Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

2024大模型落地实例

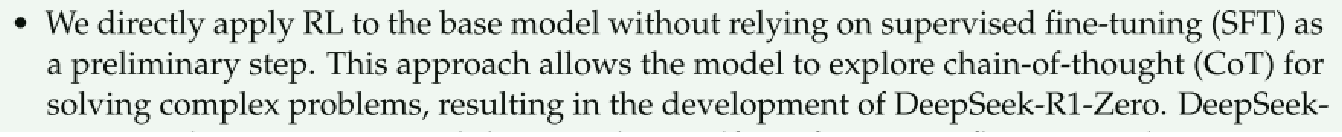
(点击题目可下载)

基于v3大模型的强化学习自然而然产生的r1模型的推理能力

1. 本文贡献了一种后训练方法,诞生了两个成果

Post-Training: Large-Scale Reinforcement Learning on the Base Model

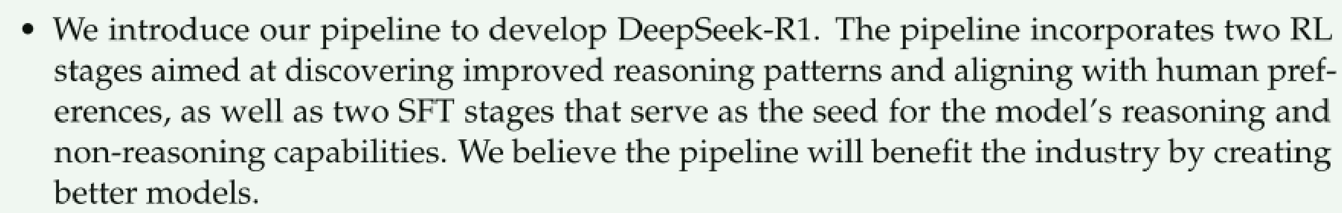
在V3基础上进行的大规模强化学习



- We directly apply RL to the base model without relying on supervised fine-tuning (SFT) as a preliminary step. This approach allows the model to explore chain-of-thought (CoT) for solving complex problems, resulting in the development of DeepSeek-R1-Zero. DeepSeek-

1. 在v3大模型基础上直接采用强化学习，而没有用SFT监督微调（没有人工标注监督），但是诞生了可以解决复杂问题的思维链，r1-zero模型展现了很多自我验证，反射，以及产生长思维链

   重点是，大模型推理能力，可以纯粹用rl来激励，不需要sft的协助

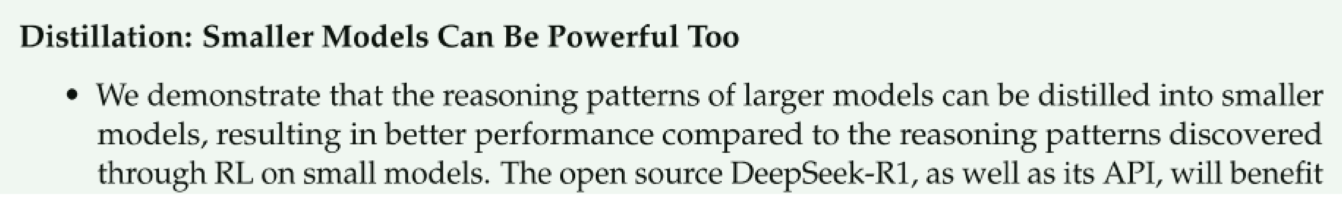


- We introduce our pipeline to develop DeepSeek-R1. The pipeline incorporates two RL stages aimed at discovering improved reasoning patterns and aligning with human preferences, as well as two SFT stages that serve as the seed for the model's reasoning and non-reasoning capabilities. We believe the pipeline will benefit the industry by creating better models.

如何去训练r1模型，后面介绍一种pipeline

2. 小模型可以通过模型蒸馏，蒸馏r1的内容，Qwen-7b llama3的推理能力可以超越参数比他们大的模型,变得更强

除此以外,泛化到普通任务上,有很强的性能,例如写作,问题回答上,在较为复杂的问题理解能力性能大大超过了v3



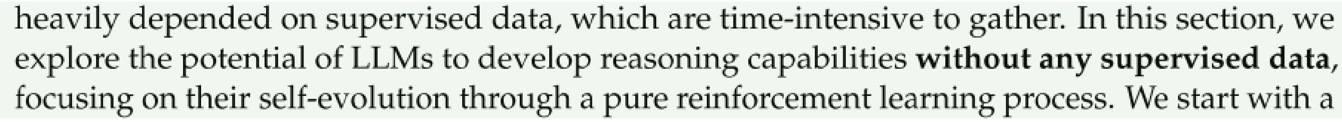
Distillation: Smaller Models Can Be Powerful Too

- We demonstrate that the reasoning patterns of larger models can be distilled into smaller models, resulting in better performance compared to the reasoning patterns discovered through RL on small models. The open source DeepSeek-R1, as well as its API, will benefit

2. 实现方法

2.1之前的工作：

用了大量有标注的数据，或者监督的数据sft，来训练推理能力,而r1只用了强化学习rl就得到了推理能力

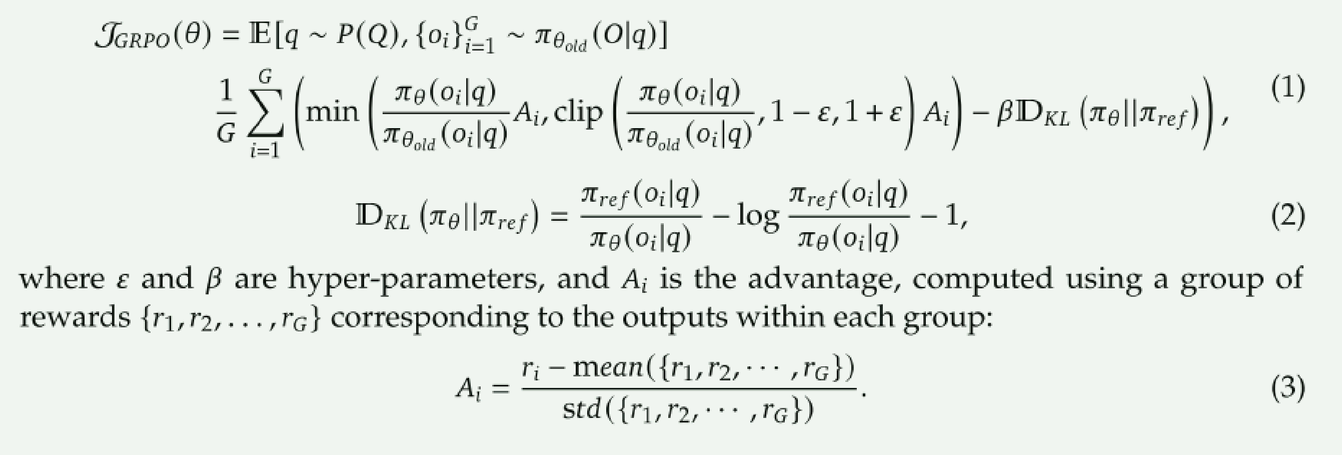


heavily depended on supervised data, which are time-intensive to gather. In this section, we explore the potential of LLMs to develop reasoning capabilities **without any supervised data**, focusing on their self-evolution through a pure reinforcement learning process. We start with a

2.2 v3作为基座模型

v3质量很好，作为基座模型训练很良好,在此基础上用纯粹的强化学习来进行训练

2.2.1 PRO RL算法



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

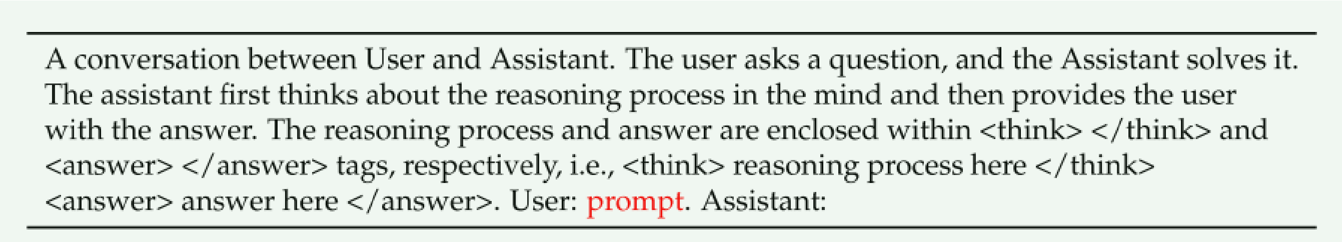
$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \tag{1}$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{2}$$

where ε and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \tag{3}$$

GRPO，主要算法，不同于ppo算法，不需要价值函数模型actor- critic模型或者value model减少了训练硬件要求，只用一个就可以，采取与平均值之间的距离的方式来判断



A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

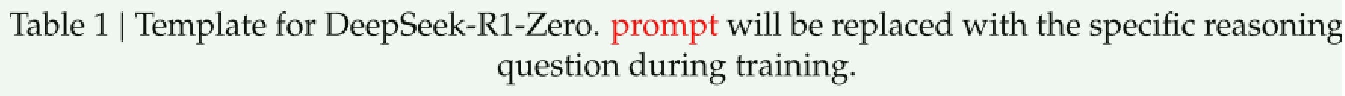
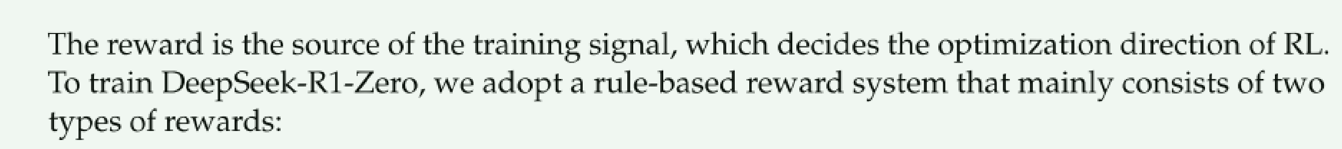


Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

给出一个用户与语言助手对话的样本模板,prompt是参数,一个具体问题,后面回答是需要微调的部分

2.2.2 奖励模型

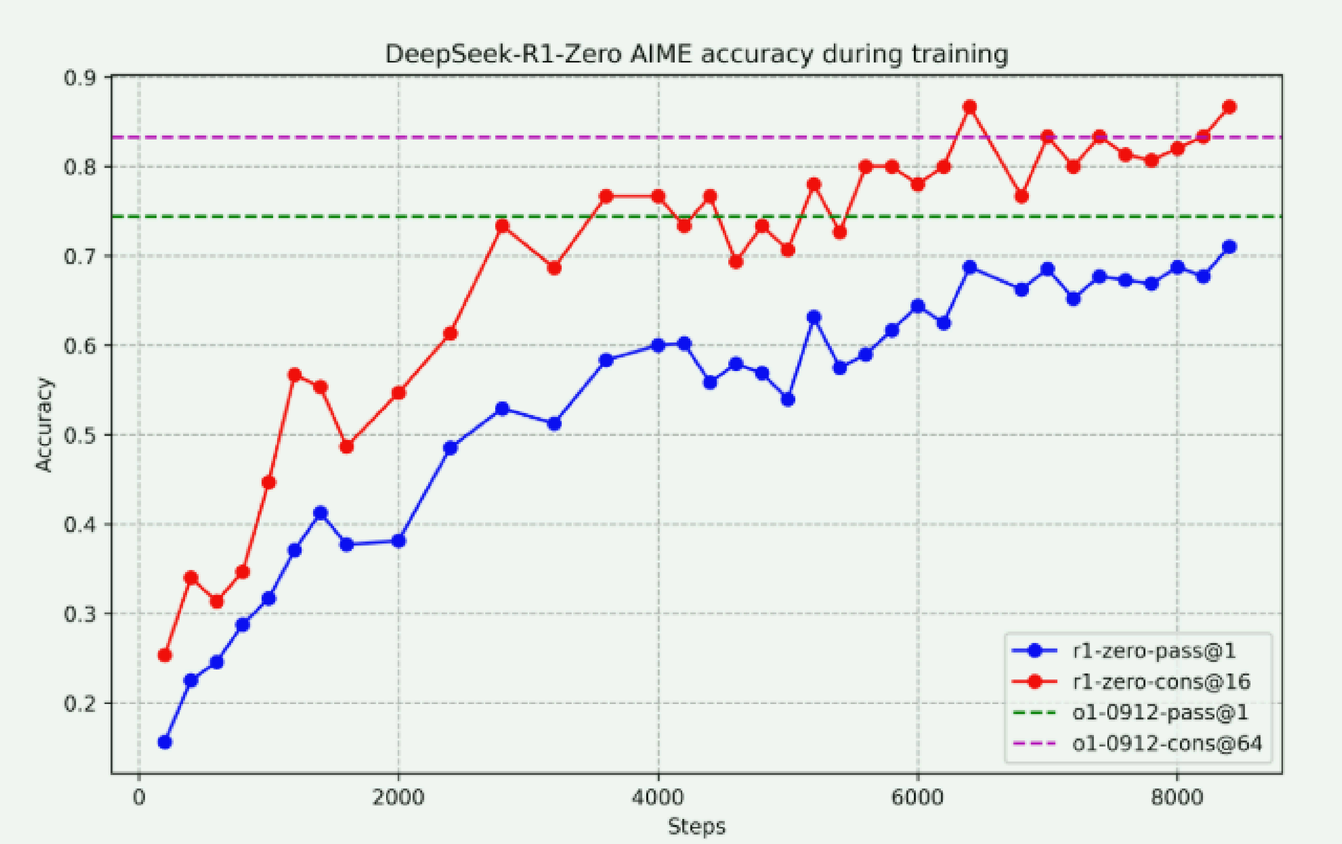


The reward is the source of the training signal, which decides the optimization direction of RL. To train DeepSeek-R1-Zero, we adopt a rule-based reward system that mainly consists of two types of rewards:

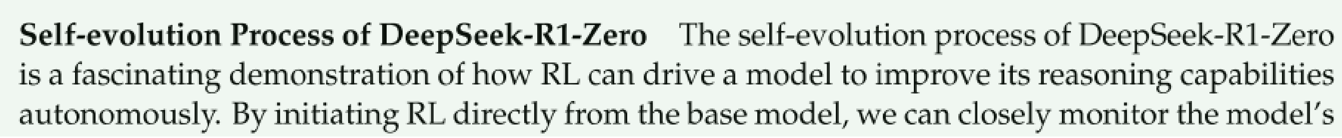
奖励模型来决定优化方向,包括两个部分,验证准确度奖励和格式奖励

2.2.3 zero的表现情况,自我进化过程,ahamemonet

r1训练随着次数,他的连贯性和通过率渐渐超过chatgpt-o1

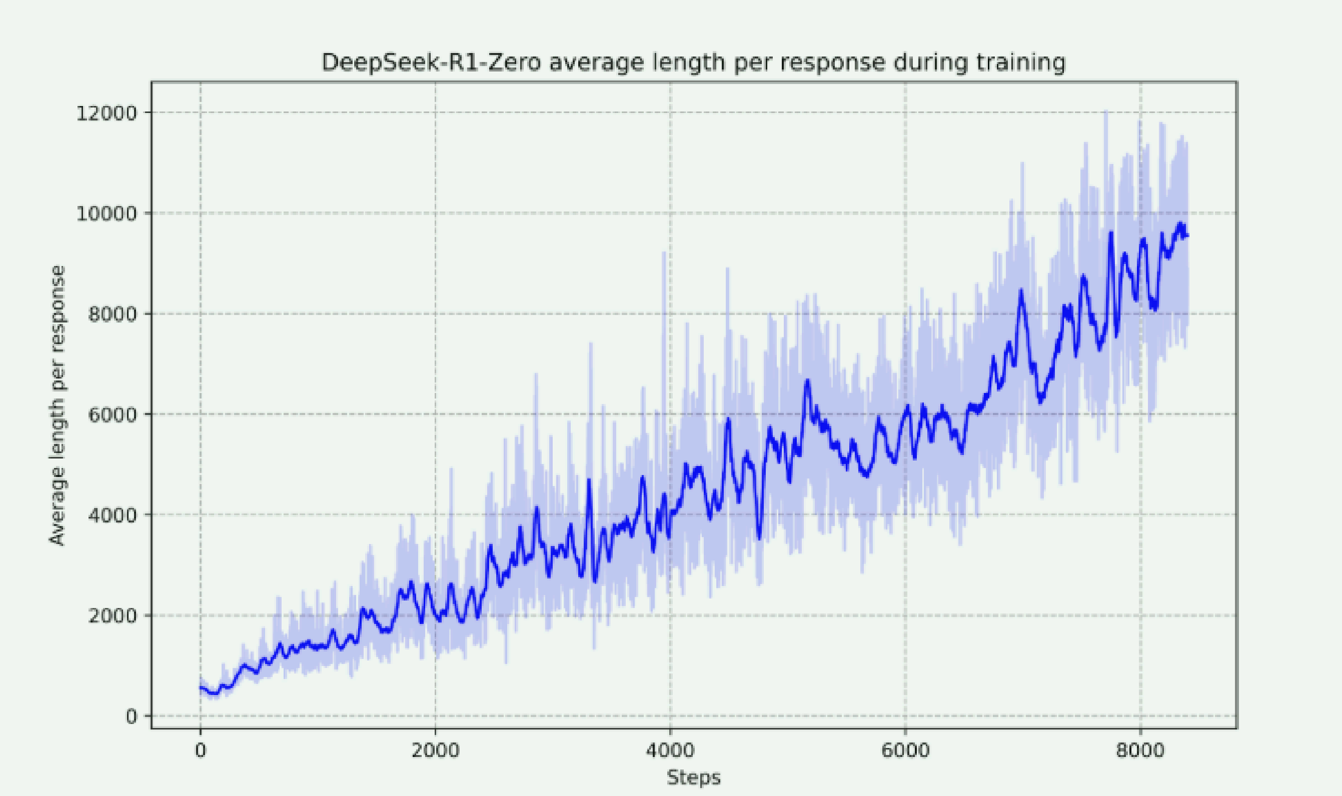


随着rl强化学习,"自然而然的出现了一种推理能力"

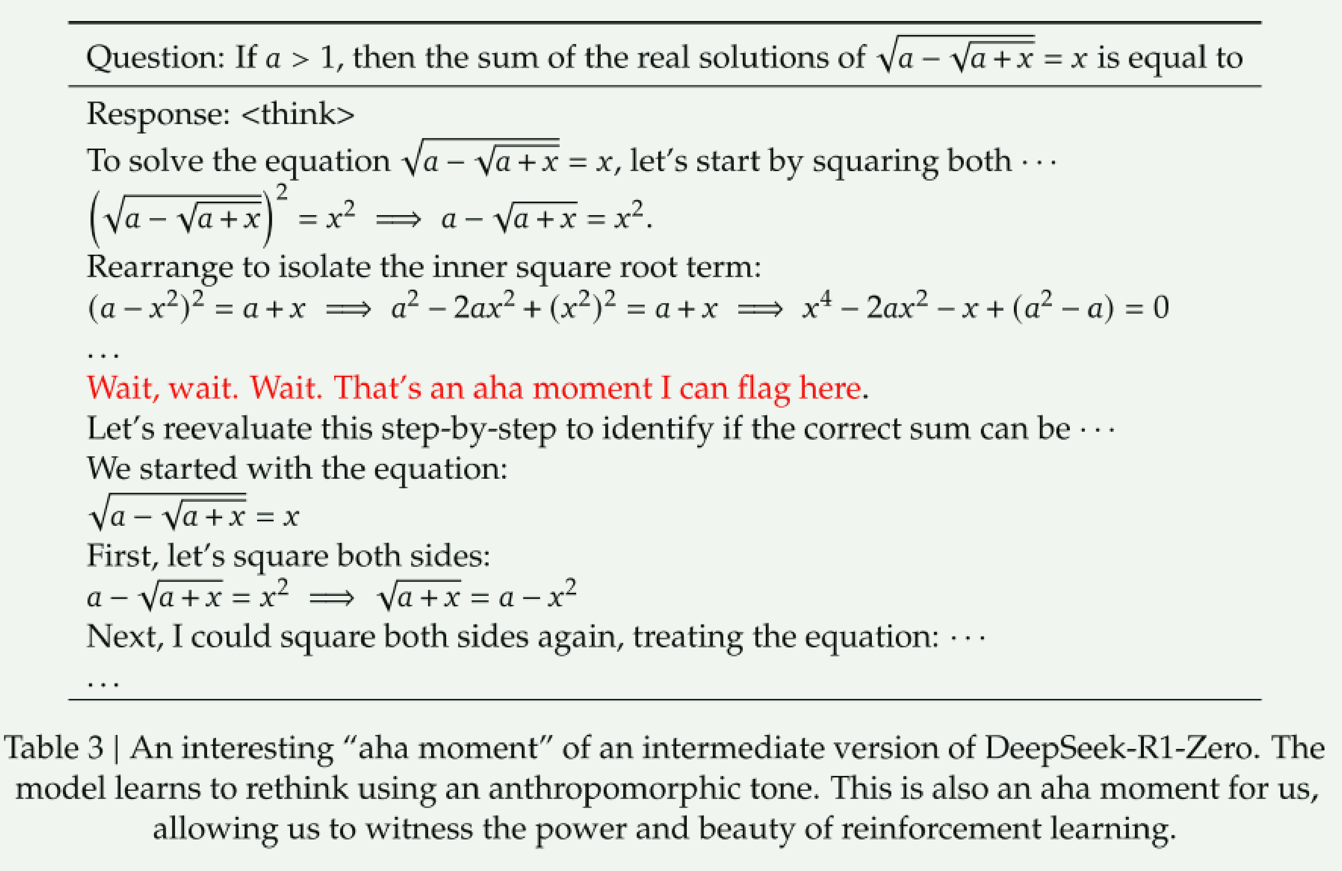


Self-evolution Process of DeepSeek-R1-Zero The self-evolution process of DeepSeek-R1-Zero is a fascinating demonstration of how RL can drive a model to improve its reasoning capabilities autonomously. By initiating RL directly from the base model, we can closely monitor the model's

推理链,随着推理次数越来越长,出现了很多自我涌现的行为,包括反思,或者重新再评估



One of the most remarkable aspects of this self-evolution is the emergence of sophisticated behaviors as the test-time computation increases. Behaviors such as reflection—where the model revisits and reevaluates its previous steps—and the exploration of alternative approaches to problem-solving arise spontaneously. These behaviors are not explicitly programmed but instead



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$

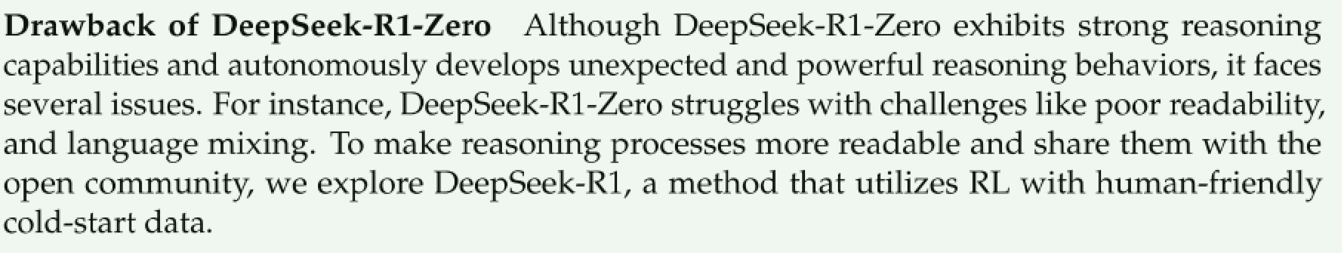
Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

出现了aha-moment

为了解决这个数学问题,在思考过程中,ai出现了突变的思考,单纯的运行了很多次,一直反馈,后出现了这句话

本质上是概率训练,有几率输出了wait,后面接着wait,然后出现了这些,但是为什么会出现这种话语?训练语料中的思维逻辑关系?

可以泛化到其他领域吗



Drawback of DeepSeek-R1-Zero Although DeepSeek-R1-Zero exhibits strong reasoning capabilities and autonomously develops unexpected and powerful reasoning behaviors, it faces several issues. For instance, DeepSeek-R1-Zero struggles with challenges like poor readability, and language mixing. To make reasoning processes more readable and share them with the open community, we explore DeepSeek-R1, a method that utilizes RL with human-friendly cold-start data.

zero缺点

zero语序胡乱,难以让人理解,较为分散,可能中文英文夹杂

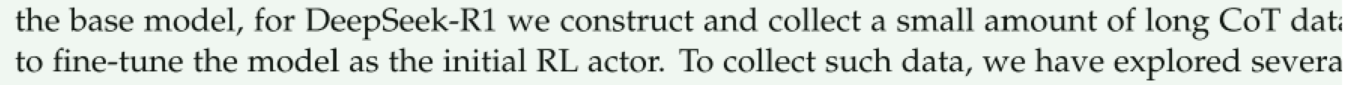
2.3 r1模型的冷启动

利用少量高质量数据,作为冷启动可以提高推理性能和加速收敛?

如何训练一个推理能力强,且泛化通用能力强的r1?

设计了一个4步骤pipeline

2.3.1 冷启动

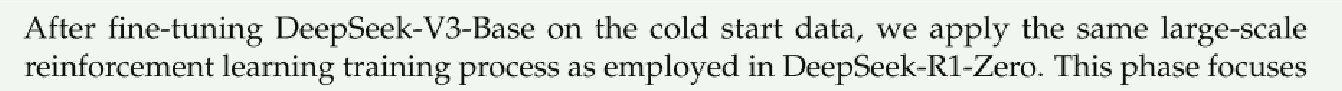


the base model, for DeepSeek-R1 we construct and collect a small amount of long CoT dat
to fine-tune the model as the initial RL actor. To collect such data, we have explored severa

r1的训练过程前期,用了少量的cot链,作为微调模型,甚至包括直接从zero模型中获得这些样本(数千条)

样本应该可读性提高,发现迭代训练是推理模型的更好的方法

2.3.2 针对推理的强化学习

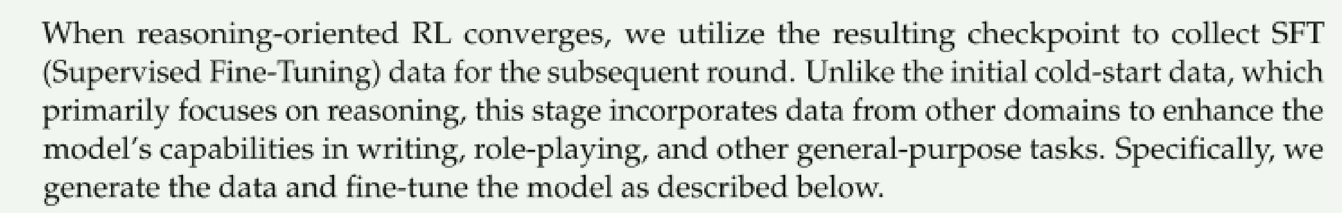


After fine-tuning DeepSeek-V3-Base on the cold start data, we apply the same large-scale reinforcement learning training process as employed in DeepSeek-R1-Zero. This phase focuses

在微调后,用与之前训练zero时一样的rl训练,来训练r1

2.3.3 sft 监督微调和拒绝抽样

感觉模型快收敛时



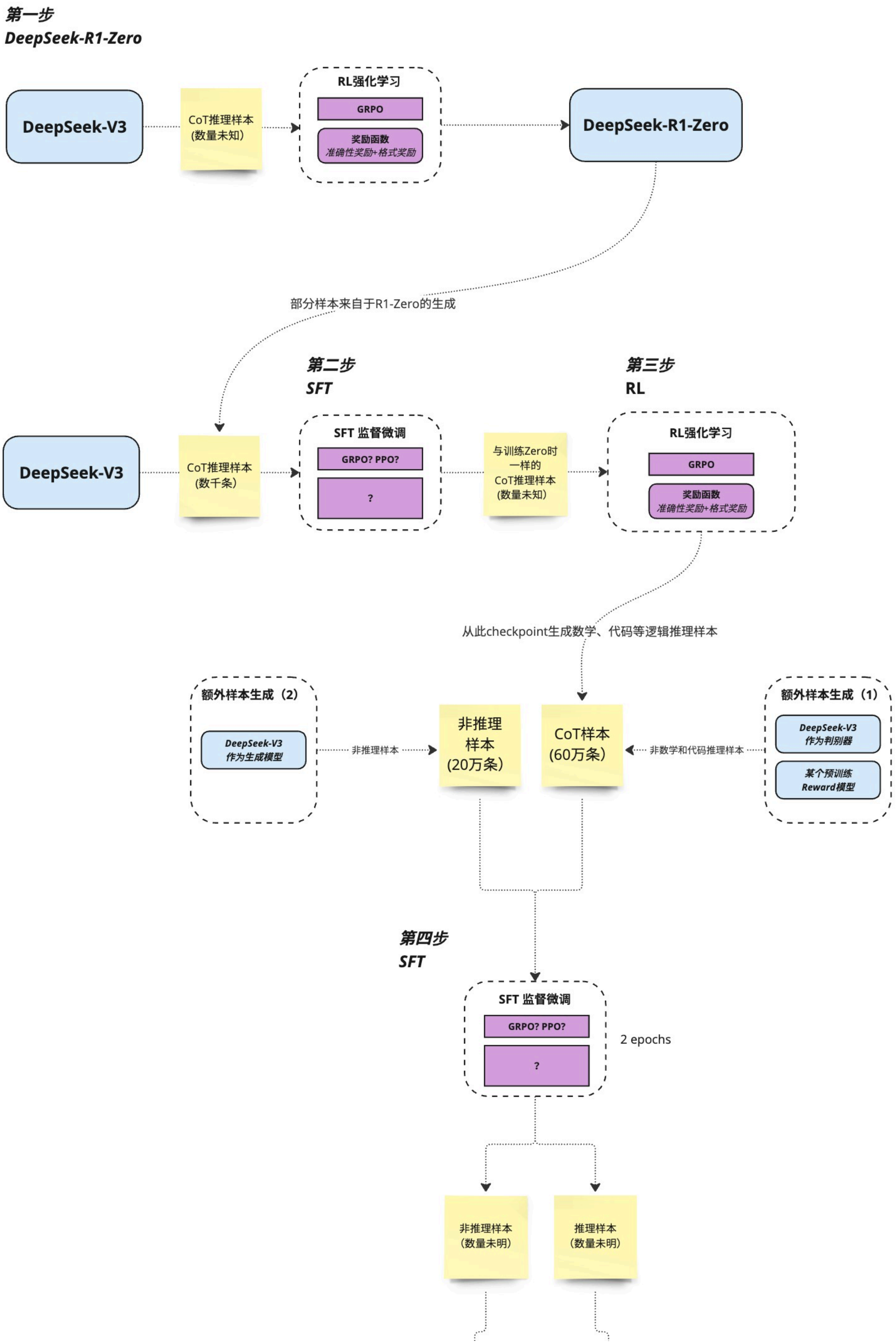
When reasoning-oriented RL converges, we utilize the resulting checkpoint to collect SFT (Supervised Fine-Tuning) data for the subsequent round. Unlike the initial cold-start data, which primarily focuses on reasoning, this stage incorporates data from other domains to enhance the model's capabilities in writing, role-playing, and other general-purpose tasks. Specifically, we generate the data and fine-tune the model as described below.

利用得到的checkpoint作为sft数据

来用于下一轮的训练,与冷启动过程不同的是,这个阶段主要基于推理能力,且这个阶段合并了其他领域数据,用来增强编写等通用任务的能力

有推理数据(拓展了样本数据集,用v3评判模型,和一个reward模型来训练)和非推理数据(v3生成)

2.3.4 训练适用于所有场景的强化学习



第一步

DeepSeek-R1-Zero

DeepSeek-V3

CoT推理样本（数量未知）

RL强化学习

GRPO

奖励函数

准确性奖励+格式奖励

DeepSeek-R1-Zero

部分样本来自于R1-Zero的生成

第二步

SFT

第三步

RL

DeepSeek-V3

CoT推理样本（数千条）

SFT 监督微调

GRPO? PPO?

?

与训练Zero时一样的CoT推理样本（数量未知）

RL强化学习

GRPO

奖励函数

准确性奖励+格式奖励

从此checkpoint生成数学、代码等逻辑推理样本

额外样本生成（2）

DeepSeek-V3 作为生成模型

非推理样本

非推理样本（20万条）

CoT样本（60万条）

非数学和代码推理样本

额外样本生成（1）

DeepSeek-V3 作为判别器

某个预训练Reward模型

第四步

SFT

SFT 监督微调

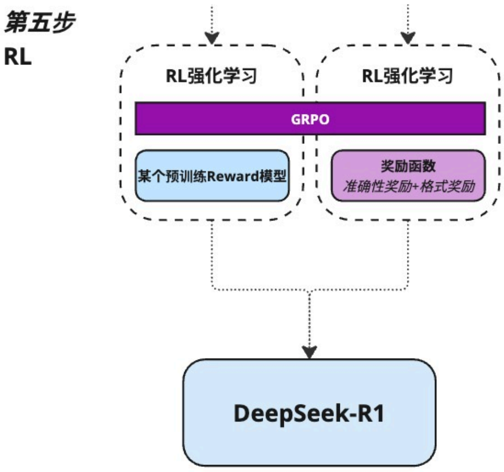
GRPO? PPO?

?

2 epochs

非推理样本（数量未明）

推理样本（数量未明）



2.4蒸馏 提升小模型推理能力

对比了蒸馏和小模型强化学习,蒸馏结果更好

非成功的尝试,比如说,用prm 过程奖励函数,蒙特卡洛树搜索,效果不好,效率不高,空间无限的,不一定不行,题目认为不成功

总结限制未来工作