deepseek



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

DeepSeek-AI
research@deepseek.com

# DeepSeek-R1 解读及技术报告中文版

AINLP
AINLP公众号和我爱自然语言处理(52nlp)网站保姆



前两天DeepSeek发布了DeepSeek R1的技术报告：

技术报告原文：github.com/deepseek-ai/...

以下是这篇论文的解读，由DeepSeek辅助完成。

近年来，大型语言模型（LLMs）在自然语言处理领域取得了显著进展，但其核心推理能力仍面临挑战。传统方法多依赖监督微调（SFT）和复杂的提示工程，而DeepSeek-AI团队的最新研究《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》提出了一种革命性路径：通过纯强化学习（RL）自主激发模型的推理能力，并结合蒸馏技术实现高效迁移。本文将从技术突破、实验成果与行业影响三个维度，深度解析这一研究的核心价值。

## 一、技术突破：从零开始的推理能力进化

### 1. DeepSeek-R1-Zero：纯RL训练的“自我觉醒”

传统LLM的推理能力通常需要大量人工标注的监督数据，但DeepSeek-R1-Zero首次验证了无需任何SFT数据，仅通过强化学习即可实现推理能力的自主进化。其核心创新在于：

- 算法框架：采用Group Relative Policy Optimization（GRPO），通过组内奖励对比优化策略，避免传统RL中复杂价值模型的依赖。
- 自我进化现象：模型在训练中自发涌现出“反思”（Re-evaluation）、“多步验证”（Multi-step Verification）等复杂推理行为。例如，在解决数学方程时，模型会主动纠正早期错误步骤（如表3的“Aha Moment”）。
- 性能飞跃：在AIME 2024数学竞赛任务中，模型Pass@1准确率从初始的15.6%提升至71.0%，多数投票（Majority Voting）后更达86.7%，与OpenAI的o1-0912模型持平。

然而，纯RL训练的代价是可读性差与多语言混杂。模型生成的推理过程常包含中英文混合、格式混乱等问题，限制了实际应用。

### 2. DeepSeek-R1：冷启动与多阶段训练的平衡之道

为解决上述问题，团队提出**“冷启动+多阶段RL”策略**：

- 两阶段强化学习：

1. 推理导向RL：结合规则奖励（答案准确性、语言一致性），优化数学、编程等结构化任务表现。
2. 通用对齐RL：融入人类偏好奖励模型（Helpfulness & Harmlessness），确保模型在开放域任务中的安全性与实用性。

性能对标：DeepSeek-R1在MATH-500（97.3% Pass@1）、Codeforces（超越96.3%人类选手）等任务上达到与OpenAI-o1-1217相当的水平，同时在MMLU（90.8%）、GPQA Diamond（71.5%）等知识密集型任务中显著超越前代模型。

---

## 二、实验验证：推理能力的全方位跃升

### 1. 基准测试：超越顶尖闭源模型

论文在20余项基准任务中对比了DeepSeek-R1与Claude-3.5、GPT-4o、OpenAI-o1系列等模型（表4），关键结论包括：

- 数学与编程：AIME 2024（79.8%）、MATH-500（97.3%）、LiveCodeBench（65.9%）等任务表现全面领先，Codeforces评分（2029）接近人类顶尖选手。
- 知识密集型任务：MMLU（90.8%）、GPQA Diamond（71.5%）等得分显著高于DeepSeek-V3，逼近OpenAI-o1-1217。
- 通用能力：AlpacaEval 2.0（87.6%胜率）、长上下文理解（如FRAMES任务82.5%）表现突出，证明RL训练可泛化至非推理场景。

### 2. 蒸馏技术：小模型的逆袭

通过将DeepSeek-R1生成的80万条数据用于微调开源模型（Qwen、Llama系列），团队实现了推理能力的高效迁移：

- 小模型性能飞跃：7B参数模型在AIME 2024上达55.5%，超越32B规模的QwQ-Preview；70B蒸馏模型在MATH-500（94.5%）等任务接近o1-mini。
- 开源贡献：发布1.5B至70B的蒸馏模型，为社区提供低成本、高性能的推理解决方案。

---

## 三、行业启示：AGI之路的新范式

### 1. 纯RL训练的价值与挑战

DeepSeek-R1-Zero的成功证明，无需人工标注的RL训练可自主挖掘模型的推理潜力。这一发现挑战了传统LLM依赖监督数据的范式，为AGI研究提供了新思路。然而，其局限性（如可读性差）也表明，完全自主进化仍需与人类先验知识结合。

### 2. 蒸馏技术的普惠意义

通过蒸馏实现推理能力迁移，不仅降低了计算成本，更使小模型在特定任务中媲美大模型。例如，7B模型在数学任务上超越GPT-4o，这为边缘计算、实时应用场景提供了可行方案。

### 3. 开源生态的推动力

## 四、未来展望：从推理到通用智能

尽管DeepSeek-R1取得了突破，其局限仍指向未来方向：

- 多语言与工程任务：当前模型优化⁺以中英文为主，其他语言支持有限；软件工程任务因评估效率问题提升缓慢。
- 长推理链的扩展：探索CoT⁺在函数调用、多轮对话等复杂场景的应用。
- 安全与可控性：RL训练中奖励模型的设计需进一步平衡性能与伦理约束。

## 结语

DeepSeek-R1的研究标志着LLM推理能力进化的一次重要跨越。通过纯强化学习与蒸馏技术，团队不仅验证了模型自主进化的可能性，更构建了从理论研究到产业落地的完整链条。这一工作为AGI的发展提供了新范式：在减少对人类先验依赖的同时，通过算法创新与开源协作，推动智能技术的普惠与深化。未来，随着更多类似研究的涌现，我们或许正站在通用人工智能的真正起点。

以下是技术报告中文版本，由DeepSeek API⁺将其全文翻译为中文，仅供学习参考：

deepseek

# DeepSeek-R1：通过强化学习激励LLMs中的推理能力

深度探索人工智能

research@deepseek.com

## 摘要

我们推出了第一代推理模型，DeepSeek-R1-Zero 和 DeepSeek-R1。DeepSeek-R1-Zero 是一个通过大规模强化学习（RL）训练而成的模型，无需监督微调（SFT）作为初步步骤，展示了卓越的推理能力。通过 RL，DeepSeek-R1-Zero 自然涌现出许多强大且有趣的推理行为。然而，它也面临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升推理性能，我们推出了 DeepSeek-R1，它在 RL 之前引入了多阶段训练和冷启动数据。DeepSeek-R1 在推理任务上实现了与 OpenAI-o1-1217 相当的性能。为了支持研究社区，我们开源了 DeepSeek-R1-Zero、DeepSeek-R1 以及基于 Qwen 和 Llama 从 DeepSeek-R1 蒸馏出的六个密集模型（1.5B、7B、8B、14B、32B、70B）。

图1 | DeepSeek-R1的基准性能。

## 1. 引言

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

最近，后训练已成为完整训练流程中的一个重要组成部分。它已被证明可以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，同时相对于预训练所需的计算资源相对较少。在推理能力方面，OpenAI的o1（OpenAI, 2024b）系列模型首次通过增加思维链推理过程的长度引入了推理时扩展。这种方法在数学、编码和科学推理等各种推理任务中取得了显著改进。然而，有效的测试时扩展仍然是研究界的一个开放性问题。之前的一些工作探索了各种方法，包括基于过程的奖励模型（Lightman等，2023；Uesato等，2022；Wang等，2023）、强化学习（Kumar等，2024）以及蒙特卡洛树搜索和束搜索等搜索算法（Feng等，2024；Trinh等，2024；Xin等，2024）。然而，这些方法都没有达到与OpenAI的o1系列模型相媲美的通用推理性能。

在本文中，我们迈出了利用纯强化学习（RL）提升语言模型推理能力的第一步。我们的目标是探索大型语言模型（LLMs）在没有监督数据的情况下发展推理能力的潜力，重点关注它们通过纯RL过程的自我进化。具体而言，我们使用DeepSeek-V3-Base作为基础模型，并采用GRPO（Shao等，2024）作为RL框架，以提升模型在推理任务中的表现。在训练过程中，DeepSeek-R1-Zero自然涌现出许多强大且有趣的推理行为。经过数千次RL步骤后，DeepSeek-R1-Zero在推理基准测试中展现出卓越的性能。例如，AIME 2024上的pass@1得分从15.6%提升至71.0%，而在多数投票机制下，得分进一步提高至86.7%，与OpenAI-o1-0912的性能相当。

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升推理性能，我们引入了 DeepSeek-R1，它结合了少量冷启动数据和多阶段训练流程。具体来说，我们首先收集数千条冷启动数据来微调 DeepSeek-V3-Base 模型。随后，我们进行类似 DeepSeek-R1-Zero 的推理导向强化学习（RL）。在 RL 过程接近收敛时，我们通过对 RL 检查点进行拒绝采样，结合来自 DeepSeek-V3 在写作、事实问答和自我认知等领域的监督数据，创建新的 SFT 数据，然后重新训练 DeepSeek-V3-Base 模型。使用新数据微调后，检查点会经历额外的 RL 过程，考虑所有场景的提示。经过这些步骤，我们获得了称为 DeepSeek-R1 的检查点，其性能与 OpenAI-o1-1217 相当。

我们进一步探索了从DeepSeek-R1到更小密集模型的蒸馏过程。以Qwen2.5-32B（Qwen, 2024b）为基础模型，直接从DeepSeek-R1进行蒸馏的效果优于在其上应用强化学习。这表明，更大基础模型发现的推理模式对于提升推理能力至关重要。我们开源了蒸馏后的Qwen和Llama（Dubey等，2024）系列。值得注意的是，我们蒸馏的14B模型大幅超越了当前最先进的开源QwQ-32B-Preview（Qwen, 2024a），而蒸馏的32B和70B模型在密集模型的推理基准测试中创下了新纪录。







发布于 2025-01-24 09:44 · IP 属地江苏

**内容所属专栏**