

DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning

<https://doi.org/10.1038/s41586-025-09422-z>

Received: 14 February 2025

Accepted: 17 July 2025

Published online: 17 September 2025

Open access

 Check for updates

General reasoning represents a long-standing and formidable challenge in artificial intelligence (AI). Recent breakthroughs, exemplified by large language models (LLMs)^{1,2} and chain-of-thought (CoT) prompting³, have achieved considerable success on foundational reasoning tasks. However, this success is heavily contingent on extensive human-annotated demonstrations and the capabilities of models are still insufficient for more complex problems. Here we show that the reasoning abilities of LLMs can be incentivized through pure reinforcement learning (RL), obviating the need for human-labelled reasoning trajectories. The proposed RL framework facilitates the emergent development of advanced reasoning patterns, such as self-reflection, verification and dynamic strategy adaptation. Consequently, the trained model achieves superior performance on verifiable tasks such as mathematics, coding competitions and STEM fields, surpassing its counterparts trained through conventional supervised learning on human demonstrations. Moreover, the emergent reasoning patterns exhibited by these large-scale models can be systematically used to guide and enhance the reasoning capabilities of smaller models.

Reasoning capability, the cornerstone of human intelligence, enables complex cognitive tasks ranging from mathematical problem-solving to logical deduction and programming. Recent advances in AI have demonstrated that LLMs can exhibit emergent behaviours, including reasoning abilities, when scaled to a sufficient size^{4,5}. However, achieving such capabilities in pre-training typically demands substantial computational resources. In parallel, a complementary line of research has demonstrated that LLMs can be effectively augmented through CoT prompting. This technique, which involves either providing carefully designed few-shot examples or using minimalist prompts such as “Let’s think step by step”^{3,6}, enables models to produce intermediate reasoning steps, thereby substantially enhancing their performance on complex tasks. Similarly, further performance gains have been observed when models learn high-quality, multistep reasoning trajectories during the post-training phase^{2,7}. Despite their effectiveness, these approaches exhibit notable limitations. Their dependence on human-annotated reasoning traces slows scalability and introduces cognitive biases. Furthermore, by constraining models to replicate human thought processes, their performance is inherently capped by the human-provided exemplars, which prevents the exploration of superior, non-human-like reasoning pathways.

To tackle these issues, we aim to explore the potential of LLMs for developing reasoning abilities through self-evolution in a RL framework, with minimal reliance on human labelling efforts. Specifically, we build on DeepSeek-V3 Base⁸ and use Group Relative Policy Optimization (GRPO)⁹ as our RL framework. The reward signal is only based on the correctness of final predictions against ground-truth answers, without imposing constraints on the reasoning process itself. Notably, we bypass the conventional supervised fine-tuning (SFT) phase before RL training. This design choice originates from our hypothesis that human-defined reasoning patterns may limit model exploration,

whereas unrestricted RL training can better incentivize the emergence of new reasoning capabilities in LLMs. Through this process, detailed in the next section, our model (referred to as DeepSeek-R1-Zero) naturally developed diverse and sophisticated reasoning behaviours. To solve reasoning problems, the model exhibits a tendency to generate longer responses, incorporating verification, reflection and the exploration of alternative approaches within each response. Although we do not explicitly teach the model how to reason, it successfully learns improved reasoning strategies through RL.

Although DeepSeek-R1-Zero demonstrates excellent reasoning capabilities, it faces challenges such as poor readability and language mixing, occasionally combining English and Chinese in a single CoT response. Furthermore, the rule-based RL training stage of DeepSeek-R1-Zero is narrowly focused on reasoning tasks, resulting in limited performance in broader areas such as writing and open-domain question answering. To address these challenges, we introduce DeepSeek-R1, a model trained through a multistage learning framework that integrates rejection sampling, RL and supervised fine-tuning, detailed in the ‘DeepSeek-R1’ section. This training pipeline enables DeepSeek-R1 to inherit the reasoning capabilities of its predecessor, DeepSeek-R1-Zero, while aligning model behaviour with human preferences through further non-reasoning data.

To enable broader access to powerful AI at a lower energy cost, we have distilled several smaller models and made them publicly available. These distilled models exhibit strong reasoning capabilities, surpassing the performance of their original instruction-tuned counterparts. We believe that these instruction-tuned versions will also greatly contribute to the research community by providing a valuable resource for understanding the mechanisms underlying long CoT reasoning models and for promoting the development of more powerful reasoning models. We release DeepSeek-R1-Zero, DeepSeek-R1, data samples and distilled models to the public as described in the ‘Code availability’ section.

A list of authors and their affiliations appears at the end of the paper.

DeepSeek-R1 通过强化学习激励大语言模型进行推理

<https://doi.org/10.1038/s41586-025-09422-z>

收稿日期：2025年2月14日

接受日期：2025年7月17日

在线发表：2025年9月17日

开放获取

 检查更新

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。近期 AI 领域的进展表明，当 LLM 扩展到足够规模时，能够表现出涌现行为，包括推理能力^{4,5}。然而，在预训练阶段获得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 CoT 提示可以有效增强 LLM。这种技术可以提供精心设计的少样本示例，也可以使用“让我们一步一步地思考”这类简约提示^{3,6}，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。同样的，在后训练阶段学习高质量、多步推理轨迹时，也观察到了进一步的性能提升^{2,7}。尽管这些方法行之有效，但仍存在显著局限。它们依赖人工标注的推理轨迹，这减缓了规模化进程，并引入了认知偏差。此外，由于模型被限制为复现人类思维过程，其性能必然受限于人工提供的示例，因而无法探索更优且非人类式的推理路径。

为解决这些问题，我们旨在探索在强化学习框架中通过自我进化来培养推理能力、且尽可能减少对人工标注依赖的大语言模型（LLM）的潜力。具体而言，我们以 DeepSeek-V3 Base⁸为基础，并采用群体相对策略优化（GRPO）⁹作为强化学习框架。奖励信号仅依据最终预测与标准答案的一致性，不对推理过程本身施加限制。值得注意的是，我们跳过了强化学习训练前常规的监督微调（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会限制模型的探索，

作者及其所属机构列表现论文末尾

而不受限制的强化学习训练能够更好地激励大语言模型中涌现新的推理能力。通过这一过程（详见下一节），我们的模型（称为 DeepSeek-R1-Zero）自然发展出多样且复杂的推理行为。为解决推理问题，模型倾向于生成更长的回答，并在每个回答中融入验证、反思以及对其他方法的探索。尽管我们没有明确教模型如何推理，它仍通过强化学习成功学会了更好的推理策略。

尽管 DeepSeek-R1-Zero 展现出出色的推理能力，但它仍面临可读性较差、语言混用等挑战，偶尔会在单条 CoT 回复中混合使用英语和中文。此外，DeepSeek-R1-Zero 的基于规则的 RL 训练阶段专注于推理任务，因此在写作和开放域问答等更广泛领域的表现有限。为应对这些挑战，我们推出 DeepSeek-R1。这是一个通过多阶段学习框架训练的模型，该框架整合了拒绝采样、RL 和监督微调，详情见“DeepSeek-R1”部分。该训练流程使 DeepSeek-R1 能够继承其前身 DeepSeek-R1-Zero 的推理能力，同时通过补充非推理数据，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能够使用强大的 AI，我们蒸馏了若干较小的模型，并将其公开。这些蒸馏模型展现出强大的推理能力，性能超过了其原始指令微调版本。我们认为，这些指令微调版本也将为研究社区作出重要贡献，为理解长 CoT 推理模型的底层机制提供宝贵资源，并推动更强大推理模型的发展。我们按照“代码可用性”一节所述，向公众发布 DeepSeek-R1-Zero、DeepSeek-R1、数据样本和蒸馏模型。

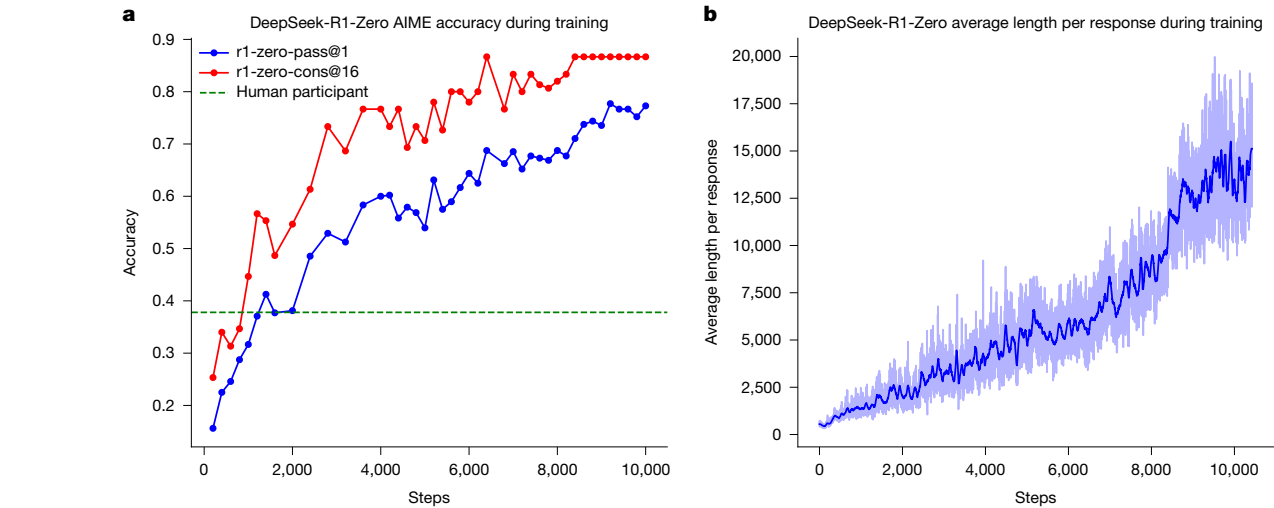


Fig. 1 | Accuracy and output length of DeepSeek-R1-Zero throughout the training process. a, AIME accuracy of DeepSeek-R1-Zero during training. AIME takes a mathematical problem as input and a number as output, illustrated in Extended Data Table 1. pass@1 and cons@16 are described in Supplementary Information, section 4.1. The baseline is the average score achieved by human

DeepSeek-R1-Zero

To implement large-scale RL of DeepSeek-R1-Zero, we use a highly efficient RL pipeline. Specifically, we use GRPO⁹ as our RL algorithm, described in Methods section ‘GRPO’. Furthermore, we use a rule-based reward system to compute accuracy and format rewards, with detailed methodologies outlined in Methods section ‘Reward design’. Furthermore, our high-performance RL infrastructure is described in Supplementary Information, section 2.1, ensuring scalable and efficient training.

Specifically, we apply the RL technique on the DeepSeek-V3 Base to train DeepSeek-R1-Zero. During training, we design a straightforward template to require DeepSeek-R1-Zero to first produce a reasoning process, followed by the final answer. The prompt template is written as below.

“A conversation between User and Assistant. The User asks a question and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the answer. The reasoning process and answer are enclosed within <think>...</think> and <answer>...</answer> tags, respectively, that is, <think> reasoning process here </think><answer> answer here </answer>. User: prompt. Assistant:”, in which the prompt is replaced with the specific reasoning question during training. We intentionally limit our constraints to this structural format, avoiding any content-specific biases to ensure that we can accurately observe the natural progression of the model during the RL process.

Figure 1a shows the performance trajectory of DeepSeek-R1-Zero on the American Invitational Mathematics Examination (AIME) 2024 benchmark throughout the RL training process, in which the average pass@1 score on AIME 2024 shows a marked increase, jumping from an initial value of 15.6% to 77.9%. Also, by using the self-consistency decoding¹⁰, the performance of the model can be further improved, achieving an accuracy of 86.7%. This performance greatly surpasses the average performance across all human competitors of the AIME. Besides the maths competitions, as shown in Supplementary Fig. 8, DeepSeek-R1-Zero also achieves remarkable performance in coding competitions and graduate-level biology, physics and chemistry problems. These results underscore the effectiveness of RL in enhancing the reasoning capabilities of LLMs.

participants in the AIME competition. **b**, The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time. Note that a training step refers to a single policy update operation.

As well as the progressive enhancement of reasoning capabilities during training, DeepSeek-R1-Zero also demonstrates self-evolutionary behaviour with RL training. As shown in Fig. 1b, DeepSeek-R1-Zero exhibits a steady increase in thinking time throughout training, driven only by intrinsic adaptation rather than external modifications. Making use of long CoT, the model progressively refines its reasoning, generating hundreds to thousands of tokens to explore and improve its problem-solving strategies.

The increase in thinking time helps with the autonomous development of sophisticated behaviours. Specifically, DeepSeek-R1-Zero increasingly exhibits advanced reasoning strategies such as reflective reasoning and systematic exploration of alternative solutions provided in Extended Data Fig. 1a, substantially boosting its performance on verifiable tasks such as maths and coding. Notably, during training, DeepSeek-R1-Zero exhibits an ‘ahamoment’, shown in Table 1, characterized by a sudden increase in the use of the word ‘wait’ during reflections, provided in Extended Data Fig. 1b. This moment marks a distinct change in reasoning patterns and clearly shows the self-evolution process of DeepSeek-R1-Zero.

The self-evolution of DeepSeek-R1-Zero underscores the power and beauty of RL: rather than explicitly teaching the model how to solve a problem, we simply provide it with the right incentives and it autonomously develops advanced problem-solving strategies. This serves as a reminder of the potential of RL to unlock higher levels of capabilities in LLMs, paving the way for more autonomous and adaptive models in the future.

DeepSeek-R1

Although DeepSeek-R1-Zero exhibits strong reasoning capabilities, it faces several issues. DeepSeek-R1-Zero struggles with challenges such as poor readability and language mixing, as DeepSeek-V3 Base is trained on several languages, especially English and Chinese. To address these issues, we develop DeepSeek-R1, whose pipeline is illustrated in Fig. 2. In the initial stage, we collect thousands of cold-start data that exhibit a conversational, human-aligned thinking process, as detailed in Supplementary Information, section 2.3.2. RL training is then applied with hyperparameters in Methods section ‘Training details of the first RL stage’, data details in Supplementary Information, section 2.3.1, to improve the model performance with the

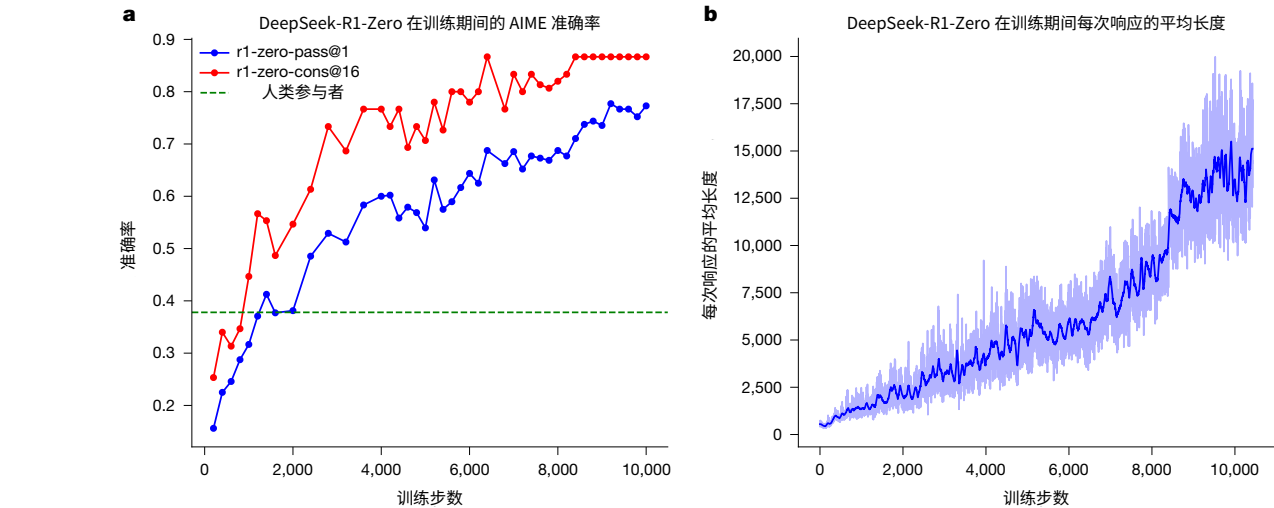


图 1 | DeepSeek-R1-Zero 在整个训练过程中的准确率和输出长度。a, DeepSeek-R1-Zero 在训练期间的 AIME 准确率。AIME 以一道数学题作为输入，并输出一个数字，如扩展数据表 1 所示。pass@1 和 cons@16 的说明见补充信息第 4.1 节。基线为人类参赛者在 AIME 竞赛中取得的平均分数。

DeepSeek-R1-Zero

为实现 DeepSeek-R1-Zero 的大规模 RL，我们采用了高效的 RL 流程。具体而言，我们使用 GRPO⁹ 作为我们的 RL 算法，相关说明见方法部分 “GRPO” 一节。此外，我们采用基于规则的奖励系统来计算准确率奖励和格式奖励，详细方法见方法部分 “奖励设计” 一节。此外，我们的高性能 RL 基础设施见补充信息第 2.1 节，以确保训练具有可扩展性和高效率。

具体而言，我们将强化学习（RL）技术应用于 DeepSeek-V3 Base，以训练 DeepSeek-R1-Zero。在训练过程中，我们设计了一个简单的模板，要求 DeepSeek-R1-Zero 先生成推理过程，然后给出最终答案。提示模板如下。

“用户与助手之间的一段对话。用户提出问题，助手负责解答。助手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> and <answer>...</answer> tags 中，也就是说，<think> reasoning 过程在此处 </think><answer> answer 答案在此处 </answer>。用户：提示。助手：”，其中提示会在训练过程中替换为具体的推理问题。我们有意将约束限制在这种结构形式上，避免引入任何特定内容的偏差，以确保能够准确观察模型在强化学习（RL）过程中的自然进展。

图 1a 展示了 DeepSeek-R1-Zero 在 RL 训练过程中于 2024 年美国数学邀请赛（AIME）基准上的性能变化轨迹，其中 AIME 2024 的平均 pass@1 分数显著提升，从最初的 15.6% 跃升至 77.9%。此外，通过采用自一致性解码¹⁰，模型的性能还可以进一步提升，准确率达到 86.7%。这一表现远超 AIME 所有人类参赛者的平均水平。除了数学竞赛之外，如补充图 8 所示，DeepSeek-R1-Zero 在编程竞赛以及研究生级别的生物学、物理学和化学问题上也取得了显著表现。这些结果凸显了 RL 在增强 LLM 推理能力方面的有效性。

b, RL 过程中 DeepSeek-R1-Zero 在训练集上的平均响应长度。DeepSeek-R1-Zero 会自然地学会通过投入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。

除了训练期间推理能力逐步增强之外，DeepSeek-R1-Zero 在 RL 训练过程中还展现出自我进化行为。如图 1b 所示，DeepSeek-R1-Zero 在整个训练过程中思考时间稳步增加，这一变化仅由内在适应驱动，而非外部修改。借助长 CoT，模型逐步改进其推理过程，生成数百至数千个 token，以探索并改进其问题求解策略。

思考时间的增加有助于复杂行为的自主发展。具体而言，DeepSeek-R1-Zero 越来越多地展现出先进的推理策略，例如扩展数据图 1a 所示的反思推理和对备选解决方案的系统探索，从而显著提升了其在数学和编程等可验证任务上的表现。值得注意的是，在训练过程中，DeepSeek-R1-Zero 出现了表 11 中所示的“顿悟时刻”，其特点是在反思过程中“wait”一词的使用突然增加（见扩展数据图 1b）。这一时刻标志着推理模式发生了显著变化，并清楚展现了 DeepSeek-R1-Zero 的自我进化过程。

DeepSeek-R1-Zero 的自我进化凸显了强化学习的力量与魅力：我们不是明确地教模型如何解决问题，而只是为它提供恰当的激励，它便自主发展出先进的问题解决策略。这提醒我们，强化学习有潜力释放大语言模型更高水平的能力，并为未来更自主、更具适应性的模型铺平道路。

DeepSeek-R1

尽管 DeepSeek-R1-Zero 展现出强大的推理能力，但仍面临若干问题。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 在可读性较差和语言混用等方面存在困难。为解决这些问题，我们开发了 DeepSeek-R1，其流程如图 2 所示。在初始阶段，我们收集了数千条冷启动数据，这些数据展现了对话式、符合人类意图的思维过程，详情见补充信息第 2.3.2 节。随后采用强化学习训练，超参数见方法部分 “第一阶段强化学习训练细节”，数据详情见补充信息第 2.3.1 节，以通过对话式思维过程和语言一致性提升模型性能。

Table 1 | An interesting ‘aha moment’ of an intermediate version of DeepSeek-R1-Zero

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to
Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...
$(\sqrt{a - \sqrt{a + x}})^2 = x^2 \Rightarrow a - \sqrt{a + x} = x^2$.
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \Rightarrow a^2 - 2ax^2 + (x^2)^2 = a + x \Rightarrow x^4 - 2ax^2 - x + (a^2 - a) = 0$
...
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step by step to identify whether the correct sum can be ...
We started with the equation:
$\sqrt{a - \sqrt{a + x}} = x$
First, let's square both sides:
$a - \sqrt{a + x} = x^2 \Rightarrow \sqrt{a + x} = a - x^2$
Next, I could square both sides again, treating the equation: ...
...
The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of RL.

conversational thinking process and language consistency. Subsequently, we apply rejection sampling and SFT once more. This stage incorporates both reasoning and non-reasoning datasets into the SFT process, as detailed in Supplementary Information, section 2.3.3, enabling the model to not only excel in reasoning tasks but also demonstrate advanced writing capabilities. To further align the model with human preferences, we implement a secondary RL stage designed to enhance the helpfulness and harmlessness of the model while simultaneously refining its reasoning capabilities. The reward model is described in Methods section ‘Reward design’ and RL hyperparameters are in Methods section ‘Training details of the second RL stage’. The total training cost is listed in Supplementary Information, section 2.4.4.

We evaluate our models on MMLU¹¹, MMLU-Redux¹², MMLU-Pro¹³, DROP¹⁴, C-Eval¹⁵, IFEval¹⁶, FRAMES¹⁷, GPQA Diamond¹⁸, SimpleQA¹⁹, C-SimpleQA²⁰, CLUEWSC²¹, AlpacaEval 2.0 (ref. 22), Arena-Hard²³, SWE-bench Verified²⁴, Aider-Polyglot²⁵, LiveCodeBench²⁶ (2024-08–2025-01), Codeforces²⁷, Chinese National High School Mathematics

Olympiad (CNMO 2024)²⁸ and AIME 2024 (ref. 29). The details of these benchmarks are provided in Supplementary Tables 15–29.

Table 2 summarizes the performance of DeepSeek-R1 across several developmental stages, as outlined in Fig. 2. A comparison between DeepSeek-R1-Zero and DeepSeek-R1 Dev1 reveals substantial improvements in instruction-following, as evidenced by higher scores on the IF-Eval and Arena-Hard benchmarks. However, owing to the limited size of the cold-start dataset, Dev1 exhibits a partial degradation in reasoning performance compared with DeepSeek-R1-Zero, most notably on the AIME benchmark. By contrast, DeepSeek-R1 Dev2 demonstrates marked performance enhancements on benchmarks that require advanced reasoning skills, including those focused on code generation, mathematical problem solving and STEM-related tasks. Benchmarks targeting general-purpose tasks, such as AlpacaEval 2.0, show marginal improvement. These results indicate that reasoning-oriented RL considerably enhances reasoning capabilities while exerting limited influence on user-preference-oriented benchmarks.

DeepSeek-R1 Dev3 integrates both reasoning and non-reasoning datasets into the SFT pipeline, thereby enhancing the proficiency of the model in both reasoning and general language-generation tasks. Compared with Dev2, DeepSeek-R1 Dev3 achieves notable performance improvements on AlpacaEval 2.0 and Aider-Polyglot, attributable to the inclusion of large-scale non-reasoning corpora and code-engineering datasets. Finally, comprehensive RL training on DeepSeek-R1 Dev3 using mixed reasoning-focused and general-purpose data produced the final DeepSeek-R1. Marginal improvements occurred in code and mathematics benchmarks, as substantial reasoning-specific RL was done in previous stages. The primary advancements in the final DeepSeek-R1 were in general instruction-following and user-preference benchmarks, with AlpacaEval 2.0 improving by 25% and Arena-Hard by 17%.

We also compare DeepSeek-R1 with other models in Supplementary Information, section 4.2. Model safety evaluations are provided in Supplementary Information, section 4.3. A comprehensive analysis of evaluation is provided in Supplementary Information, section 5, including a comparison with DeepSeek-V3, performance evaluations on both fresh test sets, a breakdown of mathematical capabilities by category and an investigation of test-time scaling behaviour. Supplementary Information, section 6 shows that the strong reasoning capability can be transferred to smaller models.

表 1| DeepSeek-R1-Zero 的一个中间版本出现的有趣“顿悟时刻”

问题: 若 $a > 1$, 则方程 $\sqrt{a - \sqrt{a + x}} = x$ 的实数解之和为
回答: <think>
为求解方程 $\sqrt{a - \sqrt{a + x}} = x$, 先将等式两边平方……
$(\sqrt{a - \sqrt{a + x}})^2 = x^2 \Rightarrow a - \sqrt{a + x} = x^2$.
移项以分离出内部的平方根项:
$(a - x^2)^2 = a + x \Rightarrow a^2 - 2ax^2 + (x^2)^2 = a + x \Rightarrow x^4 - 2ax^2 - x + (a^2 - a) = 0$
...
等等, 等等. 等等。这是一个值得在此标记的顿悟时刻。
让我们逐步重新评估, 看看正确的和是否可以……
我们从以下方程开始:
$\sqrt{a - \sqrt{a + x}} = x$
首先, 我们将等式两边平方:
$a - \sqrt{a + x} = x^2 \Rightarrow \sqrt{a + x} = a - x^2$
接下来, 我可以再次将等式两边平方, 并将方程视为: ……
...
模型学会以拟人化的语气进行反思。这对我们来说也是一个顿悟时刻, 让我们见证了强化学习 (RL) 的力量与魅力。

和语言一致性。随后, 我们再次应用拒绝采样和 SFT。该阶段将推理和非推理数据集都纳入 SFT 流程, 详情见补充信息第 2.3.3 节, 使模型不仅能够出色地完成推理任务, 还能展现高级写作能力。为使模型进一步符合人类偏好, 我们实施了第二阶段强化学习, 旨在提升模型的有效性和无害性, 同时进一步完善其推理能力。奖励模型见方法部分“奖励设计”, 强化学习超参数见方法部分“第二阶段强化学习训练细节”。总训练成本见补充信息第 2.4.4 节。

我们在以下基准上评估模型¹¹, MMLU-Redux¹², MMLU-Pro¹³, DROP¹⁴, C-Eval¹⁵, IFEval¹⁶, FRAMES¹⁷, GPQA Diamond¹⁸, SimpleQA¹⁹, C-SimpleQA²⁰, CLUEWSC²¹, AlpacaEval 2.0 (参见 22), Arena-Hard²³, SWE-bench Verified²⁴, Aider-Polyglot²⁵, LiveCodeBench²⁶ (2024-08–2025-01), Codeforces²⁷, 中国全国高中数学

奥林匹克竞赛 (CNMO 2024)²⁸ 和 AIME 2024 (参见 29)。这些基准的详细信息见补充表 15–29。

表 2 总结了 DeepSeek-R1 在多个开发阶段的表现, 如图所示。2。DeepSeek-R1-Zero 与 DeepSeek-R1 Dev1 的比较表明, Dev1 在遵循指令方面取得了显著改进, 这体现在其 IF-Eval 和 Arena-Hard 基准测试得分更高。然而, 由于冷启动数据集规模有限, 与 DeepSeek-R1-Zero 相比, Dev1 的推理表现有所下降, 在 AIME 基准测试上尤为明显。相比之下, DeepSeek-R1 Dev2 在需要高级推理能力的基准测试中表现显著提升, 包括代码生成、数学问题求解和 STEM 相关任务。针对通用任务的基准测试 (如 AlpacaEval 2.0) 仅有小幅改进。这些结果表明, 以推理为导向的 RL 能显著增强推理能力, 同时对以用户偏好为导向的基准测试影响有限。

DeepSeek-R1 Dev3 将推理和非推理数据集都整合进 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, 这归因于纳入了大规模非推理语料和代码工程数据集。最后, 使用混合的推理导向和通用数据对 DeepSeek-R1 Dev3 进行全面的 RL 训练, 产出了最终版本 DeepSeek-R1。代码和数学基准测试仅有小幅提升, 因为此前阶段已进行了大量针对推理的 RL 训练。最终版 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%, Arena-Hard 提升了 17 %。

我们还在补充信息第 4.2 节将 DeepSeek-R1 与其他模型进行了比较。模型安全性评估见补充信息第 4.3 节。补充信息第 5 节对评估进行了全面分析, 包括与 DeepSeek-V3 的比较、在两个全新测试集上的性能评估、按类别划分的数学能力以及对测试时扩展行为的考察。补充信息第 6 节表明, 强大的推理能力可以迁移到较小的模型。

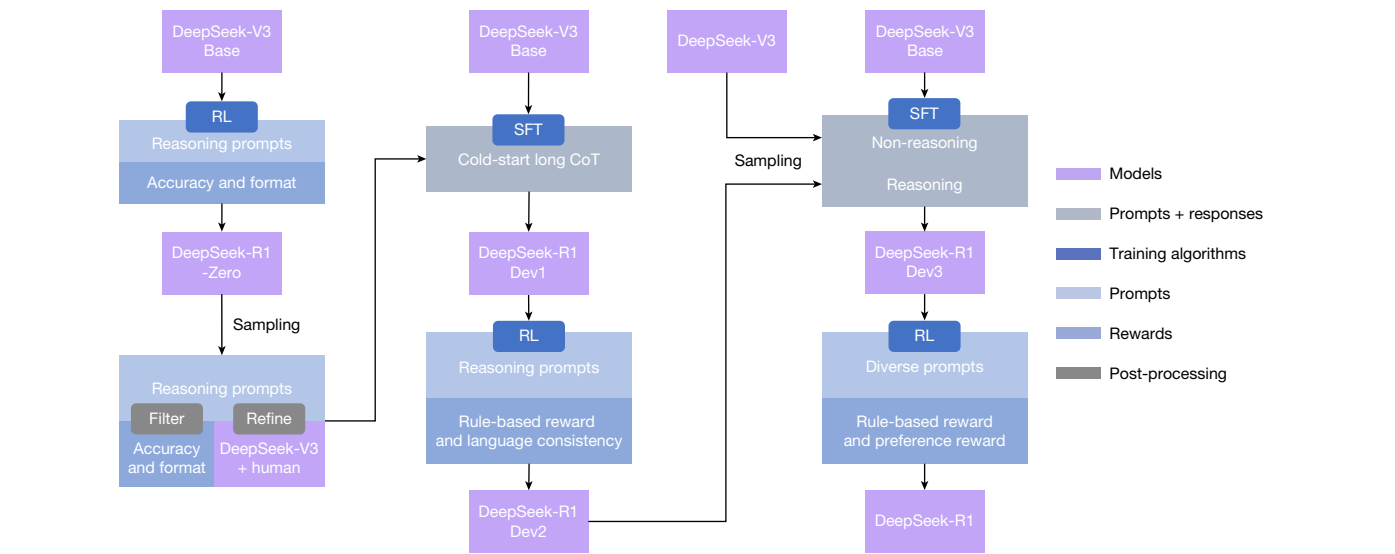


Fig. 2| The multistage pipeline of DeepSeek-R1. A detailed background on DeepSeek-V3 Base and DeepSeek-V3 is provided in Supplementary Information, section 1.1. The models DeepSeek-R1 Dev1, Dev2 and Dev3 represent intermediate checkpoints in this pipeline.

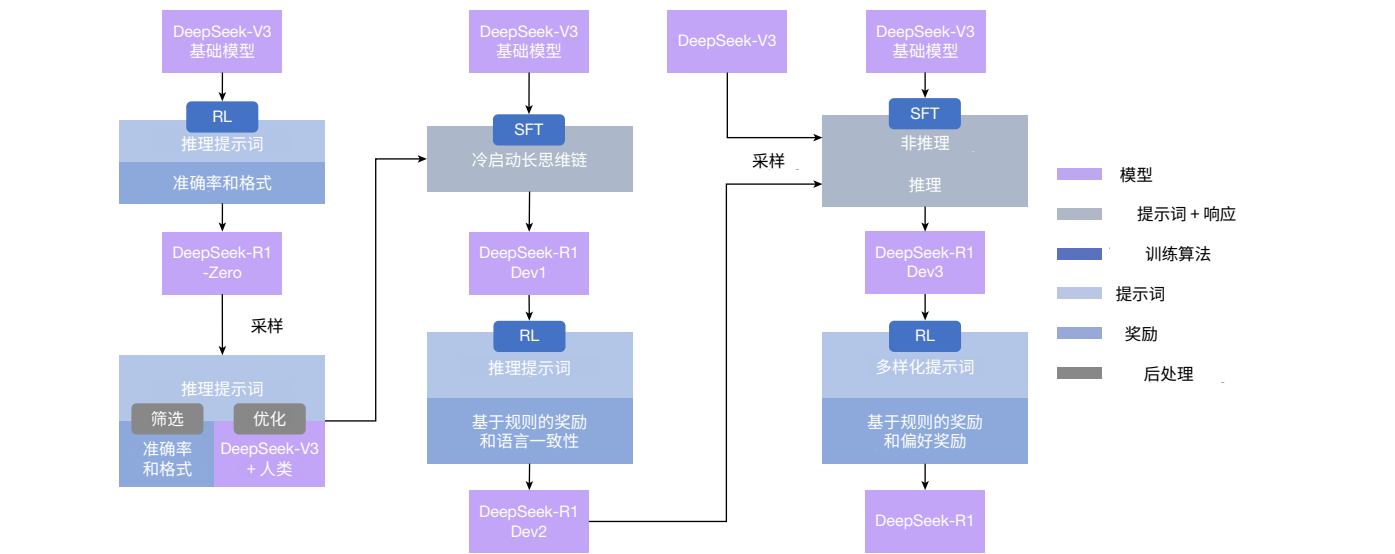


图 2| DeepSeek-R1 的多阶段流程。 DeepSeek-V3 Base 和 DeepSeek-V3 的详细背景介绍见补充信息第 1.1 节。DeepSeek-R1 Dev1、Dev2 和 Dev3 模型代表该流程中的中间检查点。

Table 2 | Experimental results at each stage of DeepSeek-R1

	Benchmark (metric)	R1-Zero	R1Dev1	R1Dev2	R1Dev3	R1
English	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
	GPQA Diamond (Pass@1)	75.8	66.1	70.7	71.2	71.5
	SimpleQA (Correct)	30.3	17.8	28.2	24.9	30.1
	FRAMES (Acc.)	82.3	78.5	81.8	81.9	82.5
	AlpacaEval 2.0 (LC-winrate)	24.7	50.1	55.8	62.1	87.6
	Arena-Hard (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 (Percentile)	80.4	84.5	90.5	92.1	96.3
	Codeforces (Rating)	1,444	1,534	1,687	1,746	2,029
	SWE-bench Verified (Resolved)	43.2	39.6	44.6	45.6	49.2
	Aider-Polyglot (Acc.)	12.2	6.7	25.6	44.8	53.3
Maths	AIME 2024 (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
	CNMO 2024 (Pass@1)	88.1	58.0	73.9	77.3	78.8
Chinese	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 (Correct)	66.4	58.8	64.2	66.9	63.7

Numbers in bold denote that the performance is statistically significant (t-test with $P < 0.01$).

Ethics and safety statement

With the advancement in the reasoning capabilities of DeepSeek-R1, we deeply recognize the potential ethical risks. For example, R1 can be subject to jailbreak attacks, leading to the generation of dangerous content such as explosive manufacturing plans, whereas the enhanced reasoning capabilities enable the model to provide plans with better operational feasibility and executability. Besides, a public model is also vulnerable to further fine-tuning that could compromise inherent safety protections.

In Supplementary Information, section 4.3, we present a comprehensive safety report from several perspectives, including performance on open-source and in-house safety evaluation benchmarks, and safety levels across several languages and against jailbreak attacks. These comprehensive safety analyses conclude that the inherent safety level of the DeepSeek-R1 model, compared with other state-of-the-art models, is generally at a moderate level (comparable with GPT-4o (2024-05-13)³⁰). Besides, when coupled with the risk control system, the safety level of the model is increased to a superior standard.

Conclusion, limitation and future work

We present DeepSeek-R1-Zero and DeepSeek-R1, which rely on large-scale RL to incentivize model reasoning behaviours. Our results demonstrate that pre-trained checkpoints inherently have substantial potential for complex reasoning tasks. We believe that the key to unlocking this potential lies not in large-scale human annotation but in the provision of hard reasoning questions, a reliable verifier and sufficient computational resources for RL. Sophisticated reasoning behaviours, such as self-verification and reflection, seemed to emerge organically during the RL process.

Even if DeepSeek-R1 achieves frontier results on reasoning benchmarks, it still faces several capability limitations, as outlined below.

Structure output and tool use

At present, the structural output capabilities of DeepSeek-R1 remain suboptimal compared with existing models. Moreover, DeepSeek-R1 cannot make use of tools, such as search engines and calculators, to improve the performance of output. However, as it is not hard to build a RL environment for structure output and tool use, we believe that the issue will be addressed in the next version.

Token efficiency

Unlike conventional test-time computation scaling approaches, such as majority voting or Monte Carlo tree search (MCTS), DeepSeek-R1 dynamically allocates computational resources during inference according to the complexity of the problem at hand. Specifically, it uses fewer tokens to solve simple tasks but generating more tokens for complex tasks. Nevertheless, there remains room for further optimization in terms of token efficiency, as instances of excessive reasoning—manifested as overthinking—are still observed in response to simpler questions.

Language mixing

DeepSeek-R1 is at present optimized for Chinese and English, which may result in language-mixing issues when handling queries in other languages. For instance, DeepSeek-R1 might use English for reasoning and responses, even if the query is in a language other than English or Chinese. We aim to address this limitation in future updates. The limitation may be related to the base checkpoint, DeepSeek-V3 Base, which mainly uses Chinese and English, so that it can achieve better results with the two languages in reasoning.

Prompting engineering

When evaluating DeepSeek-R1, we observe that it is sensitive to prompts. Few-shot prompting consistently degrades its performance. Therefore, we recommend that users directly describe the problem and specify the output format using a zero-shot setting for optimal results.

表格 2 | DeepSeek-R1 各阶段的实验结果

	基准测试（指标）	R1-Zero	R1 Dev1	R1 Dev2	R1 Dev3	R1
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（提示词严格性）	46.6	71.7	72.0	78.1	83.3
	GPQA Diamond (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
	AlpacaEval 2.0（LC 胜率）	24.7	50.1	55.8	62.1	87.6
	Arena-Hard（GPT-4-1106）	53.6	77.0	73.2	75.6	92.3
代码	LiveCodeBench（Pass@1-COT）	50.0	57.5	63.5	64.6	65.9
	Codeforces（百分位）	80.4	84.5	90.5	92.1	96.3
	Codeforces（评分）	1,444	1,534	1,687	1,746	2,029
	SWE-bench Verified（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
数学	AIME 2024（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
	CNMO 2024（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

粗体数字表示性能具有统计显著性（t检验， $P < 0.01$ ）。

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方案等危险内容；与此同时，增强的推理能力使模型能够提供操作可行性和可执行性更强的方案。此外，公开模型也容易受到进一步微调的影响，进而损害其固有的安全防护。

在补充信息第4.3节中，我们从多个角度提供了全面的安全报告，包括在开源和内部安全评估基准上的表现，以及模型在多种语言中的安全水平和抵御越狱攻击的能力。这些全面的安全分析得出结论：与其他最先进的模型相比，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。³⁰）。此外，结合风险控制系统后，模型的安全水平提升到了更高的标准。

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。我们的结果表明，预训练检查点本身就具有处理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RL计算资源。在RL过程中，自我验证和反思等复杂推理行为似乎自然涌现。

尽管 DeepSeek-R1 在推理基准测试中取得了前沿成果，但它仍面临若干能力限制，如下所述。

结构化输出与工具使用

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 无法使用搜索引擎和计算器等工具来提升输出性能。不过，鉴于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。

令牌效率

与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不同，DeepSeek-R1 会根据当前问题的复杂程度，在推理过程中动态分配计算资源。具体而言，它解决简单任务时使用较少的令牌，而解决复杂任务时生成更多令牌。尽管如此，令牌效率仍有进一步优化的空间，因为在回答较简单的问题时，仍能观察到过度推理（表现为想得太多）的情况。

语言混用

DeepSeek-R1 目前针对中文和英文进行了优化，因此在处理其他语言的查询时可能出现语言混用问题。例如，即使查询使用的不是英语或中文，DeepSeek-R1 也可能用英语进行推理并作答。我们计划在未来的更新中解决这一局限。这一局限可能与基础检查点 DeepSeek-V3 Base 有关，该检查点主要使用中文和英文，因此在使用这两种语言进行推理时能取得更好的结果。

提示工程

在评估 DeepSeek-R1 时，我们发现它对提示词很敏感。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。

Software-engineering tasks

Owing to the long evaluation times, which affect the efficiency of the RL process, large-scale RL has not been applied extensively in software-engineering tasks. As a result, DeepSeek-R1 has not demonstrated a huge improvement over DeepSeek-V3 on software-engineering benchmarks. Future versions will address this by implementing rejection sampling on software-engineering data or incorporating asynchronous evaluations during the RL process to improve efficiency.

Beyond specific capability limitations, the pure RL methodology itself also presents inherent challenges:

Reward hacking

The success of pure RL depends on reliable reward signals. In this study, we ensure reward reliability through a reasoning-domain rule-based reward model. However, such dependable reward models are difficult to construct for certain tasks, such as writing. If the reward signal is assigned by a model instead of predefined rules, it becomes more susceptible to exploitation as training progresses, which means that the policy model may find shortcuts to hack the reward model. Consequently, for complex tasks that cannot be effectively evaluated by a reliable reward model, scaling up pure RL methods remains an open challenge.

In this work, for tasks that cannot obtain a reliable signal, DeepSeek-R1 uses human annotation to create supervised data and only conducts RL for hundreds of steps. We hope that, in the future, a robust reward model can be obtained to address such issues.

With the advent of pure RL methods such as DeepSeek-R1, the future holds immense potential for solving any task that can be effectively evaluated by a verifier, regardless of its complexity for humans. Machines equipped with such advanced RL techniques are poised to surpass human capabilities in these domains, driven by their ability to optimize performance iteratively through trial and error. However, challenges remain for tasks for which constructing a reliable reward model is inherently difficult. In such cases, the lack of a robust feedback mechanism may slow progress, suggesting that future research should focus on developing innovative approaches to define and refine reward structures for these complex, less verifiable problems.

Furthermore, making use of tools during the reasoning process holds notable promise. Whether it is using tools such as compilers or search engines to retrieve or compute necessary information or using external tools such as biological or chemical reagents to validate final results in the real world, this integration of tool-augmented reasoning could greatly enhance the scope and accuracy of machine-driven solutions.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-025-09422-z>.

- Brown, T. B. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33 (eds Larochelle, H. et al.) (ACM, 2020).
- OpenAI et al. GPT4 technical report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2024).
- Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* 35 (eds Koyejo, S. et al.) 24824–24837 (ACM, 2022).
- Wei, J. et al. Emergent abilities of large language models. In *Transactions on Machine Learning Research* (eds Kamath, G. et al.) (2022).
- Kaplan, J. et al. Scaling laws for neural language models. Preprint at <https://doi.org/10.48550/arXiv.2001.08361> (2020).
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems* 35 (eds Oh, A. H. et al.) 22199–22213 (ACM, 2022).

- Chung, H. W. et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**, 1–53 (2024).
- DeepSeek-AI et al. DeepSeek-V3 technical report. Preprint at <https://doi.org/10.48550/arXiv.2412.19437> (2025).
- Shao, Z. et al. DeepSeekMath: pushing the limits of mathematical reasoning in open language models. Preprint at <https://doi.org/10.48550/arXiv.2402.03300> (2024).
- Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. In *11th International Conference on Learning Representations* (ICLR, 2023).
- Hendrycks, D. et al. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations* (ICLR, 2021).
- Gema, A. P. et al. Are we done with MMLU? In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Chiruzzo, L. et al.) Vol. 1 (Long Papers), 5069–5096 (ACL, 2025).
- Wang, Y. et al. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 37* (eds Globersons, A. et al.) 95266–95290 (ACM, 2024).
- Dua, D. et al. DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long and Short Papers) (eds Burstein, J. et al.) 2368–2378 (ACL, 2019).
- Huang, Y. et al. C-EVAL: a multi-level multi-discipline Chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems* 36 (eds Oh, A. et al.) 62991–63010 (ACM, 2023).
- Zhou, J. et al. Instruction-following evaluation for large language models. Preprint at <https://doi.org/10.48550/arXiv.2311.07911> (2023).
- Krishna, S. et al. Fact, fetch, and reason: a unified evaluation of retrieval-augmented generation. In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long Papers) 4745–4759 (ACL, 2025).
- Rein, D. et al. GPQA: a graduate-level Google-proof Q&A benchmark. Preprint at <https://doi.org/10.48550/arXiv.2311.2022> (2023).
- OpenAI. Introducing SimpleQA; <https://openai.com/index/introducing-simpleqa/> (2024).
- He, Y. et al. Chinese SimpleQA: a Chinese factuality evaluation for large language models. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics* Vol. 1 (Long Papers), 19182–19208 (ACL, 2025).
- Xu, L. et al. CLUE: a Chinese Language Understanding Evaluation benchmark. In *Proc. 28th International Conference on Computational Linguistics* (eds Scott, D. et al.) 4762–4772 (International Committee on Computational Linguistics, 2020).
- Dubois, Y., Galambosi, B., Liang, P. & Hashimoto, T. B. Length-controlled AlpacaEval: a simple way to debias automatic evaluators. Preprint at <https://doi.org/10.48550/arXiv.2404.04475> (2025).
- Li, T. et al. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. Preprint at <https://doi.org/10.48550/arXiv.2406.11939> (2024).
- OpenAI. Introducing SWE-bench verified; <https://openai.com/index/introducing-swe-bench-verified/> (2024).
- Aider. Aider LLM leaderboards; <https://aider.chat/docs/leaderboards/> (2024).
- Jain, N. et al. LiveCodeBench: holistic and contamination free evaluation of large language models for code. In *13th International Conference on Learning Representations* (ICLR, 2024).
- Mirzayanov, M. Codeforces; <https://codeforces.com/> (2025).
- Chinese Mathematical Society (CMS). *Chinese National High School Mathematics Olympiad*; <https://www.cms.org.cn/Home/comp/comp/cid/12.html> (2024).
- Mathematical Association of America. *American Invitational Mathematics Examination*; <https://maa.org/maa-invitational-competitions> (2024).
- OpenAI. Hello GPT-4o; <https://openai.com/index/hello-gpt-4o/> (2024).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

Daya Guo¹, Dejian Yang¹, Haowei Zhang¹, Junxiao Song¹, Peiyi Wang¹, Qihao Zhu¹, Runxin Xu¹, Ruoyu Zhang¹, Shirong Ma¹, Xiao Bi¹, Xiaokang Zhang¹, Xingkai Yu¹, Yu Wu¹, Z. F. Wu¹, Zhibin Gou¹, Zhihong Shao¹, Zhuoshu Li¹, Ziyi Gao¹, Aixin Liu¹, Bing Xue¹, Bingxuan Wang¹, Bochao Wu¹, Bei Feng¹, Chengda Lu¹, Chenggang Zhao¹, Chengqi Deng¹, Chong Ruan¹, Damai Dai¹, Deli Chen¹, Dongjie Ji¹, Erhang Li¹, Fangyun Lin¹, Fucong Dai¹, Fuli Luo^{1,2}, Guangbo Hao¹, Guanting Chen¹, Guowei Li¹, H. Zhang¹, Hanwei Xu¹, Honghui Ding¹, Huazuo Gao¹, Hui Qu¹, Hui Li¹, Jianzhong Guo¹, Jiashi Li¹, Jingchang Chen¹, Jingyang Yuan¹, Jinhao Tu^{1,3}, Junjie Qiu¹, Junlong Li¹, J. L. Cai¹, Jiaqi Ni¹, Jian Liang¹, Jin Chen¹, Kai Dong¹, Kai Hu^{1,4}, Kaichao You¹, Kaige Gao¹, Kang Guan¹, Kexin Huang^{1,5}, Kuai Yu¹, Lean Wang¹, Lecong Zhang¹, Liang Zhao¹, Litong Wang¹, Liyue Zhang¹, Lei Xu¹, Leyi Xia¹, Mingchuan Zhang¹, Minghua Zhang¹, Minghui Tang¹, Mingxu Zhou¹, Meng Li¹, Miaojun Wang¹, Mingming Li¹, Ning Tian¹, Panpan Huang¹, Peng Zhang¹, Qiancheng Wang¹, Qinyu Chen¹, Qiushi Du¹, Ruiqi Ge¹, Ruisong Zhang¹, Ruizhe Pan¹, Runji Wang¹, R. J. Chen¹,

软件工程任务

由于评估耗时较长，影响了 RL 流程的效率，大规模 RL 尚未广泛应用于软件工程任务。因此，DeepSeek-R1 在软件工程基准测试上的表现并未较 DeepSeek-V3 有显著提升。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。

除了特定能力方面的局限之外，纯 RL 方法本身也面临固有挑战：

奖励投机

纯 RL 的成功取决于可靠的奖励信号。在本研究中，我们通过推理领域的基于规则的奖励模型来确保奖励的可靠性。然而，对于写作等某些任务，构建此类可靠的奖励模型十分困难。如果奖励信号由模型而非预先定义的规则给出，那么随着训练的推进，它更容易受到利用，这意味着策略模型可能会找到捷径来操纵奖励模型。因此，对于无法通过可靠奖励模型有效评估的复杂任务，如何扩展纯 RL 方法仍是一项悬而未决的挑战。

在这项工作中，对于无法获得可靠信号的任务，DeepSeek-R1 使用人工标注来创建监督数据，并且仅进行数百步的 RL。我们希望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，未来充满巨大潜力：任何能够由验证器有效评估的任务，无论对人类而言多么复杂，都有望得到解决。具备此类先进 RL 技术的机器，凭借通过反复试错迭代优化性能的能力，有望在这些领域超越人类。然而，对于那些难以从根本上构建可靠奖励模型的任务，挑战依然存在。在这种情况下，缺乏稳健的反馈机制可能会减缓进展，这表明未来研究应着重开发创新方法，为这些复杂且较难验证的问题定义并完善奖励结构。

此外，在推理过程中使用工具也展现出显著潜力。无论是使用编译器或搜索引擎等工具来检索或计算所需信息，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，这种将工具融入推理的做法都能大幅拓展机器驱动解决方案的适用范围，并提高其准确率。

在线内容

任何方法、其他参考文献、Nature Portfolio 报告摘要、源数据、扩展数据、补充信息、致谢、同行评审信息；作者贡献和利益冲突详情；以及数据和代码可用性声明， 均可通过以下链接获取 <https://doi.org/10.1038/s41586-025-09422-z>。

- Brown, T. B. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33 (eds Larochelle, H. et al.) (ACM, 2020).
- OpenAI et al. GPT4 technical report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2024).
- Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* 35 (eds Koyejo, S. et al.) 24824–24837 (ACM, 2022).
- Wei, J. et al. Emergent abilities of large language models. In *Transactions on Machine Learning Research* (eds Kamath, G. et al.) (2022).
- Kaplan, J. et al. Scaling laws for neural language models. Preprint at <https://doi.org/10.48550/arXiv.2001.08361> (2020).
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems* 35 (eds Oh, A. H. et al.) 22199–22213 (ACM, 2022).

- Chung, H. W. et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**, 1–53 (2024).
- DeepSeek-AI et al. DeepSeek-V3 technical report. Preprint at <https://doi.org/10.48550/arXiv.2412.19437> (2025).
- Shao, Z. et al. DeepSeekMath: pushing the limits of mathematical reasoning in open language models. Preprint at <https://doi.org/10.48550/arXiv.2402.03300> (2024).
- Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. In *11th International Conference on Learning Representations* (ICLR, 2023).
- Hendrycks, D. et al. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations* (ICLR, 2021).
- Gema, A. P. et al. Are we done with MMLU? In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Chiruzzo, L. et al.) Vol. 1 (Long Papers), 5069–5096 (ACL, 2025).
- Wang, Y. et al. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 37* (eds Globersons, A. et al.) 95266–95290 (ACM, 2024).
- Dua, D. et al. DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long and Short Papers) (eds Burstein, J. et al.) 2368–2378 (ACL, 2019).
- Huang, Y. et al. C-EVAL: a multi-level multi-discipline Chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems* 36 (eds Oh, A. et al.) 62991–63010 (ACM, 2023).
- Zhou, J. et al. Instruction-following evaluation for large language models. Preprint at <https://doi.org/10.48550/arXiv.2311.07911> (2023).
- Krishna, S. et al. Fact, fetch, and reason: a unified evaluation of retrieval-augmented generation. In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long Papers) 4745–4759 (ACL, 2025).
- Rein, D. et al. GPQA: a graduate-level Google-proof Q&A benchmark. Preprint at <https://doi.org/10.48550/arXiv.2311.2022> (2023).
- OpenAI. Introducing SimpleQA; <https://openai.com/index/introducing-simpleqa/> (2024).
- He, Y. et al. Chinese SimpleQA: a Chinese factuality evaluation for large language models. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics* Vol. 1 (Long Papers), 19182–19208 (ACL, 2025).
- Xu, L. et al. CLUE: a Chinese Language Understanding Evaluation benchmark. In *Proc. 28th International Conference on Computational Linguistics* (eds Scott, D. et al.) 4762–4772 (International Committee on Computational Linguistics, 2020).
- Dubois, Y., Galambosi, B., Liang, P. & Hashimoto, T. B. Length-controlled AlpacaEval: a simple way to debias automatic evaluators. Preprint at <https://doi.org/10.48550/arXiv.2404.04475> (2025).
- Li, T. et al. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. Preprint at <https://doi.org/10.48550/arXiv.2406.11939> (2024).
- OpenAI. Introducing SWE-bench verified; <https://openai.com/index/introducing-swe-bench-verified/> (2024).
- Aider. Aider LLM leaderboards; <https://aider.chat/docs/leaderboards/> (2024).
- Jain, N. et al. LiveCodeBench: holistic and contamination free evaluation of large language models for code. In *13th International Conference on Learning Representations* (ICLR, 2024).
- Mirzayanov, M. Codeforces; <https://codeforces.com/> (2025).
- Chinese Mathematical Society (CMS). *Chinese National High School Mathematics Olympiad*; <https://www.cms.org.cn/Home/comp/comp/cid/12.html> (2024).
- Mathematical Association of America. *American Invitational Mathematics Examination*; <https://maa.org/maa-invitational-competitions> (2024).
- OpenAI. Hello GPT-4o; <https://openai.com/index/hello-gpt-4o/> (2024).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

Daya Guo¹, Dejian Yang¹, Haowei Zhang¹, Junxiao Song¹, Peiyi Wang¹, Qihao Zhu¹, Runxin Xu¹, Ruoyu Zhang¹, Shirong Ma¹, Xiao Bi¹, Xiaokang Zhang¹, Xingkai Yu¹, Yu Wu¹, Z. F. Wu¹,Zhibin Gou¹, Zhihong Shao¹, Zhuoshu Li¹, Ziyi Gao¹, Aixin Liu¹, Bing Xue¹, Bingxuan Wang¹,Bochao Wu¹, Bei Feng¹, Chengda Lu¹, Chenggang Zhao¹, Chengqi Deng¹, Chong Ruan¹,Damai Dai¹, Deli Chen¹, Dongjie Ji¹, Erhang Li¹, Fangyun Lin¹, Fucong Dai¹, Fuli Luo^{1,2},Guangbo Hao¹, Guanting Chen¹, Guowei Li¹, H. Zhang¹, Hanwei Xu¹, Honghui Ding¹,Huazuo Gao¹, Hui Qu¹, Hui Li¹, Jianzhong Guo¹, Jiashi Li¹, Jingchang Chen¹, Jingyang Yuan¹,Jinhao Tu^{1,3}, Junjie Qiu¹, Junlong Li¹, J. L. Cai¹, Jiaqi Ni¹, Jian Liang¹, Jin Chen¹, Kai Dong¹,Kai Hu^{1,4}, Kaichao You¹, Kaige Gao¹, Kang Guan¹, Kexin Huang^{1,5}, Kuai Yu¹, Lean Wang¹,Lecong Zhang¹, Liang Zhao¹, Litong Wang¹, Liyue Zhang¹, Lei Xu¹, Leyi Xia¹,Mingchuan Zhang¹, Minghua Zhang¹, Minghui Tang¹, Mingxu Zhou¹, Meng Li¹,Miaojun Wang¹, Mingming Li¹, Ning Tian¹, Panpan Huang¹, Peng Zhang¹, Qiancheng Wang¹,Qinyu Chen¹, Qiushi Du¹, Ruiqi Ge¹, Ruisong Zhang¹, Ruizhe Pan¹, Runji Wang¹, R. J. Chen¹,

R. L. Jin¹, Ruyi Chen¹, Shanghao Lu¹, Shangyan Zhou¹, Shanhuang Chen¹, Shengfeng Ye¹, Shiyu Wang¹, Shuiping Yu¹, Shunfeng Zhou¹, Shuting Pan¹, S. S. Li¹, Shuang Zhou¹, Shaoqing Wu¹, Tao Yun¹, Tian Pei¹, Tianyu Sun¹, T. Wang¹, Wangding Zeng¹, Wen Liu¹, Wenfeng Liang^{1,2,3}, Wenjun Gao¹, Wenqin Yu^{1,5}, Wentao Zhang¹, W. L. Xiao¹, Wei An¹, Xiaodong Liu¹, Xiaohan Wang¹, Xiaokang Chen¹, Xiaotao Nie¹, Xin Cheng¹, Xin Liu¹, Xin Xie¹, Xingchao Liu¹, Xinyu Yang¹, Xinyuan Li^{1,5}, Xuecheng Su¹, Xuheng Lin¹, X. Q. Li¹, Xiangyue Jin¹, Xiaojin Shen¹, Xiaosha Chen¹, Xiaowen Sun¹, Xiaoxiang Wang¹, Xinnan Song¹, Xinyi Zhou¹, Xianzu Wang¹, Xinxia Shan¹, Y. K. Li¹, Y. Q. Wang¹, Y. X. Wei¹, Yang Zhang¹, Yanhong Xu¹, Yao Li¹, Yao Zhao¹, Yaofeng Sun¹, Yaohui Wang¹, Yi Yu¹, Yichao Zhang¹, Yifan Shi¹, Yiliang Xiong¹, Ying He¹, Yishi Piao¹, Yisong Wang¹, Yixuan Tan¹, Yiyang Ma¹, Yiyuan Liu¹, Yongqiang Guo¹, Yuan Ou¹, Yuduan Wang¹, Yue Gong^{1,5}, Yuheng Zou¹, Yujia He^{1,5}, Yunfan Xiong¹, Yuxiang Luo¹, Yuxiang You¹, Yuxuan Liu¹, Yuyang Zhou¹, Y. X. Zhu¹,

Yanping Huang¹, Yaohui Li¹, Yi Zheng¹, Yuchen Zhu¹, Yunxian Ma¹, Ying Tang¹, Yukun Zha¹, Yuting Yan¹, Z. Z. Ren¹, Zehui Ren¹, Zhangli Sha¹, Zhe Fu¹, Zhean Xu¹, Zhenda Xie¹, Zhengyan Zhang¹, Zhewen Hao¹, Zhicheng Ma¹, Zhigang Yan¹, Zhiyu Wu¹, Zihui Gu¹, Zijia Zhu¹, Zijun Liu^{1,6}, Zilin Li¹, Ziwei Xie¹, Ziyang Song^{1,7}, Zizheng Pan¹, Zhen Huang¹, Zhipeng Xu¹, Zhongyu Zhang¹ & Zhen Zhang¹

¹DeepSeek-AI Team, Hangzhou, China. ²Present address: Individual Researcher, Beijing, China. ³Present address: Jianping High School, Shanghai, China. ⁴Present address: University of Science and Technology of China, Hefei, China. ⁵Present address: Peking University, Beijing, China. ⁶Present address: Tsinghua University, Beijing, China. ⁷Present address: Citadel Securities, Hong Kong SAR, China. [✉]e-mail: wenfeng.liang@deepseek.com

R. L. Jin¹, Ruyi Chen¹, Shanghao Lu¹, Shangyan Zhou¹, Shanhuang Chen¹, Shengfeng Ye¹, Shiyu Wang¹, Shuiping Yu¹, Shunfeng Zhou¹, Shuting Pan¹, S. S. Li¹, Shuang Zhou¹,Shaoqing Wu¹, Tao Yun¹, Tian Pei¹, Tianyu Sun¹, T. Wang¹, Wangding Zeng¹, Wen Liu¹, Wenfeng Liang^{1,2,3}, 温军 Gao¹, Wenqin Yu^{1,5}, Wentao Zhang¹, W. L. Xiao¹, Wei An¹,Xiaodong Liu¹, Xiaohan Wang¹, Xiaokang Chen¹, Xiaotao Nie¹, Xin Cheng¹, Xin Liu¹, Xin Xie¹,Xingchao Liu¹, Xinyu Yang¹, Xinyuan Li^{1,5}, Xuecheng Su¹, Xuheng Lin¹, X. Q. Li¹, Xiangyue Jin¹,Xiaojin Shen¹, Xiaosha Chen¹, Xiaowen Sun¹, Xiaoxiang Wang¹, Xinnan Song¹, Xinyi Zhou¹,Xianzu Wang¹, Xinxia Shan¹, Y. K. Li¹, Y. Q. Wang¹, Y. X. Wei¹, Yang Zhang¹, Yanhong Xu¹,Yao Li¹, Yao Zhao¹, Yaofeng Sun¹, Yaohui Wang¹, Yi Yu¹, Yichao Zhang¹, Yifan Shi¹,Yiliang Xiong¹, Ying He¹, Yishi Piao¹, Yisong Wang¹, Yixuan Tan¹, Yiyang Ma¹, Yiyuan Liu¹,Yongqiang Guo¹, Yuan Ou¹, Yuduan Wang¹, Yue Gong^{1,5}, Yuheng Zou¹, Yujia He^{1,5},Yunfan Xiong¹, Yuxiang Luo¹, Yuxiang You¹, Yuxuan Liu¹, Yuyang Zhou¹, Y. X. Zhu¹,

Yanping Huang¹, Yaohui Li¹, Yi Zheng¹, Yuchen Zhu¹, Yunxian Ma¹, Ying Tang¹, Yukun Zha¹, Yuting Yan¹, Z. Z. Ren¹, Zehui Ren¹, Zhangli Sha¹, Zhe Fu¹, Zhean Xu¹, Zhenda Xie¹,Zhengyan Zhang¹, Zhewen Hao¹, Zhicheng Ma¹, Zhigang Yan¹, Zhiyu Wu¹, Zihui Gu¹, Zijia Zhu¹,Zijun Liu^{1,6}, Zilin Li¹, Ziwei Xie¹, Ziyang Song^{1,7}, Zizheng Pan¹, Zhen Huang¹, Zhipeng Xu¹,Zhongyu Zhang¹ & Zhen Zhang¹

¹DeepSeek-AI Team, Hangzhou, China. ²Present address: Individual Researcher, Beijing, China. ³Present address: Jianping High School, Shanghai, China. ⁴Present address: University of Science and Technology of China, Hefei, China. ⁵Present address: Peking University, Beijing, China. ⁶Present address: Tsinghua University, Beijing, China. ⁷Present address: Citadel Securities, Hong Kong SAR, China. [✉]e-mail: wenfeng.liang@deepseek.com

Methods

GRPO

GRPO⁹ is the RL algorithm that we use to train DeepSeek-R1-Zero and DeepSeek-R1. It was originally proposed to simplify the training process and reduce the resource consumption of proximal policy optimization (PPO)³¹, which is widely used in the RL stage of LLMs³². The pipeline of GRPO is shown in Extended Data Fig. 2.

For each question q , GRPO samples a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and then optimizes the policy model π_{θ} by maximizing the following objective:

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

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

in which π_{ref} is a reference policy, ϵ and β are hyperparameters and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \dots, r_G\}$ corresponding to the outputs in each group:

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

We give a comparison of GRPO and PPO in Supplementary Information, section 1.3.

Reward design

The reward is the source of the training signal, which decides the direction of RL optimization. For DeepSeek-R1-Zero, we use rule-based rewards to deliver precise feedback for data in mathematical, coding and logical reasoning domains. For DeepSeek-R1, we extend this approach by incorporating both rule-based rewards for reasoning-oriented data and model-based rewards for general data, thereby enhancing the adaptability of the learning process across diverse domains.

Rule-based rewards. Our rule-based reward system mainly consists of two types of reward: accuracy rewards and format rewards.

Accuracy rewards evaluate whether the response is correct. For example, in the case of maths problems with deterministic results, the model is required to provide the final answer in a specified format (for example, within a box), enabling reliable rule-based verification of correctness. Similarly, for code competition prompts, a compiler can be used to evaluate the responses of the model against a suite of predefined test cases, thereby generating objective feedback on correctness.

Format rewards complement the accuracy reward model by enforcing specific formatting requirements. In particular, the model is incentivized to encapsulate its reasoning process within designated tags, specifically <think> and </think>. This ensures that the thought process of the model is explicitly delineated, enhancing interpretability and facilitating subsequent analysis.

$$\text{Reward}_{\text{rule}} = \text{Reward}_{\text{acc}} + \text{Reward}_{\text{format}} \quad (4)$$

The accuracy, reward and format reward are combined with the same weight. Notably, we abstain from applying neural reward models—whether outcome-based or process-based—to reasoning tasks. This decision is predicated on our observation that neural reward models are susceptible to reward hacking during large-scale RL.

Moreover, retraining such models necessitates substantial computational resources and introduces further complexity into the training pipeline, thereby complicating the overall optimization process.

Model-based rewards. For general data, we resort to reward models to capture human preferences in complex and nuanced scenarios. We build on the DeepSeek-V3 pipeline and use a similar distribution of preference pairs and training prompts. For helpfulness, we focus exclusively on the final summary, ensuring that the assessment emphasizes the use and relevance of the response to the user while minimizing interference with the underlying reasoning process. For harmlessness, we evaluate the entire response of the model, including both the reasoning process and the summary, to identify and mitigate any potential risks, biases or harmful content that may arise during the generation process. **Helpful reward model.** For helpful reward model training, we first generate preference pairs by prompting DeepSeek-V3 using the Arena-Hard prompt format, listed in Supplementary Information, section 2.2, for which each pair consists of a user query along with two candidate responses. For each preference pair, we query DeepSeek-V3 four times, randomly assigning the responses as either Response A or Response B to mitigate positional bias. The final preference score is determined by averaging the four independent judgments, retaining only those pairs for which the score difference (Δ) exceeds 1 to ensure meaningful distinctions. Furthermore, to minimize length-related biases, we ensure that the chosen and rejected responses of the whole dataset have comparable lengths. In total, we curated 66,000 data pairs for training the reward model. The prompts used in this dataset are all non-reasoning questions and are sourced either from publicly available open-source datasets or from users who have explicitly consented to share their data for the purpose of model improvement. The architecture of our reward model is consistent with that of DeepSeek-R1, with the addition of a reward head designed to predict scalar preference scores.

$$\text{Reward}_{\text{helpful}} = \text{RM}_{\text{helpful}}(\text{Response}_{\text{A}}, \text{Response}_{\text{B}}) \quad (5)$$

The helpful reward models were trained with a batch size of 256, a learning rate of 6×10^{-6} and for a single epoch over the training dataset. The maximum sequence length during training is set to 8,192 tokens, whereas no explicit limit is imposed during reward model inference. **Safety reward model.** To assess and improve model safety, we curated a dataset of 106,000 prompts with model-generated responses annotated as ‘safe’ or ‘unsafe’ according to predefined safety guidelines. Unlike the pairwise loss used in the helpfulness reward model, the safety reward model was trained using a pointwise methodology to distinguish between safe and unsafe responses. The training hyperparameters are the same as the helpful reward model.

$$\text{Reward}_{\text{safety}} = \text{RM}_{\text{safety}}(\text{Response}) \quad (6)$$

For general queries, each instance is categorized as belonging to either the safety dataset or the helpfulness dataset. The general reward, $\text{Reward}_{\text{general}}$, assigned to each query corresponds to the respective reward defined in the associated dataset.

Training details

Training details of DeepSeek-R1-Zero. To train DeepSeek-R1-Zero, we set the learning rate to 3×10^{-6} , the Kullback–Leibler (KL) coefficient to 0.001 and the sampling temperature to 1 for rollout. For each question, we sample 16 outputs with a maximum length of 32,768 tokens before the 8.2k step and 65,536 tokens afterward. As a result, both the performance and response length of DeepSeek-R1-Zero exhibit a substantial jump at the 8.2k step, with training continuing for a total of 10,400 steps, corresponding to 1.6 training epochs. Each training step consists of 32 unique questions, resulting in a training batch size of 512 per step. Every 400 steps, we replace the reference model with

方法

GRPO

GRPO⁹ 是我们用于训练 DeepSeek-R1-Zero 和 DeepSeek-R1 的强化学习（RL）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）的资源消耗³¹, 近端策略优化广泛用于大型语言模型（LLM）的强化学习阶段³²。GRPO 的流程如扩展数据图 2 所示。

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

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

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

其中 π_{ref} 是一个参考策略， ϵ 和 β 是超参数，且 A_i 是优势值，使用一组奖励计算得出 $\{r_1, r_2, \dots, r_G\}$ 对应于每组中的输出：

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

我们在补充信息第 1.3 节中比较了 GRPO 和 PPO。

奖励设计

奖励是训练信号的来源，它决定了 RL 优化的方向。对于 DeepSeek-R1-Zero，我们使用基于规则的奖励，为数学、编码和逻辑推理领域的数​​据提供精确反馈。对于 DeepSeek-R1，我们通过将面向推理数据的基于规则的奖励与面向通用数据的基于模型的奖励相结合，扩展了这一方法，从而增强了学习过程在不同领域中的适应性。

基于规则的奖励。 我们的基于规则的奖励系统主要由两种奖励组成：准确率奖励和格式奖励。

准确率奖励用于评估回答是否正确。例如，对于结果确定的数学问题，模型需要按照指定格式（例如放在方框内）给出最终答案，以便通过可靠的基于规则的方式验证其正确性。同样，对于代码竞赛提示，可以使用编译器将模型的回答与一套预定义的测试用例进行比对评估，从而生成有关正确性的客观反馈。

格式奖励通过强制执行特定格式要求来补充准确率奖励模型。具体而言，模型会受到激励，将其推理过程封装在指定标签中，具体为 <think> and </think>。这确保模型的思维过程得到明确界定，增强了可解释性，并便于后续分析。

$$\text{Reward}_{\text{rule}} = \text{Reward}_{\text{acc}} + \text{Reward}_{\text{format}} \quad (4)$$

准确率、奖励和格式奖励以相同权重合并。值得注意的是，我们不会在推理任务中应用神经奖励模型——无论是基于结果还是基于过程的模型。这一决定基于我们的观察：在大规模强化学习中，神经奖励模型容易受到奖励投机的影响。

此外，重新训练此类模型需要大量计算资源，并会给训练流程带来额外复杂性，从而使整体优化过程更加复杂。

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情境中的人类偏好。我们以 DeepSeek-V3 流程为基础，并使用分布相似的偏好对和训练提示。对于有用性，我们只关注最终摘要，确保评估侧重于响应对用户的用途和相关性，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响应，包括推理过程和摘要，以识别并缓解生成过程中可能出现的风险、偏见或有害内容。

有用性奖励模型。 在训练有用性奖励模型时，我们首先使用补充信息第 2.2 节所列的 Arena-Hard 提示格式提示 DeepSeek-V3，生成偏好对；每个偏好对由一个用户查询和两个候选响应组成。对于每个偏好对，我们向 DeepSeek-V3 提问四次，并随机将响应指定为响应 A 或响应 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值确定；仅保留分数差（ Δ ）超过 1 的偏好对，以确保差异具有意义。此外，为尽量减少与长度相关的偏差，我们确保整个数据集中被选中的响应和被拒绝的响应长度相近。我们总共整理了 66,000 个数据对，用于训练奖励模型。该数据集使用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，要么是明确同意分享数据以改进模型的用户。我们的奖励模型架构与 DeepSeek-R1 一致，并增加了一个奖励头，用于预测标量偏好分数。

$$\text{Reward}_{\text{helpful}} = \text{RM}_{\text{helpful}}(\text{Response}_{\text{A}}, \text{Response}_{\text{B}}) \quad (5)$$

这些有用性奖励模型的训练批量大小为 256，学习率为 6×10^{-6} 并在训练数据集上训练了一个 epoch。训练期间的最大序列长度设为 8,192 个 token，而奖励模型推理期间未施加明确的限制。

安全奖励模型。 为评估和提升模型安全性，我们整理了一个包含 106,000 条提示的数据集，其中模型生成的回复根据预先制定的安全准则标注为“安全”或“不安全”。不同于有用性奖励模型所采用的成对损失，安全奖励模型使用逐点方法进行训练，以区分安全和不安全的回复。其训练超参数与有用性奖励模型相同。

$$\text{Reward}_{\text{safety}} = \text{RM}_{\text{safety}}(\text{Response}) \quad (6)$$

对于一般查询，每个实例都会被归类到安全数据集或有用性数据集。 $\text{Reward}_{\text{general}}$ 分配给每个查询的通用奖励对应于相关数据集中定义的相应奖励。

训练细节

DeepSeek-R1-Zero 的训练细节。 为训练 DeepSeek-R1-Zero，我们将学习率设为 3×10^{-6} ，Kullback–Leibler（KL）系数设为 0.001，采样温度设为 1，用于生成 rollout。对于每个问题，在第 8.2k 步之前，我们采样 16 个输出，最大长度为 32,768 个 token；此后最大长度为 65,536 个 token。因此，在第 8.2k 步时，DeepSeek-R1-Zero 的性能和响应长度均大幅跃升；训练总计持续 10,400 步，相当于 1.6 个训练轮次。每个训练步骤包含 32 个不同的问题，因此每步的训练批次大小为 512。每隔 400 步，我们将参考模型替换为

Article

the latest policy model. To accelerate training, each rollout generates 8,192 outputs, which are randomly split into 16 minibatches and trained for only a single inner epoch.

Training details of the first RL stage. In the first stage of RL, we set the learning rate to 3×10^{-6} , the KL coefficient to 0.001, the GRPO clip ratio ϵ to 10 and the sampling temperature to 1 for rollout. For each question, we sample 16 outputs with a maximum length of 32,768. Each training step consists of 32 unique questions, resulting in a training batch size of 512 per step. Every 400 steps, we replace the reference model with the latest policy model. To accelerate training, each rollout generates 8,192 outputs, which are randomly split into 16 minibatches and trained for only a single inner epoch. However, to mitigate the issue of language mixing, we introduce a language consistency reward during RL training, which is calculated as the proportion of target language words in the CoT.

$$\text{Reward}_{\text{language}} = \frac{\text{Num}(\text{Words}_{\text{target}})}{\text{Num}(\text{Words})} \quad (7)$$

Although ablation experiments in Supplementary Information, sction 2.6 show that such alignment results in a slight degradation in the performance of the model, this reward aligns with human preferences, making it more readable. We apply the language consistency reward to both reasoning and non-reasoning data by directly adding it to the final reward.

Note that the clip ratio plays a crucial role in training. A lower value can lead to the truncation of gradients for a large number of tokens, thereby degrading the performance of the model, whereas a higher value may cause instability during training. Details of RL data used in this stage are provided in Supplementary Information, section 2.3.

Training details of the second RL stage. Specifically, we train the model using a combination of reward signals and diverse prompt distributions. For reasoning data, we follow the methodology outlined in DeepSeek-R1-Zero, which uses rule-based rewards to guide learning in mathematical, coding and logical reasoning domains. During the training process, we observe that CoT often exhibits language mixing, particularly when RL prompts involve several languages. For general data, we use reward models to guide training. Ultimately, the integration of reward signals with diverse data distributions enables us to develop a model that not only excels in reasoning but also assigns priority to helpfulness and harmlessness. Given a batch of data, the reward can be formulated as

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

in which

$$\text{Reward}_{\text{reasoning}} = \text{Reward}_{\text{rule}} \quad (9)$$

$$\text{Reward}_{\text{general}} = \text{Reward}_{\text{reward_model}} + \text{Reward}_{\text{format}} \quad (10)$$

The second stage of RL retains most of the parameters from the first stage, with the key difference being a reduced temperature of 0.7, as we find that higher temperatures in this stage lead to incoherent generation. The stage comprises a total of 1,700 training steps, during which general instruction data and preference-based rewards are incorporated exclusively in the final 400 steps. We find that more

training steps with the model-based preference reward signal may lead to reward hacking, which is documented in Supplementary Information, section 2.5.

Data availability

We provide the data samples used in our rejection sampling and RL prompts at <https://github.com/deepseek-ai/DeepSeek-R1> (<https://doi.org/10.5281/zenodo.15753193>)³³. Comprehensive statistics and details of our complete data-generation methodology are presented in Supplementary Information, section 2.3.

Code availability

Trained weights of DeepSeek-R1-Zero and DeepSeek-R1 are available under an MIT license at <https://github.com/deepseek-ai/DeepSeek-R1> (<https://doi.org/10.5281/zenodo.15753193>)³³. The inference script is released at <https://github.com/deepseek-ai/DeepSeek-V3> (<https://doi.org/10.5281/zenodo.15753347>)³⁴. Neural networks were developed with PyTorch³⁵ and the distributed framework is based on our internal framework HAI-LLM (<https://www.high-flyer.cn/en/blog/hai-llm>). The inference framework is based on vLLM³⁶. Data analysis used Python v.3.8 (<https://www.python.org/>), NumPy v.1.23.1 (<https://github.com/numpy/numpy>), Matplotlib v.3.5.2 (<https://github.com/matplotlib/matplotlib>) and TensorBoard v.2.9.1 (<https://github.com/tensorflow/tensorboard>).

- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. Preprint at <https://doi.org/10.48550/arXiv.1707.06347> (2017).
- Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35* (eds Koyejo, S. et al.) 27730–27744 (ACM, 2022).
- Nano et al. deepseek-ai/DeepSeek-R1: v1.0.0. *Zenodo* <https://doi.org/10.5281/zenodo.15753192> (2025).
- Yu, X. et al. deepseek-ai/DeepSeek-V3: v1.0.0. *Zenodo* <https://doi.org/10.5281/zenodo.15753346> (2025).
- Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (eds Wallach, H. M. et al.) 8026–8037 (ACM, 2019).
- Kwon, W. et al. Efficient memory management for large language model serving with PagedAttention. In *Proc. ACM SIGOPS 29th Symposium on Operating Systems Principles* 611–626 (ACM, 2023).

Acknowledgements The research is supported by DeepSeek-AI.

Author contributions All authors have contributed to the publication, being variously involved in collecting and curating data, designing the experiments and building the LLM training framework. The authors also participated in implementing and testing the experimental set-up, refining the RL process and performing the analysis of results. The scientific findings were discussed and approved by all contributors. This article was written by a subgroup of authors designated by the collaboration and it underwent an internal collective review process. All authors reviewed and approved the final version of the manuscript. Core contributors: D.G., D.Y., H.Z., J.S., R.Z., R.X., Q.Z., S.M., P.W., X.B., X.Z., X.Y., Y.W., Z.F.W., Z.G., Z.S., Z.L., Z.G. These authors, designated as core contributors, contributed equally to this work, and are presented in alphabetical order. The remaining authors have also contributed meaningfully to the study, and their names are likewise presented in alphabetical order.

Competing interests The authors declare no competing interests and will not file patents related to the content of this manuscript.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09422-z>.

Correspondence and requests for materials should be addressed to Wenfeng Liang. **Peer review information** *Nature* thanks Edward Beeching, Yarin Gal, José Hernández-Orallo, Daphne Ippolito, Subbarao Kambhampati, Lewis Tunstall, Yiming Zhang and Lexin Zhou for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

文章

最新的策略模型。为加快训练，每次 rollout 会生成 8,192 个输出，并随机拆分为 16 个小批次，仅训练一个内层轮次。

第一阶段强化学习的训练细节。 在强化学习的第一阶段，我们将学习率设为 3×10^{-6} ，KL 系数设为 0.001，GRPO 裁剪值设为 $\text{ratio} \epsilon$ to 10，采样温度设为 1 进行 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 个不同的问题，因此每步的训练批次大小为 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 rollout 生成 8,192 个输出，并将其随机拆分为 16 个小批次，仅训练一个内部 epoch。然而，为缓解语言混用问题，我们在强化学习训练期间引入语言一致性奖励，该奖励按思维链中目标语言词语所占比例计算。

$$\text{Reward}_{\text{language}} = \frac{\text{Num}(\text{Words}_{\text{target}})}{\text{Num}(\text{Words})} \quad (7)$$

尽管补充信息第 2.6 节的消融实验表明，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，使内容更易读。我们将语言一致性奖励直接加到最终奖励中，应用于推理和非推理数据。

需要注意的是，裁剪比例在训练中起着至关重要的作用。较低的取值可能导致大量 token 的梯度被截断，从而降低模型性能；较高的取值则可能导致训练不稳定。本阶段使用的 RL 数据详情见补充信息第 2.3 节。

第二阶段 RL 的训练细节。 具体而言，我们使用奖励信号与多样化提示分布相结合的方式训练模型。对于推理数据，我们遵循 DeepSeek-R1-Zero 中概述的方法，该方法使用基于规则的奖励来指导模型在数学、编程和逻辑推理领域的学习。在训练过程中，我们观察到 CoT 经常出现语言混用，尤其是在 RL 提示涉及多种语言时。对于通用数据，我们使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，使我们能够开发出一个不仅擅长推理，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，奖励可以表示为

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

其中

$$\text{Reward}_{\text{reasoning}} = \text{Reward}_{\text{rule}} \quad (9)$$

$$\text{Reward}_{\text{general}} = \text{Reward}_{\text{reward_model}} + \text{Reward}_{\text{format}} \quad (10)$$

RL 的第二阶段保留了第一阶段的大部分参数，主要区别是将温度降低至 0.7，因为我们发现，该阶段较高的温度会导致生成内容不连贯。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，更多

使用基于模型的偏好奖励信号进行训练可能会导致奖励投机，相关内容见补充信息第 2.5 节。

数据可用性

我们在以下地址提供拒绝采样和 RL 提示中使用的数据样本：<https://github.com/deepseek-ai/DeepSeek-R1> (<https://doi.org/10.5281/zenodo.15753193>)³³。完整数据生成方法的全面统计信息和详细说明见补充信息第 2.3 节。

代码可用性

DeepSeek-R1-Zero 和 DeepSeek-R1 的训练权重以 MIT 许可证发布于 <https://github.com/deepseek-ai/DeepSeek-R1> (<https://doi.org/10.5281/zenodo.15753193>)³³。推理脚本发布于 <https://github.com/deepseek-ai/DeepSeek-V3> (<https://doi.org/10.5281/zenodo.15753347>)³⁴。神经网络使用 PyTorch 开发³⁵，分布式框架基于我们的内部框架 HAI-LLM (<https://www.high-flyer.cn/en/blog/hai-llm>)。推理框架基于 vLLM³⁶。数据分析使用了 Python v.3.8 (<https://www.python.org/>)、NumPy v.1.23.1 (<https://github.com/numpy/numpy>)、Matplotlib v.3.5.2 (<https://github.com/matplotlib/matplotlib>) 和 TensorBoard v.2.9.1 (<https://github.com/tensorflow/tensorboard>)。

- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. Preprint at <https://doi.org/10.48550/arXiv.1707.06347> (2017).
- Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35* (eds Koyejo, S. et al.) 27730–27744 (ACM, 2022).
- Nano et al. deepseek-ai/DeepSeek-R1: v1.0.0. *Zenodo* <https://doi.org/10.5281/zenodo.15753192> (2025).
- Yu, X. et al. deepseek-ai/DeepSeek-V3: v1.0.0. *Zenodo* <https://doi.org/10.5281/zenodo.15753346> (2025).
- Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (eds Wallach, H. M. et al.) 8026–8037 (ACM, 2019).
- Kwon, W. et al. Efficient memory management for large language model serving with PagedAttention. In *Proc. ACM SIGOPS 29th Symposium on Operating Systems Principles* 611–626 (ACM, 2023).

致谢 本研究得到 DeepSeek-AI 的支持。

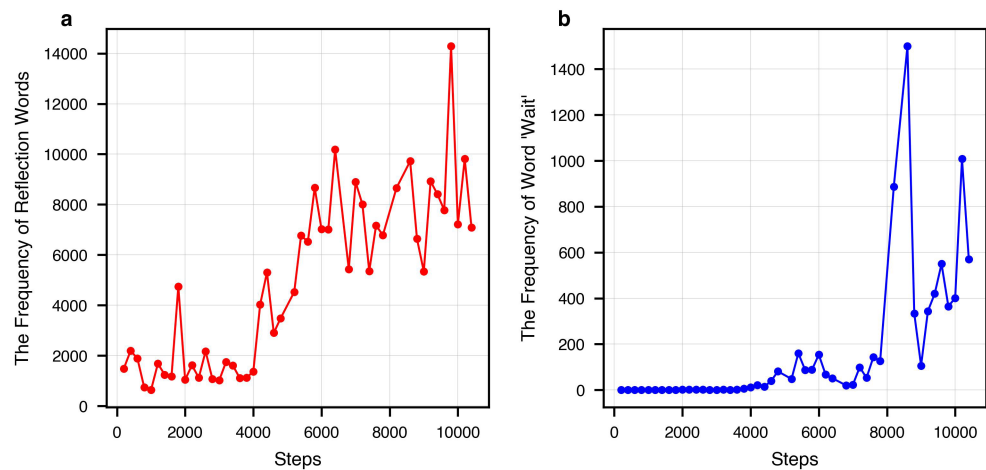
作者贡献 所有作者均为本文的发表作出了贡献，分别参与了数据收集与整理、实验设计以及 LLM 训练框架的构建。作者还参与了实验设置的实施与测试、RL 流程的改进以及结果分析。所有贡献者共同讨论并批准了科学研究结果。本文由合作团队指定的部分作者撰写，并经过内部集体评审。所有作者均审阅并批准了手稿的最终版本。核心贡献者：D.G.、D.Y.、H.Z.、J.S.、R.Z.、R.X.、Q.Z.、S.M.、P.W.、X.B.、X.Z.、X.Y.、Y.W.、Z.F.W.、Z.G.、Z.S.、Z.L.、Z.G.。这些作者被指定为核心贡献者，对本研究作出了同等贡献，姓名按字母顺序排列。其余作者也为本研究作出了重要贡献，姓名同样按字母顺序排列。

利益冲突 作者声明不存在任何利益冲突，也不会申请与本文内容相关的专利。

补充信息

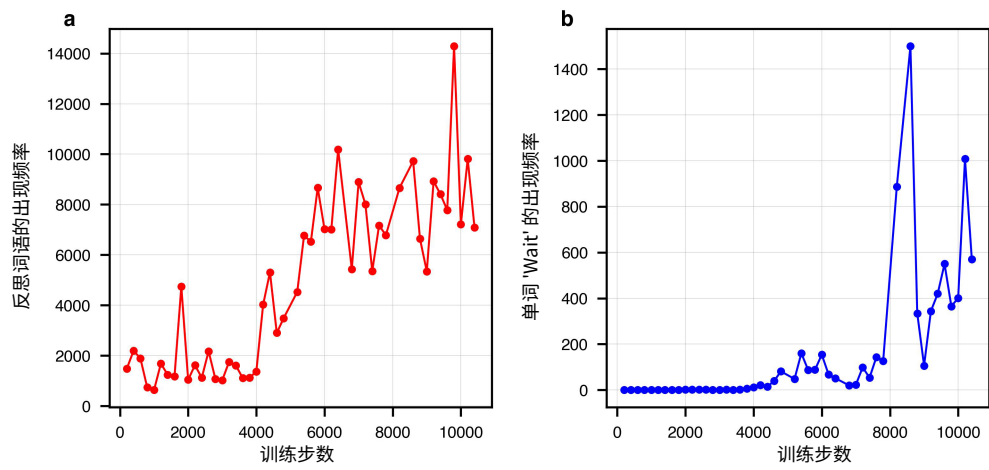
补充信息 在线版本包含补充材料，可通过以下网址获取

<https://doi.org/10.1038/s41586-025-09422-z>。
通信及材料索取 请联系 Wenfeng Liang。
同行评审信息 *Nature* 感谢 Edward Beeching, Yarin Gal, José Hernández-Orallo, Daphne Ippolito, Subbarao Kambhampati, Lewis Tunstall, Yiming Zhang 和 Lexin Zhou 对本文同行评审工作的贡献。同行评审报告可供查阅。
重印与许可信息 可在以下网址获取 <http://www.nature.com/reprints>。



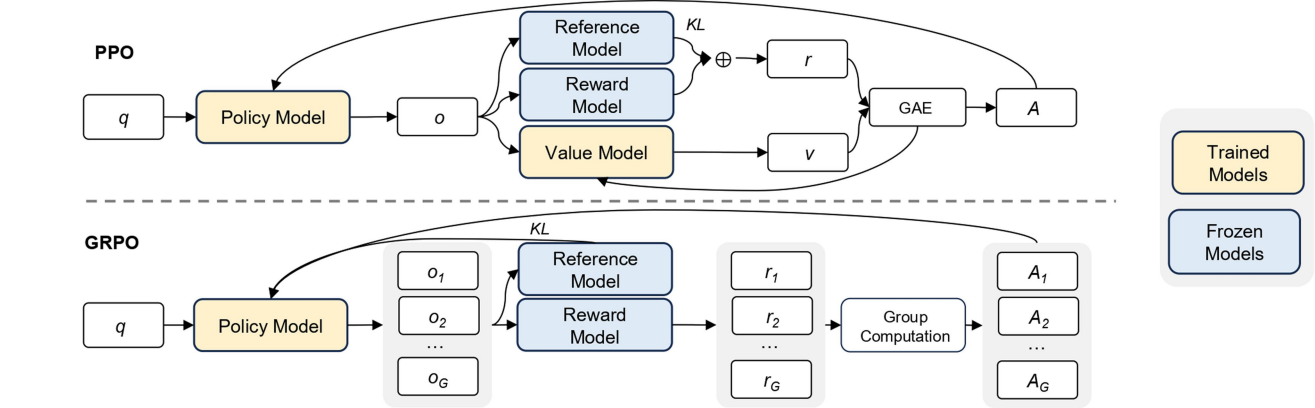
Extended Data Fig. 1| Evolution of reasoning-related linguistic features in model outputs across training steps. **a**, Frequency of representative reflective terms in model-generated outputs throughout the training process. Reflective terms—including ‘wait’, ‘mistake’, ‘however’, ‘but’, ‘retry’, ‘error’, ‘verify’, ‘wrong’, ‘evaluate’ and ‘check’—were identified and curated by a panel of three human experts. Each expert independently proposed a set of words indicative of reflective reasoning, which were subsequently consolidated

through consensus into a final vocabulary list. **b**, Frequency of the term ‘wait’ in model outputs over the course of training. This term was virtually absent during the initial training stages, appeared sporadically between steps 4,000 and 7,000 and exhibited a marked increase in frequency after step 8,000. These trends suggest the emergence of temporal reasoning or self-monitoring behaviour as training progresses.



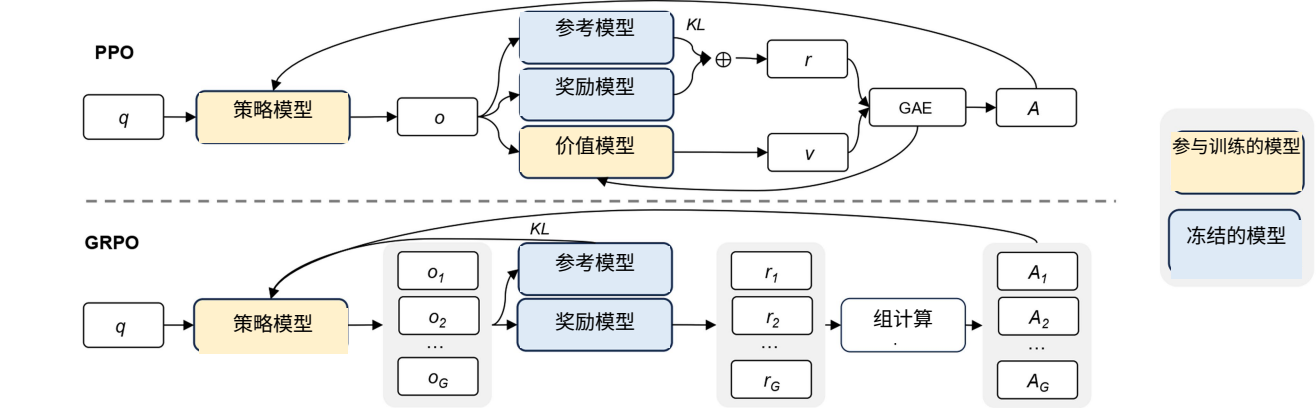
扩展数据图 1 | 模型输出中与推理相关的语言特征随训练步数的变化。a，训练过程中模型生成的输出中具有代表性的反思词语的出现频率。反思词语包括 ‘wait’ 、 ‘mistake’ 、 ‘however’ 、 ‘but’ 、 ‘retry’ 、 ‘error’ 、 ‘verify’ 、 ‘wrong’ 、 ‘evaluate’ 和 ‘check’ ，由三位人类专家组成的小组识别并整理。每位专家独立提出一组表示反思性推理的词语，随后通过协商一致将其汇总为最终词汇表。

b，模型输出中 ‘wait’ 一词在训练过程中的出现频率。在训练初期，该词几乎没有出现；在 4,000 至 7,000 步之间零星出现，并在 8,000 步之后显著增多。这些趋势表明，随着训练推进，模型开始呈现时间推理或自我监控行为。



Extended Data Fig. 2 | Illustration of the proposed GRPO for RL-based training. In the proposed framework, a LLM is used as a policy model to generate responses $\{o_1, o_2, \dots, o_G\}$ conditioned on a given query q . Each response within the group is evaluated by a reward model—either learned (model-based) or manually specified (rule-based)—to assign a scalar reward signal. Subsequently, GRPO computes the relative advantages of each group member based on their assigned rewards. Rather than relying on an explicit value function, as in PPO,

GRPO directly estimates advantages from the intra-group reward distribution. The policy parameters are then updated to maximize the expected reward while simultaneously minimizing divergence from a reference policy, typically quantified through the KL divergence. By eliminating the need for a separate value network, GRPO offers a simplified yet effective alternative to traditional actor-critic methods such as PPO.



扩展数据图。 2 | 基于强化学习的 GRPO 提案示意图
训练。在所提出的框架中，LLM 被用作策略模型，根据给定查询生成响应 $\{o_1, o_2, \dots, o_G\}$ ，条件是给定查询 q 。组内的每个响应都会由奖励模型（通过学习获得的模型型奖励模型或人工指定的规则型奖励模型）进行评估，以分配一个标量奖励信号。随后，GRPO 根据分配的奖励计算组内各成员的相对优势。与 PPO 不同，它不依赖显式的价值函数，

GRPO 直接根据组内奖励分布估计优势。随后更新策略参数，以最大化期望奖励，同时尽量减少与参考策略的差异，通常用 KL 散度量化。通过免去单独的价值网络，GRPO 为 PPO 等传统的演员-评论家方法提供了一种更简化且有效的替代方案。

Extended Data Table 1 | An illustrative example from the AIME dataset

PROMPT
Let $b \geq 2$ be an integer. Call a positive integer n <i>b-eautiful</i> if it has exactly two digits when expressed in base b , and these two digits sum to $\sqrt{n}$. For example, 81 is 13-eautiful because $81 = \underline{6} \underline{3}_{13}$ and $6 + 3 = \sqrt{81}$. Find the least integer $b \geq 2$ for which there are more than ten b -eautiful integers. Please reason step by step, and put your final answer within <code>\boxed{}</code> .
Evaluation
Parse the final answer within <code>\boxed{}</code> and use a rule-based grader to determine if it equals the ground truth. Round numerical values as needed, and use ‘SymPy’ to parse expressions.

The presented question is sourced from the 2024 AIME. The model is tasked with solving the problem and formatting its answer in a required format (for example, ANSWER). For evaluation, a rule-based grading system is used to determine correctness. The output of the model is considered correct if and only if it exactly matches the ground-truth solution.

扩展数据表 1 | AIME 数据集中的一个示例

PROMPT
Let $b \geq 2$ be an integer. Call a positive integer n <i>b-eautiful</i> if it has exactly two digits when expressed in base b , and these two digits sum to $\sqrt{n}$. For example, 81 is 13-eautiful because $81 = \underline{6} \underline{3}_{13}$ and $6 + 3 = \sqrt{81}$. Find the least integer $b \geq 2$ for which there are more than ten b -eautiful integers. Please reason step by step, and put your final answer within <code>\boxed{}</code> .
Evaluation
Parse the final answer within <code>\boxed{}</code> and use a rule-based grader to determine if it equals the ground truth. Round numerical values as needed, and use ‘SymPy’ to parse expressions.

所示问题取自 2024 年 AIME。模型需要解答该问题，并按照规定格式（例如 ANSWER）组织答案。评估时，使用基于规则的评分系统判断答案是否正确。仅当模型输出与标准答案完全一致时，才视为正确。

扩展数据表 1：图片表格的中文译文

提示词

设 $b \geq 2$ 为一个整数。如果一个正整数 n 用 b 进制表示时恰有两位，且这两位数字之和等于 $\sqrt{n}$ ，则称 n 为 b -美丽数。例如，81 是 13-美丽数，因为 $81 = 63$ （13 进制），且 $6 + 3 = \sqrt{81}$ 。求使得 b -美丽数多于十个的最大整数 $b \geq 2$ 。请逐步推理，并将最终答案写在 `\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。必要时对数值进行四舍五入，并使用 SymPy 解析表达式。

非官方英中翻译与排版修订样稿。原文：Guo et al., Nature 645, 633–638 (2025). DOI: 10.1038/s41586-025-09422-z。© The Author(s) 2025。许可：CC BY 4.0, <https://creativecommons.org/licenses/by/4.0/>。译文使用 gpt-6-luna；数学对象保留原字形，图内文字按审阅坐标替换，图片表格提供中文转录（基数下标改写为“13进制”说明，数学含义不变）；这些改动不表示作者或出版商认可。