DeepSeek-R1：通过强化学习激发LLM的推理能力

DeepSeek AI

research@deepseek.com

摘要

我们介绍了我们的第一代推理模型DeepSeek-R1-Zero和DeepSeek-R1。DeepSeek-R1-Zero是一个通过大规模强化学习（RL）训练的模型，没有作为初步步骤的监督微调（SFT），表现出卓越的推理能力。通过RL,DeepSeek-R1-Zero自然会出现许多强大而有趣的推理行为。然而，它遇到了可读性差和语言混合等挑战。为了解决这些问题并进一步提高推理性能，我们引入了DeepSeek-R1，它在RL之前结合了多阶段训练和冷启动数据。DeepSeekR1在推理任务上的性能与OpenAI-ol-1217相当。为了支持研究社区，我们开源了DeepSeek-R1-Zero、DeepSeek-R1，以及基于Qwen和Llama从DeepSeek-R2中提取的六个密集模型（1.5B、7B、8B、14B、32B、70B）。

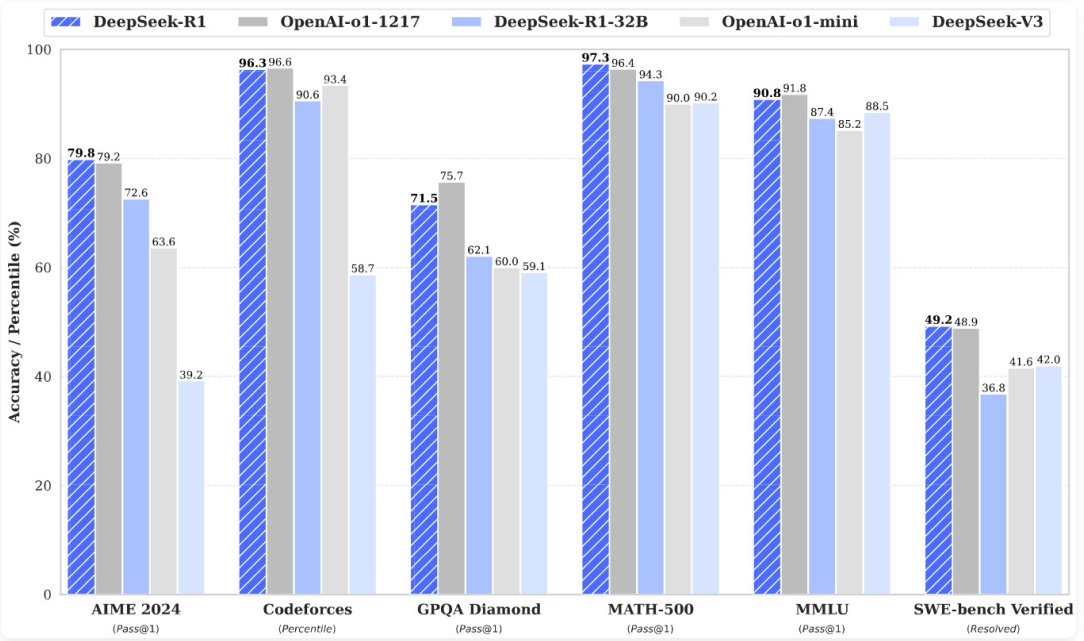

图1|DeepSeek-R1的基准性能。

内容

2.方法

2.1.概述

之前的工作在很大程度上依赖于大量的监督数据来提高模型性能。在这项研究中，我们证明了即使不使用监督微调（SFT）作为冷启动，通过大规模强化学习（RL）也可以显著提高推理能力。此外，通过包含少量冷启动数据，可以进一步提高性能。在以下部分中，我们将介绍：（1）DeepSeek-R1-Zero，它直接将RL应用于基础模型，而不需要任何SFT数据，以及（2）DeepSeek-R1，它从一个检查点开始应用RL，该检查点经过数千个长思维链（CoT）示例的微调。3）将DeepSeek-R1的推理能力提取到小型密集模型中。

2.2.DeepSeek-R1-Zero：基于基础模型的强化学习

强化学习在推理任务中表现出了显著的有效性，正如我们之前的工作所证明的那样（Shao等人，2024；Wang等人，2023）。然而，这些工作在很大程度上依赖于监督数据，而这些数据的收集需要耗费大量时间。在本节中，我们将探讨LLM在没有任何监督数据的情况下发展推理能力的潜力，重点关注它们通过纯强化学习过程的自我进化。我们首先简要概述了我们的RL算法，然后介绍了一些令人兴奋的结果，并希望这能为社区提供有价值的见解。

2.2.1.强化学习算法

集团相关策略优化为了节省培训成本RL·，我们采用了组相对策略优化（GRPO）（Shao等人，2024），它放弃了通常与策略模型大小相同的批评模型，而是根据组分数估计基线。具体来说，对于每个问题q，GRPO对一组输出进行采样$o1, o_2, \cdots, o_G$从旧政策$\pi_{\theta_{old}}$然后优化策略模型$_{\pi_\theta}$通过最大化以下目标：

$$\mathcal{I}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{\sigma_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta\left(\sigma_i|q\right)}{\pi_{\theta_{old}}\left(\sigma_i|q\right)}A_i, \text{clip}\left(\frac{\pi_\theta\left(\sigma_i|q\right)}{\pi_{\theta_{old}}\left(\sigma_i|q\right)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right),$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1,$$

哪里ε和β是超参数，以及A_i是使用一组奖励计算的优势吗$r_1, r_2, \ldots, r_G$对应于每个组内的输出：

$$A_i = \frac{r_i - \mathrm{m}ean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}.$$