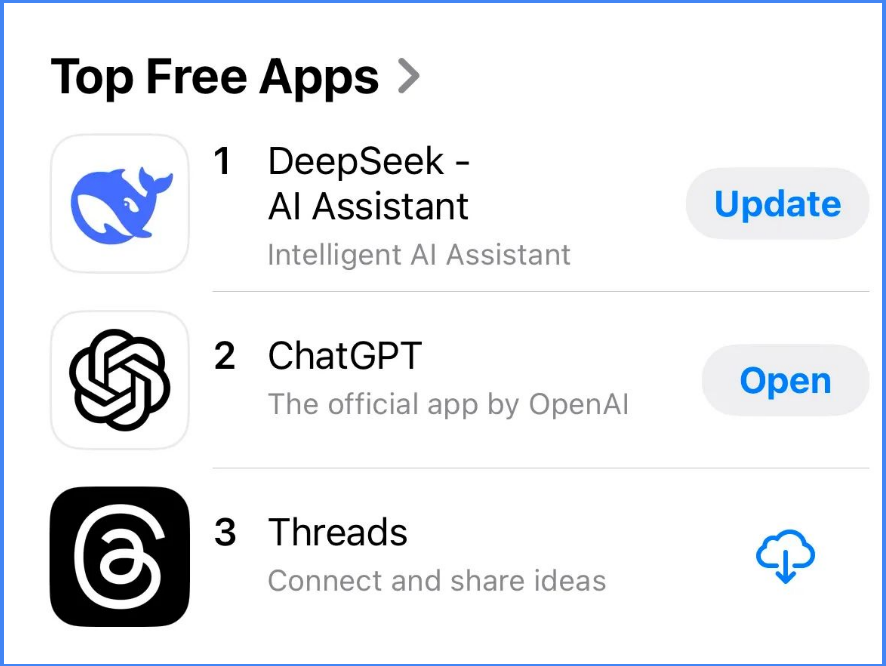
Deepseek-R1 Explained

Joyee Wang & Wanqiao Xu
Feb. 12, 2025

"Deepseek moment"



Outline

Capabilities

- Reasoning
- Benchmark Performance

Training Pipeline

- "Normal" pipeline
- Reward rules
- Deepseek-R1 Zero
- Deepseek-R1

Takeaways

- Direct RL training
- Rule-based rewards
- Efficiency

Outline

Capabilities

- Reasoning
- Benchmark Performance

Training Pipeline

- “Normal” pipeline
- Reward rules
- Deepseek-R1 Zero
- Deepseek-R1

Takeaways

- Direct RL training
- Rule-based rewards
- Efficiency

Reasoning models think before they answer

Normal LLM: **Think GPT-4o**

Reasoning LLM: **Think OpenAI’s o1**

https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1

Sample AIME 2024 problem

Problem

Let $ABCD$ be a tetrahedron such that $AB = CD = \sqrt{41}$, $AC = BD = \sqrt{80}$, and $BC = AD = \sqrt{89}$. There exists a point I inside the tetrahedron such that the distances from I to each of the faces of the tetrahedron are all equal. This distance can be written in the form $\frac{m\sqrt{n}}{p}$, where m, n, and p are positive integers, m and p are relatively prime, and n is not divisible by the square of any prime. Find $m + n + p$.

Solution 1

Notice that $41 = 4^2 + 5^2, 89 = 5^2 + 8^2$, and $80 = 8^2 + 4^2$, let $A\ (0,0,0)$, $B\ (4,5,0)$, $C\ (0,5,8)$, and $D\ (4,0,8)$. Then the plane BCD has a normal

$$\mathbf{n} := \frac{1}{4}\overrightarrow{BC} \times \overrightarrow{CD} = \frac{1}{4}\begin{pmatrix}-4\\0\\8\end{pmatrix} \times \begin{pmatrix}4\\-5\\0\end{pmatrix} = \begin{pmatrix}10\\8\\5\end{pmatrix}.$$

Hence, the distance from A to plane BCD, or the height of the tetrahedron, is

https://artofproblemsolving.com/wiki/index.php/2024_AIME_I_Problems/Problem_14

It learns...!

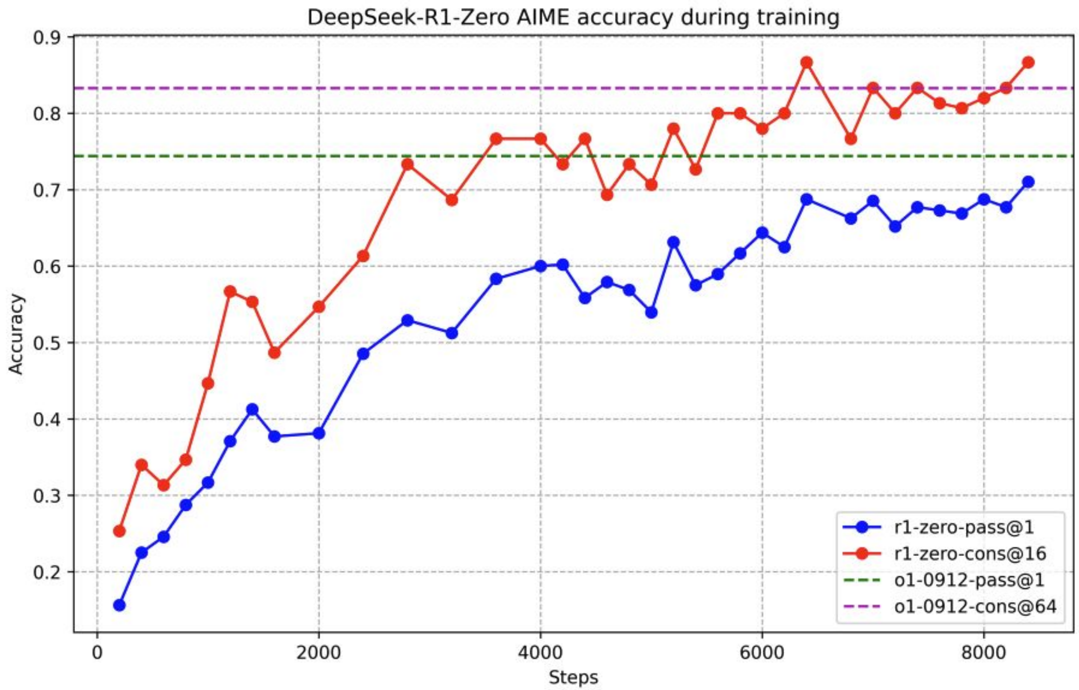


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

It learns...to think more over time!

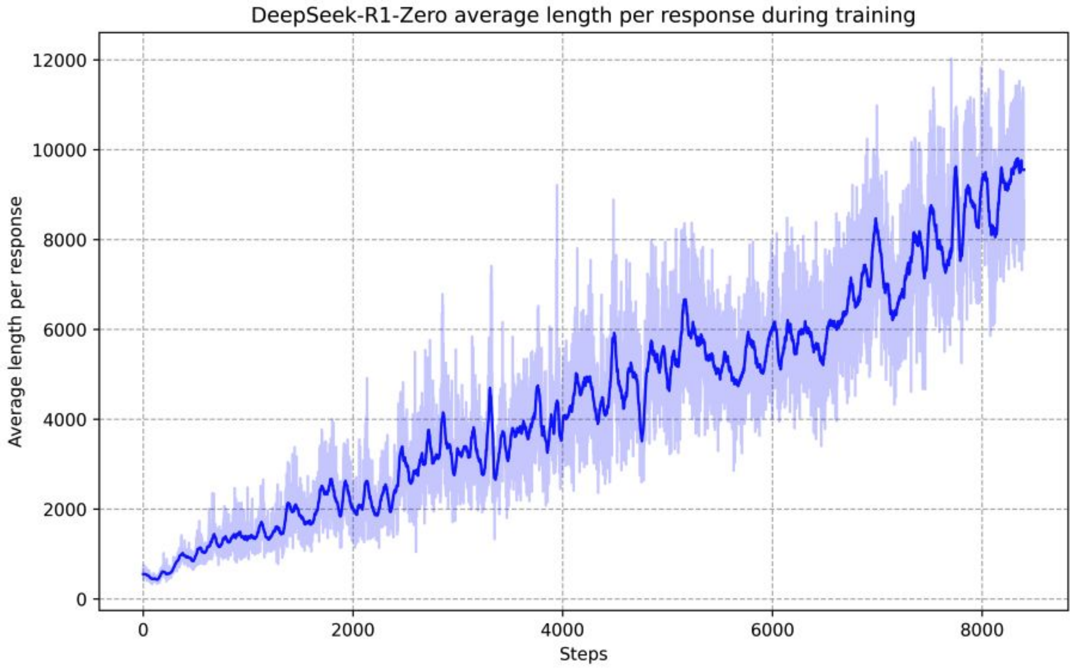
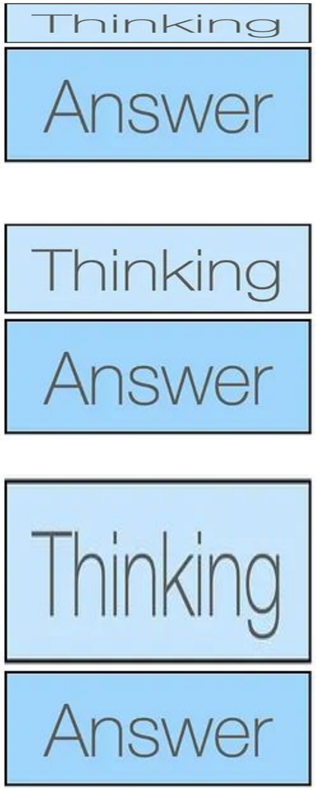


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.



https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1

Performance comparable to OpenAI's o1

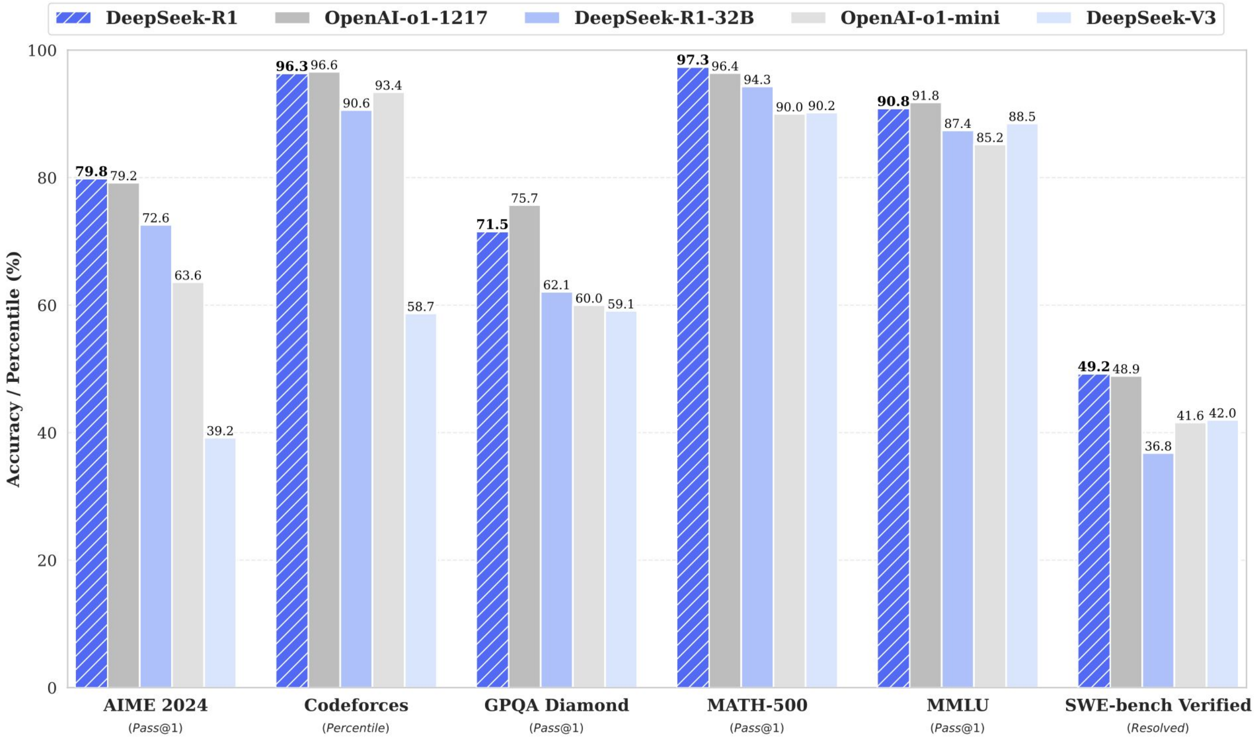


Figure 1 | Benchmark performance of DeepSeek-R1.

Outline

Capabilities

- Reasoning
- Benchmark Performance

Training Pipeline

- “Normal” pipeline
- Reward rules
- Deepseek-R1 Zero
- Deepseek-R1

Takeaways

- Direct RL training
- Rule-based rewards
- Efficiency

The recipe to create a "normal" LLM

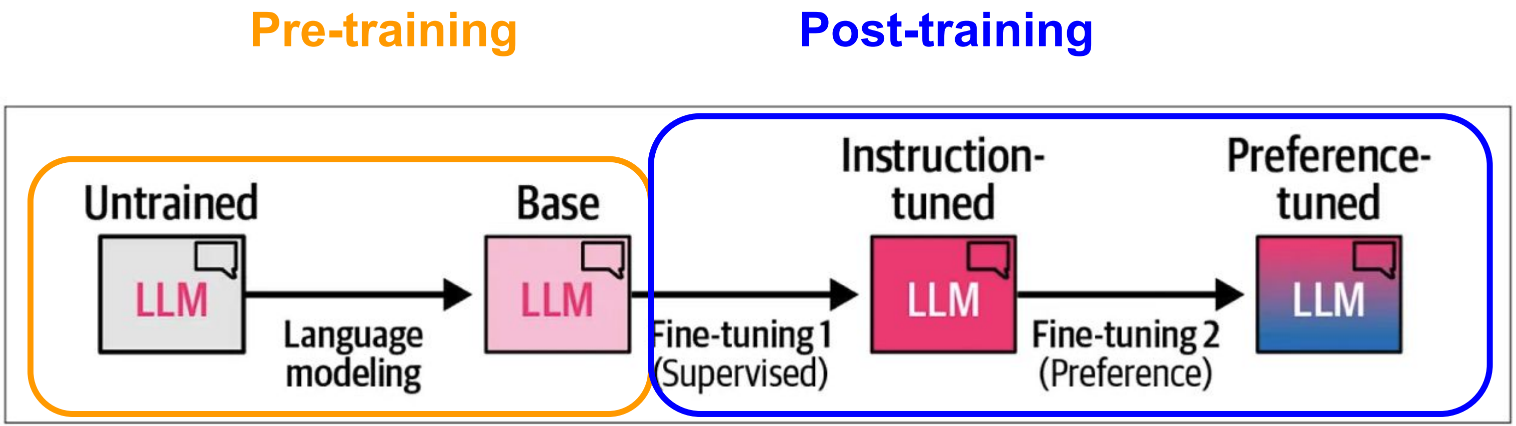


Figure 12-3. The three steps of creating a high-quality LLM.

What did Deepseek-R1 do differently?

Pre-training

Post-training

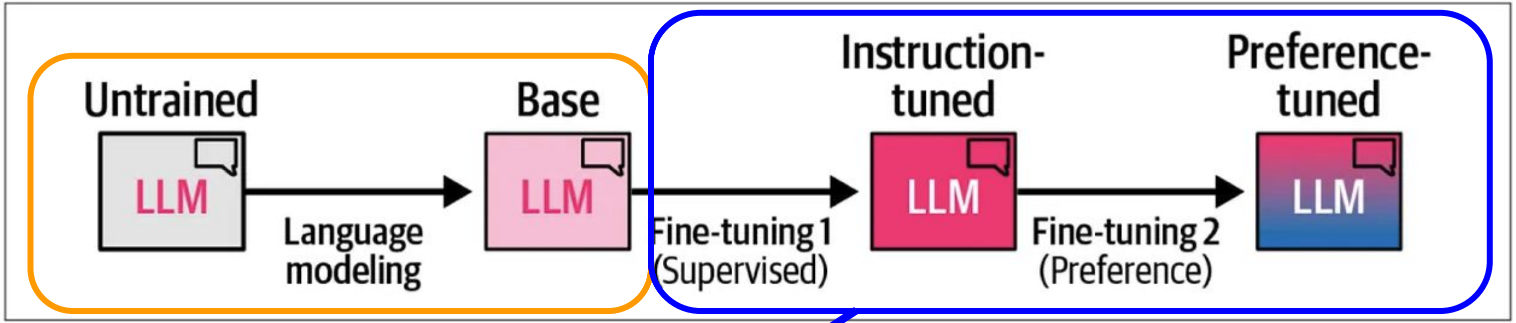
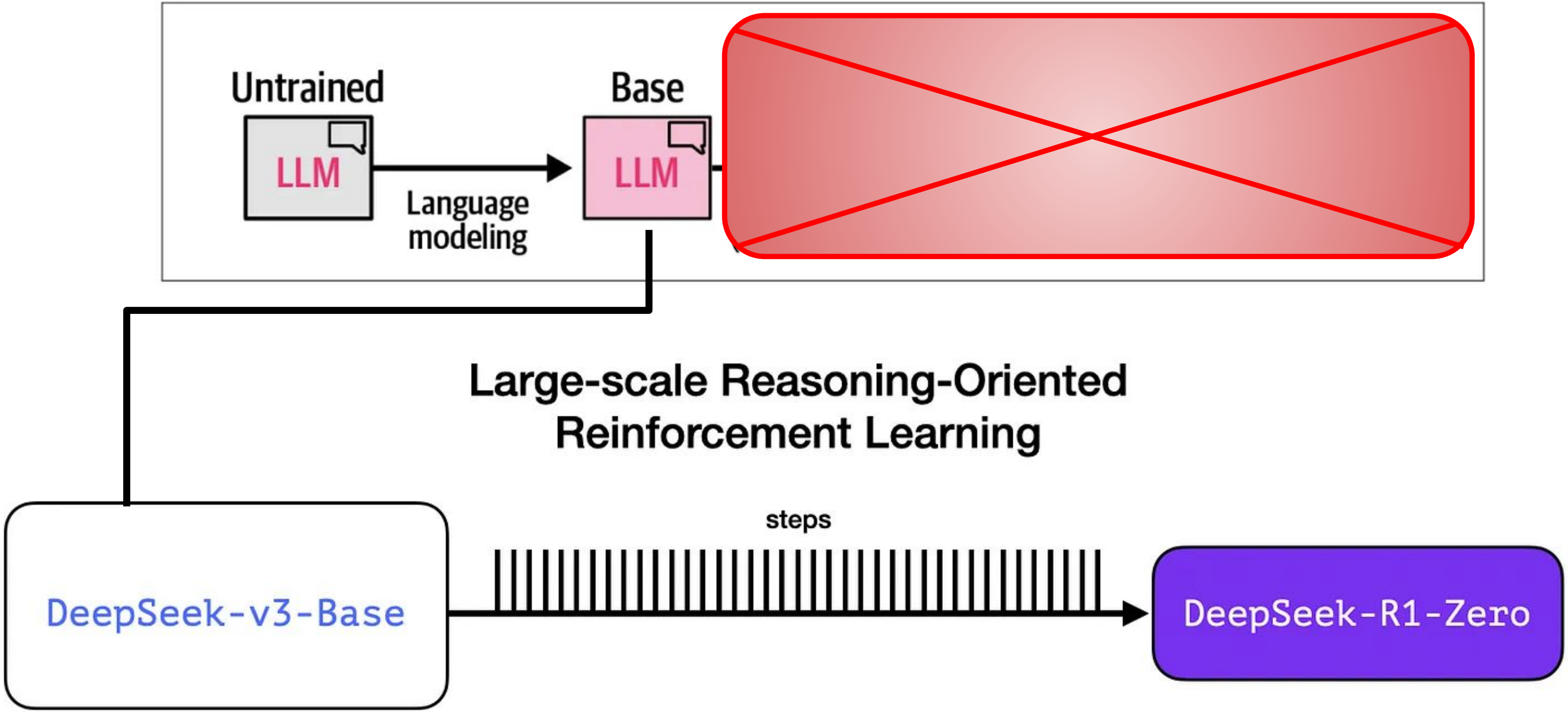


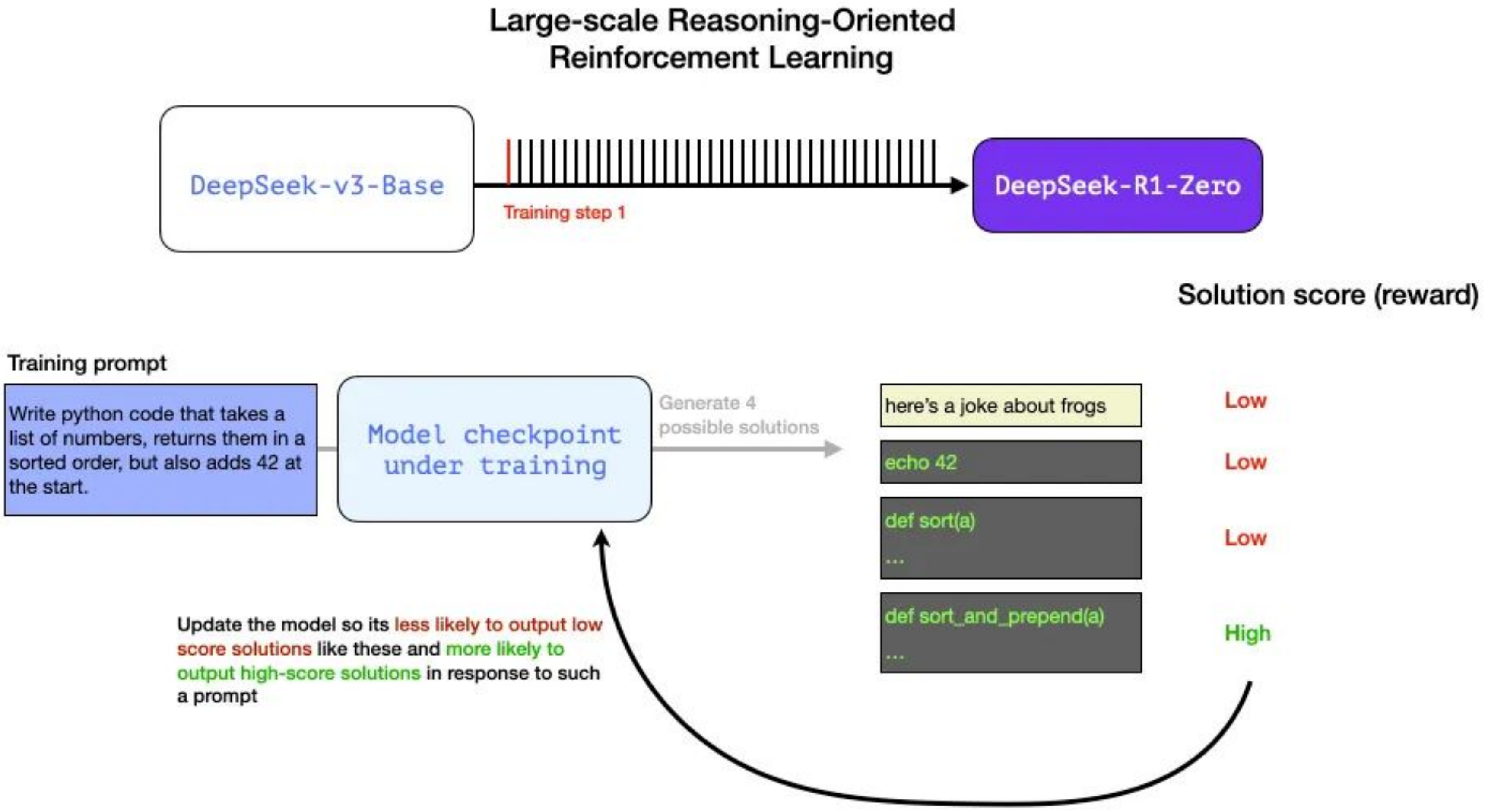
Figure 12-3. The three steps of creating a high-quality LLM.

They used a different **post-training** recipe on the base model!

Deepseek-R1 Zero



https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1



https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1

Rule-Based Rewards

1. **Accuracy** rewards: check math answers and coding test cases **Verifiable!**
2. **Format** rewards: whether the model puts its thinking process within 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.

Group Relative Policy Optimization (GRPO) vs Proximal Policy Optimization (PPO)

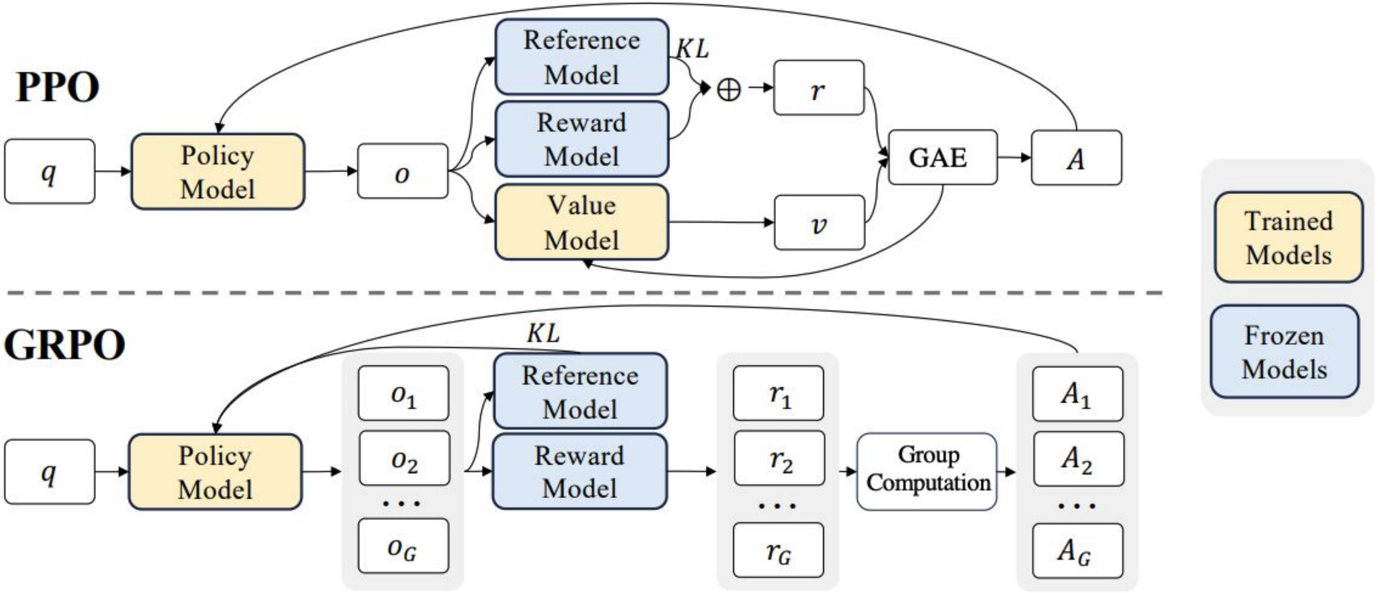


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

RLHF objective

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]$$

PPO

$$L(s,a,\theta_k,\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s,a), \;\; \operatorname{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, 1-\epsilon, 1+\epsilon\right) A^{\pi_{\theta_k}}(s,a)\right)$$

GRPO

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

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

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:

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

Problem with R1-Zero

- Mixes language in response - English + Chinese
- Chain of thoughts are not always readable/comprehensible
- Non-reasoning capabilities are limited

Two natural questions:

- Can **cold starting with high quality data** improve training efficiency and boost performance further?
- How to **train a user-friendly model** that produces readable CoT and strong general capabilities?

Deepseek-R1

Cold start SFT: with small amount (thousands) of long CoT data generated by R1-Zero

Better readability ✅ Better performance ✅

RL (same as R1-Zero)

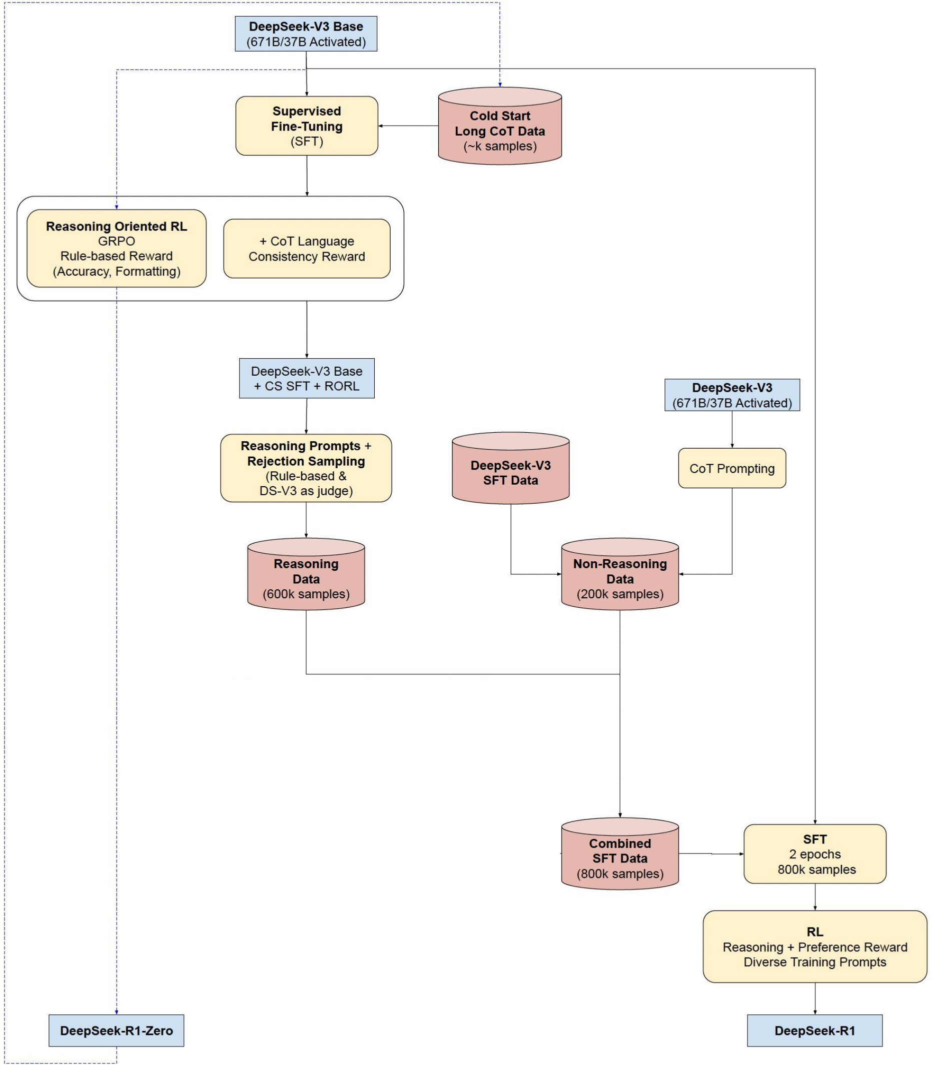
Second SFT:

600K self-generated reasoning data after filtering

200K non-reasoning data for general-purpose tasks and other domains

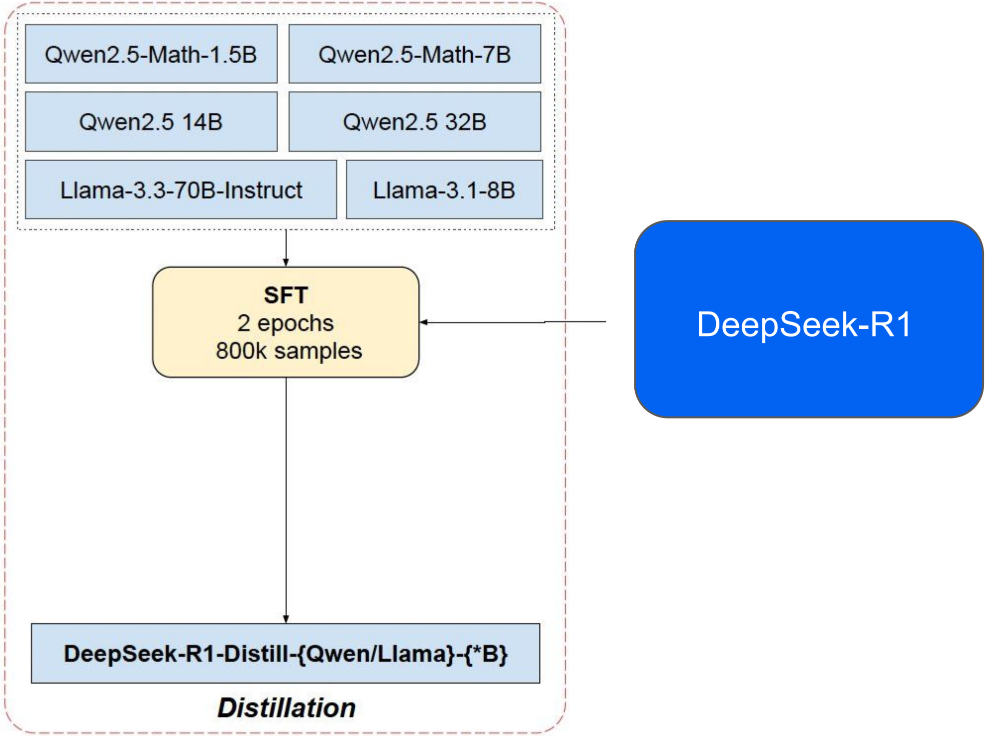
Second RL

Deepseek-R1



https://x.com/SirrahChan/status/1881488738473357753

Distilling Deepseek-R1 into Llama and Qwen



https://x.com/SirrahChan/status/1881488738473357753

Distilled models...also very very good!

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.

Outline

Capabilities

- Reasoning
- Benchmark Performance

Training Pipeline

- “Normal” pipeline
- Reward rules
- Deepseek-R1 Zero
- Deepseek-R1

Takeaways

- Direct RL training
- Rule-based rewards
- Efficiency

Direct RL on base model leads to reasoning!

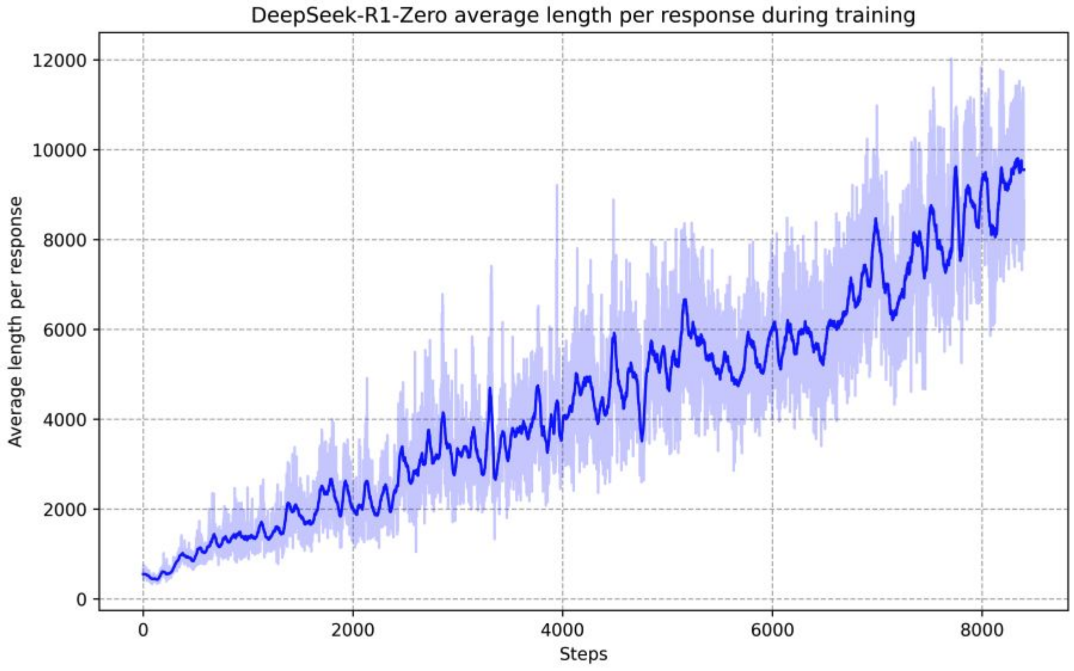
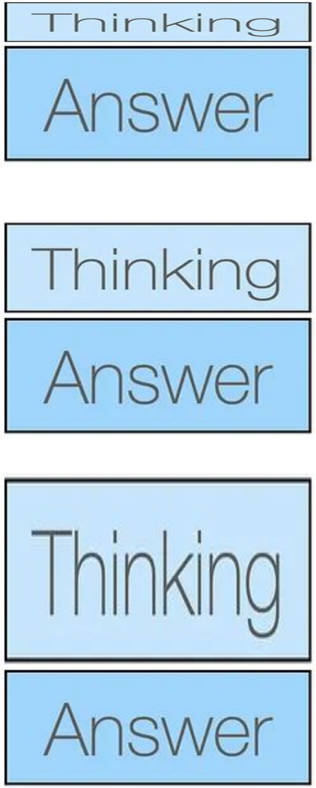


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.



https://newsletter.languagemodels.co/p/the-illustrated-deepseek-r1

Aha moment: self-correction

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

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.

Rule-Based Rewards

1. **Accuracy** rewards: check math answers and coding test cases **Verifiable!**
2. **Format** rewards: whether the model puts its thinking process within 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.

Unsuccessful attempts

Not verifiable!

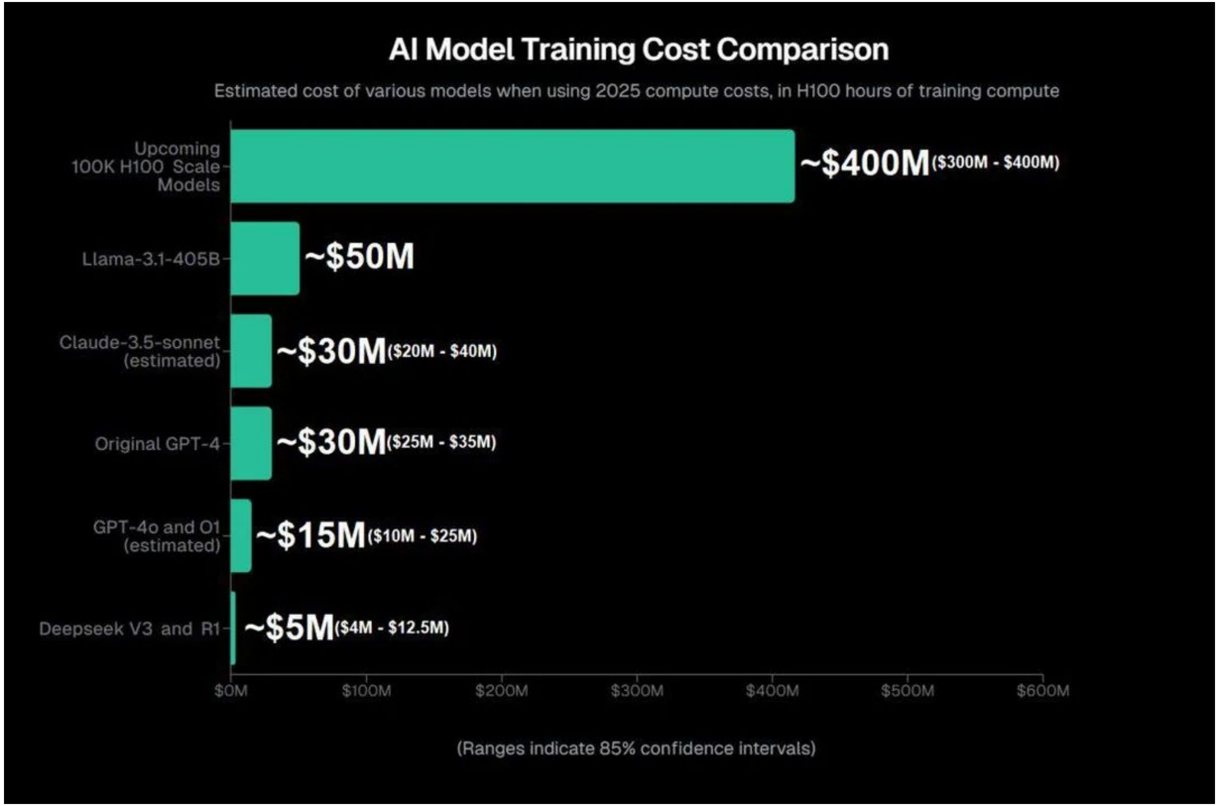
1. **Process Reward Model (PRM)**
    a. Challenging to define a step in general reasoning
    b. Challenging to determine whether a step is correct
    c. Significant reward hacking
2. **Monte Carlo tree search (MCTS)**
    a. Exponential search space
    b. Reliable value model (critic) is hard to obtain

$5.5M training cost of Deepseek-V3

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

Disproportionate cost reduction



https://www.reddit.com/r/singularity/comments/1id60qi/big_misconceptions_of_training_costs_for_deepseek/