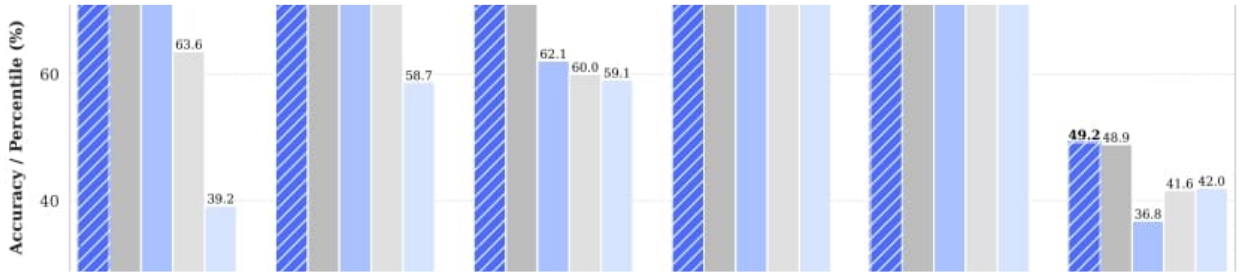
一文读懂 DeepSeek R1：强化学习如何重塑大语言模型推理能力?

DeepSeek R1 强化学习 大模型

Jan 24, 2025 14 min read



最近，AI领域又迎来了一项重磅研究成果——`DeepSeek R1`。这一推理模型在性能上取得了重大突破，甚至能与`OpenAI`的`o1-1217`相媲美。它的出现，不仅为大语言模型（LLMs）的发展开辟了新路径，也为整个AI研究领域注入了新的活力。今天，就让我们深入解读一下`DeepSeek R1`背后的研究论文，看看它究竟有哪些创新点和过人之处。

DeepSeek R1：挑战与突破并存

在AI发展的浪潮中，LLMs正快速迭代，不断缩小与通用人工智能（AGI）之间的差距。后训练作为训练流程的关键一环，能有效提升模型在推理任务中的准确率，还能让模型更好地契合社会价值和用户偏好。此前，OpenAI的o1系列模型通过增加思维链推理过程的长度，在推理任务上取得了显著进展，但如何实现有效的测试时扩展，仍然是学界亟待解决的难题。

在这样的背景下，DeepSeek R1的研究团队另辟蹊径，尝试运用纯强化学习（RL）来提升语言模型的推理能力。他们的目标很明确：探索LLMs在没有任何监督数据的情况下，通过纯RL过程自我进化出推理能力的潜力。

研究团队以DeepSeek-V3-Base为基础模型，采用GRPO（Group Relative Policy Optimization）作为RL框架。在训练过程中，他们惊喜地发现，DeepSeek R1-Zero（不依赖监督微调的纯RL模型）展现出了强大且有趣的推理行为。经过数千次RL训练步骤，DeepSeek R1-Zero在推理基准测试中的表现大幅提升。以AIME 2024测试为例，其单样本通过率（pass@1）从最初的15.6% 飙升至71.0%；若采用多数投票策略，这一成绩更是能提升到86.7%，与OpenAI-o1-0912的水平相当。

不过，`DeepSeek R1-Zero`也并非十全十美，它存在可读性差、语言混合等问题。为了解决这些问题，并进一步提升推理性能，研究团队推出了`DeepSeek R1`。`DeepSeek R1`通过引入少量冷启动数据和多阶段训练流程，成功克服了`DeepSeek R1-Zero`的部分缺陷，最终在性能上达到了与`OpenAI-o1-1217`相媲美的水平。

技术亮点：创新架构与训练策略

DeepSeek R1-Zero：强化学习的深度探索

`DeepSeek R1-Zero`的训练过程可谓独树一帜。团队采用GRPO算法，这一算法舍弃了与策略模型大小相同的评论家模型，通过群组分数来估计基线，大大节省了训练成本。

在奖励建模方面，团队采用了基于规则的奖励系统，主要包含准确性奖励和格式奖励。准确性奖励用于评估模型的回答是否正确，比如在数学问题中，模型需按指定格式给出最终答案，以便进行正确性验证；格式奖励则要求模型将思考过程放在“”和“”标签之间。这种奖励机制简单直接，有效避免了神经奖励模型可能出现的奖励作弊问题，同时也降低了训练的复杂性。

为了引导模型的训练，团队设计了一个简洁的模板。该模板要求模型先进行推理，再给出最终答案，并且尽量避免对内容进行特定限制，以便观察模型在RL过程中的自然发展。

在训练过程中，DeepSeek R1-Zero展现出了令人惊叹的自我进化能力。随着训练步数的增加，它在AIME 2024测试中的准确率稳步提升。不仅如此，模型还学会了自我反思和探索多种解题方法。在遇到复杂问题时，它会重新评估之前的步骤，尝试不同的解题思路，这种“顿悟时刻”充分体现了强化学习的魅力，让模型能够自主发展出先进的解题策略。

DeepSeek R1：融入冷启动数据的优化升级

`DeepSeek R1`的训练流程分为四个阶段，旨在解决`DeepSeek R1-Zero`存在的问题，并进一步提升模型性能。

在冷启动阶段，团队构建并收集了少量高质量的长思维链（CoT）数据，对DeepSeek-V3-Base模型进行微调，以此作为RL训练的初始演员。这些冷启动数据经过精心设计，具有良好的可读性，能够有效避免模型在训练初期出现不稳定的情况。

在推理导向的强化学习阶段，团队采用了与DeepSeek R1-Zero相同的大规模RL训练过程，但在此基础上引入了语言一致性奖励，以缓解思维链中出现的语言混合问题。虽然这一奖励机制会导致模型性能略有下降，但却使模型的输出更符合人类的阅读习惯。

当推理导向的RL训练接近收敛时，团队利用拒绝采样和监督微调（SFT）来收集更多数据。他们不仅从推理任务中收集数据，还纳入了写作、角色扮演等其他领域的数据，以增强模型的通用能力。在这个过程中，团队对数据进行了严格筛选，过滤掉了语言混合、冗长段落和代码块等难以阅读的内容。

为了使模型更好地符合人类偏好，团队还进行了全场景的强化学习。在这个阶段，他们综合运用多种奖励信号和多样化的提示分布，对模型进行进一步训练。对于推理数据，仍然采用基于规则的奖励；对于通用数据，则借助奖励模型来捕捉人类偏好。通过这种方式，模型在保证推理能力的同时，更加注重对用户的帮助和无害性。

模型蒸馏：赋予小模型强大推理能力

为了让更高效的小模型也具备强大的推理能力，研究团队从DeepSeek R1向小模型进行知识蒸馏。他们直接使用DeepSeek R1生成的800k样本对Qwen和Llama等开源模型进行微调。实验结果令人惊喜，经过蒸馏的小模型在推理能力上有了显著提升。例如，DeepSeek-R1-Distill-Qwen-7B在AIME 2024测试中取得了55.5%的成绩，超越了QwQ-32B-Preview；DeepSeek-R1-Distill-Qwen-32B在多个测试中表现优异，其成绩与o1-mini相当。这一成果表明，将大模型的推理模式蒸馏到小模型中是一种非常有效的方法，能够让小模型在保持高效性的同时，获得强大的推理能力。

实验结果：全方位超越与领先

研究团队对`DeepSeek R1`及蒸馏后的小模型进行了广泛的实验评估，涵盖了多个基准测试，包括`MMLU`、`MMLU-Pro`、`GPQA Diamond`、`AIME 2024`、`LiveCodeBench`等，同时还与多个强大的基线模型进行了对比。

在教育导向的知识基准测试中，`DeepSeek R1`的表现优于`DeepSeek-V3`，尤其在`STEM`相关问题上，通过大规模RL训练取得了显著的准确率提升。在`FRAMES`等长上下文依赖的问答任务中，`DeepSeek R1`也展现出了强大的文档分析能力。

在数学任务和编码算法任务中，`DeepSeek R1`的性能与`OpenAI-o1-1217`相当，大幅超越了其他模型。在写作任务和开放域问答任务中，`DeepSeek R1`在`AlpacaEval 2.0`和`ArenaHard`测试中表现出色，其生成的总结长度简洁，避免了长度偏差，进一步证明了其在多任务处理上的稳健性。

蒸馏后的小模型同样表现优异，`DeepSeek-R1-Distill-Qwen-7B`在`AIME 2024`测试中超越了`GPT-4o-0513`等非推理模型；`DeepSeek-R1-Distill-Qwen-14B`在所有评估指标上均超过了`QwQ-32B-Preview`；`DeepSeek-R1-Distill-Qwen-32B`和`DeepSeek-R1-Distill-Llama-70B`在大多数基准测试中显著超过`o1-mini`。这些结果充分展示了蒸馏技术的有效性，以及`DeepSeek R1`强大的推理能力和泛化能力。

未来展望：持续创新与拓展

DeepSeek R1的出现无疑为LLMs的发展带来了新的思路和方法，但研究团队并没有满足于此。他们在论文中指出了未来的研究方向，旨在进一步提升DeepSeek R1的性能和应用范围。

在通用能力方面，DeepSeek R1在函数调用、多轮对话、复杂角色扮演和json输出等任务上还有提升空间。团队计划探索如何利用长思维链来优化这些任务的处理能力。

在语言混合问题上，目前DeepSeek R1主要针对中文和英文进行了优化，在处理其他语言的查询时可能会出现语言混合的情况。未来，团队将致力于解决这一问题，使模型能够更好地处理多种语言的任务。

在提示工程方面，DeepSeek R1对提示较为敏感，少样本提示会导致其性能下降。团队建议用户采用零样本设置来描述问题和指定输出格式，以获得最佳效果。未来，他们也将进一步研究如何优化模型对提示的适应性，提高模型在不同提示条件下的稳定性。

在软件工程任务方面，由于评估时间较长，影响了RL过程的效率，DeepSeek R1在软件工程基准测试上的提升有限。未来版本将通过对软件工程数据进行拒绝采样或在RL过程中引入异步评估来提高效率，从而提升模型在软件工程任务中的表现。

DeepSeek R1的研究成果为LLMs的推理能力提升提供了重要的参考和借鉴，其创新的训练方法和优秀的实验结果让人对AI的未来发展充满期待。相信在研究团队的不断努力下，DeepSeek R1将在未来取得更大的突破，为AI领域带来更多的惊喜。作为AI爱好者，我们不妨持续关注DeepSeek R1的发展动态，见证AI技术的不断进步。

相关地址

- DeepSeek R1 论文地址
- DeepSeek R1 代码地址

Share

更多文章