中信建投证券
CHINA SECURITIES

证券研究报告•行业动态研究

DeepSeek R1深度解析及算力影响几何

分析师：于芳博
yufangbo@csc.com.cn
SAC编号：S1440522030001

分析师：庞佳军
pangjiajun@csc.com.cn
SAC 编号:S1440524110001

分析师：辛侠平
xinxiaping@csc.com.cn
SAC编号：S1440524070006

研究助理：孟龙飞
meng longfei@csc.com.cn
010-56135277

发布日期：2025年2月3日

摘要

- **核心观点： Deepseek发布深度推理能力模型。R1-Zero采用纯粹的强化学习训练，证明了大语言模型仅通过强化学习也可以有强大的推理能力，DeepSeek-R1经历微调和强化学习取得了与OpenAI-o1-1217相媲美甚至超越的成绩。DeepSeek R1训练和推理算力需求较低，主要原因是DeepSeek R1实现算法、框架和硬件的优化协同。过去的预训练侧的scaling law正逐步迈向更广阔的空间，在深度推理的阶段，模型的未来算力需求依然会呈现爆发式上涨，充足的算力需求对于人工智能模型的性能进步依然至关重要。**

- **Deepseek发布深度推理能力模型，性能和成本方面表现出色。**Deepseek发布两款具备深度推理能力的大模型R1-Zero和DeepSeek-R1。R1-Zero采用纯粹的强化学习训练，模型效果逼近OpenAI o1模型，证明了大语言模型仅通过RL，无SFT，大模型也可以有强大的推理能力。但是R1-Zero也存在可读性差和语言混合的问题，在进一步的优化过程中，DeepSeek-V3-Base经历两次微调和两次强化学习得到R1模型，主要包括冷启动阶段、面向推理的强化学习、拒绝采样与监督微调、面向全场景的强化学习四个阶段，R1在推理任务上表现出色，特别是在AIME 2024、MATH-500和Codeforces等任务上，取得了与OpenAI-o1-1217相媲美甚至超越的成绩。

- **国产模型迈向深度推理，策略创新百花齐放。在**Deepseek R1-Zero模型中，采用的强化学习策略是GRPO策略，取消价值网络，采用分组相对奖励，专门优化数学推理任务，减少计算资源消耗； KIMI 1.5采用Partial rollout的强化学习策略，同时采用模型合并、最短拒绝采样、DPO 和long2short RL策略实现短链推理； Qwen2.5扩大监督微调数据范围以及两阶段强化学习，增强模型处理能力。

- **DeepSeek R1通过较少算力实现高性能模型表现，主要原因是DeepSeek R1实现算法、框架和硬件的优化协同。** DeepSeek R1在诸多维度上进行了大量优化，算法层面引入专家混合模型、多头隐式注意力、多token预测，框架层面实现FP8混合精度训练，硬件层面采用优化的流水线并行策略，同时高效配置专家分发与跨节点通信，实现最优效率配置。当前阶段大模型行业正处于从传统的生成式模型向深度推理模型过渡阶段，算力的整体需求也从预训练阶段逐步过渡向后训练和推理侧，通过大量协同优化，DeepSeek R1在特定发展阶段通过较少算力实现高性能模型表现，算力行业的长期增长逻辑并未受到挑战。过去的预训练侧的scaling law正逐步迈向更广阔的空间，在深度推理的阶段，模型的未来算力需求依然会呈现爆发式上涨，充足的算力需求对于人工智能模型的性能进步依然至关重要。

- **风险提示：大模型技术发展不及预期、商业化落地不及预期、政策监管力度不及预期、数据数量与数据质量不及预期。**

中信建投证券
CHINA SECURITIES

第一章

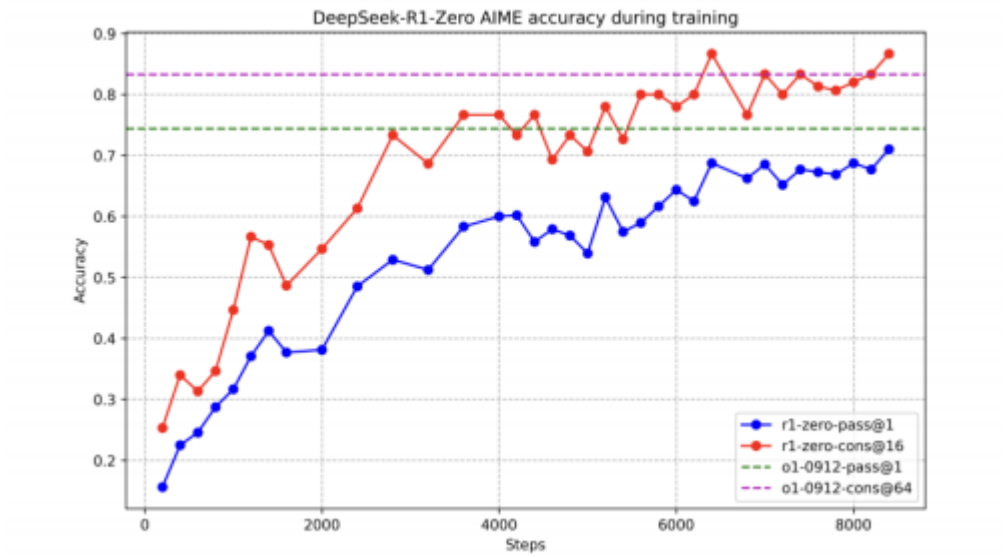
中信建投证券
CHINA SECURITIES

R1-Zero验证了大模型仅通过RL就可实现强大推理能力

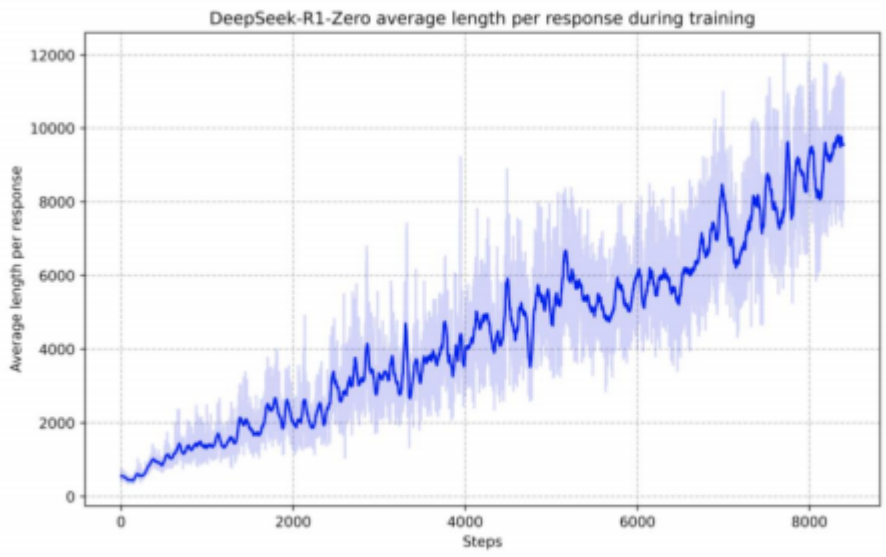
- **Deepseek发布两款具备深度推理能力的大模型R1-Zero和DeepSeek-R1。**
- **R1-Zero的训练，证明了仅通过RL，无SFT，大模型也可以有强大的推理能力。**在AIME 2024上，R1-Zero的pass@1指标从15.6%提升至71.0%，经过投票策略（majority voting）后更是提升到了86.7%，与OpenAI-o1-0912相当。

- ✓ **架构思路：**没有任何SFT数据的情况下，通过纯粹的强化学习。
- ✓ **算法应用：**直接在DeepSeek-V3-Base模型上应用GRPO算法进行强化学习训练。
- ✓ **奖励机制：**使用基于规则的奖励机制，包括准确性奖励和格式奖励，来指导模型的学习。
- ✓ **训练模板：**采用了简洁的训练模板，要求模型首先输出推理过程（置于标签内），然后给出最终答案（置于标签内）。

图：R1-Zero在AIME 2024基准测试上的性能测试



图：强化学习过程中的scaling law

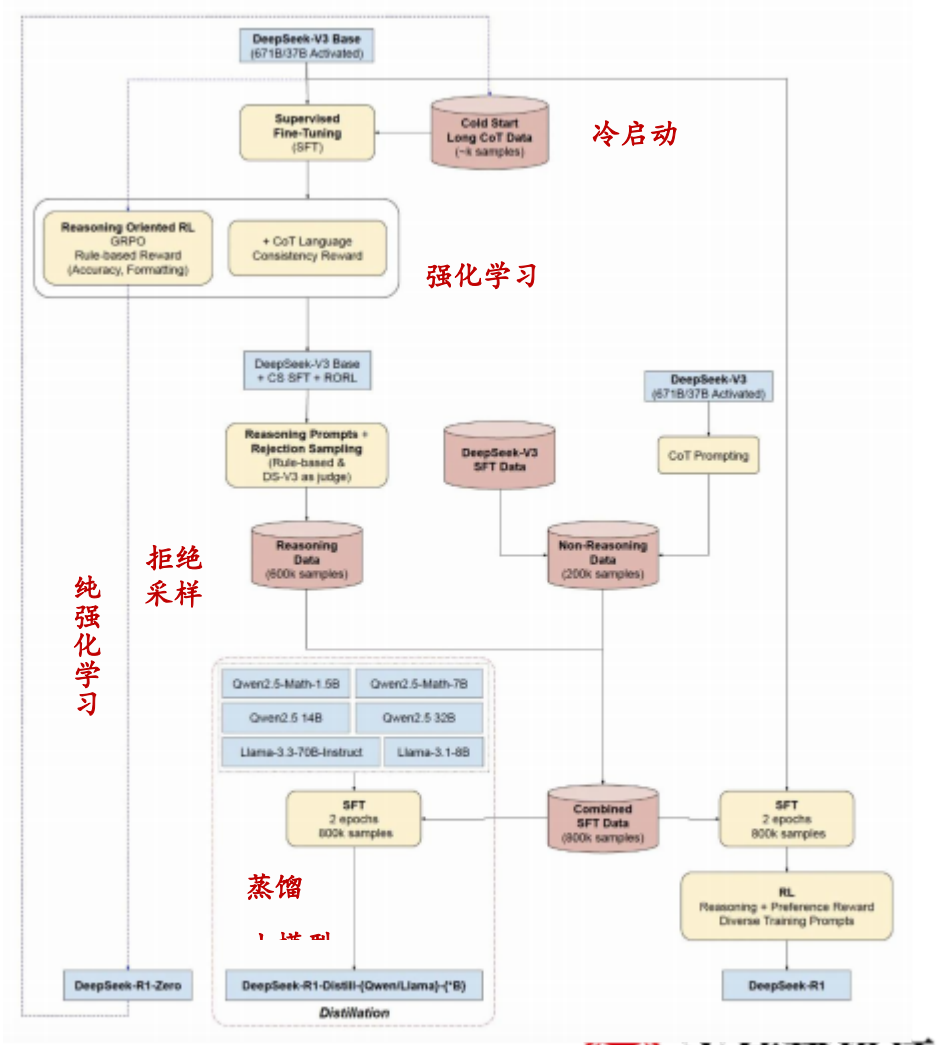


资料来源：DeepSeek-R1：Incentivizing Reasoning Capability in LLMs via Reinforcement Learning，中信建投

DeepSeek-R1：长CoT数据微调基础上应用强化学习

- **为了解决R1-Zero可读性差和语言混合的问题，构建了R1。**
- **架构思路：** 在DeepSeek-V3-Base模型的基础上，经历两次微调和两次强化学习得到R1模型。
- Step 1.**冷启动阶段：**使用数千个高质量的长Cot人工标注样本对DeepSeek-V3-Base模型进行微调，作为强化学习的初始模型。
- Step 2.**面向推理的强化学习：**在冷启动阶段之后，R1采用了与R1-Zero类似的强化学习训练，但针对推理任务进行了特别优化。为了解决训练过程中可能出现的语言混杂问题，R1引入了语言一致性奖励，该奖励根据CoT中目标语言单词的比例来计算。
- Step 3.**拒绝采样与监督微调：** 当面向推理的强化学习收敛后，R1利用训练好的RL模型进行拒绝采样，生成新的SFT数据。
- Step 4.**面向全场景的强化学习：**在收集了新的SFT数据后，R1会进行第二阶段的强化学习训练，这一次，训练的目标不再局限于推理任务，而是涵盖了所有类型的任务。此外，R1采用了不同的奖励信号和提示分布，针对不同的任务类型进行了优化。

图：DeepSeek-R1训练过程

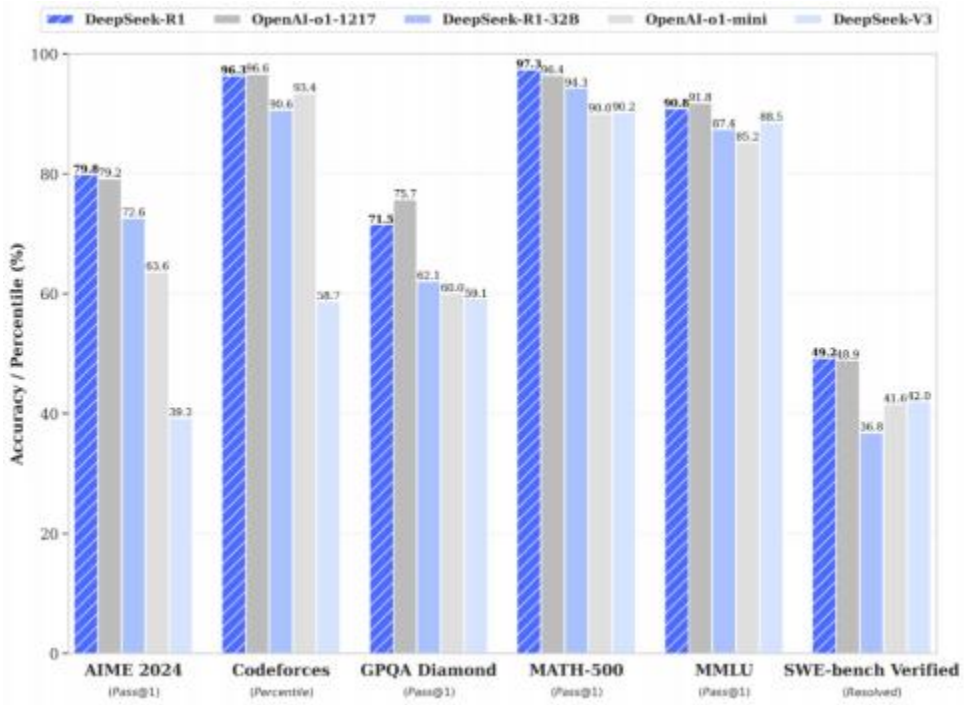


中信建投证券
CHINA SECURITIES

R1模型推理任务表现出色

- R1在推理任务上表现出色，特别是在AIME 2024（美国数学邀请赛）、MATH-500（数学竞赛题）和Codeforces（编程竞赛）等任务上，取得了与OpenAI-o1-1217相媲美甚至超越的成绩。在MMLU（90.8%）、MMLU-Pro（84.0%）和GPQA Diamond（71.5%）等知识密集型任务基准测试中，性能显著超越了DeepSeek-V3模型。在针对长上下文理解能力的FRAMES数据集上，R1的准确率达到了82.5%，优于DeepSeek-V3模型。在开放式问答任务AlpacaEval 2.0和Arena-Hard基准测试中，R1分别取得了87.6%的LC-winrate和92.3%的GPT-4-1106评分，展现了其在开放式问答领域的强大能力。

图表：R1在数学、代码、自然语言推理等任务的性能测试结果



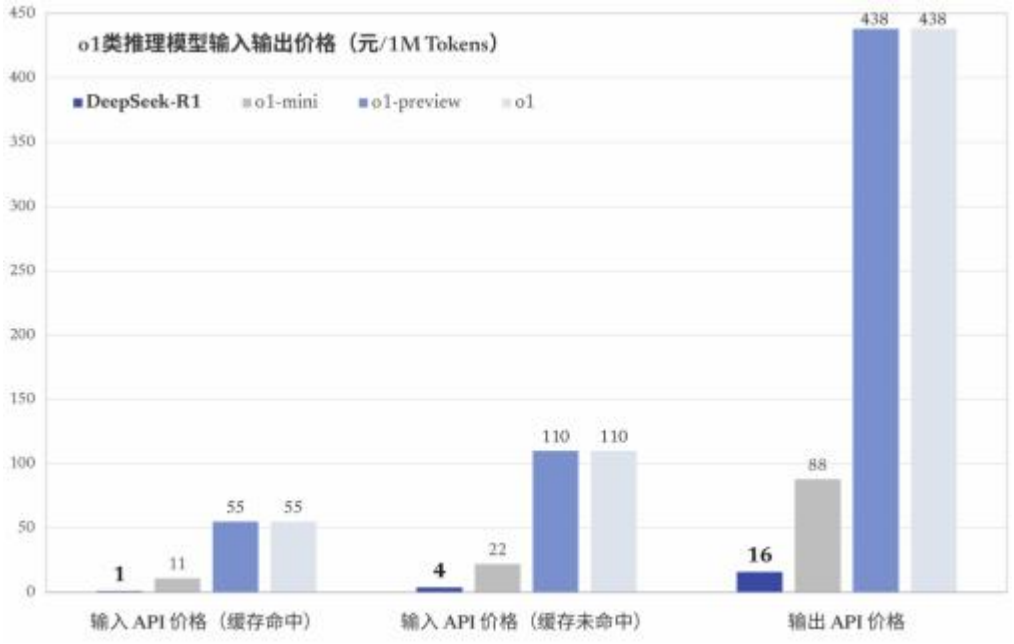
	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**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
	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**
Code	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

资料来源：DeepSeek-R1：Incentivizing Reasoning Capability in LLMs via Reinforcement Learning，中信建投

通过蒸馏实现推理能力迁移

- DeepSeek团队进一步探索了将R1的推理能力蒸馏到更小的模型中的可能性。他们使用R1生成的800K数据，对Qwen和Llama系列的多个小模型(1.5B、7B、8B、14B、32B、70B)进行了微调。经过R1蒸馏的小模型，在推理能力上得到了显著提升，甚至超越了在这些小模型上直接进行强化学习的效果。
- **推理成本来看，**R1模型价格只有OpenAI o1模型的几十分之一。训练成本来看，DeepSeek-V3在一个配备2048个NVIDIA H800 GPU的集群上进行训练，预训练阶段在不到两个月内完成，并消耗了2664K GPU小时，总训练成本为557.6万美元。

图：o1类推理模型输入输出价格(元/1M Tokens)



图：蒸馏模型表现

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820.0**
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633.0