

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

A Novel Approach to Enhancing Large Language Models

Presented by: Ali Hamza

HSE University
Department of Computer Science



February 3, 2025

Introduction: Evolution of LLMs

- **Rapid Advancements:**

- Recent years have seen rapid advancements in Large Language Models (LLMs) (Anthropic, 2024; Google, 2024; OpenAI, 2024a).
- These models are progressively narrowing the gap toward Artificial General Intelligence (AGI).

- **Capabilities:**

- LLMs demonstrate remarkable capabilities in understanding and generating human-like text.
- They are increasingly being applied to complex tasks such as reasoning, coding, and scientific analysis.

Introduction: Post-Training for Reasoning

- **Post-Training:**

- Post-training has emerged as a critical component of the LLM training pipeline (OpenAI, 2024b).
- It enhances reasoning, aligns models with social values, and adapts them to user preferences.

- **Chain-of-Thought (CoT) Reasoning:**

- OpenAI's o1 series introduced inference-time scaling via CoT reasoning.
- This approach has achieved significant improvements in tasks such as mathematics, coding, and scientific reasoning.

Introduction: Challenges in Reasoning

- **Test-Time Scaling:**

- Effective test-time scaling remains an open challenge in the research community.
- Current methods struggle to generalize across diverse reasoning tasks.

- **Prior Approaches:**

- Process-based reward models (Lightman et al., 2023; Uesato et al., 2022; Wang et al., 2023).
- Reinforcement learning (Kumar et al., 2024).
- Search algorithms like Monte Carlo Tree Search (Feng et al., 2024; Trinh et al., 2024; Xin et al., 2024).

- **Limitations:**

- None of these methods has achieved general reasoning performance comparable to OpenAI's o1 series.

The Paper Contributions

- **Pure Reinforcement Learning (RL):**

- The research took the first step toward improving reasoning capabilities in LLMs using pure RL, without relying on supervised data.
- The approach focuses on self-evolution through RL.

- **Base Model and Framework:**

- The Paper use DeepSeek-V3-Base as the base model.
- Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is employed as the RL framework.

- **Open Source Models:**

- DeepSeek-R1-Zero (No supervised fine-tuning), DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

Contributions: Distillation to Smaller Models

- **Distillation:**

- The research explored distillation from DeepSeek-R1 to smaller dense models (e.g., Qwen2.5-32B).
- Distillation outperforms the application of RL directly to smaller models.

- **Key Insight:**

- Reasoning patterns discovered by larger base models are crucial for improving reasoning capabilities in smaller models.

- **Results:**

- The distilled models (14B, 32B, 70B) set new records in reasoning benchmarks.
- Outperform state-of-the-art open source models like QwQ-32B-Preview (Qwen, 2024a).

Core Concepts

- **Group Relative Policy Optimization (GRPO):**
- **Cold Start:**

Group Relative Policy Optimization (GRPO)

- **Motivation:** Traditional RL methods require a critic model, which is computationally expensive. GRPO eliminates the need for a critic by estimating the baseline from group scores.
- **Algorithm:**
 - For each question q , sample a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the old policy $\pi_{\theta_{old}}$.
 - Optimize the policy model π_{θ} by maximizing:

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

$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right),$$
$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}.$$

- **Advantages:**
 - Reduces computational cost by eliminating the critic model.
 - Encourages exploration while maintaining stability.

Cold Start: Overview

- **Definition:**

- Cold start refers to the initial phase of training where the model is fine-tuned on a small amount of high-quality, structured data before applying reinforcement learning (RL).

- **Purpose:**

- Provides a stable starting point for RL.
- Ensures the model produces readable, structured outputs.

- **Key Benefit:**

- Addresses the instability and poor readability issues observed in DeepSeek-R1-Zero.

Why Cold Start is Needed

- **DeepSeek-R1-Zero Limitations:**

- **Poor Readability:** Outputs often mix languages or lack proper formatting.
- **Instability:** Training RL directly from the base model can lead to unstable behavior.

- **Solution:**

- Fine-tune the base model on cold-start data before RL training.
- Ensures stability and readability from the start.

Cold Start Data Collection

- **Methods:**

- **Few-Shot Prompting:** Use examples of long Chain-of-Thought (CoT) reasoning.
- **Direct Prompting:** Generate detailed answers with reflection and verification.
- **Refining DeepSeek-R1-Zero Outputs:** Gather and refine outputs for readability.
- **Human Annotation:** Post-process data to ensure quality and consistency.

- **Output Format:**

- Structured format:

—special_token—|reasoning_process|—special_token—|summary|

- **Reasoning Process:** Detailed CoT reasoning.
- **Summary:** Concise summary of the reasoning results.

Advantages of Cold Start

- **Improved Readability:**

- Structured outputs are reader-friendly.
- Summaries make it easier for users to understand the reasoning process.

- **Better Performance:**

- Cold start data, designed with human priors, leads to better reasoning performance.
- Provides a strong foundation for iterative training during RL.

Cold Start in DeepSeek-R1

- **Implementation:**

- Thousands of cold-start examples are collected.
- Used to fine-tune the DeepSeek-V3-Base model.
- Fine-tuned model serves as the initial RL actor.

- **Impact:**

- Ensures stability and readability from the start.
- Helps DeepSeek-R1 achieve performance on par with OpenAI-o1-1217.

Example of Cold Start Data Format

- **Structured Output:**

—special_token—

i|think|

To solve the equation $x^2 + 5x + 6 = 0$, we can factor it as $(x + 2)(x + 3) = 0$.
Thus, the solutions are $x = -2$ and $x = -3$.

i|/think|

—special_token—

i|summary|

The solutions to the equation $x^2 + 5x + 6 = 0$ are $x = -2$ and $x = -3$.

i|/summary|

- **Key Features:**

- Clear separation of reasoning process and summary.
- Ensures readability and user-friendliness.

Methodology: DeepSeek-R1-Zero

- **Pure RL Approach:**

- No supervised fine-tuning (SFT) used.
- RL applied directly to the base model (DeepSeek-V3-Base).

- **Key Components:**

- Group Relative Policy Optimization (GRPO) for RL.
- Reward modeling: Accuracy and format rewards.
- Training template: Ensures structured reasoning and answers.

DeepSeek-R1-Zero: Results

| Model | AIME 2024 | | MATH-500 | GPQA Diamond | LiveCode Bench | CodeForces |
|------------------|-----------|---------|----------|-----------------|-------------------|------------|
| | pass@1 | cons@64 | pass@1 | pass@1 | pass@1 | rating |
| OpenAI-o1-mini | 63.6 | 80.0 | 90.0 | 60.0 | 53.8 | 1820 |
| OpenAI-o1-0912 | 74.4 | 83.3 | 94.8 | 77.3 | 63.4 | 1843 |
| DeepSeek-R1-Zero | 71.0 | 86.7 | 95.9 | 73.3 | 50.0 | 1444 |

Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

DeepSeek-R1-Zero: Results

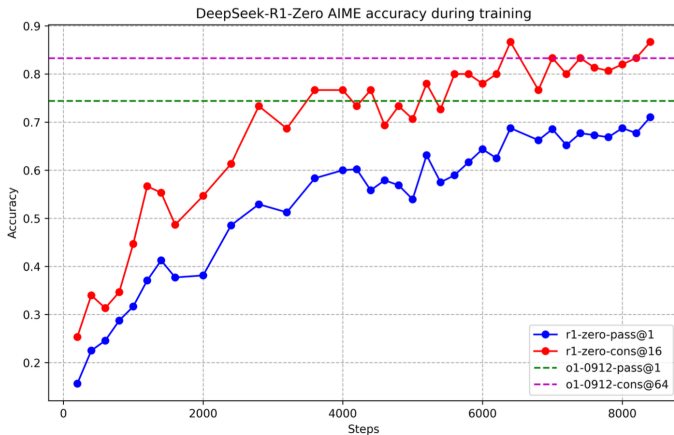


Figure: AIME accuracy of DeepSeek-R1-Zero during training

DeepSeek-R1-Zero: "Aha moment"

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 ...

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

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

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

Next, I could square both sides again, treating the equation: ...

— An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone.

Methodology: DeepSeek-R1

- **Cold-Start Data:**

- Thousands of long Chain-of-Thought (CoT) examples collected for fine-tuning.
- Improves readability and reasoning performance.

- **Multi-Stage Training:**

- Initial fine-tuning with cold-start data.
- Reasoning-oriented RL to enhance reasoning capabilities.
- Rejection sampling and supervised fine-tuning (SFT) for alignment.

- **Final RL Stage:**

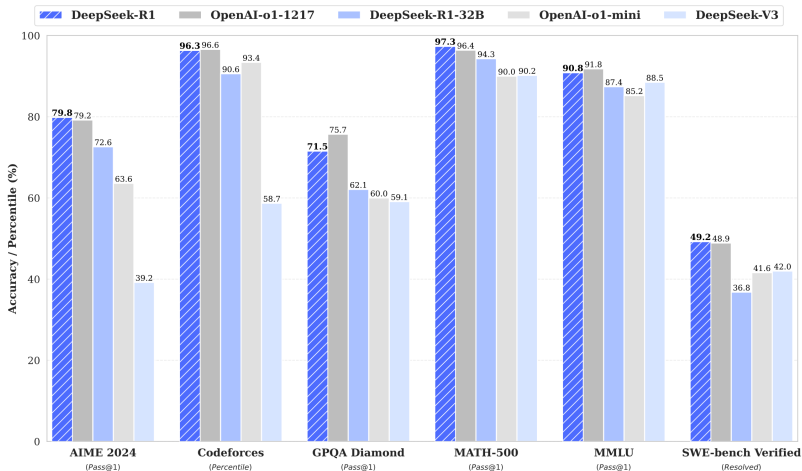
- Aligns the model with human preferences for helpfulness and harmlessness.

DeepSeek-R1: Resluts

| 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 |

— Comparison between DeepSeek-R1 and other representative models.

Results Visualization



Benchmark performance of DeepSeek-R1

DeepSeek-R1: Results

| 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 |

— Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

- **Distillation vs. RL:**

- Distillation is more efficient than large-scale RL for smaller models.
- RL requires significant computational resources but is essential for advancing reasoning capabilities.

- **Unsuccessful Attempts:**

- Process Reward Models (PRMs) and Monte Carlo Tree Search (MCTS) faced challenges in scalability and reward hacking.

Conclusion and Future Work

- **Conclusion:**

- DeepSeek-R1-Zero and DeepSeek-R1 demonstrate the potential of RL in enhancing reasoning capabilities.
- Distillation enables smaller models to achieve competitive performance.

- **Future Work:**

- Improve general capabilities (e.g., function calling, multi-turn tasks).
- Address language mixing and prompt sensitivity.
- Enhance performance on software engineering tasks.

Acknowledgments

- Core contributors and team members from DeepSeek-AI.
- Open-source community for supporting research and development.

Questions?

Thank you for your attention!

Demonstration

- **Demonstration Website:**

- Explore the live demonstration of DeepSeek-R1's reasoning capabilities.
- [Click here to visit the demonstration website.](#)

- **Demonstration Notes:**

- Detailed notes and explanations for the demonstration are available.
- [Click here to access the demonstration notes.](#)