学习目标下的自适应采样框架。*arXiv* 预印本 *arXiv:2510.04996*，2025年。Yan, J., Li, Y., Hu, Z., Wang, Z., Cui, G., Qu, X., Cheng, Y. 和 Zhang, Y. 在离策略指导下学习推理。*arXiv* 预印本 *arXiv:2504.14945*，2025年。Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H. 等。Qwen2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*，2024年。Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C. 等。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025年。Yao, J., Hao, Y., Zhang, H., Dong, H., Xiong, W., Jiang, N. 和 Zhang, T. 通过拒绝采样与强化学习中的梯度方差最小化优化思维链推理器。*arXiv* 预印本 *arXiv:2505.02391*，2025年。Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Fan, T., Liu, G., Liu, L., Liu, X. 等。DAPO：大规模开源大语言模型强化学习系统。*arXiv* 预印本 *arXiv:2503.14476*, 2025年。Zhang, C., Shen, W., Zhao, L., Zhang, X., Qi, L., Dou, W. 和 Bian, J. RLHF 中的策略过滤：微调大语言模型以实现代码生成。Zhang, K., Lv, A., Li, J., Wang, Y., Wang, F., Hu, H. 和 Yan, R. Stephint：多层级逐步提示增强强化学习推理。*arXiv* 预印本 *arXiv:2507.02841*，2025a。Zhang, X., Wu, S., Zhu, Y., Tan, H., Yu, S., He, Z. 和 Jia, J. Scaf-GRPO：用于增强大语言模型推理的脚手架式组相对策略优化。*arXiv* 预印本 *arXiv:2510.19807*, 2025b。

Zhang, Y., Yao, W., Yu, C., Liu, Y., Yin, Q., Yin, B., Yun, H. 和 Li, L. 通过自适应推演和响应复用提升 RLVR 的采样效率。*arXiv* 预印本 *arXiv:2509.25808*，2025c。

A. 特权提示示意图

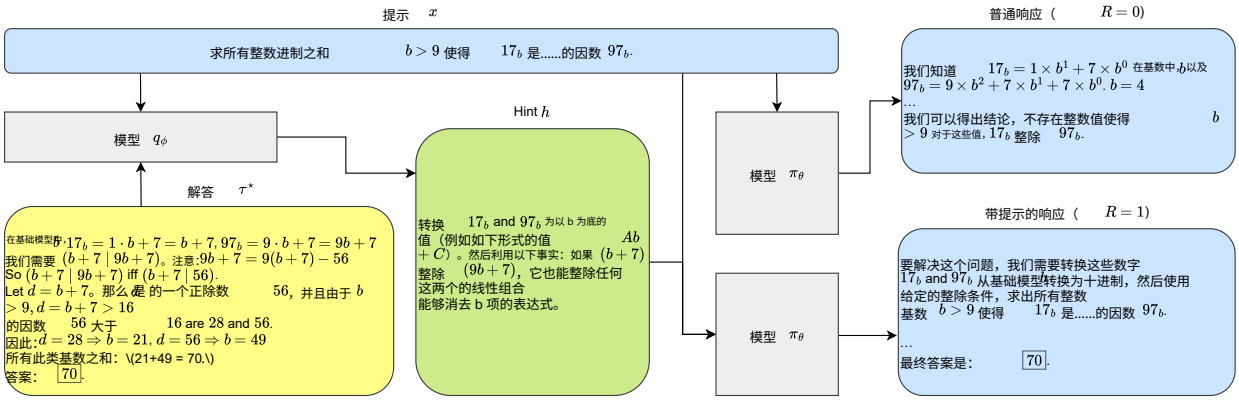

图 *A.1.* 单个提示的特权提示示例。给定数学提示 x，提示生成器 q_ϕ 在训练期间使用参考解答 $\tau^\star$ 生成程序性提示 (h)，概括中间推理，同时不泄露最终答案。在原始提示上展开策略 π_θ，可能得到奖励为零的错误解答；而以 (x, h) 为条件，则会改变展开分布，并使模型能够得出奖励为正的正确解答。任务回报保持不变；在部署时会移除提示，因此模型只根据原始提示运行。

案例研究：$17_b \mid 97_b$ 的渐进式特权提示

提示。 求所有满足 17_b 是 97_b 的因数的整数进制 $b > 9$ 之和。

参考解答（仅在训练期间可用）。 在基础模型 b，

$$17_b = 1 \cdot b + 7 = b + 7, \qquad 97_b = 9 \cdot b + 7 = 9b + 7.$$

我们需要 $b + 7 \mid 9b + 7$。注意，

$$9b + 7 = 9(b + 7) - 56,$$

因此，$b + 7 \mid 9b + 7$ 当且仅当 $b + 7 \mid 56$。令 $d = b + 7$。那么 d 是 56 的一个正因数，并且由于 $b > 9$，我们有 $d = b + 7 > 16$。大于 16 的 56 的因数是 28 和 56，因此 $b \in \{21, 49\}$，总和为 70。

特权提示（仅用于训练；测试时绝不显示）。

- **第 1 级（最简）。** 将以 b 为底的 b 进制数改写为普通整数，然后将整除条件转化为关于一个简单线性表达式的陈述。
- **第 2 级（中等）。** 如果 $(b + 7)$ 整除 $(9b + 7)$，那么它也整除这些表达式中任意一个能消去 b 项的线性组合。
- **第 3 级（详细）。** 从 $(9b + 7)$ 中减去 $(b + 7)$ 的某个常数，以消去 b，例如，$(9b + 7) - 9(b + 7)$。条件变为“$(b + 7)$ 整除一个常数”。枚举因数，并仅保留 $b > 9$。

代表性模型行为（示意性）。

- **无提示（$h = \varnothing$）：** 模型可能会错误展开 97_b (e.g.，将其视为多位数多项式），并错误地断定不存在有效的 $b > 9$，从而得出错误的终局决策，奖励因此为 0。
- **借助更强提示（等级 2/3）：** 模型会被引导至关键消去步骤 $9b + 7 - 9(b + 7) = -56$，此后便可枚举因数并恢复正确的基数及最终和（70）。

案例研究：在一个简单的可验证数学任务中使用特权提示。 为了具体说明特权提示这一想法，我们针对一道可验证的整除问题进行了一个小型案例研究。关键在于，提示属于训练时特权上下文：它们不会修改验证器或终局奖励，而是会重塑展开分布，使策略在有限采样下更有可能生成信息丰富的轨迹（例如，执行正确的代数化简）。实际上，如果没有提示，模型可能会反复走上错误的表示路径（例如，错误地展开以 b 为底的数），并在同一个 rollout 组中连续获得相同的零奖励，导致 GRPO 优势趋于零。借助

渐进式提示，模型会被引导至正确的约简路径，从而提高一个组中出现混合结果的概率，进而产生非退化的更新。

本示例符合 SAGE 的操作作用：提示不会改变验证器奖励，但会提高有限组中至少有一条 rollout 遵循有用 轨迹（此处为消去与因数枚举路径）的概率。这会增加组内结果多样性，降低退化的全零组出现频率，从而防止标准化GRPO优势在有限采样下于困难提示上发生坍缩。

B. 提示生成所用提示词及提示的用法

提示生成的系统提示词

你是一名辅导助手，负责生成渐进式提示，帮助学生解决难题，同时不直接揭示解答。

任务：
给定一个问题及其解答，生成 3 个提示等级，逐步引导学生独立解决问题。

提示等级：
- 等级 1：最少提示——指出关键概念或方法，但不提供具体细节
- 等级 2：中等提示——对方法或中间步骤提供更多指引
- 等级 3：详细提示——提供充分指导，但仍要求学生完成解答

指导原则：
- 切勿透露最终答案
- 每个层级都应建立在前一个层级的基础上
- 提示应启发问题求解，而不只是提供可照抄的步骤
- 根据学生的水平与解答之间的差距调整提示难度

输出格式：

“‘json { "level_1": "简短提示文本", "level _2": "中等长度提示文本", "level 3": "详细提 示文本"_} “‘

用于提示生成的用户提示

问题：{问
题}解答：
{解答}

强化学习系统提示词

请逐步推理，并将最终答案放在 \\boxed{} 中。

用于强化学习的问题提

{问题}以下是一个提
示，可帮助你：{提示}

C. 详细实现设置

C.1. 训练设置

监督微调。 我们使用 OpenRLHF（Hu 等人，2024年）进行监督微调，并将学习率设为 5e-5、批次大小设为 64、预热比例设为 10%，轮次数设为 3。我们在最终检查点上进行评估。

GRPO。 GRPO 与 SAGE 采用相同的训练设置，如 §5所述。

LUFFY。 我们使用开源实现[4] 复现 LUFFY，并将批次大小设为 128，*PPO*小批量大小 设为 64。这两个超参数在所有强化学习方法中保持不变。

Scaf-GRPO。 我们使用开源实现[5] 复现 Scaf-GRPO。

C.2. 评估设置

对于表 1 中的主要结果，我们将所有模型的最大响应长度设置为 8192，温度=*0.6* 和 *top p=0.95*。对于其余情况，我们将最大响应长度设置为 2048。

表*C.1*。 图 3 （左）中的详细数值。默认情况下，我们训练 200 个步骤，但对于由 * 表示的方法训练 400 个步骤。图 $l = 2$ 中的等级 3 结果来自由 * 表示的方法。不同于 SAGE 的默认训练设置，我们在此处设置 *PPO*小批量大小=*32*。

方法	提示等级	AIME24	AIME25	AMC23	MATH-500	Minerva	奥林匹克竞赛	Avg.
Qwen3-4B-Instruct	-	30.8	28.5	75.5	87.8	42.8	54.3	53.3
无提示	0	34.0	29.2	78.8	88.6	40.7	53.1	54.1
GPT-5 提示	1	34.0	29.8	83.0	89.6	42.2	56.9	55.9
离线自提示	1	35.8	28.3	80.2	89.9	42.4	56.9	55.6
在线自提示	1	35.0	28.5	82.7	90.0	42.7	61.0	56.7
GPT-5 提示	2	34.8	29.6	81.1	89.6	42.9	60.9	56.5
离线自提示	2	33.1	27.3	78.1	89.0	42.5	59.2	54.9
在线自提示	2	34.2	27.7	81.7	89.2	43.5	61.8	56.4
GPT-5 提示*	2	34.8	29.6	81.1	89.6	42.9	60.9	56.5
离线自提示*	2	34.0	26.0	79.8	88.5	43.7	60.4	55.4
在线自提示*	2	38.5	30.8	82.7	90.3	44.4	62.9	58.3
GPT-5 提示	3	37.9	26.9	78.3	89.2	43.1	62.1	56.3
离线自提示	3	36.7	27.5	83.1	89.6	43.3	61.5	57.0
在线自提示	3	36.7	27.2	83.1	89.9	43.2	62.6	57.1

D. 证明

本附录为第 3.1 节中的结果（以及补充分析）提供详细证明。

[4]https://github.com/ElliottYan/LUFFY [5]
https://github.com/JIA-Lab-research/Scaf-GRPO

D.1. 预备知识：伯努利组与样本方差

固定一个上下文 $(x,\ h)$，并抽取 $G \geq 2$ 个独立同分布的推演，其终端奖励为 $R_i \in \{0,1\}$。令

$$\bar{R} \coloneqq \frac{1}{G}\sum_{i=1}^{G} R_i, \qquad s^2 \coloneqq \frac{1}{G}\sum_{i=1}^{G}(R_i-\bar{R})^2, \qquad s \coloneqq \sqrt{s^2}.$$

使用标准化 GRPO 时，优势值为

$$A_i \coloneqq \frac{R_i-\bar{R}}{s+\epsilon},$$

其中 $\epsilon > 0$ 是一个数值稳定项（与正文中的用法相同）。

我们将反复用到这样一个事实：由于 $R_i \in \{0,1\}$,

$$\sum_{i=1}^{G}(R_i-\bar{R})^2 = \sum_{i=1}^{G} R_i^2 - G\bar{R}^2 = \sum_{i=1}^{G} R_i - G\bar{R}^2 = G\bar{R} - G\bar{R}^2 = G\bar{R}(1-\bar{R}), \tag{14}$$

因此

$$s^2 = \bar{R}(1-\bar{R}). \tag{15}$$

特别地，$s^2 = 0$ 当且仅当 $\bar{R} \in \{0,1\}$，也就是说，当且仅当所有 R_i 都相同。

D.2. 推论3.1的证明

证明。 回顾优势能量

$$E \coloneqq \frac{1}{G}\sum_{i=1}^{G} A_i^2, \qquad A_i = \frac{R_i-\bar{R}}{s+\epsilon}.$$

对于 $\epsilon > 0$，我们计算

$$E = \frac{1}{G}\sum_{i=1}^{G}\frac{(R_i-\bar{R})^2}{(s+\epsilon)^2} = \frac{1}{(s+\epsilon)^2}\cdot\frac{1}{G}\sum_{i=1}^{G}(R_i-\bar{R})^2 = \frac{s^2}{(s+\epsilon)^2},$$

这正是式（6）。

由于 $s \geq 0$ 和 $\epsilon > 0$，我们有 $0 \leq \frac{s}{s+\epsilon} < 1$，因此

$$0 \leq E = \left(\frac{s}{s+\epsilon}\right)^2 < 1,$$

因此 $E \in [0,1)$（在极限情况下为 $\epsilon \to 0^+, E \to \mathbb{I}[s>0]$）。

s 中的单调性可通过求导得出：定义 $f(s) \coloneqq \frac{s^2}{(s+\epsilon)^2}$，其中 $s \geq 0$。则

$$f'(s) = \frac{2s(s+\epsilon)^2 - s^2\cdot 2(s+\epsilon)}{(s+\epsilon)^4} = \frac{2s\epsilon}{(s+\epsilon)^3} \geq 0,$$

因此 E 关于 s是非递减的。

最后，如果该组是退化的，那么 $s=0$，因此 $E=0$。这表明，标准化信号能量恰好在组内方差趋于零时也趋于零。

□

D.3. 命题 3.2的证明

证明。 记 $p \coloneqq p_\theta(x,h) = \Pr[R(x,\tau) = 1 \mid x,h]$。则 $R_1,\ldots,R_G$ 为独立同分布的伯努利分布(p)。

如第 D.1节所述， $s = 0$当且仅当所有奖励均相同。恰好存在两种退化情形：(i) 全零：$R_1 = \cdots = R_G = 0$；(ii) 全一：$R_1 = \cdots = R_G = 1$。因此

$$\Pr[s > 0 \mid x, h] = 1 - \Pr[\text{all-zero}] - \Pr[\text{all-one}] = 1 - (1-p)^G - p^G,$$

这证明了式（7）。

为确定最大化点，定义 $u(p) := 1 - (1-p)^G - p^G$ 在 $[0,1]$上。我们有对称性 $u(p) = u(1-p)$。此外，对于$G \geq 2$,

$$u''(p) = -G(G-1)\Big(p^{G-2} + (1-p)^{G-2}\Big) < 0,$$

因此 u 严格凹，故存在唯一最大化点。根据对称性，该唯一最大化点必定为 $p = \frac{1}{2}$。最后 -

, 在稀疏区间 $p \ll 1$,

$$u(p) = 1 - (1-p)^G - p^G = 1 - \Big(1 - Gp + O(p^2)\Big) - O(p^G) = Gp + O(p^2),$$

因此 $\Pr[s > 0 \mid x, h] \approx Gp$。 □

D.4. 命题 3.3 的证明

证明。 固定 x 和 $G \geq 2$。回顾 $u(p) = 1 -(1-p)^G - p^G$ 以及

$$J_x(\theta, q) = \mathbb{E}_{h \sim q(\cdot|x)}\Big[u\big(p_\theta(x,h)\big)\Big].$$

我们首先验证关于 u 的断言。对称性： $u(1-p) = 1 - p^G - (1-p)^G = u(p)$。对于严格凹性，对$G \geq 2$计算：

$$u''(p) = -G(G-1)\Big(p^{G-2} + (1-p)^{G-2}\Big) < 0,$$

so u 在 $[0,1]$ 上严格凹。根据对称性和严格凹性，唯一最大化点为 $p = \frac{1}{2}$ -.

对于固定的 θ, 定义可测函数

$$v_\theta(h) := u(p_\theta(x,h)) \in [0,1].$$

于是 $J_x(\theta, q) = \mathbb{E}_{h \sim q}[v_\theta(h)]$. 在所有由可容许提示空间支持的概率分布 $q(\cdot \mid x)$ 中，最大化点必须将概率质量集中在 v_θ 的（本质）最大化点上：事实上，由于目标函数关于 q 是线性的，因此可以选择一个将全部质量置于

$$\arg\max_h v_\theta(h) = \arg\max_h u(p_\theta(x,h)).$$

由于 u 在 $p = \frac{1}{2}$处取得唯一最大值，并且随着 p 偏离1 2 而严格递减（这是严格凹性和对称性的结果）， $v_\theta(h)$ 的最大化点恰好是校准提示 ，它们使 $p_\theta(x, h)$ 尽可能接近1 2 （在可实现时等于1 2 ）。这证明了最优 $q^\star_\theta(\cdot \mid x)$ 将其质量置于校准提示上的结论。

一般而言， $p_\theta(x, h)$ 会随 θ 变化，因为在 $\pi_\theta(\cdot \mid x, h)$ 下的展开分布会随 θ 变化。因此，校准提示的集合

$$\mathcal{H}^\star_\theta(x) := \arg\max_h u(p_\theta(x,h))$$

通常会随 θ变化。除非 $p_\theta(x, h)$（因此 $v_\theta(h)$）在 θ 上对 q-几乎所有 h都保持不变，否则，在训练早期最优（或近似最优）的固定分布 q 之后会逐渐偏离最优状态，使得相较于适应 θ的选择， $J_x(\theta, q)$降低。这从形式上说明了为何使用策略在线更新提示生成器能够减少门控失配。 □

D.5. 备注1中詹森不等式的证明

证明。 固定 $G \geq 2$ 并定义 $u(p) = 1-(1-p)^G - p^G$。上文已证明，由于 $u''(p) \leq 0$， u 在 $[0,1]$ 上是凹函数。令 $Z := p_\theta(x,\ h) \in [0,1]$ 为通过采样 $h \sim q$所诱导的随机成功概率。于是，对凹函数 u 应用詹森不等式可得

$$\mathbb{E}_h\big[u(Z)\big] \leq u(\mathbb{E}_h[Z])\,,$$

这恰好是式（11）。等式仅当 Z 几乎必然为常数时成立（或当 u 在支撑集上为仿射时；除退化情形外，这对 $G \geq 2$ 不会发生）。这表明，在平均成功概率固定时，提示间的额外变异性只会降低期望开门频率。 □

D.6. 门概率更精细的小p展开

为完整起见，我们还记录一个使 Gp 标度显式化的精确分解。令 $p = \Pr[R=1 \mid x,h]$。则

$$\begin{aligned}\Pr[s>0 \mid x,h] &= 1-(1-p)^G - p^G \\ &= \sum_{k=1}^{G-1} \binom{G}{k} p^k (1-p)^{G-k}.\end{aligned} \tag{16}$$

当 $p \ll 1$时，主导项为 $k=1$:

$$\Pr[s>0 \mid x,h] = Gp(1-p)^{G-1} + O(p^2) \approx Gp,$$

而被忽略的 p^G 项在 G中呈指数级小。

D.7. 非标准化 GRPO 信号能量

非标准化优势值也广泛应用于类GRPO算法（Liu等人，2025年），这也为理解其行为提供了参考。

设定。定义非标准化（中心化）优势值

$$\tilde{A}_i := R_i - \bar{R}, \qquad \bar{R} = \frac{1}{G}\sum_{i=1}^{G} R_i,$$

并定义（非标准化）能量

$$\tilde{E} := \frac{1}{G}\sum_{i=1}^{G} \tilde{A}_i^2.$$

注意，根据定义， $\tilde{E} = s^2$ 。

命题 D.1 (期望 的伯努利奖励下的非标准化能量)。 在给定 $(x,\ h)$ 且成功概率为 $p = p_\theta(x,h)$,

$$\mathbb{E}\big[\tilde{E} \mid x,h\big] = \frac{G-1}{G}\,p(1-p). \tag{17}$$

证明。 令 $S = \sum_{i=1}^{G} R_i$ 使得 $\bar{R} = S/G$。利用恒等式 (14),

$$\tilde{E} = \frac{1}{G}\sum_{i=1}^{G}(R_i - \bar{R})^2 = \bar{R}(1-\bar{R}).$$

取期望:

$$\mathbb{E}[\tilde{E}] = \mathbb{E}[\bar{R}] - \mathbb{E}[\bar{R}^2].$$

¯我们有 $\mathbb{E}[R] = p$。此外,

$$\mathbb{E}[\bar{R}^2] = \mathrm{Var}(\bar{R}) + (\mathbb{E}[\bar{R}])^2 = \frac{1}{G^2}\mathrm{Var}(S) + p^2 = \frac{1}{G^2}\cdot Gp(1-p) + p^2 = \frac{p(1-p)}{G} + p^2.$$

因此

$$\mathbb{E}[\tilde{E}] = p - \left(\frac{p(1-p)}{G} + p^2\right) = \frac{G-1}{G}\,p(1-p).$$

□

命题 D.2 (均值中心化更新的最优校准难度**)。** 对于固定的 $G \geq 2$，方程(17)右侧在 $p = \frac{1}{2}$ 处取得唯一最大值。

证明。 令 $f(p) = p(1-p) = p - p^2$。则 $f'(p) = 1 - 2p$ 且 $f''(p) = -2 < 0$，因此 f是严格凹函数，并在 $p = \frac{1}{2}$ 处取得唯一最大值。前置因子 $(G-1)/G$ 不会影响最大化点。 □

备注 D.3（固定均值下，提示随机性会带来方差惩罚）。 令 $Z = p_\theta(x,\ h)$ 因 $h \sim q$ 和$\bar{p} = \mathbb{E}[Z]$而成为随机变量。那么

$$\mathbb{E}\big[Z(1-Z)\big] = \bar{p}(1-\bar{p}) - \mathrm{Var}(Z).$$

因此，在平均成功率固定的情况下，提示之间 $p_\theta(x,\ h)$ 的额外变异性会降低期望能量。

证明。计算资源

$$\mathbb{E}[Z(1-Z)] = \mathbb{E}[Z] - \mathbb{E}[Z^2] = \bar{p} - \big(\mathrm{Var}(Z) + \bar{p}^2\big) = \bar{p}(1-\bar{p}) - \mathrm{Var}(Z).$$

□