

DeepSeek-R1: 通过强化学习激励LLMs中的推理能力

DeepSeek-人工智能

research@deepseek.com

摘要

我们介绍了我们的第一代重新定位模型，DeepSeek-r1-零和DeepSeek-r1。DeepSeek-R1-Zero是一个通过大规模强化学习（RL）训练的模型，没有经过监督微调（SFT）作为初步步骤，展示了显著的推理能力。通过RL，DeepSeek-R1-零natu集会出现了许多强大而有趣的推理行为。然而，它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。DeepSeek-R1在推理任务上取得了与OpenAI-o1-1217相当的性能。为了支持研究社区，我们开源了DeepSeek-R1-零、DeepSeek-R1、r1和6个密集模型（1.5B、7B、8B、14B、32B、70B）。

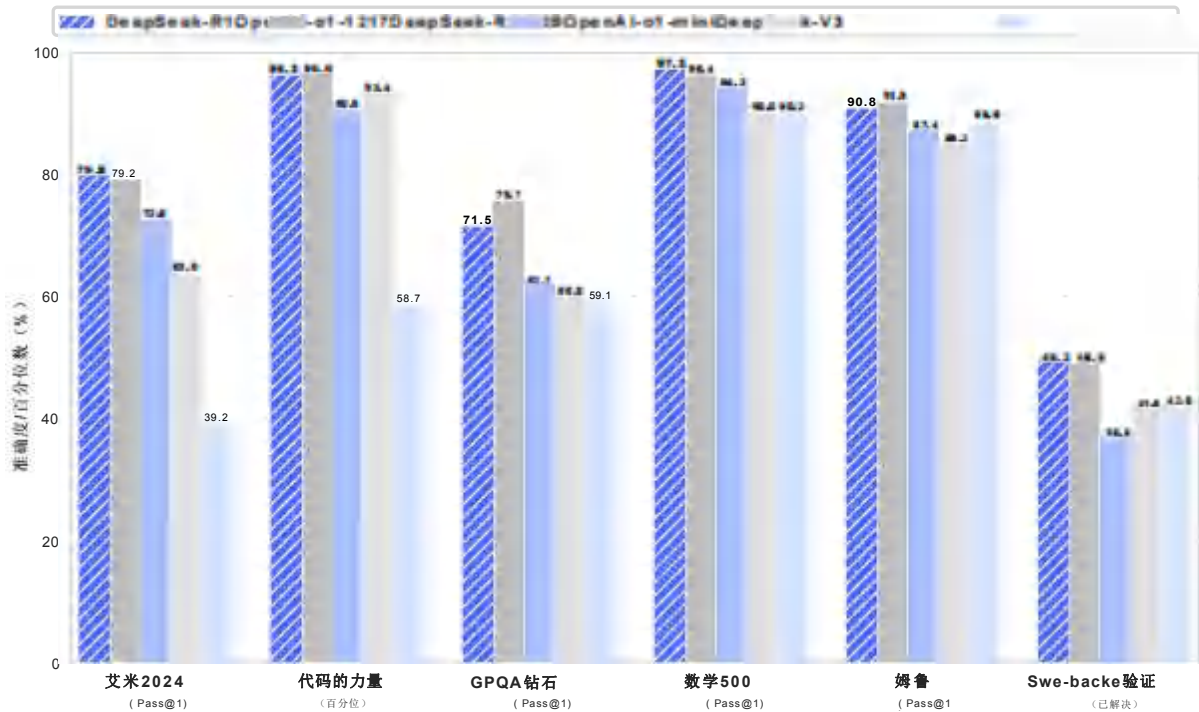


图1 DeepSeek-R1的基准测试性能。



人工智能产业链联盟

星主： AI产业链盟主

 知识星球

微信扫描预览星球详情



内容

	1介绍	3
1.1 贡献		4
1.2 评价结果汇总表		4
	2方法	5
2.1 概述		5
2.2 在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 培训模板		6
	2.2.4 性能、DeepSeek的自我进化过程和Aha时刻	6
2.3 DeepSeek-R1：在冷启动条件下的强化学习		9
2.3.1 冷启动		9
2.3.2 面向推理的强化学习		10
2.3.3 拒绝采样和监督微调		10
2.3.4 强化学习的所有场景		11
2.4 蒸馏：授权具有推理能力的小模型		11
	3实验	11
3.1 DeepSeek-R1评估		13
3.2 蒸馏模型评价		14
	4讨论	14
4.1 蒸馏法v .s。强化学习		14
4.2 失败的尝试		15
	5.结论、局限性和未来的工作	16
A的贡献和致谢	20	

1.介绍

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

最近，培训后已成为整个培训管道的一个重要组成部分。研究表明，它可以提高推理任务的准确性，与社会价值观相一致，并适应用户的偏好，同时比训练前需要相对最小的人工计算资源。在推理能力方面，OpenAI的o1 (OpenAI, 2024b)系列模型是第一个通过增加思维链推理过程的长度来引入推理时间sc传导的模型。这种方法在各种推理任务上取得了显著的改进，如数学、编码和科学推理。然而，有效的测试时间缩放挑战仍然是研究界的一个有待解决的问题。一些初级研究已经探索了各种方法，包括基于过程的奖励模型(莱特曼 以及其他他人 2023;Uesato等人, 2022;王等人, 2023)、强化学习(Kumar等人, 2024)，以及诸如蒙特卡罗树搜索和波束搜索(Feng et al., 2024;特林 以及其他他人 2024;辛等人, 2024)。然而，这些方法都没有达到可与OpenAI的o1系列模型相媲美的一般推理性能。

在本文中，我们利用纯强化学习（RL）为提高语言模型推理能力迈出了第一步。我们的目标是探索llm在没有任何监督数据的情况下发展推理能力的潜力，重点关注于它们通过纯RL过程的自我进化。特别地，我们使用DeepSeek-V3-Base作为基础模型，并使用GRPO (Shao 等人, 2024)作为RL框架来提高推理中模型的性能。在训练过程中，DeepSeek-R1-Zero自然出现了许多强大而有趣的推理行为。经过数千个RL步骤后，DeepSeek-R1-Zero在推理基准rks上表现出了优异的性能。例如，AIME 2024的@1通过分数从15.6%上升到71.0%，通过多数投票，分数进一步提高到86.7%，与OpenAI-o1-0912的性能相匹配。

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高ce推理性能，我们引入了DeepSeek-R1，它包含了少量的冷启动数据和一个多阶段的训练管道。具体来说，我们首先收集数千个冷启动数据来微调DeepSeek-V3-Base模型。接下来，我们执行面向推理的RL，就像DeepSeek-R1-Zero一样。在RL过程接近收敛时，我们通过在RL检查点进行拒绝采样来创建新的SFT数据，结合DeepSeek-V3在写作、事实QA和自我认知等领域的监督数据，然后对DeepSeek-V3-Base模型进行再训练。在对新数据进行微调后，检查对象将经历一个额外的RL过程，并考虑到来自所有场景的提示。在这些步骤之后，我们获得了一个被称为DeepSeek-R1的检查点，它实现了与OpenAI-o1-1217相同的性能。

我们进一步探索了从DeepSeek-R1到更小的密集模型的蒸馏。使用Qwen2.5-32B (Qwen, 2024b)作为基础模型，从DeepSeek-R1中直接蒸馏优于应用RL。这表明，由更大的基础模型发现的推理模式有利于提高推理能力。我们开源了蒸馏过的Qwen和美洲驼 (Dubey 以及其他他人 2024) 一系列值得注意的是，我们提炼出的14B模型优于最先进的开源QwQ-32B-预览版(Qwen, 2024a)，蒸馏后的32B和70B模型设置为a关于密集模型之间的推理基准测试的新记录。

1.1.贡献

训练后：在基础模型上的大规模强化学习

我们直接将RL应用于基础模型，而不依赖于监督微调（SFT）作为一个初步步骤。这种方法允许模型探索思想链（CoT）来解决复杂的问题，从而开发出DeepSeek-R1-Zero。DeepSeek-R1-Zero演示了诸如自我验证、反思和生成长cot等能力，这标志着研究界的一个重要里程碑。值得注意的是，这是第一个开放的研究，以验证LLMs的推理能力可以纯粹通过RL来激励，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的管道。该管道包含了两个rl阶段，旨在发现改进的推理模式，并与人类的偏好对齐，以及两个SFT阶段，作为该模型的推理和非推理能力的种子。我们相信，该管道将通过创造更好的模型，使该行业受益。

蒸馏效果：较小的模型也可以很强大

我们证明了较大的models的推理模式，可以提炼成更小的模型，与通过RL在小模型上发现的推理模式相比，它具有更好的性能。开源的DeepSeek-R1及其API将有利于研究界在未来提炼出更好的更小的模型。

利用DeepSeek-R1生成的推理数据，我们对几个在研究领域广泛使用的密集模型进行了微调。评价结果表明，蒸馏的小密度模型在基准上表现非常好。DeepSeek-R1-蒸馏版-Qwen-7B在AIME 2024年展会上达到55.5%，超过了QwQ-32B-预览版。此外，DeepSeek-r1-蒸馏-Qwen-32B在AIME 2024上得分72.6%，在MATH-500上分为94.3%，在LiveCodeBench上分为57.2%。这些结果明显优于在此之前的开源模型，并可与o1-mini相媲美。我们以Qwen 2.5和Llama 3系列为基础，向社区开源提取了1.5B、7B、8B、14B、32B和70 B的检查点。

1.2.评价结果汇总表

推理任务：(1)DeepSeek-R1在AIME 2024上获得了79.8%的通过率@1的分数，略高于OpenAI-o1-1217。在MATH-500上，它获得了令人印象深刻的97.3%的分数，表现与OpenAI-o1-1217相当，明显优于其他车型。(2)在与编码相关的任务中，DeepSeek-R1在代码竞争任务中展示了足够的水平，它在代码力量中获得了2029个Elo评分，在竞争中超过了96.3%的人类参与者。对于与工程相关的任务，DeepSeek-R1的表现略好于DeepSeek-V3，这可以帮助开发人员在现实世界的任务中工作。

知识：在MM LU、MMLU-Pro和GPQA钻石等基准测试中，DeepSeek-R1取得了显著的成绩，显著超过DeepSeek-V3，MM LU 90.8%，MMLU-Pro84.0%，GPQA钻石71.5%。虽然在这些基准测试上，它的性能略低于OpenAI-o1-1217，但DeepSeek-R1超过了其他闭源模型，显示了它在教育任务方面的竞争优势。在事实基准SimpleQA上，DeepSeek-R1的性能超过了DeepSeek-V3，展示了它处理基于事实的查询的能力。在这个基准测试上，OpenAI-o1超过4o时也观察到类似的趋势。

其他的：DeepSeek-R1也涉及到广泛的任务，包括创造性的写作、一般的问题回答、编辑、总结等等。它在paca2.0上达到了87.6%的胜率，在ara-naHard上的胜率为92.3%，显示了它智能处理非考试或智能查询的强大能力。此外，DeepSeek-R1在重新激发长上下文理解的任务上展示了出色的性能，在长上下文基准测试上大大优于DeepSeek-V3。

2.方法

2.1.概述

以前的工作非常依赖于大量的监督数据来提高模型的性能。在这项研究中，我们证明了通过大规模的强化学习（RL）可以显著提高推理能力，即使不使用监督微调（SFT）作为一个冷开始。此外，通过包含少量的冷启动数据，还可以进一步提高性能。在下面的章节中，我们介绍：(1)DeepSeek-R1-Zero，它将RL直接应用于没有任何SFT数据的基础模型；(2) DeepSeek-R1，它从一个经过数千个长思想链（CoT）例子微调的检查点开始应用RL。3)将推理能力从DeepSeek-R1提取到小型密集模型。

2.2.Reinforce巩固了在基础模型上的学习

强化学习在推理任务中已经证明了显著的有效性，正如我们之前的工作所述(Shao et al., 2024;王等人, 2023)。然而，这些工作严重依赖于监督数据，而监督数据是收集时间的。在本节中，我们将探讨llm在没有任何超级建议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

组相对策略优化为了节省RL的培训成本，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者模型，而是从组分数中估计基线。具体来说，对于每个问题 q ，GRPO从旧策略中抽取一组输出 $\{o_1, o_2, \dots, o_G\}$ ，然后通过最大化以下目标优化策略模型 π_θ ：

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

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

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

where ϵ and β 区域超参数，而 a_i 是优势，使用一组奖励计算 $\{r_1, r_2, \dots, r_G\}$ 对应于每一组内的输出：

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

用户和助手之间的对话。用户问一个问题，助手就会解决它。助手首先考虑思维中的推理过程，然后提供给用户与答案。推理过程和答案分别包含在<思考></think>和<回答></回答>标签，即<认为>推理过程这里</think><回答>，回答这里，</回答>。用户：提示符。助理

表1为DeepSeek-R1-零的模板。提示将被具体的推理所取代
培训期间的问题。

2.2.2.奖励建模

奖励是训练信号的来源，它决定了RL的优化方向。为了训练DeepSeek-R1- Zero，我们采用了一个基于规则的奖励系统，它主要由两种类型的奖励组成：

准确性奖励：准确性奖励模型评估反应是否正确。例如，对于具有确定性结果的数学问题，模型需要以指定的格式（例如，在一个方框内）提供最终答案，从而实现可靠的基于规则的正确性验证。类似地，对于LeetCode问题，可以使用编译器来基于预定义的测试用例生成反馈。

格式奖励：除了准确性奖励模型外，我们还采用了一种格式奖励模型，强制该模型将其思维过程置于“<trink>”和“</trink>”标签之间。

我们在开发DeepSeek-R1- Zero的过程中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遭受奖励黑客攻击，再训练奖励模型需要额外的训练资源，使整个训练管道复杂化。

2.2.3.培训模板

为了培训DeepSeek-R1-Zero，我们首先设计了一个简单的模板，以指导基础模型遵循我们指定的指令。如表1所示，这个模板需要DeepSeek-R1- Zero，首先产生一个推理过程，然后是最后的答案。我们有意地将我们的约束限制在这种结构格式上，避免任何内容特定的偏差——例如强制反思推理或促进特定问题解决的策略——以确保我们能够准确地观察模型在RL过程中的自然进展。

2.2.4.DeepSeek-r1-零的性能、自我进化过程和Aha时刻

DeepSeek-R1-零图2 描述了DeepSeek-R1-Zero在整个RL训练过程中的性能轨迹。如所示，DeepSeek-R1- Zero随着RL训练的进展而稳定和一致的提高。值得注意的是，AIME 2024年的平均通过率@1分显著增加，从最初的15.6%跃升到令人印象深刻的71.0%，达到了与OpenAI-o1-0912相当的性能水平。这一显著的改进突出了我们的RL算法在随着时间的推移而优化模型的性能方面的有效性。

表2提供了DeepSeek-R1- Zero和OpenAI的o1-0912之间的比较分析模型跨越了各种与推理相关的基准测试。研究结果显示，RL具有能力

模型	艾米2024		数学500	格帕 金刚石	实时代码 长凳	代码的力量
	pass@1	cons@64	pass@1	pass@1	pass@1	等级
OpenAI-o1-迷你	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-R1-零	71.0	86.7	95.9	73.3	50.0	1444

表2|DeepSeek-R1-Zero和OpenAI o1模型与推理相关的|比较
基准

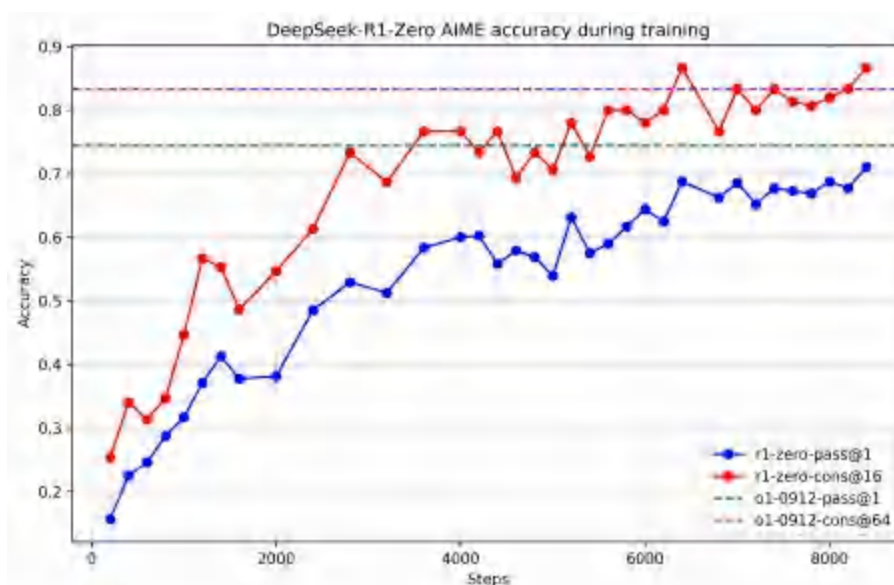


图2|DeepSeek-R1-Zero在训练期间的AIME精度。对于每个问题，我们进行抽样16个响应，并计算总体平均精度，以确保一个稳定的评价。

DeepSeek-R1-Zero，以获得健壮的推理能力，而不需要任何受监督的微调数据。这是一个值得注意的成就，因为它强调了模型单独通过RL有效学习和归纳的能力。此外，DeepSeek-e1-zero的性能可以通过多数投票的应用进一步提高。例如，当在AIME基准测试中采用多数投票时，DeepSeek-R1-Zero的性能从71.0%上升到86.7%，从而超过了OpenAI-o1-0912的性能。DeepSeek-R1-Zero实现这种竞争表现的能力，无论是否有多数投票，都突出了其强大的基础能力和在推理任务方面进一步进步的潜力。

DeepSeek-R1-Zero的自我进化过程是一个有趣的演示，展示了RL如何驱动一个模型来自主提高其推理能力。通过直接从基础模型启动RL，我们可以密切监控模型的进程，而不受监督微调阶段的影响。这种方法提供了模型如何随时间发展的清晰视图，特别是在它处理复杂推理任务的能力方面。

如图3所示，DeepSeek-R1-Zero的思维时间显示出持续的改善-

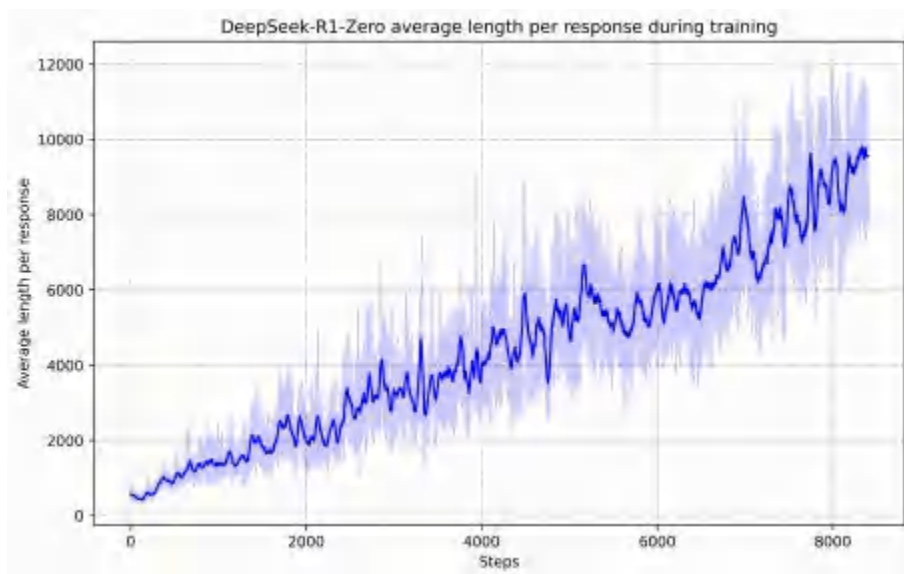


图3|在RL期间，DeepSeek-R1-零在训练集上的平均响应长度
过程DeepSeek-R1- Zero自然地学会了用更多的思考时间来解决推理任务。

并贯穿整个培训过程。这种改进不是外部调整的结果，而是模型内部的内在发展。DeepSeek-R1-Zero通过利用扩展的测试时间的比较，自然获得了解决日益复杂的推理任务的能力。这种计算范围从数百到数千个推理标记，允许模型更深入地探索和完善其思维过程。

这种自我进化最引人注目的一个方面是，随着测试时间计算的增加，复杂的行为的出现。诸如反思等行为——模型重新访问并重新评估其之前的步骤——以及对解决问题的探索是自发产生的。这些行为不是明确编程的，而是由于模型与强化学习环境的交互作用而产生的。这种自发的开发显著提高了DeepSeek-R1-Zero的推理能力，使其能够以更高的效率和准确性处理更具挑战性的任务。

在DeepSeek-R1-零的训练过程中观察到的一个特别有趣的现象是一个“啊哈时刻”的出现。此时此刻，如表3所示，出现在模型的中间版本中。在这一阶段，DeepSeek-R1- Zero通过重新评估其初始方法来学习为一个问题分配更多的思考时间。这种行为不仅证明了模型不断增长的推理能力，也是一个迷人的例子，说明强化学习如何导致意想不到的和复杂的结果。

这不仅是模型的“时刻”，也是观察模型行为的“时刻”。它强调了强化学习的力量和美丽比明确地教授模型如何解决问题，我们简单地提供正确的激励，它自主地开发adv和解决问题的策略。“啊哈时刻”有力地提醒了RL解锁新关卡的潜力 人工系统中的智能，为更自主和自适应的模型铺平了道路 未来。

问：如果 $x > 1$ ，那么 $\sqrt{a - \sqrt{a + x}} = x$ 的实解的和等于

Response:< think>

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

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

重新排列以隔离内平方根项：

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

等等，等等。等等。这是我可以在这里标记的一个时刻。

让我们一步一步重新计算这个，以确定正确的和是否可以...我们从方程开始：

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

首先，让我们把两边：

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

接下来，我可以再次调整两边，处理等式：...。

表3|DeepSeek-R1-Zero中间版本的一个有趣的“哈哈时刻”。这个模型学会了使用一种拟人化的语气来重新思考。这也是我们顿悟的时刻，让我们见证了强化学习的力量和美。

尽管DeepSeek-R1-Zero具有强大的推理能力，自主地呈现出意想不到的强大推理行为，但它面临着几个问题。例如，DeepSeek-R1-Zero就面临着可读性差和语言混合等挑战。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RL和人类友好的冷启动数据的方法。

2.3.DeepSeek-R1：在冷启动条件下的强化学习

受DeepSeek-R1-Zero的前景良好的研究结果的启发，出现了两个自然问题：1)通过合并少量高质量数据作为冷启动，能否进一步提高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，它不仅能产生清晰连贯的CoT，还能展示出强大的通用能力？为了解决这些问题，我们设计了一个训练DeepSeek-R1的管道。该管道由四个阶段的s组成，概述如下。

2.3.1.冷启动

与DeepSeek-R1-Zero不同，为了防止基础模型的早期不稳定的旧开始阶段，对于DeepSeek-R1，我们构建并收集少量的长CoT数据，微调模型作为初始RL参与者。为了收集这些数据，我们探索了几种方法：以长冠状体的少量热提示为例，直接提示模型通过反思和验证生成详细的答案，以可读的格式收集DeepSeek-R1-Zero输出，并通过人类注释者的后处理来细化结果。

在这项工作中，我们收集了数千个冷启动数据，以微调DeepSeek-V3-Base作为RL的起点。与DeepSeek-R1-Zero相比，冷启动数据的优势

包括

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，或缺乏标记格式来突出显示用户的答案。相反，当为DeepSeek-R1创建冷启动数据时，我们设计了一个可读的模式，在每个报告末包含一个摘要

|会了对读者不友好的回答。在这里，我们将输出格式定义为

特殊标记|<推理过程>|特殊标记|<摘要>，其中推理过程是查询的CoT，该摘要用于总结推理结果。

潜力：通过仔细设计具有人类先验的冷启动数据的模式，我们观察到与DeepSeek-R1-Zero相比有更好的性能。我们相信迭代训练是一种更好的建立推理模型的方法。

2.3.2.面向推理的学习者学习

在冷启动数据的基础上，我们应用了DeepSeek-R1-Zero中采用的样本尺度强化学习训练过程。这个阶段的重点是增强模型的推理能力，特别是在推理密集型任务中，如编码、数学、科学和逻辑推理，这些任务涉及到定义明确的问题和明确的解决方案。在训练过程中，我们观察到CoT经常呈现出语言混合，特别是当RL提示涉及多种语言时。为了缓解语言混合的问题，我们在RL训练中引入了一种语言一致性奖励，它被计算为目标语言词在CoT中的比例。虽然消融实验表明，这种对齐如何导致模型表现的轻微下降，但这种奖励与人类的偏好相一致，使其更容易理解。最后，我们将推理任务的准确性和对语言一致性的奖励结合起来，直接相加，形成最终的奖励。然后，我们在微调模型上应用RL训练，直到它在推理任务上实现收敛。

2.3.3.拒绝采样和监督微调

当面向推理的RL收敛时，我们利用生成的checkpoints为后续一轮收集SFT（监督微调）数据。与主要专注于推理的初始冷启动数据不同，这个阶段整合了来自其他领域的的数据，以增强该模型在写作、角色扮演和其他通用任务方面的能力。具体来说，我们生成数据并对模型进行微调，如下所述。

推理数据我们通过从上述RL训练的检查点中进行拒绝采样来管理推理提示并生成推理轨迹。在前一个阶段，我们只包括了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，我们通过合并额外的数据来扩展数据集，其中一些使用生成重构模型，通过在DeepSeek-v3中的地面真相和模型预测进行判断。此外，由于模型输出有时是混乱的，并且难以读取，所以我们已经用混合语言、长参数和代码块过滤掉了思想链。对于每个提示，我们采样多个响应，并只保留正确的响应。我们总共收集了大约600 k个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QA、自我认知和翻译，我们采用DeepSeek-V3管道和重用DeepSeek-V3 SFT数据集的部分。对于某些非推理任务，我们调用DeepSeek-V3，在通过提示回答问题之前生成一个潜在的思维链。然而，对于更简单的查询，如“hello”，我们不在推理解析中提供CoT。最后，我们总共收集了大约200 k个与推理无关的训练样本。

我们使用上述约800k个样本的管理数据集，对两个时代的DeepSeek-V3-Base进行了微调。

2.3.4.对所有场景和操作系统的强化学习

为了进一步使模型与人类偏好相一致，我们实现了一个s边界强化学习阶段，旨在提高模型的帮助性和无害性，同时同时完善其推理能力。具体来说，我们使用奖励信号和不同的提示分布的组合来训练这个模型。对于推理数据，我们参考DeepSeek-R1-Zero中概述的主题学，它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。对于一般的数据，我们诉诸于奖励模型来捕捉人类在复杂和微妙的场景中的偏好。我们建立了DeepSeek-V3管道，并采用了类似的偏好对和训练提示的分布。对于帮助ss，我们只关注最终的总结，确保评估强调对用户的响应的效用和相关性，同时最大限度地减少对底层推理过程的干扰。对于h无害性，我们评估了模型的整个反应，包括推理过程和总结，以识别和减轻在生成过程中可能出现的任何潜在风险、偏见或有害内容离子过程。最终，奖励信号和不同的数据分布的整合使我们能够训练出一个擅长推理的模型，同时优先考虑帮助和无害。

2.4.蒸馏：授权具有推理能力的小模型

为了使更高效的小型模型具备像DeepSeek-R1这样的推理能力，我们直接微调了开源模型，如Qwen (Qwen, 2024b)和Llama (AI@ Meta, 2024)，使用由DeepSeek-R1管理的800 k样本，详见2.3.3。我们的研究表明，这种直接的蒸馏方法显著提高了sm等位模型的推理能力。我们在这里使用的基本模型是Qwen2.5-Math-1.5B、Qwen2.5-Math-7B、Qwen 2.5-14B、Qwen2.5-32B、Llama-3.1-8B和Llama-3.3-70B-Dudult。我们选择Llama-3.3是因为它的推理能力略好于Llama-3.1。

对于蒸馏后的模型，我们只应用SFT和do，而不包括RL阶段，即使合并RL可以大大提高模型的性能。我们在这里的主要目标是演示提取技术的有效性，将RL阶段的探索留给更广泛的研究领域。

3.实验

我们在MMLU上评估模型的基准。(2020), MMLU-Redux (Gema 以及其他人 2024)、MMLU-Pro (Wang 等人, 2024)、C-Eval (Huang 等人, 2023)和CMMLU (Li 等人, 2023), IFEval (Zhou 等人, 2023)、框架(Krishna 等人, 2024), GPQA钻石(Rein 等人, 2023), SimpleQA (OpenAI, 2024c), C-SimpleQA (He 等人, 2024), 法官验证(OpenAI,

2024d), Aider 1 (Jain等人, 2024)(2024-08 – 2025-01), Codeforces 2, 中国国家高中数学奥林匹克竞赛 (CNMO 2024) 3, 和2024年美国数学运动学邀请赛 (AIME 2024) (MAA, 2024).除了标准的基准测试之外,我们还使用llm作为法官来评估我们在开放式生成任务上的模型。具体来说,我们坚持AlpacaEval 2.0的原始配置(Dubois等人, 2024和硬竞技场 以及其他 2024),它利用GPT-4-Turbo-1106作为两两比较的法官。在这里,我们只将最终的摘要输入给评估,以避免长度偏差。对于蒸馏模型,我们报告了AIME2024、MATH-500、GPQ A钻石、编码力和现场编码台的代表性结果。

在DeepSeek-V3的评估之后,标准基准测试如MMLU、DROP、GPQA钻石和简单GPQA框架的提示进行评估。对于MMLU-Redux,我们采用零-`eval`提示格式(Lin, 2024)在azero镜头设置。在MMLU-Pro、C-Eval和CLUE-WSC方面,由于最初的提示很少有镜头,所以我们将提示稍微修改为零镜头设置。少量喷射可能会影响DeepSeek-R1的性能。其他数据集遵循原始的评估协议,由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8种主流编程语言(Python、Java、C++、c#、JavaScript、TypeScript、PHP和Bash)。使用CoT格式评估现场模型的性能,数据收集于2024年8月至2025年1月之间。代码力数据集使用来自10个Div.2竞赛的问题以及专家制作的测试用例进行评估,然后计算预期的评级和竞争对手的百分比。SWE-Bench验证结果通过无代理框架获得(Xia et al., 2024).与辅助程序相关的基准测试使用“差异”格式来衡量。每个基准测试的DeepSeek-R1输出最多被限制为32,768个令牌。

基线我们对几个强大的基线进行了全面的评估,包括DeepSeek-V3、Claude-Sonnet-3.5-1022、GPT-4o-0513、OpenAI-o1-mini和OpenAI-o1-1217。由于访问OpenAI-o1-1217 API在中国大陆具有挑战性,我们根据官方报告报告其权限。对于蒸馏出来的模型,我们还比较了开源模型QwQ-32B-预览版(Qwen, 2024a)。

我们设置了模型的最大生成长度为32,768个令牌。我们发现,使用贪婪解码来评估长输出推理模型会在不同的检查点上产生更高的重复率和显著的变异性 y 。因此,我们默认通过@`k`评估(Chen等人, 2021),并使用非零温度报告通过@1。具体来说,我们使用0.6的采样温度和顶`p`值为0.95来为每个问题生成`k`个响应(通常在4到64之间,取决于测试集的大小)。然后通过@1,然后计算为

$$\text{pass}@1 = \frac{1}{k} \sum_{i=1}^k p_i,$$

其中, p_i 表示第`i`个响应的正确性。这种方法提供了更可靠的性能估计。对于AIME 2024,我们也报告了共识(多数投票)的结果(王以及其他 2022),使用64个样本,记为`cons@64`。

1<https://aider.chat>

2<https://codeforces.com>

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

3.1. DeepSeek-R1 E估值

基准 (公制)	Claude-3.5- 十四行诗-	GPT-4o 0513	DeepSeek V3	OpenAI o1-minio1-	OpenAI 1-1217	DeepSeek R1
	1022					
架构	-	-	能源部	-	-	能源部
激活参数#总参数	-	-	37B	-	-	37B
	-	-	671B	-	-	671B
MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8	90.8
MMLU-Redux (E	88.9	88.0	89.1	86.7	-	92.9
M) MMLU-Pro (78.0	72.6	75.9	80.3	-	84.0
EM)	88.3	83.7	91.6	83.9	90.2	92.2
英语 DROP (三射F 1)	86.5	84.3	86.1	84.8	-	83.3
IF-Eval (及时严格)	65.0	49.9	59.1	60.0	75.7	71.5
GPQA钻石 (通过@1)	28.4	38.2	24.9	7.0	47.0	30.1
SimpleQA (正确)	72.5	80.5	73.3	76.9	-	82.5
框架 (ACC)	52.0	51.1	70.0	57.8	-	87.6
	85.2	80.4	85.5	92.0	-	92.3
AlpacaEval2.0 (LC-winrate) ArenaHard (GPT-4-1106)						
编码 现场代码 (1通过) 代码力	38.9	32.9	36.2	53.8	63.4	65.9
量 (百分位)	20.3	23.6	58.7	93.4	96.6	96.3
编码器 (额定值)	717	759	1134	1820	2061	2029
软件验证 (重新用)	50.8	38.8	42.0	41.6	48.9	49.2
辅助器 (Acc)	45.3	16.0	49.6	32.9	61.7	53.3
数学 AIME2024 (通过@1)	16.0	9.3	39.2	63.6	79.2	79.8
MATH-500 (通过@1)	78.3	74.6	90.2	90.0	96.4	97.3
CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	78.8
CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	92.8
ChineseC-Eval (EM)	76.7	76.0	86.5	68.9	-	91.8
C-SimpleQA (正确)	55.4	58.7	68.0	40.3	-	63.7

表4|DeepSeek-R1与其他代表性模型比较。

对于以教育为导向的知识基准，如MMLU、MMLU-Pro和GPQA钻石，DeepSeek-R1显示了比DeepSeek-V3更优越的性能。这种识别证明主要归因于与stem相关的问题的准确性的提高，其中显著的增益是通过大规模的强化学习获得的。此外，DeepSeek-R1擅长于框架，这是一个依赖于长上下文的QA任务，展示了其强大的文档分析能力。这突出了推理模型在人工智能驱动的搜索和数据分析任务中的潜力。在事实基准测试SimpleQA上，DeepSeek-R1的性能优于DeepSeek-V3，展示了它在处理基于事实的查询方面的能力。在这个基准测试上，OpenAI-o1超过GPT-4 o也观察到类似的趋势。然而，DeepSeek-R1在中文SemeleQA基准上的表现不如DeepSeek-V3，主要是因为它在安全RL后拒绝回答某些查询。如果没有安全的RL，DeepSeek-R1可以达到70%以上的准确率。

DeepSeek-R1也在IF-Eval上提供了令人印象深刻的结果，IF-Eval是一个旨在评估变形模型在结构中遵循格式的能力的基准。这些改进可以与在监督微调(SFT)和RL训练的最后阶段中所包含的指令跟踪数据联系起来。此外，在AlpacaEval2.0和ArenaHard上观察到显著的表现，表明DeepSeek-R1在写作任务和开放领域问题回答方面的优势。它与DeepSeek-V3之间的显著优势强调了大规模RL的泛化优势，它不仅提高了推理能力，而且还提高了不同领域的性能。此外，DeepSeek-R1生成的摘要长度非常简洁，在ArenaHard上平均有689个标记，在AlpacaEval 2.0上有2218个字符。而这就表明：

DeepSeek-R1避免了在基于gpt的评估过程中引入长度偏差，进一步巩固了其在多个任务中的鲁棒性。

在数学任务上，DeepSeek-R1的性能展示了与OpenAI-o1-1217相当的性能，大大超过了其他模型。在编码算法任务中也观察到类似的趋势，如实时编码台和代码力，其中以推理为中心的模型主导了这些基准。在面向工程的编码任务上，OpenAI-o1-1217在Aider上的性能优于DeepSeek-R1，但在SWE验证上取得了相当的性能。我们相信DeepSeek-R1的工程性能将在下一个版本中得到改善，因为目前相关的RL训练数据的数量仍然非常有限。

3.2. 蒸馏模型评估

模型	艾米2024		数学500	格帕 金刚石	实时代码 长凳	代码的力量
	pass@1	cons@64	pass@1	pass@1	pass@1	等级
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-迷你	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B预览	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-蒸馏-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distilo-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distilo-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distilo-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-蒸馏-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-蒸馏-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

表5|DeepSeek-R1蒸馏模型与其他可比模型的比较
推理相关的b实质。

如表5所示，简单地提取DeepSeek-R1的输出，使高效的DeepSeek-R1-7B（即DeepSeek-R1-提取-Qwen-7B，缩写类似）在整体上优于像GPT-4o-0513这样的非推理模型。DeepSeek-R1-14 B在所有评估指标上都超过了QwQ-32B-预览，而DeepSeek-R1-32 B和DeepSeek-R1-70 B在大多数基准指标上都显著超过了o1-mini。这些结果表明了强大的蒸馏潜力。此外，我们发现将RL应用于这些蒸馏模型会产生显著的进一步收益。我们认为这值得进一步探索，因此这里仅提供这些SFT蒸馏模型的结果。

4. 讨论

4.1. 蒸馏法v.s. “强化性学习”

在第3.2节，我们可以看到，通过提炼DeepSeek-R1，小模型可以取得令人印象深刻的结果。然而，还有一个问题：该模型通过本文讨论的不进行蒸馏的大规模RL训练能否达到相当的性能？

为了回答这个问题，我们使用数学、代码和STEM数据在Qwen-32B-Base上进行了大规模的RL训练，训练超过10K步，从而得到DeepSeek-R1-Zero-Qwen-32B。实验结果见表6，证明了32 B的基础模型，经过了大规模的研究

模型	艾米2024		数学500	GPQA钻石	实时代码库
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B预览	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-0-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distro-Qwen-32B	72.6	83.3	94.3	62.1	57.2

表6|在推理相关基准上的蒸馏模型和RL模型比较。

RL训练，达到了与QwQ-32b-预览相同的性能。然而，从DeepSeek-R1中分离出来的Deen-32B，在所有基准上的表现明显优于DeepSeek-R1- Zero- Qwen-32 B。

因此，我们可以得出两个结论：首先，将更强大的模型提炼成更小的模型会得到很好的结果，而依赖于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，虽然蒸馏策略既是经济的又是有效的，但超越智能的边界可能仍然需要更强大的基础模型和更大规模的强化学习。

4.2. 尝试失败

在开发DeepSeek-R1的早期阶段，我们在开发过程中也遇到了失败和挫折。我们在这里分享了我们的失败经验，以提供见解，但这并不意味着这些方法无法开发有效的推理模型。

过程奖励模型（PRM） PRM是一种合理的方法，可以指导模型更好地解决推理任务(Lightman et al., 2023; Uesato等人, 2022; 王等人, 2023).然而，在实践中，PRM有三个主要的限制，可能阻碍其最终的过程。首先，在一般推理中明确定义一个精细的步骤具有挑战性。其次，确定当前的中间步骤是否正确是一项具有挑战性的任务。使用模型进行自动标注可能不会产生令人满意的结果，而手动标注则不利于扩大规模。第三，一旦引入了基于模型的PRM，它不可避免地会导致奖励性黑客攻击(Gao et al., 2022)，再训练奖励模型需要额外的训练资源，这使整个训练过程变得复杂化。总之，虽然PRM显示了一个良好的能力来重新排序由模型生成的顶级n响应或协助引导搜索(Snell 以及其他2024)，与我们在实验中在大规模强化学习过程中引入的数字计算开销相比，它的优势有限。

蒙特卡罗树搜索（MCTS）的灵感来自于AlphaGo (Silver等人, 2017b)和AlphaZero(帮助-ver等人, 2017a)，我们探索了使用蒙特卡罗树搜索（MCTS）来增强测试时间计算的伸缩性。这种方法涉及到将答案分解成更小的部分，以便允许模型系统地探索解决方案空间。为了促进这一点，我们提示模型生成多个对应于搜索所需的特定推理步骤的多个标签。在训练中，我们首先使用收集到的提示，通过预先训练的价值模型引导的MCTS找到答案。随后，我们使用得到的问答对来训练参与者模型和价值模型，迭代地细化过程。

然而，这种方法在扩大培训规模时遇到了几个挑战。首先，与国际象棋不同，搜索空间是相对明确定义的，令牌生成呈现了一个

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一个最大的扩展限制，但这可能会导致模型陷入本地选择ima之中。其次，价值模型直接影响生成的质量，因为它负责搜索过程的每个步骤。训练一个细粒度的价值模型本质上是困难的，这使得该模型的迭代改进具有挑战性。虽然AlphaGo的核心成功依赖于训练一个价值模型来逐步提高其性能，但由于令牌生成的复杂性，这一原则被证明在我们的设置中难以复制。

综上所述，虽然MCTS在与预先训练的价值模型配对时可以提高推理过程中的性能，但通过自搜索迭代提高模型性能仍然是一个重大挑战。

5.结论、限制性和未来的工作

在这项工作中，我们分享了我们通过强化学习来增强模型重新推理能力的旅程。DeepSeek-R1-Zero代表了一种纯RL方法，而不依赖于冷启动数据，在各种任务中实现了强大的性能。DeepSeek-R1更强大，它利用冷启动数据和迭代RL微调。最终，DeepSeek-R1在一系列任务上实现了可与OpenAI-o1-1217相媲美的性能。

我们进一步探索了对小密集模型的推理能力。我们使用DeepSeek-r1作为教师模型，生成了800K的训练样本，并对几个小型的密集模型进行了微调。结果是有希望的：DeepSeek-R1-提取-Qwen-1.5B在数学基准上优于GPT-4o和Claude-3.5-十四行诗，在AIME上为28.9%，在math上为83.9%。其他密集模型也取得了令人印象深刻的结果，显著优于基于相同底层检查点的其他指令调优模型。

在未来，我们计划投资于DeepSeek-R1的以下方向的研究。

一般能力：目前，DeepSeek-R1在功能调用、多回合、复杂角色扮演和JSON输出等任务上的能力都低于DeepSeek-V3。接下来，我们计划探索可以利用多长时间的CoT来增强这些领域的任务。

语言混合：DeepSeek-R1目前针对中文和英语进行了优化，这可能会导致在处理其他语言中的查询时出现语言混合问题。例如，DeepSeek-R1可能使用英语进行推理和响应，即使查询是英语或中文以外的语言。我们的目标是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很敏感。少镜头的激励不断地降低了它的表现。因此，我们建议用户直接描述问题，并使用azero镜头设置指定输出格式以获得最佳结果。

软件工程任务：由于评估时间长，影响了RL过程的效率，因此大规模的RL在软件工程任务中尚未得到广泛的应用。因此，DeepSeek-R1在软件工程基准上并没有证明比DeepSeek-V3的巨大改进。未来的版本将通过在软件工程数据上实现拒绝抽样或在RL过程中合并异步评估来提高效率来解决这个问题。

参考文献

人工智能@Meta。大羊驼3.1模型卡，2024年。URLhttps://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md。

人类的。克劳德·3.5十四行诗，2024年。人为的网址。[新](#)
5十四行诗。

M. Chen, 特沃雷克, 陈俊, 袁, 德奥利维拉·平托, 卡普兰, 爱德华兹, 布尔达, 约瑟夫, 布罗克曼, A. 雷, 普里, 克鲁格, 彼得罗夫, 赫拉夫, 萨斯特里, P.米什金, B. 陈, 格雷, 莱德, 巴甫洛夫, A. 鲍尔, L.凯泽, M.巴伐利亚人, C. 冬天, P.蒂勒特, F. P., 这样, D. 卡明斯, 普拉珀特, 尚齐斯, 巴恩斯, A. 赫伯特-沃斯, W. H.古斯, A. 尼科尔的变体 A. 佩诺, 特扎克, 唐, 巴布什金, 巴拉吉, 耆因, 桑德斯, C. 海塞 A.卡尔, 莱克, 阿齐亚姆, 米斯拉, 森川, A. 雷德福, 奈特, 布伦戴奇, 穆拉蒂, 梅耶, P.韦林德, B. 麦格鲁 D. 阿莫迪, 麦坎德利什, 苏茨科弗, 还有

W. Zaremba.评估经过代码训练的大型语言模式。CoRR, abs/2107.03374,2021年。URL [..03374](#)。

A. 迪贝 A. Jauhri, A. 潘迪市, A. 硫酸吗啡缓释胶囊剂 A. 达赫勒公司, A. Letman, A. 马瑟 A. 舍尔滕 A. 阳 A. 范先生, 等人。美洲驼3群的模型。arXiv预印本arXiv: 2407.21783,2024年。

Y. Dubois, B. 加兰博西, p.梁, 和T. B. 桥本龙太郎长度控制的时空: 一种简单的方法来消除自动评估器。arXiv预印本arXiv: 2404.04475,2024年。

X.冯, 万, 文, 麦卡, 文, 张, J. 王类阿尔法零树搜索可以指导大型语言模型的解码和训练, 2024年。URL <https://arxiv.org/abs/2407.17179>。

L. Gao, J.舒尔曼, 和J. 希尔顿奖励模型过度优化的缩放定律, 2022年。URL [..10760](#)。

A.盖马, 李, 洪, 洪, A. Devoto, A.曼西诺, 萨克塞纳, 何, 赵, X.杜, 马达尼, 巴拉勒, 麦克哈迪, 哈里斯, 卡杜尔, 范克里肯, 和 P. Minervini .我们用mmlu了吗? CoRR, abs/2406.04127,2024年。URL<https://doi.org/10.48550/arXiv.2406.04127>。

谷歌我们的下一代型号: 双子座1.5号, 2024年。URL<https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024>。




何Y., 李Y., 刘Y., 谭勇, 王Y., 黄Y., 布Y., 郭Y., C. 胡先生: B. 郑, 等人。简单: 大型语言的中文事实性评估。arXiv preprintarXiv:2411.07140, 2024。

D. Hendrycks, C. 伯恩斯, 巴萨特, A. Zou, M. Mazeika, D. 宋, 和J.斯坦哈特。测量大量的多任务语言理解。arXiv预印本arXiv: 2009.03300,2020年。

黄Y., 白, 朱, 张, 张, 苏, 刘、吕, 张, 磊等。C-Eval: amulti级多学科中国基础模型评估套件。arXiv preprintarXiv:2305.08322, 2023。

N. Jain, 韩韩, A. 顾, 李, 燕, 张, 王, A. 太阳能-莱萨马, 森和斯托伊卡。大型语言模型的整体和无污染评估。CoRR, abs/2403.07974,2024年。URL [..48550/ arXiv .2403.07974](#)。

- S. 奎师那, S. 奎师那, A. 莫哈纳尼, 施瓦茨, A. 流浪者, 乌帕德海伊和法鲁基先生。事实、获取和推理: 检索增强生成的统一版本。CoRR, abs/2409.12941, 2024。doi: 10.48550/ARXIV.2409.12941。URL <https://arxiv.org/abs/2409.12941> .
- A. 库马尔, 诉庄, 阿加瓦尔, 苏, 科雷耶斯, A. 辛格, 鲍姆利, 伊克巴尔, C. 毕晓普, R. Roelofs, 等人。通过强化学习来训练语言模型的自我纠正。arXiv预印本arXiv: 2409.12917, 2024年。
- 李H., 张, 国藤, 杨, 赵, 龚, 段, T. 鲍尔温苹果CMMLU: 测量大量的多任务语言理解。arXiv预印本arXiv: 2306.09212, 2023年。
- T. 李, w.l. 蒋, 弗里克, 邓拉普, 吴, 朱, 冈萨雷斯, 和我。Stoica 从众包数据到高质量的基准: 竞技场硬和基准建设者的管道。arXiv预印本arXiv: 2406.11939, 2024年。
- H. 莱特曼, 诉科萨拉朱, Y. 布尔达, H. 爱德华兹, B. 贝克, T. 李, J. 莱克, J. 舒尔曼, I. 苏斯克弗, 和K. 科布的变体让我们逐步验证一下。arXiv预印本arXiv: 2305.20050, 2023年。
- B.Y. 林。《ZeroEval: 评估语言模型的统一框架》, 2024年7月。URL <https://wild-eval.com/WildEval/ZeroEval> .
- “妈妈”。美国数学邀请考试。在美国数学邀请赛-AIME2024, 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。学习与llms的推理, 2024b。URL <https://openai.com/index/learn-in-reasoning-with-llms/> 表示理由。
- OpenAI。介绍SimpleQA, 2024c。URL <https://openai.com/index/introducing-simpleqa/> .
- OpenAI。我们发布了一个经过人工验证的软件板凳子集, 2024d。URL <https://openai.com/index/introducing-swe-bench-verified/> .
- Qwen .Qwq: 深入反思未知的边界, 2024a。URL <https://qwenlm.github.io/blog/qwq-32b-preview/> .
- 奎温。Qwen2.5: 一个基础模型, 2024b。URL <https://qwenlm.github.io/blog/qwen-2.5/> .
- D. 驾驭 B.L. Hou, A. 斯蒂克兰, 佩蒂, 庞, 迪拉尼, 迈克尔, 和S. R. 鲍曼。GPQA: 一个研究生水平的谷歌验证的q和一个基准。arXiv预印本arXiv: 2311.12022, 2023年。
- 邵, 王, 朱, 徐, 宋, 张, 李, 吴, D. 郭志不清。深度数学: 在开放语言模型中推动数学推理的极限。arXiv preprint arXiv: 2402.03300, 2024。
- D. 西尔弗, 休伯特, 施里特维泽, 安东诺格卢, 赖, A. 盖兹, 兰开特, 西弗雷, D. 库马兰, T. 格雷佩尔, T. P. 利利克拉普, K. 西蒙尼扬, 和 D. 哈的。用一般的强化学习算法自行掌握国际象棋和小国际象棋。CoRR, abs/1712.01815, 2017a。URL <https://arxiv.org/abs/1712.01815> .

- D. 西尔弗, 施里特维泽, 西蒙尼扬, 安东诺格卢, A. 黄 A. 盖兹, 休伯特, 贝克, 赖, A. 博尔顿, 陈, 李乐, 辉, 谢, 谢, 范登, 格雷佩尔, 和 D. 哈萨比斯。掌握了没有人类知识的围棋游戏。Nat., 550年(7676): 354-359年, 2017年b. doi: 10.1038/自然24270. URL .. <https://doi.org/10.1038/nature24270>.
- C. 斯奈尔, 李, 徐k, 和 A. 库马尔公司。最优缩放最测试时间计算比缩放模型参数更有效。URL ..033  14.
- T. Trinh, 吴, 乐, H.H和T.龙。解决人类演示演示解决解决奥没有人人类演示。自然》, 2024. doi: 10.1038/s41586-023-06747-5.
- J. Uesato, 库什曼, 库马尔, 宋, 西格尔, 王, 克雷斯韦尔, 欧文, 和 I.希金斯。用基于过程和结果的反馈来解决数学单词问题。arXiv预印本arXiv: 2211.14275,2022年。 -
- 王P., L.李, 邵子, R.徐, D. 李戴勇, D. 陈, 吴, 祖穗。数学牧人: 在数学推理中llms的无标签逐步验证器。arXiv预印本arXiv: 2312.08935,2023年。 -
- Wang X., J. Wei, D. 舒尔曼, 勒, 志志, 纳朗, A. 杂, 和 D. 周自一致性改进了语言模型中的思维推理链。arXiv preprintarXiv:2203.11171, 2022.
- 王Y., 马九, 张, 倪, A. 钱德拉, 郭, 任, A. 阿鲁拉杰, 何泽, 李, 库, 王, A. 庄, 范, 岳, 陈晓。Mmlu- pro: 一个更健壮和更具挑战性的多任务语言理解基准测试。CoRR, abs/2406.01574,2024年。URL ..48550/ arXiv .2406.01574 . 
- C.夏a, 邓, 邓恩和L. 张的。无代理: 揭开基于llm的软件工程代理的神秘面纱。arXiv预印本, 2024年。 -
- H.辛, 仁, 宋, 邵, 赵, 王, B. 刘, 张, 陆晓, 杜, 高, 朱, D. 杨、郭台铭、吴、罗、 C. 鲁安深度搜索-证明程序-v1.5: 利用证明辅助反馈进行强化学习和蒙特-卡洛树搜索, 2024年。URL ..08152 . 
- 周J., 陆, 米什拉, 梵天, 巴苏, 漆, 周, 侯伦。对大型语言模型的指令-后续评估。arXiv预印本arXiv: 2311.07911,2023年。 -

附录

A. 贡献和致谢

核心贡献者

郭大亚
杨德健
张俊晓松
张若雨
徐润欣
朱启豪
马世荣
王佩一
小比
小康张兴凯于
于武
Z.F.吴
郭志斌
邵卓淑丽志宏
高子怡

贡献者

刘爱心
冰雪
王子王波超武
北峰
陆成达
成钢赵成琪邓
冲阮达迈戴
陈德丽
纪洁洁
李二杭
林方云
戴福聪
Fuli Luo*
光博郝官厅陈
国维李
张H.
徐汉伟
洪辉丁华祖高
许曲

李慧
李建忠
景昌陈泾阳源金
兔
李俊俊
J.L.材
倪佳琪
吉安梁金
陈凯东凯
胡*
高凯歌
康管可心黄快
玉
王莱恩
乐从张良昭
利通王立悦
张磊徐
夏乐毅
明川张明华张
唐明辉
周明旭
李梦
王妙君
李明明
宁天
黄盼盼
彭张
前程王勤于陈
杜秋实
Ruiqi Ge*
张丽思潘
王润基
Chen
R.L.金

陈如意
尚浩鲁
上岩周山黄陈胜
丰
王施余
于水平
周舒婷潘
S.S.李
双周少卿吴胜
风叶涛云
田培
孙天宇
T.王
王定曾文刘
文山亮文君高
Wenqin Yu*
张文涛肖
魏安
小东刘晓涵王
小康陈晓涛聂
辛程
刘欣
谢欣
星超刘新余
杨新源李
学诚苏许恒林
X.Q.列支敦士登
金湘月
沈小京
陈小莎
孙孝文
潇湘王新南之歌
信义周显祖
王新霞山
Y.K.列支敦士登
Y.Q.王

Y.X.魏
杨张彦徐姚
李
赵耀
耀凤孙耀辉王
毅宇
伊超张一凡史
宜良雄英和
朴义士
白松王一轩谭
益阳Ma*刘
怡源
永强郭元欧
玉端王跃宫
玉恒邹玉佳
云帆熊玉祥罗
尤玉祥
刘宇轩
周玉阳
Y.X.朱
黄耀辉李
义正
朱云贤马英
汤
玉昆扎玉
婷燕
Z.Z.任
任泽辉
张丽莎哲夫
徐哲安
谢天达
张哲文豪正言
志诚马志刚
严志宇吴
顾子辉

朱紫嘉
刘子军*
李子林
谢紫薇
资阳宋正正
潘

黄真
徐志鹏
张中字

在每个角色中，作者都按名字的字母顺序列出。标有*的名字表示已经离开我们团队的人。

AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩



Zero

AI音乐创作人

水墨动漫联盟创始人

百脑共创联合创始人

人工智能产业链联盟创始人

中关村人才协会秘书长助理

河北北大企业家分会秘书长

墨攻星辰智能科技有限公司CEO

河北清华发展研究院智能机器人中心线上负责人

中关村人才协会数字体育与电子竞技专委会秘书长助理



主要业务:AI商业化答疑及课程应用场景探索, 各类AI产品学习手册, 答疑及课程



欢迎扫码交流

提供: 学习手册/工具/资源链接/商业化案例/
行业报告/行业最新资讯及动态



人工智能产业链联盟创始人

邀请你加入星球, 一起学习

人工智能产业链联盟报 告库



星主: 人工智能产业链联盟创始人

每天仅需0.5元, 即可拥有以下福利!
每周更新各类机构的最新研究成果。立志将人工智能产业链联盟打造成市面上最全的AI研究资料库, 覆盖券商、产业公司、科研院所等...

知识星球

微信扫码加入星球 ▶

