DeepSeek-R1：通过以下方式激励 LLM 中的推理能力
强化学习

深度搜索-AI

research@deepseek.com

抽象

我们推出了我们的第一代推理模型，DeepSeek-R1-Zero 和 DeepSeek-R1。DeepSeek-R1-Zero 是一个通过大规模强化学习 (RL) 训练的模型，未作为初步步骤进行监督微调 (SFT)，展现出卓越的推理能力。通过强化学习，DeepSeek-R1-Zero 自然出现了众多强大而引人入胜的推理行为。然而，它也面临着可读性差和语言混合等挑战。为了解决这些问题并进一步提升推理性能，我们引入了 DeepSeek-R1，该模型在 RL 之前采用了多阶段训练和冷启动数据。DeepSeek-R1 在推理任务上的表现与 OpenAI-o1-1217 可比。为了支持研究社区，我们开源了 DeepSeek-R1-Zero、DeepSeek-R1，以及基于 Qwen 和 Llama 从 DeepSeek-R1 蒸馏的六个稠密模型（1.5B、7B、8B、14B、32B、70B）。

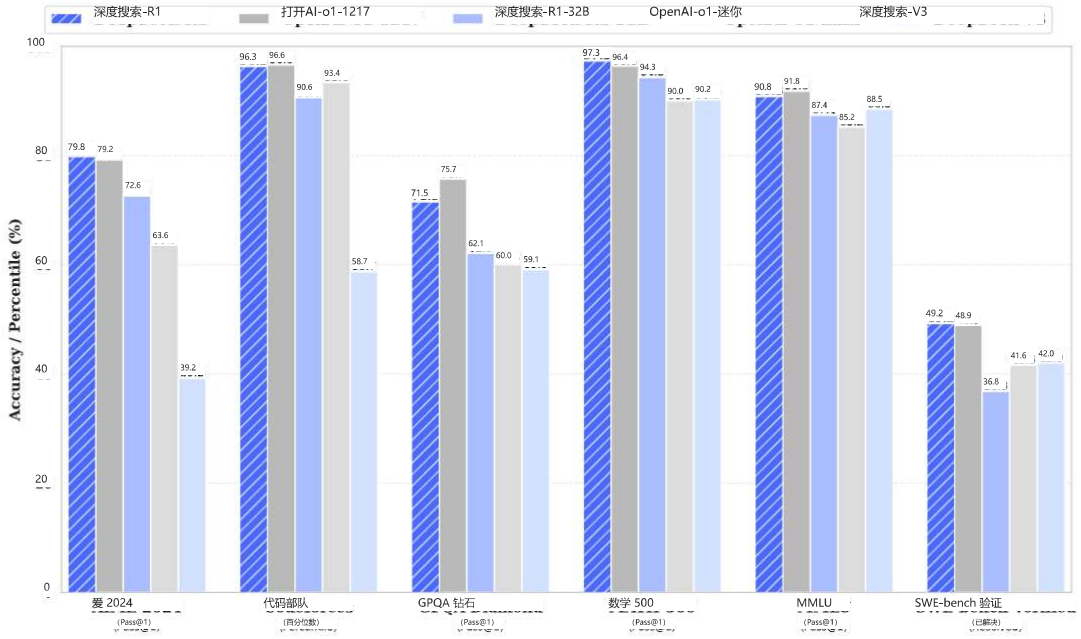

图 1 |DeepSeek-R1 的基准测试性能。

内容

1. 引言

近年来，大型语言模型（LLMs）正在迅速迭代和演化（Anthropic，2024；Google，2024；OpenAI，2024a），逐渐缩小与人工通用智能（AGI）之间的差距。

最近，后训练已成为完整训练流程中的一个重要组成部分。研究表明，它能够提高推理任务的准确性，符合社会价值观，并适应用户偏好，同时相较于预训练所需的计算资源相对较少。在推理能力方面，OpenAI 的 o1 系列模型首次引入了推理时扩展的概念，通过增加链式思维推理过程的长度实现。这种方法在数学、编程和科学推理等各种推理任务中取得了显著的改善。然而，有效的测试时扩展仍然是研究界的一个未解之题。一些先前的研究探索了各种方法，包括基于过程的奖励模型（Lightman 等，2023；Uesato 等，2022；Wang 等，2023）、强化学习（Kumar 等，2024），以及蒙特卡洛树搜索和束搜索等搜索算法（Feng 等，2024；Trinh 等，2024；Xin 等，2024）。然而，这些方法都未能达到与 OpenAI 的 o1 系列模型相当的通用推理性能。

在本文中，我们迈出了提升语言模型推理能力的第一步，采用纯强化学习（RL）。我们的目标是探索大型语言模型（LLM）在没有任何监督数据的情况下开发推理能力的潜力，重点关注其通过纯RL过程的自我进化。具体而言，我们使用DeepSeek-V3-Base作为基础模型，并采用GRPO（Shao等，2024）作为RL框架，以提高模型在推理方面的表现。在训练过程中，DeepSeek-R1-Zero自然而然地形成了许多强大且有趣的推理行为。经过数千步的RL训练，DeepSeek-R1-Zero在推理基准测试中表现出色。例如，AIME 2024的pass@1分数从15.6%提高至71.0%，并且通过多数投票，分数进一步提升至86.7%，与OpenAI-o1-0912的表现相匹配。

然而，DeepSeek-R1-Zero 遇到了可读性差和语言混合等挑战。为了应对这些问题并进一步增强推理性能，我们引入了 DeepSeek-R1，该模型结合了一小部分冷启动数据和多阶段训练流程。具体而言，我们首先收集了数千个冷启动数据来微调 DeepSeek-V3-Base 模型。在此之后，我们进行面向推理的强化学习，类似于 DeepSeek-R1-Zero。在强化学习过程接近收敛时，我们通过对 RL 检查点进行拒绝采样生成新的 SFT 数据，并结合来自 DeepSeek-V3 的监督数据，覆盖写作、事实问答和自我认知等领域，然后对 DeepSeek-V3-Base 模型进行再训练。经过新数据的微调后，该检查点经过额外的强化学习过程，考虑到所有场景的提示。经过这些步骤，我们得到了一个称为 DeepSeek-R1 的检查点，其性能达到了与 OpenAI-o1-1217 同等的水平。

我们进一步探索从 DeepSeek-R1 到较小的稠密模型的蒸馏。以 Qwen2.5-32B（Qwen，2024b）为基础模型，从 DeepSeek-R1 直接蒸馏的效果超过了在其上应用强化学习。这表明，较大基模发现的推理模式对于提升推理能力至关重要。我们开源了蒸馏版的 Qwen 和 Llama（Dubey et al., 2024）系列。值得注意的是，我们蒸馏的 14B 模型在性能上大幅超越了当前最先进的开源 QwQ-32B-Preview（Qwen，2024a），而蒸馏的 32B 和 70B 模型在稠密模型的推理基准测试中创下了新纪录。

1.1. 贡献

训练后：在基础模型上进行大规模强化学习

我们直接将强化学习（RL）应用于基础模型，而无需依赖监督微调（SFT）作为初步步骤。这种方法使得模型能够探索链式思维（CoT）以解决复杂问题，从而开发出了DeepSeek-R1-Zero。DeepSeek-R1-Zero展示了自我验证、反思和生成长链式思维的能力，标志着研究领域的一个重要里程碑。值得注意的是，这是首次公开研究验证大型语言模型的推理能力可以仅通过强化学习激励，而不需要监督微调。这一突破为未来在这一领域的进展铺平了道路。

我们介绍了开发DeepSeek-R1的流程。该流程包含两个强化学习阶段，旨在发现更好的推理模式并与人类偏好对齐，以及两个监督微调阶段，为模型的推理和非推理能力打下基础。我们相信该流程将通过创建更好的模型来推动行业的发展。

蒸馏：较小的模型也可以很强大

我们证明了更大模型的推理模式可以被提炼到较小的模型中，从而使其性能优于通过强化学习在小模型上发现的推理模式。开源的DeepSeek-R1及其API将使研究社区未来能够提炼出更好的小模型。

使用 DeepSeek-R1 生成的推理数据，我们对几个在研究界广泛使用的密集模型进行了微调。评估结果表明，经过蒸馏的小型密集模型在基准测试中表现出色。DeepSeek-R1-Distill-Qwen-7B 在 AIME 2024 中获得了 55.5%的分数，超越了 QwQ-32B-Preview。此外，DeepSeek-R1-Distill-Qwen-32B 在 AIME 2024 中得分 72.6%，在 MATH-500 中得分 94.3%，在 LiveCodeBench 中得分 57.2%。这些结果显著优于之前的开源模型，并且与 o1-mini 相当。我们向社区开源了基于 Qwen2.5 和 Llama3 系列的 1.5B、7B、8B、14B、32B 和 70B 检查点。

1.2. 评估结果总结

• 推理任务：（1）DeepSeek-R1在AIME 2024上以79.8%的Pass@1得分稍微超过OpenAI-o1-1217。在MATH-500上，它获得了97.3%的令人印象深刻的得分，表现与OpenAI-o1-1217相当，显著超越了其他模型。（2）在编码相关任务中，DeepSeek-R1在代码竞赛任务中表现出色，达到了2,029的Elo评级，在Codeforces上超越了96.3%的参赛人类参与者。在工程相关任务中，DeepSeek-R1的表现略优于DeepSeek-V3，这可能帮助开发者完成实际任务。

• 知识：在MMLU、MMLU-Pro和GPQA Diamond等基准测试中，DeepSeek-R1取得了出色的成绩，显著优于DeepSeek-V3，分别在MMLU上得分90.8%、MMLU-Pro上84.0%、GPQA Diamond上71.5%。虽然在这些基准测试中，其表现稍逊于OpenAI-o1-1217，但DeepSeek-R1超越了其他闭源模型，显示出在教育任务中的竞争优势。在事实基准测试SimpleQA中，DeepSeek-R1的表现也优于DeepSeek-V3，展示了其处理基于事实查询的能力。在该基准测试中，OpenAI-o1的表现也优于4o，呈现出类似的趋势。

DeepSeek-R1 在各种任务中表现出色，包括创意写作、一般问答、编辑、摘要等。它在 AlpacaEval 2.0 上实现了 87.6% 的长度控制胜率，在 ArenaHard 上胜率达到 92.3%，展示了其智能处理非考试导向查询的强大能力。此外，DeepSeek-R1 在需要长上下文理解的任务上表现出色，明显优于 DeepSeek-V3 在长上下文基准测试上的表现。

2. 方法

2.1. 概述

以往的研究在提高模型性能方面主要依赖于大量的监督数据。在本研究中，我们展示了通过大规模的强化学习（RL）显著提升推理能力，即使在没有使用监督微调（SFT）的情况下作为冷启动。此外，通过加入少量冷启动数据，性能还可以进一步提升。在接下来的部分中，我们介绍：（1）DeepSeek-R1-Zero，该方法直接将强化学习应用于基础模型，而不使用任何SFT数据；（2）DeepSeek-R1，该方法从经过数千个长的链式思维（CoT）示例微调的检查点开始应用强化学习；（3）从DeepSeek-R1中提炼推理能力到小型稠密模型。

2.2. DeepSeek-R1-Zero：基础模型上的强化学习

强化学习在推理任务中表现出了显著的有效性，正如我们之前的研究所证明的（Shao et al., 2024; Wang et al., 2023）。然而，这些研究严重依赖于监督数据，而收集这些数据的过程非常耗时。在本节中，我们探讨了大型语言模型（LLMs）在没有任何监督数据的情况下发展推理能力的潜力，重点关注它们通过纯粹的强化学习过程自我进化的能力。我们将首先简要概述我们的强化学习算法，然后展示一些令人兴奋的结果，希望这能为社区提供有价值的见解。

2.2.1. 强化学习算法

为了节省强化学习的训练成本，我们采用了群体相对策略优化（GRPO）（Shao et al., 2024），该方法省去了通常与策略模型大小相同的评论员模型，而是从群体得分中估计基线。具体而言，对于每个问题 q，GRPO 从旧策略 $\pi\theta old$ 中采样一组输出 $\{o1, o2, \cdots, oG\}$，然后通过最大化以下目标来优化策略模型 $\pi\theta$：

$$JGRPO\ (\theta) = E[q \sim P\ (Q),\ \{oi\}Gi=1 \sim \pi\theta oId\ (O|q)\ ]$$

$$\frac{1}{G}\sum_{i=1} \text{分钟}\left(\frac{\pi\theta\ (oi\,|q)}{\pi\theta oId\ (oi\,|q)}Ai,\ \text{剪辑}\left(\frac{\pi\theta\ (oi\,|q)}{\pi\theta oId\ (oi\,|q)},\ 1-\varepsilon,\ 1+\varepsilon\right)\right) - \beta DKL\pi\theta\,||\pi re\,f, \quad (1)$$

$$DKL\pi\theta\,||\pi re\,f = \frac{\pi re\,f\ (oi\,|q)}{\pi\theta\ (oi\,|q)} - \text{对数}\frac{\pi re\,f\ (oi\,|q)}{\pi\theta\ (oi\,|q)} - 1, \quad (2)$$

其中 ε 和 β 是超参数，Ai 是优势，通过一组奖励 $\{r1, r2, \ldots, rG\}$ 计算得出，这些奖励对应于每个组中的输出。

$$Ai = \frac{ri - mean\ (\{r1,\ r2,\ \cdots,\ rG\})}{std\ (\{r1,\ r2,\ \cdots,\ rG\})}. \quad (3)$$

用户和 Assistant 之间的对话。用户提出问题，Assistant 解决。Assistant 首先在脑海中思考推理过程，然后为用户提供答案。推理过程和答案分别包含在 and <answer> </answer> tags 中<think></think>，即 <think> 这里的推理过程 </think> <answer> 在此处回答</answer>。用户：提示符。 助理：

表 1 |DeepSeek-R1-Zero 的模板。prompt 将替换为具体推理问题。

2.2.2. 奖励建模

奖励是训练信号的来源，它决定了强化学习的优化方向。为了训练DeepSeek-R1-Zero，我们采用了一种基于规则的奖励系统，主要由两种类型的奖励组成：

- 准确性奖励：准确性奖励模型评估响应是否正确。例如，在具有确定性结果的数学问题中，模型需要以指定格式（例如，放在框内）提供最终答案，从而实现可靠的基于规则的正确性验证。类似地，对于 LeetCode 问题，可以使用编译器根据预定义的测试用例生成反馈。

- 格式奖励：除了准确性奖励模型，我们还采用格式奖励模型，要求模型将其思考过程置于‘<think>’和‘</think>’标签之间。

我们在开发 DeepSeek-R1-Zero 时没有使用结果或过程神经奖励模型，因为我们发现神经奖励模型在大规模强化学习过程中可能会遭受奖励黑客攻击，并且重新训练奖励模型需要额外的训练资源，这使得整个训练流程更加复杂。

2.2.3. 训练模板

为了训练DeepSeek-R1-Zero，我们首先设计了一个简单的模板，指导基础模型遵循我们指定的指令。如表1所示，该模板要求DeepSeek-R1-Zero首先产生推理过程，然后给出最终答案。我们故意限制我们的约束在这一结构格式上，避免任何与内容相关的偏见——例如，强制要求反思性推理或促进特定的解决问题策略——以确保我们能够准确观察模型在强化学习过程中的自然进展。

2.2.4. DeepSeek-R1-Zero 的性能、自进化过程和顿悟矩

DeepSeek-R1-Zero的性能图表展示了其在AIME 2024基准测试中的表现轨迹，贯穿整个强化学习训练过程。如图所示，DeepSeek-R1-Zero在训练过程中表现出稳定且持续的性能提升。值得注意的是，AIME 2024上的平均通过率@1分数显著提高，从最初的15.6%跃升至令人印象深刻的71.0%，达到了与OpenAI-o1-0912相当的性能水平。这一显著的改进突显了我们的强化学习算法在优化模型性能方面的有效性。

表2提供了DeepSeek-R1-Zero和OpenAI的o1-0912模型在各种推理相关基准上的比较分析。研究结果表明，强化学习赋予了

型	爱 2024		数学 500	GPQA 认证钻石	LiveCode 实时代码板凳	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	额定值
OpenAI-o1-迷你	63.6	80.0	90.0	60.0	53.8	1820
打开AI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
深度搜索-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

表 2 |DeepSeek-R1-Zero 和 OpenAI o1 模型在推理相关方面的比较基准。

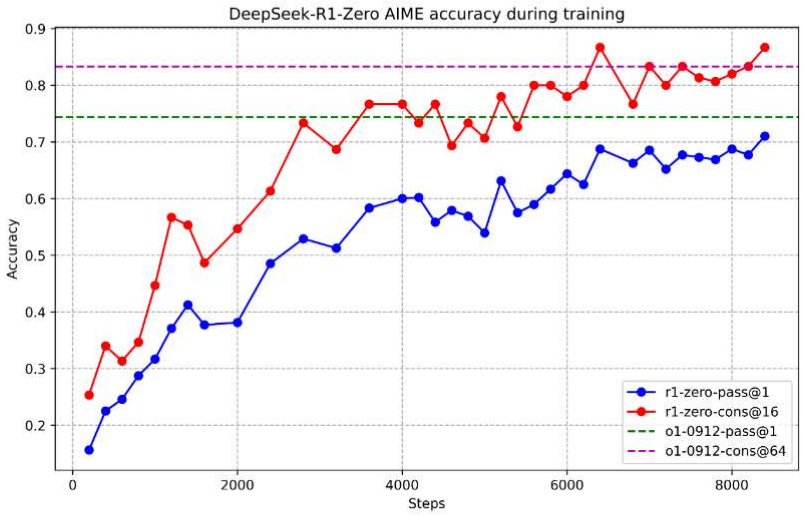

图2 | DeepSeek-R1-Zero在训练过程中的AIME准确率。对于每个问题，我们采样16个响应并计算总体平均准确率，以确保评估的稳定性。

DeepSeek-R1-Zero在没有任何监督微调数据的情况下，取得了强大的推理能力。这一成就是值得注意的，因为它强调了模型通过强化学习（RL）进行有效学习和泛化的能力。此外，通过应用多数投票，DeepSeek-R1-Zero的表现还可以进一步增强。例如，当在AIME基准上采用多数投票时，DeepSeek-R1-Zero的表现从71.0%提升至86.7%，从而超过了OpenAI-o1-0912的表现。DeepSeek-R1-Zero在有无多数投票情况下都能实现如此竞争力的表现，突显了其强大的基础能力以及在推理任务上进一步发展的潜力。

DeepSeek-R1-Zero 的自我进化过程 DeepSeek-R1-Zero 的自我进化过程

这是一个引人入胜的示范，展示了强化学习如何推动模型自主提高其推理能力。通过直接从基础模型启动强化学习，我们可以在没有监督微调阶段影响的情况下，密切监控模型的进展。这种方法清晰地展示了模型随时间演变的过程，特别是在处理复杂推理任务时的能力。

如图 3 所示，DeepSeek-R1-Zero 的思考时间显示出持续的改善——

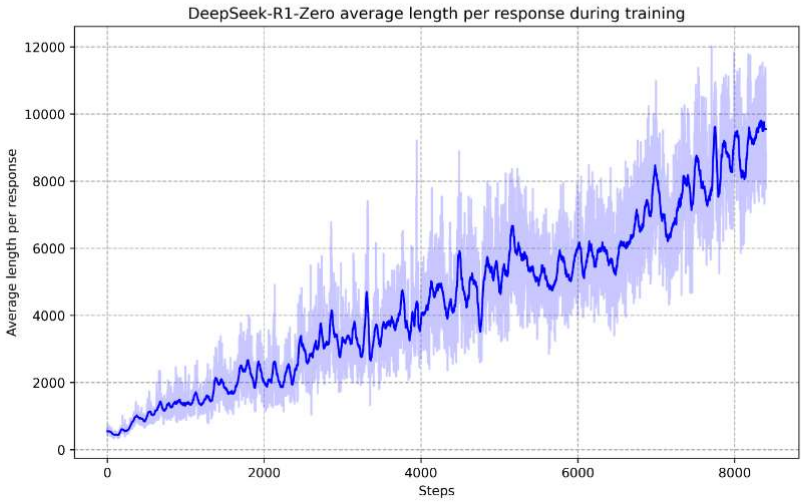

图3 | DeepSeek-R1-Zero在训练集上进行强化学习过程中的平均响应长度。DeepSeek-R1-Zero自然地学习在更多思考时间内解决推理任务。

在训练过程中，这种改进不是外部调整的结果，而是模型内部的自我发展。DeepSeek-R1-Zero 自然获得了解决日益复杂的推理任务的能力，利用扩展的测试时间计算。这种计算范围从生成数百到数千个推理令牌，使模型能够更深入地探索和完善其思维过程。

这种自我演化最显著的方面之一是，随着测试时间计算的增加，复杂行为的出现。诸如反思——模型重新审视和重新评估其先前步骤的行为——以及探索替代问题解决方法的能力自发产生。这些行为并不是显式编程的，而是模型与强化学习环境互动的结果。这种自发的发展显著增强了DeepSeek-R1-Zero的推理能力，使其能够更高效、更准确地应对更具挑战性的任务。

在DeepSeek-R1-Zero的训练过程中观察到的一个特别有趣的现象是“顿悟时刻”的出现。正如表3所示，这一时刻发生在模型的一个中间版本中。在这个阶段，DeepSeek-R1-Zero通过重新评估其初始方法，学会了在解决问题时分配更多的思考时间。这种行为不仅证明了模型日益增长的推理能力，而且是强化学习如何导致意想不到和复杂结果的迷人例子。

这一时刻不仅是模型的“顿悟时刻”，也是观察其行为的研究人员的“顿悟时刻”。它突显了强化学习的力量与美：我们并不是明确教模型如何解决问题，而是给它提供正确的激励，让它自主发展出高级的解决问题策略。“顿悟时刻”是一个强有力的提醒，展示了强化学习在释放人工系统新智能层次方面的潜力，为未来更自主和适应性强的模型铺平了道路。

√ a − √a + x = x，让我们先把两者平方 · · ·

–问题：如果 a > 1，则 √ a − √a + x = x 的实数解之和等于 –

响应： <think> 求解方程 √ a − √a + x2= x2 =⇒ a − √a + x = x2。重新排列以隔离内部平方根项： （a − x2）2 = a + x =⇒ a2 − 2ax2 + （x2）2 = a + x =⇒ x4 − 2ax2 − x + （a2 − a） = 0 .

等等。等。这是我可以在这里标记的顿悟时刻。让我们一步一步地重新评估一下，以确定正确的总和是否可以 · · ·我们从方程式开始：√ a − √a + x = x

首先，让我们将两边平方：a − √a + x = x2 =⇒ √a + x = a − x2 接下来，我可以再次将两边平方，处理方程式： · · · . . .

表3 | DeepSeek-R1-Zero 中间版本的一个有趣的“顿悟时刻”。该模型学会了以人性化的语气重新思考。这对我们来说也是一个顿悟时刻，让我们见证了强化学习的强大与美妙。

DeepSeek-R1-Zero的缺点 尽管DeepSeek-R1-Zero展现出强大的推理能力，并自主发展出意想不到且强有力的推理行为，但它面临一些问题。例如，DeepSeek-R1-Zero在可读性差和语言混用等挑战上存在困难。为了使推理过程更易读并与公众分享，我们探索了DeepSeek-R1，这是一种利用强化学习和人类友好的冷启动数据的方法。

2.3. DeepSeek-R1： 使用冷启动进行强化学习

受DeepSeek-R1-Zero的良好结果启发，两个自然问题出现：1）通过加入少量高质量数据作为冷启动，是否可以进一步提高推理性能或加快收敛？2）我们如何训练一个用户友好的模型，既能生成清晰连贯的思维链（CoT），又能展示出强大的通用能力？为了解决这些问题，我们设计了一个培训DeepSeek-R1的流程。该流程分为四个阶段，具体如下。

2.3.1. 冷启动

与DeepSeek-R1-Zero不同，为了防止RL训练初始模型的早期不稳定冷启动阶段，我们为DeepSeek-R1构建并收集了一小部分长的连锁思维（CoT）数据，以微调模型作为初始RL演员。为了收集这些数据，我们探索了几种方法：使用带有长CoT示例的少量提示，直接提示模型生成详细的答案并进行反思和验证，收集DeepSeek-R1-Zero以可读格式输出的结果，以及通过人工注释对结果进行后处理和精炼。

在这项工作中，我们收集了数千个冷启动数据，以对DeepSeek-V3-Base进行微调，作为强化学习的起点。与DeepSeek-R1-Zero相比，冷启动数据的优势在于。

包括：

• 可读性：DeepSeek-R1-Zero的一项关键限制是其内容通常不适合阅读。回应可能混合多种语言，或者缺乏markdown格式来突出显示用户的答案。相比之下，在为DeepSeek-R1创建冷启动数据时，我们设计了一种可读的模式，包括在每个回应末尾的总结，并过滤掉不适合阅读的回应。在这里，我们定义输出格式为 |special_token|<推理过程>|special_token|，其中推理过程是针对查询的链式思维，而总结则用于总结推理结果。

潜力：通过仔细设计冷启动数据的模式，并结合人类先验，我们观察到相较于DeepSeek-R1-Zero表现更好。我们相信迭代训练是一种更好的推理模型方式。

2.3.2. 面向推理的强化学习

在对DeepSeek-V3-Base进行冷启动数据的微调后，我们应用与DeepSeek-R1-Zero相同的大规模强化学习训练过程。此阶段专注于增强模型的推理能力，特别是在编码、数学、科学和逻辑推理等推理密集型任务中，这些任务涉及具有明确解答的定义清晰的问题。在训练过程中，我们观察到CoT经常表现出语言混合，尤其是当强化学习提示涉及多种语言时。为了缓解语言混合的问题，我们在强化学习训练中引入语言一致性奖励，该奖励的计算是根据CoT中目标语言词汇的比例。尽管消融实验表明这样的对齐会导致模型性能略微下降，但该奖励与人类偏好一致，使得模型的可读性更高。最后，我们通过直接相加推理任务的准确性和语言一致性奖励来形成最终奖励。然后，我们在微调后的模型上应用强化学习训练，直到其在推理任务上达到收敛。

2.3.3. 抑制采样和监督微调

当面向推理的强化学习收敛时，我们利用生成的检查点收集下一轮的SFT（监督微调）数据。这一阶段与最初主要关注推理的冷启动数据不同，它纳入了来自其他领域的数据，以增强模型在写作、角色扮演和其他通用任务中的能力。具体来说，我们生成这些数据并按照如下所述对模型进行微调。

我们策划推理提示并生成推理轨迹，通过对上述强化学习训练的检查点进行拒绝采样。在之前的阶段，我们只包含了可以使用基于规则的奖励进行评估的数据。然而，在这一阶段，我们通过加入额外的数据来扩展数据集，其中一些数据使用生成奖励模型，通过将真实值和模型预测输入到DeepSeek-V3进行判断。此外，由于模型输出有时混乱且难以阅读，我们过滤掉了混合语言、长段落和代码块的推理过程。对于每个提示，我们采样多个响应，仅保留正确的响应。总共，我们收集了大约60万个与推理相关的训练样本。

对于非推理数据，如写作、事实问答、自我认知和翻译，我们采用 DeepSeek-V3 管道，并重用 DeepSeek-V3 的部分 SFT 数据集。对于某些非推理任务，我们在回答问题之前调用 DeepSeek-V3 生成潜在的思维链。不过，对于较简单的查询，如“你好”，我们不提供思维链作为回应。最终，我们收集了大约 20 万条与推理无关的训练样本。

我们使用上述约80万样本的精选数据集对DeepSeek-V3-Base进行了两轮微调。

2.3.4. 适用于所有场景的强化学习

为了进一步使模型与人类偏好保持一致，我们实施了一个次强化学习阶段，旨在提高模型的有用性和无害性，同时精炼其推理能力。具体而言，我们使用奖励信号和多样化的提示分布进行模型训练。对于推理数据，我们遵循DeepSeek-R1-Zero中概述的方法，利用基于规则的奖励来指导数学、代码和逻辑推理领域的学习过程。对于一般数据，我们采用奖励模型来捕捉复杂和微妙场景中的人类偏好。我们在DeepSeek-V3管道的基础上，采取类似的偏好对和训练提示分布。对于有用性，我们专注于最终摘要，确保评估强调响应对用户的实用性和相关性，同时尽量减少对基础推理过程的干扰。对于无害性，我们评估模型的整个响应，包括推理过程和摘要，以识别和减轻在生成过程中可能出现的任何潜在风险、偏见或有害内容。最终，奖励信号和多样数据分布的整合使我们能够训练出一个在推理方面表现出色，同时优先考虑有用性和无害性的模型。

2.4. 蒸馏：赋予小模型推理能力

为了使更高效的小型模型具备像 DeepSeek-R1 这样的推理能力，我们直接对开源模型如 Qwen（Qwen, 2024b）和 Llama（AI@Meta, 2024）进行了微调，使用了通过 DeepSeek-R1 精心策划的 80 万个样本，具体细节见 §2.3.3。我们的研究发现，这种简单的蒸馏方法显著提升了小型模型的推理能力。我们使用的基础模型包括 Qwen2.5-Math-1.5B、Qwen2.5-Math-7B、Qwen2.5-14B、Qwen2.5-32B、Llama-3.1-8B 和 Llama-3.3-70B-Instruct。我们选择 Llama-3.3，因为它的推理能力比 Llama-3.1 略强。

对于蒸馏模型，我们仅应用SFT，而不包括RL阶段，尽管加入RL可能会显著提高模型性能。我们在这里的主要目标是展示蒸馏技术的有效性，将RL阶段的探索留给更广泛的研究社区。

3. 实验

基准 我们在 MMLU（Hendrycks 等，2020），MMLU-Redux（Gema 等，2024），MMLU-Pro（Wang 等，2024），C-Eval（Huang 等，2023），和 CMMLU（Li 等，2023），IFEval（Zhou 等，2023），FRAMES（Krishna 等，2024），GPQA Diamond（Rein 等，2023），SimpleQA（OpenAI，2024c），C-SimpleQA（He 等，2024），SWE-Bench Verified（OpenAI）上评估模型。

2024d）、Aider 1、LiveCodeBench（Jain 等，2024）（2024-08 – 2025-01）、Codeforces 2、中国全国高中数学奥林匹克（CNMO 2024）3，以及美国邀请数学考试 2024（AIME 2024）（MAA，2024）。除了标准基准外，我们还在开放式生成任务上评估我们的模型，使用 LLM 作为评判者。具体而言，我们遵循 AlpacaEval 2.0（Dubois 等，2024）和 Arena-Hard（Li 等，2024）的原始配置，这些配置利用 GPT-4-Turbo-1106 进行成对比较的评判。在这里，我们仅输入最终摘要进行评估，以避免长度偏差。对于精简模型，我们报告 AIME 2024、MATH-500、GPQA Diamond、Codeforces 和 LiveCodeBench 的代表性结果。

根据DeepSeek-V3的设置，使用simple-evals框架中的提示对标准基准测试如MMLU、DROP、GPQA Diamond和SimpleQA进行评估。对于MMLU-Redux，我们在零-shot设置中采用Zero-Eval提示格式（Lin, 2024）。在MMLU-Pro、C-Eval和CLUE-WSC方面，由于原始提示是few-shot的，我们稍微修改提示以适应零-shot设置。在few-shot中使用的CoT可能会影响DeepSeek-R1的性能。其他数据集按照其创建者提供的默认提示遵循原始评估协议。对于代码和数学基准，HumanEval-Mul数据集涵盖了八种主流编程语言（Python、Java、C++、C#、JavaScript、TypeScript、PHP和Bash）。在LiveCodeBench上的模型性能评估使用CoT格式，数据收集时间为2024年8月至2025年1月。Codeforces数据集采用来自10个Div.2竞赛的问题和专家精心制作的测试用例进行评估，随后计算预期等级和竞争者的百分比。SWE-Bench的验证结果通过无代理框架获得（Xia等人，2024）。与AIDER相关的基准使用“diff”格式进行测量。DeepSeek-R1的输出在每个基准上限制为最大32,768个标记。

基准线 我们针对多个强基准进行全面评估，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困难，我们基于官方报告提供其性能表现。对于蒸馏模型，我们还比较了开源模型QwQ-32B-Preview（Qwen，2024a）。

评估设置 我们将模型的最大生成长度设置为32,768个token。我们发现，使用贪婪解码来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的结果变异性显著。因此，我们默认使用pass@k评估（Chen et al.，2021），并报告使用非零温度的pass@1。具体来说，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个回应（通常在4到64之间，具体取决于测试集的大小）。然后计算pass@1。

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

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

1https://aider.chat 2https://codeforces.com
3https://www.cms.org.cn/Home/comp/comp/cid/12.html

3.1. DeepSeek-R1 评估

	基准 （度量）	Claude-3.5- Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	建筑	-	-	教育部	-	-	教育部
	# 激活的 Params	-	-	编号 378	-	-	编号 378
	# 总参数	-	-	671B 系列	-	-	671B 系列
英语	MMLU （Pass@1）	88.3	87.2	88.5	85.2	91.8	90.8
	MMLU-Redux （EM）	88.9	88.0	89.1	86.7	-	92.9
	MMLU-Pro （EM）	78.0	72.6	75.9	80.3	-	84.0
	DROP （3 次 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 钻石 （Pass@1） 65.0		49.9	59.1	60.0	75.7	71.5
	SimpleQA（正确）	28.4	38.2	24.9	7.0	47.0	30.1
	FRAMES （Acc.）	72.5	80.5	73.3	76.9	-	82.5
	AlpacaEval2.0 （LC-胜率）	52.0	51.1	70.0	57.8	-	87.6
	竞技场 （GPT-4-1106）	85.2	80.4	85.5	92.0	-	92.3
法典	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 （评级）	717	759	1134	1820	2061	2029
	SWE 验证 （已解决）	50.8	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot （Acc.）	45.3	16.0	49.6	32.9	61.7	53.3
数学 AIME 2024 （Pass@	数学 500 （Pass@1）	16.0	9.3	39.2	63.6	79.2	79.8
	CNMO 2024 （Pass@1）	78.3	74.6	90.2	90.0	96.4	97.3
中文 CLUEWSC （EM）		13.1	10.8	43.2	67.6	-	78.8
	C-Eval （EM）	85.4	87.9	90.9	89.9	-	92.8
	C-SimpleQA（正确）	76.7	76.0	86.5	68.9	-	91.8
		55.4	58.7	68.0	40.3	-	63.7

表4 | DeepSeek-R1与其他代表性模型的比较。

对于以教育为导向的知识基准，如MMLU、MMLU-Pro和GPQA Diamond，DeepSeek-R1相比于DeepSeek-V3表现出色。这个提升主要归功于在STEM相关问题上的更高准确性，通过大规模的强化学习实现了显著的进步。此外，DeepSeek-R1在FRAMES这个长期上下文依赖的问答任务上也表现优异，展示了其强大的文档分析能力。这突显了推理模型在人工智能驱动的搜索和数据分析任务中的潜力。在事实基准SimpleQA上，DeepSeek-R1超越了DeepSeek-V3，展现了其处理基于事实查询的能力。在这个基准上也观察到类似的趋势，即OpenAI-o1超过了GPT-4o。然而，DeepSeek-R1在中文SimpleQA基准上的表现不如DeepSeek-V3，这主要是由于其在安全强化学习后倾向于拒绝回答某些查询。如果没有安全强化学习，DeepSeek-R1的准确率可能超过70%。

DeepSeek-R1在IF-Eval上也表现出色，该基准旨在评估模型遵循格式指令的能力。这些改善可以归因于在监督微调（SFT）和强化学习（RL）训练的最后阶段引入了遵循指令的数据。此外，在AlpacaEval2.0和ArenaHard上观察到的显著表现表明DeepSeek-R1在写作任务和开放领域问答方面的优势。它显著超越了DeepSeek-V3，凸显了大规模强化学习的泛化优势，不仅提升了推理能力，还提高了在多个领域的表现。此外，DeepSeek-R1生成的摘要长度简洁，在ArenaHard上平均为689个词，在AlpacaEval 2.0上为2,218个字符。这表明。

DeepSeek-R1避免在基于GPT的评估中引入长度偏差，进一步巩固了它在多个任务中的鲁棒性。

在数学任务上，DeepSeek-R1的表现与OpenAI-o1-1217相当，远超其他模型。在编码算法任务中，如LiveCodeBench和Codeforces，注重推理的模型在这些基准测试中占据主导地位。在面向工程的编码任务中，OpenAI-o1-1217在Aider上优于DeepSeek-R1，但在SWE Verified上表现相当。我们相信，随着相关强化学习训练数据的增加，DeepSeek-R1在工程性能上将在下一个版本中有所提升。

3.2. 蒸馏模型评估

型	爱 2024		数学 500	GPQA 认证钻石	LiveCode 实时代码板凳	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	额定值
GPT-4o-0513 9.3		13.4	74.6	49.9	32.9	759
克劳德-3.5-十四行诗-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-迷你	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-预览版	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-蒸馏-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-蒸馏-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-蒸馏-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-蒸馏-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-蒸馏-骆驼-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-蒸馏-骆驼-70B	70.0	86.7	94.5	65.2	57.5	1633

表 5 |DeepSeek-R1 蒸馏模型与其他类似模型的比较
与推理相关的基准测试。

如表5所示，仅仅对DeepSeek-R1的输出进行蒸馏，就能够让高效的DeepSeek-R1-7B（即DeepSeek-R1-Distill-Qwen-7B，下面简要表示）在各项评估指标上超越非推理模型如GPT-4o-0513。DeepSeek-R1-14B在所有评估指标上超过了QwQ-32B-Preview，而DeepSeek-R1-32B和DeepSeek-R1-70B在大多数基准测试中显著超越o1-mini。这些结果展示了蒸馏的强大潜力。此外，我们发现对这些蒸馏模型应用强化学习可以带来显著的进一步提升。我们认为这值得进一步探索，因此在此仅呈现简单的SFT蒸馏模型的结果。

4. 讨论

4.1. 蒸馏 vs. 强化学习

在第3.2节中，我们可以看到通过对DeepSeek-R1进行蒸馏，小模型能够取得令人印象深刻的结果。然而，仍然有一个问题：没有蒸馏的情况下，模型能否通过论文中讨论的大规模强化学习训练实现可比的性能？

为了解答这个问题，我们使用数学、代码和STEM数据对Qwen-32B-Base进行了大规模强化学习训练，训练超过了10000步，结果得到了DeepSeek-R1-Zero-Qwen-32B。实验结果如表6所示，表明经过大规模训练后，32B基础模型的表现。

型	爱 2024		数学 500	GPQA 钻石	LiveCodeBench 函数
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-预览版	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-零-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-蒸馏-Qwen-32B	72.6	83.3	94.3	62.1	57.2

表6 | 蒸馏模型与强化学习模型在推理相关基准上的比较。

RL训练的表现达到与QwQ-32B-Preview相当。然而，从DeepSeek-R1提炼出来的DeepSeek-R1-Distill-Qwen-32B在所有基准测试中表现显著优于DeepSeek-R1-Zero-Qwen-32B。

因此，我们可以得出两个结论：首先，将更强大的模型蒸馏成较小的模型能取得优秀的结果，而依赖于本文提到的大规模强化学习的小模型需要巨大的计算能力，甚至可能无法达到蒸馏的效果。其次，尽管蒸馏策略既经济又有效，但超越智能的界限可能仍需要更强大的基础模型和更大规模的强化学习。

4.2. 尝试失败

在DeepSeek-R1开发的早期阶段，我们也遇到了失败和挫折。我们在这里分享我们的失败经验，以提供一些见解，但这并不意味着这些方法无法开发出有效的推理模型。

过程奖励模型（PRM）是一种合理的方法，可以引导模型朝着更好的解决推理任务的方法前进（Lightman et al., 2023；Uesato et al., 2022；Wang et al., 2023）。然而，在实践中，PRM存在三个主要局限性，可能会妨碍其最终成功。首先，在一般推理中，明确地定义一个细粒度的步骤是具有挑战性的。其次，确定当前的中间步骤是否正确也是一项具有挑战性的任务。使用模型进行自动标注可能无法产生令人满意的结果，而手动标注则不利于规模化。第三，一旦引入基于模型的PRM，就不可避免地会导致奖励黑客行为（Gao et al., 2022），而重新训练奖励模型需要额外的训练资源，并使整个训练流程变得复杂。总之，尽管PRM在重新排序模型生成的前N个响应或辅助引导搜索方面表现出良好的能力（Snell et al., 2024），但与其在我们实验中引入的大规模强化学习过程中的额外计算开销相比，其优势是有限的。

受到AlphaGo（Silver等人，2017b）和AlphaZero（Silver等人，2017a）的启发，我们探索了使用蒙特卡洛树搜索（MCTS）来增强测试时的计算可扩展性。这种方法涉及将答案分解成更小的部分，以便模型系统地探索解决方案空间。为此，我们提示模型生成多个标签，这些标签对应于搜索所需的特定推理步骤。在训练过程中，我们首先使用收集到的提示，通过受训的价值模型引导MCTS找到答案。随后，我们使用生成的问答对来训练演员模型和价值模型，迭代地完善这一过程。

然而，这种方法在扩大训练规模时遇到了几个挑战。首先，与国际象棋不同，国际象棋的搜索空间相对明确，而令牌生成则呈现出一个

为了应对指数级增长的搜索空间，我们对每个节点设置了最大扩展限制，但这可能导致模型陷入局部最优解。其次，价值模型直接影响生成的质量，因为它指导着搜索过程的每一步。训练一个细粒度的价值模型本质上是困难的，这使得模型迭代改进变得具有挑战性。尽管AlphaGo的核心成功依赖于训练一个价值模型以逐步提高其性能，但由于标记生成的复杂性，这一原则在我们的设置中难以复制。

总之，尽管与预训练的价值模型结合使用时，MCTS可以在推理过程中提高性能，但通过自我搜索迭代地提升模型性能仍然是一个重大挑战。

5. 结论、局限性和未来工作

在这项工作中，我们分享了我们通过强化增强模型推理能力的历程
学习。 DeepSeek-R1-Zero 代表了一种不依赖冷启动的纯 RL 方法
数据，在各种任务中实现强大的性能。DeepSeek-R1 功能更强大，利用冷启动数据以及迭代 RL 微调。最终，DeepSeek-R1 在一系列任务上实现了与 OpenAI-o1-1217 相当的性能。

我们进一步探索了将推理能力提炼到小型密集模型。我们使用 DeepSeek-R1 作为教师模型来生成 800K 训练样本，并对几个小型密集模型进行了微调。结果是有希望的：DeepSeek-R1-Distill-Qwen-1.5B 在数学基准测试中优于 GPT-4o 和 Claude-3.5-Sonnet，在 AIME 上为 28.9%，在 MATH 上为 83.9%。其他密集模型也取得了令人印象深刻的结果，明显优于基于相同底层检查点的其他教学调整模型。

未来，我们计划投资于 DeepSeek-R1 的以下方向的研究。

• 通用功能：目前，DeepSeek-R1 在函数调用、多轮次、复杂角色扮演和 JSON 输出等任务方面的能力不如 DeepSeek-V3。展望未来，我们计划探索可以利用 CoT 来增强这些领域的任务多长时间。

• 语言混合：DeepSeek-R1 目前针对中文和英文进行了优化，这可能会导致在处理其他语言的查询时出现语言混合问题。例如，DeepSeek-R1 可能会使用英语进行推理和响应，即使查询使用的是英语或中文以外的语言。我们的目标是在未来的更新中解决此限制。

• 提示工程：在评估 DeepSeek-R1 时，我们观察到它对提示很敏感。小样本提示会持续降低其性能。因此，我们建议用户直接描述问题并使用零样本设置指定输出格式以获得最佳结果。

• 软件工程任务：由于评估时间长，影响了 RL 过程的效率，大规模 RL 尚未在软件工程任务中得到广泛应用。因此，DeepSeek-R1 在软件工程基准测试中没有表现出比 DeepSeek-V3 有的巨大改进。未来的版本将通过对软件工程数据实施拒绝抽样或在 RL 过程中结合异步评估来解决这个问题，以提高效率。

引用

AI@Meta。调用 3.1 模型卡，2024 年。网址 https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md 的

人类的。克劳德 3.5 十四行诗，2024 年。网址 https://www.anthropic.com/news/claude-3-5-十四行诗。

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， H. 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， M. Plappert， F. Chantz是，E. Barnes， A. Herbert-Voss， WH Guss， A. Nichol， A. Paino， N. Tezak， J. Tang， I. Babuschkin， S. Balaji， S. Jain， W. Saunders， C. Hesse， AN 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。M. Chen， J. Tworek， H. Jun， Q. Yuan， H. P. de Oliv.eira Pinto， J. Kaplan， H. Ed.沃德， Y. Burda， N. Joseph， G. Brockm

A. Dubey， A. Jauhri， A. Pandey， A. Kadian， A. Al-Dahle， A. Letman， A. Mathur， A. Schelten， A. Yang， A. Fan， et al.美洲驼 3 群模型。arXiv 预印本 arXiv：2407.21783， 2024。

Y. Dubois、B. Galambosi、P. Liang 和 T. B. Hashimoto。长度控制羊驼：一个简单的消除自动评估器偏差的方法。arXiv 预印本 arXiv：2404.04475,2024 年。

X. Feng， Z. Wan， M. 温， SM McAleer， Y. 温， W. Zhang， 和 J. Wang.类似 alphazero 的树搜索可以指导大型语言模型解码和训练， 2024.网址 https：//arxiv.org/abs/2309.17179.

L. Gao、J. Schulman 和 J. Hilton。奖励模型过度优化的缩放定律，2022 年。网址 https://arxiv.org/abs/2210.10760。

AP Gema、JO Leang、G. Hong、A. Devoto、ACM Mancino、R. Saxena、X. He、Y. Zhao、X. Du、MRG Madani、C. Barale、R. McHardy、J. Harris、J. Kaddour、E. van Krieken 和 P. Minervini。我们完成了 mmlu 吗？CoRR，abs/2406.04127,2024 年。网址 https://doi.org/10.48550/arXiv.2406.04127。

谷歌。 我们的下一代型号：Gemini 1.5,2024 年。网址 https://blog.google/technology/ai/google-gemini-next-generation-model-2024 年 2 月。

Y. He， S. Li， J. Liu， Y. Tan， W. Wang， H. Huang， X. Bu， H. Guo， C. 胡， B. Zheng， et al.中文 simpleqa：大型语言模型的中文事实性评估。arXiv 预印本 arXiv：2411.07140， 2024。

D. Hendrycks、C. Burns、S. Basart、A. Zou、M. Mazeika、D. Song 和 J. Steinhardt。 测量海量多任务语言理解。arXiv 预印本 arXiv：2009.03300,2020 年。

Y. Huang， Y. Bai， Z. Zhu， J. Zhang， J. Zhang， T. Su， J. Liu， C. Lv， Y. Zhang， J. Lei， et al. C-Eval：用于基础模型的多级多学科中文评估套件。arXiv 预印本 arXiv：2305.08322， 2023。

N. Jain、K. Han、A. Gu、W. Li、F. Yan、T. Zhang、S. Wang、A. Solar-Lezama、K. Sen 和 I. Stoica。Livecodebench：代码大型语言模型的整体和无污染评估。CoRR，abs/2403.07974,2024 年。URL https://doi.org/10.48550/arXiv.2403.07974。

S. Krishna、K. Krishna、A. Mohananey、S. Schwarcz、A. Stambler、S. Upadhyay 和 M. Faruqui。事实、获取和原因：检索增强生成的统一评估。CoRR，abs/2409.12941,2024 年。doi：10.48550/ARXIV.2409.12941。网址 https://doi.org/10.485 50/arXiv.2409.12941。

A. Kumar， V. Zhuang， R. Agarwal， Y. Su， JD Co-Reyes， A. Singh， K. Baumli， S. Iqbal， C. Bishop， R. Roelofs， et al.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。

H. Li， Y. Zhang， F. Koto， Y. Yang， H. Zhao， Y. Gong， N. Duan， 和 T. Baldwin.CMMLU：衡量中文中的大量多任务语言理解。arXiv 预印本 arXiv：2306.09212，2023。

李 T. Li， WLChiang， E. Frick， L. Dunlap， T. Wu， B. Zhu， JE Gonzalez 和 I. Stoica。从众包数据到高质量基准：Arena-hard 和 benchbuilder 管道。arXiv 预印本 arXiv：2406.11939,2024 年。

H. Lightman， V. Kosaraju， Y. Burda， H. Edwards， B. Baker， T. Lee， J. Leike， J. Schulman， I. Sutskever 和 K. Cobbe。让我们一步一步验证。arXiv 预印本 arXiv：2305.20050,2023 年。

B. Y. Lin. ZeroEval：评估语言模型的统一框架，2024 年 7 月。网址 https://github.com/WildEval/ZeroEval。

马阿。美国数学邀请赛 - AIME。在美国数学邀请赛考试 - AIME 2024 中，2024 年 2 月。网址 https://maa.org/math -competitions/american-invitational-mathematics-examination-aime。打开人工智能。你好 GPT-4o，2024a. 网址 https://openai.com/index/hello-gpt-4o/。打开人工智能。学习使用 llms 推理，2024b. 网址 https://openai.com/index/learnin

g-to-reason-with-llms/的 URL 中。开放人工智能。介绍 SimpleQA，2024c。URL https://openai.com/index/introducing -simpleqa/ 中。

开放人工智能。介绍 SWE-bench 验证 我们将发布更多经过人工验证的 swe- bench 子集，2024d. URL https://openai.com/index/introducing-swe-bench -verified/。

Qwen.Qwq：深刻反思未知的界限，2024a。网址 https://qwenlm.github.io/blog/qwq-32b-preview/.Qwen.Qwen2.5： 基础模型的聚会，2024b。网址 https://qwenlm.github.io/blog/qwen2.5 的

D. Rein、BL Hou、AC Stickland、J. Petty、RY Pang、J. Dirani、J. Michael 和 SR Bowman。GPQA：研究生水平的谷歌证明问答基准。arXiv 预印本 arXiv：2311.12022,2023 年。

Z. Shao， P. Wang， Q. Zhu， R. Xu， J. Song， M. Zhang， Y. Li， Y. Wu， 和 D. Guo.Deepseekmath：在开放语言模型中突破数学推理的极限。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、TP Lillicrap、K. Simonyan 和 D. Hassabis。使用通用强化学习算法通过自博掌握国际象棋和将棋。CoRR，abs/1712.01815,2017a. 网址 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, TP Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, 和 D. Hassabis.在没有人类知识的情况下掌握围棋游戏。自然, 550（7676）：354–359, 2017b. doi: 10.1038/NATURE24270.网址 https://doi.org/10.1038/nature24270。

C. Snell、J. Lee、K. Xu 和 A. Kumar。以最佳方式扩展 llm 测试时间计算可能比扩展模型参数更有效，2024 年。网址 https://arxiv.org/abs/2408.033 14。

T. Trinh、Y. Wu、Q. Le、H.He 和 T. Luong。在没有人类的情况下求解奥林匹克几何
示威。 自然, 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。通过基于过程和结果的反馈解决数学单词问题。arXiv 预印本 arXiv：2211.14275,2022 年。

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, 和 Z. Sui.数学牧羊人：数学推理中 llms 的无标签分步验证器。arXiv 预印本 arXiv：2312.08935, 2023。

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, 和 D. 周.自洽性改善了语言模型中的思维链推理。arXiv 预印本 arXiv：2203.11171, 2022.

Y. Wang, X. 马, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. 任, A. Arulraj, X. He, Z. 江, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, 和 W. Chen.Mmlu-pro：一种更强大且更具挑战性的多任务语言理解基准。CoRR, abs/2406.01574, 2024.网址 https://doi.org/10.48550/arXiv.2406.01574。

C. S. Xia、Y. 邓、S. Dunn 和 L. Zhang。无代理：揭秘基于 LLM 的软件
工程代理。arXiv 预印本，2024 年。

H. Xin, Z. Z. 任, 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.网址 https://arxiv.org/abs/2408.08152.

周建、卢淑、S. 米什拉、梵天、巴苏、栾妍、周大和侯立。指令跟随
大型语言模型的评估。arXiv 预印本 arXiv：2311.07911,2023 年。

附录

A. 贡献和鸣谢

主要贡献者： 郭达雅， 德建， 杨浩伟， 张俊霄， 宋若宇， 张润欣， 徐启浩， 朱世荣， 马培义， 王小碧

张小康， 张兴凯， 于宇， 吴志峰， 吴志斌， 苟志宏， 邵卓树， 李子怡， 高子怡

贡献者：爱欣、刘冰、薛冰、冰轩、王伯超、吴北、冯成达、卢成刚、赵成奇、邓崇阮、戴大麦、陈东杰、纪尔航、李方云、林福聪、戴梅斯、罗光波、郝冠廷、陈国伟、李 H. 张汉伟、徐洪辉、丁华作、高辉曲

惠 李建中 郭佳时 李景昌 陈景阳 袁金浩 涂俊杰 邱俊龙 李俊龙 蔡佳琪 倪建梁 金 陈凯 董凯 胡* 开超 尤凯歌 高康 关克欣 黄葵 于莱恩 王乐聪 张亮 赵立通 王立月 张雷 徐乐怡 夏明川 张明华 张明辉 唐明旭 周孟丽

王妙军 李明明 田盼盼 黄鹏 张倩成 王钦宇 陈秋石 杜瑞琪 葛瑞松* 张瑞松 张瑞哲 潘润吉 王 R.J. 陈 R.L. 金

陈睿 尚浩 卢尚安
徐善黄 陈胜峰 易
世玉 王淑英 于顺
峰 徐射击 潘S.S,
李

双周 少青 吴胜
峰 叶涛 云田培
天宇 孙T. 王望
丁 曾温 刘文峰
梁文军 高文勤
于文涛* 张文涛
小伟 安晓东 刘
晓汉 王小康 陈
小涛 聂欣 程欣
刘欣欣 谢兴超
刘欣宇 杨欣元
李学成 苏旭恒
林晓Q. 李翔月
金小金 沈小莎
陈晓文 孙晓翔
王欣南 宋欣义
周贤祖 王欣霞
单 李

Y.H. 魏 杨 张彦
宏 徐彦宏 李姚
赵耀峰 孙耀辉
王毅 于超 张一
凡 石一亮 熊英
何一诗 飘一松
王义轩 谭义阳
马* 刘永强 郭元
欧玉端 王跃 龚
宇恒 邹玉佳 何
云帆 熊玉贤g 罗
玉祥 尤玉轩 刘
玉阳 周 朱燕平
黄耀辉 李毅 郑
玉辰 朱云贤 马
英 唐玉坤 扎玉
婷 闫Z.Z. 任泽
辉 任 张丽 沙哲
傅哲安 徐振达
谢正岩 张哲文
郝志成 马志刚
严志玉 吴子慧顾

王耀清

Zijia Zhu
Zijun Liu*
Zilin Li Ziwei
谢紫阳 宋子政
Pan
黄臻 黄志鹏 徐忠
玉 张臻 张臻

在每个角色中，作者按名字的字母顺序列出。标有 * 的姓名表示已离开我们团队的个人。