

# DeepSeek-R1: 通过强化学习激励LLMs中的推理能力

深度探索人工智能

research@deepseek.com

## 摘要

我们推出了第一代推理模型，DeepSeek-R1-Zero 和 DeepSeek-R1。DeepSeek-R1-Zero 是一个通过大规模强化学习（RL）训练而成的模型，无需监督微调（SFT）作为初步步骤，展示了卓越的推理能力。通过 RL，DeepSeek-R1-Zero 自然涌现出许多强大且有趣的推理行为。然而，它也面临诸如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升推理性能，我们推出了 DeepSeek-R1，它在 RL 之前引入了多阶段训练和冷启动数据。DeepSeek-R1 在推理任务上实现了与 OpenAI-o1-1217 相当的性能。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-R1 蒸馏出的六个密集模型（1.5B、7B、8B、14B、32B、70B）。

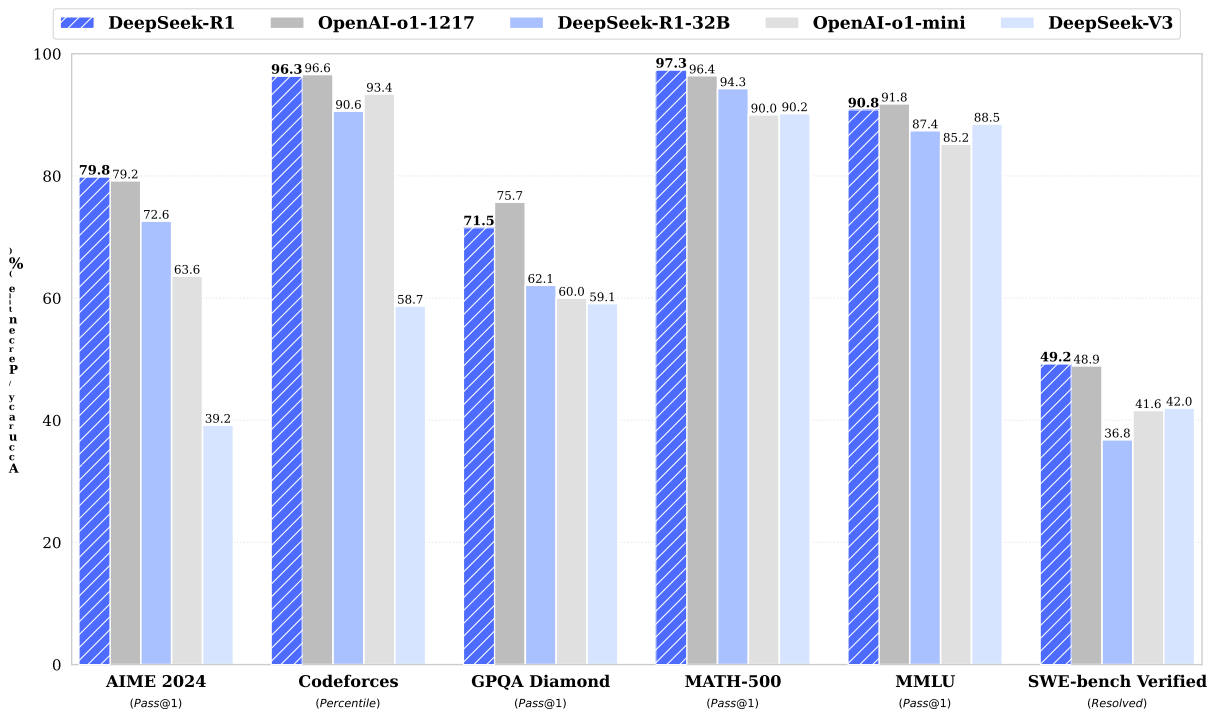


图1 | DeepSeek-R1的基准性能。

# 目录

<b>1 引言</b>	<b>3</b>
1.1 贡献	4
1.2 评估结果总结	4
<b>2 方法</b>	<b>5</b>
2.1 概述	5
2.2 DeepSeek-R1-Zero: 基础模型上的强化学习	5
2.2.1 强化学习算法	5
2.2.2 奖励建模	6
2.2.3 训练模板	6
2.2.4 DeepSeek-R1-Zero 的性能、自我进化过程和顿悟时刻	6
2.3 DeepSeek-R1: 冷启动强化学习	9
2.3.1 冷启动	9
2.3.2 面向推理的强化学习	10
2.3.3 拒绝采样与监督微调	10
2.3.4 全场景强化学习	11
2.4 蒸馏: 赋予小模型推理能力	11
<b>3 实验</b>	<b>11</b>
3.1 DeepSeek-R1 评估	12
3.2 蒸馏模型评估	14
<b>4 讨论</b>	<b>14</b>
4.1 蒸馏与强化学习	15
4.2 未成功的尝试	15
<b>5 结论、局限性与未来工作</b>	<b>16</b>
<b>A 贡献与致谢</b>	<b>20</b>

# 1. 引言

近年来，大型语言模型（LLMs）经历了快速的迭代和进化（Anthropic, 2024; Google, 2024; OpenAI, 2024a），逐步缩小了与人工通用智能（AGI）之间的差距。

最近，后训练已成为完整训练流程中的一个重要组成部分。它已被证明可以提高推理任务的准确性，与社会价值观保持一致，并适应用户偏好，同时相对于预训练所需的计算资源相对较少。在推理能力方面，OpenAI的o1（OpenAI, 2024b）系列模型首次通过增加思维链推理过程的长度引入了推理时扩展。这种方法在数学、编码和科学推理等各种推理任务中取得了显著改进。然而，有效的测试时扩展仍然是研究界的一个开放性问题。之前的一些工作探索了各种方法，包括基于过程的奖励模型（Lightman等, 2023; Uesato等, 2022; Wang等, 2023）、强化学习（Kumar等, 2024）以及蒙特卡洛树搜索和束搜索等搜索算法（Feng等, 2024; Trinh等, 2024; Xin等, 2024）。然而，这些方法都没有达到与OpenAI的o1系列模型相媲美的通用推理性能。

在本文中，我们迈出了利用纯强化学习（RL）提升语言模型推理能力的第一步。我们的目标是探索大型语言模型（LLMs）在没有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RL过程的自我进化。具体而言，我们使用DeepSeek-V3-Base作为基础模型，并采用GRPO（Shao等, 2024）作为RL框架，以提升模型在推理任务中的表现。在训练过程中，DeepSeek-R1-Zero自然涌现出许多强大且有趣的推理行为。经过数千次RL步骤后，DeepSeek-R1-Zero在推理基准测试中展现出卓越的性能。例如，AIME 2024上的pass@1得分从15.6%提升至71.0%，而在多数投票机制下，得分进一步提高至86.7%，与OpenAI-o1-0912的性能相当。

然而，DeepSeek-R1-Zero 遇到了诸如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升推理性能，我们引入了 DeepSeek-R1，它结合了少量冷启动数据和多阶段训练流程。具体来说，我们首先收集数千条冷启动数据来微调 DeepSeek-V3-Base 模型。随后，我们进行类似 DeepSeek-R1-Zero 的推理导向强化学习（RL）。在 RL 过程接近收敛时，我们通过对 RL 检查点进行拒绝采样，结合来自 DeepSeek-V3 在写作、事实问答和自我认知等领域的监督数据，创建新的 SFT 数据，然后重新训练 DeepSeek-V3-Base 模型。使用新数据微调后，检查点会经历额外的 RL 过程，考虑所有场景的提示。经过这些步骤，我们获得了称为 DeepSeek-R1 的检查点，其性能与 OpenAI-o1-1217 相当。

我们进一步探索了从DeepSeek-R1到更小密集模型的蒸馏过程。以Qwen2.5-32B（Qwen, 2024b）为基础模型，直接从DeepSeek-R1进行蒸馏的效果优于在其上应用强化学习。这表明，更大基础模型发现的推理模式对于提升推理能力至关重要。我们开源了蒸馏后的Qwen和Llama（Dubey等, 2024）系列。值得注意的是，我们蒸馏的14B模型大幅超越了当前最先进的开源QwQ-32B-Preview（Qwen, 2024a），而蒸馏的32B和70B模型在密集模型的推理基准测试中创下了新纪录。

## 1.1. 贡献

训练后：在基础模型上进行大规模强化学习

- 我们直接将强化学习（RL）应用于基础模型，而不依赖于监督微调（SFT）作为初步步骤。这种方法使模型能够探索思维链（CoT）以解决复杂问题，从而开发出DeepSeek-R1-Zero。DeepSeek-R1-Zero展示了自我验证、反思和生成长思维链等能力，标志着研究社区的一个重要里程碑。值得注意的是，这是首个公开的研究，验证了LLMs的推理能力可以纯粹通过RL激励，而不需要SFT。这一突破为这一领域的未来进展铺平了道路。
- 我们介绍了开发DeepSeek-R1的流程。该流程包含两个RL阶段，旨在发现改进的推理模式并与人类偏好对齐，以及两个SFT阶段，作为模型推理和非推理能力的基础。我们相信该流程将通过创建更好的模型为行业带来益处。

蒸馏：小型模型也能强大

- 我们证明了较大模型的推理模式可以被提炼到较小的模型中，与通过强化学习在小模型上发现的推理模式相比，性能更优。开源的DeepSeek-R1及其API将有益于研究社区在未来提炼出更好的小型模型。
- 利用DeepSeek-R1生成的推理数据，我们对研究社区中广泛使用的多个密集模型进行了微调。评估结果表明，经过蒸馏的较小密集模型在基准测试中表现尤为出色。DeepSeek-R1-Distill-Qwen-7B在AIME 2024上达到了55.5%，超越了QwQ-32B-Preview。此外，DeepSeek-R1-Distill-Qwen-32B在AIME 2024上得分为72.6%，在MATH-500上得分为94.3%，在LiveCodeBench上得分为57.2%。这些结果显著优于之前的开源模型，并与o1-mini相当。我们向社区开源了基于Qwen2.5和Llama3系列的1.5B、7B、8B、14B、32B和70B蒸馏检查点。

## 1.2. 评估结果总结

- 推理任务：（1）DeepSeek-R1在AIME 2024上取得了79.8%的Pass@1分数，略微超过了OpenAI-o1-1217。在MATH-500上，它获得了令人印象深刻的97.3%的分数，与OpenAI-o1-1217表现相当，并显著优于其他模型。（2）在与编码相关的任务中，DeepSeek-R1在代码竞赛任务中展示了专家水平，它在Codeforces上获得了2,029的Elo评分，超过了96.3%的人类参赛者。对于工程相关任务，DeepSeek-R1的表现略优于DeepSeek-V3，这可能有助于开发人员在现实世界任务中取得更好的成果。
- 知识：在MMLU、MMLU-Pro和GPQA Diamond等基准测试中，DeepSeek-R1取得了出色的成绩，显著超越了DeepSeek-V3，得分分别为MMLU 90.8%、MMLU-Pro 84.0%和GPQA Diamond 71.5%。虽然在这些基准测试中其表现略低于OpenAI-o1-1217，但DeepSeek-R1超越了其他闭源模型，展示了其在教育任务中的竞争优势。在事实基准测试SimpleQA上，DeepSeek-R1的表现优于DeepSeek-V3，展示了其处理基于事实的查询的能力。类似趋势在OpenAI-o1超越4o的基准测试中也有所体现。

- 其他方面：DeepSeek-R1 在多种任务中也表现出色，包括创意写作、通用问答、编辑、摘要等。它在 AlpacaEval 2.0 上实现了 87.6% 的长度控制胜率，在 Are-naHard 上实现了 92.3% 的胜率，展示了其智能处理非应试查询的强大能力。此外，DeepSeek-R1 在需要长上下文理解的任务上表现出色，在长上下文基准测试中显著优于 DeepSeek-V3。

## 2. 方法

### 2.1. 概述

先前的工作严重依赖大量监督数据来提升模型性能。在本研究中，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）也能显著提升推理能力。此外，加入少量冷启动数据可以进一步提升性能。在接下来的章节中，我们将介绍：（1）DeepSeek-R1-Zero，它直接将RL应用于基础模型，不使用任何SFT数据；（2）DeepSeek-R1，它从经过数千个长链思维（CoT）示例微调的检查点开始应用RL；（3）将DeepSeek-R1的推理能力蒸馏到小型密集模型中。

### 2.2. DeepSeek-R1-Zero：基础模型上的强化学习

强化学习在推理任务中展现了显著的有效性，正如我们之前的工作所证明的那样（Shao等，2024；Wang等，2023）。然而，这些工作严重依赖于监督数据，而这些数据的收集非常耗时。在本节中，我们探索了大型语言模型（LLMs）在没有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强化学习算法，随后展示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. Reinforcement Learning Algorithm

群体相对策略优化 为了节省强化学习的训练成本，我们采用了群体相对策略优化（GRPO）（Shao等，2024），该方法放弃了通常与策略模型大不相同的评论家模型，转而从群体分数中估计基线。具体来说，对于每个问题 $q$ ，GRPO从旧策略 $\pi_{\theta_{old}}$ 中采样一组输出 $\{o_1, o_2, \dots, o_G\}$ ，然后通过最大化以下目标来优化策略模型 $\pi_{\theta}$ ：

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \frac{1}{G} \sum_{i=1}^G \left( \min \left( \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left( \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

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

其中 $\epsilon$ 和 $\beta$ 是超参数， $A_i$ 是优势，通过使用一组奖励 $\{r_1, r_2, \dots, r_G\}$ 来计算，这些奖励对应于每个组内的输出：

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

---

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:

---

表1 | DeepSeek-R1-Zero的模板。在训练过程中，prompt将被替换为具体的推理问题。

### 2.2.2. Reward Modeling

奖励是训练信号的来源，它决定了强化学习的优化方向。为了训练DeepSeek-R1-Zero，我们采用了一种基于规则的奖励系统，主要包括两种类型的奖励：

- 准确性奖励：准确性奖励模型评估响应是否正确。例如，在具有确定性结果的数学问题中，模型需要以指定格式（例如，在方框内）提供最终答案，从而实现基于规则的可靠正确性验证。同样，对于LeetCode问题，可以使用编译器根据预定义的测试用例生成反馈。
- 格式化奖励：除了准确性奖励模型外，我们还采用了一种格式化奖励模型，该模型强制模型将其思考过程放在‘<思考>’和‘</思考>’标签之间。

我们在开发DeepSeek-R1-Zero时没有应用结果或过程神经奖励模型，因为我们发现神经奖励模型在大规模强化学习过程中可能会受到奖励黑客攻击，并且重新训练奖励模型需要额外的训练资源，这会使整个训练流程复杂化。

### 2.2.3. Training Template

为了训练DeepSeek-R1-Zero，我们首先设计了一个简单的模板，指导基础模型遵循我们指定的指令。如表1所示，该模板要求DeepSeek-R1-Zero首先生成推理过程，然后给出最终答案。我们有意将约束限制在这种结构格式上，避免任何内容特定的偏见——例如强制进行反思性推理或推广特定的问题解决策略——以确保我们能够准确观察模型在强化学习（RL）过程中的自然进展。

### 2.2.4. Performance, Self-evolution Process and Aha Moment of DeepSeek-R1-Zero

DeepSeek-R1-Zero 的性能 图2展示了DeepSeek-R1-Zero在AIME 2024基准测试上通过强化学习（RL）训练过程中的性能轨迹。如图所示，随着RL训练的推进，DeepSeek-R1-Zero表现出稳定且持续的性能提升。值得注意的是，AIME 2024上的平均pass@1分数显著增加，从最初的15.6%跃升至令人印象深刻的71.0%，达到了与OpenAI-o1-0912相当的性能水平。这一显著改进凸显了我们的RL算法在优化模型性能方面的有效性。

表2提供了DeepSeek-R1-Zero与OpenAI的o1-0912模型在各种推理相关基准上的比较分析。研究结果表明，RL赋能

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

表2 | DeepSeek-R1-Zero与OpenAI o1模型在推理相关基准上的比较。

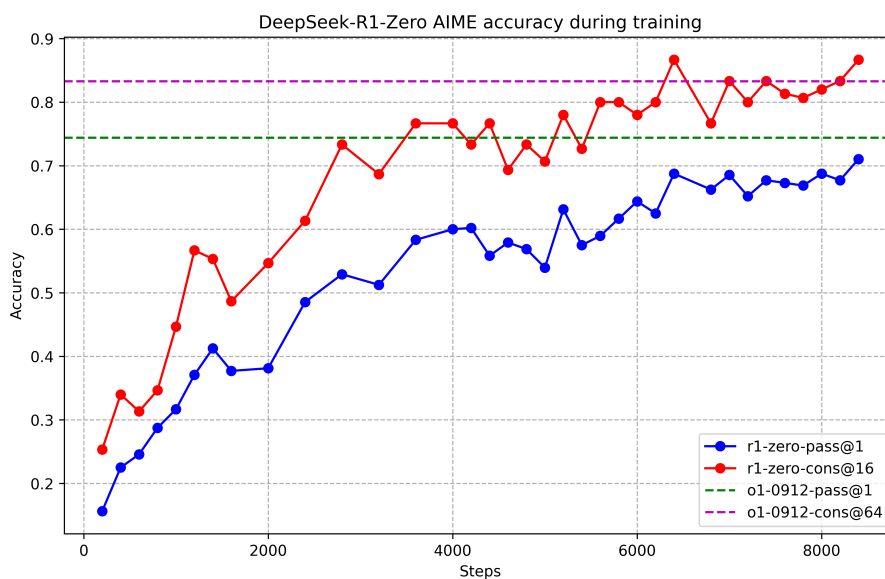


图2 | DeepSeek-R1-Zero在训练期间的AIME准确率。对于每个问题，我们采样16个响应并计算总体平均准确率，以确保评估的稳定性。

DeepSeek-R1-Zero 无需任何监督微调数据即可获得强大的推理能力。这是一项值得注意的成就，因为它强调了模型仅通过强化学习就能有效学习和泛化的能力。此外，通过应用多数投票，DeepSeek-R1-Zero 的性能可以进一步增强。例如，在 AIME 基准测试中使用多数投票时，DeepSeek-R1-Zero 的性能从 71.0% 提升到 86.7%，从而超过了 OpenAI-o1-0912 的表现。DeepSeek-R1-Zero 无论是否使用多数投票都能取得如此有竞争力的表现，这突显了其强大的基础能力及其在推理任务中进一步发展的潜力。

DeepSeek-R1-Zero 的自我进化过程 DeepSeek-R1-Zero 的自我进化过程是一个令人着迷的演示，展示了强化学习 (RL) 如何驱动模型自主提升其推理能力。通过直接从基础模型启动 RL，我们可以在不受监督微调阶段影响的情况下，密切监控模型的进展。这种方法清晰地展示了模型随时间的演变，尤其是在处理复杂推理任务方面的能力提升。

如图3所示，DeepSeek-R1-Zero的思考时间呈现出持续提升-

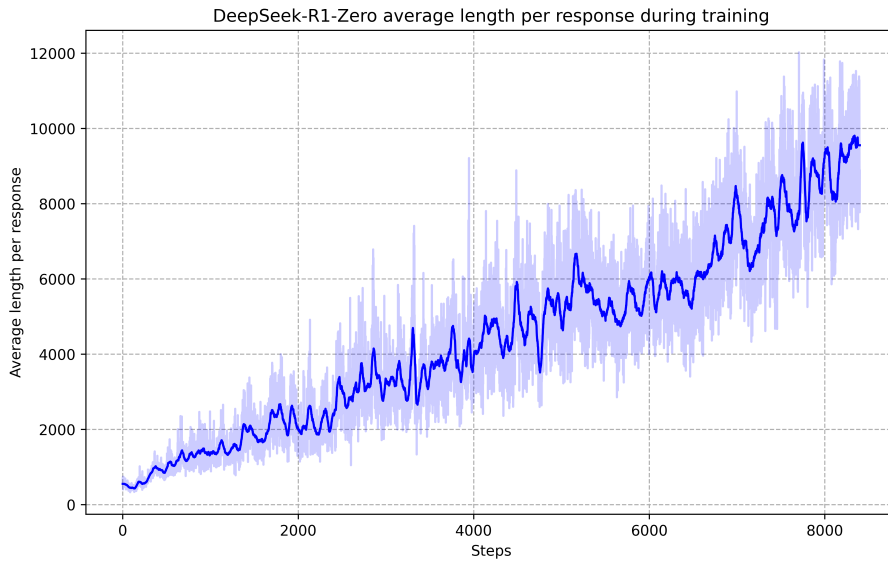


图3 | DeepSeek-R1-Zero在RL过程中训练集上的平均响应长度。DeepSeek-R1-Zero自然地学会了通过更多的思考时间来解决推理任务。

在整个训练过程中，模型的改进并非源于外部调整，而是模型内部固有的发展。DeepSeek-R1-Zero 通过利用扩展的测试时间计算，自然地获得了解决日益复杂推理任务的能力。这种计算范围从生成数百到数千个推理标记，使模型能够更深入地探索和完善其思维过程。

这种自我进化最显著的方面之一是随着测试时间计算的增加，复杂行为的出现。诸如反思——模型重新审视和重新评估其先前步骤——以及探索解决问题的替代方法等行为会自发产生。这些行为并非明确编程，而是模型与强化学习环境交互的结果。这种自发发展显著增强了DeepSeek-R1-Zero的推理能力，使其能够更高效、更准确地应对更具挑战性的任务。

**DeepSeek-R1-Zero 的顿悟时刻** 在训练DeepSeek-R1-Zero 的过程中，观察到了一个特别有趣的现象，即“顿悟时刻”的出现。如表 3 所示，这一时刻出现在模型的中间版本中。在此阶段，DeepSeek-R1-Zero 学会了通过重新评估其初始方法，为问题分配更多的思考时间。这种行为不仅证明了模型推理能力的提升，也是强化学习如何带来意外且复杂结果的一个引人入胜的例证。

这一时刻不仅是模型的“顿悟时刻”，也是观察其行为的研究人员的“顿悟时刻”。它突显了强化学习的力量与美感：我们无需明确教导模型如何解决问题，只需提供适当的激励，它便能自主发展出高级的问题解决策略。这一“顿悟时刻”有力地提醒了强化学习在解锁人工智能系统新智能水平方面的潜力，为未来更加自主和自适应的模型铺平了道路。



---

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 ...

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

Rearrange to isolate the inner square root term:

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

...

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 ...

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: ...

...

---

表3 | DeepSeek-R1-Zero中间版本的一个有趣的“顿悟时刻”。模型学会了以拟人化的语气重新思考。这也是我们的一个顿悟时刻，让我们见证了强化学习的力量与美感。

DeepSeek-R1-Zero的缺点 尽管DeepSeek-R1-Zero展现出强大的推理能力，并自主发展出意想不到且强大的推理行为，但它仍面临一些问题。例如，DeepSeek-R1-Zero在可读性差和语言混合等挑战上表现不佳。为了使推理过程更具可读性并与开放社区共享，我们探索了DeepSeek-R1，这是一种利用人类友好冷启动数据进行强化学习的方法。

### 2.3. DeepSeek-R1：冷启动强化学习

受到DeepSeek-R1-Zero的显著成果启发，两个自然问题随之而来：1) 通过引入少量高质量数据作为冷启动，能否进一步提升推理性能或加速收敛？2) 我们如何训练一个用户友好的模型，不仅能生成清晰连贯的思维链（CoT），还能展现出强大的通用能力？为了解决这些问题，我们设计了一个训练DeepSeek-R1的流程。该流程包含四个阶段，概述如下。

#### 2.3.1. Cold Start

与DeepSeek-R1-Zero不同，为了防止从基础模型开始的RL训练早期不稳定冷启动阶段，对于DeepSeek-R1，我们构建并收集了一小部分长链思维（CoT）数据来微调模型，作为初始RL演员。为了收集此类数据，我们探索了几种方法：使用带有长链思维示例的少样本提示，直接提示模型生成带有反思和验证的详细答案，以可读格式收集DeepSeek-R1-Zero的输出，并通过人工注释者的后处理来精炼结果。

在本工作中，我们收集了数千条冷启动数据，以微调DeepSeek-V3-Base作为强化学习的起点。与DeepSeek-R1-Zero相比，冷启动数据的优势

包含:

- 可读性: DeepSeek-R1-Zero 的一个关键限制是其内容通常不适合阅读。响应可能会混合多种语言, 或者缺乏用于突出显示答案的 markdown 格式。相比之下, 在为 DeepSeek-R1 创建冷启动数据时, 我们设计了一种可读的模式, 包括在每个响应的末尾添加摘要, 并过滤掉不便于阅读的响应。在这里, 我们将输出格式定义为 `|special_token|<reasoning_process>|special_token|<summary>`, 其中推理过程是查询的 CoT, 而摘要用于总结推理结果。
- 潜力: 通过精心设计具有人类先验知识的冷启动数据模式, 我们观察到相对于 DeepSeek-R1-Zero 有更好的性能。我们相信迭代训练是推理模型的更好方式。

### 2.3.2. Reasoning-oriented Reinforcement Learning

在对冷启动数据进行微调 DeepSeek-V3-Base 后, 我们采用了与 DeepSeek-R1-Zero 相同的大规模强化学习训练过程。这一阶段重点提升模型的推理能力, 特别是在编码、数学、科学和逻辑推理等推理密集型任务中, 这些任务涉及具有明确解决方案的明确定义问题。在训练过程中, 我们观察到 CoT 经常出现语言混合现象, 尤其是在 RL 提示涉及多种语言时。为了缓解语言混合问题, 我们在 RL 训练中引入了语言一致性奖励, 该奖励计算为 CoT 中目标语言词汇的比例。尽管消融实验表明这种对齐会导致模型性能略有下降, 但这种奖励符合人类偏好, 使其更具可读性。最后, 我们将推理任务的准确性与语言一致性奖励直接相加, 形成最终奖励。然后, 我们在微调后的模型上应用强化学习 (RL) 训练, 直到其在推理任务上达到收敛。

### 2.3.3. Rejection Sampling and Supervised Fine-Tuning

当面向推理的强化学习收敛时, 我们利用生成的检查点来收集 SFT (监督微调) 数据, 用于后续轮次。与最初主要关注推理的冷启动数据不同, 此阶段会融入来自其他领域的的数据, 以增强模型在写作、角色扮演和其他通用任务中的能力。具体而言, 我们按照以下描述生成数据并微调模型。

**推理数据** 我们精心策划推理提示, 并通过从上述 RL 训练的检查点执行拒绝采样来生成推理轨迹。在前一阶段, 我们仅包含可以使用基于规则的奖励进行评估的数据。然而, 在这一阶段, 我们通过整合额外数据来扩展数据集, 其中一些数据通过将真实值和模型预测输入 DeepSeek-V3 进行判断, 使用生成式奖励模型。此外, 由于模型输出有时混乱且难以阅读, 我们过滤掉了混合语言、长段落和代码块的思维链。对于每个提示, 我们采样多个响应并仅保留正确的响应。总共, 我们收集了约 60 万条与推理相关的训练样本。

非推理数据 对于非推理数据，如写作、事实问答、自我认知和翻译，我们采用DeepSeek-V3管道并复用部分DeepSeek-V3的SFT数据集。对于某些非推理任务，我们调用DeepSeek-V3在回答问题前生成潜在的思维链。然而，对于更简单的查询，如“你好”，我们不会提供思维链作为回应。最终，我们收集了大约20万个与推理无关的训练样本。

我们使用上述约80万个样本的精选数据集对DeepSeek-V3-Base进行了两个周期的微调。

### 2.3.4. Reinforcement Learning for all Scenarios

为了进一步使模型与人类偏好保持一致，我们实施了一个次级强化学习阶段，旨在提高模型的有用性和无害性，同时精炼其推理能力。具体来说，我们结合奖励信号和多样化的提示分布来训练模型。对于推理数据，我们遵循DeepSeek-R1-Zero中概述的方法，该方法利用基于规则的奖励来指导数学、代码和逻辑推理领域的学习过程。对于一般数据，我们依靠奖励模型来捕捉复杂和微妙场景中的人类偏好。我们在DeepSeek-V3流程的基础上，采用了类似的偏好对分布和训练提示。对于有用性，我们专注于最终总结，确保评估强调响应对用户的实用性和相关性，同时最小化对底层推理过程的干扰。对于无害性，我们评估模型的整个响应，包括推理过程和总结，以识别和减轻生成过程中可能出现的任何潜在风险、偏见或有害内容。最终，奖励信号和多样化数据分布的整合使我们能够训练出一个在推理方面表现出色，同时优先考虑有用性和无害性的模型。

## 2.4. 蒸馏：赋予小模型推理能力

为了让更高效的小型模型具备类似DeepSeek-R1的推理能力，我们直接使用DeepSeek-R1整理的80万样本对开源模型如Qwen (Qwen, 2024b) 和Llama (AI@Meta, 2024) 进行了微调，具体细节见§2.3.3。我们的研究表明，这种简单的蒸馏方法显著提升了小型模型的推理能力。我们在此使用的基础模型包括Qwen2.5-Math-1.5B、Qwen2.5-Math-7B、Qwen2.5-14B、Qwen2.5-32B、Llama-3.1-8B和Llama-3.3-70B-Instruct。我们选择Llama-3.3是因为其推理能力略优于Llama-3.1。

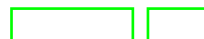
对于蒸馏模型，我们仅应用SFT，不包括RL阶段，尽管引入RL可以显著提升模型性能。我们这里的主要目标是展示蒸馏技术的有效性，将RL阶段的探索留给更广泛的研究社区。

## 3. 实验

基准测试 我们在MMLU (Hendrycks等, 2020)、MMLU-Redux (Gema等, 2024)、MMLU-Pro (Wang等, 2024)、C-Eval (Huang等, 2023)、CMMLU (Li等, 2023)、IFEval (Zhou等, 2023)、FRAMES (Krishna等, 2024)、GPQA Diamond (Rein等, 2023)、SimpleQA (OpenAI, 2024c)、C-SimpleQA (He等, 2024)、SWE-Bench Verified (OpenAI,

2024d], Aider<sup>1</sup>, LiveCodeBench (Jain 等, 2024) (2024-08 – 2025-01), Codeforces<sup>2</sup>, 中国国家高中数学奥林匹克 (CNMO 2024)<sup>3</sup>, 以及美国数学邀请赛 2024 (AIME 2024) (MAA, 2024)。除了标准基准测试外, 我们还使用 LLMs 作为评判者, 在开放式生成任务上评估我们的模型。具体来说, 我们遵循 AlpacaEval 2.0 (Dubois 等, 2024) 和 Arena-Hard (Li 等, 2024) 的原始配置, 这些配置利用 GPT-4-Turbo-1106 作为成对比较的评判者。在这里, 我们只将最终摘要提供给评估, 以避免长度偏差。对于蒸馏模型, 我们报告了在 AIME 2024、MATH-500、GPQA Diamond、Codeforces 和 LiveCodeBench 上的代表性结果。

评估提示 遵循 DeepSeek-V3 的设置, 使用 simple-evals 框架中的提示对标准基准测试 (如 MMLU、DROP、GPQA Diamond 和 SimpleQA) 进行评估。对于 MMLU-Redux, 我们在零样本设置中采用 Zero-Eval 提示格式 (Lin, 2024)。对于 MMLU-Pro、C-Eval 和 CLUE-WSC, 由于原始提示是少样本的, 我们略微修改提示以适应零样本设置。少样本中的 CoT 可能会影响 DeepSeek-R1 的性能。其他数据集遵循其原始评估协议, 并使用其创建者提供的默认提示。对于代码和数学基准测试, HumanEval-Mul 数据集涵盖了八种主流编程语言 (Python、Java、C++、C#、JavaScript、TypeScript、PHP 和 Bash)。使用 CoT 格式评估模型在 LiveCodeBench 上的表现, 数据收集时间为 2024 年 8 月至 2025 年 1 月。Codeforces 数据集使用 10 场 Div.2 比赛的题目以及专家编写的测试用例进行评估, 之后计算预期评级和竞争对手的百分比。SWE-Bench 的验证结果通过无代理框架 (Xia 等, 2024) 获得。AIDER 相关基准测试使用 “diff” 格式进行测量。DeepSeek-R1 的输出在每个基准测试中最多限制为 32,768 个标记。



基线 我们对多个强大的基线进行了全面评估, 包括 DeepSeek-V3、Claude-Sonnet-3.5-1022、GPT-4o-0513、OpenAI-o1-mini 和 OpenAI-o1-1217。由于在中国大陆访问 OpenAI-o1-1217 API 较为困难, 我们根据官方报告记录了其性能。对于蒸馏模型, 我们还比较了开源模型 QwQ-32B-Preview (Qwen, 2024a)。



生成设置 对于我们所有的模型, 最大生成长度设置为 32,768 个标记。对于需要采样的基准测试, 我们使用温度为 0.6, top-p 值为 0.95, 并为每个查询生成 64 个响应以估计 pass@1。

### 3.1. DeepSeek-R1 评估

对于以教育为导向的知识基准, 如 MMLU、MMLU-Pro 和 GPQA Diamond, DeepSeek-R1 相比 DeepSeek-V3 表现出更优越的性能。这一改进主要归因于在 STEM 相关问题上准确性的提升, 通过大规模强化学习 (RL) 实现了显著进步。此外, DeepSeek-R1 在 FRAMES 这一依赖长上下文的问答任务上表现出色, 展示了其强大的文档分析能力。这凸显了推理模型在 AI 驱动中的潜力。

---

<sup>1</sup><https://aider.chat>

<sup>2</sup><https://codeforces.com>

<sup>3</sup><https://www.cms.org.cn/Home/comp/comp/cid/12.html>

Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1	
Architecture	-	-	MoE	-	-	MoE	
# Activated Params	-	-	37B	-	-	37B	
# Total Params	-	-	671B	-	-	671B	
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	<b>91.8</b>	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	<b>92.9</b>
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	<b>84.0</b>
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	<b>92.2</b>
	IF-Eval (Prompt Strict)	<b>86.5</b>	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	<b>75.7</b>	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	<b>47.0</b>	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	<b>82.5</b>
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	<b>87.6</b>
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	<b>92.3</b>
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	<b>65.9</b>
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	<b>96.6</b>	96.3
	Codeforces (Rating)	717	759	1134	1820	<b>2061</b>	2029
	SWE Verified (Resolved)	<b>50.8</b>	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	<b>61.7</b>	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	<b>79.8</b>
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	<b>97.3</b>
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	<b>78.8</b>
Chinese	CLU EWSC (EM)	85.4	87.9	90.9	89.9	-	<b>92.8</b>
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	<b>91.8</b>
	C-SimpleQA (Correct)	55.4	58.7	<b>68.0</b>	40.3	-	63.7

表4 | DeepSeek-R1与其他代表性模型比较。

搜索和数据分析任务。在事实基准测试SimpleQA上，DeepSeek-R1优于DeepSeek-V3，展示了其处理基于事实查询的能力。类似的趋势在OpenAI-o1超越GPT-4o的基准测试中也有所体现。然而，DeepSeek-R1在中文SimpleQA基准测试上表现不如DeepSeek-V3，主要是由于其在安全强化学习（RL）后倾向于拒绝回答某些查询。如果没有安全RL，DeepSeek-R1的准确率可以超过70%。

DeepSeek-R1在IF-Eval基准测试中也取得了令人印象深刻的结果，该基准旨在评估模型遵循格式指令的能力。这些改进可以归因于在监督微调（SFT）和强化学习（RL）训练的最后阶段加入了指令遵循数据。此外，在AlpacaEval2.0和ArenaHard上观察到了显著的表现，表明DeepSeek-R1在写作任务和开放域问答方面的优势。其显著优于DeepSeek-V3的表现突显了大规模RL的泛化优势，不仅提升了推理能力，还提高了跨多个领域的性能。此外，DeepSeek-R1生成的摘要长度简洁，在ArenaHard上平均为689个token，在AlpacaEval2.0上平均为2,218个字符。这表明DeepSeek-R1在基于GPT的评估中避免了引入长度偏差，进一步巩固了其在多项任务中的鲁棒性。

在数学任务上，DeepSeek-R1表现出与OpenAI-o1-1217相当的性能，远超其他模型。在编码算法任务上，如LiveCodeBench和Codeforces，也观察到类似的趋势，专注于推理的模型在这些基准测试中占据主导地位。在面向工程的编码任务上，OpenAI-o1-1217在Aider上优于DeepSeek-R1，但在SWE Verified上表现相当。我们认为工程

DeepSeek-R1 的性能将在下一个版本中提升，因为目前相关的 RL 训练数据量仍然非常有限。

### 3.2. 蒸馏模型评估

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	<b>1820</b>
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	<b>72.6</b>	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	<b>86.7</b>	<b>94.5</b>	<b>65.2</b>	<b>57.5</b>	1633

表5 | DeepSeek-R1蒸馏模型与其他可比模型在推理相关基准上的比较。

如表5所示，仅通过蒸馏DeepSeek-R1的输出，就能使高效的DeepSeek-R1-7B（即DeepSeek-R1-Distill-Qwen-7B，下文简称相同）全面超越非推理模型如GPT-4o-0513。DeepSeek-R1-14B在所有评估指标上均超过了QwQ-32B-Preview，而DeepSeek-R1-32B和DeepSeek-R1-70B在大多数基准测试中显著超越了o1-mini。这些结果展示了蒸馏的强大潜力。此外，我们发现对这些蒸馏模型应用强化学习（RL）能带来显著的进一步增益。我们认为这值得进一步探索，因此在此仅展示简单SFT蒸馏模型的结果。

## 4. 讨论

### 4.1. 蒸馏 vs. 强化学习

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	<b>72.6</b>	<b>83.3</b>	<b>94.3</b>	<b>62.1</b>	<b>57.2</b>

表6 | 蒸馏模型与强化学习模型在推理相关基准上的比较。

在第3.2节中，我们可以看到，通过蒸馏DeepSeek-R1，小模型可以取得令人印象深刻的结果。然而，仍然有一个问题：模型是否可以通过论文中讨论的大规模RL训练而不进行蒸馏，达到类似的性能？

为了回答这个问题，我们在Qwen-32B-Base上使用数学、代码和STEM数据进行了大规模RL训练，训练超过10K步，最终得到了DeepSeek-R1-Zero-Qwen-32B。实验结果如图6所示，表明32B基础模型在经过大规模训练后，{v\*} □

RL训练，达到了与QwQ-32B-Preview相当的性能。然而，从DeepSeek-R1蒸馏而来的DeepSeek-R1-Distill-Qwen-32B在所有基准测试中表现显著优于DeepSeek-R1-Zero-Qwen-32B。因此，我们可以得出两个结论：首先，将更强大的模型蒸馏为较小的模型能产生出色的结果，而依赖本文中提到的大规模RL的较小模型需要巨大的计算能力，甚至可能无法达到蒸馏的性能。其次，尽管蒸馏策略既经济又有效，但要突破智能的界限，可能仍需要更强大的基础模型和更大规模的强化学习。

## 4.2. 未成功的尝试

在开发DeepSeek-R1的早期阶段，我们也经历了许多失败和挫折。我们在此分享失败的经验，以提供见解，但这并不意味着这些方法无法开发出有效的推理模型。

过程奖励模型（PRM） PRM 是一种合理的方法，用于引导模型朝着更好的解决推理任务的方向发展（Lightman 等，2023；Uesato 等，2022；Wang 等，2023）。然而，在实践中，PRM 有三个主要局限性，可能会阻碍其最终的成功。首先，在一般推理中明确定义细粒度步骤是具有挑战性的。其次，确定当前中间步骤是否正确是一项艰巨的任务。使用模型进行自动标注可能无法产生令人满意的结果，而手动标注则不利于扩展。第三，一旦引入基于模型的 PRM，不可避免地会导致奖励欺骗（Gao 等，2022），并且重新训练奖励模型需要额外的训练资源，并使整个训练流程复杂化。总之，虽然 PRM 展示了在重新排序模型生成的前 N 个响应或辅助引导搜索方面的良好能力（Snell 等，2024），但在我们的大规模强化学习实验中，与其引入的额外计算开销相比，其优势有限。

蒙特卡洛树搜索（MCTS）受AlphaGo（Silver等，2017b）和AlphaZero（Silver等，2017a）启发，我们探索了使用蒙特卡洛树搜索（MCTS）来增强测试时计算的可扩展性。该方法涉及将答案分解为较小的部分，以便模型能够系统地探索解决方案空间。为此，我们提示模型生成多个标签，这些标签对应于搜索所需的特定推理步骤。在训练过程中，我们首先使用收集的提示通过由预训练的价值模型引导的MCTS找到答案。随后，我们使用生成的问题-答案对来训练演员模型和价值模型，并迭代优化这一过程。

然而，这种方法在扩大训练规模时遇到了几个挑战。首先，与国际象棋不同，国际象棋的搜索空间相对明确，而令牌生成则呈现出一个指数级更大的搜索空间。为了解决这个问题，我们为每个节点设置了最大扩展限制，但这可能导致模型陷入局部最优。其次，价值模型直接影响生成的质量，因为它指导搜索过程的每一步。训练一个细粒度的价值模型本身就很困难，这使得模型难以迭代改进。尽管AlphaGo的核心成功依赖于训练价值模型以逐步提升其性能，但由于令牌生成的复杂性，这一原则在我们的设置中难以复制。

总之，虽然MCTS在与预训练的价值模型配对时可以在推理过程中提高性能，但通过自我搜索迭代提升模型性能仍然是一个

significant challenge.

## 5. 结论、局限性与未来工作

在这项工作中，我们分享了通过强化学习（RL）增强模型推理能力的历程。DeepSeek-R1-Zero 代表了一种不依赖冷启动数据的纯RL方法，在各种任务中表现出色。DeepSeek-R1则更为强大，它结合了冷启动数据和迭代RL微调。最终，DeepSeek-R1在一系列任务上的表现与OpenAI-o1-1217相当。

我们进一步探索将推理能力蒸馏到小型密集模型中。我们使用DeepSeek-R1作为教师模型生成800K数据，并对多个小型密集模型进行微调。结果令人鼓舞：DeepSeek-R1-Distill-Qwen-1.5B在数学基准测试中表现优于GPT-4o和Claude-3.5-Sonnet，AIME得分为28.9%，MATH得分为83.9%。其他密集模型也取得了令人印象深刻的成果，显著优于基于相同基础检查点的其他指令调优模型。

未来，我们计划在以下方向为DeepSeek-R1进行投资研究。

- 通用能力：目前，DeepSeek-R1在函数调用、多轮对话、复杂角色扮演和json输出等任务上的能力尚不及DeepSeek-V3。未来，我们计划探索如何利用长链思维（CoT）来提升这些领域的任务表现。
- 语言混合：DeepSeek-R1目前针对中文和英文进行了优化，这可能导致在处理其他语言的查询时出现语言混合问题。例如，即使查询使用的是非英文或中文的语言，DeepSeek-R1也可能使用英文进行推理和响应。我们计划在未来的更新中解决这一限制。
- 提示工程：在评估DeepSeek-R1时，我们观察到它对提示非常敏感。少样本提示会持续降低其性能。因此，我们建议用户直接描述问题，并在零样本设置中指定输出格式以获得最佳结果。
- 软件工程任务：由于评估时间较长，影响了RL过程的效率，大规模RL在软件工程任务中尚未得到广泛应用。因此，DeepSeek-R1在软件工程基准测试中并未显示出比DeepSeek-V3有显著提升。未来版本将通过实施软件工程数据的拒绝采样或在RL过程中引入异步评估来提高效率。

## 参考文献

AI@Meta. Llama 3.1 模型卡, 2024年。URL [https://github.com/meta-llama/llama-models/blob/main/models/llama3\\_1/MODEL\\_CARD.md](https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md)。

Anthropic. Claude 3.5 十四行诗, 2024。网址 <https://www.anthropic.com/news/claude-3-5-sonnet>。

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, 等。llama 3 模型群。arXiv 预印本 arXiv:2407.21783, 2024。

Y. Dubois, B. Galambosi, P. Liang, 和 T. B. Hashimoto。长度控制的alpacaeval：一种简单的去偏自动评估器方法。arXiv预印本 arXiv:2404.04475, 2024。



X. Feng, Z. Wan, M. Wen, S. M. McAleer, Y. Wen, W. Zhang, 和 J. Wang. Alphazero-like 树搜索可以指导大型语言模型解码和训练, 2024. URL <https://arxiv.org/abs/2309.17179>.

L. Gao, J. Schulman, 和 J. Hilton. 奖励模型过度优化的缩放定律, 2022. URL <https://arxiv.org/abs/2210.10760>.

A. P. Gema, J. O. J. Leang, G. Hong, A. Devoto, A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, M. R. G. Madani, C. Barale, R. McHardy, J. Harris, J. Kaddour, E. van Krieken, 和 P. Minervini. 我们完成MMLU了吗? CoRR, abs/2406.04127, 2024. URL <https://doi.org/10.48550/arXiv.2406.04127>.

谷歌。我们的下一代模型: Gemini 1.5, 2024年。网址 <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024>.

Y. He, S. Li, J. Liu, Y. Tan, W. Wang, H. Huang, X. Bu, H. Guo, C. Hu, B. Zheng, 等. 中文简单问答: 针对大语言模型的中文事实性评估. arXiv预印本 arXiv:2411.07140, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, 和 J. Steinhardt. 测量大规模多任务语言理解. arXiv 预印本 arXiv:2009.03300, 2020.

Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, 等. C-Eval: 一个多层次多学科的中文基础模型评估套件. arXiv预印本 arXiv:2305.08322, 2023.

N. Jain, K. Han, A. Gu, W. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, 和 I. Stoica. Live codebench: 对大型语言模型进行代码评估的全面且无污染的方法. CoRR, abs/2403.07974, 2024. URL {v\*}.

S. Krishna, K. Krishna, A. Mohananeey, S. Schwarcz, A. Stambler, S. Upadhyay, 和 M. Faruqui. 事实、获取与推理: 检索增强生成的统一评估. CoRR, abs/2409.12941, 2024. doi: 10.48550/ARXIV.2409.12941. URL {v\*} {v\*}.

A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs, 等. 通过强化学习训练语言模型进行自我纠正. arXiv 预印本 arXiv:2409.12917, 2024.

H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, N. Duan, 和 T. Baldwin. CMMLU: 衡量中文大规模多任务语言理解. arXiv 预印本 arXiv:2306.09212, 2023.

T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, 和 I. Stoica. 从众包数据到高质量基准: Arena-hard 和 Benchbuilder 管道. arXiv 预印本 arXiv:2406.11939, 2024.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, 和 K. Cobbe. 让我们一步步验证. arXiv预印本 arXiv:2305.20050, 2023.

B. Y. Lin. ZeroEval: 评估语言模型的统一框架, 2024年7月. URL <https://github.com/WildEval/ZeroEval>.

- MAA. 美国数学邀请赛 - AIME。在美国数学邀请赛 - AIME 2024, 2024年2月。URL <https://maa.org/math-competitions/american-invitational-mathematics-examination-aime>。
- OpenAI. 你好 GPT-4o, 2024a。URL <https://openai.com/index/hello-gpt-4o/>。
- OpenAI. 学习用llms推理, 2024b。URL <https://openai.com/index/learn-how-to-reason-with-llms/>。
- OpenAI. 介绍SimpleQA, 2024c。网址 <https://openai.com/index/introducing-simpleqa/>。
- OpenAI. 介绍SWE-bench验证版, 我们正在发布一个经过人工验证的swe-bench子集, 更多内容, 2024年。URL <https://openai.com/index/introducing-swe-bench-verified/>。
- Qwen. Qwq: 深刻反思未知的边界, 2024a。URL <https://qwenlm.github.io/blog/qwq-32b-preview/>。
- Qwen. Qwen2.5: 基础模型的聚会, 2024b。URL <https://qwenlm.github.io/blog/qwen2.5>。
- D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, 和 S. R. Bowman。GPQA: 一个研究生级别的防谷歌问答基准。arXiv预印本 arXiv:2311.12022, 2023。
- Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. Li, Y. Wu, 和 D. Guo. Deepseekmath: 推动开放语言模型数学推理的极限。arXiv 预印本 arXiv:2402.03300, 2024。
- D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. P. Lillicrap, K. Simonyan, 和 D. Hassabis。通过自我对弈掌握国际象棋和将棋的通用强化学习算法。CoRR, abs/1712.01815, 2017a。URL <http://arxiv.org/abs/1712.01815>。
- D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. P. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, 和 D. Hassabis。无需人类知识掌握围棋游戏。《自然》, 550(7676):354-359, 2017b。doi: 10.1038/NATURE24270。URL <https://doi.org/10.1038/nature24270>。
- C. Snell, J. Lee, K. Xu, 和 A. Kumar。在2024年提出, 优化扩展LLM测试时计算可能比扩展模型参数更有效。URL <https://arxiv.org/abs/2408.03314>。
- T. Trinh, Y. Wu, Q. Le, H. He, 和 T. Luong. 无需人类示范解决奥林匹克几何问题。《自然》, 2024。doi: 10.1038/s41586-023-06747-5。
- J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, 和 I. Higgins。使用过程和结果反馈解决数学文字问题。arXiv 预印本 arXiv:2211.14275, 2022。
- 王鹏, 李磊, 邵志远, 徐瑞, 戴德明, 李阳, 陈东, 吴宇, 隋志远. Math-shepherd: 一种用于数学推理中大型语言模型的无标签逐步验证器。arXiv预印本 arXiv:2312.08935, 2023。

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, 和 W. Chen. Mmlu-pro: 一个更强大且更具挑战性的多任务语言理解基准. CoRR, abs/2406.01574, 2024. URL <https://doi.org/10.48550/arXiv.2406.01574>. C. S. Xia, Y. Deng, S. Dunn, 和 L. Zhang. Agentless: 揭秘基于LLM的软件工程代理. arXiv 预印本, 2024. H. Xin, Z. Z. Ren, J. Song, Z. Shao, W. Zhao, H. Wang, B. Liu, L. Zhang, X. Lu, Q. Du, W. Gao, Q. Zhu, D. Yang, Z. Gou, Z. F. Wu, F. Luo, 和 C. Ruan. Deepseek-prover-v1.5: 利用证明助手反馈进行强化学习和蒙特卡洛树搜索, 2024. URL <https://arxiv.org/abs/2408.08152>. J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Lu an, D. Zhou, 和 L. Hou. 大语言模型的指令遵循评估. arXiv 预印本 arXiv:2311.07911, 2023.

---

## 附录

### A. 贡献与致谢

#### 核心贡献者

郭大亚 杨德建  
张浩伟 宋俊晓  
张若愚 徐润新  
朱启豪 马世荣  
王培艺 毕晓 张  
晓康 于兴凯 吴  
宇 吴子峰 吴志  
斌 苟志宏 邵卓  
书 李子怡 高子  
义

惠利 建忠 郭嘉实  
李景昌 陈景阳 袁  
金浩 涂俊杰 邱俊  
龙 李J.L. 蔡佳琪 倪  
健 梁金 陈凯 董凯  
胡凯\* 游凯超 高凯  
歌 关康 黄可欣 于  
快 王乐聪 张亮 赵  
亮 王立通 张丽月  
徐磊 夏乐毅 张明  
川 张明华 唐明辉  
周明旭 李萌 王妙  
君 李明明 田宁 黄  
盼盼 张鹏 王千成  
陈钦宇 杜秋实 葛  
瑞琪\* 张瑞松 潘瑞  
哲 王润基 陈R.J. 金  
R.L.

贡献者 刘爱新 薛  
冰 王炳轩 吴伯超  
冯蓓 卢成达 赵成  
刚 邓成奇 阮冲 戴  
大麦 陈德利 纪东  
杰 李二航 林芳云  
戴福聪 罗富立\*  
郝广博 陈冠廷 李  
国伟 H. 张 徐汉威  
丁红辉 高华作 屈  
辉

如意 陈尚浩 卢尚岩 周山黄 陈盛丰 叶诗雨 王水平 于顺丰 周舒婷 潘 S.S. 李双周 邵清 吴盛丰 叶涛 云田 培天宇 孙 T. 王王鼎 曾文 刘文峰 梁文俊 高文钦 于\* 文涛 张 W.L. 肖伟安 晓东 刘晓涵 王小康 陈晓涛 聂鑫 程鑫 刘鑫 谢兴超 刘新宇 杨新元 李学成 苏徐恒 林 X.Q. 李向月 金晓金 沈晓莎 陈晓文 孙晓翔 王新南 宋新怡 周宪祖 王新霞 单 Y.K. 李 Y.Q. 王

Y.X. 魏阳 张艳红 徐瑶 李瑶 赵耀峰 孙耀辉 王毅 余一超 张一帆 史一良 熊英 何一石 朴一松 王艺璇 谭一阳 马\* 刘一源 郭永强 欧元 王玉端 龚悦 邹宇恒 何雨佳 熊云帆 罗宇翔 游宇翔 刘宇轩 周宇阳 Y.X. 朱延平 黄耀辉 李毅 郑宇辰 朱云仙 马英 唐宇坤 查宇婷 严 Z.Z. 任泽辉 任张力 沙哲夫 徐哲安 谢振达 张正言 郝哲文 马志成 严志刚 吴志宇 顾子辉

朱子嘉 刘子  
君\* 李子琳  
谢子维 宋子  
阳 潘子正

黄震 徐志鹏 张  
忠宇 张震

在每个角色中，作者按名字的字母顺序排列。标有 \* 的姓名表示已离开我们团队的个人。

本文由 AINLP 公众号通过大模型 API 进行翻译，更多资源请扫码关注！



长按扫码关注我们

# AINLP

我爱自然语言处理

一个能聊天有趣有AI的NLP公众号