

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

- Presenter: Abdullah Mamun
- *Date: Jan 29, 2025*



What is DeepSeek R1?



- DeepSeek R1 is an open-source large language model designed for conversational AI, coding, and problem-solving.
- It recently surpassed OpenAI's flagship reasoning model, o1, in performance across several benchmarks.

Benchmark performance

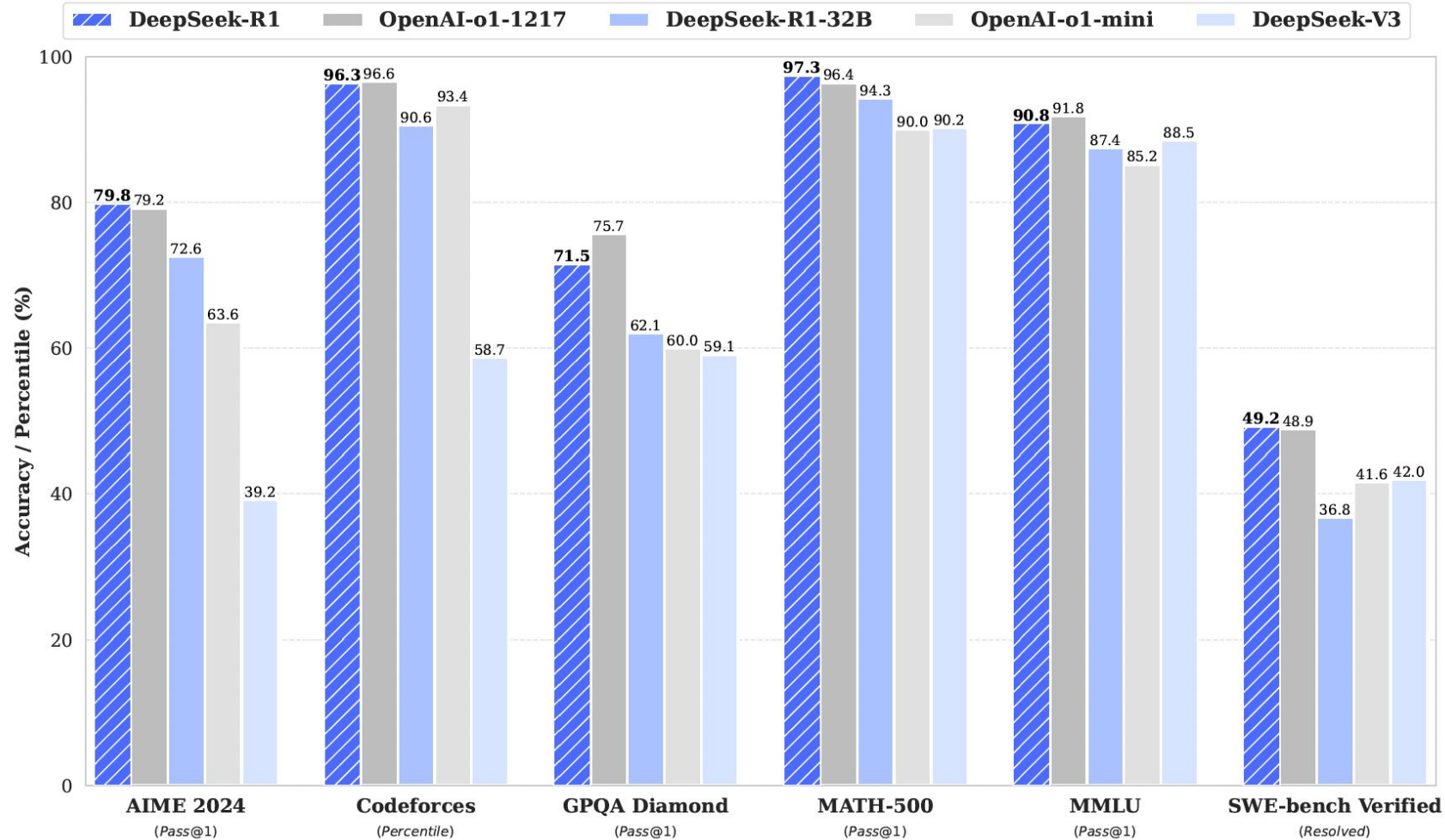


Figure 1 | Benchmark performance of DeepSeek-R1.

DeepSeek R1 vs DeepSeek V3 vs DeepSeek R1-Zero

DeepSeek R1:

- Used for complex reasoning tasks.
- Incorporates supervised fine-tuning and reinforcement learning to boost performance.

DeepSeek R1-Zero:

- Similar architecture but prioritizes zero-shot capabilities without fine-tuning.

DeepSeek V3:

- Scalable NLP with selective parameter activation for task-specific processing.

Contribution 1

Post-Training: Large-Scale Reinforcement Learning on the Base Model

- We directly apply reinforcement learning (RL) to the base model without relying on supervised fine-tuning (SFT) as a preliminary step. This approach allows the model to explore chain-of-thought (CoT) for solving complex problems, resulting in the development of DeepSeek-R1-Zero. DeepSeek-R1-Zero demonstrates capabilities such as self-verification, reflection, and generating long CoTs, marking a significant milestone for the research community. Notably, it is the first open research to validate that reasoning capabilities of LLMs can be incentivized purely through RL, without the need for SFT. This breakthrough paves the way for future advancements in this area.
- We introduce our pipeline to develop DeepSeek-R1. The pipeline incorporates two RL stages aimed at discovering improved reasoning patterns and aligning with human preferences, as well as two SFT stages that serve as the seed for the model's reasoning and non-reasoning capabilities. We believe the pipeline will benefit the industry by creating better models.

Contribution 2

Distillation: Smaller Models Can Be Powerful Too

- We demonstrate that the reasoning patterns of larger models can be distilled into smaller models, resulting in better performance compared to the reasoning patterns discovered through RL on small models. The open source DeepSeek-R1, as well as its API, will benefit the research community to distill better smaller models in the future.
- Using the reasoning data generated by DeepSeek-R1, we fine-tuned several dense models that are widely used in the research community. The evaluation results demonstrate that the distilled smaller dense models perform exceptionally well on benchmarks. DeepSeek-R1-Distill-Qwen-7B achieves 55.5% on AIME 2024, surpassing QwQ-32B-Preview. Additionally, DeepSeek-R1-Distill-Qwen-32B scores 72.6% on AIME 2024, 94.3% on MATH-500, and 57.2% on LiveCodeBench. These results significantly outperform previous open-source models and are comparable to o1-mini. We open-source distilled 1.5B, 7B, 8B, 14B, 32B, and 70B checkpoints based on Qwen2.5 and Llama3 series to the community.

Summary of Evaluation

Reasoning Tasks:

- **AIME 2024:** DeepSeek-R1 achieves 79.8% Pass@1, slightly surpassing OpenAI-o1-1217.
- **MATH-500:** Scores 97.3%, matching OpenAI-o1-1217 and significantly outperforming other models.
- **Coding-related tasks:** Attains a 2,029 Elo rating on Codeforces, outperforming 96.3% of human participants in code competitions.
- **Engineering-related tasks:** Performs slightly better than DeepSeek-V3, aiding real-world development tasks.

Knowledge Tasks:

- **MMLU Benchmarks:**
 - Scores: 90.8% on MMLU, 84.0% on MMLU-Pro, 71.5% on GPQA Diamond.
 - Outperforms DeepSeek-V3 but trails OpenAI-o1-1217.
 - Excels in educational tasks compared to other closed-source models.

SimpleQA: Outperforms DeepSeek-V3, demonstrating strong factual query handling.

- OpenAI-o1 still leads on this benchmark.

Template for DeepSeek R1-Zero

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.

Results: DeepSeek-R1-Zero vs OpenAI-o1-mini

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

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

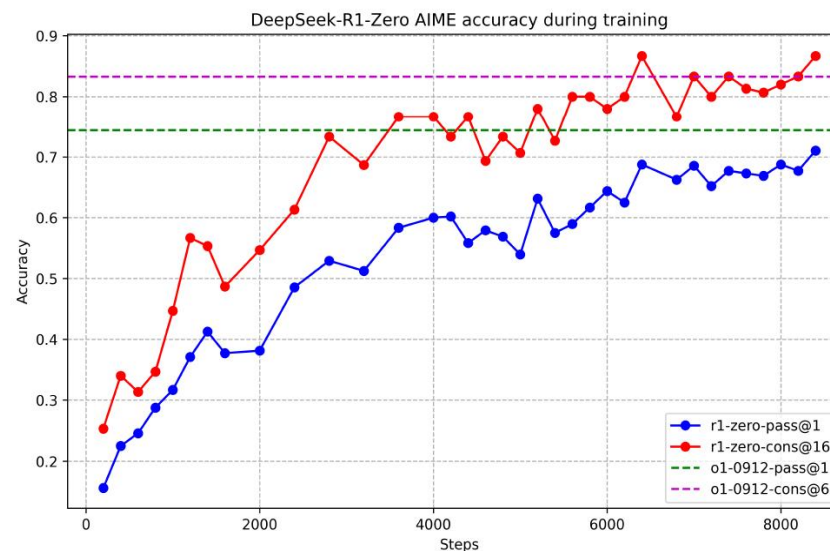


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

The Aha moment!

Aha Moment of DeepSeek-R1-Zero:

- Observed in an intermediate model version.
- Model reevaluates and allocates more thinking time.
- Shows potential advanced reasoning via reinforcement learning (RL).
- RL enables problem-solving without explicit teaching.
- Highlights RL's potential for adaptive intelligence.

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 ...

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

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: ...

...

Limitations

- **Current Limitations of DeepSeek-R1:**
- **General Capability:** Falls short of DeepSeek-V3 in tasks like function calling, multi-turn conversations, complex role-playing, and JSON output.
- **Language Mixing:** Struggles with non-English/Chinese queries, often defaulting to English for reasoning and responses.
- **Prompt Sensitivity:** Performance degrades with few-shot prompting; better results seen with zero-shot setups.
- **Software Engineering:** Limited improvement over DeepSeek-V3 due to long evaluation times and lack of large-scale RL applications.



<https://abdullah-mamun.com>
a.mamun@asu.edu