

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

DeepSeek-AI

research@deepseek.com

## Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

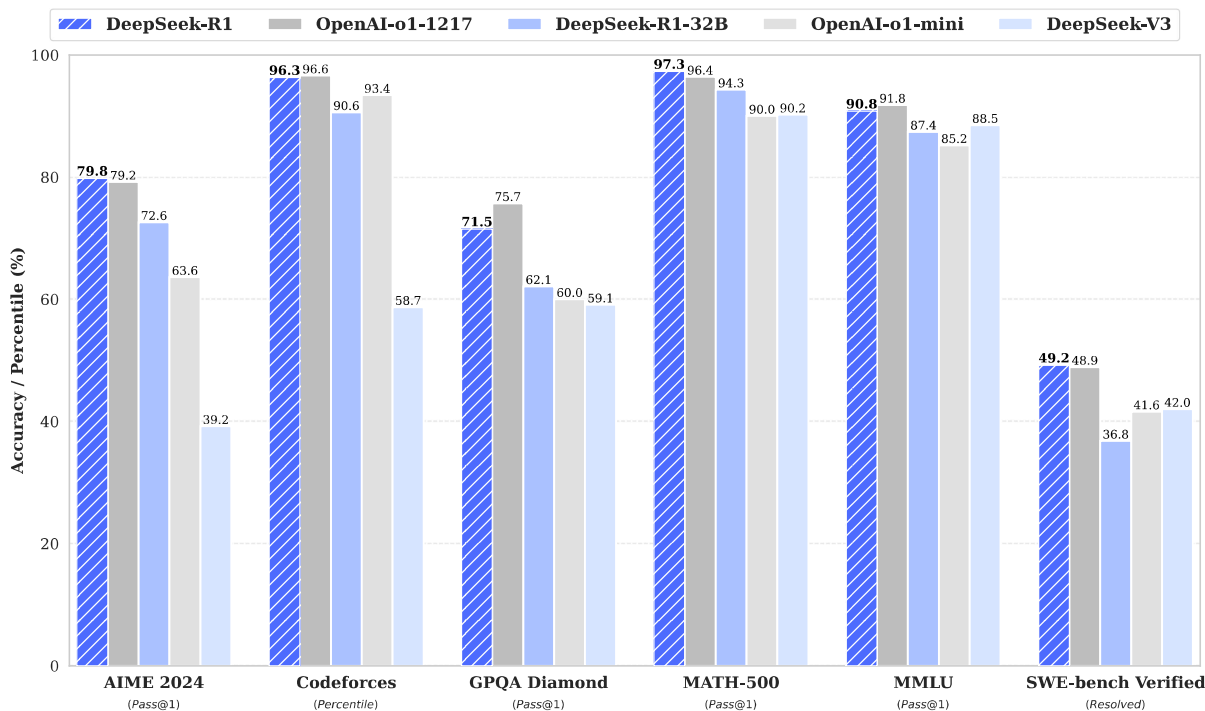


Figure 1 | Benchmark performance of DeepSeek-R1.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Contributions . . . . .	4
1.2	Summary of Evaluation Results . . . . .	4
<b>2</b>	<b>Approach</b>	<b>5</b>
2.1	Overview . . . . .	5
2.2	DeepSeek-R1-Zero: Reinforcement Learning on the Base Model . . . . .	5
2.2.1	Reinforcement Learning Algorithm . . . . .	5
2.2.2	Reward Modeling . . . . .	6
2.2.3	Training Template . . . . .	6
2.2.4	Performance, Self-evolution Process and Aha Moment of DeepSeek-R1-Zero . . . . .	6
2.3	DeepSeek-R1: Reinforcement Learning with Cold Start . . . . .	9
2.3.1	Cold Start . . . . .	9
2.3.2	Reasoning-oriented Reinforcement Learning . . . . .	10
2.3.3	Rejection Sampling and Supervised Fine-Tuning . . . . .	10
2.3.4	Reinforcement Learning for all Scenarios . . . . .	11
2.4	Distillation: Empower Small Models with Reasoning Capability . . . . .	11
<b>3</b>	<b>Experiment</b>	<b>11</b>
3.1	DeepSeek-R1 Evaluation . . . . .	13
3.2	Distilled Model Evaluation . . . . .	14
<b>4</b>	<b>Discussion</b>	<b>14</b>
4.1	Distillation v.s. Reinforcement Learning . . . . .	14
4.2	Unsuccessful Attempts . . . . .	15
<b>5</b>	<b>Conclusion, Limitations, and Future Work</b>	<b>16</b>
<b>A</b>	<b>Contributions and Acknowledgments</b>	<b>20</b>

- **Others:** DeepSeek-R1 also excels in a wide range of tasks, including creative writing, general question answering, editing, summarization, and more. It achieves an impressive length-controlled win-rate of 87.6% on AlpacaEval 2.0 and a win-rate of 92.3% on ArenaHard, showcasing its strong ability to intelligently handle non-exam-oriented queries. Additionally, DeepSeek-R1 demonstrates outstanding performance on tasks requiring long-context understanding, substantially outperforming DeepSeek-V3 on long-context benchmarks.

## 2. Approach

### 2.1. Overview

Previous work has heavily relied on large amounts of supervised data to enhance model performance. In this study, we demonstrate that reasoning capabilities can be significantly improved through large-scale reinforcement learning (RL), even without using supervised fine-tuning (SFT) as a cold start. Furthermore, performance can be further enhanced with the inclusion of a small amount of cold-start data. In the following sections, we present: (1) DeepSeek-R1-Zero, which applies RL directly to the base model without any SFT data, and (2) DeepSeek-R1, which applies RL starting from a checkpoint fine-tuned with thousands of long Chain-of-Thought (CoT) examples. 3) Distill the reasoning capability from DeepSeek-R1 to small dense models.

### 2.2. DeepSeek-R1-Zero: Reinforcement Learning on the Base Model

Reinforcement learning has demonstrated significant effectiveness in reasoning tasks, as evidenced by our previous works (Shao et al., 2024; Wang et al., 2023). However, these works heavily depended on supervised data, which are time-intensive to gather. In this section, we explore the potential of LLMs to develop reasoning capabilities **without any supervised data**, focusing on their self-evolution through a pure reinforcement learning process. We start with a brief overview of our RL algorithm, followed by the presentation of some exciting results, and hope this provides the community with valuable insights.

#### 2.2.1. Reinforcement Learning Algorithm

**Group Relative Policy Optimization** In order to save the training costs of RL, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question  $q$ , GRPO samples a group of outputs  $\{o_1, o_2, \dots, o_G\}$  from the old policy  $\pi_{\theta_{old}}$  and then optimizes the policy model  $\pi_{\theta}$  by maximizing the following objective:

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

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

where  $\varepsilon$  and  $\beta$  are hyper-parameters, and  $A_i$  is the advantage, computed using a group of rewards  $\{r_1, r_2, \dots, r_G\}$  corresponding to the outputs within each group:

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