



首发于  
LLM应用技术指北

# DeepSeek-R1论文速读



Meta

关注LLM4Code和LLM infra

关注他

158 人赞同了该文章

春节将至，DeepSeek又出王炸！DeepSeek-R1系列重磅开源。本文对其技术报告做简单解读。

话不多说，show me the benchmark。从各个高难度benchmark结果来看，DeepSeek-R1已经比肩OpenAI-o1-1217，妥妥的第一梯队推理模型。同时蒸馏Qwen2.5-32B而来的DeepSeek-R1-32B也取得非常惊艳的效果，和OpenAI-o1-mini旗鼓相当。

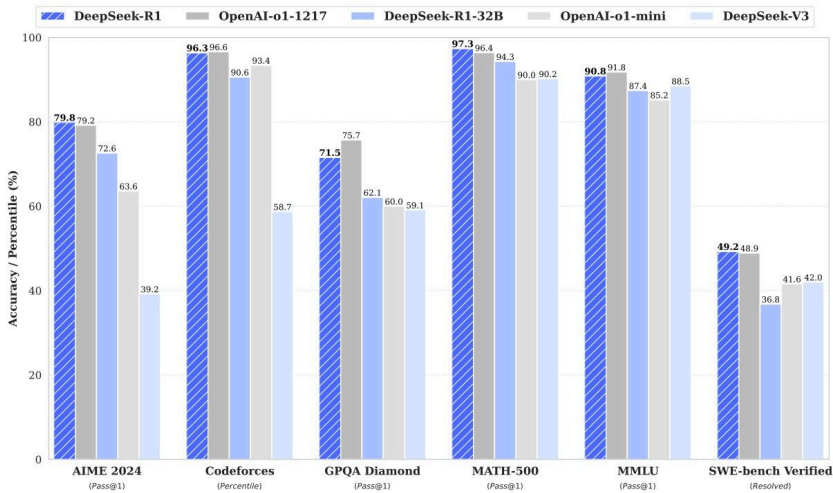


Figure 1 | Benchmark performance of DeepSeek-R1.

知乎 @Meta



天下苦SFT久矣，已有的公开研究无一例外都是采用SFT+RL的方式，首先需要大量的SFT数据进行指令微调。而DeepSeek不走寻常路。他们发现即使不使用SFT，也可以通过大规模强化学习（RL）显著提高推理能力。此外，通过包含少量冷启动数据进行SFT就可以进一步提高性能。

本文的几个主要贡献：

- DeepSeek-R1-Zero：不用SFT直接进行RL，也能取得不错的效果。
- DeepSeek-R1：加入少量（数千量级）CoT数据进行SFT作为冷启动，然后再进行RL，可以取得更优的性能。同时回答更符合人类偏好。
- 用DeepSeek-R1的样例去蒸馏小模型，能取得惊人的效果。

下面会逐一介绍。

## DeepSeek-R1-Zero

直接从DeepSeek-V3-Base开搞，仍然用DeepSeek独家定制的GRPO，使用如下平平无奇的PE模版。

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.

知乎 @Meta

RM方面，考虑到是推理任务，没有训练常规的稠密奖励模型，而是采用了两种奖励方式结合：

就是这么看起来似乎暴力又简单的方法，效果却出奇地好。

看起来随着训练步数的增加，性能稳步提升，达到和OpenAI-o1-0912接近的水平。

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

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

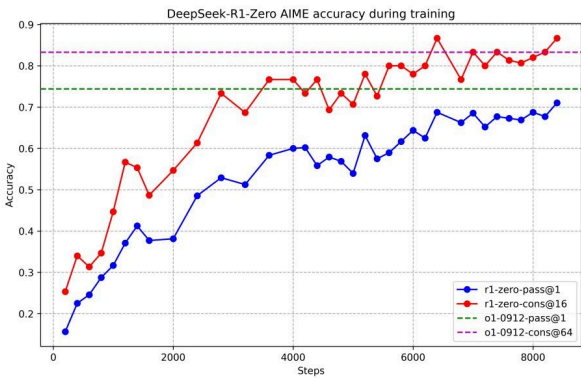


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

并且观察到了明显的“进化”现象，随着训练步数的增加，输出平均长度也在增加。意味着LLM<sup>+</sup>似乎自己已经潜移默化地学会了进行更多的思考和推理，达到更好的效果。

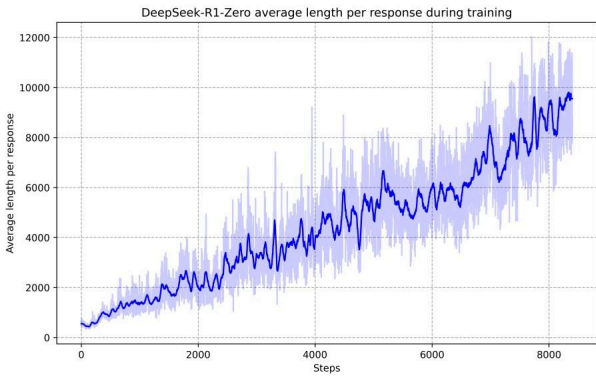



Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

甚至出现了自主的“Aha Moment”情况，突然就能开始反思。DeepSeek顿悟了，DeepSeek的兄弟们也顿悟了，我似乎也顿悟了。怎么只是平平无奇的奖励信号，就能让它学会这么多？还得是RL！！

Response: <think>  
To solve the equation  $\sqrt{a - \sqrt{a + x}} = x$ , let's start by squaring both ...  
 $(\sqrt{a - \sqrt{a + x}})^2 = x^2 \implies a - \sqrt{a + x} = x^2$ .  
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: ...  
...

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. 

真就这么完美？SFT完全不必要了？显然也不是，DeepSeek的兄弟们也发现了一些问题，比如，DeepSeek-R1-Zero 生成的答案**可读性**相对差、存在混合语言输出情况（这个似乎QwQ也比较明显）。为了让模型说人话，还是得加点SFT，这就到DeepSeek-R1的舞台了。

## DeepSeek-R1

DeepSeek-R1-Zero已经证明了，完全不进行SFT, 直接RL就能显著提升LLM的推理能力；但是同时输出的可读性、混合语言输出问题还是老大难。可别忘了SFT不就是为了遵循指令，让LLM模仿说人话吗？那把SFT阶段再加上不就得了。既然完全不SFT也能有非常好的效果，那少加一点是不是就能让LLM学会说人话了，同时推理能力也能进一步提升呢？DeepSeek-R1采用如下4个阶段，又把能力进一步加强。

### • 少量数据冷启动

采用一定的手段收集少量高质量数据：比如对于长CoT数据，使用**few-shot**<sup>+</sup>，直接提示DeepSeek-R1-Zero通过反思和验证生成详细答案，然后通过人工注释者的后处理来细化结果。

总共收集了数千个样本，相比完全不用SFT，收集的样本进行SFT，可以显著增强可读性；同时后续的实验也证明了通过少量数据冷启动也能进一步提升推理能力。

### • 对推理场景进行RL

然后对数学、代码等推理场景进行RL。这里没啥好说的，和DeepSeek-R1-Zero一样的方式。针对DeepSeek-R1-Zero输出中语言混合的情况，额外增加一个奖励：语言一致性奖励，统计输出中目标语言的占比作为奖励信号。将原始的准确性奖励与语言一致性奖励求和作为最终奖励，进行过程反馈。

### • 拒绝采样和SFT

这一步主要是为了提升模型的通用能力，通过构建两部分的数据进行SFT来实现。

- 推理数据：采用拒绝采样的方式从前一阶段得到的模型生成推理过程，同时额外引入一些无法用规则进行奖励的数据（这部分数据使用DeepSeek-V3通过LLM-as-judge的方式进行奖励，比较GroudTruth与实际输出）。同时，过滤掉了包含混合语言、长段落、代码块的CoT数据。总计有60w样本。
- 非推理数据：使用DeepSeek-V3生成、使用DeepSeek-V3的SFT数据，共计20w推理无关的样本。

这一阶段总共生成了80w样本，用DeepSeek-V3-Base 进行了2个**epoch**<sup>+</sup>的SFT。

### • 适配所有场景的RL阶段

最后为了同时平衡推理能力和通用能力，又进行了一次RL。对于不同的数据类型，采用不同的prompt和奖励。



首发于  
LLM应用技术指南

的方式。对于有用性，专注于评估最终的summary，确保评估对用户的实用性和相关性，同时尽量减少对底层推理过程的干扰。对于无害性，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得最终的模型既能保持推理能力，又能满足有用性和无害性，取得比较好的用户体验。

实验结果自然是遥遥领先，和OpenAI-o1-1217不分伯仲。

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	91.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2
	IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	75.7
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	47.0
	FRAMES (Acc)	72.5	80.5	73.3	76.9	-
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-
Code	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-
	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6
	Codeforces (Rating)	717	759	1134	1820	2061
	SWE Verified (Resolved)	50.8	38.8	42.0	41.6	48.9
Math	Aider-Polyglot (Acc)	45.3	16.0	49.6	32.9	61.7
	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4
Chinese	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-
	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-
	C-Eval (EM)	76.7	76.0	86.5	68.9	-
Chinese	C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-

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

蒸馏小模型

直接用DeepSeek-R1阶段三：“拒绝采样和SFT”时的数据对小模型进行SFT，**不包含RL阶段**，就能取得比较好的效果。

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.

一些讨论

- 蒸馏 v.s. RL

从实验结果来看，蒸馏是又便宜又实用。用小的模型哼哧哼哧一顿SFT+RL操作，最后的效果还远不如直接蒸馏更好性能模型的输出直接SFT。



首发于  
LLM应用技术指北

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. 知乎@Meta

一些暂未成功的尝试

这里DeepSeek团队也是诚意满满，分享了几个业界呼声很高，但是他们暂时没尝试成功的方法。

**PRM<sup>+</sup>**: 指出PRM的几个主要限制，影响了它的规模化应用。1. 在一般推理过程中明确定义细粒度的步骤比较困难。2. 对步骤打标难以扩展，采用自动标注难以获得较高准确率，手动标注又难以规模化应用。3. 引入基于模型的PRM，不可避免地会遇到reward hacking<sup>+</sup>，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。

**MCTS<sup>+</sup>**: 他们也尝试了实用MCTS，但是过程中遇到了一些问题，1是搜索空间过大，虽然设置了最大扩展限制使得不会无限搜索，但是容易陷入局部最优；2是value model<sup>+</sup>直接影响生成的质量，而训练一个细粒度的value model本质上是困难的，这使得模型比较难以迭代改进。

一些未来的改进方向

- 通用能力**: DeepSeek-R1的通用能力仍然不及DeepSeekV3<sup>+</sup>。接下来，DeepSeek团队计划探索如何利用长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，但是在处理其他语言以及语言遵循方面还是会有问题。
- PE**: DeepSeek-R1对Prompt非常敏感。few-shot提示会持续降低其性能。这里建议用户直接描述问题并指定输出格式（采用zero-shot<sup>+</sup>，不要加示例），以获得最佳结果。
- 软件工程任务**: 由于长时间的评估会影响RL过程的效率，大规模RL尚未在软件工程任务中广泛应用。因此，DeepSeek-R1在软件工程基准测试上未显示出比DeepSeek-V3更大的改进。未来版本将通过在软件工程数据上实施拒绝采样或在RL过程中引入异步评估来提高效率。




参考

1. [github.com/deepseek-ai/...](https://github.com/deepseek-ai/)

编辑于 2025-01-23 10:34 · IP 属地浙江

内容所属专栏

 **LLM应用技术指北**  
不定期分享LLM应用相关技术


订阅专栏

LLM DeepSeek-R1

赞同 158 1 条评论 分享 喜欢 收藏 申请转载 ...

理性发言，友善互动

1 条评论 默认 最新

 **水dong方块** ...  
感觉deepseek 就是chatgpt o1的上限了, 就是学到了math 这些数据集的上限.  
昨天 09:40 · 美国 回复 喜欢

推荐阅读