强化学习实验报告

Mengqi Liao

2025 年 2 月 2 日

1 实验内容

1.1 概述

经过微调后，我们获得了一个能够解决 24 点游戏的大模型。然而，该模型的推理效率较低，往往会进行大量无效的探索。即使在微调时将推理长度控制在不超过 1024 个 token，并在生成阶段允许生成长度达到 4096 个 token，仍有约 20%的测试集实例无法在规定的生成长度内完成答案的推理。

近期，DeepSeek-R1[1] 的卓越表现证明了强化学习在提升大模型推理能力方面的巨大潜力。为了解决推理效率问题，我们尝试实现 DeepSeek-R1 的强化学习流程，以进一步增强大模型的推理能力。

1.2 GRPO 与改进

大模型通常通过人类反馈强化学习（HFRL）来对齐人类偏好，常用的强化学习算法包括 PPO 等。然而，PPO 在进行在线学习时需要加载演员（actor）模型、参考模型、奖励模型以及评论家（critic）模型四个模型，这使得训练开销较大且过程较为复杂。相比之下，DeepSeek-R1 使用了 GRPO 算法[1] 进行强化学习，只需加载演员模型和参考模型，从而显著降低了强化学习的复杂性。GRPO 的优化目标是最大化以下公式：

[1] https://github.com/deepseek-ai/DeepSeek-R1

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}\hat{A}_{i,t}, \right.\right.$$
$$\left. \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, 1-\epsilon, 1+\epsilon \right)\hat{A}_{i,t} \right)$$
$$\left. - \beta D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}] \right]. \tag{1}$$

该公式与 PPO 的优化目标非常相似，但最大的区别在于优势值 $\hat{A}i,t$ 的计算方式。GRPO 不依赖 critic 模型来计算优势值，而是通过组内采样输出的归一化平均奖励作为基线，计算优势值为 $\hat{A}_{i,t} = \tilde{r}i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$，其中 $\mathbf{r} = \{r_1, \ldots, r_G\}$ 表示每组采样结果 $\{o_i\}_{i=1}^{G}$ 对应的奖励。公式中的 $D_{\text{KL}}[\pi_\theta|\pi_{\text{ref}}]$ 是 KL 散度约束项，用于限制模型的优化方向，防止其偏离参考模型过远。

在标准的 GRPO 中，优势值的计算是通过对组内奖励进行归一化处理，然而，这种方法仅依赖当前组数据的均值和标准差，可能导致更新过程对单次采样结果过于敏感，影响模型的稳定性。我们采用滑动均值和标准差来平滑奖励归一化的过程。具体来说当前奖励的均值和标准差不直接用于计算，而是通过动量更新历史均值和标准差：

$$\text{mean} = \text{mean} \times \text{momentum} + \text{mean}(\mathbf{r}) \times (1 - \text{momentum}) \tag{2}$$

$$\text{std} = \text{std} \times \text{momentum} + \text{std}(\mathbf{r}) \times (1 - \text{momentum}) \tag{3}$$

$$\hat{A}_{i,t} = \tilde{r}i = \frac{r_i - \text{mean}}{\text{std}} \tag{4}$$

这种方法使得归一化的均值和标准差在不同批次之间具有一定的连续性，减小了单次采样中异常值对归一化过程的影响。此外，动量值（momentum）从一个较低的初始值开始，每次更新后动量值都会加增加，但不会超过一个设定的最大值。动态调整动量的方式可以在训练早期更快地适应奖励分布的变化，而在训练后期通过较大的动量保持稳定。

1.3 基于规则的奖励模型

在 DeepSeek-R1 的强化学习 pipeline 中，另一个重要的变化是放弃使用训练的奖励模型，转而采用基于规则的奖励模型。这是一个关键设计选

择，旨在解决训练奖励模型时可能出现的 reward hacking 问题。Reward Hacking 是指在强化学习中，智能体找到了一种利用奖励函数漏洞的方式，以最大化奖励，但其行为不符合预期的目标。在这种情况下，训练的奖励模型可能会导致智能体行为偏离预期目标，甚至学习出完全无意义的策略。

我们的基于规则的奖励模型的奖励分为三部分，分别为输出答案奖励 r_{oc}，过程正确性奖励 r_{pr} 和推理长度奖励 r_{l}。当模型输出最终表达式时将获得输出答案奖励，如果表达式正确则 $r_{\text{oc}} = 1$ ，否则为 0.35 （用于鼓励模型输出最终答案）。$r_{\text{pr}} \in [0, 1]$ 则是验证所有推理步骤的正确性得到的正确率。长度奖励用于优化推理路径的搜索效率。该奖励函数定义如下：

$$r_{\text{l}}(l) = \begin{cases} 1 & \text{当 } l \leq t \\ \dfrac{1}{1 + e^{k(l-t-m)}} & \text{当 } l > t \end{cases} \tag{5}$$

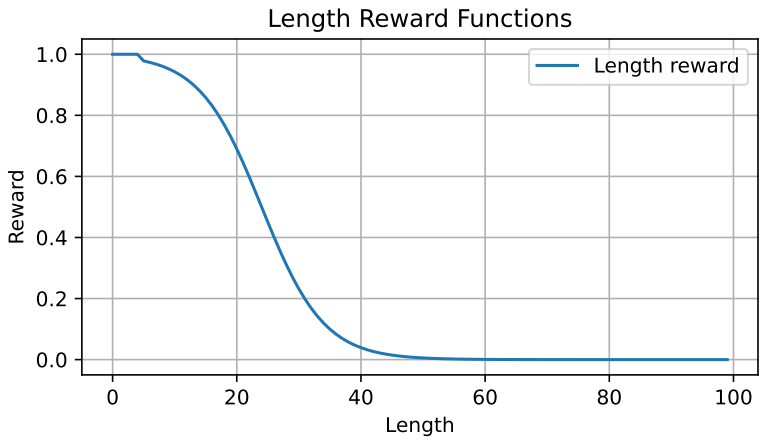

图 1: 长度奖励值和推理步骤长度的关系.

其中 $t = 4$ 为触发奖励衰减的步长（每一行为一步）阈值，$m = 20$ 控制Sigmoid曲线中点的偏移量，$k = 0.2$ 调节衰减速率的陡峭系数。该函数具有双重调节特性：(1) 对于有效步长（$l \leq 4$）给予全额奖励；(2) 当步长超过阈值时，通过平移后的Sigmoid函数（中点在$t + m = 24$处）实现渐进式惩罚。这种设计既能鼓励模型在早期寻找有效路径，又避免对合理长度的中间推理过程施加过早的强惩罚。最终奖励使用公式 $r = \lambda_{\text{oc}} \times r_{\text{oc}} + \lambda_{\text{pr}} \times r_{\text{pr}} + \lambda_{\text{l}} \times r_{\text{l}}$ 来计算。

2 实验

2.1 实验设置

我们以使用 Format v3 混合数据微调的模型作为基座模型开始强化学习。训练过程中，学习率固定为 2e-6，β 设置为 0.01，最大 momentum 设定为 0.2。batch size 为 12，组大小 G 设置为 8。奖励模型的系数 λ_{oc}、λ_{pr} 和 λ_{l} 分别设置为 0.55、0.3 和 0.15。在采样过程中，我们将温度设置为 0.7，并采用核采样，top-p 参数为 0.9。训练采样和评估时，最大生成长度均设置为 1024。

2.2 实验结果

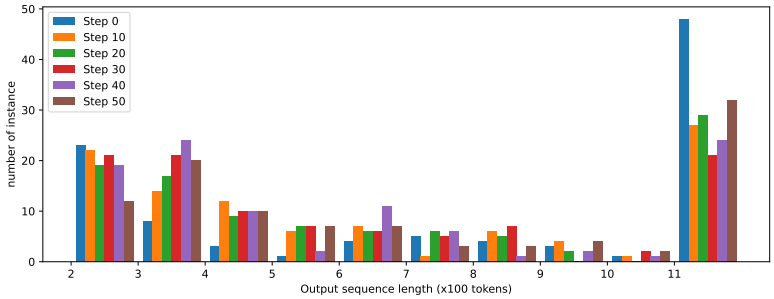

图 2: 训练不同步数的模型输出序列的 token 长度的分布

图 2 展示了模型在不同训练步数（Step 0 至 Step 50）时输出序列的 token 长度分布。从图中可以看出，经过强化学习后，生成长度在 300 至 600 的样本显著增加，尤其是长度在 300 至 400 的样本随着训练步数的增加呈现出明显的增长趋势。同时，输出长度超过窗口大小的样本（图的最右侧）显著减少。这表明强化学习有效地提升了模型的推理效率，减少了无效的冗长生成。

图 3 展示了随着训练的进行，模型的准确率、未完成率和错误率的变化趋势。从图中可以看出，准确率和完成率在训练的前 30 步表现出持续提升，并在约 30 步时达到最佳状态，其中准确率接近 80%。然而，随着训练的进一步进行，准确率开始逐渐下降，同时未完成率和错误率均出现了反弹。这一现象可能表明，随着训练步数的增加，模型过度拟合奖励信号，

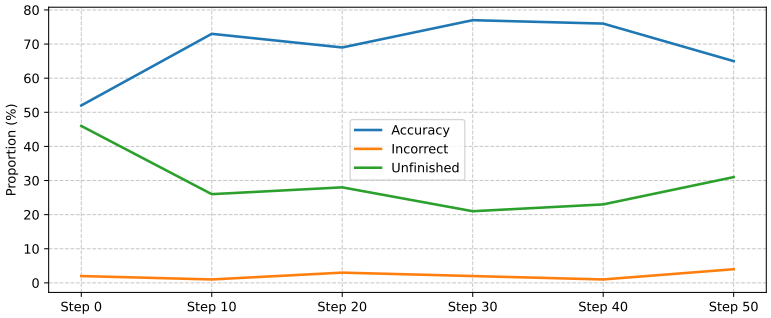

图 3: 随着训练的进行精确率、未完成率和错误率的变化

从而导致生成质量下降。更细粒度的奖励，比如生序列中的不同 token 给予不同的奖励可能会取得更好的效果。

参考文献

[1] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.