# DeepSeek-R1

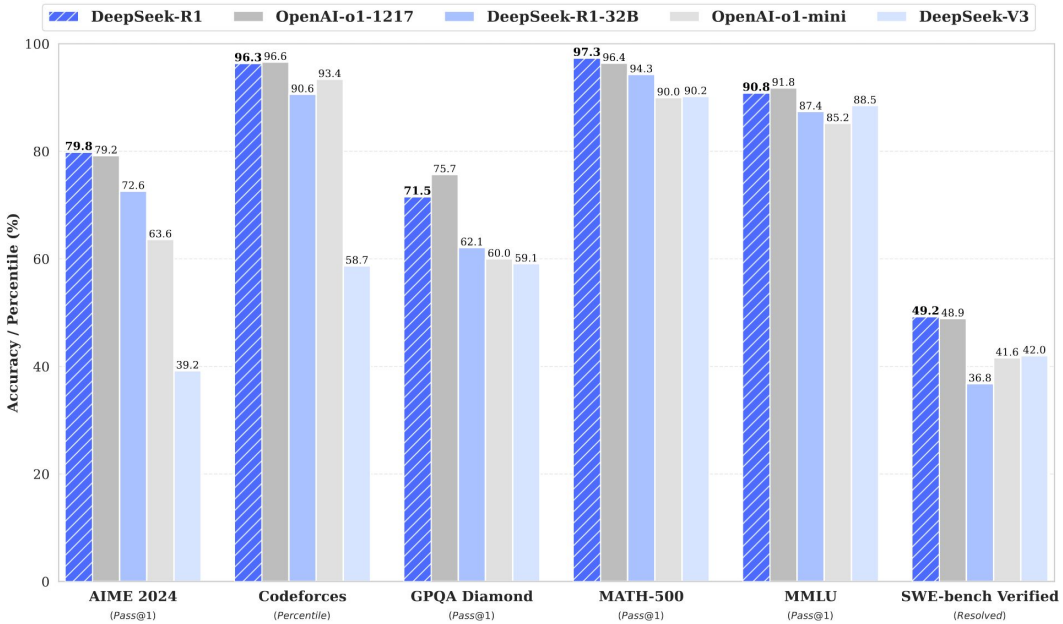


Figure 1 | Benchmark performance of DeepSeek-R1.

# Contributions

1. Post-Training
    a. DeepSeek-R1-Zero
    b. DeepSeek-R1
2. Distillation

# Contributions

1. Post-Training
    a. **DeepSeek-R1-Zero** ← RL over SFT (no fine tuning cold start)
    b. DeepSeek-R1
2. Distillation

# Contributions

1. Post-Training
    a. **DeepSeek-R1-Zero** ← RL over SFT (no fine tuning cold start)
    b. DeepSeek-R1 ← Improved readability via fine tuning pipeline
2. Distillation

# Contributions

1. Post-Training
    a. **DeepSeek-R1-Zero** ← RL over SFT (no fine tuning cold start)
    b. DeepSeek-R1 ← Improved readability via fine tuning pipeline
2. Distillation ← Distill reasoning into smaller models

# DeepSeek-R1-Zero: RL on the Base Model

- Reasoning capabilities with no supervised data
- Group Relative Policy Optimization (GRPO)
- Two types of rewards:
  - Accuracy rewards
    - ex) correct final answer to math question
  - Format rewards
    - Enforces model put scratchpad work between <think></think> tags

| |
|---|
| 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, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant: |

Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

# DeepSeek-R1-Zero: RL on the Base Model

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

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

# DeepSeek-R1-Zero: GRPO

question

group of outputs $\{o1, o2, \cdots, oG\}$ from the old policy $\pi_{\theta old}$

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

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

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

# DeepSeek-R1-Zero: GRPO

question

group of outputs $\{o1, o2, \cdots, oG\}$ from the old policy $\pi_{\theta old}$

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

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

Adjust policy according to $A_i$, Regularized by past policy

Constrain ratio

Trust region

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

# DeepSeek-R1-Zero: GRPO

question

group of outputs $\{o1, o2, \cdots, oG\}$ from the old policy $\pi_{\theta old}$

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

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

Adjust policy according to $A_i$, Regularized by past policy

Constrain ratio

Trust region

Advantage estimation
Usually defined as
A(s,a)=Q(s,a) - V(s)

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

Use reward samples as baseline

# DeepSeek-R1-Zero: Results

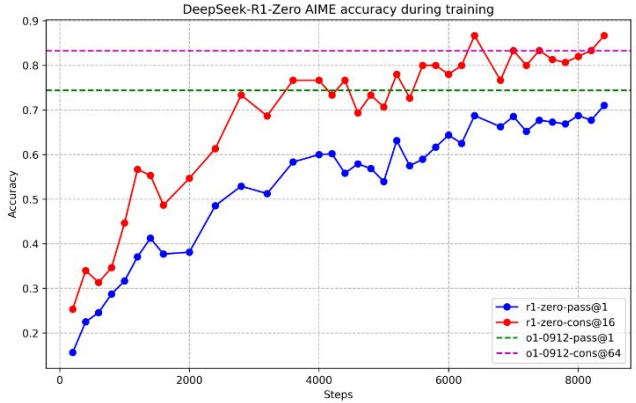


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

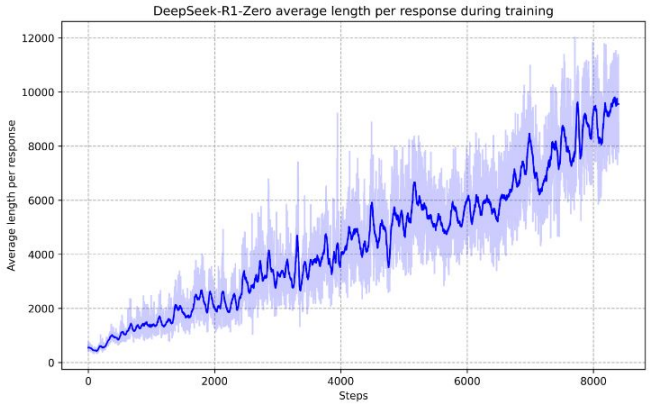


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

# DeepSeek-R1-Zero: Results

“rather than explicitly teaching the model on how to solve a problem, we simply provide it with the right incentives, and it autonomously develops advanced problem-solving strategies.”

| |
|---|
| Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to |
| Response: <think><br>To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$<br>$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.<br>Rearrange to isolate the inner square root term:<br>$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$<br>$\ldots$<br>Wait, wait. Wait. That's an aha moment I can flag here.<br>Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$<br>We started with the equation:<br>$\sqrt{a - \sqrt{a + x}} = x$<br>First, let's square both sides:<br>$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$<br>Next, I could square both sides again, treating the equation: $\cdots$<br>$\ldots$ |

Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. 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 reinforcement learning.

# DeepSeek-R1: Addressing drawbacks

- Readable CoT reasoning
  - Readability
  - Language mixing
- Improve results with small but high quality SFT dataset
- Four stage pipeline:
  1) Cold start fine tuning
  2) RL for reasoning
  3) Rejection sampling and SFT
  4) Final RL step

# DeepSeek-R1: Pipeline

1. **Cold Start Fine Tuning**
    a. Collect few thousands of long CoT data to finetune base model
        i. Few-shot prompting other models with long CoT example
        ii. Prompt models to generate answers with reflections
        iii. gather/filter R1-zero outputs that are readable and fix via human annotators
    b. Starting point for RL actor

Advantages:

- Readable reasoning scratchpad
- Cold-start with human priors leads to better performance against DeepSeek-R1-Zero

# DeepSeek-R1: Pipeline

1. Cold Start Fine Tuning
2. **RL for reasoning**
    a. Same RL as presented in DeepSeek-R1-Zero
    b. Negative reward for language mixing
3. Rejection sampling and SFT
4. Final RL step

# DeepSeek-R1: Pipeline

1. Cold Start Fine Tuning
2. RL for reasoning
3. **Rejection sampling and SFT**
    a. Collect 800k datapoints for finetuning:
        *i. Reasoning Data*
            1. 600k samples
            2. Rejection sampling from step 2 checkpoint
                a. filter unreadable/incorrect reasoning out
            3. Data from ground-truth + model predictions fed to LLM evaluator
        *ii. Non-Reasoning Data*
            1. 200k samples
            2. Traditional datasets on writing, factual QA, self-cognition, and translation, etc.
4. Final RL step

# DeepSeek-R1: Pipeline

1. Cold Start Fine Tuning
2. RL for reasoning
3. Rejection sampling and SFT
4. **Final RL step**
    a. Combination of:
        i. Step 2 again
        ii. Traditional RLHF

# DeepSeek-R1: Results

### 3.1. DeepSeek-R1 Evaluation

| | Benchmark (Metric) | Claude-3.5-Sonnet-1022 | GPT-4o 0513 | DeepSeek V3 | OpenAI o1-mini | OpenAI o1-1217 | DeepSeek R1 |
|---|---|---|---|---|---|---|---|
| | Architecture | - | - | MoE | - | - | MoE |
| | # Activated Params | - | - | 37B | - | - | 37B |
| | # Total Params | - | - | 671B | - | - | 671B |
| English | MMLU (Pass@1) | 88.3 | 87.2 | 88.5 | 85.2 | **91.8** | 90.8 |
| | MMLU-Redux (EM) | 88.9 | 88.0 | 89.1 | 86.7 | - | **92.9** |
| | MMLU-Pro (EM) | 78.0 | 72.6 | 75.9 | 80.3 | - | **84.0** |
| | DROP (3-shot F1) | 88.3 | 83.7 | 91.6 | 83.9 | 90.2 | **92.2** |
| | IF-Eval (Prompt Strict) | **86.5** | 84.3 | 86.1 | 84.8 | - | 83.3 |
| | GPQA Diamond (Pass@1) | 65.0 | 49.9 | 59.1 | 60.0 | **75.7** | 71.5 |
| | SimpleQA (Correct) | 28.4 | 38.2 | 24.9 | 7.0 | **47.0** | 30.1 |
| | FRAMES (Acc.) | 72.5 | 80.5 | 73.3 | 76.9 | - | **82.5** |
| | AlpacaEval2.0 (LC-winrate) | 52.0 | 51.1 | 70.0 | 57.8 | - | **87.6** |
| | ArenaHard (GPT-4-1106) | 85.2 | 80.4 | 85.5 | 92.0 | - | **92.3** |
| Code | LiveCodeBench (Pass@1-COT) | 38.9 | 32.9 | 36.2 | 53.8 | 63.4 | **65.9** |
| | Codeforces (Percentile) | 20.3 | 23.6 | 58.7 | 93.4 | **96.6** | 96.3 |
| | Codeforces (Rating) | 717 | 759 | 1134 | 1820 | **2061** | 2029 |
| | SWE Verified (Resolved) | **50.8** | 38.8 | 42.0 | 41.6 | 48.9 | 49.2 |
| | Aider-Polyglot (Acc.) | 45.3 | 16.0 | 49.6 | 32.9 | **61.7** | 53.3 |
| Math | AIME 2024 (Pass@1) | 16.0 | 9.3 | 39.2 | 63.6 | 79.2 | **79.8** |
| | MATH-500 (Pass@1) | 78.3 | 74.6 | 90.2 | 90.0 | 96.4 | **97.3** |
| | CNMO 2024 (Pass@1) | 13.1 | 10.8 | 43.2 | 67.6 | - | **78.8** |
| Chinese | CLUEWSC (EM) | 85.4 | 87.9 | 90.9 | 89.9 | - | **92.8** |
| | C-Eval (EM) | 76.7 | 76.0 | 86.5 | 68.9 | - | **91.8** |
| | C-SimpleQA (Correct) | 55.4 | 58.7 | **68.0** | 40.3 | - | 63.7 |

Table 4 | Comparison between DeepSeek-R1 and other representative models.

# Distillation

- Directly finetune open source smaller models using the 800k samples from Step 3 in the pipeline

### 3.2. Distilled Model Evaluation

| Model | AIME 2024 | | MATH-500 | GPQA Diamond | LiveCode Bench | CodeForces |
|---|---|---|---|---|---|---|
| | pass@1 | cons@64 | pass@1 | pass@1 | pass@1 | rating |
| **GPT-4o-0513** | 9.3 | 13.4 | 74.6 | 49.9 | 32.9 | 759 |
| **Claude-3.5-Sonnet-1022** | 16.0 | 26.7 | 78.3 | 65.0 | 38.9 | 717 |
| **OpenAI-o1-mini** | 63.6 | 80.0 | 90.0 | 60.0 | 53.8 | **1820** |
| **QwQ-32B-Preview** | 50.0 | 60.0 | 90.6 | 54.5 | 41.9 | 1316 |
| **DeepSeek-R1-Distill-Qwen-1.5B** | 28.9 | 52.7 | 83.9 | 33.8 | 16.9 | 954 |
| **DeepSeek-R1-Distill-Qwen-7B** | 55.5 | 83.3 | 92.8 | 49.1 | 37.6 | 1189 |
| **DeepSeek-R1-Distill-Qwen-14B** | 69.7 | 80.0 | 93.9 | 59.1 | 53.1 | 1481 |
| **DeepSeek-R1-Distill-Qwen-32B** | **72.6** | 83.3 | 94.3 | 62.1 | 57.2 | 1691 |
| **DeepSeek-R1-Distill-Llama-8B** | 50.4 | 80.0 | 89.1 | 49.0 | 39.6 | 1205 |
| **DeepSeek-R1-Distill-Llama-70B** | 70.0 | **86.7** | **94.5** | **65.2** | **57.5** | 1633 |

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

# Distillation vs. RL

- large-scale RL training on Qwen-32B-Base using math, code, and STEM data, training for over 10K steps, resulting in DeepSeek-R1-Zero-Qwen-32B

| Model | AIME 2024 | | MATH-500 | GPQA Diamond | LiveCodeBench |
|---|---|---|---|---|---|
| | pass@1 | cons@64 | pass@1 | pass@1 | pass@1 |
| **QwQ-32B-Preview** | 50.0 | 60.0 | 90.6 | 54.5 | 41.9 |
| **DeepSeek-R1-Zero-Qwen-32B** | 47.0 | 60.0 | 91.6 | 55.0 | 40.2 |
| **DeepSeek-R1-Distill-Qwen-32B** | **72.6** | **83.3** | **94.3** | **62.1** | **57.2** |

Table 6 | Comparison of distilled and RL Models on Reasoning-Related Benchmarks.

# Distillation vs. RL

“... we can draw two conclusions: First, distilling more powerful models into smaller ones yields excellent results, whereas smaller models relying on the large-scale RL mentioned in this paper require enormous computational power and may not even achieve the performance of distillation. Second, while distillation strategies are both economical and effective, **advancing beyond the boundaries of intelligence may still require more powerful base models and large scale reinforcement learning.**”