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

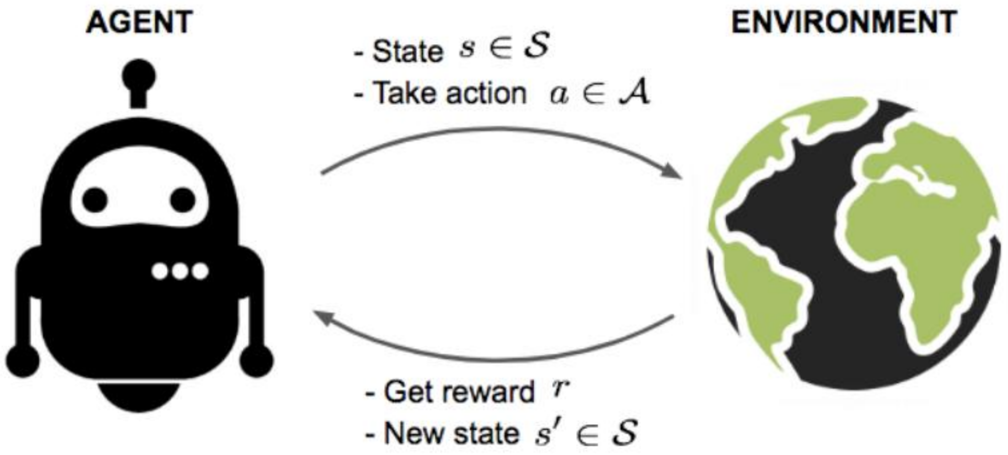
Chenye Wang

Feb 18, 2025

[1] https://github.com/deepseek-ai/DeepSeek-R1

Introduction to Reinforcement Learning

- Reinforcement Learning is concerned with how an intelligent **agent** should take actions in an **environment** to maximize the cumulative reward.


AGENT
ENVIRONMENT
- State $s \in \mathcal{S}$
- Take action $a \in \mathcal{A}$
- Get reward r
- New state $s' \in \mathcal{S}$

RL example

Agent: the cat

State: the position of the cat (x, y) in the grid

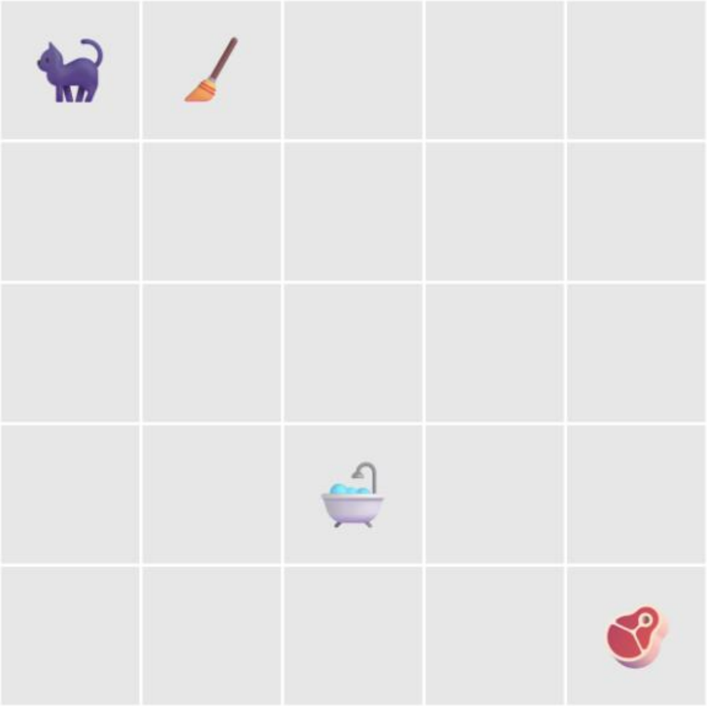
Action: at each position, the cat can move to one of the 4-directionally connected cells. If a move is invalid, the cell will not move and remain in the same position. Every time the cat makes a move, it results in a new state and a reward.

Reward model:

- A move to another empty cell results in a reward of 0.
- A move towards the broom, will result in a reward of -1.
- A move towards the bathtub will result in a reward of -10 and the cat fainting (episode over). The cat will be respawned at the initial position again.
- A move towards the meat will result in a reward of +100

Policy: a policy rules how the agent selects the action to perform given the state it is in: $a_t \sim \pi(\cdot\,|s_t)$

The goal in RL is to select a policy that maximizes the expected return when the agent acts according to it.



RL: connection to language models

Agent: the language model itself

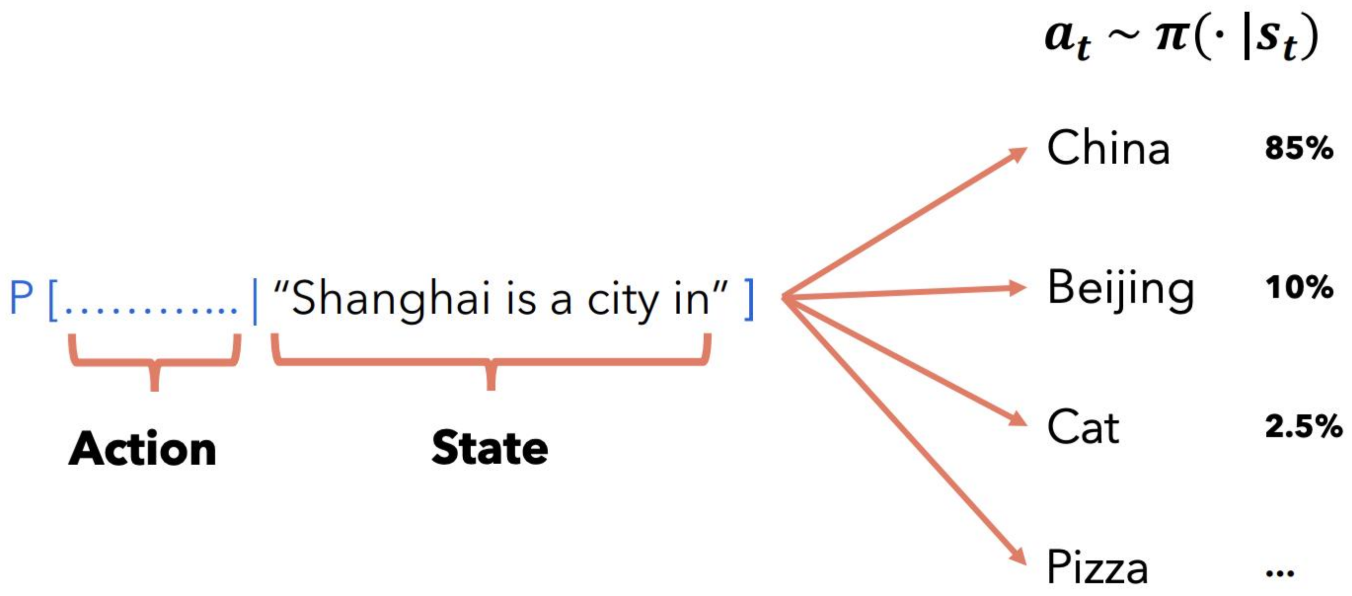
State: the prompt (input tokens)

Action: which token is selected as the next token

Reward model: the language model should be rewarded for generating “good responses” and should not receive any reward for generating “bad responses”.

Policy: In the case of language models, the policy is the language model itself! Because it models the probability of the action space given the current state of the agent: $a_t \sim \pi(\cdot\,|s_t)$

Let’s look at how we can define the reward model for our language model



$a_t \sim \pi(\cdot\,|s_t)$

P [............| “Shanghai is a city in”]

Action

State

China **85%**

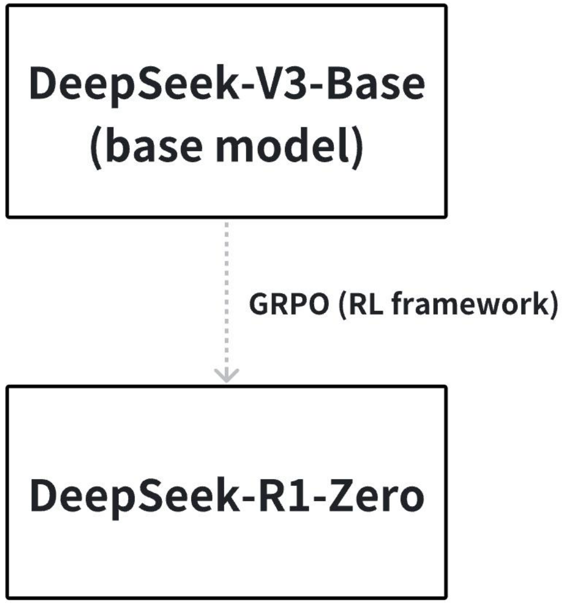
Beijing **10%**

Cat **2.5%**

Pizza **...**

DeepSeek-R1-Zero

- Our goal is to explore the potential of LLMs to develop reasoning capabilities without any supervised fine-tuning(SFT) data(time-intensive to gather), focusing on their **self-evolution through a pure RL process**.
- This approach allows the model to explore chain-of-thought (CoT) for solving complex problems.



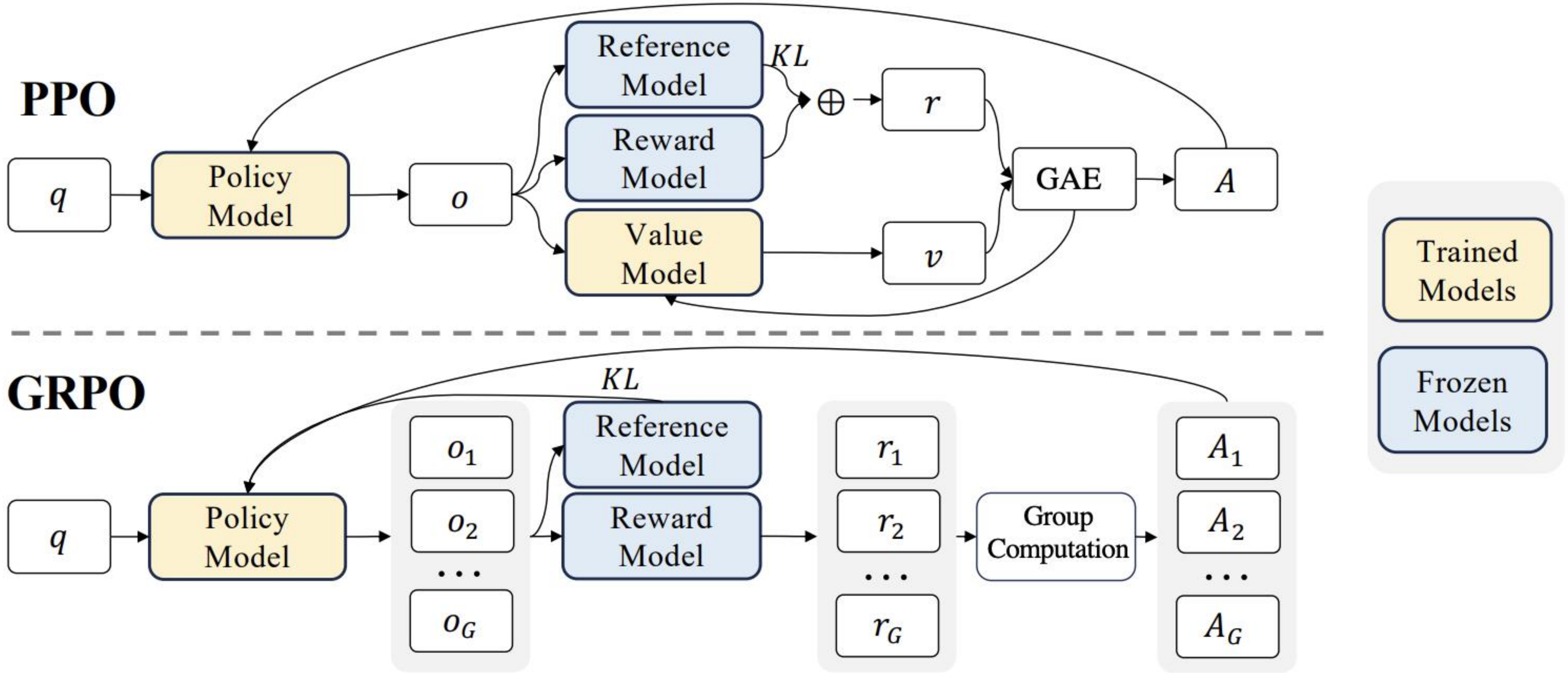
DeepSeek-V3-Base (base model)

GRPO (RL framework)

DeepSeek-R1-Zero

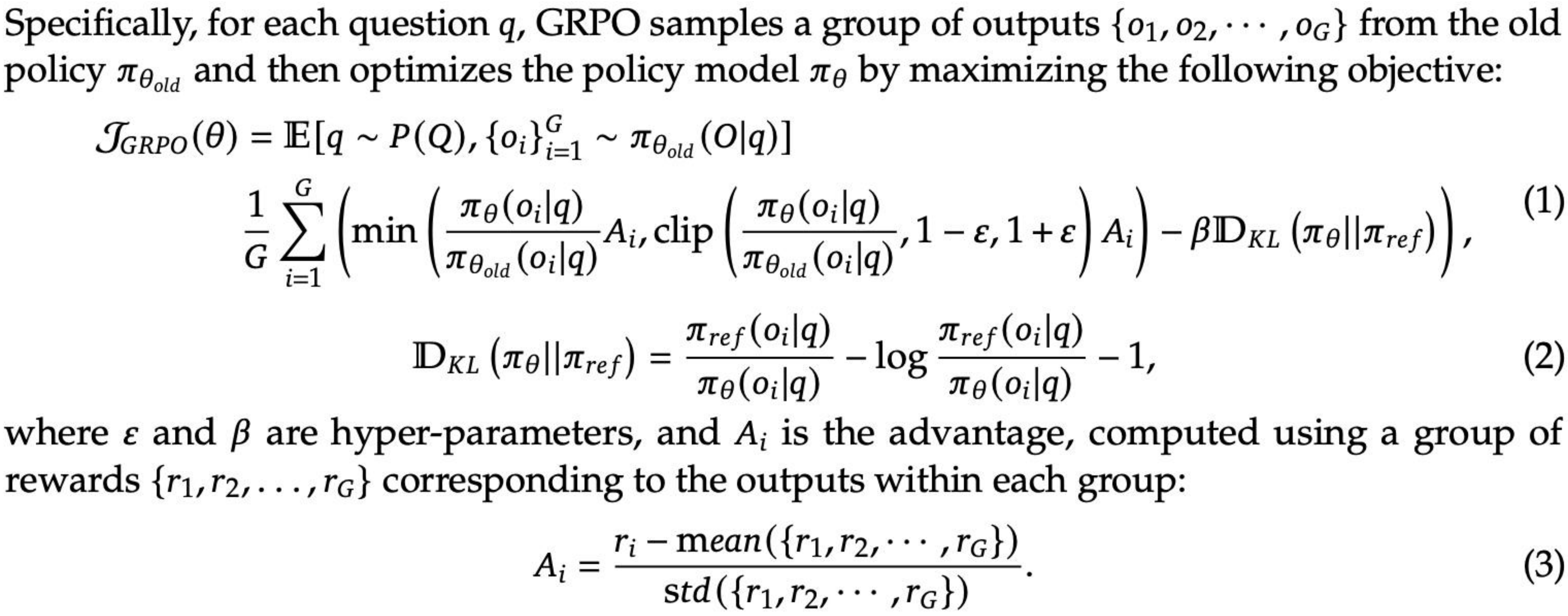
RL Algorithm -- Group Relative Policy Optimization

- Value model brings a substantial memory and computational burden.



[1] DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

RL Algorithm -- Group Relative Policy Optimization



Specifically, for each question q, GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing the following objective:

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

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

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

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

Reward Modeling

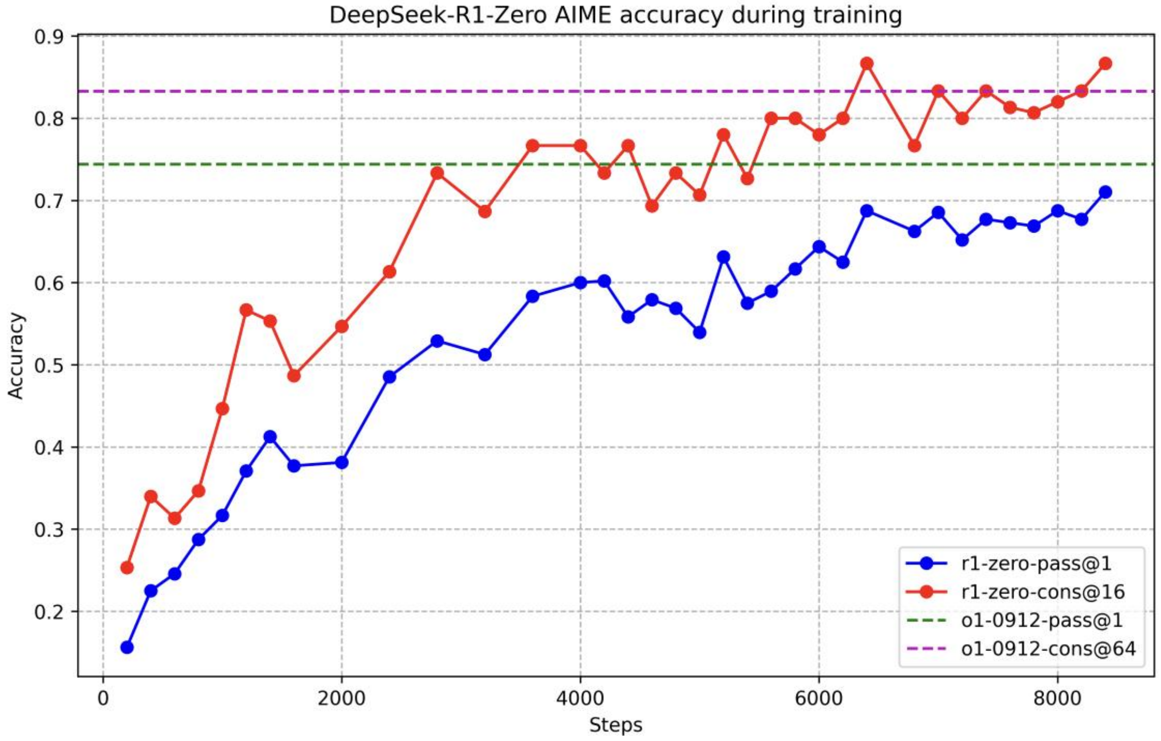
- The reward is the source of the training signal, which decides the optimization direction of RL.
- DeepSeek-R1-Zero adopts a *rule-based* reward system that mainly consists of two types of rewards:
  - **Accuracy rewards**: math problems, LeetCode problems
  - **Format rewards**: avoid content-specific biases, align with human preferences

Performance

- The findings underscore the model's ability to learn and generalize effectively through RL alone.

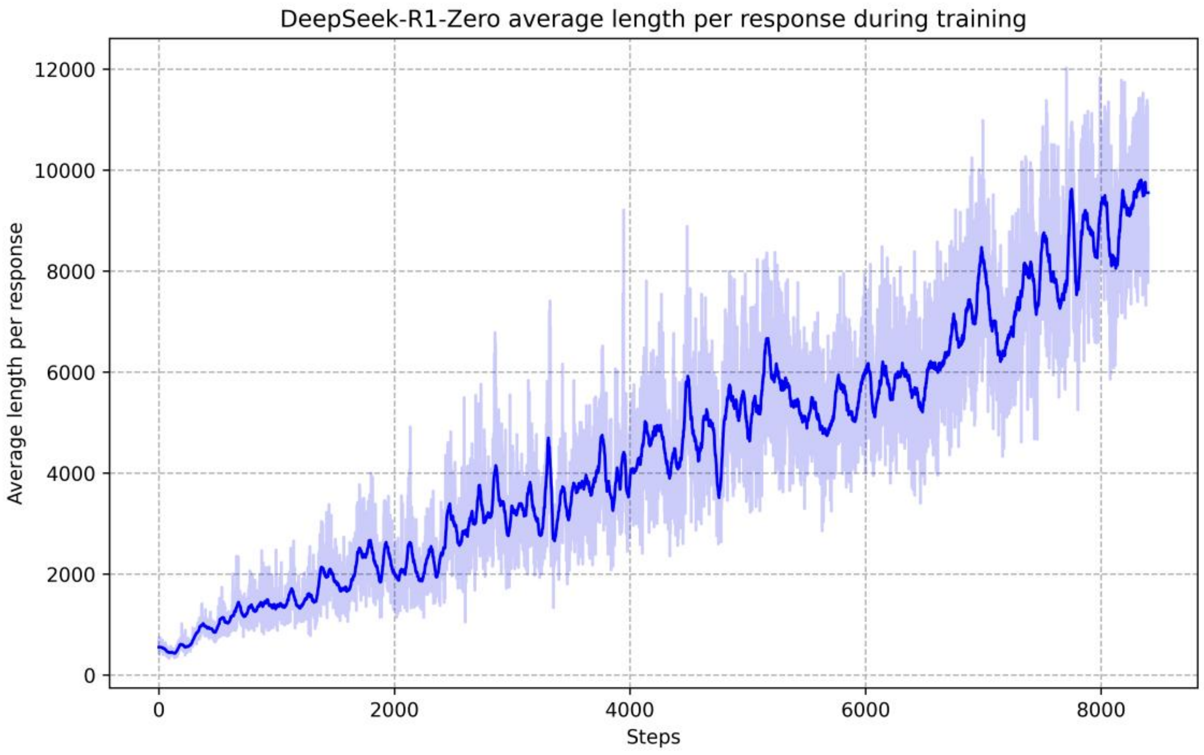
Model	**AIME 2024**		**MATH-500**	**GPQA Diamond**	**LiveCode Bench**	**CodeForces**
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

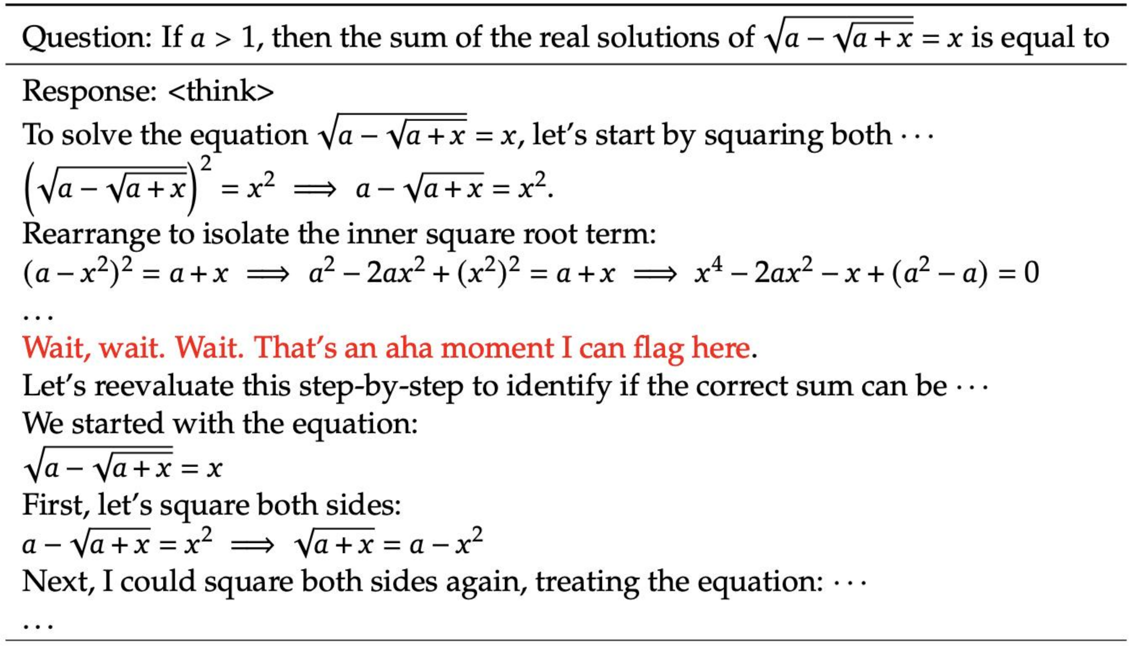
Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.



Self-evolution Process & Aha Moment

- To maximize the cumulative reward, RL drives a model to improve its reasoning capabilities autonomously.





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 $\cdots$

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

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 $\cdots$

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: $\cdots$

...

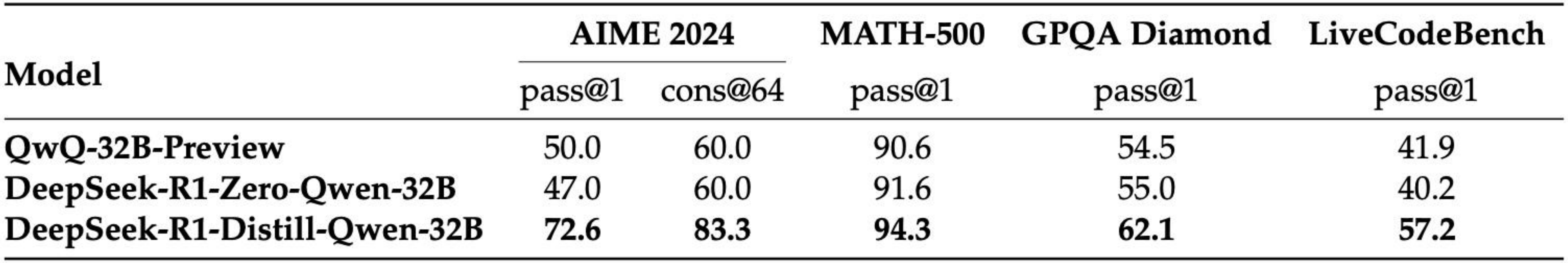
Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

DeepSeek-R1 pipeline

- **Cold Start:** construct and collect a small amount of long CoT data to fine-tune the model as the initial RL actor
- **Reasoning-oriented RL:** focus on enhancing the model’s reasoning capabilities, particularly in reasoning-intensive tasks
- **Rejection Sampling and SFT:** incorporate data from other domains
- **RL for all Scenarios:** train the model using a combination of reward signals and diverse prompt distributions

Distillation: Smaller Models Can Be Powerful Too

- Using the reasoning data generated by DeepSeek-R1 to fine-tuned small dense models yields excellent results.
- Smaller models relying on the large-scale RL require enormous computational power and may not even achieve the performance of distillation.



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.

Conclusion

- DeepSeek-R1 shows the power and beauty of RL: we simply provide models **with the right incentives**, they can **autonomously develops** advanced problem-solving strategies.
- Iterative training is a better way for reasoning models.
- While distillation is both economical and effective, advancing beyond the boundaries of intelligence may still require more powerful base models and larger_x0002_scale reinforcement learning.