# DeepSeek-R1：通过强化学习激励大语言模型的推理能力

DeepSeek-AI research@deepseek.com

2025 年 10 月 23 日

## 摘要

我们介绍了我们的第一代推理模型 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，无需在强化学习前进行监督微调（SFT），展现出卓越的推理能力。通过强化学习，DeepSeek-R1-Zero 自然涌现出多种强大且引人入胜的推理行为。然而，该模型在可读性差和语言混杂方面仍面临挑战。为解决这些问题并进一步提升推理性能，我们引入了 DeepSeek-R1，其在强化学习前加入了多阶段训练和冷启动数据。DeepSeek-R1 在推理任务上的性能与 OpenAI-o1-1217 相当。为支持研究社区，我们开源了 DeepSeek-R1-Zero、DeepSeek-R1 以及六个基于 Qwen 和 Llama 模型从 DeepSeek-R1 提炼出的密集模型（1.5 B, 7 B, 8 B, 14 B, 32 B, 70 B）。

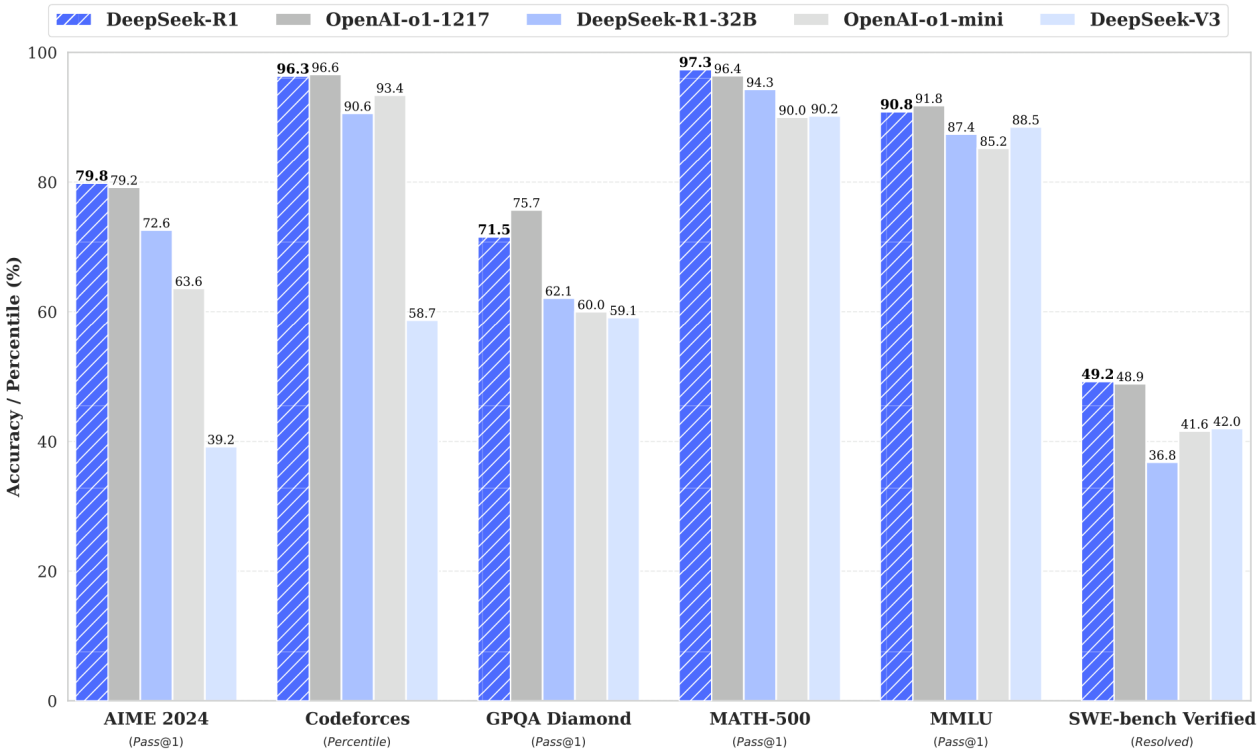


图 1 | DeepSeek-R1 的基准性能。

# 目录

# 1. 引言

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

近年来，微调训练已成为完整训练流程中的一个重要组成部分。研究表明，它能够在提升推理任务准确率、与社会价值观对齐以及适应用户偏好方面发挥重要作用，同时相比预训练阶段所需计算资源相对较少。在推理能力方面，OpenAI 的 o1（OpenAI, 2024b）系列模型首次引入了推理时的扩展方法，即通过延长思维链（Chain-of-Thought）推理过程的长度来实现。该方法在数学、编程和科学推理等多种推理任务中取得了显著提升。然而，如何实现有效的测试时扩展仍然是研究社区尚未解决的开放问题。已有若干工作探索了多种方法，包括基于过程的奖励模型（Lightman et al., 2023；Uesato et al., 2022；Wang et al., 2023）、强化学习（Kumar et al., 2024）以及蒙特卡洛树搜索和束搜索等搜索算法（Feng et al., 2024；Trinh et al., 2024；Xin et al., 2024）。然而，这些方法均未能实现与 OpenAI o1 系列模型相媲美的通用推理性能。

在本文中，我们首次尝试通过纯强化学习（RL）来提升语言模型的推理能力。我们的目标是探索大型语言模型在无需任何监督数据的情况下发展推理能力的潜力，重点在于其通过纯强化学习过程实现的自我演化。具体而言，我们以 DeepSeek-V3-Base 作为基础模型，并采用 GRPO（Shao 等，2024）作为强化学习框架，以提升模型在推理任务上的表现。在训练过程中，DeepSeek-R1-Zero 自然涌现出大量强大且有趣的推理行为。经过数千个强化学习步骤后，DeepSeek-R1-Zero 在推理基准测试中展现出超常性能。例如，AIME 2024 的 pass@1 得分从 15.6% 提升至 71.0%，通过多数投票，得分进一步提升至 86.7%，与 OpenAI-o1-0912 的性能相当。

然而，DeepSeek-R1-Zero 遇到了可读性差、语言混杂等挑战。为解决这些问题并进一步提升推理性能，我们引入了 DeepSeek-R1，该模型融合了少量冷启动数据和多阶段训练流程。具体而言，我们首先收集数千条冷启动数据，对 DeepSeek-V3-Base 模型进行微调。随后，我们执行类似 DeepSeek-R1Zero 的推理导向强化学习。当强化学习过程接近收敛时，我们通过在强化学习检查点上进行拒绝采样生成新

的 SFT 数据，并结合 DeepSeek-V3 在写作、事实问答和自我认知等领域的监督数据，然后重新训练 DeepSeek-V3-Base 模型。在使用新数据进行微调后，检查点再经历一次额外的强化学习过程，考虑所有场景下的提示。经过这些步骤，我们得到了一个称为 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 到更小的密集模型的蒸馏。以 Qwen2.532B（Qwen, 2024b）作为基础模型，直接从 DeepSeek-R1 进行蒸馏的表现优于对其应用强化学习。这表明，较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 et al., 2024）系列。值得注意的是，我们的蒸馏 14 B 模型在性能上显著优于最先进的开源模型 QwQ-32B-Preview（Qwen, 2024a），而蒸馏后的 32 B 和 70 B 模型在密集模型的推理基准上创下了新纪录。

### 1.1. 主要贡献

## 训练后：在基础模型上进行大规模强化学习

- 我们直接将强化学习（RL）应用于基础模型，而不依赖于监督微调（SFT）作为前期步骤。该方法使模型能够探索思维链（CoT）以解决复杂问题，从而发展出 DeepSeek-R1-Zero。DeepSeek-R1-Zero 展现了自我验证、反思以及生成长思维链等能力，标志着研究社区的一个重要里程碑。值得注意的是，它是首个公开研究，验证了无需监督微调（SFT），仅通过强化学习（RL）即可激励大语言模型（LLMs）的推理能力。这一突破为该领域的未来进展铺平了道路。

- 我们介绍了开发 DeepSeek-R1 的流程。该流程包含两个强化学习阶段，旨在发现更优的推理模式并符合人类偏好，以及两个监督微调阶段，作为模型推理能力和非推理能力的初始基础。我们相信该流程将通过构建更优的模型而造福行业。

## 蒸馏：小模型也可以非常强大

- 我们证明了较大模型的推理模式可以被蒸馏到较小模型中，其性能优于通过在小模型上进行强化学习发现的推理模式。开源的 DeepSeek-R1 及其 API 将有助于研究社区未来蒸馏出更优的小型模型。

- 利用 DeepSeek-R1 生成的推理数据，我们对研究社区中广泛使用的若干密集模型进行了微调。评估结果表明，蒸馏后的较小密集模型在基准测试中表现优异。DeepSeek-R1-Distill-Qwen-7B 在 AIME 2024 上达到 55.5%，超越了 QwQ-32B-Preview。此外，DeepSeek-R1-Distill-Qwen-32B 在 AIME 2024 上得分为 72.6%，在 MATH-500 上得分为 94.3%，在 LiveCodeBench 上得分为 57.2%。这些结果显著优于以往的开源模型，性能可与 o1-mini 相媲美。我们基于 Qwen2.5 和 Llama3 系列向社区开源了 1.5B、7B、8B、14B、32B 和 70B 的蒸馏检查点。

### 1.2. 评估结果总结

- 推理任务：（1）DeepSeek-R1 在 AIME 2024 上达到 79.8% Pass@1，略微超过 OpenAI-o1-1217。在 MATH-500 上，其表现优异，得分为 97.3%，与 OpenAI-o1-1217 相当，并显著优于其他模型。（2）在与编程相关的任务中，DeepSeek-R1 在编程竞赛任务上展现出专家级水平，其在 Codeforces 上的 Elo 排名达到 2,029，超越了竞赛中 96.3% 的人类参与者。在与工程相关的任务中，DeepSeek-R1 的表现略优于 DeepSeek-V3，这将有助于开发者在实际任务中应用。

- 知识：在 MMLU、MMLU-Pro 和 GPQA Diamond 等基准测试中，DeepSeek-R1 取得了卓越的表现，显著优于 DeepSeek-V3，在 MMLU 上得分为 90.8%，在 MMLU-Pro 上得分为 84.0%，在 GPQA Diamond 上得分为 71.5%。尽管其在这些基准上的表现略低于 OpenAI-o1-1217，但 DeepSeek-R1 仍超越了其他闭源模型，展示了其在教育任务中的竞争优势。在事实类基准 SimpleQA 上，DeepSeek-R1 优于 DeepSeek-V3，体现了其处理基于事实查询的能力。在该基准上，OpenAI-o1 也优于 4o，呈现出类似的趋势。- 其他任务：DeepSeek-R1 在多种任务上也表现出色，包括创意写作、通用问答、编辑、摘要等。它在 AlpacaEval 2.0 上实现了令人印象深刻的长度控制胜率 87.6%，在 ArenaHard 上的胜率为 92.3%，展现了其在处理非考试导向查询时的强智能能力。此外，DeepSeek-R1 在需要长上下文理解的任务上表现出色，显著优于 DeepSeek-V3 在长上下文基准测试中的表现。

## 2. 方法

### 2.1. 概述

先前的工作严重依赖大量监督数据来提升模型性能。在本研究中，我们证明，即使不使用监督微调（SFT）作为冷启动，通过大规模强化学习（RL）也可以显著提升推理能力。此外，通过引入少量冷启动数据，性能可以进一步提升。在以下部分中，我们介绍了：（1）DeepSeek-R1-Zero，该方法直接将强化学习应用于基础模型，而无需任何监督微调数据；（2）DeepSeek-R1，该方法从一个使用数千个长链思维（CoT）示例进行微调的检查点开始应用强化学习；（3）将 DeepSeek-R1 的推理能力蒸馏到小型密集模型中。

### 2.2. DeepSeek-R1-Zero：基于基模型的强化学习

强化学习在推理任务中已展现出显著的有效性，如我们之前的工作所示（Shao 等，2024；Wang 等，2023）。然而，这些工作严重依赖于监督数据，而收集监督数据耗时较长。在本节中，我们探讨了大语言模型在无需任何监督数据的情况下发展推理能力的潜力，重点关注其通过纯粹的强化学习过程实现的自我演化。我们

首先简要介绍我们的强化学习算法，随后展示一些令人兴奋的结果，希望这能为社区提供有价值的见解。

### 2.2.1. 强化学习算法

组相对策略优化为节省强化学习的训练成本,我们采用组相对策略优化(GRPO)(Shao等，2024），该方法省去了通常与策略模型大小相同的批评者模型，并从组得分中估计基线。具体而言，对于每个问题 $q$，GRPO 从旧策略 $\pi_{\theta_{\text{old}}}$ 中采样一组输出 $\{o_1, o_2, \cdots, o_G\}$，然后通过最大化以下目标函数来优化策略模型 $\pi_\theta$：

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right),$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1,$$

其中 $\varepsilon$ 和 $\beta$ 是超参数，$A_i$ 是优势，通过对应每组输出的一组奖励 $\{r_1, r_2, \ldots, r_G\}$ 计算得到：

$$\mathcal{J}\_GRPO(\theta) = \mathbb{E}\left[q \sim P(Q), \{o\_i\}\_i = 1^G \sim \pi\_\theta\_\text{old}\ (O \mid q)\right] \frac{1}{G}\sum\_i = 1^G \left(\min\left(\frac{\pi\_\theta\,(o\_i \mid q)}{\pi\_\theta\_\text{old}\ (o\_i \mid}\right.\right.$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}.$$

$$A_i = \frac{r_i - \text{mean}\left(\{r_1, r_2, \cdots, r_G\}\right)}{\text{std}\left(\{r_1, r_2, \cdots, r_G\}\right)}.$$

用户与助手之间的对话。用户提出一个问题,助手解决该问题。助手首先在思维过程中进行推理，然后向用户提供答案。推理过程和答案分别被包含在 <think> </think> 和 <answer> </answer> 标签内，即 <think> 这里的推理过程 </think> <answer> 这里的答案 </answer>。用户：提示。助手：

表 1 | DeepSeek-R1-Zero 模板。提示将在训练过程中被具体的推理问题替代。

### 2.2.2. 奖励建模

奖励是训练信号的来源，它决定了强化学习的优化方向。为了训练 DeepSeek-R1-Zero，我们采用一种基于规则的奖励系统，该系统主要由两类奖励构成：

- 精度奖励：精度奖励模型评估响应是否正确。例如，在具有确定性结果的数学问题中，模型需要以指定格式（例如，置于方框内）提供最终答案，从而实现可

靠基于规则的正确性验证。类似地，对于 LeetCode 问题，可以使用编译器根据预定义的测试用例生成反馈。

- 格式奖励：除了准确率奖励模型外，我们还采用一个格式奖励模型，强制模型将其思考过程放在'<think>' 和'</think>' 标签之间。

我们未在开发 DeepSeek-R1-Zero 时采用基于神经网络的奖励模型，因为我们发现，在大规模强化学习过程中，神经奖励模型可能存在奖励黑客问题，且重新训练奖励模型需要额外的训练资源，会使得整个训练流程更加复杂。

### 2.2.3. 训练模板

为了训练 DeepSeek-R1-Zero，我们首先设计了一个简单的模板，引导基础模型遵循我们指定的指令。如表 1 所示，该模板要求 DeepSeek-R1-Zero 首先生成推理过程，然后给出最终答案。我们特意将约束限制在这一结构格式上，避免任何内容相关的偏见——例如强制要求反思性推理或推广特定的解题策略——以确保我们能够准确观察模型在强化学习过程中的自然演进。

### 2.2.4. DeepSeek-R1-Zero 的性能、自演化过程及顿悟时刻

DeepSeek-R1-Zero 的性能如图 2 所示，展示了 DeepSeek-R1-Zero 在 RL 训练过程中对 AIME 2024 基准测试的性能轨迹。如图所示，随着 RL 训练的进行，DeepSeek-R1-Zero 的性能表现出稳定且持续的提升。值得注意的是，DeepSeek-R1-Zero 在 AIME 2024 上的平均 pass@1 得分显著提高，从初始的 15.6% 增加到令人印象深刻的 71.0%，达到了与 OpenAI-o1-0912 相当的性能水平。这一显著提升凸显了我们的 RL 算法在优化模型性能方面的有效性。

表 2 提供了 DeepSeek-R1-Zero 与 OpenAI 的 o1-0912 模型在多种推理相关基准测试中的对比分析。结果表明，强化学习能够

| 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 |

表 2 | DeepSeek-R1-Zero 和 OpenAI o1 模型在推理相关基准测试上的对比。

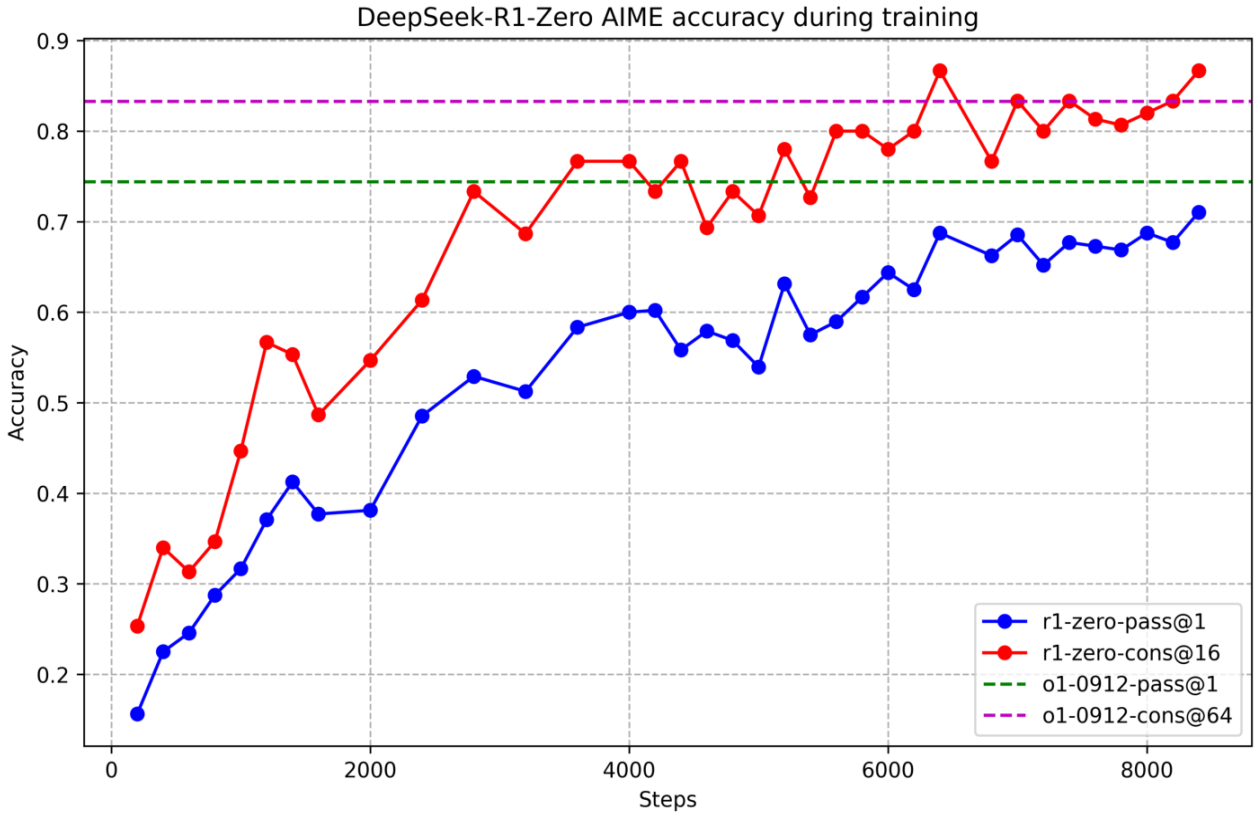


图 2 | DeepSeek-R1-Zero 在训练过程中的 AIME 准确率。对于每个问题，我们采样 16 个回答，并计算整体平均准确率以确保评估的稳定性。

DeepSeek-R1-Zero 在无需任何监督微调数据的情况下即可获得强大的推理能力。这一成果值得注意，因为它凸显了模型仅通过强化学习即可有效学习并泛化的能力。此外，通过应用多数投票机制，可以进一步提升 DeepSeek-R1-Zero 的性能。例如，在 AIME 基准上采用多数投票时，DeepSeek-R1-Zero 的性能从 71.0% 提升至 86.7%，从而超越了 OpenAI-o1-0912 的表现。DeepSeek-R1-Zero 在有无多数投票情况下均能取得如此具有竞争力的性能，凸显了其强大的基础能力以及在推理任务中进一步发展的潜力。

DeepSeek-R1-Zero 模型的自进化过程深度求解器 R1-Zero 模型的自进化过程是强化学习如何自主驱动模型提升其推理能力的一个引人入胜的展示。通过直接从基础模型启动强化学习，我们可以在不受到监督微调阶段影响的情况下，密切监控模型的演进过程。该方法清晰地展示了模型随时间演化的过程，尤其是在处理复杂推理任务方面的能力提升。

如图 3 所示，DeepSeek-R1-Zero 的思考时间表现出一致的提升。

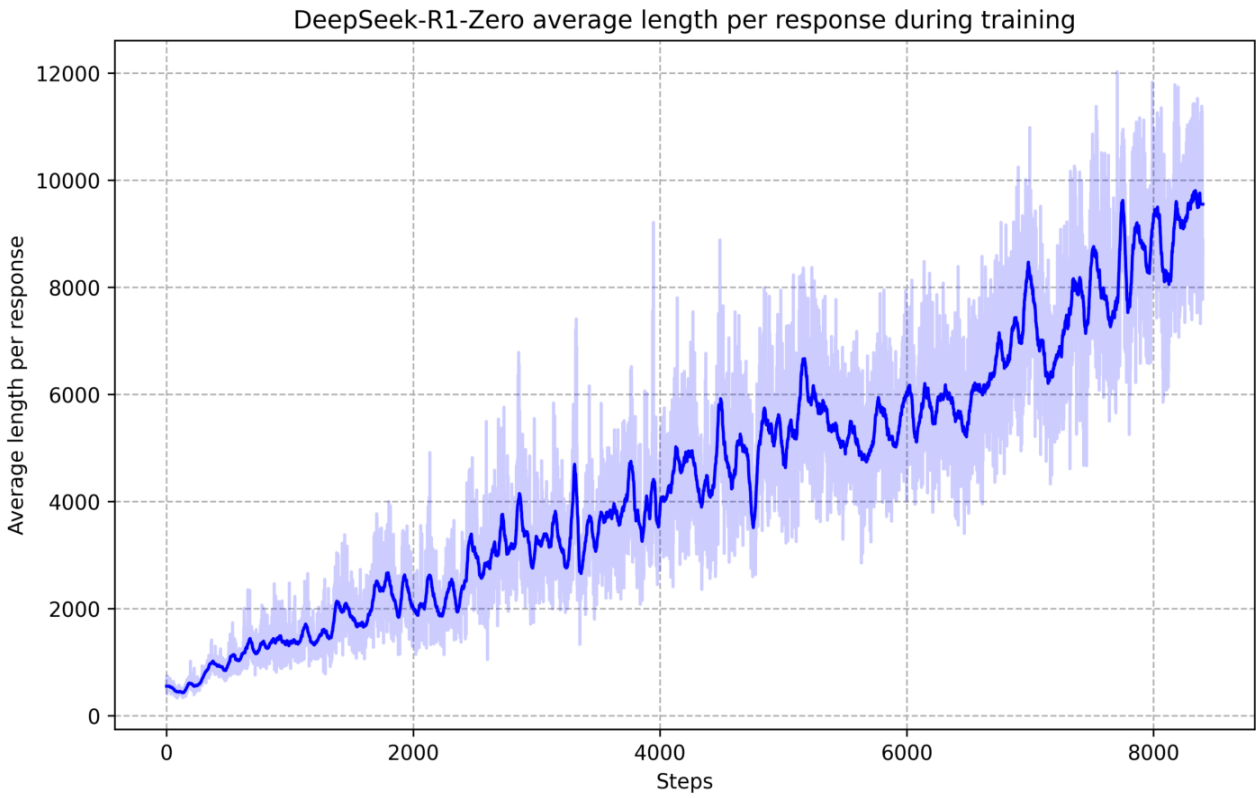


图 3 | DeepSeek-R1-Zero 在 RL 过程中训练集上的平均响应长度。DeepSeek-R1-Zero 自然地通过增加思考时间来学习解决推理任务。

在整个训练过程中持续进行。这一改进并非来自外部调整，而是模型内部的内在发展。DeepSeek-R1-Zero 通过利用扩展的推理阶段计算，自然地获得了解决日益复杂推理任务的能力。这种计算范围从生成数百到数千个推理标记，使模型能够更深入地探索和优化其思维过程。

自我演化的一个最显著特征是，随着测试时间计算量的增加，逐渐涌现出复杂的行为。例如，模型会进行反思——回顾并重新评估其先前的步骤——以及自发探索解决问题的替代方法。这些行为并非显式编程实现，而是源于模型与强化学习环境的交互作用。这种自发性的发展显著增强了 DeepSeek-R1-Zero 的推理能力，使其能够更高效、更准确地应对更具挑战性的任务。

DeepSeek-R1-Zero 模型训练中的“顿悟时刻”在 DeepSeek-R1-Zero 模型训练过程中观察到一个特别引人入胜的现象，即“顿悟时刻”的出现。如表 3 所示，这一时刻出现在模型的中间版本中。在该阶段，DeepSeek-R1-Zero 通过重新评估其初始方法，学会为一个问题分配更多的思考时间。这种行为不仅体现了模型推理能力的不断提升，也是强化学习能够导致出人意料且复杂结果的迷人例证。

这一时刻不仅对模型本身是“顿悟时刻”，也对观察其行为的研究人员具有重要意义。它凸显了强化学习的力量与美感：我们并非显式地教导模型如何解决问题，而是仅仅为其提供正确的激励，它便能自主发展出先进的问题解决策略。这一“顿悟时刻”有力地提醒我们，强化学习有潜力在人工系统中解锁新的智能水平，为未

来更加自主和适应性的模型铺平道路。问题：若 $a > 1$，则方程 $\sqrt{a - \sqrt{a + x}} = x$ 的实数解之和等于

响应：< 思考 >

为了解方程 $\sqrt{a - \sqrt{a + x}} = x$，我们从两边平方开始 $\cdots$

$(\sqrt{a - \sqrt{a + x}})^2 = x^2 \Longrightarrow a - \sqrt{a + x} = x^2.$

整理以分离内部平方根项：

$(a - x^2)^2 = a + x \Longrightarrow a^2 - 2ax^2 + (x^2)^2 = a + x \Longrightarrow x^4 - 2ax^2 - x + (a^2 - a) = 0$

等等，等等。等等。这里我可以标记一个“啊哈”时刻。

让我们逐步重新评估这一步，以确定正确的和是否为 $\cdots$

我们从以下方程开始：

$\sqrt{a - \sqrt{a + x}} = x$

首先，我们对两边平方：

$a - \sqrt{a + x} = x^2 \Longrightarrow \sqrt{a + x} = a - x^2$

接下来，我可以再次对两边进行平方，处理这个方程：$\cdots$

表 3 | 深度求索-R1-零模型一个中间版本中出现的一个有趣的“顿悟时刻”。模型学会了以拟人化的语气重新思考。这同样也是我们的一次顿悟，让我们见证了强化学习的力量与美感。

DeepSeek-R1-Zero 的局限性尽管 DeepSeek-R1-Zero 具有强大的推理能力，并能自主发展出意外且强大的推理行为，但它仍面临若干问题。例如，DeepSeek-R1-Zero 在可读性差和语言混杂等挑战上表现不佳。为使推理过程更加可读，并与开放社区共享，我们探索了 DeepSeek-R1，这是一种利用人类友好型冷启动数据的强化学习方法。

## 2.3. DeepSeek-R1：冷启动下的强化学习

受 DeepSeek-R1-Zero 预期成果的启发，两个自然的问题浮现：1）通过在冷启动阶段引入少量高质量数据，是否能够进一步提升推理性能或加快收敛速度？2）我们如何训练一个用户友好的模型，该模型不仅能够生成清晰且连贯的思维链（CoT），还能展现出强大的通用能力？为回答这些问题，我们设计了一条训练 DeepSeek-R1 的流程。该流程包含四个阶段，具体如下。

### 2.3.1. 冷启动

不同于 DeepSeek-R1-Zero，为防止基于基础模型的强化学习训练出现早期不稳定的冷启动阶段，我们为 DeepSeek-R1 构建并收集了一小部分长链思维（CoT）数据，用于微调模型作为初始的强化学习策略。为收集此类数据，我们探索了多种方法：使用少量样本提示结合长链思维作为示例，直接提示模型生成带有反思与

验证的详细答案，以可读格式收集 DeepSeek-R1Zero 的输出，并通过人工标注者的后处理对结果进行优化。

在本工作中，我们收集了数千条冷启动数据，以对 DeepSeek-V3-Base 进行微调，作为强化学习的起点。与 DeepSeek-R1-Zero 相比，冷启动数据的优势在于

# 包含：

- 可读性：DeepSeek-R1-Zero 的一个关键限制是其内容通常不适合阅读。回复可能混合多种语言，或缺乏标记语言格式以突出答案供用户查看。相比之下，在为 DeepSeek-R1 创建冷启动数据时，我们设计了一种可读性模式，即每个回复末尾包含摘要，并过滤掉不可读的回复。在此，我们定义输出格式为 | special_token | <reasoning_process> | special_token | <summary>，其中推理过程是查询的思维链（CoT），摘要用于总结推理结果。

- 潜在优势：通过精心设计融合人类先验的冷启动数据模式，我们观察到其在与 DeepSeek-R1-Zero 对比时表现更优。我们认为迭代训练是推理模型更好的方法。

### 2.3.2. 以推理为导向的强化学习

在基于冷启动数据对 DeepSeek-V3-Base 进行微调后，我们采用了与 DeepSeek-R1-Zero 中相同的大型强化学习训练过程。这一阶段的重点是提升模型的推理能力，尤其是在编码、数学、科学和逻辑推理等推理密集型任务中，这些任务涉及具有明确问题和清晰解法的场景。在训练过程中，我们观察到思维链（CoT）常常出现语言混合现象，特别是在强化学习提示涉及多种语言时。为缓解语言混合问题，我们在强化学习训练中引入了语言一致性奖励，该奖励计算为思维链中目标语言词的比例。尽管消融实验表明，这种对齐会导致模型性能轻微下降，但该奖励符合人类偏好，使输出更易读。最后，我们通过直接将推理任务的准确率与语言一致性奖励相加，形成最终的奖励。然后，我们对微调后的模型进行强化学习训练，直至其在推理任务上达到收敛。

### 2.3.3. 拒绝采样与监督微调

当基于推理的强化学习收敛时，我们利用得到的检查点来收集用于后续轮次的监督微调（SFT）数据。与最初冷启动数据主要侧重于推理不同，此阶段引入了来自其他领域的数据，以提升模型在写作、角色扮演及其他通用任务中的能力。具体而言，我们按照以下方式生成数据并微调模型。

推理数据我们通过从上述 RL 训练的检查点中进行拒绝采样，精心策划推理提示并生成推理轨迹。在上一阶段，我们仅包含能够使用基于规则的奖励进行评估的数据。然而，在本阶段，我们通过引入额外数据来扩展数据集，其中部分数据采用

生成式奖励模型，将真实值和模型预测输入 DeepSeek-V3 进行判断。此外，由于模型输出有时混乱且难以阅读，我们已过滤掉混合语言、长段落和代码块的思维链。对于每个提示，我们采样多个响应并仅保留正确的响应。总计我们收集了约 600k 个与推理相关的训练样本。非推理数据对于非推理数据，如写作、事实问答、自我认知和翻译，我们采用 DeepSeek-V3 流程，并复用 DeepSeek-V3 的 SFT 数据集部分。对于某些非推理任务，我们通过提示调用 DeepSeek-V3 生成潜在的思维链，然后再回答问题。然而，对于简单的查询，如“hello”，我们不会在回复中提供思维链。最终，我们收集了大约 20 万条与推理无关的训练样本。

我们使用上述约 80 万样本的精选数据集，对 DeepSeek-V3-Base 进行两轮微调。

### 2.3.4. 所有场景的强化学习

为使模型更符合人类偏好，我们实施了一个二次强化学习阶段，旨在同时提升模型的有用性和无害性，并优化其推理能力。具体而言，我们结合奖励信号和多样化的提示分布来训练模型。在推理数据方面，我们遵循 DeepSeek-R1-Zero 中提出的方法，采用基于规则的奖励来指导数学、代码和逻辑推理领域的学习过程。在通用数据方面，我们采用奖励模型来捕捉复杂且细微场景中的人类偏好。我们基于 DeepSeek-V3 的流水线，并采用相似的偏好对和训练提示分布。在有用性方面，我们仅关注最终摘要，确保评估侧重于响应对用户的有效性和相关性，同时最小化对底层推理过程的干扰。在无害性方面，我们对模型的整个响应进行评估，包括推理过程和摘要，以识别并缓解生成过程中可能出现的风险、偏见或有害内容。最终，奖励信号与多样化数据分布的结合，使我们能够训练出在推理方面表现优异，同时优先保证有用性和无害性的模型。

## 2.4. 混合蒸馏：用推理能力赋能小模型

为了使更高效的小型模型具备类似 DeepSeek-R1 的推理能力，我们直接使用 DeepSeek-R1 精选的 80 万样本对 Qwen（Qwen, 2024b）和 Llama（AI@Meta, 2024）等开源模型进行微调，具体细节见 §2.3.3。我们的研究结果表明，这种简单的蒸馏方法显著提升了小型模型的推理能力。本文所使用的基模型包括 Qwen2.5-Math-1.5B、Qwen2.5-Math-7B、Qwen2.514B、Qwen2.5-32B、Llama-3.1-8B 和 Llama-3.3-70B-Instruct。我们选择 Llama-3.3 是因为其推理能力略优于 Llama-3.1。

对于蒸馏模型，我们仅应用 SFT，不包含强化学习阶段，即使引入强化学习阶段可以显著提升模型性能。我们的主要目标在于展示蒸馏技术的有效性，将强化学习阶段的探索留给更广泛的研究社区。

# 3. 实验

基准测试我们在 MMLU (Hendrycks 等, 2020)、MMLU-Redux (Gema 等, 2024)、MMLU-Pro (Wang 等, 2024)、C-Eval (Huang 等, 2023)、CMMLU (Li 等, 2023)、IFEval (Zhou 等, 2023)、FRAMES (Krishna 等, 2024)、GPQA Diamond (Rein 等, 2023)、SimpleQA (OpenAI, 2024c)、C-SimpleQA (He 等, 2024)、SWE-Bench Verified (OpenAI, 2024d)，Aider [1]，LiveCodeBench（Jain 等，2024）（2024-08 - 2025-01），Codeforces [2]，中国全国高中数学联赛（CNMO 2024）[3]，以及 2024 年美国数学邀请赛（AIME 2024）（MAA，2024）。除了标准基准外，我们还使用大语言模型作为评判者，在开放性生成任务上评估我们的模型。具体而言，我们遵循 AlpacaEval 2.0（Dubois 等，2024）和 Arena-Hard（Li 等，2024）的原始配置，这些配置利用 GPT-4-Turbo-1106 作为评判者进行成对比较。在此，我们仅将最终总结输入评估以避免长度偏差。对于蒸馏模型，我们在 AIME 2024、MATH-500、GPQA Diamond、Codeforces 和 LiveCodeBench 上报告代表性结果。

评估提示在 DeepSeek-V3 的设置下，使用 simpleevals 框架的提示对 MMLU、DROP、GPQA Diamond 和 SimpleQA 等标准基准进行评估。对于 MMLU-Redux，我们在零样本设置下采用 Zero-Eval 提示格式（Lin, 2024）。对于 MMLU-Pro、C-Eval 和 CLUE-WSC，由于原始提示是少样本提示，我们将其略微修改为零样本设置。少样本中的思维链（CoT）可能会损害 DeepSeek-R1 的性能。其他数据集则遵循其原始评估协议，并使用其创建者提供的默认提示。对于代码和数学基准，HumanEval-Mul 数据集涵盖了八种主流编程语言（Python、Java、C++、C#、JavaScript、TypeScript、PHP 和 Bash）。模型在 LiveCodeBench 上的性能评估采用 CoT 格式，数据收集时间为 2024 年 8 月至 2025 年 1 月。Codeforces 数据集采用 10 场 Div. 2 比赛中的题目以及专家设计的测试用例进行评估，随后计算出预期的排名和参赛者的百分比。SWE-Bench 的验证结果通过无代理框架（Xia et al., 2024）获得。与 AIDER 相关的基准采用“diff”格式进行测量。DeepSeek-R1 在每个基准上的输出被限制在最多 32,768 个 token。

基线模型我们针对若干强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。由于在大陆地区访问 OpenAI-o1-1217 API 存在困难，我们根据官方报告报告其性能。对于蒸馏模型，我们还与开源模型 QwQ-32B-Preview（Qwen, 2024a）进行了对比。

评估设置我们将模型的最大生成长度设置为 32,768 个标记。我们发现，使用贪婪解码来评估长输出推理模型会导致较高的重复率以及在不同检查点之间显著的差异性。因此，我们默认采用 pass@ $k$ 评估（Chen 等，2021），并报告使用非零温度的 pass@ 1。具体而言，我们使用采样温度为 0.6 和 top- $p$ 值为 0.95，为每个问题生成 $k$ 个响应（通常在 4 到 64 之间，具体取决于测试集大小）。pass@ 1 的计算方式如下：

$$\text{pass@1} = \frac{1}{k}\sum_{i=1}^{k} p_i,$$

其中 $p_i$ 表示第 $i$ 个响应的正确性。该方法提供了更可靠的性能估计。对于 AIME 2024，我们还报告了使用 64 个样本的共识（多数投票）结果（Wang 等，2022），记为 cons@64。

$$\text{pass@ } 1 = \frac{1}{k}\sum\_i = 1^k p\_i,$$

## 3.1. DeepSeek-R1 评估

| | 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 |

[01] https://aider.chat

[2] https://codeforces.com

[3] https://www.cms.org.cn/Home/comp/comp/cid/12.html

表 4 | 深度求解器-R1 与其他代表性模型的对比。

对于面向教育的知识基准，如 MMLU、MMLU-Pro 和 GPQA Diamond，DeepSeek-R1 相比 DeepSeek-V3 表现出更优的性能。这一提升主要归因于在与 STEM 相关问题上的准确率提高，通过大规模强化学习实现了显著进步。此外，DeepSeek-R1 在 FRAMES 这一依赖长上下文的问答任务上表现出色，展现了其强大的文档分析能力。这凸显了推理模型在人工智能驱动的搜索与数据分析任务中的潜力。在事实性基准 SimpleQA 上，DeepSeek-R1 超越了 DeepSeek-V3，证明了其处理基于事实的查询的能力。在该基准上，OpenAI-o1 也优于 GPT-4o。然而，DeepSeek-R1 在中文 SimpleQA 基准上的表现逊于 DeepSeek-V3，主要原因是其在安全强化学习后倾向于拒绝回答某些问题。如果没有安全强化学习，DeepSeek-R1 的准确率可超过 70%。

DeepSeek-R1 在 IF-Eval 上也取得了令人印象深刻的结果，IF-Eval 是一个旨在评估模型遵循格式指令能力的基准测试。这些改进可以追溯到在监督微调（SFT）和强化学习（RL）训练的最后阶段引入了遵循指令的数据。此外，在 AlpacaEval2.0 和 ArenaHard 上观察到显著的性能表现，表明 DeepSeek-R1 在写作任务和开放域问答方面具有优势。其显著优于 DeepSeek-V3 的表现凸显了大规模强化学习的泛化优势，不仅提升了推理能力，也改善了在不同领域中的表现。此外，DeepSeek-R1 生成的摘要长度简洁，ArenaHard 上的平均长度为 689 个标记，AlpacaEval 2.0 上的平均长度为 2,218 个字符。这表明 DeepSeek-R1 在基于 GPT 的评估中避免了引入长度偏差，进一步巩固了其在多个任务上的鲁棒性。

在数学任务上，DeepSeek-R1 的表现与 OpenAI-o1-1217 相当，显著优于其他模型。在编码算法任务上，如 LiveCodeBench 和 Codeforces，也观察到类似的趋势，以推理为导向的模型在这些基准测试中占据主导地位。在面向工程的编码任务上，OpenAI-o1-1217 在 Aider 上优于 DeepSeek-R1，但在 SWE Verified 上表现相当。我们相信，DeepSeek-R1 在工程方面的表现将在下一个版本中得到提升，因为目前相关的强化学习训练数据量仍然非常有限。

### 3.2. 模型蒸馏评估

| 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 |

表 5 | 在与推理相关的基准测试中，DeepSeek-R1 蒸馏模型与其他可比模型的对比。

如表 5 所示，仅通过蒸馏 DeepSeek-R1 的输出，即可使高效的 DeepSeek-R1-7B（即 DeepSeek-R1-Distill-Qwen-7B，下文简写类似）在各项指标上超越非推理模型如 GPT-4o-0513。DeepSeek-R1-14B 在所有评估指标上均优于 QwQ-32BPreview，而 DeepSeek-R1-32B 和 DeepSeek-R1-70B 在多数基准测试中显著超越 o1-mini。这些结果展示了蒸馏的强大潜力。此外，我们发现将强化学习应用于这些蒸馏模型可带来显著的进一步提升。我们认为这值得进一步探索，因此本文仅展示简单 SFT 蒸馏模型的结果。

# 4. 讨论

## 4.1. 混合与强化学习

在第 3.2 节中，我们可以看到通过蒸馏 DeepSeek-R1，小模型能够取得令人印象深刻的结果。然而，仍有一个问题有待解答：论文中讨论的通过大规模强化学习训练是否也能在不进行蒸馏的情况下实现相近的性能？

为回答这个问题，我们使用数学、代码和 STEM 数据对 Qwen-32B-Base 进行大规模强化学习训练，训练超过 10000 步，得到 DeepSeek-R1-Zero-Qwen-32B。实验结果如表 6 所示，表明 32B 基础模型在大规模训练后，

| 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** |

表 6 | 在推理相关基准上蒸馏模型与强化学习模型的对比。

RL 训练达到了与 QwQ-32B-Preview 相当的性能。然而，从 DeepSeek-R1 蒸馏而来的 DeepSeek-R1-Distill-Qwen-32B 在所有基准测试中均显著优于 DeepSeek-R1-Zero-Qwen-32B。

因此，我们可以得出两个结论：首先，将更强大的模型蒸馏为更小的模型能够取得优异的效果，而依赖于本文提到的大规模强化学习的较小模型则需要巨大的计算资源，甚至可能无法达到蒸馏的效果。其次，尽管蒸馏策略在经济性和有效性方面表现良好，但要超越智能的边界，仍可能需要更强大的基础模型和大规模强化学习。

## 4.2. 失败的尝试

在开发 DeepSeek-R1 的早期阶段，我们也遇到了各种失败和挫折。我们在此分享我们的失败经验以提供借鉴，但这并不意味着这些方法无法用于构建有效的推理模型。

过程奖励模型（PRM）PRM 是一种合理的方法，能够引导模型朝着更优的路径解决推理任务（Lightman 等，2023；Uesato 等，2022；Wang 等，2023）。然而，在实践中，PRM 存在三个主要局限性，可能会阻碍其最终成功。首先，一般推理中显式定义细粒度步骤具有挑战性。其次，判断当前中间步骤是否正确是一项具有挑战性的任务。使用模型进行自动化标注可能无法获得令人满意的结果，而人工标注则不利于规模化。第三，一旦引入基于模型的 PRM，不可避免地会导致奖励黑客行为（Gao 等，2022），并且重新训练奖励模型需要额外的训练资源，还会使整个训练流程复杂化。综上所述，尽管 PRM 展现出良好的能力，能够对模型生成的前 N 个响应进行重排序或协助引导搜索（Snell 等，2024），但在我们的实验中，其优势相较于在大规模强化学习过程中引入的额外计算开销而言仍然有限。

受 AlphaGo（Silver 等，2017b）和 AlphaZero（Silver 等，2017a）启发，我们探索使用蒙特卡洛树搜索（MCTS）来提升测试时计算可扩展性。该方法涉及将答案分解为更小的部分，以允许模型系统地探索解空间。为实现这一点，我们提示模型生成多个标签，这些标签对应于搜索过程中所需的特定推理步骤。在训练阶段，我们首先利用收集到的提示，通过由预训练价值模型引导的 MCTS 找到答案。随后，我们使用生成的问题-答案对来训练策略模型和价值模型，迭代优化该过程。

然而，当扩展训练规模时，这种方法面临若干挑战。首先，与棋类游戏不同，棋类游戏的搜索空间相对明确，而令牌生成则呈现了搜索空间呈指数级增长。为解决这一问题，我们为每个节点设置最大扩展限制，但这可能导致模型陷入局部最优。其次，价值模型直接决定了生成质量，因为它指导搜索过程中的每一步。训练一个细粒度的价值模型本质上是困难的，这使得模型难以迭代改进。尽管 AlphaGo 的核心成功依赖于训练价值模型以逐步提升其性能，但由于标记生成的复杂性，这一原则在我们的设置中难以复制。

总之，尽管 MCTS 在与预训练的价值模型结合时能够提升推理过程中的性能，

但通过自搜索持续提升模型性能仍是一个重大挑战。

# 5. 结论、局限性与未来工作

在本工作中，我们分享了通过强化学习提升模型推理能力的探索历程。DeepSeek-R1-Zero 代表了一种不依赖冷启动数据的纯强化学习方法，在多种任务上均取得了优异表现。DeepSeek-R1 更为强大，结合了冷启动数据与迭代强化学习微调。最终，DeepSeek-R1 在一系列任务上的表现与 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 800 K 训练样本，并对若干小型密集模型进行微调。结果令人鼓舞：DeepSeek-R1-Distill-Qwen-1.5B 在数学基准测试中优于 GPT-4o 和 Claude-3.5-Sonnet，其在 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印象深刻的结果，显著优于基于相同基础检查点的其他指令微调模型。

未来，我们计划在以下方向上对 DeepSeek-R1 进行研究投资。

- 通用能力：目前，DeepSeek-R1 在函数调用、多轮对话、复杂角色扮演以及 JSON 输出等任务上的能力尚不及 DeepSeek-V3。未来，我们计划探索如何利用更长的思维链（CoT）来提升这些领域的任务表现。

- 语言混合：DeepSeek-R1 当前针对中文和英文进行了优化，处理其他语言查询时可能会出现语言混合问题。例如，DeepSeek-R1 可能会在推理和回复中使用英文，即使查询语言并非英文或中文。我们计划在未来的更新中解决这一限制。

- 提示工程：在评估 DeepSeek-R1 时，我们观察到其对提示词较为敏感。少样本提示会持续降低其性能。因此，我们建议用户直接描述问题并使用零样本设置指定输出格式，以获得最佳效果。

- 软件工程任务：由于评估时间较长，影响了强化学习过程的效率，大规模强化学习尚未在软件工程任务中得到广泛应用。因此，DeepSeek-R1 在软件工程基准上的表现并未显著优于 DeepSeek-V3。未来版本将通过在软件工程数据上实施拒绝采样，或在强化学习过程中引入异步评估来提升效率。