

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

论文解读DeepSeek-R1



时空猫的问题盒

北京大学 微电子学与固体电子学硕士

关注他

4 人赞同了该文章

收起

文章推荐

文章名称: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

文章链接: [github.com/deepseek-ai/...](https://github.com/deepseek-ai/DeepSeek-R1)

hf链接: [huggingface.co/deepseek...](https://huggingface.co/deepseek-ai/DeepSeek-R1)

大家好, 今天我们来讨论一篇名为《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》的论文。这篇论文的核心是探讨如何通过强化学习 (Reinforcement Learning, RL) 来提升大型语言模型 (Large Language Models, LLMs) 的推理能力。让我们一起来看看这篇论文的摘要部分, 并尽量用通俗易懂的方式进行解释。

摘要

首先, 论文介绍了两个一代推理模型: **DeepSeek-R1-Zero**和**DeepSeek-R1**。DeepSeek-R1-Zero是一个通过大规模强化学习 (Reinforcement Learning, RL) 训练的模型, 而不是通过监督微调 (Supervised Fine-Tuning, SFT) 作为预备步骤。这意味着, DeepSeek-R1-Zero在没有经过传统的监督学习训练的情况下, 通过强化学习自然地发展出了许多强大且引人入胜的推理行为。简单来说, 这个模型通过不断尝试和调整, 自己学会了如何更好地进行推理。

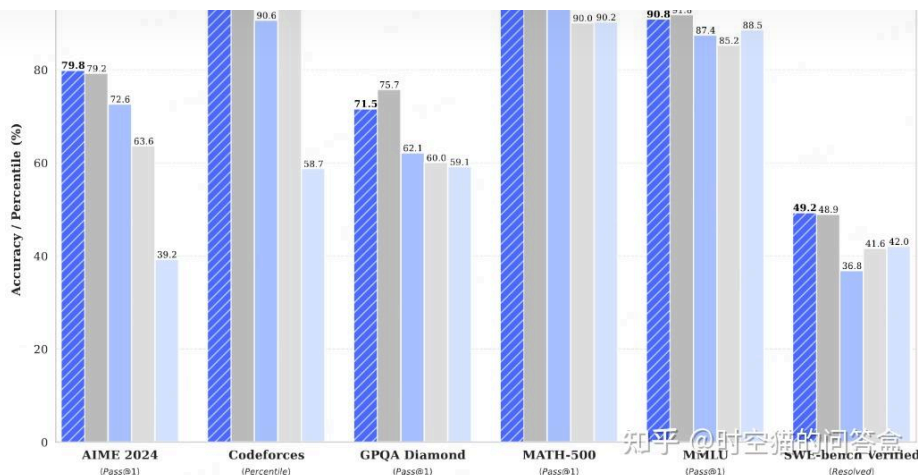
然而, DeepSeek-R1-Zero在实际应用中遇到了一些问题, 比如文本的可读性不佳, 以及语言混乱。为了解决这些问题, 并进一步提升推理性能, 作者引入了**DeepSeek-R1模型⁺**。DeepSeek-R1采用了多阶段训练和冷启动数据 (cold-start data) 在强化学习之前进行准备。这种方法帮助模型在推理任务上的表现达到了与OpenAI的o1-1217模型相当的水平。

最后, 为了支持研究社区, 作者开源了DeepSeek-R1-Zero和DeepSeek-R1模型, 以及六个基于Qwen和Llama的密集模型 (1.5B, 7B, 8B, 14B, 32B, 70B)。这些模型的参数量不同, 从1.5亿到70亿, 为不同的研究和应用提供了多样化的选择。

1 Introduction

在过去的几年里, 大型语言模型 (Large Language Models, LLMs) 经历了快速的迭代和演变 (Anthropic, 2024; Google, 2024; OpenAI, 2024a), 逐步缩小了与人工通用智能 (Artificial General Intelligence, AGI) 之间的差距。这表明LLMs在智能水平上的提升, 越来越接近于AGI的能力。

近期, 后训练 (post-training) 成为完整训练流水线的重要组成部分。研究表明, 后训练能够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⁺** (pre-training) 所需的计算资源较少。在推理能力的背景下, OpenAI的o1系列模型、强化学习 (Reinforcement Learning, RL) 以及搜索算法如**蒙特卡洛树搜索⁺** (Monte Carlo Tree Search) 和梯度搜索 (Beam Search) 等方法都有所应用。然而, 这些方法尚未达到OpenAI o1系列模型在通用推理性能上的水平。



本文首次尝试通过纯粹的强化学习框架来提升语言模型的推理能力。在训练过程中，DeepSeek-R1-Zero自然地展现出了许多强大且引人入胜的推理行为。经过数千步的强化学习训练，DeepSeek-R1-Zero在推理基准测试中表现出色。例如，其在AIME 2024⁺上的Pass@1得分从15.6%提升到71.0%，通过多数投票，得分进一步提高到86.7%，与OpenAI-o1-0912的表现相当。

然而，DeepSeek-R1-Zero在实际应用中遇到了一些问题，如文本可读性差和语言混乱。为了解决这些问题并进一步提升推理性能，作者引入了DeepSeek-R1模型。DeepSeek-R1采用了少量的冷启动数据（cold-start data）和多阶段训练流水线。具体来说，首先收集数千条冷启动数据用于微调DeepSeek-V3-Base模型，然后进行类似DeepSeek-R1-Zero的推理导向强化学习。在RL过程接近收敛时，通过对RL检查点进行拒绝采样（rejection sampling）结合DeepSeek-V3在写作、事实问答（factual QA）和自我认知等领域的监督数据，重新训练DeepSeek-V3-Base模型。在微调新数据后，检查点经过额外的RL过程，考虑所有场景的提示。经过这些步骤，获得的检查点被称为DeepSeek-R1，其性能与OpenAI-o1-1217相当。

进一步探索从DeepSeek-R1到更小密集模型的蒸馏。使用Qwen2.5系列。值得注意的是，我们的蒸馏14B模型在推理基准测试中大幅超越了开源的QwQ-32B-Preview（Qwen, 2024a），而蒸馏的32B和70B模型在密集模型中的推理基准测试中创造了新纪录。

在 1.1. Contributions部分，论文详细介绍了其贡献：

1. 后训练：大规模基础模型的强化学习

- 直接将强化学习应用于基础模型，而不依赖于监督微调作为预备步骤。这种方法使模型能够探索链式思维⁺（Chain-of-Thought, CoT）来解决复杂问题，从而开发出DeepSeek-R1-Zero。DeepSeek-R1-Zero展示了自我验证、反思和生成链式思维的能力，这是研究社区的重要里程碑。值得注意的是，这是第一次开放研究验证，LLMs的推理能力可以通过纯粹的RL来激励，而无需SFT。这一突破为未来的进步铺平了道路。

2. 蒸馏：小型模型也能强大

- 证明了大型模型的推理模式可以被蒸馏到小型模型中，从而在小型模型上通过RL发现的推理模式表现更佳。开源的DeepSeek-R1及其API将为研究社区提供在未来蒸馏更好的小型模型的基础。

1.2. Summary of Evaluation Results

- 推理任务：DeepSeek-R1在AIME 2024上的Pass@1得分为79.8%，略高于OpenAI-o1-1217。在MATH-500上，其得分为97.3%，与OpenAI-o1-1217持平，显著超过其他模型。在编程相关任务中，DeepSeek-R1在Codeforces上达到了2,029 Elo评分，超过了96.3%的人类参赛者。在工程相关任务中，DeepSeek-R1略优于DeepSeek-V3⁺，有助于开发者在现实任务中。

- 知识：在MMLU、MMLU-Pro和GPQA Diamond⁺等基准测试中，DeepSeekR1取得了出色的成绩，显著超过DeepSeek-V3，分别为90.8%、84.0%和71.5%。虽然其表现略低于OpenAI-o1-1217，但DeepSeek-R1超过其他闭源模型，展示了其在教育任务中的竞争力。在事实基准SimpleQA上，DeepSeek-R1超过DeepSeek-V3，展示了其处理事实查询的能力。

考试导向查询中的强大智能处理能力。此外，DeepSeek-R1在需要长上下文理解的任务中表现出色，显著超过DeepSeek-V3。

通过这些贡献和评估结果，论文展示了通过创新的训练方法和蒸馏技术⁺，如何显著提升大型语言模型的推理能力，并为研究社区提供了丰富的资源和基础。

2 Approach

在 2. Approach部分，论文详细介绍了提升大型语言模型推理能力的方法。首先，我们看到 2.1. Overview部分，研究者们指出，传统方法依赖大量监督数据来提升模型性能。然而，本研究通过大规模强化学习（Reinforcement Learning, RL）展示了可以在没有监督微调（Supervised Fine-Tuning, SFT）的情况下显著提升推理能力。研究者们提出了两个模型：DeepSeek-R1-Zero 和 DeepSeek-R1，以及将推理能力蒸馏到小型模型的方法。

在 2.2. DeepSeek-R1-Zero: Reinforcement Learning on the Base Model部分，研究者们探讨了直接在基础模型上应用强化学习的可能性。他们使用了一种称为Group Relative Policy Optimization（GRP优化）的算法，这种算法通过组分数来估计基线，而不是使用与策略模型同等大小的批评模型。这种方法可以节省RL训练的成本。此外，研究者们采用了基于规则的奖励系统，包括准确性奖励和格式奖励，以指导模型生成正确且格式化的推理过程。

在 2.2.3. Training Template部分，研究者们设计了一个简单的模板，指导模型生成推理过程和最终答案。这种设计允许观察模型在RL过程中的自然进化，而不受内容偏见的影响。

在 2.2.4. Performance, Self-evolution Process and Aha Moment of DeepSeek-R1-Zero部分，研究者们展示了DeepSeek-R1-Zero在AIME 2024基准测试中的性能提升，从15.6%的初始 Pass@1得分增加到71.0%。通过多数投票，性能进一步提高到86.7%。这表明模型在没有监督微调的情况下，通过RL自然发展出强大的推理能力。此外，模型在自我演化过程中展示了反思和探索替代解决方案的能力，这些能力是在RL过程中自发出现的。

然而，DeepSeek-R1-Zero也存在一些问题，如文本可读性差和语言混乱。为了解决这些问题，研究者们引入了 2.3. DeepSeek-R1: Reinforcement Learning with Cold Start部分的方法。通过使用少量的冷启动数据和多阶段训练流水线，DeepSeek-R1模型在推理任务上的表现得到了显著提升，并且生成的推理过程更加可读。

在 2.3.1. Cold Start部分，研究者们收集了大量的冷启动数据，用于微调DeepSeek-V3-Base模型，作为RL的初始演员。这些数据的优点包括可读性和潜力，通过设计可读的模式和过滤不可读的输出来实现。

在 2.3.2. Reasoning-oriented Reinforcement Learning部分，研究者们应用了与DeepSeek-R1-Zero相同的大规模RL训练过程，但引入了语言一致性奖励来减少语言混乱。最终奖励是推理任务准确性和语言一致性奖励的直接求和。

在 2.3.3. Rejection Sampling and Supervised Fine-Tuning部分，研究者们使用RL收敛后的检查点收集SFT数据，以增强模型在写作、角色扮演和其他通用任务上的能力。他们收集了大约600k的推理相关训练样本和200k的非推理训练样本。

在 2.3.4. Reinforcement Learning for all Scenarios部分，研究者们实施了第二阶段的RL，以进一步提升模型的有用性和无害性，同时保持其推理能力。他们使用奖励信号和多样化的提示分布来训练模型。

最后，在 2.4. Distillation: Empower Small Models with Reasoning Capability部分，研究者们展示了如何通过简单的SFT方法将推理能力蒸馏到小型模型中，如Qwen和Llama系列模型。这些小型模型在推理基准测试中表现出色，展示了蒸馏技术的有效性。

通过这些方法，论文展示了通过创新的训练方法和蒸馏技术，如何显著提升大型语言模型的推理能力，并为研究社区提供了丰富的资源和基础。

3 Experiment

Understanding)、MMLU-Redux、MMLU-Pro、C-Eval、CMMLU、IFEval、FRAMES、GPQA Diamond、SimpleQA、SWE-Bench Verified和Arena-Hard。这些基准测试利用GPT-4-Turbo-1106作为评估的对比标准。为了避免长度偏差，研究者们只将最终的摘要结果输入到评估中。对于蒸馏模型⁺，研究者在AIME 2024、MATH-500、GPQA Diamond、Codeforces和LiveCodeBench等基准测试中报告了代表性结果。

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

Table 4 | Comparison between DeepSeek-R1 and other representative models.

在评估提示方面，研究者们遵循DeepSeek-V3的设置，使用simpleevals框架中的提示来评估标准基准测试，如MMLU、DROP、GPQA Diamond和SimpleQA。对于MMLU-Redux，研究者们采用了Zero-Eval提示格式。AIDER相关的基准测试使用“diff”格式进行测量。DeepSeek-R1的输出在每个基准测试中被限制在最大32,768个token。

在基线比较方面，研究者们对多个强大的基线模型进行了全面评估，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，研究者们根据官方报告来报告其性能。对于蒸馏模型，研究者们还与开源模型QwQ-32B-Preview进行了比较。

在生成设置方面，所有模型的最大生成长度被设置为32,768个token。对于需要采样的基准测试，研究者们使用温度为0.6、top-p值为0.95，并为每个查询生成64个响应来估计Pass@1。

3_1 DeepSeek-R1 Evaluation

在 3.1. DeepSeek-R1 Evaluation部分，研究者们展示了DeepSeek-R1在教育导向的知识基准测试中的优越性能，如MMLU、MMLU-Pro和GPQA Diamond，相比于DeepSeek-V3。这种改进主要归因于STEM相关问题的准确性提高，这是通过大规模强化学习 (RL) 实现的。此外，DeepSeek-R1在FRAMES基准测试中表现出色，这是一个依赖长上下文的QA任务，展示了其强大的文档分析能力。这突显了推理模型在AI驱动搜索和数据分析任务中的潜力。在事实基准SimpleQA上，DeepSeek-R1超过了DeepSeek-V3，展示了其处理事实查询的能力。类似的趋势也观察到，OpenAI-o1在这个基准测试中超过了GPT-4o。然而，DeepSeek-R1在中文SimpleQA基准测试中表现不如DeepSeek-V3，主要是因为其在安全RL后倾向于拒绝回答某些查询。没有安全RL，DeepSeek-R1可以实现超过70%的准确性。

DeepSeek-R1还在IF-Eval基准测试中取得了令人印象深刻的结果，这是一个设计用于评估模型遵循格式指令能力的基准测试。这些改进可以归因于最终阶段的监督微调 (SFT) 和RL训练中包含的指令遵循数据。此外，令人瞩目的表现也观察到在AlpacaEval2.0和ArenaHard上，表明DeepSeek-R1在写作任务和开放域问答中的优势。其显著超过DeepSeek-V3的表现强调了大规模

表明DeepSeek-R1在GPT基础评估中避免引入长度偏差，进一步巩固了其在多个任务中的稳健性。

在数学任务上，DeepSeek-R1的表现与OpenAI-o1-1217持平，大幅超过其他模型。在编码算法任务上，如LiveCodeBench和Codeforces，推理导向的模型主导了这些基准测试。在工程导向的编码任务上，OpenAI-o1-1217在Aider上超过了DeepSeek-R1，但在SWE Verified上的表现相当。我们相信DeepSeek-R1的工程性能将在下一个版本中得到改善，因为当前相关的RL训练数据量仍然非常有限。

3_2 Distilled Model Evaluation

在 3.2. Distilled Model Evaluation部分，研究者们展示了仅仅蒸馏DeepSeek-R1的输出就可以使得高效的DeepSeekR1-7B（即DeepSeek-R1-Distill-Qwen-7B，以下简称）在所有评估指标上超过非推理模型如GPT-4o-0513。DeepSeek-R1-14B在所有评估指标上超过了QwQ-32BPreview，而DeepSeek-R1-32B和DeepSeek-R1-70B在大多数基准测试上显著超过了o1-mini。这些结果展示了蒸馏的强大潜力。此外，研究者们发现将RL应用于这些蒸馏模型可以获得显著的进一步提升。他们认为这值得进一步探索，因此在这里只呈现了简单SFT蒸馏模型的结果。

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	1820
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	72.6	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	86.7	94.5	65.2	57.5	1633

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

通过这些评估，论文展示了通过创新的训练方法和蒸馏技术，如何显著提升大型语言模型的推理能力，并为研究社区提供了丰富的资源和基础。

4 Discussion

在 4. Discussion部分，论文探讨了蒸馏（Distillation）与强化学习（Reinforcement Learning, RL）在提升大型语言模型推理能力方面的比较，以及在开发DeepSeek-R1过程中遇到的一些挑战和失败尝试。

4_1 蒸馏与强化学习的比较

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	72.6	83.3	94.3	62.1	57.2

Table 6 | Comparison of distilled and RL Models on Reasoning-Related Benchmarks.

论文中的表6（Table 6）比较了蒸馏模型和基于RL的模型在推理相关基准测试上的表现。在第3.2节中，我们看到通过蒸馏DeepSeek-R1，小型模型可以取得令人印象深刻的结果。然而，仍然存在一个问题：是否可以通过论文中讨论的大规模RL训练，而不进行蒸馏，实现相似的性能？

为了回答这个问题，研究者在数学、编码和STEM数据上对Qwen-32B-Base进行了大规模RL训练，训练步数超过10,000步，得到了DeepSeek-R1-Zero-Qwen-32B模型。实验结果（如图6所示）表明，经过大规模RL训练的32B基础模型，其性能与QwQ-32B-Preview相当。然而，从DeepSeek-R1蒸馏出来的DeepSeek-R1Distill-Qwen-32B在所有基准测试中都显著优于DeepSeek-R1-Zero-Qwen-32B。因此，我们可以得出两个结论：

- 虽然蒸馏策略既经济又有效，但超越智能边界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，研究者们也遇到了失败和挫折。他们分享了这些失败经验，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 (Process Reward Model, PRM)

在实践中，PRM存在三个主要限制，可能会阻碍其最终成功：

- 定义细粒度步骤困难：在一般推理中，难以明确定义细粒度步骤。
- 判断中间步骤正确性困难：确定当前中间步骤是否正确是一个具有挑战性的任务。自动标注使用模型可能不会得到满意的结果，而手动标注不利于扩展。
- 引入模型奖励可能导致奖励作弊：一旦引入基于模型的PRM，就不可避免地会导致奖励作弊，重新训练奖励模型需要额外的训练资源，并复杂化整个训练流水线。因此，虽然PRM在重排模型生成的顶N响应或辅助引导搜索方面表现出良好能力，但其优势有限，与在大规模强化学习过程中引入的额外计算开销相比。

蒙特卡罗*树搜索 (Monte Carlo Tree Search, MCTS) 和AlphaZero

研究者们探索了使用蒙特卡罗树搜索 (MCTS) 来提高测试时间计算可扩展性。这种方法涉及将答案分解为较小的部分，以便模型可以系统地探索解决方案空间。为了实现这一点，他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。在训练中，首先使用收集的提示通过由预训练值模型引导的MCTS找到答案。然后，使用生成的问题-答案对训练行动模型和值模型，并迭代改进过程。

然而，这种方法在扩展训练时遇到了几个挑战：

- 搜索空间更大：与国际象棋不同，其搜索空间相对较好定义，令牌生成的搜索空间呈指数增长。为了解决这个问题，他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。
- 值模型直接影响生成质量：值模型直接影响每个搜索步骤的质量。训练一个精细的值模型本身就很困难，这使得模型难以迭代改进。虽然AlphaGo的核心成功依赖于训练一个值模型来逐步提高其性能，但这一原则在我们的设置中难以复制，因为令牌生成的复杂性。

总结来说，虽然MCTS可以在配备预训练值模型的情况下提高推理性能，但通过自我搜索迭代提升模型性能仍然是一个挑战。

通过这些讨论，论文不仅展示了蒸馏和强化学习在提升大型语言模型推理能力方面的优缺点，还分享了在研究过程中遇到的挑战和失败经验，为未来的研究提供了宝贵的经验教训。

5 Conclusion, Limitation, and Future Work

在本文中，我们分享了通过强化学习 (Reinforcement Learning, RL) 提升模型推理能力的旅程。DeepSeek-R1-Zero代表了一种纯粹的RL方法，不依赖冷启动数据，其在各种任务上表现出色。DeepSeek-R1更为强大，它结合了冷启动数据和迭代的RL微调。最终，DeepSeek-R1在多个任务上的表现与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，并微调多个小型密集模型。结果令人鼓舞：DeepSeek-R1-Distill-Qwen-1.5B在数学基准测试上超过了GPT-4o和Claude-3.5-Sonnet，分别在AIME上达到28.9%，在MATH上达到83.9%。其他密集模型也取得了令人瞩目的成绩，显著超过基于相同底层检查点的其他指令微调模型。

在未来，我们计划在以下方向上投入研究以进一步提升DeepSeek-R1：

- 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不如DeepSeekV3。未来，我们计划探索如何利用长链式思维 (Chain-of-Thought,


现语言混合问题。例如，DeepSeek-R1可能在推理和响应中使用英文，即使查询不是英文或中文。我们计划在未来更新中解决这一限制。

3. 提示工程：在评估DeepSeek-R1时，我们观察到它对提示非常敏感。少量提示一致地降低了其性能。因此，我们建议用户直接描述问题并使用零示例设置指定输出格式，以获得最佳结果。
4. 软件工程任务：由于评估时间较长，影响RL过程的效率，大规模RL尚未广泛应用于软件工程任务。因此，DeepSeek-R1在软件工程基准测试上并没有显著超过DeepSeek-V3。未来版本将通过拒绝采样软件工程数据或在RL过程中引入异步评估来改善效率，从而解决这一问题。

通过这些探索和计划，我们希望在未来的版本中进一步提升DeepSeek-R1的能力，使其在更广泛的任务中表现出色。

编辑于 2025-01-21 15:37 · IP 属地上海

内容所属专栏

 **LLMs**
常见的LLMs

订阅专栏

LLM（大型语言模型）



理性发言，友善互动



发布



还没有评论，发表第一个评论吧

推荐阅读

Interventional Few-Shot Learning

Intongyi Yao^{1,2}, Hanwang Zhang¹, Qianru Sun¹, Xian-Sheng Hua¹
¹Singapore Technological University, ²Singapore Management University, ³Alibaba Group

【Causal Inference论文阅读笔记】Interventional Few-...

套娃的套娃

强化学习论文阅读笔记（1）Deconfounded Value...

写文章的想法主要是做个笔记，因为之前看完论文过了一周就忘掉了，感觉没什么作用。而在知乎上写着说不定有人能看到笔记，可以一起交流学习，并且在这里做笔记可以强迫自己写的容易懂，...
君莫笑

近两年FPN论文

1、AugFPN: Improved Scale Feature Learning for Object Detection的论文，在论文中讨论问题（1）不同层之间的gap，直接对不同层...
Fight...