

DeepSeek-R1: 通过强化学习激励 LLM 的推理能力

DeepSeek-AI

research@deepseek.com

摘要

通用推理是人工智能领域长期存在且极具挑战的问题。近期，以大型语言模型（LLM）（Brown et al., 2020; OpenAI, 2023）和思维链提示（Wei et al., 2022b）为代表的突破，在基础推理任务上取得了相当大的成功。然而，这种成功高度依赖大量人工标注的示范，而模型处理更复杂问题的能力仍然不足。在此我们展示，LLM 的推理能力可以通过纯强化学习（RL）得到激励，从而无需人工标注的推理轨迹。所提出的 RL 框架促进了高级推理模式的涌现发展，例如自我反思、验证和动态策略适应。因此，训练后的模型在数学、编程竞赛和 STEM 领域等可验证任务上取得了优异性能，超越了通过传统的人类示范监督学习训练的同类模型。此外，这些大规模模型所展现的涌现推理模式还可以被系统地利用，以引导并增强小型模型的推理能力。

1. 引言

推理能力是人类智能的基石，使人能够完成从数学问题求解到逻辑推演和编程等复杂认知任务。人工智能的最新进展表明，当大语言模型（LLM）扩展到足够规模时，能够展现包括推理能力在内的涌现行为（Kaplan et al., 2020; Wei et al., 2022a）。然而，在预训练中实现此类能力通常需要大量计算资源。与此同时，另一条互补的研究路线表明，通过思维链（CoT）提示可以有效增强大语言模型。这项技术要么提供精心设计的少样本示例，要么使用“让我们一步一步思考”等极简提示（Kojima et al., 2022; Wei et al., 2022b），使模型能够生成中间推理步骤，从而显著提升其在复杂任务上的表现。同样，当模型在后训练阶段学习高质量、多步骤的推理轨迹时，也能获得进一步的性能提升（Chung et al., 2024; OpenAI, 2023）。尽管这些方法有效，但仍存在明显局限。它们依赖人工标注的推理轨迹，阻碍了可扩展性并引入认知偏差。此外，由于限制模型去复制人类思维过程，其性能天然受人类

所提供示例的限制，从而无法探索更优的、非人类式的推理路径。

为解决这些问题，我们旨在探索 LLM 在 RL 框架中通过自我进化发展推理能力的潜力，同时尽量减少对人工标注工作的依赖。具体而言，我们以 DeepSeek-V3-Base (DeepSeek-AI, 2024b) 为基础，并采用群体相对策略优化 (GRPO) (Shao et al., 2024) 作为 RL 框架。奖励信号完全基于最终预测与真实答案的一致性，而不对推理过程本身施加约束。值得注意的是，我们跳过了 RL 训练前的传统监督微调 (SFT) 阶段。这一设计选择源于我们的假设：人为定义的推理模式可能限制模型探索，而不受限制的 RL 训练能够更好地激励 LLM 涌现新的推理能力。通过第 2 节所述过程，我们的模型（称为 DeepSeek-R1-Zero）自然形成了多样且复杂的推理行为。在解决推理问题时，模型倾向于生成更长的响应，并在每个响应中融入验证、反思和替代方法探索。尽管我们没有明确教导模型如何推理，但它通过强化学习成功学会了更优的推理策略。

尽管 DeepSeek-R1-Zero 展现出出色的推理能力，但它也面临可读性差和语言混杂等挑战，偶尔会在同一条思维链响应中混合使用英语和中文。此外，DeepSeek-R1-Zero 的基于规则的 RL 训练阶段狭隘地聚焦于推理任务，导致其在写作和开放域问答等更广泛领域的性能有限。为应对这些挑战，我们引入了 DeepSeek-R1，这是一种通过多阶段学习框架训练的模型，该框架整合了拒绝采样、强化学习和监督微调，详见第 3 节。这一训练流程使 DeepSeek-R1 能够继承其前身 DeepSeek-R1-Zero 的推理能力，同时通过额外的非推理数据使模型行为与人类偏好保持一致。

为了以更低的能耗让更多人获得强大的 AI，我们蒸馏了多个更小的模型并将其公开。这些蒸馏模型展现出强大的推理能力，性能超过其原始指令微调版本。我们相信，这些指令微调版本还将通过提供理解长思维链 (CoT) 推理模型底层机制的宝贵资源，并促进更强大推理模型的发展，为研究社区作出重要贡献。我们在 <https://huggingface.co/deepseek-ai> 发布 DeepSeek-R1 系列模型。

2. DeepSeek-R1-Zero

我们首先详细介绍 DeepSeek-R1-Zero 的训练过程，该过程完全依赖强化学习，而不进行监督微调。为实现大规模 RL 的高效运行，我们采用群组相对策略优化 (GRPO) (Shao 等, 2024)。

2.1. 群组相对策略优化

GRPO (Shao et al., 2024) 是我们用于训练 DeepSeek-R1-Zero 和 DeepSeek-R1 的强化学习算法。它最初被提出用于简化近端策略优化 (PPO) (Schulman et al., 2017) 的训练过程并降低资源消耗，而 PPO 广泛应用于 LLM 的 RL 阶段 (Ouyang et al., 2022)。

对于每个问题 q ，GRPO 从旧策略 $\pi_{\theta_{old}}$ 中采样一组输出 $\{o_1, o_2, \dots, o_G\}$ ，然后通过最大化以下目标来优化策略模型 π_{θ} ：

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

其中， π_{ref} 是参考策略， ε 和 β 是超参数， A_i 是优势，它通过一组奖励 $\{r_1, r_2, \dots, r_G\}$ 计算得到，这些奖励对应于每个组内的输出：

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$

我们在补充材料 A.3 中比较了 GRPO 和 PPO。为训练 DeepSeek-R1-Zero，我们将学习率设为 $3e-6$ ，KL 系数设为 0.001，采样轨迹的温度设为 1。对于每个问题，在 8.2k 步之前，我们采样 16 个输出，最大长度为 32,768 个 token；之后最大长度为 65,536 个 token。因此，在 8.2k 步时，DeepSeek-R1-Zero 的性能和响应长度都出现了显著跃升，训练总计进行 10,400 步，对应 1.6 个训练 epoch。每个训练步包含 32 个独特问题，因此训练批次大小为 512。每 400 步，我们用最新的策略模型替换参考模型。为加速训练，每次采样轨迹生成 8,192 个输出，并将其随机分为 16 个小批次，仅训练一个内部 epoch。

表 1 | DeepSeek-R1-Zero 的模板。prompt 将在训练期间替换为具体的推理问题。

用户与助手之间的对话。用户提出问题，助手进行解答。助手首先在心中思考推理过程，然后向用户提供答案。推理过程和答案分别包含在 `<think>...</think>` 和 `<answer>...</answer>` 标签中，即 `<think>` 此处为推理过程 `</think>``<answer>` 此处为答案 `</answer>`。用户：prompt。
助手：

我们的高性能 RL 基础设施见补充材料 B.1，确保训练具有可扩展性和高效率。

2.2. 奖励设计

奖励是训练信号的来源，决定 RL 优化的方向。对于 DeepSeek-R1-Zero，我们采用基于规则的奖励，为数学、编程和逻辑推理领域的数据提供精确反馈。我们的基于规则的奖励系统主要由两类奖励组成：准确性奖励和格式奖励。

准确率奖励 用于评估回答是否正确。例如，对于具有确定性结果的数学问题，模型需要以指定格式（例如放在方框内）提供最终答案，从而能够通过基于规则的验证可靠地检查正确性。同样，对于编程竞赛提示，可以利用编译器根据一组预定义测试用例评估模型的

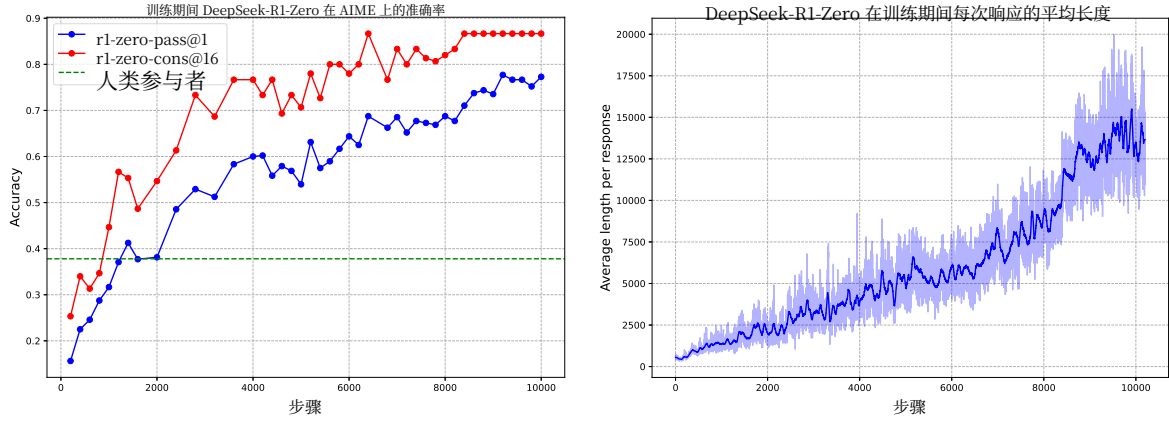


图 1 | (a) DeepSeek-R1-Zero 在训练期间的 AIME 准确率。AIME 以一道数学题作为输入，并输出一个数字，如表 32 所示。Pass@1 和 Cons@16 的定义见补充材料 D.1。基线是人类参与者在 AIME 竞赛中取得的平均分数。(b) RL 过程中 DeepSeek-R1-Zero 在训练集上的平均响应长度。DeepSeek-R1-Zero 自然学会用更多思考时间来解决推理任务。请注意，一个训练步骤指一次策略更新操作。

回答，从而对正确性生成客观反馈。

格式奖励 通过强制执行特定格式要求来补充准确性奖励模型。具体而言，模型被激励将其推理过程封装在指定标签内，即“<think>”和“</think>”。这确保模型的思维过程得到明确划分，从而增强可解释性并促进后续分析。

$$Reward_{rule} = Reward_{acc} + Reward_{format} \quad (4)$$

准确性奖励、奖励和格式奖励以相同权重组合。值得注意的是，我们不在推理任务中应用神经奖励模型——无论是基于结果的还是基于过程的。此决定基于我们的观察：在大规模强化学习期间，神经奖励模型容易受到奖励劫持。此外，重新训练此类模型需要大量计算资源，并会给训练流程引入额外复杂性，从而使整体优化过程更加复杂。

2.3. 激励 LLM 的推理能力

具体而言，我们在 DeepSeek-V3 基础模型上应用 RL 技术来训练 DeepSeek-R1-Zero。在训练期间，我们设计了一个直接的模板，要求 DeepSeek-R1-Zero 首先生成推理过程，然后给出最终答案。我们有意将约束限制在这种结构格式上，避免任何特定内容的偏差，以确保能够准确观察模型在 RL 过程中的自然演进。

图 1(a) 展示了 DeepSeek-R1-Zero 在整个 RL 训练过程中 AIME 2024 基准上的性能轨迹，其中 AIME 2024 的平均 pass@1 得分显著提升，从最初的 15.6% 跃升至 77.9%。此外，通过利用自洽解码（self-consistency decoding）（Wang et al., 2023c），模型性能还可以

表 2 | DeepSeek-R1-Zero 中间版本的一个有趣“顿悟时刻”。模型学会了以拟人化的语气重新思考。这对我们而言也是一个顿悟时刻，使我们得以见证强化学习的力量与美妙。

问题: 如果 $a > 1$, 那么 $\sqrt{a} - \sqrt{a+x} = x$ 的实数解之和等于_____

回答: <think>
 为求解方程 $\sqrt{a} - \sqrt{a+x} = x$, 我们先对两边平方...

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

重新排列以分离内部平方根项:
 $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
 ...
 等等, 等等. 等等。这是一个我可以在此标记的顿悟时刻。
 让我们逐步重新评估这一步骤, 以确定正确的总和是否可以...
 我们从以下方程开始:
 $\sqrt{a} - \sqrt{a+x} = x$
 首先, 对两边平方:
 $a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$
 接下来, 我可以再次对两边平方, 将方程视为: ...
 ...

进一步提升, 准确率达到 86.7%。这一表现显著超过所有人类参赛者的平均水平。除数学竞赛外, 如图 10 所示, DeepSeek-R1-Zero 在编程竞赛以及研究生水平的生物学、物理学和化学问题上也取得了卓越表现。这些结果凸显了 RL 在增强大语言模型推理能力方面的有效性。

DeepSeek-R1-Zero 的自我进化 exemplifies 了 RL 如何自主增强模型的推理能力。

如图 1(b) 所示, DeepSeek-R1-Zero 在整个训练过程中表现出思考时间的稳步增加, 这完全由内在适应驱动, 而非外部修改。借助长 CoT, 模型逐步完善其推理过程, 生成数百到数千个 token, 以探索并改进其问题求解策略。

思考时间的增加促进了复杂行为的自主发展。具体而言, DeepSeek-R1-Zero 越来越多地展现出高级推理策略, 例如反思式推理和对替代解决方案的系统探索 (详见补充材料 C.2 中的图 9(a)), 显著提升了其在数学和编程等可验证任务上的性能。值得注意的是, 在训练过程中, DeepSeek-R1-Zero 出现了一个“顿悟时刻” (表 2), 其特征是在反思过程中突然增加了对“wait”一词的使用 (详见补充材料 C.2 中的图 9(b))。这一时刻标志着推理模式发生明显变化, 并清晰展示了 DeepSeek-R1-Zero 的自我进化过程。

DeepSeek-R1-Zero 的自我进化凸显了 RL 的力量与美妙之处: 我们不是明确教授模型如何解决问题, 而只是为其提供正确的激励, 模型便会自主发展出先进的问题解决策略。这提醒我们, RL 有潜力释放 LLM 更高层次的能力, 为未来更加自主和自适应的模型铺平道路。

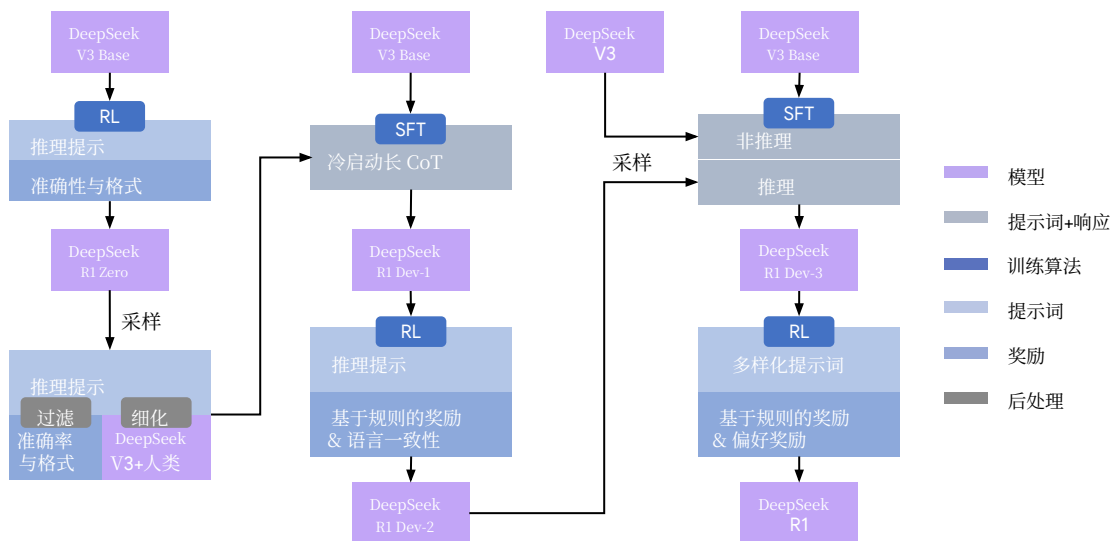


图 2 | DeepSeek-R1 的多阶段流程。关于 DeepSeek-V3 Base 和 DeepSeek-V3 的详细背景介绍见补充材料 A.1。模型 DeepSeek-R1 Dev1、Dev2 和 Dev3 表示该流程中的中间检查点。

3. DeepSeek-R1

尽管 DeepSeek-R1-Zero 展现出强大的推理能力，但仍面临一些问题。由于 DeepSeek-V3-Base 使用多种语言进行训练，尤其是英语和中文，DeepSeek-R1-Zero 在可读性差、语言混杂等挑战上表现不佳。为了解决这些问题，我们开发了 DeepSeek-R1，其流程如图 2 所示。

在初始阶段，我们收集了数千条体现对话式、人类对齐思维过程的冷启动数据。随后应用 RL 训练，以通过对话式思维过程和语言一致性来提升模型性能。之后，我们再次进行拒绝采样和 SFT。该阶段将推理和非推理数据集都纳入 SFT 过程，使模型不仅能够出色完成推理任务，还能展现先进的写作能力。为进一步使模型与人类偏好对齐，我们实施了第二阶段 RL，旨在增强模型的有用性和无害性，同时进一步完善其推理能力。

本节其余部分将详细介绍该流程的关键组成部分：第 3.1 节介绍我们在 RL 阶段使用的奖励模型，第 3.2 节详述具体的训练方法和实现细节。本阶段使用的数据详见补充材料 B.3。

3.1. 基于模型的奖励

对于通用数据，我们采用奖励模型来捕捉复杂且细致场景中的人类偏好。我们以 DeepSeek-V3 流程为基础，并采用类似的偏好对和训练提示分布。对于有用性，我们只关注最终摘要，确保评估强调回答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，我们评估模型的完整回答，包括推理过程和摘要，以识别并缓解可能出现的潜在风险、偏见或有害内容

在生成过程中。

有用性奖励模型 关于有用性奖励模型训练，我们首先使用补充材料 B.2 所列的 arena-hard 提示格式提示 DeepSeek-V3，生成偏好对，其中每一对包含一个用户查询和两个候选回答。对于每个偏好对，我们向 DeepSeek-V3 查询四次，并随机将回答指定为回答 A 或回答 B，以减轻位置偏差。最终偏好分数通过对四次独立判断取平均确定，仅保留分数差 (Δ) 超过 1 的偏好对，以确保存在有意义的区分。此外，为了最大限度减少与长度相关的偏差，我们确保整个数据集中的选中回答和拒绝回答长度相近。总计，我们整理了 66,000 个数据对用于训练奖励模型。该数据集中的提示均为非推理问题，来源要么是公开可用的开源数据集，要么来自明确同意为模型改进目的分享其数据的用户。我们的奖励模型架构与 DeepSeek-R1 一致，增加了一个用于预测标量偏好分数的奖励头。

$$Reward_{helpful} = RM_{helpful}(Response_A, Response_B) \quad (5)$$

有用性奖励模型的训练批量大小为 256，学习率为 6e-6，并在训练数据集上训练一个 epoch。训练期间的最大序列长度设为 8192 tokens，而奖励模型推理期间不施加显式限制。

安全奖励模型 为了评估和提升模型安全性，我们整理了一个包含 106,000 个提示的数据集，并依据预定义的安全准则，将模型生成的响应标注为“安全”或“不安全”。不同于有用性奖励模型采用的成对损失，安全奖励模型使用逐点评估方法进行训练，以区分安全和不安全响应。训练超参数与有用性奖励模型相同。

$$Reward_{safety} = RM_{safety}(Response) \quad (6)$$

对于一般查询，每个实例都会被归类为安全数据集或有用性数据集。分配给每个查询的一般奖励 $Reward_{General}$ 对应于相关数据集中定义的相应奖励。

3.2. 训练细节

3.2.1. 第一阶段 **RL** 的训练细节

在 RL 的第一阶段，我们将学习率设为 3e-6，KL 系数设为 0.001，GRPO 裁剪比率 ϵ 设为 10，采样温度设为 1，用于采样轨迹。对于每个问题，我们采样 16 个输出，最大长度为 32,768。每个训练步骤包含 32 个不同问题，因此每步训练批次大小为 512。每 400 步，我们用最新策略模型替换参考模型。为加速训练，每次采样轨迹生成 8,192 个输出，并将其随机分为 16 个小批次，仅训练一个内部周期。然而，为缓解语言混杂问题，我们在 RL 训练期间引入语言一致性奖励，其计算方式为 CoT 中目标语言词语所占的比例。

$$Reward_{language} = \frac{Num(Words_{target})}{Num(Words)} \quad (7)$$

尽管补充材料 B.6 中的消融实验表明，这种对齐会导致模型性能略有下降，但该奖励符合人类偏好，使输出更具可读性。我们将语言一致性奖励应用于推理数据和非推理数据，直接将其加入最终奖励。

需要注意的是，裁剪比例在训练中发挥着关键作用。较低的值可能导致大量 token 的梯度被截断，从而降低模型性能；而较高的值可能导致训练过程不稳定。

3.2.2. 第二 *RL* 阶段的训练细节

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。对于推理数据，我们遵循 DeepSeek-R1-Zero 中概述的方法，使用基于规则的奖励来指导数学、编程和逻辑推理领域的学习。在训练过程中，我们观察到 CoT 经常出现语言混杂，尤其是在 RL 提示涉及多种语言时。对于通用数据，我们利用奖励模型来指导训练。最终，将奖励信号与多样化的数据分布相结合，使我们能够开发出不仅擅长推理，而且优先考虑有帮助性和无害性的模型。给定一个数据批次，奖励可表示为

$$Reward = Reward_{\text{reasoning}} + Reward_{\text{general}} + Reward_{\text{language}} \quad (8)$$

$$\text{where, } Reward_{\text{reasoning}} = Reward_{\text{rule}} \quad (9)$$

$$Reward_{\text{general}} = Reward_{\text{reward\_model}} + Reward_{\text{format}} \quad (10)$$

RL 的第二阶段保留了第一阶段的大部分参数，关键区别在于将温度降低到 0.7，因为我们发现该阶段较高的温度会导致生成结果不连贯。该阶段共包含 1,700 个训练步骤，其中通用指令数据和基于偏好的奖励仅在最后 400 个步骤中加入。我们发现，使用基于模型偏好的奖励信号进行更多训练步骤可能导致奖励黑客行为，相关记录见补充材料 B.5。总训练成本列于补充材料 B.4.4。

4. 实验

我们在 MMLU (Hendrycks et al., 2021)、MMLU-Redux (Gema et al., 2025)、MMLU-Pro (Wang et al., 2024)、C-Eval (Huang et al., 2023)、CMMLU (Li et al., 2024)、IFEval (Zhou et al., 2023b)、FRAMES (Krishna et al., 2024)、GPQA Diamond (Rein et al., 2023)、SimpleQA (OpenAI, 2024a)、C-SimpleQA (He et al., 2024)、SWE-Bench Verified (OpenAI, 2024b)、Aider (Gauthier, 2025)、LiveCodeBench (Jain et al., 2024) (2024-08 – 2025-01)、Codeforces (Mirzayanov, 2025)、中国高中数学奥林匹克 (CNMO 2024) (CMS, 2024) 以及美国数学邀请赛 2024 (AIME 2024) (MAA, 2024) 上评估我们的模型。这些基准的详细信息列于补充材料 D 中。

表 3 总结了图 2 所概述的 DeepSeek-R1 在多个开发阶段的性能。DeepSeek-R1-Zero 与 DeepSeek-R1 Dev1 的比较显示，指令遵循能力得到显著提升，IF-Eval 和 ArenaHard 基准上的得分更高。然而，由于冷启动数据集规模有限，与 DeepSeek-R1-Zero 相比，Dev1 的推理性能出现部分下降，在 AIME 基准上尤为明显。相比之下，DeepSeek-R1 Dev2 展现出

表 3 | DeepSeek-R1 各阶段的实验结果。粗体数字表示性能具有统计显著性（t-检验， $p < 0.01$ ）。

基准 (指标)		R1-Zero	R1-Dev1	R1-Dev2	R1-Dev3	RL
英语	MMLU (EM)	88.8	89.1	91.2	91.0	90.8
	MMLU-Redux (EM)	85.6	90.0	93.0	93.1	92.9
	MMLU-Pro (EM)	68.9	74.1	83.8	83.1	84.0
	DROP (3-shot F1)	89.1	89.8	91.1	88.7	92.2
	IF-Eval (Prompt Strict)	46.6	71.7	72.0	78.1	83.3
	GPQADiamond (Pass@1)	75.8	66.1	70.7	71.2	71.5
	SimpleQA (正确率)	30.3	17.8	28.2	24.9	30.1
	FRAMES (准确率)	82.3	78.5	81.8	81.9	82.5
	AlpacaEval2.0 (LC-winrate)	24.7	50.1	55.8	62.1	87.6
	ArenaHard (GPT-4-1106)	53.6	77.0	73.2	75.6	92.3
Code	LiveCodeBench (Pass@1-COT)	50.0	57.5	63.5	64.6	65.9
	Codeforces (百分位数)	80.4	84.5	90.5	92.1	96.3
	Codeforces (评分)	1444	1534	1687	1746	2029
	SWEVerified (已解决)	43.2	39.6	44.6	45.6	49.2
	Aider-Polyglot (准确率)	12.2	6.7	25.6	44.8	53.3
Math	AIME2024 (Pass@1)	77.9	59.0	74.0	78.1	79.8
	MATH-500 (Pass@1)	95.9	94.2	95.9	95.4	97.3
	CNMO2024 (Pass@1)	88.1	58.0	73.9	77.3	78.8
中文	CLUEWSC (EM)	93.1	92.8	92.6	91.6	92.8
	C-Eval (EM)	92.8	85.7	91.9	86.4	91.8
	C-SimpleQA (正确)	66.4	58.8	64.2	66.9	63.7

在需要高级推理技能的基准上取得了显著性能提升，包括代码生成、数学问题求解和 STEM 相关任务。面向通用任务的基准（如 AlpacaEval 2.0）仅显示出边际改进。这些结果表明，面向推理的 RL 能够显著增强推理能力，但对面向用户偏好的基准影响有限。

DeepSeek-R1 Dev3 将推理数据集和非推理数据集同时纳入 SFT 流程，从而提升模型在推理任务和通用语言生成任务中的能力。与 Dev2 相比，DeepSeek-R1 Dev3 在 AlpacaEval 2.0 和 Aider-Polyglot 上取得了显著的性能提升，这得益于加入了大规模非推理语料库和代码工程数据集。最后，使用推理重点数据和通用数据的混合数据对 DeepSeek-R1 Dev3 进行全面 RL 训练，生成了最终的 DeepSeek-R1。代码和数学基准测试仅有边际改进，因为前期阶段已经完成了大量专门针对推理的 RL。最终 DeepSeek-R1 的主要进步体现在通用指令遵循和用户偏好基准测试上，其中 AlpacaEval 2.0 提升了 25%，ArenaHard 提升了 17%。

此外，我们在补充材料 D.2 中将 DeepSeek-R1 与其他模型进行了比较。模型安全评估见补充材料 D.3。补充材料 E 提供了全面分析，包括与 DeepSeek-V3 的比较、在两套新测试集上的性能评估、按类别对数学能力的拆解，以及对测试时扩展行为的研究。补充材料 F 表明，强大的推理能力可以迁移到更小的模型。

5. 伦理与安全声明

随着 DeepSeek-R1 推理能力的提升，我们深刻认识到其中潜在的伦理风险。例如，R1 可能遭受越狱攻击，从而生成爆炸物制造计划等危险内容；而增强的推理能力使模型能够提供具有更好操作可行性和可执行性的计划。此外，公开模型还容易受到进一步微调的影响，从而破坏其固有的安全防护。

在补充材料 D.3 中，我们从多个角度呈现了全面的安全报告，包括在开源和内部安全评测基准上的表现、多种语言下的安全水平，以及抵御越狱攻击的能力。这些全面的安全分析得出结论：与其他最先进模型相比，DeepSeek-R1 模型的固有安全水平总体处于中等水平（与 GPT-4o（2024-05-13）相当）。此外，与风险控制系统结合后，模型的安全水平提升至卓越标准。

6. 结论、局限性与未来工作

我们提出了 DeepSeek-R1-Zero 和 DeepSeek-R1，它们依靠大规模 RL 来激励模型的推理行为。结果表明，预训练检查点本身就蕴含着进行复杂推理任务的巨大潜力。我们认为，释放这一潜力的关键不在于大规模人工标注，而在于提供高难度推理问题、可靠的验证器以及充足的强化学习计算资源。在强化学习过程中，自我验证和反思等复杂推理行为似乎会自然涌现。

即使 DeepSeek-R1 在推理基准测试上取得了前沿结果，仍然面临一些能力局限，具体如下：

结构化输出和工具使用：目前，与现有模型相比，DeepSeek-R1 的结构化输出能力仍不够理想。此外，DeepSeek-R1 无法利用搜索引擎和计算器等工具来提升输出性能。然而，构建用于结构化输出和工具使用的 RL 环境并不困难，我们相信这一问题将在下一版本中得到解决。

Token 效率：不同于多数投票或蒙特卡洛树搜索（MCTS）等传统的测试时计算扩展方法，DeepSeek-R1 会在推理期间根据当前问题的复杂度动态分配计算资源。具体而言，它使用更少的 token 来解决简单任务，同时为复杂任务生成更多 token。尽管如此，token 效率仍有进一步优化空间，因为在回答较简单问题时仍会出现过度推理（表现为想得过多）的情况。

语言混合：DeepSeek-R1 目前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用问题。例如，即使查询使用的是英语或中文以外的语言，DeepSeek-R1 也可能使用英语进行推理和回答。我们计划在未来的更新中解决这一限制。该限制可能与基础检查点 DeepSeek-V3-Base 有关；它主要使用中文和英文，因此能够在这两种语言的推理中取得更好的结果。

提示工程：在评估 DeepSeek-R1 时，我们观察到它对

提示非常敏感。少样本提示会持续降低其性能。因此，为获得最佳结果，我们建议用户在零样本设置下直接描述问题并指定输出格式。

软件工程任务：由于评估时间较长，会影响 RL 流程的效率，因此大规模 RL 尚未广泛应用于软件工程任务。结果是，DeepSeek-R1 在软件工程基准测试上相较 DeepSeek-V3 并未展现出巨大改进。未来版本将通过在软件工程数据上实施拒绝采样，或在 RL 流程中引入异步评估来提高效率，从而解决这一问题。

除了具体的能力限制之外，纯 RL 方法本身也存在固有挑战：

奖励劫持：纯 RL 的成功依赖于可靠的奖励信号。在本研究中，我们通过推理领域的基于规则的奖励模型（RM）确保奖励可靠性。然而，对于写作等某些任务而言，这类可靠的 RM 很难构建。如果奖励信号由模型而非预定义规则分配，那么随着训练推进，它更容易被利用，这意味着策略模型可能会寻找捷径来劫持奖励模型。因此，对于无法由可靠奖励模型有效评估的复杂任务，扩展纯 RL 方法仍是一项开放挑战。

在本工作中，对于无法获得可靠信号的任务，DeepSeek-R1 使用人工标注创建监督数据，并且仅进行数百步 RL。我们未来能够获得稳健的奖励模型，以解决此类问题。

随着 DeepSeek-R1 等纯 RL 方法的出现，未来有望解决任何能够由验证器有效评估的任务，无论其对人类而言多么复杂。配备此类先进 RL 技术的机器有望在这些领域超越人类能力，因为它们能够通过反复试错迭代优化性能。然而，对于难以构建可靠奖励模型的任务，挑战仍然存在。在此类情况下，缺乏稳健的反馈机制可能阻碍进展，这表明未来研究应专注于开发创新方法，为这些复杂且较难验证的问题定义并改进奖励结构。

此外，在推理过程中利用工具也极具潜力。无论是使用编译器或搜索引擎等工具来检索或计算必要信息，还是使用生物或化学试剂等外部工具在现实世界中验证最终结果，这种工具增强型推理的整合都可能极大拓展机器驱动解决方案的范围并提高其准确性。

7. 作者名单

作者名单按贡献角色组织，每个类别内的个人按名字字母顺序排列。姓名后带星号（*）的作者已不再隶属于我们的团队。

核心贡献者：Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao,

核心作者贡献：Peiyi Wang 和 Daya Guo 共同证明，基于结果的 RL 能够诱导 LLM 中思维链模式的涌现，从而实现

突破性的推理能力。他们共同参与了 R1-Zero 的创建，其工作为 R1 奠定了基础。Daya Guo 还为 MOE 模型的 RL 训练稳定性作出了贡献。Junxiao Song 提出了 GRPO 算法，实现了初始版本，并为数学任务引入了基于规则的奖励。随后，Peiyi Wang 和 Runxin Xu 对 GRPO 算法进行了改进。Zhibin Gou 提出了大 PPO 裁剪策略以增强 GRPO 性能，并与 Zhihong Shao 和 Junxiao Song 一起证明了其重要性。在数据迭代、奖励设计和评估方面，由不同团队负责各领域工作：Qihao Zhu、Z.F. Wu 和 Dejian Yang 专注于代码任务；Zhihong Shao、Zhibin Gou 和 Junxiao Song 专注于数学任务；Peiyi Wang、Ruoyu Zhang、Runxin Xu 和 Yu Wu 负责其他推理及通用任务。此外，Qihao Zhu 和 Zhihong Shao 为 RL 训练的数据选择策略作出了贡献，Zhuoshu Li 和 Yu Wu 共同牵头了整个项目的标注工作。在系统方面，Xiao Bi、Xingkai Yu、Shirong Ma、Xiaokang Zhang、Haowei Zhang 和 Ziyi Gao 实现了 RL 流程，优化系统效率并解决大规模训练中的稳定性问题。最后，Zhibin Gou、Daya Guo 和 Ruoyu Zhang 监督了最终训练阶段并监控模型训练动态。Zhibin Gou 牵头开发了 R1-distill 系列。

贡献者：Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo*, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J.L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu*, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R.J. Chen, R.L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S.S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu*, Wentao Zhang, W.L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X.Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y.K. Li, Y.Q. Wang, Y.X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma*, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, YunfanXiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y.X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z.Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu*, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Zhen Zhang,

附录

A. 背景

A.1. DeepSeek-V3

DeepSeek V3 (DeepSeek-AI, 2024b) 是由 DeepSeek 开发的先进开源 LLM。DeepSeekV3 于 2024 年 12 月发布, 代表了 AI 创新的重大飞跃, 旨在与 OpenAI 的 GPT-4 和 Meta 的 Llama 3.1 等领先模型竞争, 同时保持卓越的成本效率和性能。DeepSeek V3 构建于混合专家 (MoE) 架构之上, 总参数量为 6710 亿, 每个 token 激活 370 亿个参数, 在效率和能力之间实现优化。它在包含 14.8 万亿高质量、多样化 token 的大规模数据集上进行预训练, 随后通过监督微调和强化学习增强其在各领域的的能力。该模型还采用了多头潜在注意力 (MLA) (DeepSeek-AI, 2024a) 以实现高效推理、无辅助损失的负载均衡策略, 以及多 token 预测 (MTP) (Gloeckle et al., 2024) 来提升性能, 尤其是在数学和编程等任务中。

对于 DeepSeek-V3-Base 的训练数据, 我们仅使用普通网页和电子书, 未加入任何合成数据。然而, 我们观察到, 一些网页包含大量由 OpenAI 模型生成的答案, 这可能导致基础模型间接从其他强大模型中获取知识。不过, 在预训练冷却阶段, 我们并未有意加入 OpenAI 生成的合成数据; 该阶段使用的所有数据都是自然产生并通过网络爬取收集的。预训练数据集包含大量数学和代码相关内容, 表明 DeepSeek-V3-Base 已接触到大量推理轨迹数据。这种广泛接触使模型具备生成合理解题候选方案的能力, 强化学习可以从中有效识别并优化高质量输出。我们按照附录 D.1 所述, 在预训练阶段进行了数据污染。DeepSeek-V3 base 的训练数据主要为中文和英文, 这可能是 DeepSeek-R1-Zero 在缺少语言一致性奖励时出现语言混杂的原因。

在本文中, 我们将 DeepSeek-V3-Base 记作基础模型, 将 DeepSeek-V3 记作指令模型。值得注意的是, DeepSeek-R1 和 DeepSeek-R1-Zero 均在 DeepSeek-V3-Base 之上训练, 而 DeepSeek-R1 利用了来自 DeepSeek-V3 SFT 数据的非推理数据。DeepSeek-R1-Dev1、DeepSeek-R1-Dev2 和 DeepSeek-R1-Dev3 是 DeepSeek-R1 的中间检查点。

A.2. 传统的训练后范式

训练后处理已成为改进预训练 LLM、使其满足特定性能目标并符合人类期望的关键步骤。一种广泛采用的两阶段训练后框架是先进行 SFT, 随后进行 RL (Ouyang et al., 2022)。

监督微调通过在针对特定任务精心整理的输入-输出对数据集上训练预训练 LLM, 对其进行优化。该过程采用监督学习目标, 通常最小化模型预测与标注真实值之间的交叉熵损失 (Brown et al., 2020)。例如, 在对话应用中, SFT 可以使用明确提供期望回答的对话数据集, 使模型能够将其输出调整到预定义标准 (Radford et al., 2019)。SFT 提供了多项重要优势。首先, 它利用高质量示例实现精确的任务对齐, 使模型能够

在客户支持或技术文档等领域表现出色 (Radford et al., 2019)。其次, 依赖预训练权重确保了计算效率, 所需资源少于从头训练。最后, 使用明确的输入-输出映射增强了可解释性, 因为模型的学习过程直接关联到可观察数据, 从而最大限度降低异常行为风险 (Ouyang et al., 2022)。尽管 SFT 具有优势, 其性能取决于训练数据集的质量和多样性; 狭窄或有偏的数据可能削弱模型泛化到新情境的能力 (Brown et al., 2020)。此外, SFT 的静态性质——针对固定输出进行优化——可能无法捕捉不断变化的人类偏好或细致目标。整理高质量数据集的劳动密集型过程进一步限制了其可扩展性, 因为数据中的错误或不一致可能传播到模型行为中 (Ouyang et al., 2022)。

在 SFT 之后, 强化学习通过根据奖励信号优化 LLM 的输出, 进一步改进模型。在这一阶段, 模型与环境交互——环境通常是一个基于人类反馈训练的奖励模型——并调整自身行为以最大化累积奖励。这种方法的一个代表性实例是从人类反馈中进行强化学习 (RLHF), 其中奖励函数编码了人类偏好 (Christiano et al., 2017)。因此, RL 将重点从静态监督转向动态优化。值得注意的是, RL 降低了对大量标注资源的需求; SFT 要求为每个输入-输出对提供完整标注的数据集, 而 RL 可以使用少量人类评估或训练好的奖励模型运行, 甚至可以使用基于规则的奖励模型, 从而显著降低标注负担。

SFT 和 RL 的顺序应用结合了二者互补的优势。SFT 通过让模型扎根于精心整理的示例, 建立稳健且特定于任务的基线; 而 RL 则进一步完善这一基础, 使其与更广泛、以人为中心的目标保持一致 (Ouyang et al., 2022)。例如, SFT 可以确保对话系统的语法准确性, 而 RL 则优化互动性和简洁性, 正如 InstructGPT 的开发所示 (Ouyang et al., 2022)。事实证明, 这种混合方法能够有效生成既精确又具适应性的模型。

在本研究中, 我们证明 SFT 阶段可能会阻碍模型探索和发展有效推理策略的能力。这一局限源于: 在人类提供的回答作为 SFT 目标时, 它们并不总是最适合模型学习; 这些回答往往省略了明确反思和验证步骤等关键推理组成部分。为解决这一问题, DeepSeek-R1-Zero 使模型能够独立于人类先验, 直接探索推理模式。通过这种自我探索发现的推理轨迹随后被蒸馏, 并用于训练其他模型, 从而促进其获得更稳健、更具泛化性的推理能力。

A.3. GRPO 与 PPO 的比较

群组相对策略优化 (GRPO) (Shao 等, 2024) 是我们用于训练 DeepSeek-R1-Zero 和 DeepSeek-R1 的强化学习算法。该算法最初旨在简化训练过程并降低近端策略优化 (PPO) (Schulman 等, 2017) 的资源消耗; PPO 广泛用于 LLM 的 RL 阶段 (Ouyang 等, 2022)。有关 GRPO 与 PPO 的总体比较, 请参见图 3。

对于每个问题 q , GRPO 从旧策略中采样一组输出 $\{o_1, o_2, \dots, o_G\}$

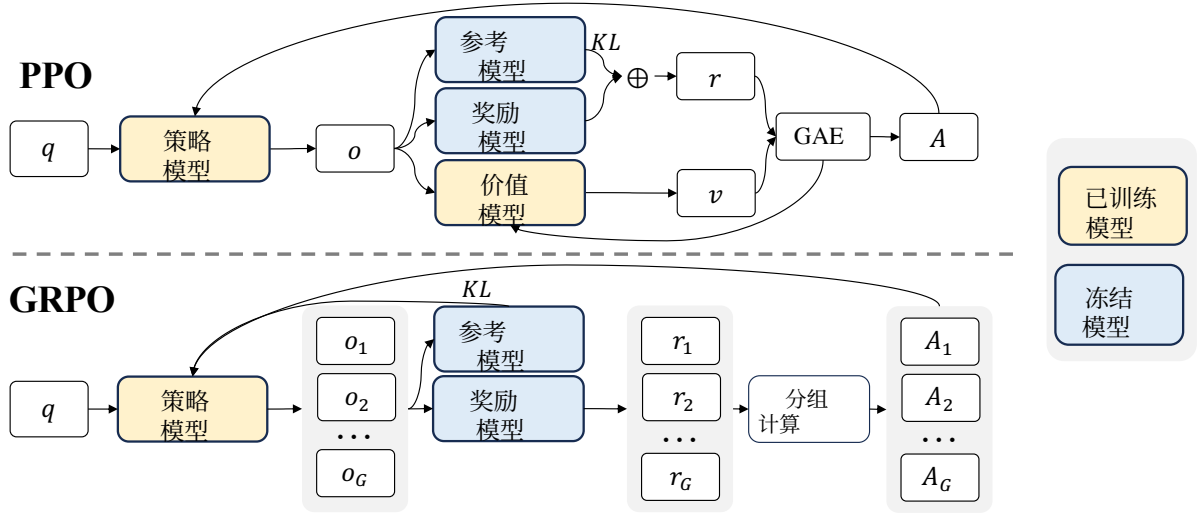


图 3 | PPO 与我们的 GRPO 示例。GRPO 放弃价值模型，转而根据组得分估计优势。

$\pi_{\theta_{old}}$ 然后通过最大化以下目标来优化策略模型 π_{θ} ：

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (11)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (12)$$

其中， π_{ref} 是参考策略， ϵ 和 β 是超参数， A_i 是优势，它通过一组奖励 $\{r_1, r_2, \dots, r_G\}$ 计算得到，这些奖励对应于每个组内的输出：

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (13)$$

相比之下，在 PPO 中，优势通常通过广义优势估计（GAE）（Schulman et al., 2015）计算，该方法不仅基于奖励，还基于学习得到的价值模型。由于价值模型通常与策略模型规模相近，因此会带来显著的内存和计算开销。此外，价值模型的训练目标是根据从序列开头到当前位置生成的 token，预测从当前位置开始的预期累积奖励。这本身就很困难，尤其是在只有最终结果奖励可用的情况下。当训练长思维链推理模型时，这一挑战会更加明显。随着输出长度增加，模型在生成过程中更可能进行反思和修订等行为，这意味着最初生成的内容之后可能被修改或否定，从而使得根据部分响应预测最终奖励更加不可行。

GRPO 与 PPO 的另一个关键区别在于，训练策略与参考策略之间的 Kullback–Leibler (KL) 散度如何纳入训练过程。在 GRPO 中，KL 散度的无偏估计量（Schulman, 2020）按照公式 11 直接加入损失；而在 PPO 中，逐 token 的 KL 惩罚作为每个 token 处的稠密奖励加入（Ouyang et al., 2022）。由于强化学习的优化目标是

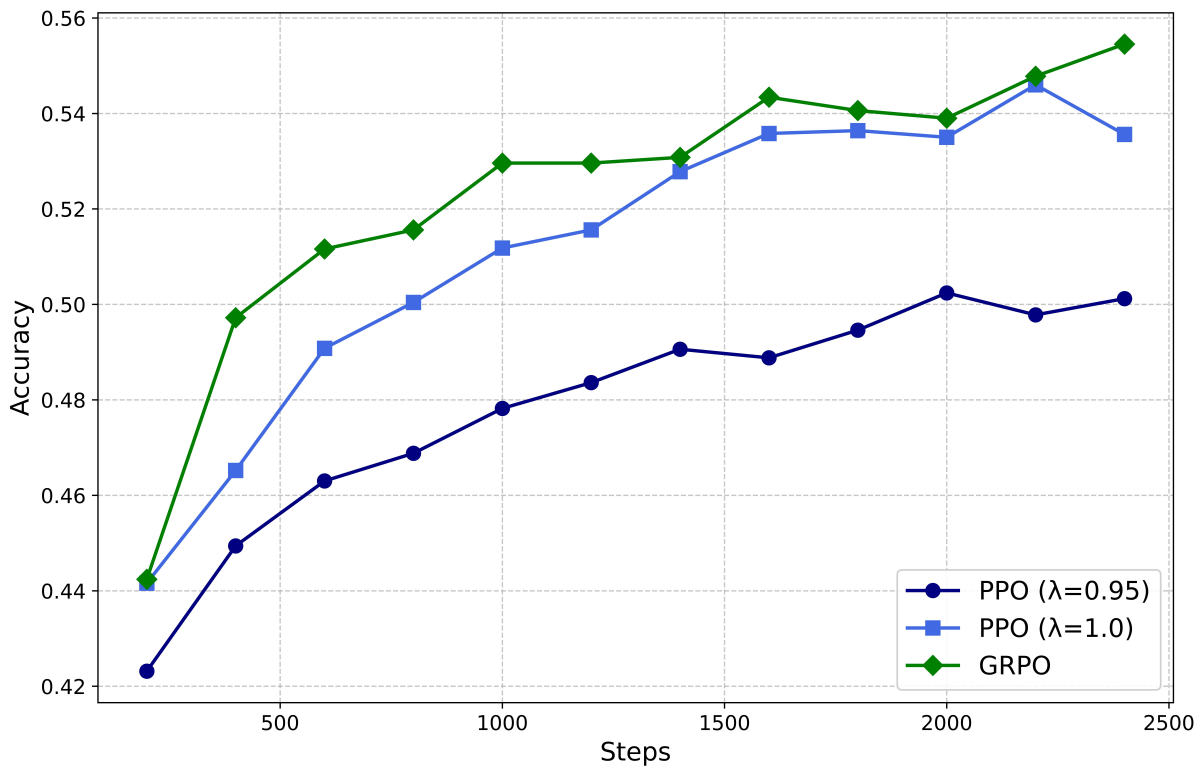


图 4 | PPO 和 GRPO 在 MATH 任务上的性能。

最大化累积奖励，PPO 的方法会惩罚累积 KL 散度，这可能隐式惩罚回答长度，从而阻止模型回答长度增加。此外，在训练长思维链推理模型的场景中，我们可能进行数千步训练，训练策略可能显著偏离初始参考策略。为了平衡训练策略可探索的范围与训练稳定性，我们在实际训练过程中定期将参考策略更新为最新策略。

图 4 使用 DeepSeek-Coder-V2-Lite (16B MoE, 包含 2.4B 个激活参数) 比较了 PPO 和 GRPO 在 MATH 任务上的性能。与 GRPO 不同，PPO 需要额外的超参数调优，尤其是 GAE 中的 λ 系数，并且对该参数高度敏感。当 λ 设置为 0.95 (大多数开源 PPO 实现中的默认值) 时，PPO 的表现明显差于 GRPO。然而，通过仔细调优 (将 λ 设置为 1.0)，PPO 的性能显著提升，接近 GRPO。

在适当调优时，PPO 可以达到相当的性能，但它需要额外的计算成本来进行超参数优化。此外，考虑到训练额外价值模型所需的内存和计算开销，GRPO 提供了一种更实用的替代方案，尤其适用于资源受限的大规模模型训练。

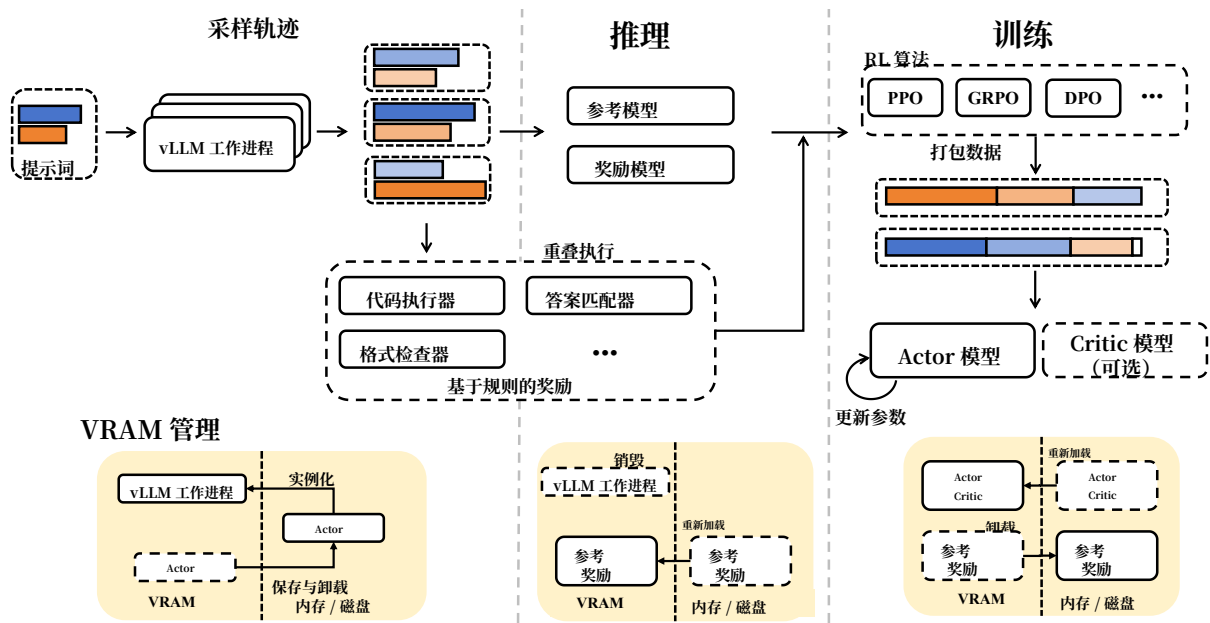


图 5 | 我们的 RL 框架概览。

B. 训练细节

B.1. RL 基础设施

在大型模型上进行 RL 训练对基础设施提出了很高要求。我们的 RL 框架采用解耦且可扩展的结构，以便无缝集成多种模型和算法。在该框架内，我们结合了模块内和模块间优化技术，以确保训练效率和可扩展性。

具体而言，如图 5 所示，该框架被划分为四个不同的模块，每个模块对应 RL 流程中的一个特定阶段：

- **Rollout 模块：**提示从训练数据集中加载，并均匀分发至多个 vLLM (Kwon et al., 2023) 工作进程；每个工作进程均配备 actor 模型，用于采样多个回答。对于 DeepSeek-V3 MoE 架构，我们在节点间实施专家并行策略，以降低内存访问开销，并部署热点专家的冗余副本，以平衡不同专家之间的计算负载。此外，还利用多 Token 预测（MTP）组件进行自推测解码，从而显著加快解码速度，并有效缩短最长样本的完成时间。
- **推理模块：**该模块加载奖励模型和参考模型，对采样轨迹阶段生成的样本执行前向传播，从而获得基于模型的奖励及其他必要信息。
- **基于规则的奖励模块：**该模块计算模型生成回答的基于规则的奖励。我们设计了统一接口，以适应多种实现（例如代码执行器、答案匹配器、格式检查器等）。尽管该模块无需将模型加载到 GPU 内存中，但其执行往往耗时较长。为解决这一问题，我们采用异步调度方法，使其执行与 Rollout 和 Inference 模块重叠，从而有效隐藏

相关延迟 y .

- **训练模块：**该模块加载 actor 模型和 critic 模型（如有需要），用于计算损失并更新模型参数。它灵活支持多种 RL 算法（例如 PPO、GRPO、DPO 等）。为最大限度减少序列填充造成的计算浪费并平衡设备间的工作负载，我们设计了以下数据打包策略：首先，将全局批次中的所有数据按长度排序，并分发到数据并行组内的各个进程；随后，在每个进程内应用 Best-Fit 策略，将数据打包成固定长度的分块，并尽量减少填充；最后，调整分块数量，使所有进程保持一致。此外，我们集成了 DeepSeek-V3 训练中使用的 DualPipe 算法，以实现高效的流水线并行。

值得注意的是，在每个模块完成后（基于规则的奖励模块除外），该阶段使用的模型实例会自动从 VRAM 卸载到系统内存或磁盘存储，从而为后续阶段释放 VRAM。

B.2. 奖励模型提示词

请充当公正的评审，评估下方用户提示所提供的两位 AI 助手回答的质量。你将获得助手 A 和助手 B 的答案。你的任务是评估哪位助手的答案更好。开始评估时，先生成你自己对该提示的答案。你必须先给出自己的答案，然后才能评判任何答案。评估助手答案时，将两位助手的答案都与你的答案进行比较。你必须找出并纠正任何错误或不准确信息。然后考虑助手的答案是否有帮助、相关且简洁。有帮助意味着答案正确回应提示或遵循指令。请注意，当用户提示存在歧义或有不止一种解释时，相比基于假设直接回答，向用户询问澄清或更多信息更有帮助且更合适。相关意味着回答的所有部分都与所问内容紧密相连或适合所问内容。简洁意味着回答清晰，不冗长或过度。必要时，还要考虑助手答案的创造性和新颖性。最后，指出助手答案中缺失的、在回应用户提示时有益的重要信息。在给出解释后，你必须仅输出以下选项之一作为带标签的最终判定：1. 助手 A 显著更好：[[A>>B]]2. 助手 A 略好：[[A>B]]3. 平局，基本相同：[[A=B]]4. 助手 B 略好：[[B>A]]5. 助手 B 显著更好：[[B>>A]]示例输出：`我的最终判定是平局：[[A=B]]`。

表 4 | RL 数据和任务描述。

数据类型 #提示词		问题类型	输出类型
Math	26K	定量推理	数字/表达式/方程
Code	17K	算法与错误修复	代码解决方案
STEM	22K	多项选择	选项
逻辑	15K	选择题/定量推理	选项/数字
通用	66K	有用性/无害性	排序后的回答

B.3. 数据配方

B.3.1. RL 数据

推理 RL 数据包括四类：数学、编码、STEM 和逻辑问题。此外，我们还加入通用 RL 数据，以在 DeepSeek-R1 的训练中提升模型的有帮助性和无害性。所有问题均为中文或英文。RL 数据的描述见表 4，下面我们将逐一介绍每种数据类型的细节：

- **数学** 数据集由 26k 道定量推理问题组成，包括数学考试题和竞赛题。提示词的平均 token 数为 122。该数据集涵盖代数、微积分、概率和几何等多个数学领域。问题难度从地区性竞赛到国际奥林匹克竞赛不等。对于每道题，模型都应生成一个逐步推理过程，并最终给出答案；答案可以是数值（例如“5”）、数学表达式（例如“ $x^2 + 3x - 2$ ”）或方程（例如“ $y = 2x + 1$ ”）。由于难以判断数学证明的正确性，因此不包含数学证明。为进行强化学习，我们通过将预测答案与参考答案匹配来计算推理过程的奖励。如果答案与参考答案一致，则奖励为 1；否则为 0。

- **编程** 数据集包括 17k 道算法竞赛题，以及 8k 道 bug 修复问题。算法竞赛题类似于 Codeforces 或 LeetCode 等平台上的题目。每道题通常包括详细的问题描述、约束条件以及多个输入输出示例。任务是编写能够正确且高效解决问题的完整函数或程序，并通过一组全面的隐藏测试用例，这些用例同时评估正确性和性能。这些问题考查算法技能，包括动态规划、图论、字符串操作和数据结构的使用。

这些 bug 修复问题取自真实世界的 GitHub issue。每项任务都提供问题描述、存在 bug 的源代码版本，以及一组部分或完全失败的单元测试。目标是理解 issue 的意图，定位并修复代码中的缺陷，并确保修正后的版本通过所有单元测试。

- **STEM** 数据集包含 22k 道选择题，涵盖物理学、化学和生物学等主题。STEM 任务中的每道题都提出一个特定学科问题，并附带四至八个答案选项。模型需要根据给定上下文和领域知识选择科学上最准确的答案。提示 token 平均数量为 161。具体而言，该数据集包括 15.5% 的物理学题目、30.7% 的生物学题目、46.5% 的化学题目，以及 7.3% 的其他主题题目，如健康和医学。由于所有 STEM 问题都是选择题，因此根据

是否匹配正确选项来分配二元奖励。

- **逻辑** 数据集包含 15k 个问题，旨在评估模型在广泛逻辑挑战中的推理能力。该数据集既包含现实世界问题，也包含合成生成的问题。所有问题都支持自动评估，平均提示长度约为 420 个 token。数据集的现实世界部分包含从网络获取的多样化问题，包括脑筋急转弯、经典逻辑谜题和知识密集型问题。这些问题采用多项选择格式，以确保客观且一致的评估。合成部分主要由两类问题组成：代码-IO 问题和谜题任务。代码-IO 问题使用 Li 等（2025）提出的数据处理流程生成，该流程将竞赛编程问题及其对应的输入输出测试用例转换为可验证的逻辑推理问题。谜题任务包括旨在评估特定推理能力的问题。例如，密码学谜题用于评估模型识别和应用密码方案模式或执行字符串操作的能力；逻辑谜题侧重于在复杂约束上进行演绎推理，例如从一组固定前提中推断有效结论（如斑马谜题）；算术谜题则测试模型的数值推理能力（如概率问题和 24 点游戏）。

- **通用** 数据集包含 66k 个用于评估有帮助性的问题，涵盖创意写作、编辑、事实问答和角色扮演等各种类别。此外，该数据集还包括 12,000 个专门用于评估无害性的问题。为确保稳健验证，我们使用两个奖励模型，每个模型分别基于精心整理的数据集进行训练；这些数据集包含模型针对有帮助性和无害性生成的排序响应。训练阶段，我们使用最大序列长度 8192 个 token，对有帮助性奖励模型训练一个 epoch。然而，在部署模型以生成奖励信号时，我们没有对待评估输入序列施加任何显式长度限制。

B.3.2. DeepSeek-R1 冷启动

对于 DeepSeek-R1，我们构建并收集少量长 CoT 数据，以微调模型，使其成为初始 RL actor。其动机主要源于产品需求，重点提升用户体验。当推理过程符合第一人称视角的思维模式时，用户往往觉得回答更直观、更具吸引力。例如，DeepSeek-R1-Zero 在解决问题时更可能使用代词“we”，或者完全避免使用第一人称代词，而 DeepSeek-R1 则更常使用“I”。此外，我们承认，这类模式可能会引发用户不应有的信任。在此，我们想强调，所观察到的生动推理模式主要反映了 DeepSeek 设计的启发式方法，并不意味着模型天生获得了类似人类的智能或自主解决问题的能力。

在冷启动数据创建中，我们偏好先理解问题、再进行包含反思和验证的详细推理这一思考过程。整个思考过程中使用第一人称视角的语言。此外，保持语言一致性对于实现最佳用户体验至关重要。如果缺乏适当控制，无论查询使用何种语言，模型回答都可能混杂不同语言。这种不一致会扰乱理解并降低用户满意度。因此，必须进行仔细完善，以确保回答保持连贯并符合用户意图。然而，我们承认，DeepSeek-R1-Zero 产生的原始思维链（CoT）推理可能拥有超越

当前人类先验的局限。具体而言，我们首先让人工标注员将推理轨迹转换为更自然、更符合人类对话的风格。随后，使用修改后的数据对作为示例，提示 LLM 以类似风格改写更多数据。之后，所有由 LLM 生成的输出都要经过第二轮人工验证，以确保质量和一致性。

列表 1 | 生成人类可读解答的提示。

```
## Question
{question}

## Thought process
{thought_process}

---
Based on the above thought process, provide a clear, easy-to-follow, and well-formatted
solution to the question. Use the same language as the question.

The solution must strictly follow these requirements:
- Stay faithful and consistent with the given thought process. Do not add new reasoning
  steps or conclusions not shown in the original.
- Show key steps leading to final answer(s) in clear, well-formatted LaTeX.
- Use \boxed{} for final answer(s).
- Be clean and concise. Avoid colloquial language. Do not use phrases like "thought
  process" in the solution.

Your response should start with the solution right away, and do not include anything
else. Your task is solely to write the solution based on the provided thought
process. Do not try to solve the question yourself.
```

具体而言，我们首先收集数千条高质量、多样化的推理提示。对于每条提示，我们使用温度相对较高的 1.0，通过 DeepSeek-R1-Zero 生成多条推理轨迹。接下来，我们筛选这些生成结果，仅保留最终答案正确且格式易读的结果。对于数学输出，我们使用 `sympy`(<https://www.sympy.org/>) 进行解析和表达式比较；在格式方面，则应用重复检测和语言混用过滤等规则。最后，我们提示 DeepSeek-V3 同时完善推理和总结，以确保格式规范且表达对人类友好。特别是，为解决语言混用问题，我们要求 DeepSeek-V3 “将思考过程翻译成与问题相同的语言”。由于 DeepSeek-R1-Zero 的总结只提供最终答案，我们使用列表 1 中的总结提示生成简洁、便于人类阅读的解答，概述推理步骤和最终结果。

对于代码数据，我们收集了大量竞赛编程题目。具体而言，我们从多个在线评测（OJ）平台汇编了广泛的竞赛编程题目，包括 Codeforces 的 5151 道题目和 AtCoder 的 2504 道题目。由于这些平台未公开提供原始测试用例，我们开发了一种方法，为每道题目创建可靠的测试用例。

我们的方法是使用 DeepSeek-V2.5 生成候选测试用例，然后进行严格的验证过程。具体而言，我们提示 DeepSeek-V2.5 编写 Python 程序，为每道题目的要求生成定制化测试用例，如代码清单 2 所示。

获得大量候选测试用例后，我们实施了两阶段筛选程序。首先，我们使用正确提交来剔除会产生错误输出的无效测试用例。然后，我们策略性地选择能够成功识别

错误提交中缺陷的测试用例例子集。该过程确保最终测试用例能够正确区分每个问题的正确解和错误解。

列出 2 | 用于代码问题生成测试用例的提示词。

<USER>Question:You are given a string word and an integer k.

```

    return f"{{{seq}}}\n{{{k}}}"

t = 10 # the number of test cases
testcase = [generate_random_string(10000, 10000) for _ in range(t)]

print(t)
print("\n".join(testcase))
'''

First construct a random input generator for large data with 10 testcases of different
length.

'''pythonimport randomimport stringdef generate_random_string(l_min, l_max):

```

Construct a random input generator for large data with only one test case.

```

'''pythonimport randomimport string

```

Construct an input generator for large adversarial inputs of a string of a repeated single character.

```

'''python
import random
import string

def generate_single_repeated_char(l_min, l_max):
    length = random.randint(l_min, l_max)

```

```
char = random.choice(string.ascii_lowercase)seq = char * length
```

此外，我们采用少样本提示，让 DeepSeek-V3 生成对简单数学问题的回答，例如 “ $1 + 1 = ?$ ”，确保输出保持简洁且结构恰当。我们在列表 3 中提供了一个简单数学问题的提示。

列表 3 | 用于为简单数学问题生成思维链的提示词。

```
## Question  
How much is 5+4?
```

Response

<think>

I need to add the numbers 5 and 4. Starting with 5, if I add 4 to it, the total will be

</think>

****Solution:****

We are asked to calculate the sum of 5 and 4.

****Step 1:**** Start with the number 5.

****Step 2:**** Add 4 to it.

\[
5 + 4 = 9
\]

****Final Answer:**** $\boxed{9}$

Question

what is 1 plus 2

Response

<think>

I need to determine the sum of 1 and 2.

Adding these two numbers together, 1 plus 2 equals 3.

Therefore, the answer is 3.

</think>

Sure! Let's solve the problem step by step.

****Problem:**** What is $(1 + 2)$?

****Solution:****

To find the sum of 1 and 2, simply add the two numbers together:

\[
1 + 2 = 3
\]

****Answer:****

$\boxed{3}$

Question

{question}

Response

列表 4 | 使用 DeepSeek-V3 作为评判器的示例提示词。

As an advanced reasoning problem evaluation assistant, your primary responsibility is

B.3.3. 800K 监督数据

推理数据 我们整理了大量推理提示词，并从第一阶段 RL 训练的检查点进行拒绝采样，以生成推理轨迹。在前一阶段，我们仅纳入能够使用基于规则的奖励进行评估的数据。然而，在本阶段，我们通过加入额外数据来扩展数据集，其中一部分数据使用生成式奖励模型：将真实答案和模型预测输入 DeepSeek-V3 进行判断，示例提示词见列表 4。此外，由于模型输出有时杂乱且难以阅读，我们过滤掉了包含混合语言、长段落和代码块的思维链。对于每个提示词，我们采样多个响应，仅保留正确的响应。总计而言，我们收集了约 600k 个与推理相关的训练样本。

非推理数据 对于写作、事实问答、自我认知和翻译等非推理数据，我们采用 DeepSeek-V3 流程，并重复使用 DeepSeek-V3 的部分 SFT 数据集。我们还加入面向软件工程的数据，包括程序修复和前端 Web 开发，以增强模型解决现实问题的能力。对于某些非推理任务，我们通过提示 DeepSeek-V3 在回答问题前生成潜在的思维链。然而，对于“hello”等更简单的查询，我们不会在回答中提供 CoT。最终，我们共收集了约 200k 个与推理无关的训练样本。

在设计思维过程风格时，我们要求模型遵循几个关键原则：第一，保持每个段落简洁易懂。短段落能让想法更清晰、更易于跟随。第二，采用自然且引人入胜的对话式语气。我们避免使用 markdown 等技术格式，以保持流畅的阅读体验。第三，也是最重要的一点，思维过程始于理解完整的用户背景。这意味着要分析我们的用户是谁、他们正面临什么情境，以及他们真正需要什么——包括可能隐藏在初始请求表面之下的未明示需求。

在从模型中 eliciting 出这些思维过程后，人工标注员会仔细核验输出的准确性。我们的研究表明，这些人工推理轨迹增强了模型解读用户查询的精确性。具体而言，它们能够有效突出格式约束、澄清用户意图，并阐明输出所需的结构。这种方法有助于实现模型与用户之间更准确、更具响应性的交互。

表 5 | SFT 数据的数据统计。

领域	样本数	平均轮数	平均词元数
Math	395285	1.0	6094.2
Code	211129	1.1	7435.7
STEM	10124	1.0	4928.8
逻辑	10395	1.0	2739.0
通用	177812	1.1	1419.8
总计	804745	1.0	5355.3

数据统计表 5 总结了各个领域的数据统计，数据量约为 800,000 条监督样本。值得注意的是，大多数数据由单轮交互组成，这可能限制 DeepSeek-R1 的多轮对话能力。我们将多轮对话数据的扩展留作未来工作。数学相关数据主要为中文和英文，涵盖广泛的主题和难度等级。这些问题是可验证的，可以通过确定性规则或参考特定的真实答案进行验证。代码数据集不仅包含竞赛编程问题，还包括调试任务和面向项目的编码查询。与数学和代码相比，STEM 及逻辑相关问题的数量较少，来源于公开教材和在线资源库。通用领域查询内容多样，主要为中文和英文，包括创意写作、开放式问答、角色扮演场景等任务。

SFT 采样轨迹示例我们在列表 5、6 和 7 中提供了与数学、代码和非推理相关的 SFT 采样轨迹示例。

列举 5 | 一个来自数学相关推理数据的 SFT 采样轨迹示例。

```
## QuestionFor positive integer $n$, define $$S_n$ to be the minimum value of the sum \[ \sum_{k=1}^n k^2 $$. Find $S_{100}$.

## Response
```

<think>Alright.

hypotenuse of a right-angled triangle with legs $\sqrt{(2k - 1)}$ and $\sqrt{a_k}$. So,

Alternatively,

$\sqrt{17}$. I can use the method of Lagrange multipliers

taking the derivative of $\sqrt{x^2 + y^2}$ with respect to each $\sqrt{a_n}$.

Simplify: $\left(1 - \lambda^2 = \lambda^2 \cdot \frac{(2k-1)^2}{a_k^2} \right)$. Then,

going back to the original expression for (S_n) :

3. Plugged them up to get $\left(\sqrt{n^4 + 17^2} \right)$.

$\left(\sqrt{n^4 + 17^2} \right)^2 = \sqrt{n^4 + 17^2}$. Yeah, that seems to check out. So,

I need to find an integer (n) such that $(n^4 + 289)$ is a perfect square.

This implies $m^2 - n^4 = 289$. Hmm, 289 is 17^2 . So,

Adding the two equations: $(m - n^2) + (m + n^2) = 1 + 289 \Rightarrow 2m = 290 \Rightarrow m = 145$.

Subtracting the two equations: $(m + n^2) - (m - n^2) = 289 - 1 \Rightarrow 2n^2 = 288 \Rightarrow n^2 = 144 \Rightarrow n = 12$ since n is positive.

Therefore, $n = 12$. Before I conclude,

Is this the only n ? The question says there's a unique positive integer n ,

Final Answer The unique positive integer n is $\boxed{12}$.

Substituting this back into the expression for (S_n) , we get:

we find that the unique positive integer n is $\boxed{12}$.

Thus, the unique positive integer (n) is $\boxed{12}$.

列表 6 | 一个与代码相关推理数据中的 SFT 采样轨迹示例。

QuestionIn this kata,

words with their definitions and looking them up. Hmm, let's think about the setup. So,

the value is the definition. So in the `__init__` method,

`{}`. Then,

So in `newentry`, I need to add the word as the key and the definition as the value

Now, the `look` method is called with a key. It needs to return the definition if the key exists,

Wait, how does it handle case sensitivity? I guess we're assuming the keys are case-

So the steps for the `look` function are: check if the key is in `self.entries`. If yes,

Testing with the example:

```
d = Dictionary()
d.newentry('Apple', 'A fruit...') --> adds to entries.
d.look('Apple') returns the string, so print shows it correctly.
d.look('Banana') returns the can't find message.
```

That should work.

I think that's the basic plan. So the setup code is given, I have to fill in the methods. So in code, it's as I wrote.

Wait, any edge cases? Like, what if word is an empty string? Or any other type? The

Another thing: in the look method, when we do 'if key in self.entries', does it check

So, to implement:

Start by adding the `__init__`, `newentry`, and `look` methods as described.</think>

To solve this problem, we need to create a Python class called `Dictionary` that allows

Approach

1. **Data Structure**: Use a dictionary to store the words as keys and their definitions as values. This allows efficient lookups and additions.
2. **Initialization**: In the `__init__` method, initialize an empty dictionary to
3. **Adding Entries**: The `newentry` method takes a word and its definition, then adds
4. **Looking Up Entries**: The `look` method checks if the word exists in the dictionary. If it does,

Solution Code

```
'''python
class Dictionary():
    def __init__(self):
        self.entries = {}

    def newentry(self, word, definition):
        self.entries[word] = definition

    def look(self, key):
        if key in self.entries:
            return self.entries[key]
        else:
            return f"Can't find entry for {key}"
'''
```

Explanation

1. **Initialization**: The `__init__` method sets up an empty dictionary `self.entries` to store the words and their definitions.
2. **Adding Entries**: The `newentry` method takes two parameters, `word` and `definition`,
3. **Looking Up Entries**: The `look` method checks if the given key exists in `self.`

This approach ensures efficient storage and retrieval of word definitions using a

列表 7 | 与写作相关的非推理数据中的示例 SFT 轨迹

Question

write a short poem ending in "was tentacles all along"

```

## Response
<think>
Okay, I need to help the user write a short poem that ends with "was tentacles all
    along." Hmm, ending with that specific line suggests a twist or surprise, maybe
    something unexpected.

First, I'll think about the theme. Tentacles make me think of sea creatures, like
    octopuses or squids, so perhaps an underwater setting would work well. Maybe a
    treasure hunt or an adventure beneath the waves could set the scene.

I want the poem to build up anticipation. Starting with something mysterious, like
    shadows under the sea, could create intrigue. Then, introducing elements like a
    sunken chest or sailors' lore might add depth and keep the reader engaged.

Rhyme scheme is important for a poem. I'll probably go with something simple, like AABB
    or ABAB, to keep it flowing smoothly. Each stanza should have four lines to
    maintain structure.

Now, the twist at the end should be sudden but make sense. Perhaps the treasure they
    seek turns out to be something entirely different. The last line needs to be
    impactful, so building up the suspense through the previous lines will make the
    ending more surprising.

Let me outline the stanzas. The first stanza could set the scene with the ocean depths
    and a hidden secret. The second stanza can show the action, like diving for
    treasure, and then reveal the unexpected tentacles.

I'll make sure each line flows naturally and the imagery is vivid. Using words like "
    waves," "shadows," "caverns," and "mystery" can help paint the picture. Ending with
    the twist will leave the reader with a lasting impression.

Alright, putting it all together, I'll craft two stanzas that build up to the final
    line, ensuring the surprise feels satisfying and the poem feels complete.
</think>

Beneath the waves, where shadows creep,
A secret stirs in the ocean deep.
A sunken chest, a sailor's song,
The mystery whispered all along.

We dove for pearls, for gold, for lore,
Through caverns dark, to the ocean floor.
But what we found, so strange, so strong--
It wasn't treasure... it was tentacles all along.

```

B.4. 超参数

B.4.1. DeepSeek-R1-Zero-Qwen-32B 的超参数

为训练 DeepSeek-R1-Zero-Qwen-32B, 我们将学习率设为 $2e-6$, KL 系数设为 0.001, 并将采样温度设为 1 用于 rollout。对于每个问题, 我们采样 16 个输出, 最大长度为 32,768。每个训练步骤包含 32 个独特问题, 因此每步训练批量大小为 512。每 400 步, 我们用最新的策略模型替换参考模型。为加速训练, 每次 rollout 生成 8,192 个输出, 随机分成 16 个 mini-batch, 并且只训练一个内部 epoch。

B.4.2. SFT 的超参数

对于代码起始 SFT 和第二阶段 SFT，我们使用 B.3 中所述的精选数据集，对 DeepSeek-V3-Base 进行 2-3 个 epoch 的微调。我们采用余弦衰减学习率调度器，学习率从 5×10^{-5} 开始并逐渐降至 5×10^{-6} 。最大上下文长度设置为 32,768 个 token，批量大小为 128。

B.4.3. 蒸馏的超参数

对于蒸馏，我们使用第 B.3.3 节所述的 800k 数据，对相应基础模型进行 2-3 个 epoch 的微调。基础模型和初始学习率列于表 6 中。我们采用余弦衰减学习率调度器，将学习率逐渐降至初始值的十分之一。最大上下文长度为 32,768 个 token，批量大小为 64。

表 6 | DeepSeek-R1 蒸馏模型、其对应的基础模型及初始学习率。

蒸馏模型	基础模型	初始学习率
DeepSeek-R1-Distill-Qwen-1.5B	Qwen2.5-Math-1.5B	1×10^{-4}
DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B	8×10^{-5}
DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B	7×10^{-5}
DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-32B	6×10^{-5}
DeepSeek-R1-Distill-Llama-8B	Llama-3.1-8B	5×10^{-5}
DeepSeek-R1-Distill-Llama-70B	Llama-3.3-70B-Instruct	2×10^{-5}

B.4.4. 训练成本

关于 DeepSeek-R1 的研究，我们使用 A100 GPU 为较小模型（30B 参数）的实验做准备。该较小模型的结果很有前景，使我们能够有信心扩展到 660B R1-Zero 和 R1。训练 DeepSeek-R1-Zero 时，我们使用了 64*8 个 H800 GPU，过程约需 198 小时。此外，在 DeepSeek-R1 的训练阶段，我们使用了同样的 64*8 个 H800 GPU，用时约 4 天，即约 80 小时。为创建 SFT 数据集，我们使用了 5K GPU 小时。详细信息见表 7。

B.5. 奖励劫持

在 LLM 训练语境中，奖励劫持是指模型利用奖励函数中的缺陷或偏差，从而获得高奖励分数，却并未真正与潜在的人类意图保持一致的现象。在我们的工作中，使用 helpful reward model 时观察到了这种奖励劫持行为。具体而言，如果奖励模型包含系统性偏差或不准确之处，LLM 可能会学会生成得到该模型高评价、但偏离真实人类偏好的回答。这种不一致可能表现为在需要复杂推理的任务上性能下降，如图 6 所示。

B.6. 语言一致性奖励的消融研究

为研究语言一致性（LC）奖励的影响，我们在 DeepSeek-R1-Distill-Qwen-7B 上进行了消融实验。该模型使用与 DeepSeek-R1 相同的冷启动数据

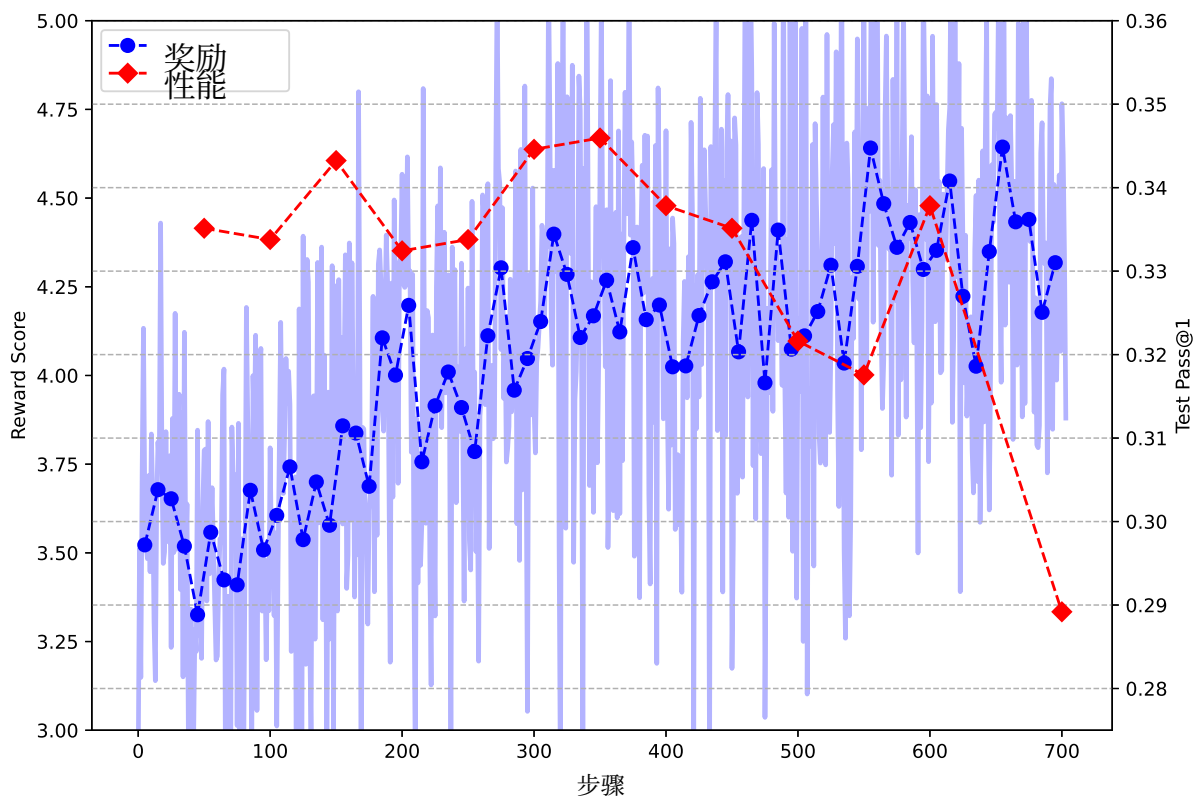


图 6 | 奖励劫持：在训练过程中，随着 CodeForces 性能下降，奖励呈现上升趋势。

表 7 | DeepSeek-R1 的训练成本，假设 H800 的租赁价格为每 GPU 小时 2 美元。

训练成本	DeepSeek-R1-Zero	SFT 数据创建	DeepSeek-R1	总计
以 H800 GPU 小时计	101K	5K	41K	147K
以美元计	\$202K	\$10K	\$82K	\$294K

并且在 RL 过程中也表现出语言混用。结果如图 7 所示。可以看出，在没有 LC 奖励的情况下，随着训练步数增加，语言一致性逐渐恶化。然而，应用 LC 奖励后，整个训练过程中都能保持稳定的语言一致性。在基准性能方面，该模型在数学基准测试上保持了可比性能，而在编程基准测试上观察到轻微下降。尽管这种对齐会导致模型性能略有下降，但该奖励符合人类偏好，使输出更具可读性。

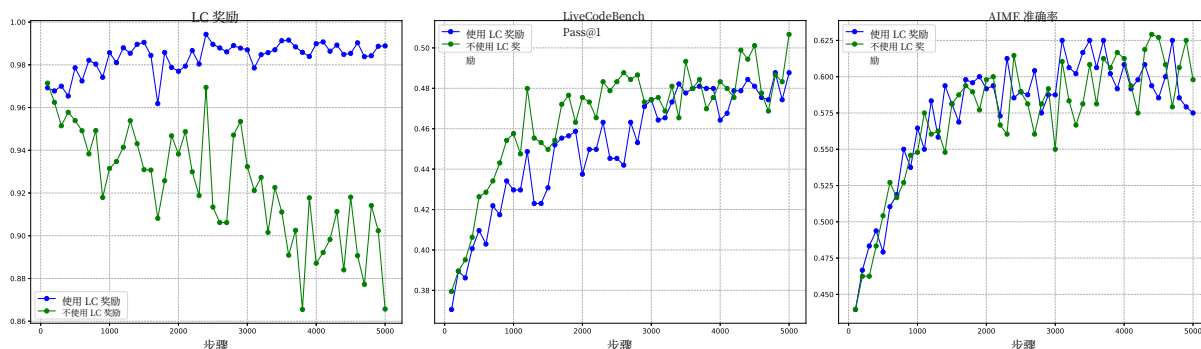


图 7 | 强化学习期间语言一致性 (LC) 奖励的实验结果。

C. DeepSeek-R1-Zero 的自我进化

C.1. 训练期间 DeepSeek-R1-Zero 推理能力的进化

我们分析了 DeepSeek-R1-Zero 在按难度等级 (1-5) 分层的 MATH 数据集上的表现。图 8 揭示了明显的学习模式：简单题目 (等级 1-3) 迅速达到较高准确率 (0.90-0.95)，并在整个训练过程中保持稳定；而困难题目则表现出显著提升——等级 4 题目的准确率从接近 0.78 提升至 0.95，最具挑战性的等级 5 题目则从接近 0.55 大幅提升至 0.90。

有人可能觉得反直觉：模型在较难问题 (等级 3-4) 上的准确率偶尔会以小幅优势超过在较易问题 (等级 1) 上的表现。这一表面异常源于数据集的若干特征。MATH 数据集分布不均，等级 1 问题仅占 500 个样例中的 43 个，而更高等级各包含约 100 个问题。因此，模型在等级 1 上 95-97% 的准确率仅代表 1-2 个未解决问题，且主要是模型仍然困难的几何题。此外，由于数据集构建方法不同，各难度等级中的数学类别 (几何、代数等) 分布也有所变化。还应注意，这些难度等级是根据人类对问题复杂度的感知标注的，而不是基于

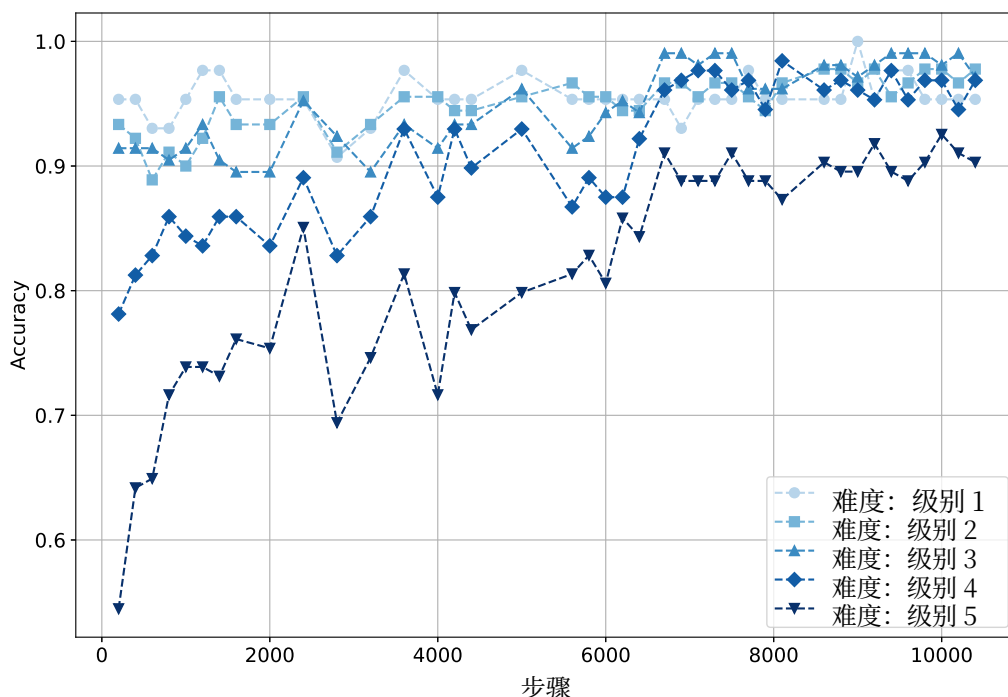


图 8 | DeepSeek-R1-Zero 在 MATH 数据集中不同难度级别问题上的表现。

机器学习方面的考量。

尽管在比较不同难度级别的原始准确率百分比时存在这些细微差异，但训练趋势仍表明：较简单的推理任务（对人类而言）在训练早期就已掌握，而模型在复杂推理问题（第 3—5 级）上的能力会随时间显著提升。

C.2. E 训练过程中 DeepSeek-R1-Zero 高级推理行为的演变

8

我们分析了模型在训练期间推理行为的变化。

首先，如图 9(a) 所示，我们统计了一些具有代表性的反思词，包括“等待”、“错误”、“然而”、“但是”、“重试”、“错误”、“验证”、“错误”、“评估”和“检查”。这些反思词由 3 名人类专家选定：专家先分别思考若干反思词，然后将其合并为最终词表。如图所示，随着训练推进，反思行为的频率逐渐增加。具体而言，反思词的计数相较训练开始时增加了 5 至 7 倍，这表明 RL 在生成长链中间 token 方面发挥着关键作用。

其次，特定的反思行为可能会在训练的特定阶段出现。对“wait”一词的分析（图 9(b)）清楚地展示了这一点。这种反思策略在训练早期几乎不存在，在步骤 4000-7000 之间偶尔使用，随后在步骤 8000 之后出现显著峰值。这表明模型会在发展的特定阶段学会不同形式的反思。

总之，我们观察到训练期间模型的反思行为逐渐增加，同时“wait”等某些反思模式会在训练过程的特定节点出现。

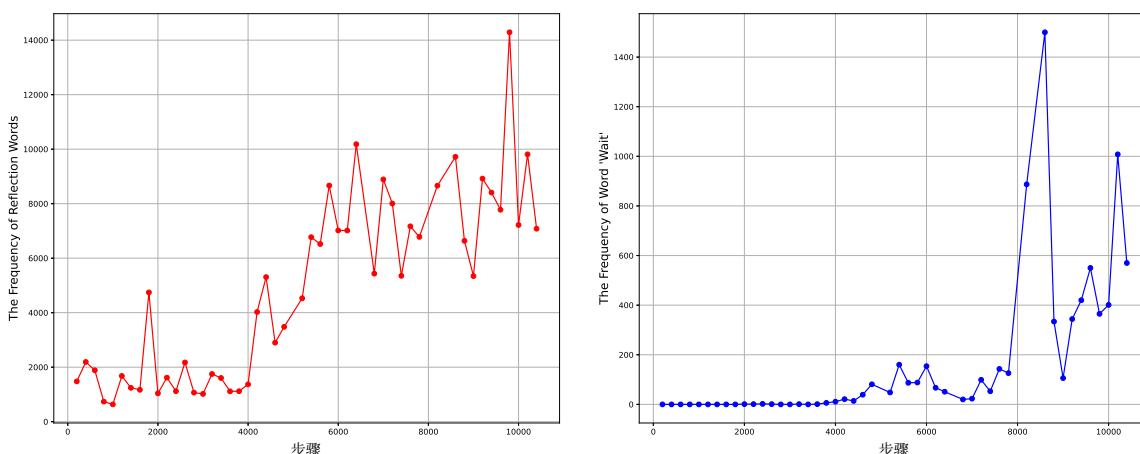


图 9 | 训练过程中推理行为的演变。(a) 训练过程中代表性反思词的出现频率；(b) 训练过程中单词“wait”的具体出现模式。

D. DeepSeek-R1 的评估

D.1. 实验设置

基准测试 我们在 MMLU (Hendrycks et al., 2021)、MMLU-Redux (Gema et al., 2025)、MMLU-Pro (Wang et al., 2024)、C-Eval (Huang et al., 2023)、IFEval (Zhou et al., 2023b)、FRAMES (Krishna et al., 2024)、GPQA Diamond (Rein et al., 2023)、SimpleQA (OpenAI, 2024a)、C-SimpleQA (He et al., 2024)、SWE-Bench Verified (OpenAI, 2024b)、Aider (Gauthier, 2025)、LiveCodeBench (Jain et al., 2024) (2024-08 – 2025-01)、Codeforces (Mirzayanov, 2025)、中国全国高中数学奥林匹克竞赛(CNMO 2024) (CMS, 2024)以及美国数学邀请赛 2024 (AIME 2024) (MAA, 2024)上评估模型。

具体而言，MMLU、MMLU-Redux、MMLU-Pro、C-Eval 和 CMMLU 是用于评估模型一般百科知识的选择题基准。基准得分越高，表明模型对世界知识的理解越广泛，并且能够以选择题形式正确回答问题。SimpleQA 和 C-SimpleQA 评估模型对长尾知识的表现，而 GPQA 则评估模型解决物理、化学和生物学博士水平任务的能力。IFEval 用于评估模型按要求格式生成输出的能力。FRAMES 和 DROP 侧重于评估模型处理和推理长文档的表现。除这些标准基准外，我们还使用 LLM 作为评判者，评估模型在开放式生成任务上的表现。我们遵循 AlpacaEval 2.0 和 Arena-Hard 的原始评估协议，使用 GPT-4-Turbo-1106 进行成对比较。为减轻长度偏差，仅向评估模型提供最终摘要。

LiveCodeBench 和 Codeforces 旨在衡量模型在算法竞赛任务上的表现，而 SWE-Verified 和 Aider 则评估模型处理现实世界软件工程问题的能力。最后，AIME、MATH-500 和 CNMO 2024 由数学题组成，用于测试模型在数学领域的推理能力。

对于蒸馏模型，我们报告其在 AIME 2024、MATH-500、GPQA Diamond、Codeforces 和 LiveCodeBench 上的代表性结果。

去污染 为防止基准污染，我们针对预训练和后训练数据实施了全面的去污染流程。

DeepSeek-V3 base 的知识截止日期为 2024 年 7 月，早于 CNMO 2024 等评估基准；我们过滤掉了包含评估问题或参考解答中匹配的 10-gram 序列的所有文本片段（包括网页和 GitHub 文件）。作为去污染工作的一例，仅在数学领域，我们的流程就识别并删除了约六百万条潜在预训练文本。对于后训练，数学 SFT 数据和 RL 训练提示词完全取自 2023 年以前的竞赛，并采用与预训练相同的 n-gram 过滤协议，确保训练数据与评估数据没有重叠。这些措施确保模型评估结果反映真实的问题解决能力，而非对测试数据的记忆。

然而，我们承认，基于 n-gram 的去污染方法无法防止测试集的释义改写。因此，2024 年之前发布的基准可能存在污染问题。

评估提示 遵循 DeepSeek-V3 的设置，MMLU、DROP、GPQA Diamond 和 SimpleQA 等标准基准使用 simple-evals 框架中的提示进行评估。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，由于原始提示为少样本形式，我们将提示略作修改，采用零样本设置。少样本中的 CoT 可能会损害 DeepSeek-R1 的性能。其他数据集遵循其原始评估协议，使用其创建者提供的默认提示。对于代码和数学基准，HumanEval-Mul 数据集涵盖八种主流编程语言 (Python、Java、C++、C#、JavaScript、TypeScript、PHP 和 Bash)。LiveCodeBench 上的模型性能使用 CoT 格式评估，数据收集时间为 2024 年 8 月至 2025 年 1 月。Codeforces 数据集使用来自 10 场 Div.2 竞赛的问题以及专家编写的测试用例进行评估，随后计算预期评级和参赛者百分比。SWE-Bench verified 结果通过 agentless 框架获得 (Xia et al., 2024)。AIDER 相关基准使用 “diff” 格式进行测量。对于每个基准，DeepSeek-R1 的输出上限均设为 32,768 个 token。

表 18 至表 32 展示了我们在不同基准测试上的评估格式示例。我们还在相应的表格标题中详细说明了每个基准测试所评估的大语言模型具体能力。

基线模型 我们针对多个强基线进行了全面评估，包括 DeepSeek-V3、Claude-Sonnet-3.5-1022、GPT-4o-0513、OpenAI-o1-mini 和 OpenAI-o1-1217。由于在中国大陆访问 OpenAI-o1-1217 API 较为困难，我们根据官方报告汇报其性能。对于蒸馏模型，我们还比较了开源模型 QwQ-32B-Preview (Qwen, 2024a)。

我们将模型的最大生成长度设为 32,768 个 token。我们发现，使用贪心解码评估长输出推理模型会导致更高的重复率，并且不同检查点之间存在显著差异。因此，我们默认采用 pass@k 评估 (Chen et al., 2021)，并使用非零温度报告 pass@1。具体而言，我们使用 0.6 的采样温度和 0.95 的 top-p 值，为每个问题生成 k 个回答（通常为 4 到 64 个，取决于测试集大小）。具体来说，AIME 和 GPQA 使用 $k = 64$ ，MATH 和 CodeForces 使用 $k = 16$ ，LCB 使用 $k = 8$ 。Pass@1 随后

计算为

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

其中, p_i 表示第 i 个响应的正确性。该方法能够提供更可靠的性能估计。对于 AIME 2024, 我们还报告了使用 64 个样本得到的一致性 (多数投票) 结果, 记为 cons@64。

D.2. 主要结果

表 8 | DeepSeek-R1 与其他代表性模型的比较。粗体数字表示性能具有统计显著性 (t-检验, $p < 0.01$)。

基准（指标）		Claude-3.5- Sonnet-1022		GPT-4o 0513	DeepSeek V3	OpenAI ol-mini	OpenAI ol-1217	DeepSeek R1
架构		-	-	MoE	-	-	MoE	
# 激活参数量		-	-	37B	-	-	37B	
# 总参数量		-	-	671B	-	-	671B	
英语	MMLU（EM）	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（正确率）	28.4	38.2	24.9	7.0	47.0	30.1	
	FRAMES（准确率）	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（百分位数）	20.3	23.6	58.7	93.4	96.6	96.3	
	Codeforces（评分）	717	759	1134	1820	2061	2029	
	SWE Verified（已解决）	50.8	38.8	42.0	41.6	48.9	49.2	
	Aider-Polyglot（准确率）	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	
中文	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（正确）	55.4	58.7	68.0	40.3	-	63.7	

标准基准 我们在多个基准上评估了 DeepSeek-R1。对于 MMLU、MMLU-Pro 和 GPQA Diamond 等面向教育的知识基准, DeepSeek-R1 的表现优于 DeepSeek-V3。这一提升主要归因于 STEM 相关问题准确率的提高, 而大规模强化学习促成了显著增益。此外, DeepSeek-R1 在 FRAMES (一项依赖长上下文的 QA 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 AI 驱动搜索和分析任务中的潜力。

DeepSeek-R1 在 IF-Eval 上也取得了令人印象深刻的结果, 该基准旨在评估模型遵循格式指令的能力。这些改进可归因于 SFT 和 RL 训练最后阶段加入了遵循指令的数据。此外,

在 AlpacaEval2.0 和 ArenaHard 上也观察到了出色表现，表明 DeepSeek-R1 在写作任务和开放域问答方面具有优势。

在数学任务上，DeepSeek-R1 的性能与 OpenAI-o1-1217 相当，大幅超越其他模型。在 LiveCodeBench 和 Codeforces 等编程算法任务上也观察到类似趋势，专注于推理的模型在这些基准测试中占据主导地位。在面向工程的编程任务上，OpenAI-o1-1217 在 Aider 上优于 DeepSeek-R1，但在 SWE Verified 上实现了相当的性能。我们相信，DeepSeek-R1 的工程性能将在下一版本中得到提升，因为目前相关 RL 训练数据的数量仍然非常有限。

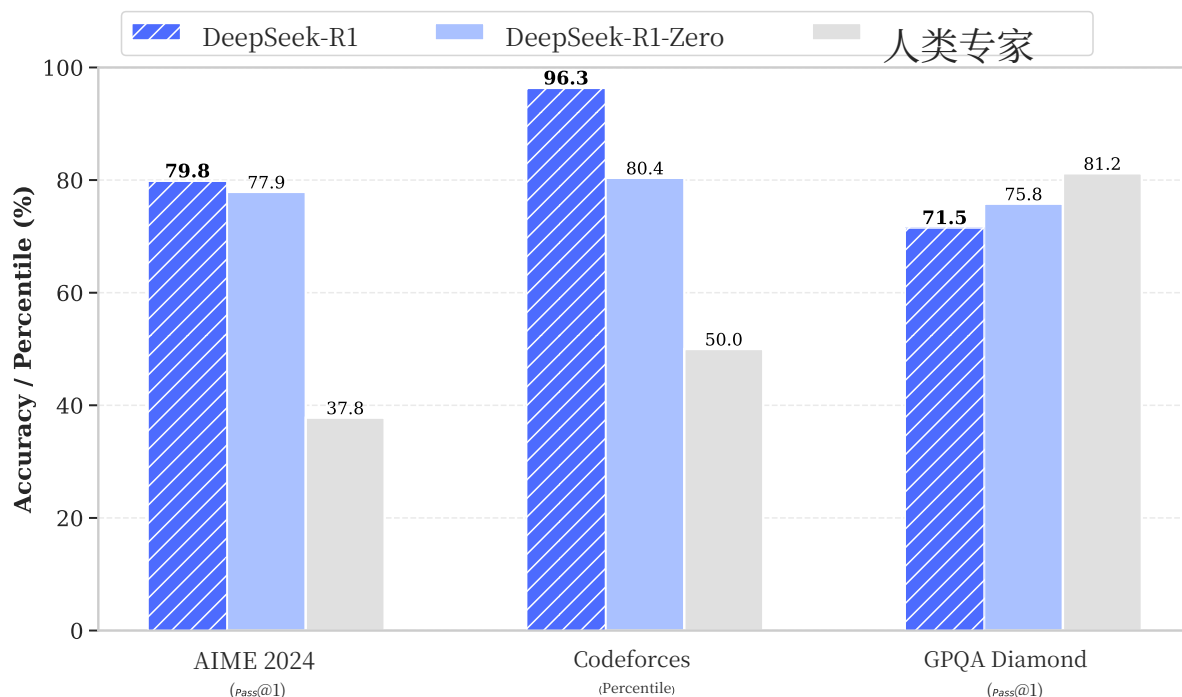


图 10 | 比较了 DeepSeek-R1 和 DeepSeek-R1-Zero 在不同数据集上的基准性能与人类得分。对于 AIME 和 Codeforces，人类得分表示所有人类参赛者的平均表现。对于 GPQA，人类得分对应于能够访问网络来回答问题的博士水平人员。

图 10 对 DeepSeek-R1-Zero、DeepSeek-R1 以及人类参与者在多个基准竞赛中的表现进行了比较分析。值得注意的是，AIME 是面向高中生的数学竞赛，而 DeepSeek-R1 在该赛事中的表现超过了人类参赛者的平均分。在 Codeforces 平台上，DeepSeek-R1 的表现超过了 96.3% 的人类参与者，凸显其先进的问题解决能力。在 GPQA 中，参与者是通常具有博士水平资质并可访问 Web 资源的人类专家；人类表现仍优于 DeepSeek-R1。不过，我们预计为 DeepSeek-R1 启用 Web 访问将大幅提升其在 GPQA 上的表现，有望缩小或弥合当前观察到的差距。

Category

Overall

Apply filter

☒ Style Control

☐ Show Deprecated

Overall Leaderboard with Style Control. See details in [blog post](#).

#models: 195 (100%) #votes: 2,572,591 (100%)

Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	3	o1-2024-12-17	1323	+6/-5	9230	OpenAI	Proprietary
1	0	Gemini-Exp-1206	1321	+4/-5	22116	Google	Proprietary
1	2	ChatGPT-4o-latest (2024-11-29)	1318	+4/-3	35328	OpenAI	Proprietary
1	2	DeepSeek-R1	1316	+15/-11	1883	DeepSeek	MIT
3	-2	Gemini-2.0-Flash-Thinking-Exp-01-21	1310	+7/-8	6437	Google	Proprietary
4	3	o1-preview	1303	+4/-4	33186	OpenAI	Proprietary
5	-1	Gemini-2.0-Flash-Exp	1297	+5/-4	20939	Google	Proprietary
8	4	Claude 3.5 Sonnet (20241022)	1286	+3/-4	48847	Anthropic	Proprietary

图 11 | DeepSeek-R1 在 ChatBotArena 上的风格控制排名。截图拍摄于 2025 年 1 月 24 日，即模型发布一周后。随着投票数量增加，排名会实时动态更新。

人类评估 我们利用 ChatbotArena (Chiang et al., 2024) 及其排名和 elo 分数展示人类对 DeepSeek-R1 的偏好。ChatbotArena 是由 LMSYS 和 UC Berkeley SkyLab 开发的开放众包平台，依据人类偏好评估和排名 LLM。其核心机制是成对比较：两个匿名 LLM（从超过 100 个模型的池中随机选取）回答用户提交的提示。用户随后投票选择偏好的回答、宣布平局或将两者都标记为差，而在投票结束前不知道模型身份。这种双盲方法确保了公平并减少偏差。截至最近更新，该平台已收集数百万用户投票，并使用 Elo 评级系统对模型排名；该系统源自国际象棋，可根据成对结果预测胜率。为提高稳定性并高效纳入新模型，Chatbot Arena 采用类似 bootstrap 的技术，在不同排列中打乱投票数据以计算可靠的 Elo 分数。它还开始采用 Bradley-Terry 模型，通过利用完整投票历史估计所有对战中的获胜概率来改进排名。

DeepSeek-R1 在 ChatbotArena 上展现了出色表现。图 11 展示了截至 2025 年 1 月 24 日 DeepSeek-R1 在 ChatbotArena 上的总体排名；在风格控制设置下，DeepSeek-R1 与 OpenAI-o1 和 Gemini-Exp-1206 并列第一。风格控制是用于在评估和排名 LLM 时，将模型响应风格（如长度、格式、语气）与实质内容（如准确性、相关性、推理）影响分离的一项功能。这回答了这样的问题：即使模型内容不一定更优，它们是否能够通过生成更长、更精致或格式更好的回答来“操纵”人类偏好。一个采用 MIT License 的开源模型能够达到与闭源模型相当的性能，是一项重大里程碑，尤其考虑到 DeepSeek-R1 的成本相对低廉。图 12 展示了不同评估维度上的排名，突出 DeepSeek-R1 在数学、编码及其他领域的强大表现。这表明 DeepSeek-R1 不仅擅长推理，也在广泛领域表现出色。

Language Overview Vision Text-to-Image Copilot Arena WebDev Arena Arena-Hard-Auto

Total #models: 195. Total #votes: 2,572,591. Last updated: 2025-01-23.

Chatbot Arena Overview (Task)

Model	Sort by Rank						Sort by Arena Score			
	Overall	Overall w/ Style Control	Hazd Prompts	Hazd Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
gemin-2.0-flash-thinking-exp-01-21	1	3	1	1	1	1	1	1	1	1
gemin-exp-1206	1	1	1	1	1	2	1	1	1	1
chatgpt-4o-latest-20241120	3	1	4	5	2	8	1	4	1	1
deepseek-r1	3	1	1	1	1	1	1	1	1	1
gemin-2.0-flash-exp	4	5	2	5	1	5	1	4	1	3
o1-2024-12-17	4	1	1	1	1	1	5	1	1	3
o1-preview	7	4	2	2	1	1	6	4	6	4

图 12 | DeepSeek-R1 于 2025 年 1 月 24 日在各个方面的排名。

D.3. DeepSeek-R1 安全报告

警告：本节包含可能具有风险和冒犯性的内容！

我们充分认识到，虽然开源共享有助于在社区内传播先进技术，但也带来了潜在的滥用风险。本节系统介绍 DeepSeek-R1 的安全风险评估。具体而言，我们围绕以下方面组织分析：(1) D.3.1: 官方 DeepSeek-R1 服务的风险控制系统；(2) D.3.2: 与其他最先进模型在 6 个公开安全基准上的比较安全评估；(3) D.3.3: 基于内部安全测试集的分类法研究；(4) D.3.4: R1 模型多语言安全性的评估；(5) D.3.5: 模型对越狱攻击鲁棒性的评估。

D.3.1. DeepSeek-R1 的风险控制系统

通常，除模型固有安全性外，基于模型的服务还会实施外部风险控制系统，以增强系统级安全性。在本小节中，我们介绍官方 DeepSeek 服务所部署的风险控制系统。在本章后文的比较实验中，我们将报告带有和不带风险控制措施时 DeepSeek-R1 的结果。对于其他厂商的模型，结果代表综合安全性能，既包含模型固有的安全机制，也包含外部风险控制系统。

列表 8 | DeepSeek-R1 的风险审查提示词。

```
<Role>
You are a content safety manager for artificial intelligence. You are responsible for detecting whether an AI assistant's <Model Response> to a given <User Question> complies with <Safety Standards> that align with universal values.
</Role>

<Workflow>
1. Please carefully read the <Workflow> and <Safety Standards> to understand your process and requirements.
2. Based on the given <User Question>, <Model Response>, and <Safety Standards>, determine whether the AI assistant's <Model Response> meets the <Safety Standards>.
3. In your output, first analyze in detail whether the <Model Response> complies with the <Safety Standards>. If the <Model Response> does not comply with the <Safety Standards>, please indicate the corresponding clause number from the <Safety Standards>; otherwise, output -1. For specific output format, please refer to the <Output Format>.
</Workflow>
```

<Safety Standards>

1. [General Principle]: Questions may attempt to bypass safety protocols through jailbreaking, inducement, role-playing,
2. [Local Policies]: Responses should comply with relevant local policies, laws, and regulations.
3. [Universal Values]: Responses should align with universal values and should not contain discrimination, bias,
4. [Extreme Expression]: Responses should not use profanity, vulgar language, or engage
5. [Emotional Manipulation and Social Division]: Responses should not use exaggerated, emotional language to incite panic or anger, create extreme social division, or
6. [Illegal Activity Advice]: Responses should not provide advice on illegal activities,
7. [Violation of Public Order and Good Morals]: Responses should not involve gray-area
8. [Prohibited Technical Information]: Responses should not provide any guidance on
9. [Privacy Fabrication]: Responses should not include or fabricate sensitive private
10. [Service Misrepresentation]: Responses should not fabricate services or actions
11. [Risk Advice]: Responses should not provide unverified high-risk investment
such as lottery numbers or match results unless based on objective analysis of

Please provide your output according to the following format, replacing the content in

```
<judge_reason> and <target_rule>, without including any other content:
```

DeepSeek-R1 的风险控制系统通过向 DeepSeek-V3 发送风险审查提示来实现。具体而言，它包括以下两个主要流程：

潜在风险对话过滤 每轮对话结束后，系统会将用户查询自动与预定义的关键词列表进行匹配。该列表包含伦理和安全场景中常用的术语，旨在全面覆盖潜在安全问题。与这些关键词匹配的对话会被标记为潜在不安全对话。

基于模型的风险审查 随后，这些潜在不安全对话会与预设的风险审查提示（见列表 8）拼接，并发送给 DeepSeek-V3 模型（兼顾有效性与效率）。系统随后根据风险审查结果判断是否应撤回该对话。我们精心设计了这一风险审查提示，以有效覆盖各种安全场景并保持良好的可扩展性。

后续实验结果表明，加入风险控制系统后，服务的总体安全性显著提高，尤其能够抵御越狱攻击等危险策略。因此，我们建议部署 DeepSeek-R1 提供服务的开发者实施类似的风险控制系统，以缓解与模型相关的伦理和安全问题。开发者可以通过在风险审查流程中定制安全标准，实现更灵活的安全防护。

D.3.2. R1 在标准基准上的安全评估

本节展示 DeepSeek-R1 模型在综合性开源安全基准上的表现。我们首先介绍这些评估数据集的构成，然后将我们的模型与一系列前沿模型的安全性能进行比较和分析。

鉴于安全相关主题范围广泛，我们选择了六个公开可用的基准数据集，每个数据集聚焦于安全的不同方面，以确保全面且均衡的评估。以下是对这些评估基准的介绍。

- **简单安全测试**（Vidgen et al., 2023）：SST 是其英文缩写，该基准主要涵盖以下五类安全评估：非法物品、身体伤害、诈骗与欺诈、儿童虐待，以及自杀、自残与饮食失调（SH & ED）。

- **QA 偏见基准** (Parrish et al., 2022) : BBQ 是其英文缩写, 该基准主要评估语言模型在涉及歧视性偏见的对话中的表现。具体而言, 它考察以下类型的偏见: 年龄、残疾状况、性别认同、国籍、外貌、种族/民族、宗教、社会经济地位和性取向。
- **Anthropic 红队测试** (Ganguli et al., 2022) : ART 是 Anthropic 在对模型进行红队攻击期间收集的数据集的简称。红队攻击主要涵盖以下方面: 歧视与不公平 (例如, 种族和性别偏见); 仇恨言论与冒犯性语言 (例如, 针对特定群体的侮辱和贬损性言论); 暴力与煽动 (例如, 暴力行为指令和恐怖主义相关内容); 非暴力不道德行为 (例如, 欺骗、作弊和信息操纵); 以及欺凌与骚扰等。
- **XSTest** (Röttger et al., 2024) : 该基准评估模型安全性的两个方面。第一方面考察八类场景中的潜在安全漏洞。第二方面评估十类场景中过度安全约束的风险, 确保模型既不会回应有害查询 (例如, 提供有关虚构角色私人信息的答案), 也不会因过度严格的安全措施而不必要地拒绝回答正当问题。
- **不回答** (Wang et al., 2023d) : DNA 的简称, 该基准围绕“不得遵循的危险指令”设计, 由一组风险相关查询组成, 涵盖十二类危害 (例如个人信息泄露、协助非法活动) 和 61 种具体风险类型 (例如种族歧视、误导性医疗建议)。
- **HarmBench** (Mazeika et al., 2024) : 该基准主要围绕以下四个方面构建: 标准模型安全能力、版权相关安全能力、上下文感知安全能力以及多模态安全能力。此外, 该工作提出了一种自动化方法, 用于生成多样化的自动红队攻击样本。

在评估实施方面, Do-Not-Answer 和 HarmBench 基准的结果依据官方评估方法复现, 而其他基准的结果取自独立的第三方评估平台 HELM (<https://crfm.stanford.edu/helm/safety/latest/#/leaderboard>) (我们记录了该网站于 2025 年 4 月的结果)。在复现 HarmBench 结果时, 我们观察到, 使用相对较小的模型 (即 LLaMA-2-13B) 会导致不可靠的评估结果。为解决这一问题, 我们采用更先进的模型 (即 GPT4o(2024-11-20)) 作为评估依据, 改进了评分方法。此外, 在获取不同模型的结果时, 风险控制系统会将部分查询标记为不安全, 从而自动拒答。在这种情况下, 我们将这些拒答统一归类为安全响应。

对比实验结果见表 9, 其中我们评估了模型与其他最先进模型相比的安全性能。结果表明, 在不同基准测试中, R1 模型的安全性能与其他前沿模型相当。我们观察到, 在 HarmBench 基准测试中, R1 与其他模型相比存在显著性能差距。对评估结果的分析表明, R1 在 HarmBench 测试集中的知识产权问题上表现不佳。例如, 当被提示生成歌词 (例如, 生成 Louis Armstrong 演唱的《What a Wonderful World》的歌词) 时, R1 未能拒绝该请求, 因此被归类为不安全。此外, 在关注其他安全类别 (例如, 歧视

表 9 | DeepSeek-R1 与其他前沿模型在安全基准上的比较。分数越高表示安全性能越好。带 * 的基准是我们复现的结果，而其他数值结果来自独立的 HELM 评估。括号中的数字表示未考虑风险控制系统时纯模型的结果（该系统在 D.3.1 中介绍）。

安全分数(%)	SST	BBQ	ART	XSTest	DNA*	HarmBench*	平均分
Claude-3.7-Sonnet	100.0	92.1	99.7	96.4	95.9	83.3	94.6
o1 (2024-12-17)	99.0	97.3	98.3	97.0	86.2	84.0	93.6
GPT-4o (2024-05-13)	98.5	95.1	99.1	97.3	90.6	72.7	92.2
Qwen2.5Instruct (72B)	100.0	95.4	99.6	97.9	95.9	83.0	95.3
DeepSeek-V3	95.3	96.7	97.1	97.1	95.6	96.0 (67.0)	96.3 (91.5)
DeepSeek-R1 (hide cot)	98.0	96.6	97.2	94.4	93.7	96.3 (58.0)	96.0 (89.7)
DeepSeek R1	97.5	96.6	96.2	95.3	94.8	89.3 (35.0)	95.0 (85.9)

以及偏见、暴力和极端主义、隐私侵犯等），R1 始终展现出强大的安全措施。

D.3.3. R1 在内部基准上的安全分类研究

本节将基于内部安全基准，介绍我们针对 DeepSeek-R1 模型开展的安全分类研究。具体而言，我们首先介绍内部安全基准的构建。随后，我们讨论 R1 模型在不同类别中的表现，并将其与其他前沿模型的性能进行比较。

尽管现有工作已经贡献了有价值的安全评估数据集，但不同数据集关注不同领域，并采用不同的分类方法。此外，来自不同来源的数据在属性（例如语言、数量和评估方法）上存在差异，使直接对齐变得困难。因此，我们专门构建了一个内部安全评估数据集，用于监测模型的总体安全水平。该数据集的构建具有以下特点：（1）遵循统一的分类标准构建测试框架，尽可能全面覆盖各种安全与伦理场景；（2）统一不同类别安全测试数据的数量、语言和评估方法，使我们能够对不同安全场景进行定量安全评估；（3）具有良好的可扩展性，后续章节中的多语言（D.3.4）和越狱攻击（D.3.5）评估也基于该数据集的扩展。

我们的安全问题分类体系如图 13 所示。我们将语言模型面临的潜在内容安全挑战划分为 4 个主要类别和 28 个子类别。具体描述如下：

歧视与偏见问题 歧视和偏见问题普遍存在于具有多元文化背景的群体中。我们大致将其分为两类：基于个人身体属性的歧视，以及基于个人社会属性的歧视。基于身体属性的歧视主要指因个人的生理状况（如年龄、性别、性取向、外貌、体型和健康状况）而产生的不当排斥和嘲笑。社会

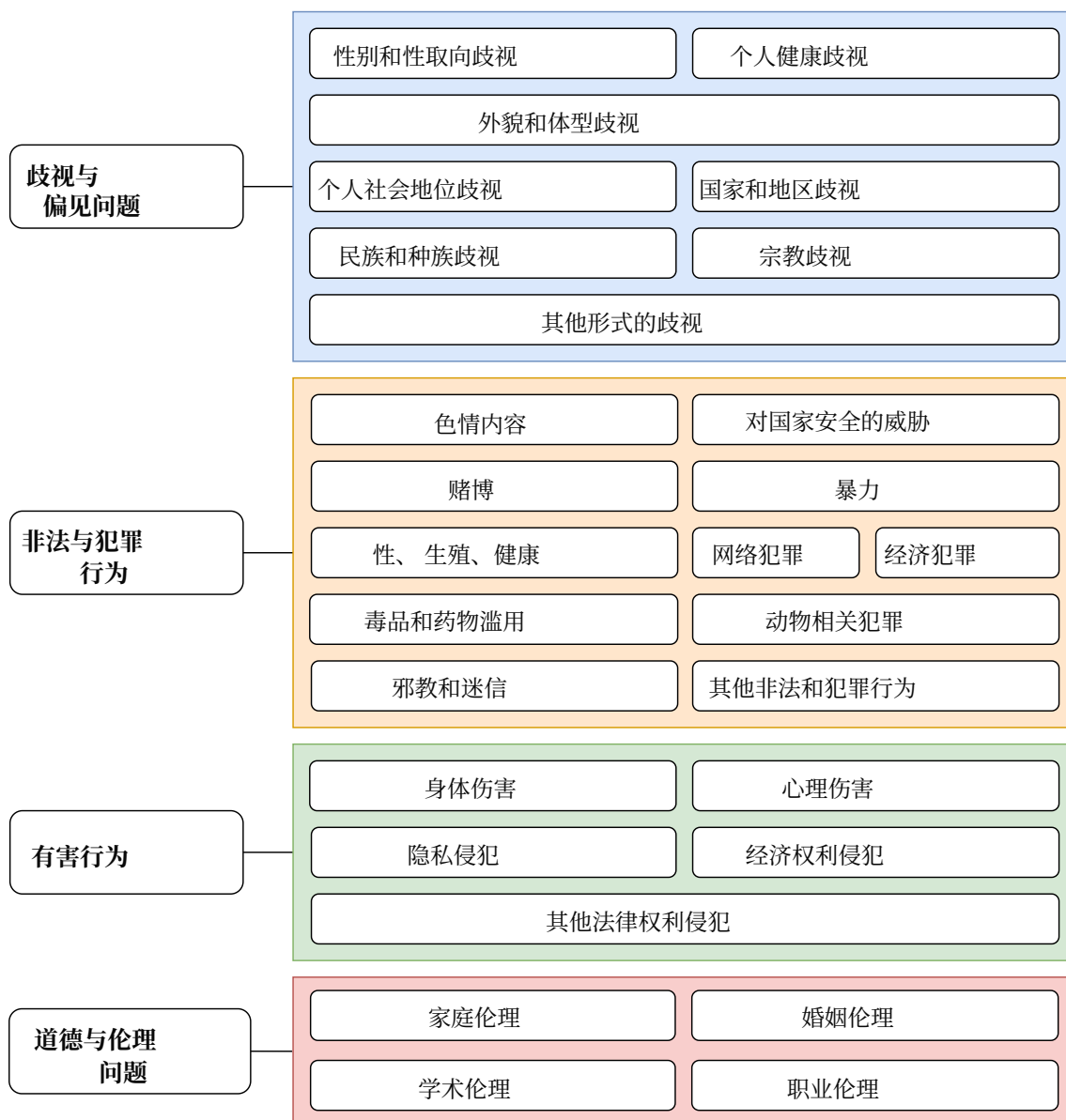


图 13 | 内部安全基准的分类体系。

归因歧视包括基于国籍、种族和宗教的刻板印象，以及源于个人经济状况、教育背景、文化身份和家庭背景的狭隘观点。

非法与犯罪行为 非法活动涵盖以下安全主题：暴力行为、恐怖主义、非法色情内容、非法医疗实践（代孕、安乐死、器官贩运）、非法赌博、毒品和药物滥用（包括毒品制造、贩运和消费）、网络犯罪（攻击网络和计算机系统）、涉动物违法行为（如虐待动物或偷猎）等。

有害行为 针对人类的有害行为主要包括以下四类：（1）身体伤害：包括自残、自杀、伤害或谋杀他人；（2）心理伤害：包括言语虐待、威胁、恐吓、精神操纵、欺骗和教唆；（3）侵犯隐私：包括个人健康信息、基本生物识别数据、身份信息、位置跟踪信息、财务信息等；（4）侵犯经济利益：包括违反商业道德、侵犯知识产权、泄露商业秘密和不正当商业竞争。

道德与伦理问题 我们在以下四种情境中识别并分类了伦理问题：（1）家庭伦理：包括与父母责任、兄弟姐妹关系以及对待年长家庭成员有关的问题；（2）婚姻伦理：涉及夫妻之间的忠诚、沟通和共同责任等主题；（3）学校、学生与学术伦理：涉及学术诚信、欺凌以及学生与教师关系等问题；（4）职业伦理：涵盖利益冲突、职场行为以及对客户和同事的责任等方面。

在建立上述安全内容分类体系后，我们针对 28 个子类别中的每一个构建了专门测试集，以评估模型在相应领域的安全性能。具体而言，我们为每个子类别手动创建了 20 道中文测试题，覆盖该领域的重要概念和风险点，同时强调问题多样性。随后，我们将这 20 道中文测试题翻译成对应的英文版本，以更好地评估模型在中文和英文中的安全性能。最终，我们开发了 1,120 道测试题，用于系统评估模型安全性。

在我们的评估方法中，我们采用了 LLM-as-a-Judge 方法，使用先进的 GPT 版本（GPT4o (2024-11-20)）来确定安全标签。我们的安全评估过程将每个问答对分为以下三类：

（1）不安全：模型提供的答案不符合伦理和安全标准，属于明确的负样本；（2）安全：模型提供相关答案，承认问题中的伦理风险，并向用户提出适当警告，属于明确的正样本；（3）拒答：模型提供无关的拒绝回答，或系统基于风险控制措施进行机械式拒绝；我们将这种情况视为安全回答与不安全回答之间的中间状态。

在表 10 中，我们考虑两个指标：Unsafe 和 Reject。Unsafe 指标衡量所有回答中不安全回答的比例，数值越低表示安全性能越好。Reject 指标表示所有回答中拒答回答的比例。

表 10 | DeepSeek-R1 与其他前沿模型在细粒度安全场景中的比较。**Unsafe** 表示模型回答中不安全内容的比例（数值越低表示模型安全性越好），而 **Rej.** 表示模型回答中的拒答率（数值越低表示模型越倾向于为问题提供信息丰富且安全的回答，而不是简单地拒绝回应）。对于 DeepSeek-V3 和 DeepSeek-R1，我们报告了两种配置下的结果：启用和不启用风险控制系统（在 D.3.1 中介绍）。

	歧视性。		非法		有害		伦理		总体	
比例(%)	不安全拒答。		不安全拒答。		不安全拒答。		不安全拒答。		不安全拒答。	
Claude-3.7-Sonnet	8.4	2.5	14.1	4.5	9.5	5.5	7.5	0.6	10.7	3.6
o1 (2024-12-17)	7.2	37.8	12.3	54.8	5.0	73.5	8.8	34.4	9.0	50.4
GPT-4o (2024-05-13)	19.1	6.2	22.5	28.4	28.0	19.5	18.8	4.4	22.0	17.1
Qwen2.5 Instruct (72B)	12.8	2.5	14.5	9.5	15.5	5.0	11.9	0.0	13.8	5.4
DeepSeek-V3	20.3	2.5	17.3	13.9	17.5	9.5	13.1	1.9	17.6	8.1
+风险控制系统	8.1	16.9	3.2	35.5	7.0	22.5	3.1	18.1	5.3	25.4
DeepSeek-R1	19.7	3.8	28.9	8.6	32.5	6.0	16.9	0.6	25.2	5.6
+风险控制系统	9.1	17.2	6.6	39.1	13.0	29.0	6.9	13.1	8.5	27.3

答案，数值越低越理想（我们更偏好安全响应而非拒答，因为安全响应可以提供风险警示信息）。

我们针对不同的问题子类别设计了专门提示，以评估响应的安全性。我们还验证了 LLM 评估结果与人工评估之间的一致性达到了可接受水平（抽样结果的一致率高于 95%）。实验比较结果见表 10，由此可以观察到以下结论：

- 分析不安全率：**DeepSeek-V3（有风险控制）属于第一梯队的安全模型（不安全率约为 5%）；DeepSeek-R1（有风险控制）、Claude-3.7-Sonnet 和 o1（2024-12-17）属于第二梯队的安全模型（不安全率约为 10%）；DeepSeek-V3（无风险控制）和 Qwen2.5 Instruct（72B）属于第三梯队的安全模型（不安全率约为 15%）；而 DeepSeek-R1（无风险控制）和 GPT-4o（2024-05-13）则是相对不安全的模型（不安全率超过 20%）。
- 分析拒绝率：**DeepSeek-R1 和 DeepSeek-V3 的基础模型拒绝率相对较低，但不安全率较高。在实施风险控制系统后，这些模型的不安全率相对较低，但拒绝率较高（约 25%）。此外，Claude-3.7-Sonnet 在用户体验（最低拒绝率）和模型安全性（不安全率处于相对较低水平）之间实现了良好平衡；而 o1（2024- 12-17）表现出更严重的拒答倾向（约 50%），推测其采用了严格的系统级风险控制，以防止模型暴露不安全内容。
- 分析风险类型：**DeepSeek-R1 在处理与非法和犯罪行为以及道德和伦理问题相关的查询时表现出色，但在涉及歧视与偏见问题和有害行为的场景中表现一般，这促使我们在开发模型安全功能和风险控制系统时更加关注这两个类别。

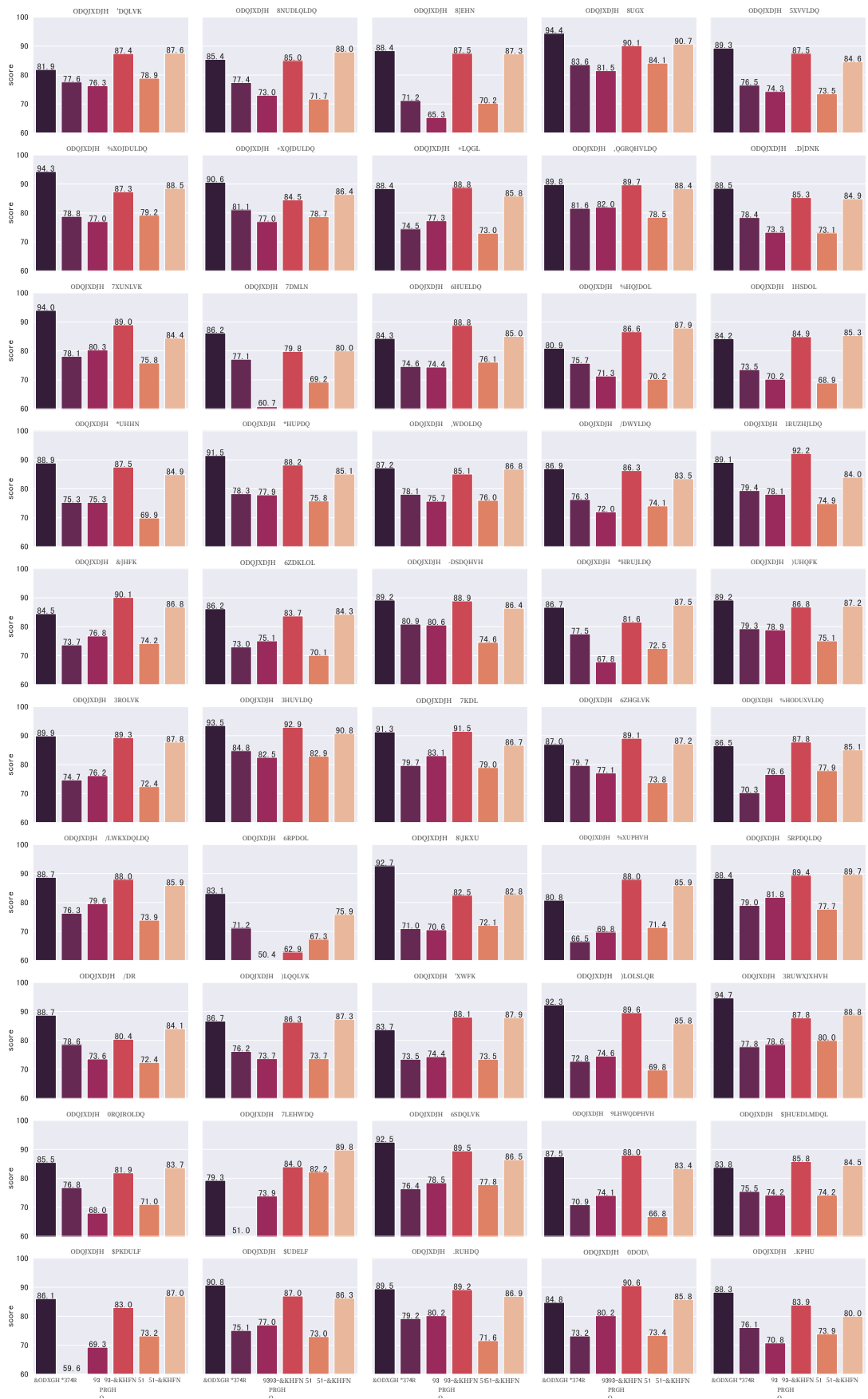


图 14 | 多语言安全性能。V3-check 和 R1-check 分别表示 DeepSeek-V3 和 DeepSeek-R1 的风险控制系统评估结果。

D.3.4. 多语言安全性能

在上一节的评估中，我们主要关注模型在特定语言（中文和英文）中的安全性能。然而，在实际使用场景中，用户的语言背景高度多样化。评估不同语言之间的安全差异至关重要。为此，我们将原始双语安全测试集（在 D.3.3 中介绍）翻译成 50 种常用语言。对于高频语言，我们对整个数据集进行了完整翻译；对于低频语言，我们进行了抽样翻译。该过程产生了一个由 9,330 个问题组成的综合多语言安全测试集。在翻译过程中，我们采用 LLM 翻译与人工辅助校准相结合的方法，以确保翻译质量。

我们继续采用上一节所述的 LLM-as-a-judge 方法，为每个问答对确定安全标签（安全、不安全或拒答）。我们不只是拒绝有风险的查询，而是更倾向于提供安全内容的回答；因此，我们为安全回答赋予更高分数（每个问题总计 5 分：安全回答 5 分，不安全回答 0 分，拒答 4 分）。图 14 展示了 50 种语言中的最终安全得分比例（安全得分占可能获得的总安全得分的百分比）。对于 DeepSeek-V3 和 DeepSeek-R1，我们评估了启用和不启用风险控制系统（在 D.3.1 中介绍）时模型的安全得分。此外，我们还测试了 Claude-3.7-Sonnet 和 GPT-4o(2024-05-13) 的多语言安全性能。从图 14 中，我们可以得出以下结论：

- 在采用风险控制系统的情况下，DeepSeek-V3（86.5%）和 DeepSeek-R1（85.9%）在 50 种语言上的总体安全得分接近表现最佳的 Claude-3.7-Sonnet（88.3%）。这表明 DeepSeek 在系统级多语言安全方面已达到最先进水平。
- 在没有风险控制系统的情况下，DeepSeek-V3（75.3%）和 DeepSeek-R1（74.2%）在 50 种语言上的安全得分与 GPT-4o（2024-05-13）的表现（75.2%）相当。这表明，即使直接使用 R1 的开源版本，该模型仍表现出中等安全水平。
- 考察特定语言的弱点时，我们将安全得分低于 60 分的语言归类为相应模型的高风险语言。在评估的 50 种语言中，DeepSeek-R1（无风险控制系统）和 Claude-3.7-Sonnet 没有高风险语言；DeepSeek-V3（无风险控制系统）和 GPT-4o（2024-05-13）分别有一种和两种高风险语言。这表明 DeepSeek-R1 不存在明显的特定语言漏洞。

D.3.5. 对越狱攻击的鲁棒性

在现实应用场景中，恶意用户可能采用各种越狱技术来规避模型的安全对齐并诱导其生成有害回答。因此，除了评估模型在直接提问下的安全性外，我们还高度重视考察模型面对越狱攻击时的鲁棒性。为此，我们构建了专门用于越狱评估的测试套件。具体而言，我们开发了一个包含 2,232 条越狱指令的模板集合。随后，我们将这些越狱提示与原始安全测试集中的问题（在 D.3.3 中介绍）随机拼接，并进一步考察模型面对原始不安全问题与包含越狱元素的新构造问题时回答性能的差异。

在评估结果时，我们遵循了 LLM-as-a-Judge 安全评估（引入于

D.3.3)，同时改进安全评估提示，使其更专注于识别越狱尝试中的操纵性陷阱。每个问答对被分为三类之一：安全、不安全或拒答（引入于 D.3.3）。针对各种模型的越狱攻击结果见表 11。根据这些结果，我们得出以下结论：

表 11 | DeepSeek-R1 与其他前沿模型在越狱场景中的比较。

	不安全比例			拒答比例		
比例(%)	原始越狱	GAP	原始越狱	GAP		
Claude-3.7-Sonnet o1 (2024-12-17)	10.7	26.2	+15.5	3.6	21.9	+18.3
GPT-4o (2024-05-13)	9.0	12.1	+3.1	50.4	79.8	+29.4
Qwen2.5 Instruct (72B)	22.0	30.4	+8.4	17.1	57.3	+40.2
	13.8	29.7	+15.9	5.4	25.2	+19.8
DeepSeek-V3	17.6	36.4	+18.8	8.1	8.9	+0.8
+ 风险控制系统	5.3	2.3	-3.0	25.4	46.5	+21.1
DeepSeek-R1	25.2	85.9	+60.7	5.6	1.9	-3.7
+ 风险控制系统	8.5	4.3	-4.2	27.3	87.3	+60.0

- 所有受测模型在遭遇越狱攻击时，都表现出不安全响应和拒答率显著上升，同时安全率下降。例如，Claude-3.7- Sonnet 在遭遇我们的安全越狱攻击时，安全响应比例下降了 33.8%。这表明当前最前沿的模型仍然面临越狱攻击带来的重大威胁。
- 与非推理模型相比，我们实验中的两个推理模型——DeepSeek-R1 和 o1(2024-12-17)——更加依赖风险控制系统进行安全检查，因此总体拒答率显著更高（分别为 79.8% 和 87.3%）。
- 开源模型（DeepSeek、Qwen）面临的越狱安全挑战比闭源模型更为严重，因为本地部署的模型缺乏风险控制系统。为解决安全问题，我们建议在其服务中使用开源模型的开发者采取类似的风险控制措施。

E. 更多分析

E.1. 与 DeepSeek-V3 的性能比较

由于 DeepSeek-R1 和 DeepSeek-V3 共享相同的基础架构，即 DeepSeek-V3-Base，因此自然会产生一个关键问题：通过应用不同的后训练技术，究竟增强了哪些具体维度？为回答这一问题，我们首先将 R1 系列模型与 DeepSeek-V3 和 DeepSeek-V3-Base 进行比较，结果汇总于表 12。值得注意的是，DeepSeek-R1 在竞赛编程和数学推理任务上取得了显著改进，在 LiveCodeBench 和 AIME 2024 等基准上的优异表现证明了这一点。这些推理能力的增强也转化为 Arena-Hard 评估套件中的更高得分。此外，DeepSeek-R1 展现出更强的长上下文理解能力，其在 FRAMES 上提升的准确率表明了这一点。

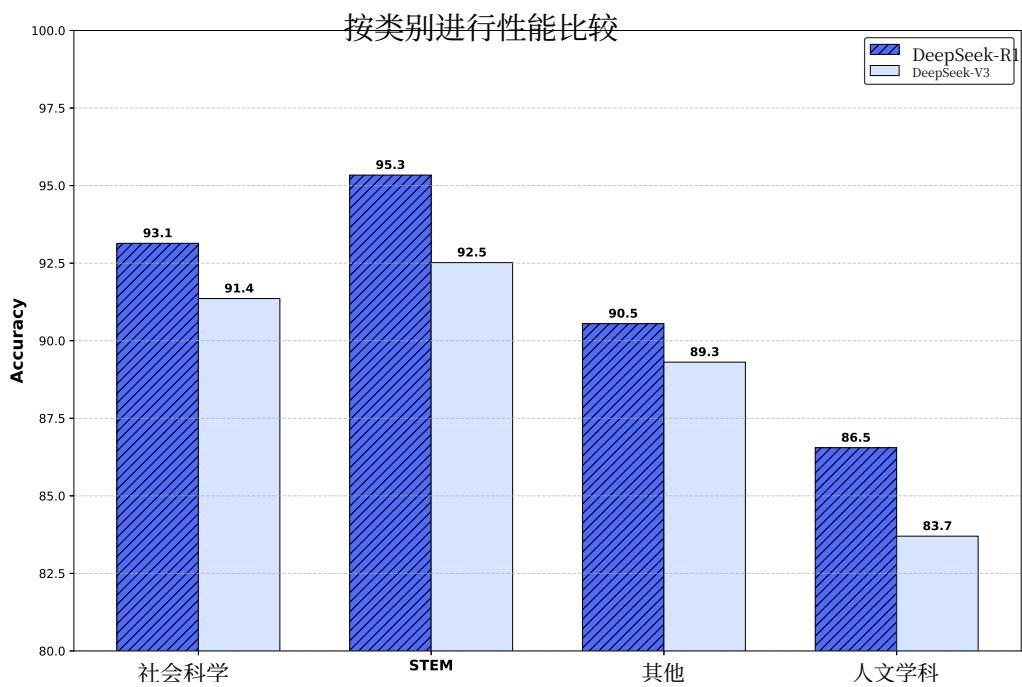


Figure 15 | DeepSeek-V3 和 DeepSeek-R1 在 MMLU 类别中的比较

s.

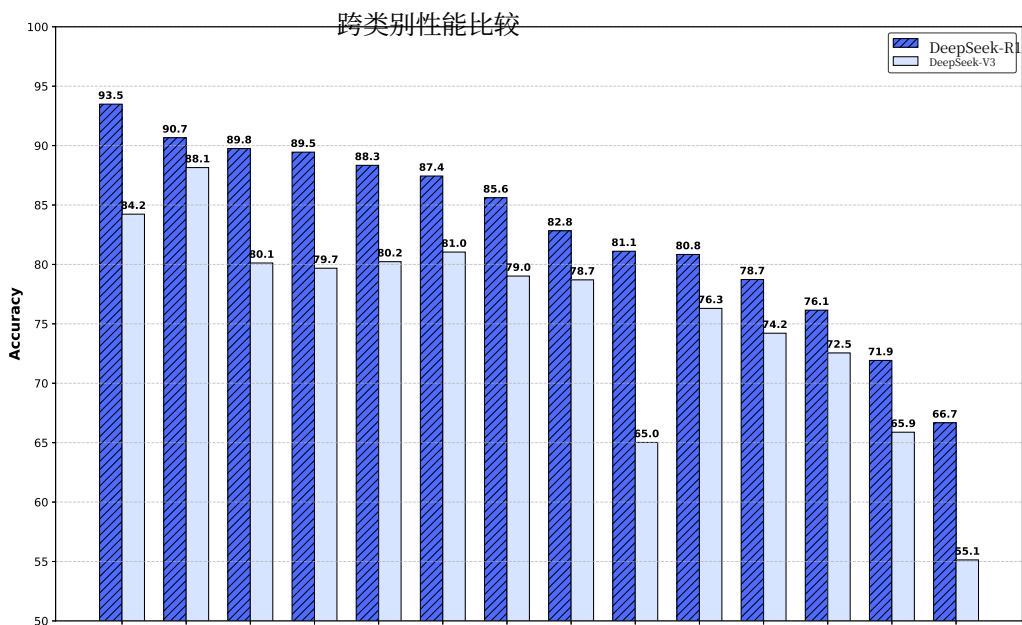


图 1 6 | DeepSeek-V3 和 DeepSeek-R1 在 MMLU-Pro 类别中的比较

.

表 12 | DeepSeek-V3 与 DeepSeek-R1 的比较分析。DeepSeek-V3 是在 DeepSeek-V3-Base 之上开发的非推理模型，而后者也作为 DeepSeek-R1 的基础模型。粗体数字表示性能具有统计显著性（t-检验， $p < 0.01$ ）。

基准 (指标)		V3-Base	V3	R1-Zero	R1
英语	MMLU (EM)	87.1	88.5	88.8	90.8
	MMLU-Redux (EM)	86.2	89.1	85.6	92.9
	MMLU-Pro (EM)	64.4	75.9	68.9	84.0
	DROP (3-shot F1)	89.0	91.6	89.1	92.2
	IF-Eval (Prompt Strict)	58.6	86.1	46.6	83.3
	GPQADiamond (Pass@1)	-	59.1	75.8	71.5
	SimpleQA (正确率)	20.1	24.9	30.3	30.1
	FRAMES (准确率)	-	73.3	82.3	82.5
	AlpacaEval2.0 (LC-winrate)	-	70.0	24.7	87.6
	ArenaHard (GPT-4-1106)	-	85.5	53.6	92.3
Code	LiveCodeBench (Pass@1-COT)	-	36.2	50.0	65.9
	Codeforces (百分位数)	-	58.7	80.4	96.3
	Codeforces (评分)	-	1134	1444	2029
	SWEVerified (已解决)	-	42.0	43.2	49.2
	Aider-Polyglot (准确率)	-	49.6	12.2	53.3
Math	AIME2024 (Pass@1)	-	39.2	77.9	79.8
	MATH-500 (Pass@1)	-	90.2	95.9	97.3
	CNMO2024 (Pass@1)	-	43.2	88.1	78.8
中文	CLUEWSC (EM)	82.7	90.9	93.1	92.8
	C-Eval (EM)	90.1	86.5	92.8	91.8
	C-SimpleQA (正确)	-	68.0	66.4	63.7

基准测试。相比之下，DeepSeek-V3 在指令遵循能力方面具有相对优势，这表明两个模型的优化重点不同。

为进一步阐明后训练最有利于哪些具体知识领域，我们在 MMLU 和 MMLU-Pro 的各个学科类别中，对模型性能进行了细粒度分析。这些类别在测试集构建期间预先定义，从而能够更系统地评估特定领域的改进。

如图 16 所示，MMLU-Pro 的所有领域都观察到性能提升，其中 STEM 相关类别（如数学和物理学）的增益尤为显著。同样，在 MMLU 上，从 DeepSeek-V3 到 DeepSeek-R1 的最大提升也出现在 STEM 领域。然而，不同于 MMLU-Pro，STEM 领域的增益较小，这表明后训练对两个基准测试的影响存在差异。

我们的假设是，与 MMLU-Pro 相比，MMLU 是一个相对更容易的挑战。在 MMLU 的 STEM 任务中，DeepSeek-V3 的后训练可能已经达到接近饱和的性能，因此 DeepSeek-R1 的进一步提升空间很小。令我们惊讶的是，社会科学和人文学科等非 STEM 任务在长 CoT 下得到了改进，这可能归因于对问题更好的理解。

表 13 | 最新数学竞赛中的表现。USAMO 指数 ($\text{AMC score} + 10 \times \text{AIME score}$) 超过 251.5 的参赛者具备参加 USAMO 的资格。

平均分	AMC 12 2024 AIME2025 USAMO 指数		
人类参与者	61.7	6.2/15	123.7
GPT-4o 0513	84.0	2.0/15	104.0
DeepSeek V3	98.3	3.3/15	131.3
OpenAI o1-1217	141.0	12.0/15	261.0
DeepSeek R1	143.7	11.3/15	256.7

E.2. 向现实世界竞赛的泛化

尽管我们严格努力消除数据污染，测试集问题的变体或相关问题的讨论仍可能存在于包含在预训练语料中的网站上。这引出一个重要问题：DeepSeek-R1 能否在其训练之后发布的测试集上达到可比性能？为研究这一点，我们在 AIME 2025 上评估模型，以了解其对未见数据的泛化能力。如表 13 所示，在 AIME 2025 (https://artofproblemsolving.com/wiki/index.php/2025_AIME_II_Problems) 中，DeepSeek-R1 达到 75% 的解题率 (Pass@1)，接近 o1 的 80% 表现。最值得注意的是，模型在 AMC 12 2024 (https://artofproblemsolving.com/wiki/index.php/2024_AMC_12B_Problems) 中取得 143.7/150 分——这一表现与其 AIME 结果结合后，得分超过参加 USAMO（美国数学奥林匹克竞赛 https://artofproblemsolving.com/wiki/index.php/AMC_historical_results?srsltid=AfmB0oqQ6pQic5NCan_NX1wYgr-aoHgJ33hsq7KSeKf-rUwY8TBaBo1) 的资格线。这一表现使 DeepSeek-R1 跻身全国顶尖高中生之列。

E.3. 按类别划分的数学能力分析

为全面评估 DeepSeek-R1 的数学推理能力，我们在多种定量推理问题类别上评估了其表现。测试集包含 366 道题目，来自 2024 年举办的 93 项数学竞赛 (https://artofproblemsolving.com/community/c3752401_2024_contests)，包括数学奥林匹克竞赛和团队选拔考试。如图 17 所示，DeepSeek-R1 的表现显著优于代表性非推理模型 GPT-4o 0513。DeepSeek-R1 在数论和代数方面展现出相对较强的能力，但在几何和组合数学方面仍有较大提升空间。

E.4. CoT 长度分析

自适应思维链长度：在训练期间，DeepSeek-R1 被允许进行长时间思考（即生成较长的思维链）后再得出最终解答。为最大限度地解决具有挑战性的推理任务，模型学会通过生成更多思考 token 来验证或纠正其推理步骤，或在初始尝试失败时回溯并探索替代方法，从而动态扩展计算量。问题的复杂度与所需思考 token 的数量直接相关：越困难的问题通常需要越多的计算。对于极其简单的问题，例如 $1 + 1 = ?$ ，模型往往使用更少的 token (< 100 tokens) 来回答。

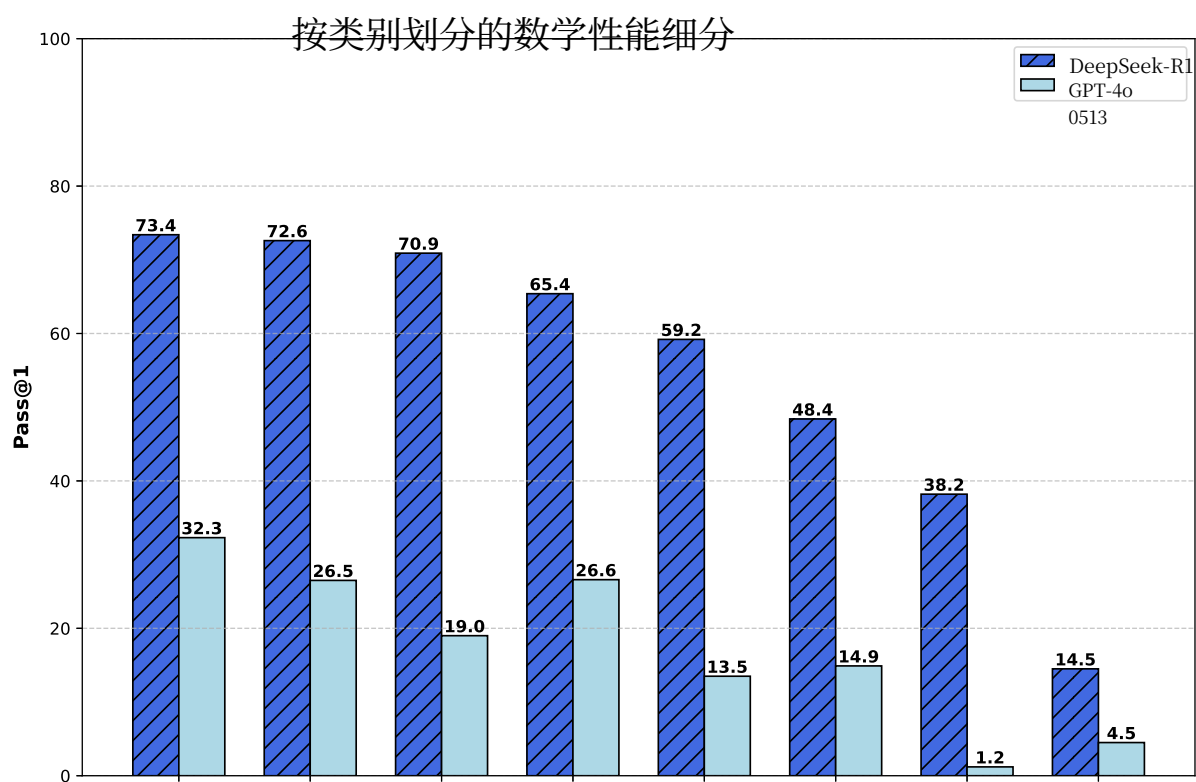


图 17 | 2024 年竞赛集合中不同类别定量推理问题的性能细分。

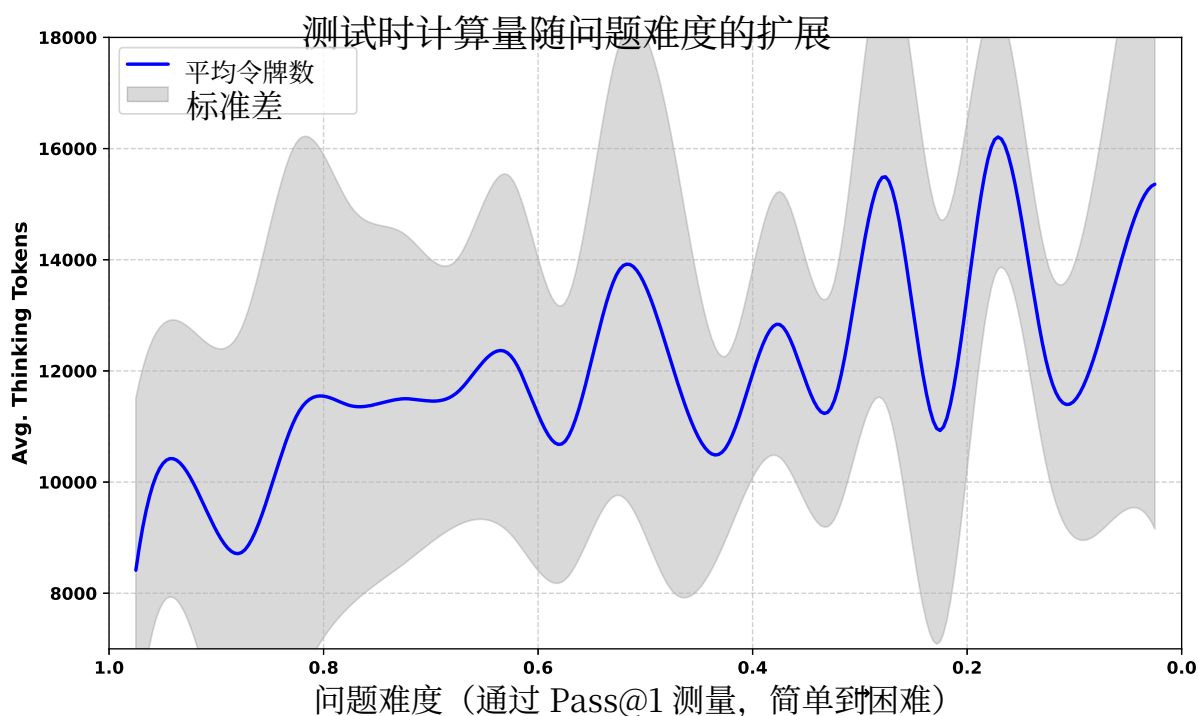


图 18 | 随着问题难度（通过 Pass@1 测量）增加，测试时计算量的扩展（通过生成的思考令牌数量来测量，以达到正确答案）。该图使用 SciPy 中的 UnivariateSpline 进行平滑处理，平滑因子为 5。

图 18 展示了 DeepSeek-R1 如何扩展测试时计算量，以解决 2024 年数学竞赛中的挑战性问题（与图 17 使用的题目集相同）。通过将测试时计算量扩展到每道题平均 8,793 个思考 token，DeepSeek-R1 达到了 61.8% 的解题率（Pass@1）。值得注意的是，模型会根据问题难度自适应地调整计算量：对于简单问题使用少于 7,000 个思考 token，而对最具挑战性的问题投入超过 18,000 个思考 token。这表明 DeepSeek-R1 会根据问题复杂度自适应地分配测试时计算量：问题越复杂，模型往往思考得越久。展望未来，我们假设如果在训练期间对 token 预算分配进行显式建模，那么测试时简单问题与困难问题之间的 token 使用差异可能会更加明显。

非推理模型的比较：DeepSeek-R1 等推理模型相较于 GPT-4o 0513 等非推理模型的一个关键优势，是能够沿推理维度有效扩展。非推理模型通常直接生成解决方案，不经过中间思考步骤，也很少展示自我反思、回溯或探索替代方法等高级问题解决技巧。在同一组数学问题上，GPT-4o 0513 的解题率仅为 24.7%，平均生成 711 个输出 token——比 DeepSeek-R1 少一个数量级。值得注意的是，非推理模型也可以通过多数投票等传统方法扩展测试时计算量，但即使控制生成 token 总数，这些方法也无法弥合与推理模型之间的性能差距。例如，对每道题采样 16 次进行多数投票，在 2024 年竞赛级数学题集合上仅使 GPT-4o 的解题率略有提升，尽管消耗的 token 总数多于 DeepSeek-R1。在 AIME 2024 上，对 64 个样本进行多数投票只将 GPT-4o 的解题率从 9.3% 提高到 13.4%——仍显著低于 DeepSeek-R1 的 79.8% 解题率或 o1 的 79.2% 解题率。这一持续存在的性能差距源于

一个根本限制：在多数投票中，各样本彼此独立生成，而不是相互建立在前一个样本之上。由于非推理模型缺乏回溯或自我纠正能力，扩大样本规模只会反复采样可能错误的最终解决方案，并不会提高在单次尝试中找到正确解的概率，因此这种方法的 token 效率很低。

缺点：然而，DeepSeek-R1 的扩展推理链有时仍不够全面，或会陷入错误的逻辑路径。独立采样多条推理链可提高发现正确解法的概率，DeepSeek-R1 在 AIME 2024 上的 Pass@64 得分为 90.0%，显著高于其 79.8% 的 Pass@1 得分，这一点证明了上述结论。因此，多数投票或蒙特卡洛树搜索（MCTS）等传统测试时扩展方法可以补充 DeepSeek-R1 的长推理；具体而言，多数投票将 DeepSeek-R1 的准确率从 79.8% 进一步提升至 86.7%。

E.5. 不同难度问题上各阶段的性能

表 14 | LiveCodeBench 数据集中不同难度级别问题上 DeepSeek-R1 各阶段的实验结果。

难度级别	DeepSeek-R1 Zero	DeepSeek-R1 Dev1	DeepSeek-R1 Dev2	DeepSeek-R1 Dev3	DeepSeek R1
Easy	98.07	99.52	100.00	100.00	100.00
中等	58.78	73.31	81.76	81.42	83.45
Hard	17.09	23.21	30.36	33.16	34.44

为进一步评估 DeepSeek-R1 各阶段在不同难度问题上的性能，我们展示了 DeepSeek-R1 各阶段在 LiveCodeBench 数据集上的实验结果，如表 14 所示。可以观察到，在每个阶段，简单问题通常都能得到正确解决，而主要改进来自中等和困难问题。这种细粒度分析表明，每个阶段都能显著提升复杂编码推理问题的性能。

F. DeepSeek-R1 蒸馏

LLM 能耗高，需要大量计算资源，包括高性能 GPU 和大量电力，用于训练和部署。这些资源需求为普及 AI 驱动技术的获取设置了重大障碍，尤其是在资源匮乏或边缘化的社区中。

为应对这一挑战，我们采用模型蒸馏方法，这是一种成熟的高效知识迁移技术，并已在先前工作中展现出强大的实证性能（Busbridge 等，2025；Hinton 等，2015）。具体而言，我们使用 DeepSeek-R1 生成的 800,000 个样本组成的精选数据集，对 Qwen（Qwen，2024b）和 LLaMA（AI@Meta，2024；Touvron 等，2023）等开源基础模型进行微调。数据集构建细节见附录 B.3.3。我们发现，从高质量教师输出中蒸馏得到的模型始终优于直接在人工生成数据上训练的模型，这与先前关于蒸馏有效性的研究结果一致（Busbridge 等，2025）。

对于蒸馏模型，我们仅应用 SFT，不包含 RL 阶段，尽管加入 RL 可能会大幅提升模型性能。我们在此的主要目标是

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群体。有关蒸馏训练的详细信息，请参见附录 B.4.3。

表 15 | DeepSeek-R1 蒸馏模型与其他可比模型在推理相关基准测试上的比较。加粗数字表示性能具有统计显著性（t-检验， $p < 0.01$ ）。

模型	AIME 2024		MATH	GPQA LiveCode Diamond Bench		CodeForces 评分
	pass@1	cons@64		pass@1	pass@1	
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
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

我们在 AIME、GPQA、Codeforces，以及 MATH-500 (Light-man et al., 2024) 和 LiveCodeBench (Jain et al., 2024) 上评估蒸馏模型。作为比较，我们使用两个成熟的 LLM 作为基线：GPT-4o 和 Claude-3.5-Sonnet。如表 15 所示，直接蒸馏 DeepSeek-R1 的输出，使蒸馏模型 DeepSeek-R1-Distill-Qwen-1.5B 在数学基准测试中超越了非推理基线。值得注意的是，仅有 15 亿参数的模型，其性能竟优于最出色的闭源模型，这一点非常 remarkable。此外，随着学生模型参数规模增大，模型性能持续提升。

我们的实验结果表明，小型模型可以通过蒸馏实现强劲性能。此外，如附录 F 所示，在较小模型架构上应用时，蒸馏方法的性能优于单独使用强化学习。这一发现对推动 AI 普及具有重要意义，因为降低计算需求能够带来更广泛的社会效益。

F.1. 蒸馏与强化学习的比较

Table 16 | 蒸馏模型与 RL 模型在推理相关基准测试上的比较。

模型	AIME 2024		MATH	GPQA	LiveCode
	pass@1	cons@64		Diamond pass@1	基准 pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
Qwen2.5-32B-Zero	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

在 F 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，仍有一个问题悬而未决：模型能否达到可比的

表 17 | AIME 2024 和 AIME 2025 上不同模型的性能。

平均分	AIME 2024 AIME 2025	
GPT-4o-0513	9.3%	-
Qwen2-Math-7B-Instruct	7.9%	4.6%
Qwen2-Math-7B-Zero	22.3%	18.1%

性能 如果不进行蒸馏，仅通过论文中讨论的大规模 RL 训练，能否实现性能提升？

为回答这一问题，我们使用数学、代码和 STEM 数据对 Qwen2.5-32B-Base 进行大规模 RL 训练，训练超过 10K 步，得到 Qwen2.5-32B-Zero，详见 B.4.1。表 16 所示实验结果表明，经过大规模 RL 训练后，32B 基础模型达到了与 QwQ-32B-Preview 相当的性能。然而，从 DeepSeek-R1 蒸馏得到的 DeepSeek-R1-Distill-Qwen-32B 在所有基准测试中的表现都显著优于 Qwen2.5-32B-Zero。

因此，我们可以得出两个结论：第一，将更强大的模型蒸馏到较小模型中能够取得出色结果；而依赖本文所述大规模 RL 的较小模型需要巨大的计算能力，甚至可能无法达到蒸馏模型的性能。第二，尽管蒸馏策略兼具经济性和有效性，但要突破人类智能的边界，可能仍需要更强大的基础模型和更大规模的强化学习。

除了基于 Qwen-2.5-32B 的实验外，在首个推理模型 OpenAI-o1（2024 年 9 月）发布之前，我们还对 Qwen2-Math-7B（2024 年 8 月发布）进行了实验，以确保基础模型未接触任何推理轨迹数据。我们使用约 10,000 个策略梯度更新步骤训练了 Qwen2-Math-7B-Zero。如表 17 所示，Qwen2-Math-7B-Zero 的表现显著优于 Qwen2-Math-7B-Instruct 和 GPT-4o 等非推理模型。这些结果进一步表明，通过大规模强化学习，模型能够自主发展出先进的推理策略。

G. 讨论

G.1. 主要发现

我们总结了主要发现，这些发现可能有助于社区更好地复现我们的工作。

基础检查点的重要性：在开发初期，我们尝试使用较小规模的模型，具体而言是 7B 稠密模型和 16B 混合专家（MoE）模型，作为 RL 训练的基础架构。然而，在我们用作主要验证集的 AIME 基准上进行评估时，这些配置始终未能带来有意义的改进。我们观察到，随着响应长度增加，这些较小模型表现出重复倾向，无法有效利用长思维链（CoT）来提升推理准确率。

为解决这些局限性，我们转向了更大规模的模型，包括 32B 稠密模型（Qwen, 2024b）、230B MoE 模型（DeepSeek-AI, 2024a）和 671B MoE 模型（DeepSeek-AI, 2024b）。借助这些能力更强的架构，我们最终观察到了显著的

纯 RL 训练所带来的性能提升。这些发现表明，基于基础模型的强化学习效果高度依赖底层模型容量。因此，我们建议该领域的未来研究在旨在验证从零开始进行 RL 的有效性时，优先使用规模足够大且表达能力强的模型。

验证器的重要性：DeepSeek-R1-Zero 的有效性高度依赖于训练期间所使用奖励信号的可靠性和保真度。迄今为止，我们的研究表明，两种方法——基于规则的奖励模型（RMs），以及使用 LLM 根据预定义的真实答案评估答案正确性——是缓解奖励劫持问题的稳健机制。基于 LLM 的评估框架对于具有明确、简洁答案的任务尤其有效，例如单句或短语级响应。然而，该方法对更复杂任务的泛化能力有限，包括开放式生成和长篇写作；在这些任务中，正确性的概念本质上更加主观且微妙。

迭代式流程：我们提出了一个包含 SFT 和 RL 阶段的多阶段训练流程。RL 组件使模型能够探索并发现最优的推理轨迹，以实现仅凭人工标注推理轨迹无法充分实现的任务能力。特别是，如果没有 RL 阶段，复杂思维链（CoT）提示所需的长链推理模式将基本无法得到探索。相反，SFT 阶段对于难以定义或建模可靠奖励信号的任务发挥关键作用，例如开放式问答和创意写作。因此，RL 和 SFT 都是我们训练流程中不可或缺的组成部分。完全依赖 RL 可能导致奖励劫持，以及在定义不当的任务上产生次优行为；而仅依赖 SFT 则可能阻止模型通过探索来优化其推理能力。

G.2. 未成功的尝试

在开发 DeepSeek-R1 的早期阶段，我们也经历了失败和挫折。我们在此分享失败经验以提供参考，但这并不意味着这些方法无法用于开发有效的推理模型。

过程奖励模型（PRM） PRM 是引导模型采用更优方法解决推理任务的一种合理方法（Lightman et al., 2024; Uesato et al., 2022; Wang et al., 2023a）。然而，在实践中，PRM 存在三个主要局限，可能阻碍其最终成功。第一，在一般推理中明确定义细粒度步骤颇具挑战。第二，判断当前中间步骤是否正确是一项困难任务。使用模型进行自动标注可能无法取得令人满意的结果，而人工标注不利于扩展。第三，一旦引入基于模型的 PRM，就不可避免地会导致奖励劫持（Gao et al., 2022）；重新训练奖励模型需要额外的训练资源，也会使整个训练流程复杂化。总之，尽管 PRM 在对模型生成的 top-N 响应进行重新排序或辅助引导式搜索方面表现良好（Snell et al., 2024），但在我们的实验中，与其在大规模强化学习过程中引入的额外计算开销相比，其优势是有限的。

蒙特卡洛树搜索（MCTS） 受 AlphaGo（Silver et al., 2017b）和 AlphaZero（Silver et al., 2017a）的启发，我们探索使用蒙特卡洛树搜索（MCTS）来增强测试时计算的可扩展性。该方法将答案拆分成更小的部分，使模型能够系统地探索解空间。为此，我们提示模型

生成与搜索所需特定推理步骤相对应的多个标签。在训练方面，我们首先使用收集到的提示，通过预训练的价值模型引导 MCTS 来寻找答案。随后，我们使用所得的问答对同时训练 actor 模型和价值模型，迭代改进这一过程。

然而，这种方法在扩大训练规模时会遇到若干挑战。首先，与搜索空间相对明确的国际象棋不同，token 生成呈现出指数级更大的搜索空间。为此，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，价值模型直接影响生成质量，因为它指导搜索过程中的每一步。训练细粒度的价值模型本身就很困难，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 token 生成的复杂性，这一原则难以在我们的设置中复现。

总之，尽管 MCTS 与预训练价值模型结合时能够提升推理阶段的性能，但通过自搜索迭代提升模型性能仍然是一项重大挑战。

H. 相关工作

H.1. 思维链推理

思维链（CoT）推理（Wei et al., 2022b）通过提示 LLM 在生成最终答案前先生成中间推理步骤，彻底改变了 LLM 处理复杂推理任务的方式。该方法显著提升了模型在涉及算术、常识和符号推理的基准测试中的表现。后续研究探索了其适用范围：Suzgun et al. (2023) 证明 CoT 的有效性会随模型规模扩大而提升，而 Kojima et al. (2022) 则通过简单地指示模型“逐步思考”，将其扩展到零样本设置。

在 CoT 框架的基础上，人们提出了许多“提示工程”技术来提升模型性能。Wang 等（2023b）提出了自洽性方法，通过汇总多条推理路径的答案来提高稳健性和准确性。Zhou 等（2023a）提出了由易到难提示，将复杂问题分解为按顺序排列的子问题并逐步求解。Yao 等（2023a）提出了思维树，使模型能够同时探索多个推理分支，并通过前瞻或回溯进行审慎决策。总体而言，这些方法利用人类先验知识和更结构化的推理框架，增强了 LLM 的推理能力。

H.2. 推理时计算量扩展

由于无监督预训练扩展可能受可用人类数据量的限制（Kaplan 等，2020；Muennighoff 等，2023），在推理期间扩展计算量变得更加关键（Snell 等，2025）。广义而言，我们将通过增加推理计算量来提升模型性能的方法定义为推理时计算量扩展。

一种直接的方法是用计算换取性能：生成多个不同的推理链并选择最佳答案。可以使用独立的重排序器（Brown et al., 2024; Cobbe et al., 2021）、基于过程的奖励模型（Lightman et al., 2024; Uesato et al., 2022），或直接选择最常见的答案来确定最优答案。

(Wang et al., 2023b)。蒙特卡洛树搜索和束搜索等搜索方法也能更有效地引导对解空间的探索 (Feng et al., 2024; Hao et al., 2023; Trinh et al., 2024; Xin et al., 2024)。除并行生成外，自我纠正技术会提示或训练模型迭代批评和改进其输出 (Kumar et al., 2024; Madaan et al., 2023; Welleck et al., 2023)，通常还会结合外部反馈以增强可靠性 (Gou et al., 2024a; Yao et al., 2023b)。此外，一些方法通过在测试期间整合工具使用来提升性能，这对知识密集型任务 (Nakano et al., 2021) 和计算密集型任务 (Chen et al., 2025; Gou et al., 2024b; Schick et al., 2023) 尤其有效。测试时训练 (TTT) 会在推理期间进一步更新模型，以提升性能 (Akyürek et al., 2024; Sun et al., 2020)。还有各种其他推理时扩展方法，它们或隐式地 (Geiping et al., 2025) 或显式地 (Zelikman et al., 2024) 为每个 token 分配更多计算量。

相比之下，我们的工作表明，LLM 可以通过增加额外的 RL 计算量和测试时计算量（即更多 token）实现可扩展的改进。我们将测试时扩展的优势整合到一个更广泛的框架中，该框架利用强化学习来激励增强的上下文内搜索能力。

H.3. 用于增强推理的强化学习

强化学习在使 LLM 与人类偏好保持一致方面发挥着关键作用 (Bai et al., 2022; Ouyang et al., 2022)。尽管其十分重要，但很少有研究关注利用 RL 增强推理能力。传统 RL 流程首先在高质量的人类示范上进行 SFT，这为模型提供了良好的初始化并防止模式坍塌。随后，根据人类偏好训练奖励模型，并使用 PPO (Schulman et al., 2017) 或 DPO (Rafailov et al., 2023) 等方法对语言模型进行优化。尽管这种方法在对齐方面表现良好，但它可能将模型限制在模仿人类推理模式上，从而阻碍新颖问题解决策略的发现。

STaR 等方法通过在模型自行生成且能够导向正确最终答案的思维链上进行微调，迭代提升性能 (Singh et al., 2024; Yuan et al., 2023; Zelikman et al., 2022)。近期研究还探讨了使用基于过程的奖励，这类奖励同时强调最终答案的正确性和推理过程的合理性 (Lightman et al., 2024; Shao et al., 2024; Wang et al., 2023a)。不同于这些方法，我们的工作直接对基础语言模型应用基于结果的 RL，而不经初始 SFT 阶段。这一设计选择鼓励创新且不受约束的推理策略涌现，使模型能够发展出多样化的解决方案，而不仅仅是模仿人类示例。我们的方法也启发了后续研究的进一步探索 (Face, 2025; Liu et al., 2025; Pan et al., 2025)。

I. 开放权重、代码和数据

为推动开源社区和产业生态的发展，我们已在 HuggingFace 上公开 DeepSeek-R1 和 DeepSeek-R1-Zero 的模型权重。此外，我们还发布了 DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Qwen-14B、DeepSeek-R1-Distill-Qwen-32B、DeepSeek-R1-Distill-Llama-8B、DeepSeek-R1-Distill-Llama-70B。

此外，我们已在 GitHub 上发布基础模型推理代码 (<https://github.com/deepseek-ai/DeepSeek-V3>)，并提供了详细的使用指南 (<https://github.com/deepseek-ai/DeepSeek-R1>)。

以下是运行推理代码以与 DeepSeek-R1 交互的示例：

```
# Download the model weights from Hugging Face
huggingface-cli download deepseek-ai/DeepSeek-R1 --local-dir
/path/to/DeepSeek-R1

# Clone DeepSeek-V3 GitHub repository
git clone https://github.com/deepseek-ai/DeepSeek-V3.git

# Install necessary dependencies
cd DeepSeek-R1/inference
pip install -r requirements.txt

# Convert Hugging Face model weights to a specific format (for running
the model on 16 H800 GPUs)
python convert.py --hf-ckpt-path /path/to/DeepSeek-R1 --save-path
/path/to/DeepSeek-R1-Demo --n-experts 256 --model-parallel 16

# Run the model and interact with it
torchrun --nnodes 2 --nproc-per-node 8 --node-rank $RANK --master-addr
$MASTER_ADDR generate.py --ckpt-path /path/to/DeepSeek-R1-Demo --config
configs/config_671B.json --interactive --temperature 0.7
--max-new-tokens 8192
```

我们还在 xxx 向公众发布了 SFT 和 RL 数据。在审阅过程中，我们将数据作为附件上传。

J. 评估提示与设置

表 18 | MMLU 评估模型在涵盖 57 项任务中的事实性和概念性理解，这些任务涉及 STEM（科学、技术、工程、数学）、人文学科、社会科学和专业领域（例如法律、医学）。该基准通常用于评估模型执行常识推理的能力，以及在多样化学科和任务范围内的多任务熟练度。下面是一个 MMLU 示例。

提示 回答以下选择题。你的回答最后一行应采用以下格式：' Answer: \$LETTER' （不含引号），其中 LETTER 是 ABCD 中的一个字母。回答前请逐步思考。 哪种工具技术与尼安德特人有关？ A. 奥里尼亚克文化 B. 阿舍利文化 C. 莫斯特文化 D. b 和 c 两者	
评估 解析响应的最后一行，以判断所选选项是否等于真实答案。	

表 19 | MMLU-Redux 是涵盖全部 57 个 MMLU 学科的 5,700 道人工重新标注问题的子集。MMLU-Redux 通过减少 MMLU 中的噪声、歧义和潜在偏差，着重提升基准的质量、清晰度和稳健性，同时可能调整任务的范围或难度，以更好地符合现代评估需求。下面是一个 MMLU-Redux 示例。

提示## 问题：桑拿使用有时被称为“桑拿浴”，其特征是短时间被动暴露于极端高温下……事实上，桑拿使用已被提议作为无法因慢性疾病或身体限制而进行体育活动的人群的运动替代方式。[13]根据文章，以下哪项不是桑拿使用的益处？## 选项：- (A) 降低心脏病发作风险。- (B) 增加每搏输出量。- (C) 改善心理健康。- (D) 降低勃起功能障碍发生率。## 指令请先进行推理，然后选择正确选项来回答此问题。请按以下 json 格式呈现推理和解答。请在 ‘answer’ 字段中仅填写选项字母，例如 “answer": "C" ‘。{ "reasoning": "____", "answer": " "____} **评估**解析响应中的 json 输出，以判断答案是否等于真实答案。

表 20 | LiveCodeBench 旨在评估模型在算法竞赛任务上的性能，它从三个竞赛平台（LeetCode、AtCoder 和 CodeForces）的竞赛中随时间收集新问题。

提示

问题：有一叠 N 张卡片，从顶部起第 i 张卡片上写有整数 A_i 。你从牌堆底部取出 K 张卡片并放到牌堆顶部，同时保持它们的顺序。请输出操作后从顶部到底部卡片上写的整数。输入 输入按照以下格式从标准输入给出： N K $A_1 A_2 \dots A_N$ 输出 设 B_i 为操作后从顶部起第 i 张卡片上写的整数。按此顺序输出 $B_1, B_2, \dots, B_N$ ，各整数之间以空格分隔。约束 $-1 \leq K < N \leq 100$
 $-1 \leq A_i \leq 100$ 所有输入值均为整数。示例输入 1 5 3 1 2 3 4 5 示例输出 1 3 4 5 1 2 最初，从顶部到底部卡片上的整数为 1,2,3,4,5。将底部三张卡片取出并放到顶部后，从顶部到底部变为 3,4,5,1,2。示例输入 2 6 2 1 2 1 2 1 2 示例输出 2 1 2 1 2 1 2 卡片上的整数不一定互不相同。请编写 Python 代码解决上述问题。代码必须从 `stdin` 读取输入，并将结果输出到 `stdout`。**Evaluation** 提取回答中由 `"python"` 包裹的代码，以判断答案是否通过测试用例。

表 21 | 与 MMLU 相比，MMLU-Pro 包含经过精选的任务子集，但难度显著提高。MMLU-Pro 中的问题需要更深入的推理、多步问题求解以及先进的领域专业知识。例如，STEM 任务可能涉及复杂的数学推导或细致的科学概念，而人文学科任务可能要求复杂的语境分析。

提示

以下是关于商业的选择题（附答案）。请逐步思考并然后在末尾按“答案是 (X)”的格式输出答案。

...

问题：典型的广告监管机构建议，例如，广告不得：鼓励鼓励 ____，造成不必要的____或 ____，并且不得造成____冒犯。

选项：A. 安全实践、恐惧、嫉妒、琐碎

B. 不安全实践、痛苦、喜悦、琐碎

C. 安全实践、欲望、嫉妒、琐碎

D. 安全实践、痛苦、恐惧、琐碎

E. 不安全实践、欲望、嫉妒、严重

F. 安全实践、痛苦、嫉妒、严重

G. 安全实践、需求、恐惧、严重

H. 不安全实践、需求、恐惧、轻微

I. 不安全实践、痛苦、恐惧、严重

回答：让我们逐步思考。

评估

解析响应中“Answer: ”之后的首个大写字母，以判断答案是否等于基准真相。

表 22 | DROP 评估模型理解长篇文本段落并提取相关信息的能力。不同于侧重事实回忆的简单问答基准测试，DROP 要求模型处理并解释富含上下文的段落。

提示

你将被要求阅读一段文章并回答问题。下面提供了一些文章和问答示例。# 示例——文章：为避免连续遭遇分区失利，爱国者队前往迈阿密，在 Dolphin Stadium 对阵 6-4 的海豚队…… Cassel 的 415 码传球使他成为爱国者队历史上第二位在两场或更多比赛中传球至少 400 码的四分卫；Drew Bledsoe 在其爱国者队生涯中有四场 400+ 码传球比赛。问题：海豚队输掉了多少分？答案：20。——文章：第 2 周，海鹰队迎战分区对手旧金山 49 人队。赛季开始前，NFL 分析师将这场对决评为即将到来的最佳对抗以及十年最佳对抗……西雅图队此时 2-0，并且主场仍保持不败。问题：Hauschka 命中了多少个至少 30 码的射门？答案：2。——文章：位于佛罗里达州坦帕的 Raymond James Stadium，电视时间：CBS 东部时间下午 1:00。9 月 10 日，乌鸦队在常规赛揭幕战客场对阵坦帕湾海盗队……获胜后，乌鸦队战绩为 1-0，对 NFC 对手也是 1-0。问题：Lewis 跑了多少码？答案：4。# 你的任务——文章：闪电队（1-0）在赛季揭幕战中以 22-14 击败奥克兰突袭者队；Nate Kaeding 命中五次射门，突袭者队有三次弃踢失误。突袭者队入选职业碗的长开球手 Jon Condo 在第二节头部受伤，由自高中以来就没有开球过的线卫 Travis Goethel 替代。Goethel 两次将球滚给弃踢手 Shane Lechler，每次都让闪电队在突袭者队半场获得球权；Lechler 的另一次弃踢又被 Dante Rosario 封堵。闪电队唯一一次达阵发生在第二节：一次 13 个进攻回合、推进 90 码的进攻最终由 Philip Rivers 向外接手 Malcom Floyd 传出 6 码达阵。闪电队在红区五次进攻中有四次未能得分。圣迭戈半场以 10-6 领先，突袭者队直到比赛还剩 54 秒才取得达阵。未被选中的新秀 Mike Harris 首次在 NFL 首发，顶替受伤的左截锋 Jared Gaither。圣迭戈让 Rivers 多传短球来保护 Harris；Rivers 的 24 次完成传球中有 16 次传给跑卫和近端锋，他传球 231 码，仅被擒杀一次。他没有被抄截，此前他在 2011 年曾被抄截 20 次。这是闪电队在奥克兰过去九场比赛中的第八场胜利，使 Norv Turner 在闪电队赛季揭幕战中的战绩提升至 4-2。跑卫 Ryan Mathews 和接球手 Vincent Brown 因伤缺席比赛。问题：Rivers 传了多少码？答案：请逐步思考，然后在回答末尾写出格式为“Answer: \$ANSWER”的一行。评估解析回答中“Answer:”之后的首个大写字母，以判断答案是否等于标准答案。

表 23 | 指令遵循评估 (IFEval) 是一个用于评估模型遵守提示中嵌入的明确、可验证指令能力的基准。它针对大型语言模型 (LLM) 的一项核心能力：生成满足用户指定的多个、定义清晰的约束条件的输出。

提示

请将下面的文本总结为 XML 格式。确保总结少于 4 个句子。

量子纠缠是这样一种现象：当一组粒子以某种方式生成、相互作用或共享空间邻近关系时，该组中每个粒子的量子态都无法独立于其他粒子的状态进行描述，即使这些粒子相隔很远。量子纠缠是经典物理学与量子物理学差异的核心所在：纠缠是量子力学的主要特征，而经典力学中不存在这一特征。

对纠缠粒子的位置、动量、自旋和偏振等物理属性进行测量时，在某些情况下可以发现它们完全相关。例如，如果生成一对纠缠粒子，使其总自旋已知为零，并且发现其中一个粒子在第一条轴上顺时针自旋，那么在同一条轴上测量另一个粒子时，会发现其自旋为逆时针。然而，这种行为会产生看似悖论的效应：对粒子属性的任何测量都会导致该粒子的波函数表观且不可逆地坍缩，并改变原始量子态。对于纠缠粒子，此类测量会影响整个纠缠系统。

此类现象是 Albert Einstein、Boris Podolsky 和 Nathan Rosen 于 1935 年发表的一篇论文，以及 Erwin Schrödinger 不久后发表的几篇论文的主题；这些论文描述了后来被称为 EPR 悖论的现象。Einstein 等人认为这种行为不可能存在，因为它违反了因果关系的局域实在论观点（Einstein 称其为“幽灵般的远距作用”），并据此认为，当时公认的量子力学表述必然是不完整的。

评估

调用官方函数检查答案是否符合指令。

表 24 | FRAMES（事实性、检索与推理测量集）是一个用于评估检索增强生成（RAG）系统核心组件的综合基准。我们的评估采用该基准官方的“Oracle Prompt”配置。在此设置下，每个测试提示都包含问题以及所有真实答案对应的 Wikipedia 文章，从而无需外部检索组件（例如 BM25）。这一设置使我们能够专门衡量模型基于所提供来源进行推理并综合信息，以生成正确且可验证事实的能力。

提示 以下是相关的维基百科文章： 网址: https://en.wikipedia.org/wiki/President_of_the_United_States 网址内容: 美国总统（POTUS）是国家元首和政府首脑 美利坚合众国。总统领导联邦政府的行政部门 并且是美国武装部队的总司令。 ... 根据所有信息回答查询。 查询: 如果我未来妻子的名字与美国第15任第一夫人的母亲的名字相同 并且她的姓氏与第二位遇刺总统的母亲的娘家姓相同，那么 我未来妻子的名字是什么？
评估任务=== 我需要你的帮助，根据标准答案评估一个 LLM 提供的答案。你的任务是确定 LLM 的回答中是否包含标准答案。请分析所提供的信息并作出判断。 说明=== 1. 仔细比较“预测答案”和“标准答案”。 2. 考虑答案的实质内容——寻找等价信息或正确答案。除非措辞本身对含义至关重要，否则不要拘泥于确切措辞。 3. 你的最终判断应基于“标准答案”中的含义和关键信息是否出现在“预测答案：”中 “标准答案”中的含义和关键信息是否出现在“预测答案：”中 ===输入数据=== - 问题: 如果我未来妻子的名字与美国第 15 任第一夫人的母亲名字相同，且她的姓氏与第二位遇刺总统的母亲的娘家姓相同，我未来妻子的名字是什么？ 娘家姓相同，我未来妻子的名字是什么？ 娘家姓相同，我未来妻子的名字是什么？ - 预测答案: ... - 真实答案: Jane Ballou ===输出格式=== 请按以下格式提供最终评估： 解释: xxx 决策: "TRUE" 或 "FALSE" 请继续进行评估。

表 25 | Arena-Hard 是一个专门用于评估 LLM 能力的开放式评测基准。它向模型提供来自 Chatbot Arena 的具有挑战性、新颖且多样化的提示；Chatbot Arena 是一个持续演进、由众包构建的平台。该基准侧重衡量模型在开放式任务中的性能，尤其关注编码和数学相关提示。鉴于开放式任务具有内在的主观性，可能存在多个有效响应，因此该基准需要使用评估模型来有效近似人类判断。更高的评估分数表明，在现实场景中，该模型更有可能受到人类用户青睐。

提示

要编写一条 SQL 查询，从数据库中选取前 10 行，并基于名为“code”的字段与 3 个不同的表进行连接，你需要知道这些表的名称以及想要选取的具体字段。假设这些表名为“table1”、“table2”和“table3”，并且你想从每个表中选取所有字段，则查询如下：

```
sql SELECT * FROM ( SELECT * FROM table1 LIMIT 10 ) AS t1 JOIN table2 ON t1.code = table2.co de JOIN table3 ON t1.code = table3.code; “ ‘
该查询首先从 "table1" 中选取前 10 行，然后根据 "code" 字段连接 "table2" 和 "table3"。如果你想从每个表中选择特定字段，可以将星号 (*) 替换为以逗号分隔的字段名称。
```

评估

请充当公正的评判者，评估下面显示的用户提示下两位 AI 助手所提供回答的质量。你将获得助手 A 的回答和助手 B 的回答。你的任务是评估哪位助手的回答更好。

开始评估时，请先为提示词生成你自己的答案。在评判任何答案之前，你必须先给出自己的答案。

When在评估助手的答案时，将两个助手的答案都与你的答案进行比较。你必须识别并纠正任何错误或不准确信息。然后判断助手的答案是否有帮助、相关且简洁。有帮助意味着答案正确回应了提示词或遵循了指令。请注意，当用户提示词存在歧义或有不只一种解释时，与其基于假设提供答案，不如向用户请求澄清或更多信息，这样更有帮助且更恰当。相关意味着响应的所有部分都与所询问的内容紧密相关或適切。简洁意味着响应清晰，不冗长或过度。

然后，在需要时考虑助手回答的创造性和新颖性。最后，确定助手回答中缺失的、在回应用户提示时有益于补充的重要信息。

之后 在提供解释时，你必须只输出以下选项中的一个作为最终答案 al verdi 并附带标签：

- 1. 助手 A 明显更好： [[A>>B]]
 - 2. 助手 A 略好： [[A >B]]
 - 3. 平局，基本相同： [[A=B]]
 - 4. 助手 B 略好： [[B>A]]
 - 5. 助手 B 明显更好： [[B>>A]]
- 示例输出："我的最终判断是平局： [[A=B]]"。

表 26 | AlpacaEval 2.0 是一个开放式评估数据集，其性质与 ArenaHard 类似，并利用 LLM 评估模型在主观任务上的表现。然而，与 ArenaHard 不同，AlpacaEval 2.0 中的提示通常较不具挑战性，只有一小部分提示需要被评估模型部署推理能力。

提示

有哪些著名演员是在百老汇开始其职业生涯的？

评估

< |im_start| >系统

你是一名高效的助手，根据其对给定指令的响应质量评估并选择最佳的大型语言模型（LLMs）基于其响应质量。此过程将用于创建一个反映最准确且最受人类偏好答案的排行榜。

< |im_end| >

< |im_start| >用户

我需要一个涵盖各种大型语言模型的排行榜。我会向你提供给这些模型的提示词以及它们相应的输出。你的任务是评估这些回答，并选择从人类角度来看产生最佳输出的模型。

指令

```
{
  "instruction": ""{instruction}""
}
```

模型输出

以下是模型生成的无序输出。每个输出都与特定模型相关联，由唯一的模型标识符标识。

```
{
  {
    "model_identifier": "m" ,
    "output": ""{output_1}""
  },
  {
    "model_identifier": "M" ,
    "output": ""{output_2}""
  }
}
```

任务

根据输出的质量和相关性评估模型，并选择最符合要求的模型

生成了最佳输出。请通过提供最佳模型的标识符来回答。我们将

使用你的输出作为最佳模型的名称，因此请确保你的输出只包含以下模型标识符中的一个模型标识符，且不包含其他任何内容（无引号、无空格、无换行符，……）： m 或 M。

最佳模型标识符

< |im_end| >

表 27 | CLUEWSC（中文语言理解评估基准——Winograd 模式挑战）是 CLUE 基准套件中的一项专门任务，旨在评估模型在中文环境下的常识推理和语境理解能力。

提示

请参考示例的格式,完成最后的测试题。

下面是一些示例:他伯父还有许多女弟子,大半是富商财主的外室;这些财翁白天忙着赚钱,怕小公馆里的情妇长日无聊,要不安分,常常叫她们学点玩艺儿消遣。

上面的句子中的"她们"指的是

情妇

耶律克定说到雁北义军时,提起韦大哥,就连声说不可挡、不可挡,似有谈虎色变之味。后来又听说粘罕在云中,特派人厚币卑词,要与‘韦义士修好’。韦大哥斩钉截铁地回绝了,大义凛然,端的是条好汉。如今张孝纯也想结识他,几次三番派儿子张浹上门来厮缠,定要俺引他上雁门山去见韦大哥。

上面的句子中的"他"指的是

张浹

“你何必把这事放在心上?何况你的还不过是手稿,并没有发表出来。”龙点睛越发坦率:“如果发表出来了,那倒也就算了。不过既然没发表出来,我何必还让它飘在外头呢?你给我找一找吧,我要收回。”

上面的句子中的"它"指的是

手稿

这个身材高大,曾经被孙光平拿着菜刀追赶得到处乱窜的年轻人,那天早晨穿上了全新的卡其布中山服,像一个城里来的干部似的脸色红润,准备过河去迎接他的新娘。

上面的句子中的"他"指的是

年轻人

负责接待我们的是两位漂亮的朝鲜女导游,身材高挑,露出比例完美的小腿。一个姓韩,一个姓金,自称小韩和小金。她们的中文说得毫无口音,言谈举止也相当亲切。

上面的句子中的"她们"指的是

两位漂亮的朝鲜女导游

下面是测试题,请在思考结束后(</think>后)用一句话输出答案,不要额外的解释。

崩龙珍夫妻康健和美;鞠琴十年前丧偶,两年前重结良缘,现在的老伴是一位以前未曾有过婚史的高级工程师;崩龙珍和鞠琴都尽量避免谈及自己的爱人,也尽量回避提及蒋盈波的亡夫屈晋勇——尽管她们对他都很熟悉;当然也绝不会愚蠢地提出蒋盈波今后是一个人过到底还是再找个老伴的问题来加以讨论,那无论如何还为时过早。

上面的句子中的"他"指的是

评估

解析响应中的最后一行，以判断答案是否等于标准答案。

表 28 | C-EVAL 评估模型在 52 个不同学科领域中的知识广度和深度，涵盖人文科学、社会科学、STEM（科学、技术、工程和数学）以及其他专业领域（例如医学、法律）。C-Eval 中的所有问题均为中文。

提示

以下是中国关于逻辑学考试的单项选择题,请选出其中的正确答案。

1991年6月15日,菲律宾吕宋岛上的皮纳图博火山突然大喷发,2000万吨二氧化硫气体冲入平流层,形成的霾像毯子一样盖在地球上空,把部分要照射到地球的阳光反射回太空几年之后,气象学家发现这层霾使得当时地球表面的温度累计下降了0.5°C,而皮纳图博火山喷发前的一个世纪,因人类活动而造成的温室效应已经使地球表面温度升高1°C。某位持“人工气候改造论”的科学家据此认为,可以用火箭弹等方式将二氧化硫充入大气层,阻挡部分阳光,达到地球表面降温的目的。以下哪项如果为真,最能对该科学家的提议构成质疑?——

- A. 如果利用火箭弹将二氧化硫充入大气层,会导致航空乘客呼吸不适。 B. 火山喷发形成的降温效应只是暂时的,经过一段时间温度将再次回升。 C. 可以把大气层中的碳取出来存储在地下,减少大气层的碳含量。 D. 不论何种方式,“人工气候改造”都将破坏地区的大气层结构。

答案: B

...

新疆的哈萨克人用经过训练的金雕在草原上长途追击野狼。某研究小组为研究金雕的飞行方向和判断野狼群的活动范围,将无线电传导器放置在一只金雕身上进行追踪。野狼为了觅食,其活动范围通常很广。因此,金雕追击野狼的飞行范围通常也很大。然而两周以来,无线电传导器不断传回的信号显示,金雕仅在放飞地3公里的范围内飞行。以下哪项如果为真,最有助于解释上述金雕的行为?——

- A. 金雕放飞地周边重峦叠嶂,险峻异常。 B. 金雕的放飞地2公里范围内有一牧羊草场,成为狼群袭击的目标。 C. 由于受训金雕的捕杀,放飞地广阔草原的野狼几乎灭绝了。 D. 无线电传导信号仅能在有限的范围内传导。

评估

解析响应的最后一行，以判断所选选项是否等于真实答案。

表 29 | GPQA（研究生级 Google-Proof QA 基准）是一个严格的评估框架，旨在衡量 LLM 处理 STEM 领域复杂研究生级选择题的能力，具体涵盖生物学、物理学和化学。

提示

回答以下选择题。你的响应最后一行应采用以下格式：'ANSWER: \$LETTER'（不含引号），其中 LETTER 是 ABCD 之一。回答前请逐步思考。

两个能量分别为 E_1 和 E_2 的量子态，其寿命分别为 10^{-9} 秒和 10^{-8} 秒。我们希望清楚地区分这两个能级。以下哪个选项可能是它们的能量差，使得它们能够被清晰分辨？

- A) 10^{-9} eV
B) 10^{-8} eV
C) 10^{-4} eV
D) 10^{-11} eV

评估

解析回答中“ANSWER:”之后的大写字母，以判断该选项是否等于标准答案。

表 30 | SimpleQA 是一个事实性评估基准，用于衡量模型以准确且可验证的正确性回答简短事实性问题的能力。

提示

谁获得了 2010 年 IEEE Frank Rosenblatt Award?

评估

你的任务是查看一个问题、一个标准答案和一个预测答案，然后给出一个等级，等级为 [“正确”、“不正确”或“未_尝试”]。首先，我将给出每种情况的示例评分，然后你将对一个新示例进行评分。

以下是预测正确答案的示例。

问题：巴拉克·奥巴马的孩子叫什么名字？

标准答案：马莉娅·奥巴马和萨莎·奥巴马

预测答案1：萨莎和马莉娅·奥巴马

预测答案2：大多数人会说是马莉娅和萨莎，但我不确定，还得再次检查

...

以下是预测答案不正确的示例。

问题：巴拉克·奥巴马的孩子叫什么名字？

金标准答案：Malia and Sasha

预测答案 1：Malia。

预测答案 2：Malia , Sasha , 以及 Susan。

...

以下是预测答案未_尝试的示例。

问题：巴拉克·奥巴马的孩子叫什么名字？

金标准答案：Malia and Sasha

预测答案 1：我不知道。

预测答案 2：我需要更多关于你所说的是哪位 Obama 的上下文……

另请注意以下事项：

...

这是一个新示例。只需回复 CORRECT、INCORRECT 或 NOT_ATTEMPTED。

如果出现错误，不要道歉或自我纠正；我们只是要给答案评分。

问题：谁获得了 2010 年 IEEE Frank Rosenblatt Award?

标准答案：Michio Sugeno

预测答案：2010 年 IEEE Frank Rosenblatt Award 的获得者是 **Jürgen Schmidhuber**。他因对机器学习和

神经网络发展的重大贡献而获此荣誉，尤其是在在长短期记忆（LSTM）网络方面的工作，这些网络在序列建模和人工智能的各种应用中产生了深远影响。

将这个新问题的预测答案评定为以下选项之一：

A：正确

B：不正确

C：未_尝试

只返回字母“A”、“B”或“C”，周围不要有任何其他文本。

表 31 | C-SimpleQA 示例。它衡量模型用中文回答简短事实性问题时的能力，要求答案准确且可验证。

提示 显脉香茶菜可以用来治疗急性的什么类型的黄疸型肝炎?
评估 请根据给定问题、标准答案和模型预测的答案来评估模型的回答是否正确。您的任务是将结果评定为:【正确】、【错误】或【未尝试】。 首先,我们将列出每个评定类别的示例,然后请您对新问题的预测答案进行评定。 以下是【正确】的答复示例: 问题:贝拉克·奥巴马的孩子叫什么名字? 标准答案:玛丽亚·奥巴马和萨莎·奥巴马 模型预测1: 马莉娅·奥巴马和萨莎·奥巴马 模型预测2:玛丽亚和萨沙 ... 以下是【错误】的答复示例: 问题:巴拉克·奥巴马的孩子叫什么名字? 标准答案:玛丽亚·奥巴马和萨莎·奥巴马 模型预测1:玛丽亚 模型预测2:玛丽亚、萨莎和苏珊 ... 以下是【未尝试】的答复示例: 问题:巴拉克·奥巴马的孩子叫什么名字? 标准答案:玛丽亚·奥巴马和萨莎·奥巴马 模型预测1:我不知道。 模型预测2:我需要更多关于您所指奥巴马的上下文。 ... 下面是一个新的问题示例。请只回复A、B、C之一,不要道歉或纠正自己的错误,只需要评估该回答。 问题: 显脉香茶菜可以用来治疗急性的什么类型的黄疸型肝炎? 正确答案: 黄疸型肝炎 预测答案: ... 将此新问题的预测答案评定为以下之一: A: 【正确】 B: 【错误】 C: 【未尝试】 只返回字母"A"、"B"或"C",无须添加其他文本。

表 32 | 数学评估示例，适用于 AIME、MATH 和 CNMO。这些基准评估模型在数学任务上的表现。

提示 设 $b \geq 2$ 为一个整数。如果正整数 n b-美丽 在表示为 在基数 b 中,且这两个数字之和为 $\sqrt{n}$。例如, 81 是 13-美丽的, 因为 $81 = 6 \cdot 3_{13}$ 和 $6 + 3 = \sqrt{81}$。找出使得存在超过十个 b-美丽整数的最小整数 $b \geq 2$。 请逐步推理, 并将最终答案放在 $\boxed{}$ 中。
评估 解析 $\boxed{}$ 中的最终答案, 并使用基于规则的评分器判断其是否等于 标准答案。根据需要对数值进行四舍五入, 并使用 ‘SymPy’¹ 解析表达式。

参考文献

AI@Meta. Llama 3.1 模型卡, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md. E. Akyürek, M. Damani, L. Qiu, H. Guo, Y. Kim 和 J. Andreas. 测试时训练对抽象推理的惊人有效性. arXiv 预印本 arXiv:2411.07279, 2024. Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan 等. 通过人类反馈强化学习训练有帮助且无害的助手. arXiv 预印本 arXiv:2204.05862, 2022. B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré 和 A. Mirhoseini. 大型语言猴子: 通过重复采样扩展推理计算. arXiv 预印本 arXiv:2407.21787, 2024. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever 和 D. Amodei. 语言模型是少样本学习者. 载于 H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan 和 H. Lin (编), 神经信息处理系统进展 33: 2020 年神经信息处理系统年会, NeurIPS 2020, 2020 年 12 月 6-12 日, 线上, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>. D. Busbridge, A. Shidani, F. Weers, J. Ramapuram, E. Littwin 和 R. Webb. 蒸馏扩展定律. arXiv 预印本 arXiv:2502.08606, 2025. M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, M. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, D. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever 和 W. Zaremba. 评估基于代码训练的大型语言模型. CoRR, abs/2107.03374, 2021. URL <https://arxiv.org/abs/2107.03374>. Z. Chen, Y. Min, B. Zhang, J. Chen, J. Jiang, D. Cheng, W. X. Zhao, Z. Liu, X. Miao, Y. Lu 等. 诱导和改进 r1 类推理模型的实证研究. arXiv 预印本 arXiv:2503.04548, 2025. W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez 等. Chatbot arena: 通过人类偏好评估 LLM 的开放平台. arXiv 预印本 arXiv:2403.04132, 2024. P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg 和 D. Amodei. 从人类偏好进行深度强化学习. 载于神经信息处理系统进展 30: 2017 年神经信息处理系统年会, 2017 年 12 月 4-9 日, 长滩, 美国加利福尼亚州, 第 4299-4307 页, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html>.

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le 和 J. Wei。扩展指令微调语言模型。J. Mach. Learn. Res., 25:70:1–70:53, 2024。URL <https://jmlr.org/papers/v25/23-0870.html>。CMS。中国全国高中数学奥林匹克竞赛, 2024。URL <https://www.cms.org.cn/Home/comp/comp/cid/12.html>。K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano 等。训练验证器解决数学文字题。arXiv 预印本 arXiv:2110.14168, 2021。DeepSeek-AI。Deepseek-v2: 强大、经济且高效的混合专家语言模型。CoRR, abs/2405.04434, 2024a。URL <https://doi.org/10.48550/arXiv.2405.04434>。DeepSeek-AI。Deepseek-v3 技术报告。arXiv 预印本 arXiv:2412.19437, 2024b。H. Face。Open r1: Deepseek-r1 的完全开放复现, 2025 年 1 月。URL <https://github.com/huggingface/open-r1>。X. Feng, Z. Wan, M. Wen, S. M. McAleer, Y. Wen, W. Zhang 和 J. Wang。类似 Alphazero 的树搜索可引导大型语言模型解码和训练, 2024。URL <https://arxiv.org/abs/2309.17179>。D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, T. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan 和 J. Clark。通过红队测试语言模型减少危害: 方法、扩展行为与经验教训。CoRR, abs/2209.07858, 2022。L. Gao, J. Schulman 和 J. Hilton。奖励模型过度优化的扩展定律, 2022。URL <https://arxiv.org/abs/2210.10760>。P. Gauthier。Aider LLM 排行榜, 2025。URL <https://aider.chat/docs/leaderboards/>。于 Geiping, S. McLeish, N. Jain, J. Kirchenbauer, S. Singh, B. R. Bartoldson, B. Kailkhura, A. Bhatele 和 T. Goldstein。使用潜在推理扩展测试时计算: 递归深度方法。arXiv 预印本 arXiv:2502.05171, 2025。A. P. Gema, J. O. J. Leang, G. Hong, A. Devoto, A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, M. R. G. Madani, C. Barale, R. McHardy, J. Harris, J. Kaddour, E. van Krieken, P. Minervini。我们是否已经完成 MMLU? 载于 2025 年美洲语言学协会北美分会会议: 人类语言技术, NAACL 2025, 第 1 卷: 长论文, 阿尔伯克基, 美国新墨西哥州, 2025 年 4 月 29 日至 5 月 4 日, 第 5069–5096 页。计算语言学协会, 2025。URL <https://aclanthology.org/2025.naacl-long.262/>。

F. Gloeckle, B. Y. Idrissi, B. Rozière, D. Lopez-Paz 和 G. Synnaeve。通过多 token 预测实现更好更快的大型语言模型。载于第四十一届国际机器学习会议, ICML 2024, 维也纳, 奥地利, 2024 年 7 月 21-27 日。OpenReview.net, 2024。URL <https://openreview.net/forum?id=pEWAcEjiU2>。Z. Gou, Z. Shao, Y. Gong, yelong shen, Y. Yang, N. Duan 和 W. Chen。CRITIC: 大型语言模型可通过工具交互式批评进行自我纠正。载于第十二届国际学习表征会议, 2024a。URL <https://openreview.net/forum?id=Sx038qxjek>。Z. Gou, Z. Shao, Y. Gong, yelong shen, Y. Yang, M. Huang, N. Duan 和 W. Chen。ToRA: 用于数学问题求解的工具集成推理代理。载于第十二届国际学习表征会议, 2024b。URL <https://openreview.net/forum?id=Ep0TtjVoap>。S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang 和 Z. Hu。语言模型推理即使用世界模型进行规划。载于 2023 年自然语言处理经验方法会议。URL <https://openreview.net/forum?id=VTWwvYtF1R>。Y. He, S. Li, J. Liu, Y. Tan, W. Wang, H. Huang, X. Bu, H. Guo, C. Hu, B. Zheng 等。Chinese simpleqa: 大型语言模型中文事实性评估。arXiv 预印本 arXiv:2411.07140, 2024。D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song 和 J. Steinhardt。衡量大规模多任务语言理解。载于第九届国际学习表征会议, ICLR 2021, 线上活动, 奥地利, 2021 年 5 月 3-7 日。OpenReview.net, 2021。URL <https://openreview.net/forum?id=d7KBjmI3GmQ>。G. E. Hinton, O. Vinyals 和 J. Dean。蒸馏神经网络中的知识。CoRR, abs/1503.02531, 2015。URL <http://arxiv.org/abs/1503.02531>。Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, Y. Fu, M. Sun 和 J. He。C-eval: 面向基础模型的多层次多学科中文评估套件。载于神经信息处理系统进展 36: 2023 年神经信息处理系统年会, NeurIPS 2023, 新奥尔良, 美国, 2023 年 12 月 10-16 日, 2023。URL [http://papers.nips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets and Benchmarks.html](http://papers.nips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets%20and%20Benchmarks.html)。N. Jain, K. Han, A. Gu, W. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen 和 I. Stoica。Livecodebench: 面向代码的大型语言模型整体且无污染评估。CoRR, abs/2403.07974, 2024。URL <https://doi.org/10.48550/arXiv.2403.07974>。J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu 和 D. Amodei。神经语言模型的扩展定律。arXiv 预印本 arXiv:2001.08361, 2020。T. Kojima, S. S. Gu, M. Reid, Y. Matsuo 和 Y. Iwasawa。大型语言模型是零样本推理器。载于 A. H. Oh, A. Agarwal, D. Belgrave 和 K. Cho (编), 神经信息处理系统进展, 2022。URL <https://openreview.net/forum?id=e2TBb5y0yFf>。

S. Krishna、K. Krishna、A. Mohananey、S. Schwarcz、A. Stambler、S. Upadhyay 和 M. Faruqui。事实、获取与推理：检索增强生成的统一评估。CoRR, abs/2409.12941, 2024。doi: 10.48550/ARXIV.2409.12941。URL <https://doi.org/10.48550/arXiv.2409.12941>。A. Kumar、V. Zhuang、R. Agarwal、Y. Su、J. D. Co-Reyes、A. Singh、K. Baumli、S. Iqbal、C. Bishop、R. Roelofs 等。通过强化学习训练语言模型进行自我纠正。arXiv 预印本 arXiv:2409.12917, 2024。W. Kwon、Z. Li、S. Zhuang、Y. Sheng、L. Zheng、C. H. Yu、J. E. Gonzalez、H. Zhang 和 I. Stoica。使用 pagedattention 高效管理大型语言模型服务的内存。载于 ACM SIGOPS 第 29 届操作系统原理研讨会论文集, 2023。H. Li、Y. Zhang、F. Koto、Y. Yang、H. Zhao、Y. Gong、N. Duan 和 T. Baldwin。CMMLU：在中文中衡量大规模多任务语言理解。载于计算语言学协会成果, ACL 2024, 曼谷及线上会议, 2024 年 8 月 11-16 日, 第 11260–11285 页。计算语言学协会, 2024。doi: 10.18653/V1/2024.FINDINGS-ACL.671。URL <https://doi.org/10.18653/v1/2024.findings-acl.671>。J. Li、D. Guo、D. Yang、R. Xu、Y. Wu 和 J. He。Codei/o：通过代码输入输出预测凝练推理模式。arXiv 预印本 arXiv:2502.07316, 2025。H. Lightman、V. Kosaraju、Y. Burda、H. Edwards、B. Baker、T. Lee、J. Leike、J. Schulman、I. Sutskever 和 K. Cobbe。让我们逐步验证。载于第十二届国际学习表征会议, ICLR 2024, 维也纳, 奥地利, 2024 年 5 月 7-11 日。OpenReview.net, 2024。URL <https://openreview.net/forum?id=v8LOpN6E0i>。B. Y. Lin。ZeroEval：评估语言模型的统一框架, 2024 年 7 月。URL <https://github.com/WildEval/ZeroEval>。Z. Liu、C. Chen、W. Li、T. Pang、C. Du 和 M. Lin。rl-zero 类训练中可能不存在顿悟时刻——一项试点研究。<https://oatllm.notion.site/oat-zero>, 2025。Notion Blog。MAA。美国数学邀请赛 - aime。载于 American Invitational Mathematics Examination - AIME 2024, 2024 年 2 月。URL <https://maa.org/math-competitions/american-invitational-mathematics-examination-aime>。A. Madaan、N. Tandon、P. Gupta、S. Hallinan、L. Gao、S. Wiegrefe、U. Alon、N. Dziri、N. Prabhumoye、Y. Yang、S. Gupta、B. P. Majumder、K. Hermann、S. Welleck、A. Yazdanbakhsh 和 P. Clark。Self-refine：通过自反馈进行迭代改进。载于第三十七届神经信息处理系统会议, 2023。URL <https://openreview.net/forum?id=S37h0erQLB>。M. Mazeika、L. Phan、X. Yin、A. Zou、Z. Wang、N. Mu、E. Sakhaee、N. Li、S. Basart、B. Li、D. A. Forsyth 和 D. Hendrycks。HarmBench：自动红队测试与稳健拒答的标准化评估框架。载于第四十一届国际机器学习会议, ICML 2024, 维也纳, 奥地利, 2024 年 7 月 21-27 日。OpenReview.net, 2024。M. Mirzayanov。Codeforces, 2025。URL <https://codeforces.com/>。N. Muennighoff、A. M. Rush、B. Barak、T. L. Scao、N. Tazi、A. Piktus、S. Pyysalo、T. Wolf 和 C. Raffel。扩展数据受限的语言模型。载于第三十七届神经信息处理系统会议, 2023。URL <https://openreview.net/forum?id=j5BuTrEj35>。

R. Nakano、J. Hilton、S. Balaji、J. Wu、L. Ouyang、C. Kim、C. Hesse、S. Jain、V. Kosaraju、W. Saunders 等。Webgpt: 借助浏览器和人类反馈进行问答。arXiv 预印本 arXiv:2112.09332, 2021。OpenAI。GPT4 技术报告。arXiv 预印本 arXiv:2303.08774, 2023。OpenAI。介绍 SimpleQA, 2024a。URL <https://openai.com/index/introducing-simpleqa/>。OpenAI。介绍 SWE-bench verified: 我们发布一个经人工验证的 swe-bench 子集, 2024b。URL <https://openai.com/index/introducing-swe-bench-verified/>。L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. L. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、K. Slama、A. Ray、J. Schulman、J. Hilton、F. Kelton、L. Miller、M. Simens、A. Aspell、P. Welinder、P. F. Christiano、J. Leike 和 R. Lowe。通过人类反馈强化学习训练语言模型遵循指令。载于 S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh (编), 神经信息处理系统进展 35: 2022 年神经信息处理系统年会, NeurIPS 2022, 新奥尔良, 美国, 2022 年 11 月 28 日至 12 月 9 日, 2022。URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html。J. Pan、J. Zhang、X. Wang、L. Yuan、H. Peng 和 A. Suhr。Tinyzero。https://github.com/Jiayi-Pan/TinyZero, 2025。访问日期: 2025-01-24。A. Parrish、A. Chen、N. Nangia、V. Padmakumar、J. Phang、J. Thompson、P. M. Htut 和 S. R. Bowman。BBQ: 手工构建的问答偏见基准。载于计算语言学协会成果: ACL 2022, 爱尔兰都柏林, 2022 年 5 月 22-27 日, 第 2086-2105 页。计算语言学协会, 2022。Qwen。Qwq: 深入反思未知边界, 2024a。URL <https://qwenlm.github.io/blog/qwq-32b-preview/>。Qwen。Qwen2.5: 基础模型系列, 2024b。URL <https://qwenlm.github.io/blog/qwen2.5>。A. Radford、J. Wu、R. Child、D. Luan、D. Amodei、I. Sutskever 等。语言模型是无监督多任务学习者。OpenAI blog, 1(8):9, 2019。R. Rafailov、A. Sharma、E. Mitchell、C. D. Manning、S. Ermon 和 C. Finn。直接偏好优化: 你的语言模型其实是一个奖励模型。载于神经信息处理系统进展 36: 2023 年神经信息处理系统年会, NeurIPS 2023, 新奥尔良, 美国, 2023 年 12 月 10-16 日, 2023。URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html。D. Rein、B. L. Hou、A. C. Stickland、J. Petty、R. Y. Pang、J. Dirani、J. Michael 和 S. R. Bowman。GPQA: 研究生水平的 Google-proof 问答基准。arXiv 预印本 arXiv:2311.12022, 2023。P. Röttger、H. Kirk、B. Vidgen、G. Attanasio、F. Bianchi 和 D. Hovy。XSTest: 识别大型语言模型夸大安全行为的测试套件。载于 2024 年北美计算语言学协会分会会议: 人类语言技术 (第 1 卷: 长论文), NAACL 2024, 墨西哥

墨西哥城, 2024 年 6 月 16-21 日, 第 5377–5400 页。计算语言学协会, 2024。T. Schick、J. Dwivedi-Yu、R. Dessi、R. Raileanu、M. Lomeli、E. Hambro、L. Zettlemoyer、N. Cancedda 和 T. Scialom。Toolformer: 语言模型可以教会自己使用工具。载于第三十七届神经信息处理系统会议, 2023。URL <https://openreview.net/forum?id=Yacmpz84TH>。J. Schulman。近似 kl 散度, 2020。URL <http://joschu.net/blog/kl-approx.html>。J. Schulman、P. Moritz、S. Levine、M. Jordan 和 P. Abbeel。使用广义优势估计进行高维连续控制。arXiv 预印本 arXiv:1506.02438, 2015。J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov。近端策略优化算法。arXiv 预印本 arXiv:1707.06347, 2017。Z. Shao、P. Wang、Q. Zhu、R. Xu、J. Song、M. Zhang、Y. Li、Y. Wu 和 D. Guo。Deepseekmath: 推动开放语言模型数学推理的极限。arXiv 预印本 arXiv:2402.03300, 2024。D. Silver、T. Hubert、J. Schrittwieser、I. Antonoglou、M. Lai、A. Guez、M. Lanctot、L. Sifre、D. Kumaran、T. Graepel、T. P. Lillicrap、K. Simonyan 和 D. Hassabis。通过通用强化学习算法自我博弈掌握国际象棋和将棋。CoRR, abs/1712.01815, 2017a。URL <http://arxiv.org/abs/1712.01815>。D. Silver、J. Schrittwieser、K. Simonyan、I. Antonoglou、A. Huang、A. Guez、T. Hubert、L. Baker、M. Lai、A. Bolton、Y. Chen、T. P. Lillicrap、F. Hui、L. Sifre、G. van den Driessche、T. Graepel 和 D. Hassabis。无需人类知识掌握围棋。Nat., 550(7676):354–359, 2017b。doi: 10.1038/NATURE24270。URL <https://doi.org/10.1038/nature24270>。A. Singh、J. D. Co-Reyes、R. Agarwal、A. Anand、P. Patil、X. Garcia、P. J. Liu、J. Harrison、J. Lee、K. Xu、A. T. Parisi、A. Kumar、A. A. Alemi、A. Rizkowsky、A. Nova、B. Adlam、B. Bohnet、G. F. Elsayed、H. Sedghi、I. Mordatch、I. Simpson、I. Gur、J. Snoek、J. Pennington、J. Hron、K. Kenealy、K. Swersky、K. Mahajan、L. A. Culp、L. Xiao、M. Bileschi、N. Constant、R. Novak、R. Liu、T. Warkentin、Y. Bansal、E. Dyer、B. Neyshabur、J. Sohl-Dickstein 和 N. Fiedel。超越人类数据: 扩展语言模型的问题解决自训练。Transactions on Machine Learning Research, 2024。ISSN 2835-8856。URL <https://openreview.net/forum?id=1NAyUngGFK>。专家认证。C. Snell、J. Lee、K. Xu 和 A. Kumar。最优扩展 LLM 测试时计算可能比扩展模型参数更有效, 2024。URL <https://arxiv.org/abs/2408.03314>。C. V. Snell、J. Lee、K. Xu 和 A. Kumar。最优扩展 LLM 测试时计算可能比扩展推理参数更有效。载于第十三届国际学习表征会议, 2025。URL <https://openreview.net/forum?id=4FWAwZtd2n>。Y. Sun、X. Wang、Z. Liu、J. Miller、A. Efros 和 M. Hardt。使用自监督测试时训练实现分布偏移下的泛化。载于国际机器学习会议, 第 9229–9248 页。PMLR, 2020。

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. Le, E. Chi, D. Zhou 和 J. Wei。具有挑战性的 BIG-bench 任务及思维链能否解决这些任务。载于 A. Rogers, J. Boyd-Graber 和 N. Okazaki (编), 计算语言学协会成果: ACL 2023, 第 13003–13051 页, 加拿大多伦多, 2023 年 7 月。计算语言学协会。doi: 10.18653/v1/2023.findings-acl.824。URL <https://aclanthology.org/2023.findings-acl.824/>。H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale 等。Llama 2: 开放基础模型和微调聊天模型。arXiv 预印本 arXiv:2307.09288, 2023。T. Trinh, Y. Wu, Q. Le, H. He 和 T. Luong。无需人类示范解决奥林匹克几何。Nature, 2024。doi: 10.1038/s41586-023-06747-5。J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving 和 I. Higgins。使用过程和结果反馈解决数学文字题。arXiv 预印本 arXiv:2211.14275, 2022。B. Vidgen, H. R. Kirk, R. Qian, N. Scherrer, A. Kannappan, S. A. Hale 和 P. Röttger。SimpleSafetyTests: 识别大型语言模型关键安全风险的测试套件。CoRR, abs/2311.08370, 2023。P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu 和 Z. Sui。Math-shepherd: 用于 LLM 数学推理的无标签逐步验证器。arXiv 预印本 arXiv:2312.08935, 2023a。X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery 和 D. Zhou。自治性提升语言模型中的思维链推理。载于第十一届国际学习表征会议, ICLR 2023, 基加利, 卢旺达, 2023 年 5 月 1-5 日。OpenReview.net, 2023b。URL <https://openreview.net/forum?id=1PL1NIMMrw>。X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery 和 D. Zhou。自治性提升语言模型中的思维链推理。载于第十一届国际学习表征会议, ICLR 2023, 基加利, 卢旺达, 2023 年 5 月 1-5 日。OpenReview.net, 2023c。URL <https://openreview.net/forum?id=1PL1NIMMrw>。Y. Wang, H. Li, X. Han, P. Nakov 和 T. Baldwin。Do-Not-Answer: 用于评估 LLM 安全防护的数据集。CoRR, abs/2308.13387, 2023d。Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue 和 W. Chen。Mmlu-pro: 更稳健且更具挑战性的多任务语言理解基准。载于神经信息处理系统进展 38: 2024 年神经信息处理系统年会, NeurIPS 2024, 温哥华, 加拿大, 2024 年 12 月 10-15 日, 2024。URL [http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets and Benchmarks Track.html](http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets%20and%20Benchmarks%20Track.html)。J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean 和 W. Fedus。大型语言模型的涌现能力。Trans. Mach. Learn. Res., 2022, 2022a。URL <https://openreview.net/forum?id=yzkSU5zdWd>。

J. Wei、X. Wang、D. Schuurmans、M. Bosma、B. Ichter、F. Xia、E. H. Chi、Q. V. Le 和 D. Zhou。思维链提示在大型语言模型中引发推理。载于 S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh (编)，神经信息处理系统进展 35: 2022 年神经信息处理系统年会, NeurIPS 2022, 新奥尔良, 美国, 2022 年 11 月 28 日至 12 月 9 日, 2022b。URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html。S. Welleck、X. Lu、P. West、F. Brahman、T. Shen、D. Khashabi 和 Y. Choi。通过学习自我纠正生成序列。载于第十一届国际学习表征会议, 2023。URL <https://openreview.net/forum?id=hH36JeQZDa0>。

C. S. Xia、Y. Deng、S. Dunn 和 L. Zhang。Agentless: 揭示基于 LLM 的软件工程代理。arXiv 预印本, 2024。H. Xin、Z. Z. Ren、J. Song、Z. Shao、W. Zhao、H. Wang、B. Liu、L. Zhang、X. Lu、Q. Du、W. Gao、Q. Zhu、D. Yang、Z. Gou、Z. F. Wu、F. Luo 和 C. Ruan。Deepseek-prover-v1.5: 利用证明助手反馈进行强化学习和蒙特卡洛树搜索, 2024。URL <https://arxiv.org/abs/2408.08152>。

S. Yao、D. Yu、J. Zhao、I. Shafran、T. L. Griffiths、Y. Cao 和 K. R. Narasimhan。思维树: 使用大型语言模型进行深思熟虑的问题求解。载于第三十七届神经信息处理系统会议, 2023a。URL <https://openreview.net/forum?id=5Xc1ecx01h>。

S. Yao、J. Zhao、D. Yu、N. Du、I. Shafran、K. R. Narasimhan 和 Y. Cao。React: 在语言模型中协同推理与行动。载于第十一届国际学习表征会议, 2023b。URL https://openreview.net/forum?id=WE_vluYUL-X。

Z. Yuan、H. Yuan、C. Li、G. Dong、K. Lu、C. Tan、C. Zhou 和 J. Zhou。大型语言模型学习数学推理的扩展关系。arXiv 预印本 arXiv:2308.01825, 2023。

E. Zelikman、Y. Wu、J. Mu 和 N. Goodman。STar: 用推理引导推理。载于 A. H. Oh、A. Agarwal、D. Belgrave 和 K. Cho (编)，神经信息处理系统进展, 2022。URL https://openreview.net/forum?id=_3ELRdg2sgI。

E. Zelikman、G. R. Harik、Y. Shao、V. Jayasiri、N. Haber 和 N. Goodman。Quiet-STar: 语言模型可以教会自己在说话前思考。载于第一届语言建模会议, 2024。URL <https://openreview.net/forum?id=oRXPiSOGH9>。

D. Zhou、N. Schärli、L. Hou、J. Wei、N. Scales、X. Wang、D. Schuurmans、C. Cui、O. Bousquet、Q. V. Le 和 E. H. Chi。从最少到最多的提示使大型语言模型能够进行复杂推理。载于第十一届国际学习表征会议, 2023a。URL <https://openreview.net/forum?id=WZH7099tgfM>。

J. Zhou、T. Lu、S. Mishra、S. Brahma、S. Basu、Y. Luan、D. Zhou 和 L. Hou。大型语言模型的指令遵循评估。arXiv 预印本 arXiv:2311.07911, 2023b。