

會進行
「深度思考」的
大型語言模型

ChatGPT o1/o3/o4、DeepSeek r1、Gemini 2 Flash Thinking、Claude 3.7 Sonnet (Extended Thinking)

1+1=



First, I'll recall that 1 is a single unit. If I have one apple and someone gives me another apple, I now have two apples. That's the concept here. Adding 1 and 1 together should give the total number of units when combined.

Wait, but maybe there's a trick here? Sometimes simple questions can be designed to test deeper understanding or to catch someone off guard. Let me think. Is there any context where $1+1$ might not equal 2? In binary, $1+1$ equals 10, but the question doesn't specify a base. However, unless stated otherwise, people usually assume base 10. Also, in some algebraic structures, like Boolean algebra, $1+1$ could be 1, but again, that's not standard arithmetic.

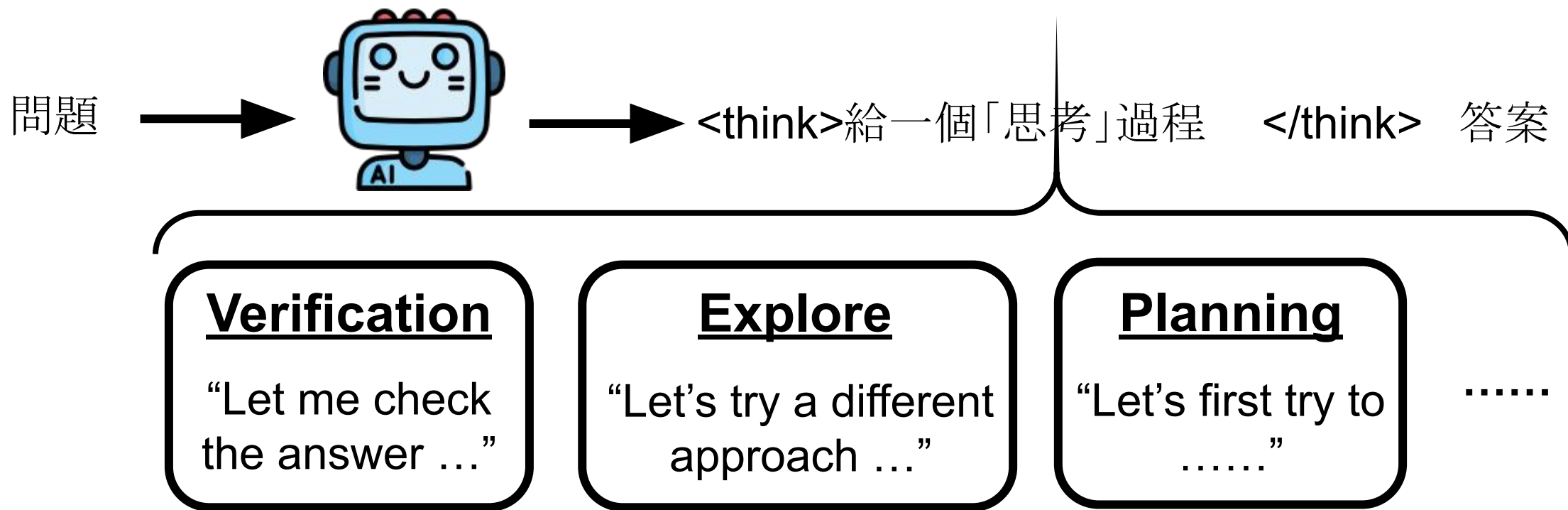
等等，這裡會不會有什麼陷阱？..... 讓我想想，有沒有什麼情況下 $1+1$ 不等於2？在二進位中， $1+1$ 等於10

The user just wrote " $1+1=$ " without any additional information. It's possible they're testing if I overcomplicate things. Let me check standard arithmetic rules. Yes, in elementary math, 1 plus 1 is definitely 2. There's no ambiguity here unless specified otherwise. I should stick with the straightforward answer unless there's a reason to think otherwise.

使用者只是寫了「 $1+1=$ 」，沒有提供額外資訊。他們可能是在測試我會不會想得太複雜。



「深度思考」語言模型的特色



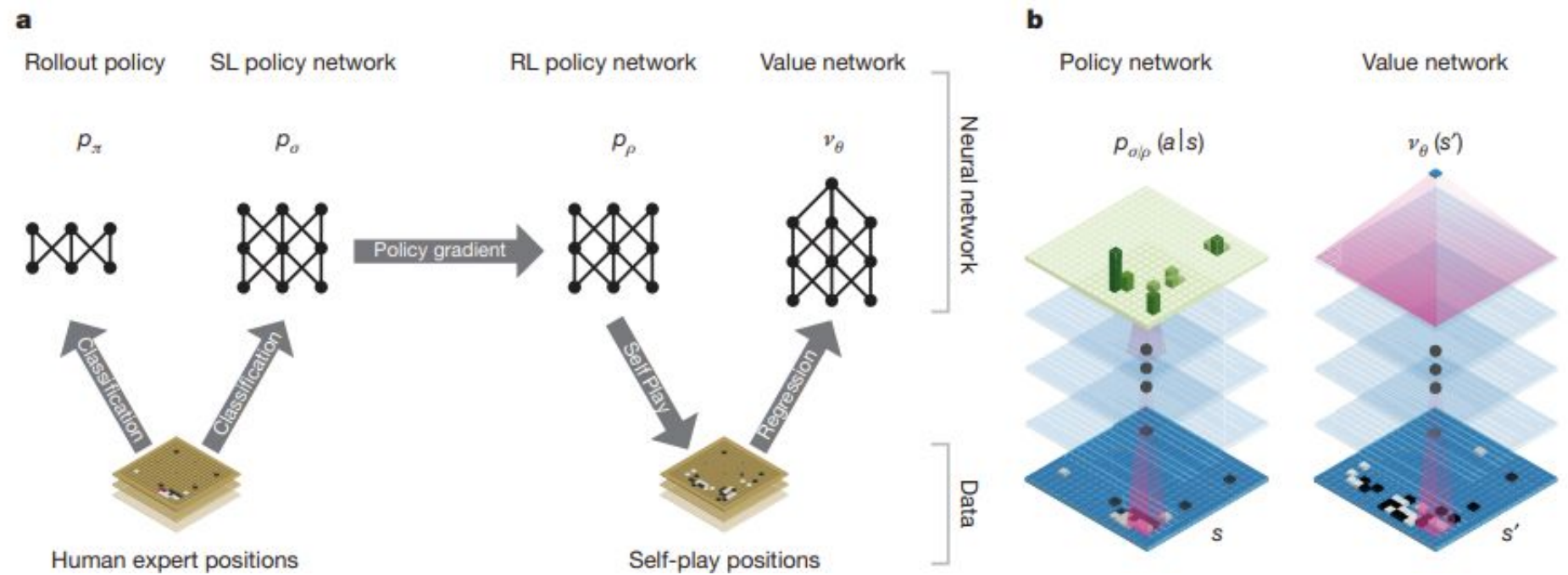
「推理」(Reasoning)

(「Inference」字面翻譯類似，但意思完全不同)

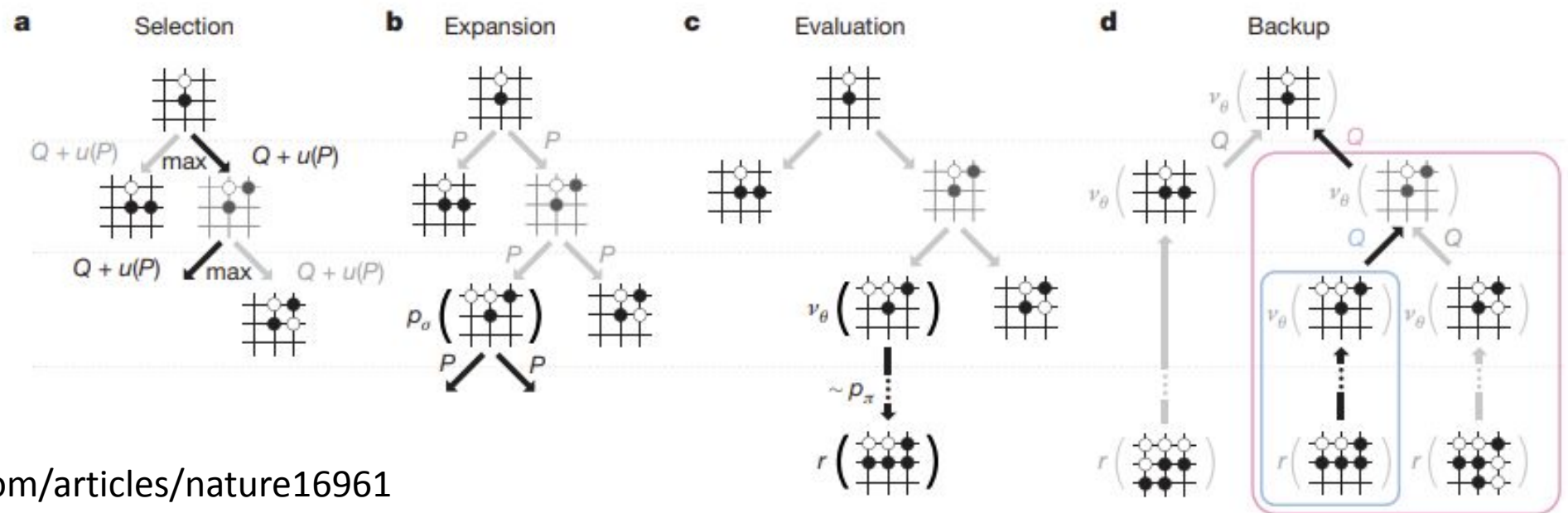
**Test-Time
Compute**

第一堂課：“深度不夠，長度來湊”

Training Time



Testing Time



Test-Time Compute

AlphaGo

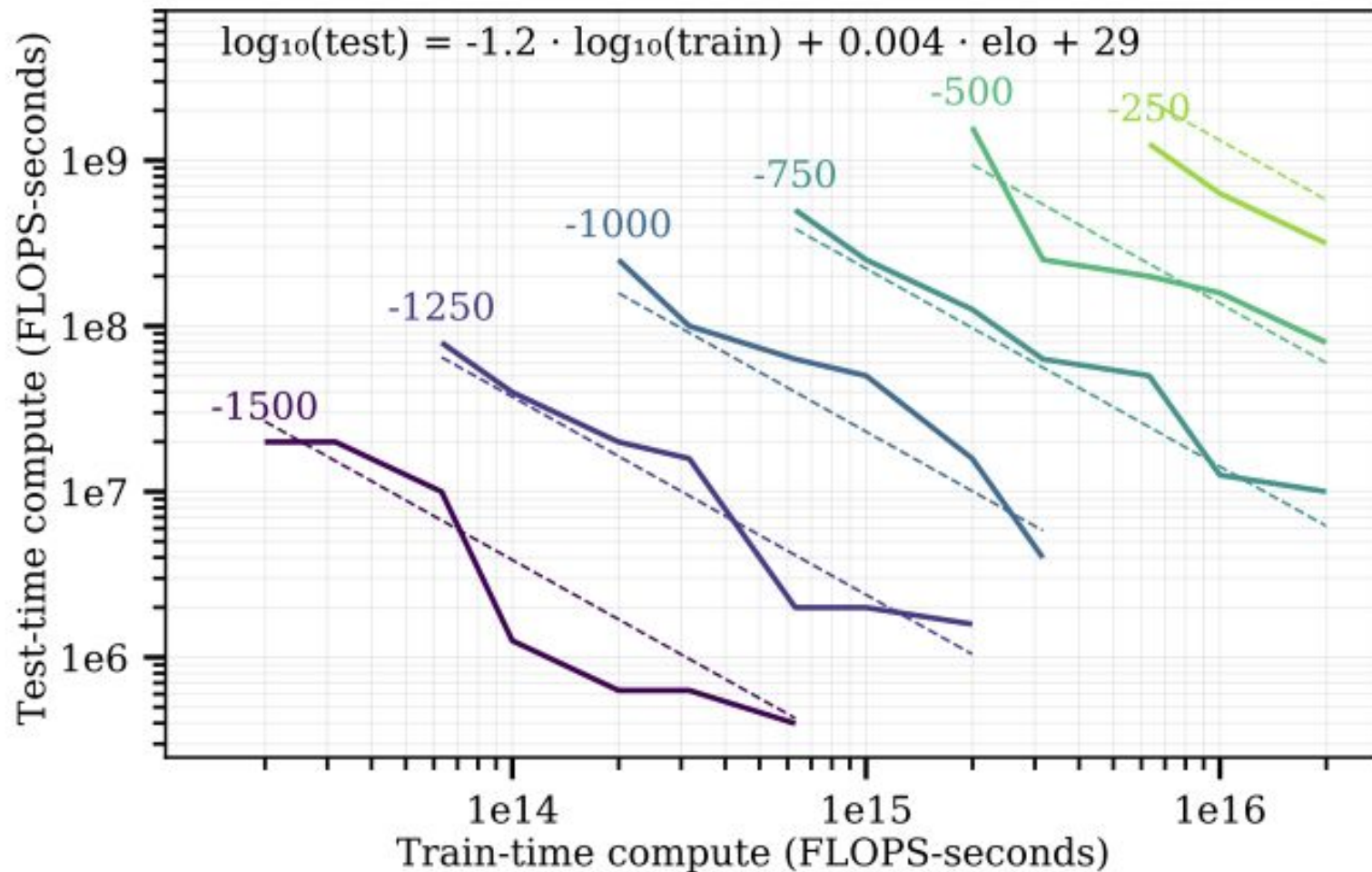
<https://www.nature.com/articles/nature16961>

「思考」越多結果越好

Test-Time Scaling

Scaling Scaling Laws with Board Games

<https://arxiv.org/abs/2104.03113>



打造「推理」語言模型的方法

不用微調參數

更強的思維鏈 (Chain-of-Thought, CoT)

給模型推論工作流程

教模型推理過程 (Imitation Learning)

以結果為導向學習推理 (Reinforcement Learning, RL)

需要微調參數

打造「推理」語言模型的方法

更強的思維鏈 (Chain-of-Thought, CoT)

給模型推論工作流程

教模型推理過程 (Imitation Learning)

以結果為導向學習推理 (Reinforcement Learning, RL)

Chain-of-Thought (CoT)

Few-shot CoT

<https://arxiv.org/abs/2201.11903>

“Short CoT”

Zero-shot CoT

<https://arxiv.org/abs/2205.11916>

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

推理模型在做的事情叫做 “Long CoT”

<https://arxiv.org/abs/2503.09567>

gpt-4o

請仔細思考並詳細回答以下問題。在回答前，請先深入分析題目的要求，訂出一個完整且清晰的解題計畫，明確列出你將如何分步完成這個問題。在執行每一個主要步驟前，請再次訂出該步驟的子計畫，仔細列出需要處理的細節，然後再按部就班地執行。每執行完一個步驟或子步驟後，請進行多次驗算，確保該步驟的答案絕對正確，並考量所有可能的解法。若在驗算過程中發現問題，請立即回到該步驟重新訂定或調整計畫。在進行以上過程時，務必將你詳細而完整的思考過程以及所有計畫、子計畫、驗算步驟，全部置於 "<think>" 和 "</think>" 這兩個符號之間。

Supervised CoT

<https://arxiv.org/abs/2410.14198>

123 x 456 =

<think>

一、題目分析：

本題要求計算一個三位數乘上一個三位數：

$$123 \times 456 = ?$$

這是一道基本的筆算乘法題，但為了符合題目要求，我們將採取逐步規劃、精密計算、多次驗算的策略來完成，並在每個階段都做出完整說明。


二、解題總計畫：

1. 分解乘法運算為部分積：將456拆解為 $400 + 50 + 6$ ，並與123逐一相乘。
2. 執行每項部分積的運算：逐項計算 123×400 、 123×50 、 123×6 。
3. 合併部分積，求總和。
4. 進行多次驗算：
 - 驗算一：使用標準直式乘法方式重新計算。
 - 驗算二：使用估算法與上下界檢查合理性。
 - 驗算三：使用Python進程式驗證（若可用）。

驗算一：使用直式乘法

SCSS

 複製

 編輯

123

驗算二：估算法檢查

X

456

估算上界：

738

6150

+

49200

估算下界：

56088

- $123 \approx 125 + 456$;

- $125 \times 450 = (100 + 25) \times 450$

$$11250 = 56250$$

實際答案 56088 在兩者

驗算三：使用乘法交換律進行其他方向拆解

- $456 = 123 \times x \Rightarrow x = \frac{456}{123} \approx 3.707$

→  結果再

不是所有模型都有能力根據複雜指令做 Long CoT

打造「推理」語言模型的方法

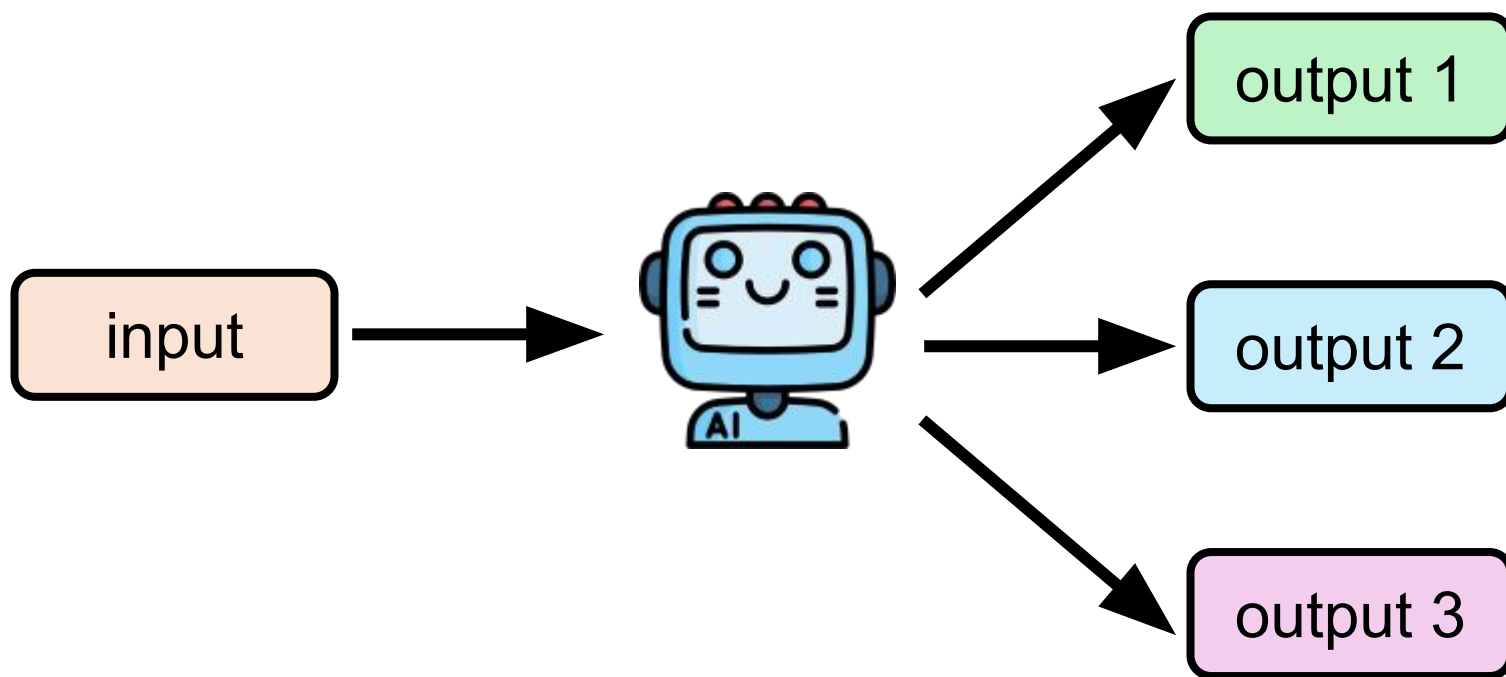
更強的思維鏈 (Chain-of-Thought, CoT)

給模型推論工作流程

教模型推理過程 (Imitation Learning)

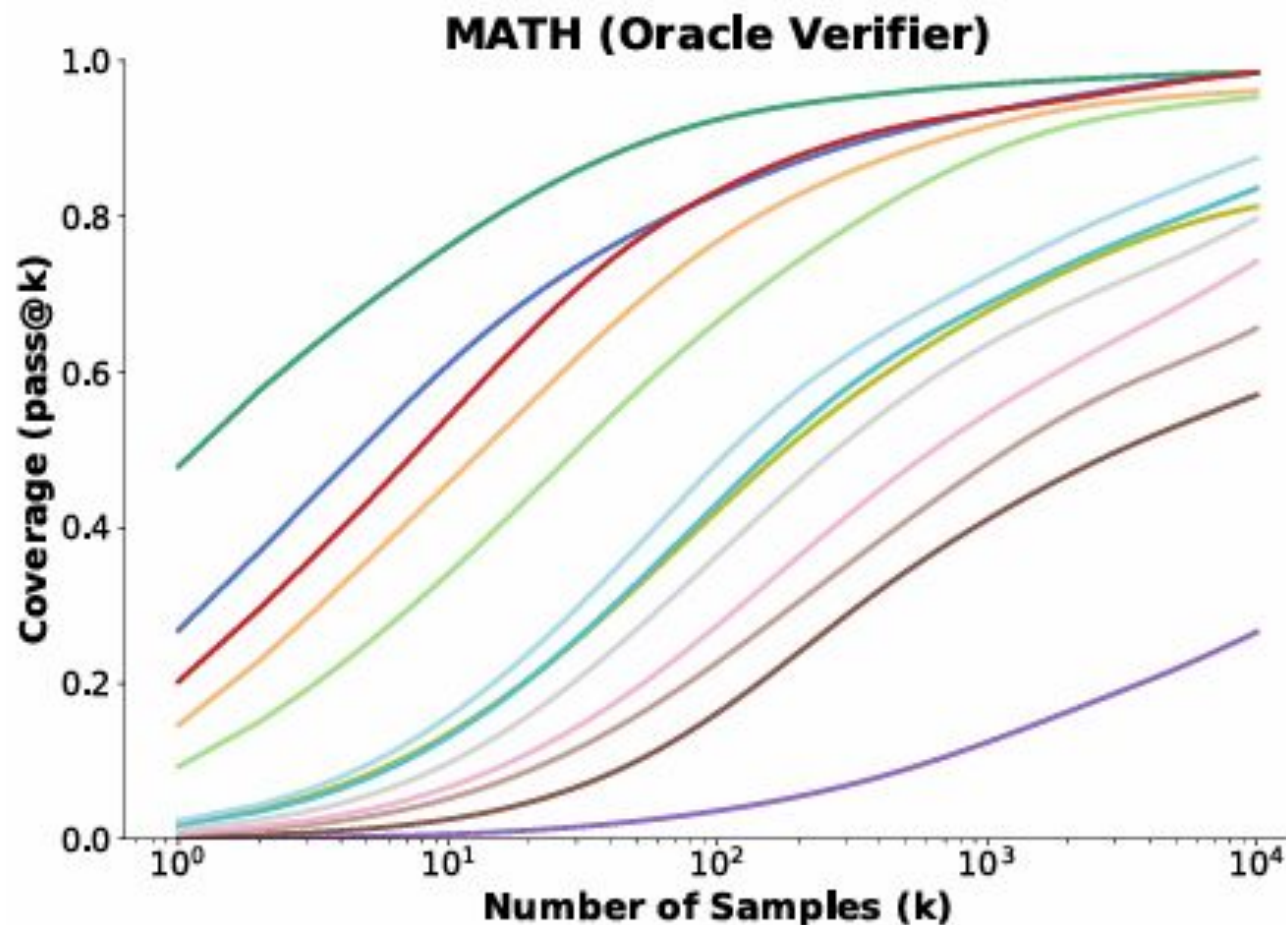
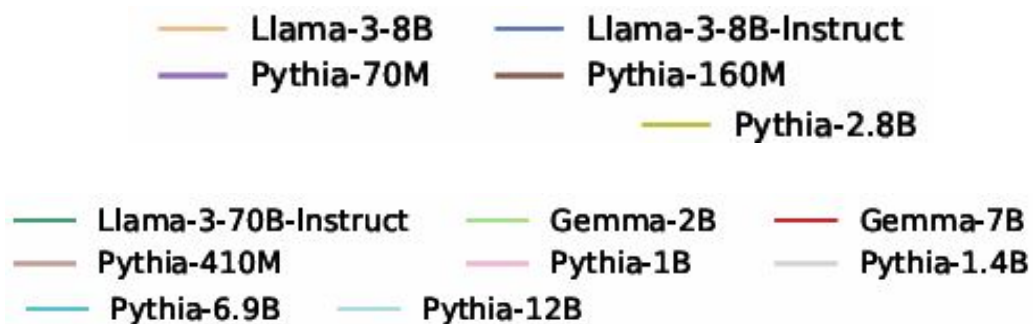
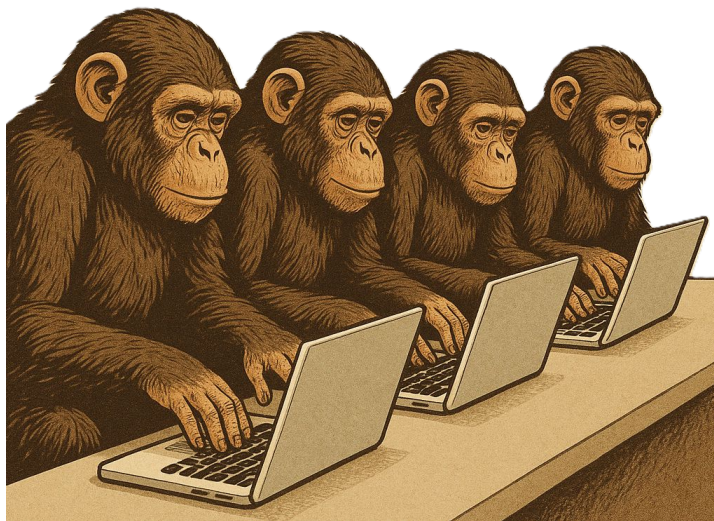
以結果為導向學習推理 (Reinforcement Learning, RL)

如何 Explore ? 同一個問題多試幾次

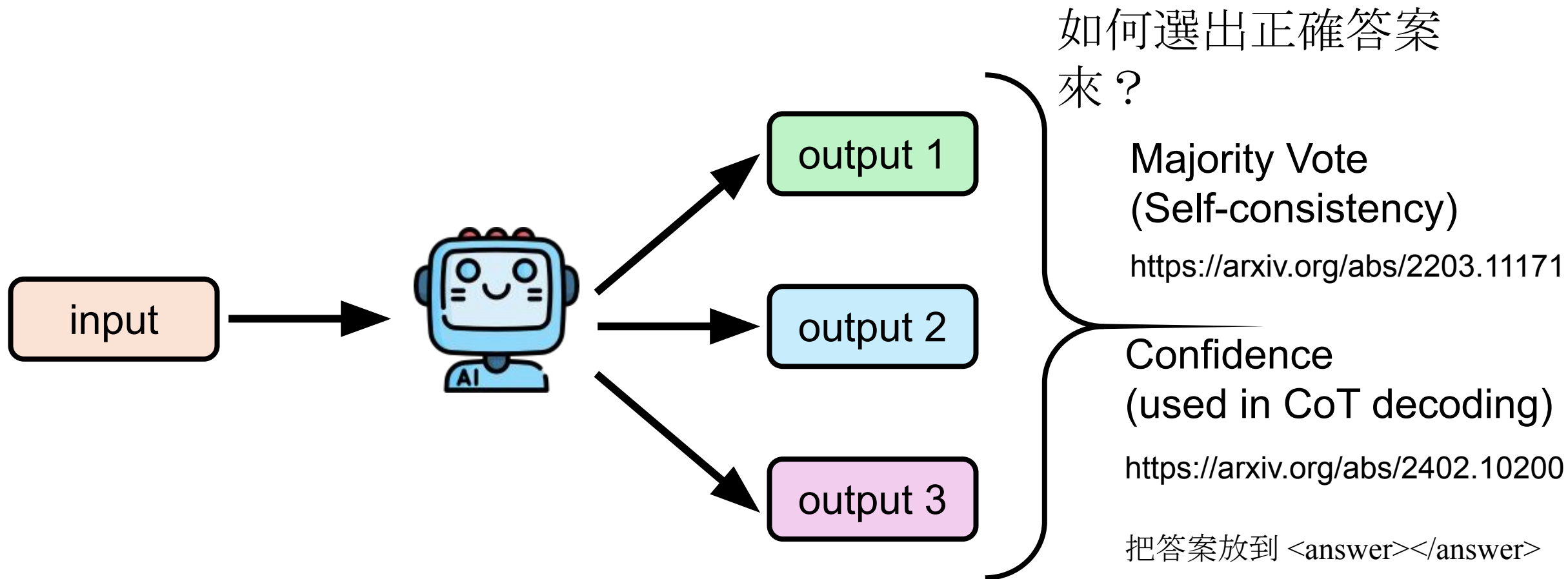


“Large Language Monkeys”

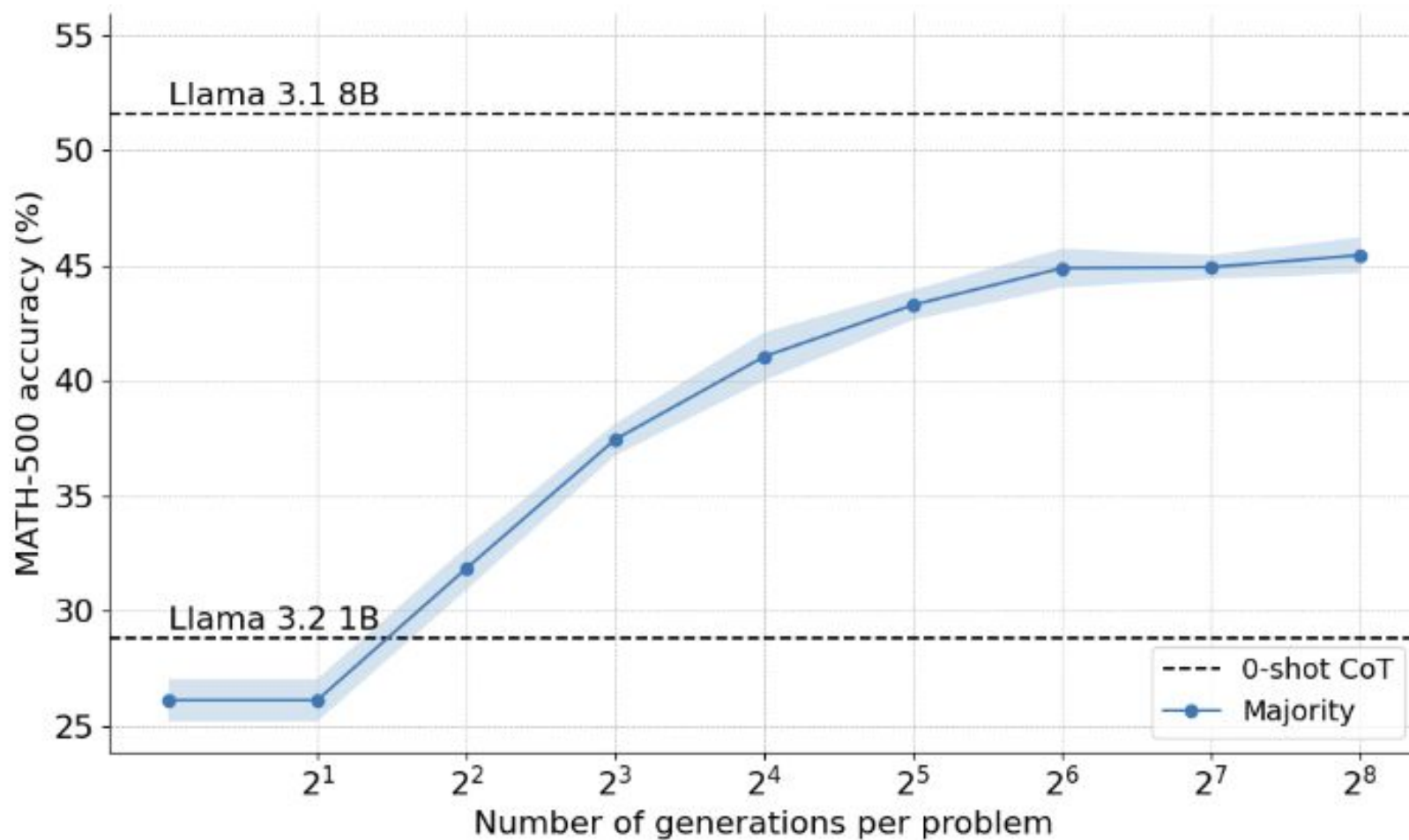
<https://arxiv.org/abs/2407.21787>



如何 Explore ? 同一個問題多試幾次

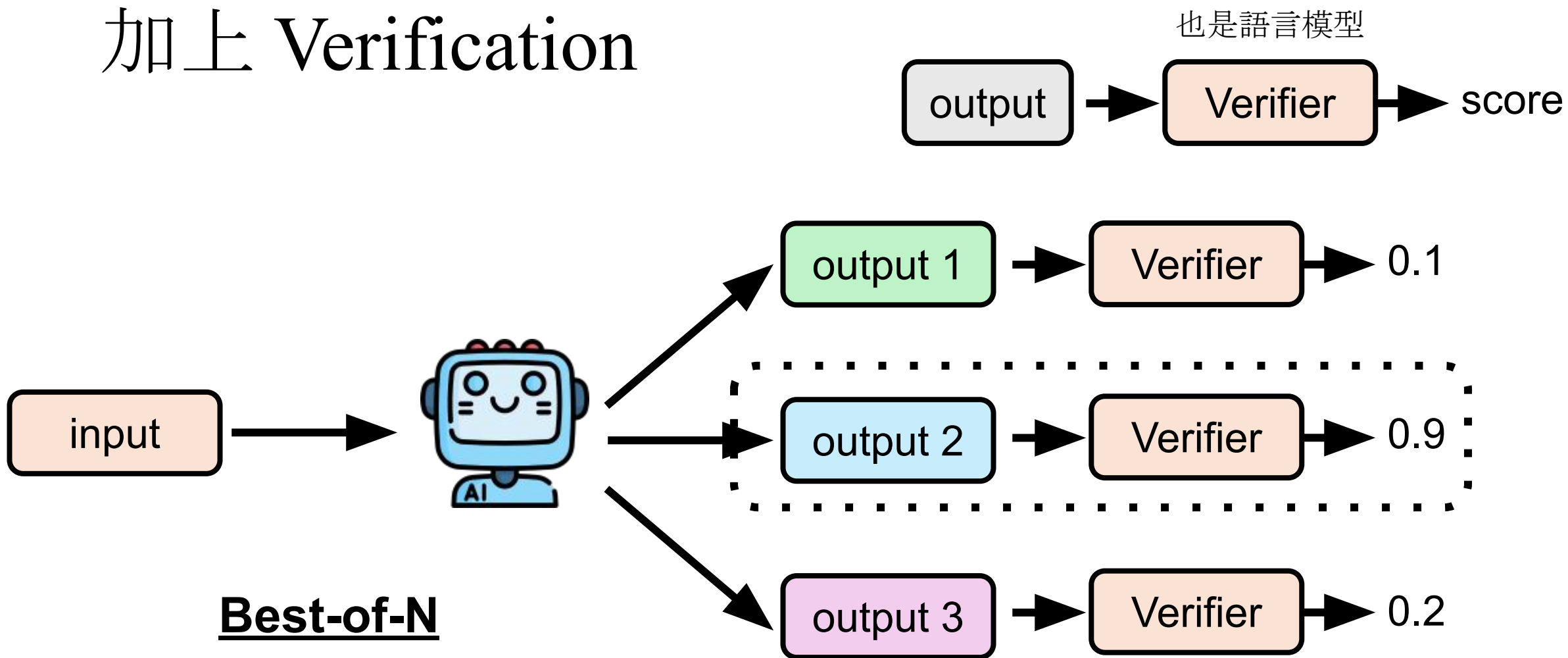


如何 Explore ? 同一個問題多試幾次



<https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute>

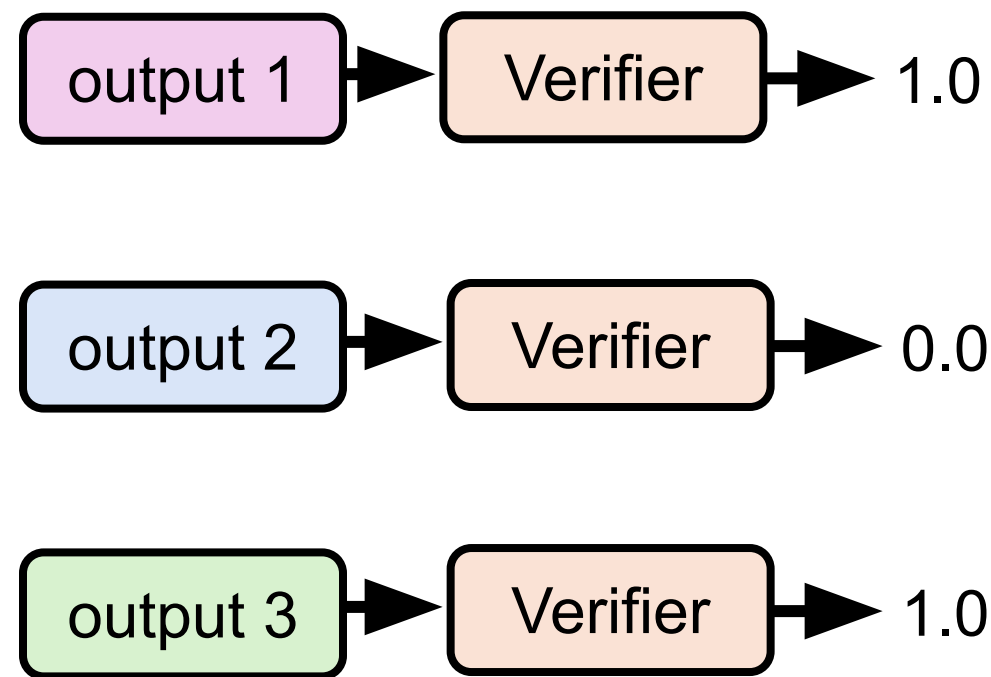
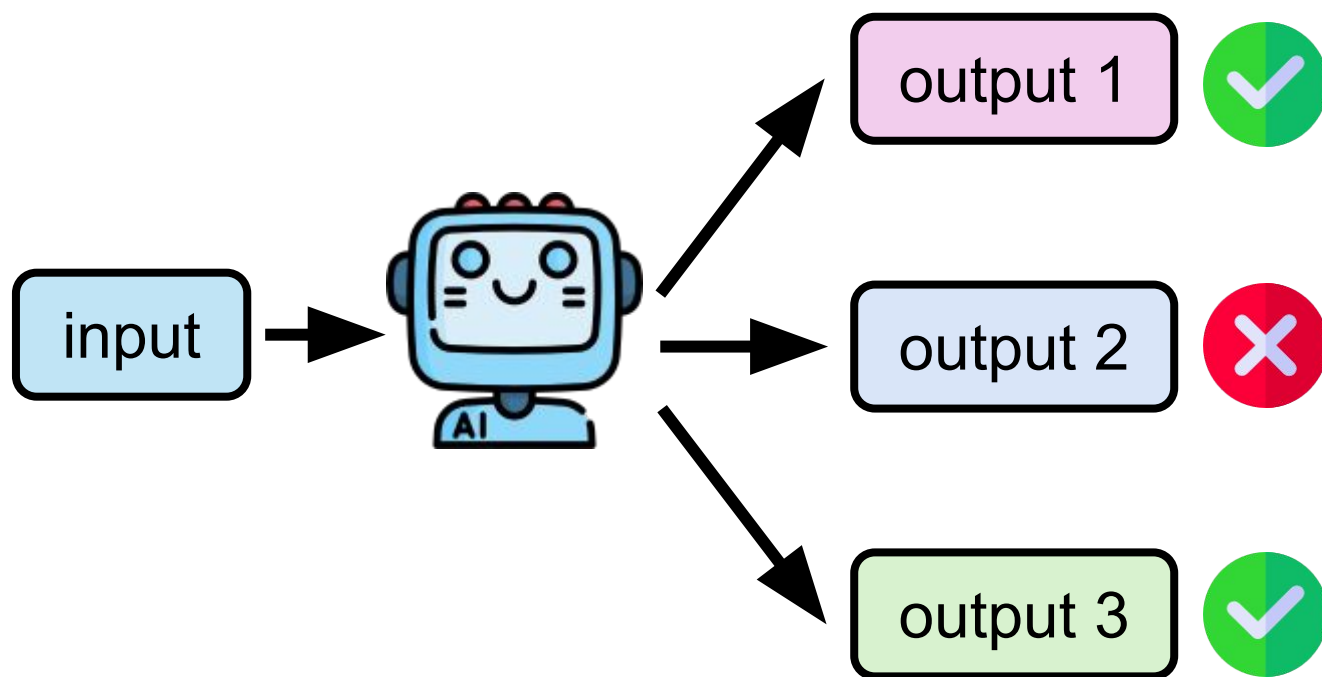
加上 Verification



<https://arxiv.org/abs/2110.14168>

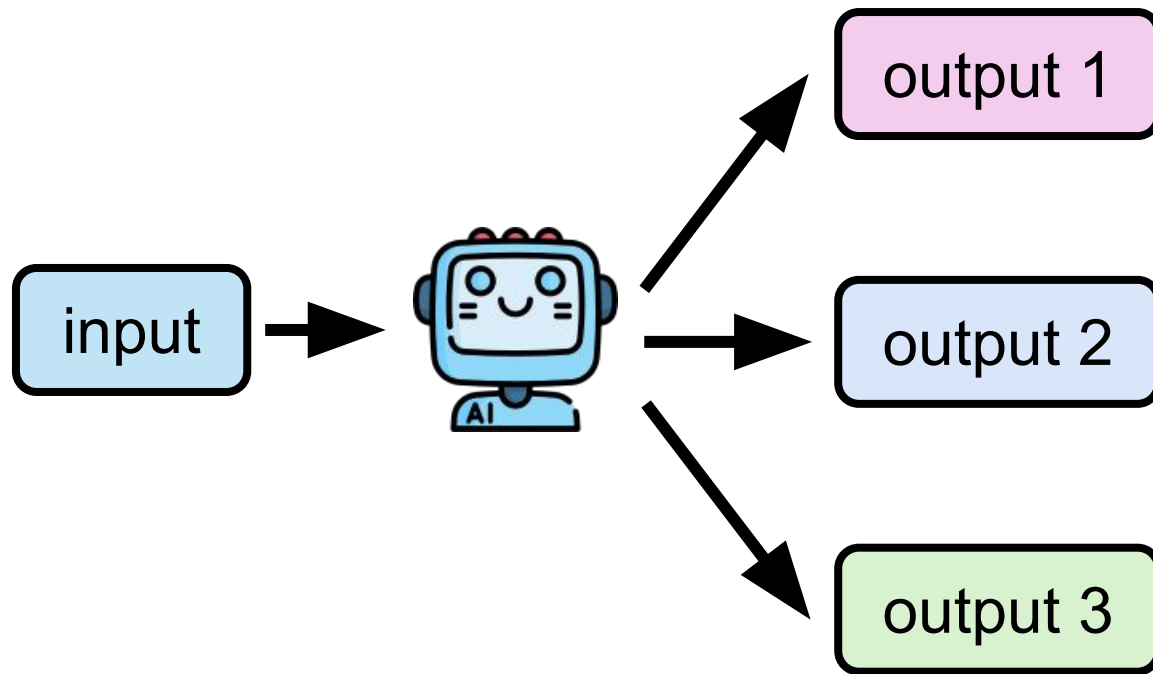
加上 Verification

Training Data: input ground truth

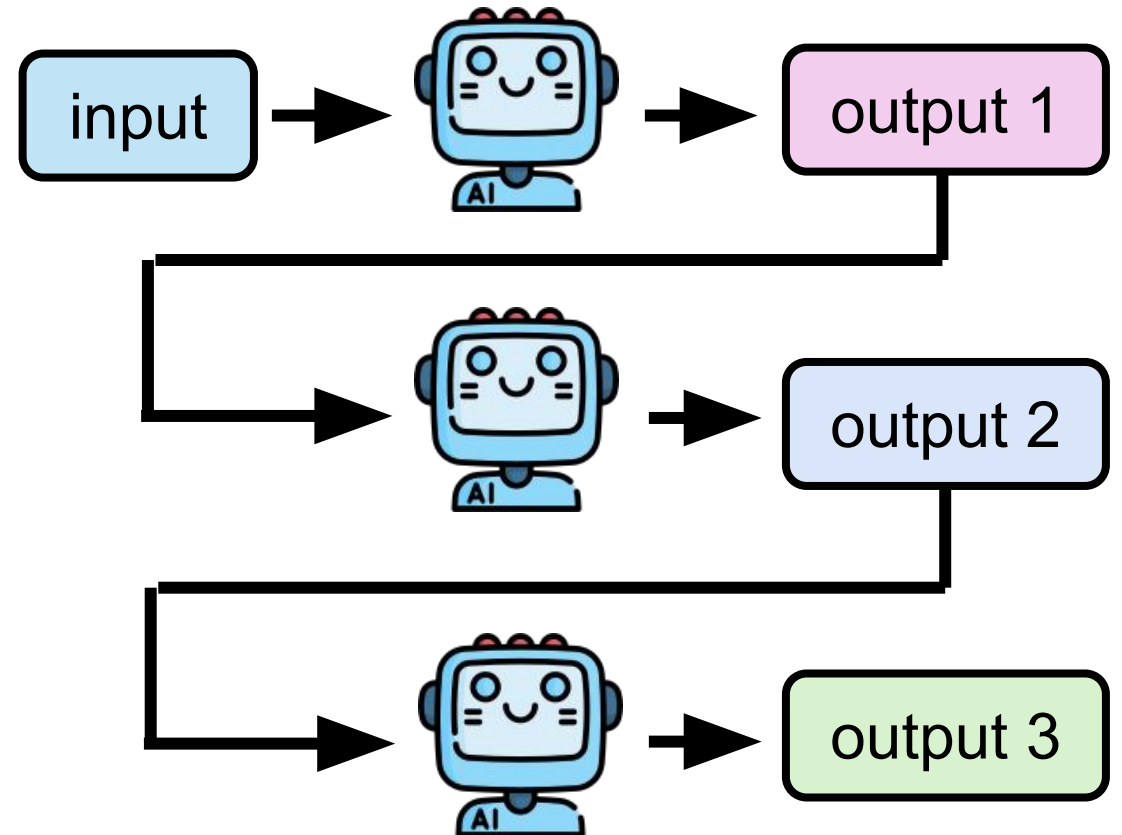


Parallel vs. Sequential

Parallel



Sequential

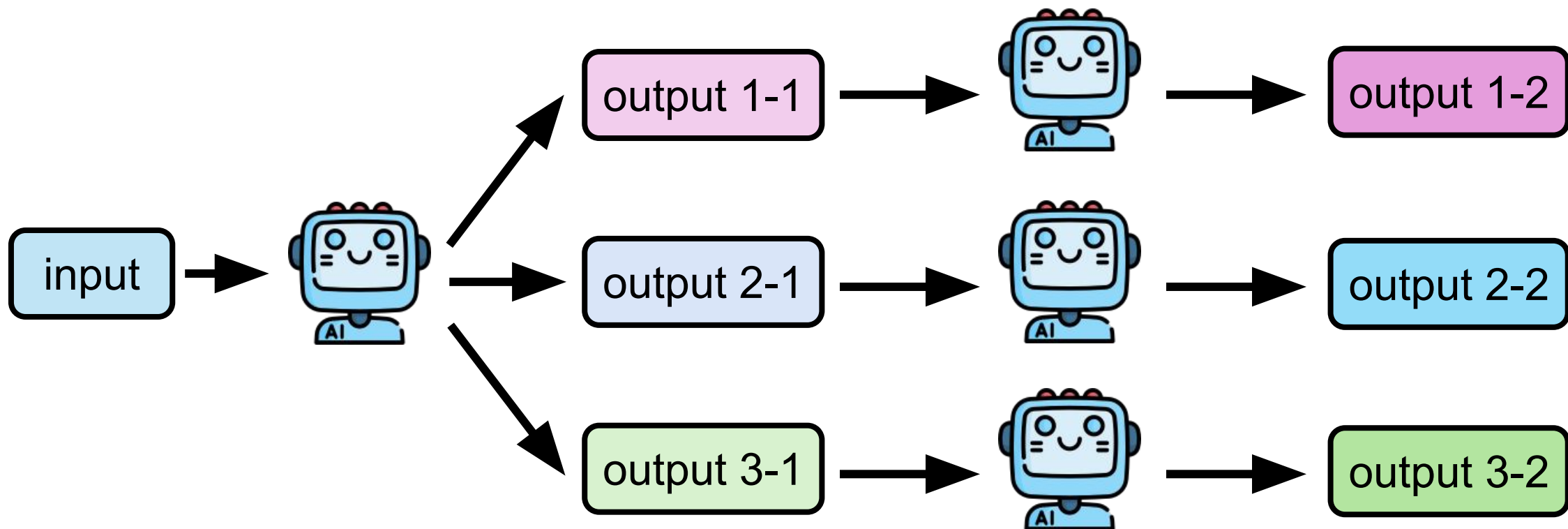


Parallel vs. Sequential

Parallel + Sequential

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

<https://arxiv.org/abs/2408.03314>



會對中間的步驟進行驗證

$$123 \times 456 = ?$$



🕒 已深度思考 (用时 59 秒) ^

Okay, so I need to multiply 123 by 456. Hmm, let's see. I remember that multiplying larger numbers can be broken down into smaller parts, right? Maybe I can use the distributive property here. Let me think.

planning

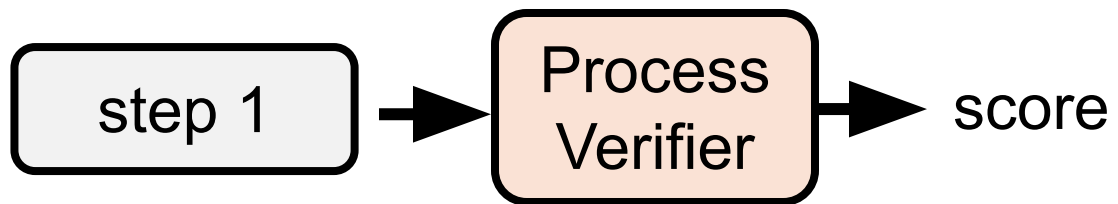
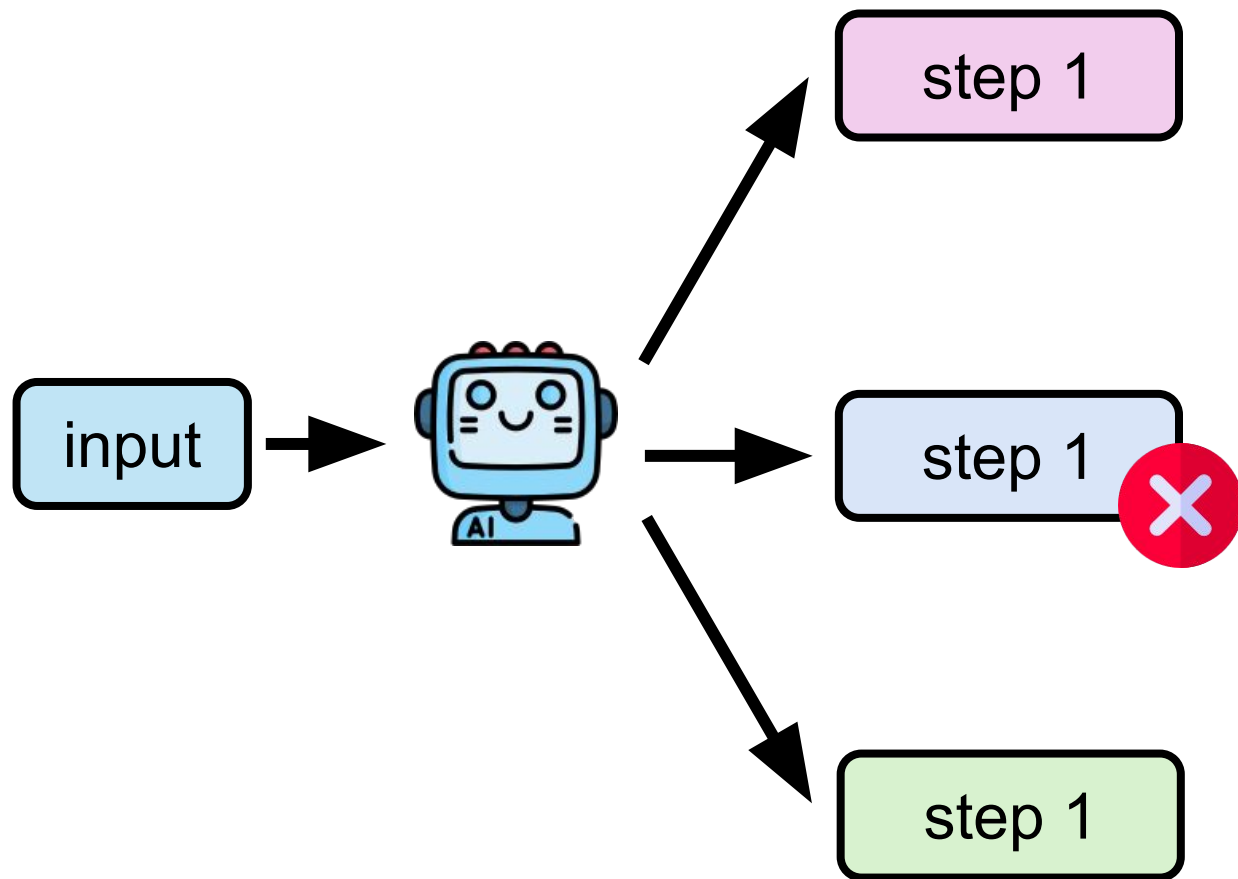
First, I should probably write it out step by step. 123 times 456. Let me break down 456 into 400, 50, and 6.

That way, I can multiply 123 by each of those parts separately and then add them up. So, 123×400 , 123×50 , and 123×6 . Then add all those results together. That should work.

Verification (for a step)

Starting with 123×400 . Well, 123 times 4 is 492, right? Because 100×4 is 400, 20×4 is 80, and 3×4 is 12.

Adding those together: $400 + 80 = 480$, plus 12 is 492. So, 123×400 would be 492 with two zeros added at the end, which makes 49,200. Got that part.



Let's Verify Step by Step
<https://arxiv.org/abs/2305.20050>

請逐步解決輸入的數學問題，每一步驟的開頭輸出
<step>，結尾都輸出 </step>。答案放在 <answer> 和
</answer> 之間。

範例：

<step> 步驟一 </step>\n

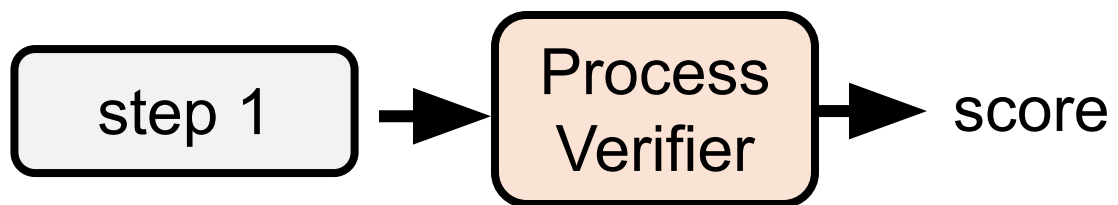
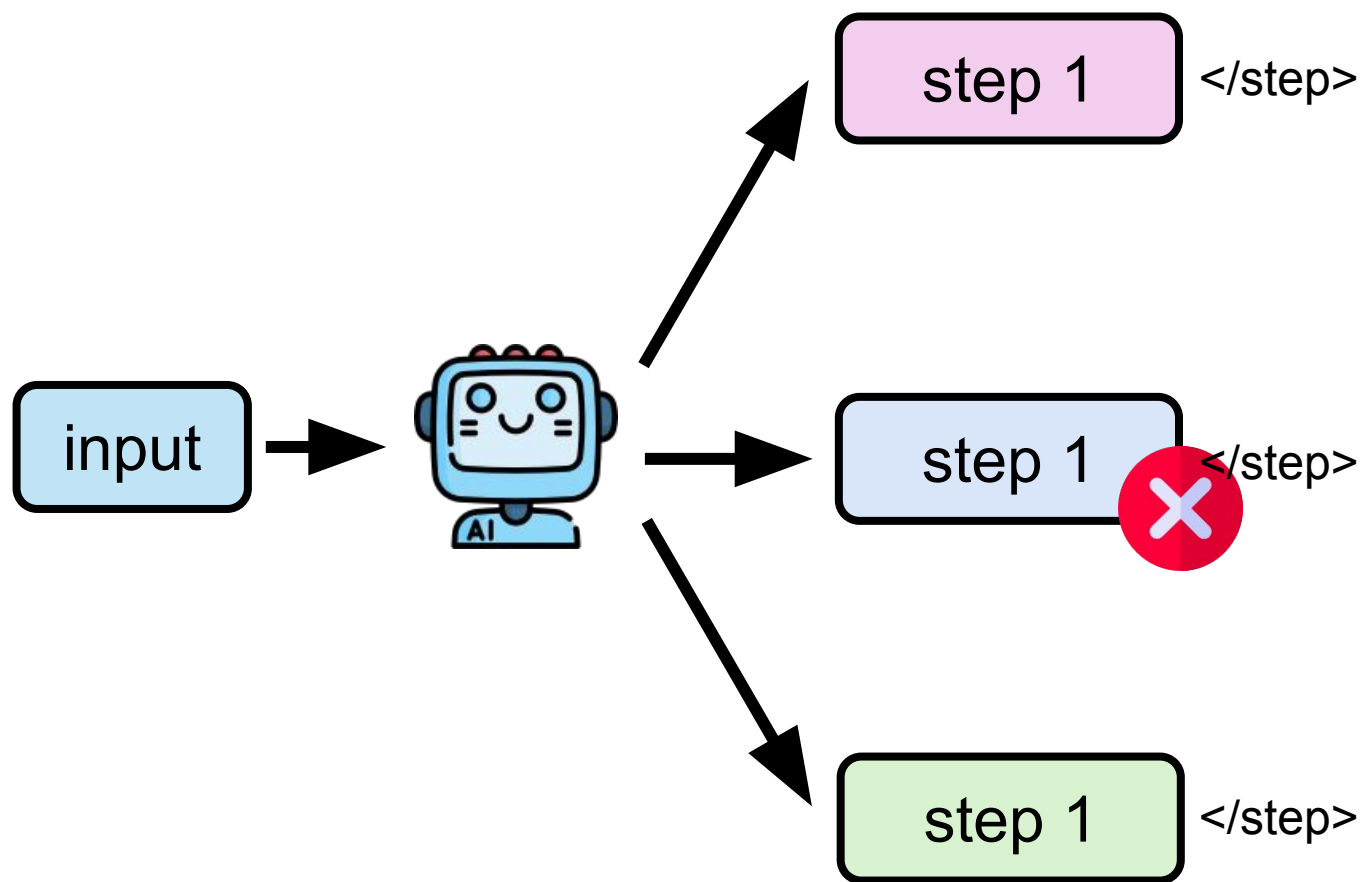
<step> 步驟二 </step>\n

<answer> 答案 </answer>

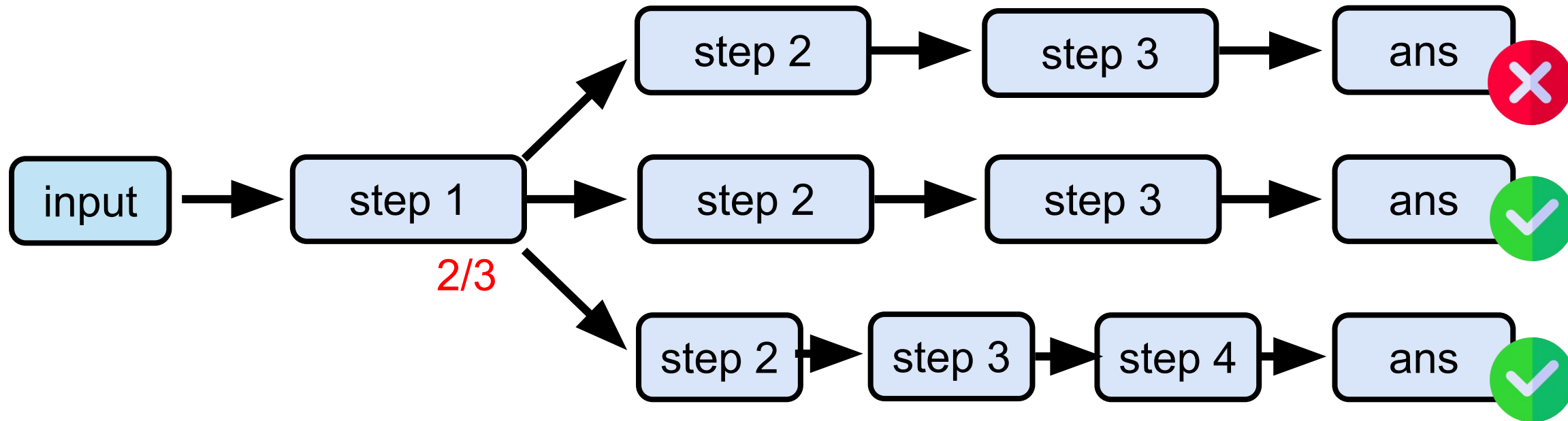
問題：

123 x 456 =

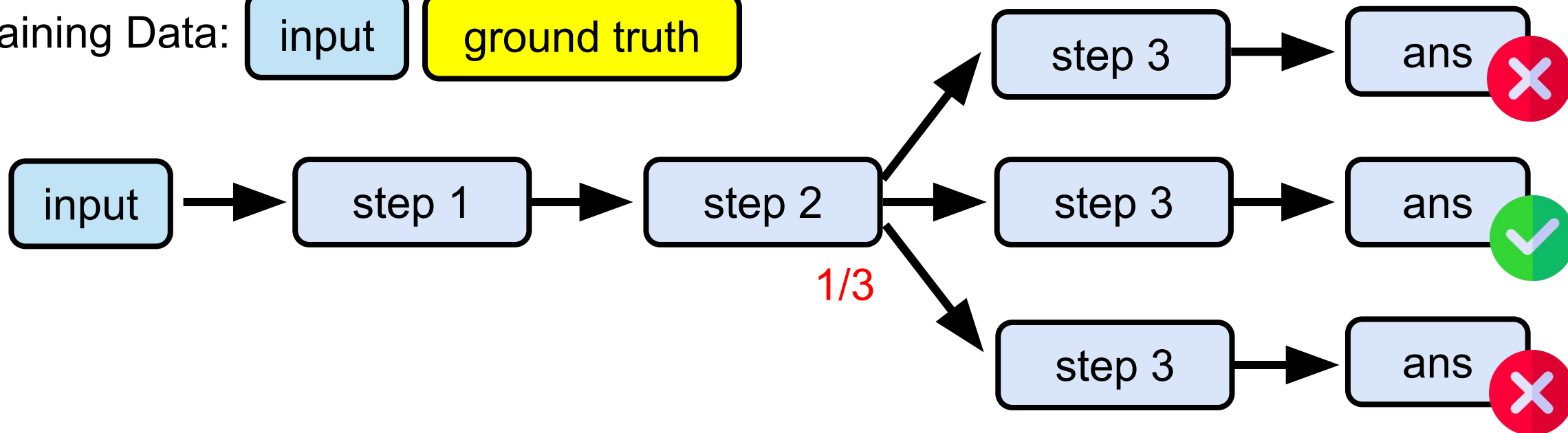
<step> 首先，我們將456拆解成400、50和6，再分別乘以123：</step>

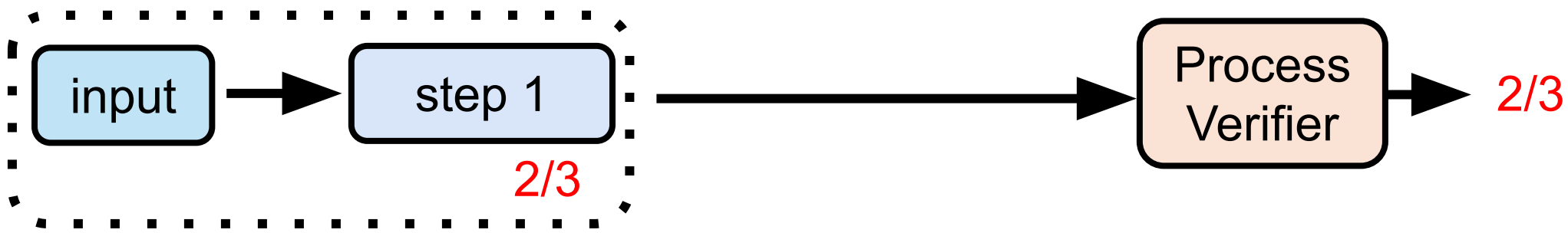


Let's Verify Step by Step
<https://arxiv.org/abs/2305.20050>

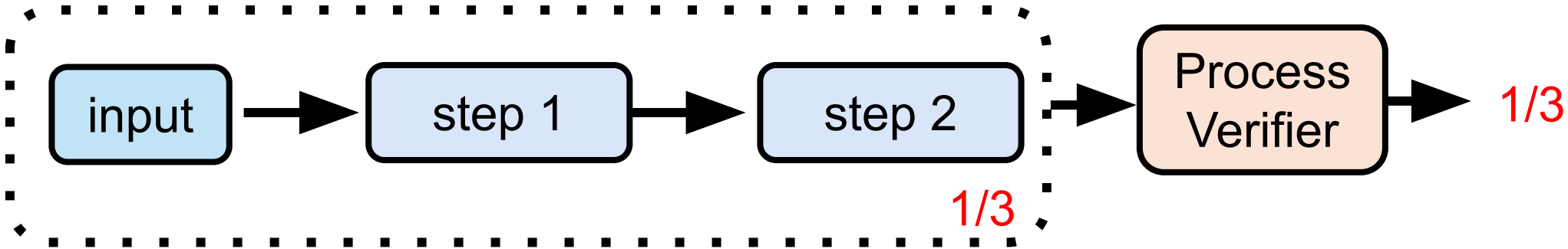


Training Data: **input** **ground truth**





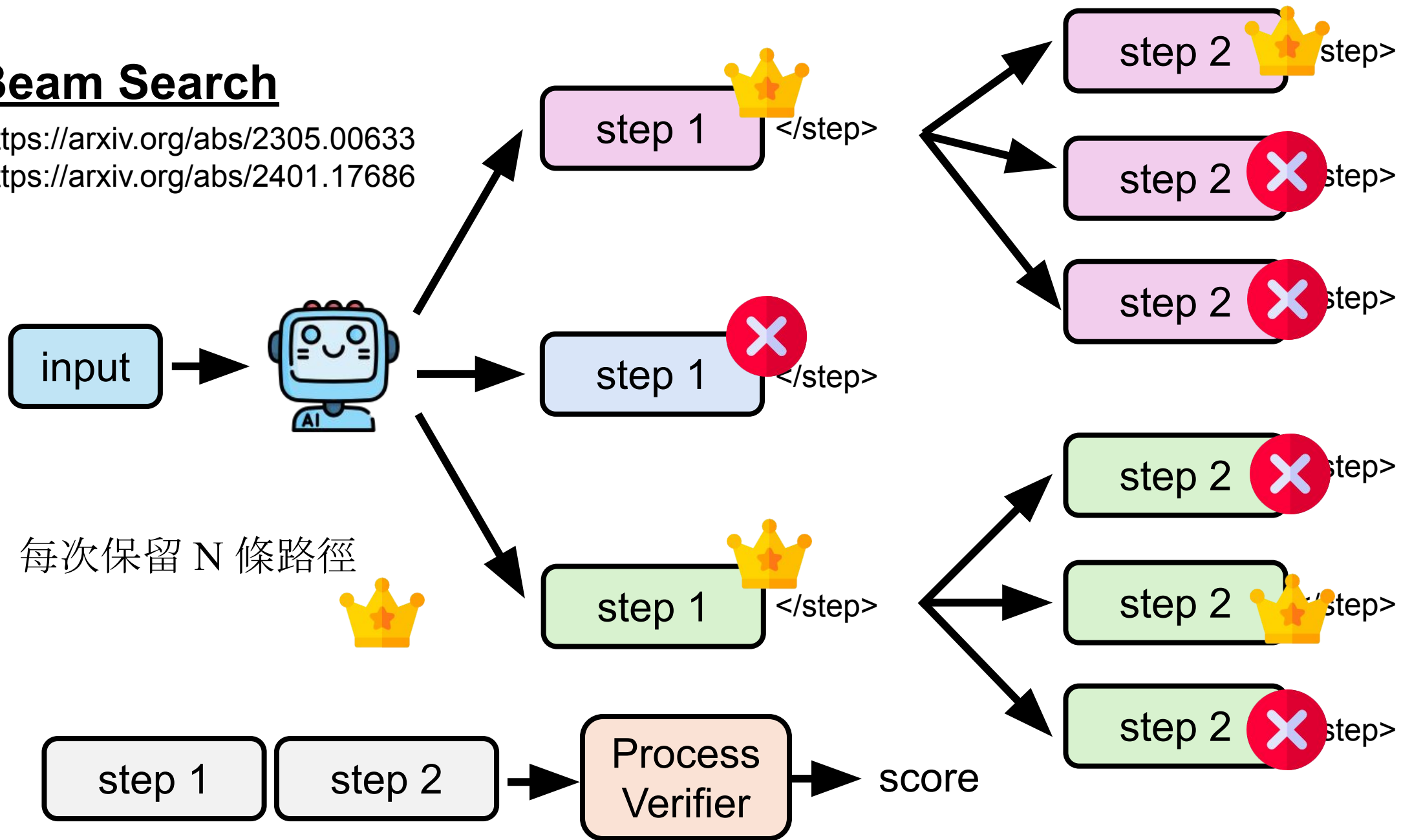
Math-Shepherd: Verify and Reinforce LLMs
Step-by-step without Human Annotations
<https://arxiv.org/abs/2312.08935>

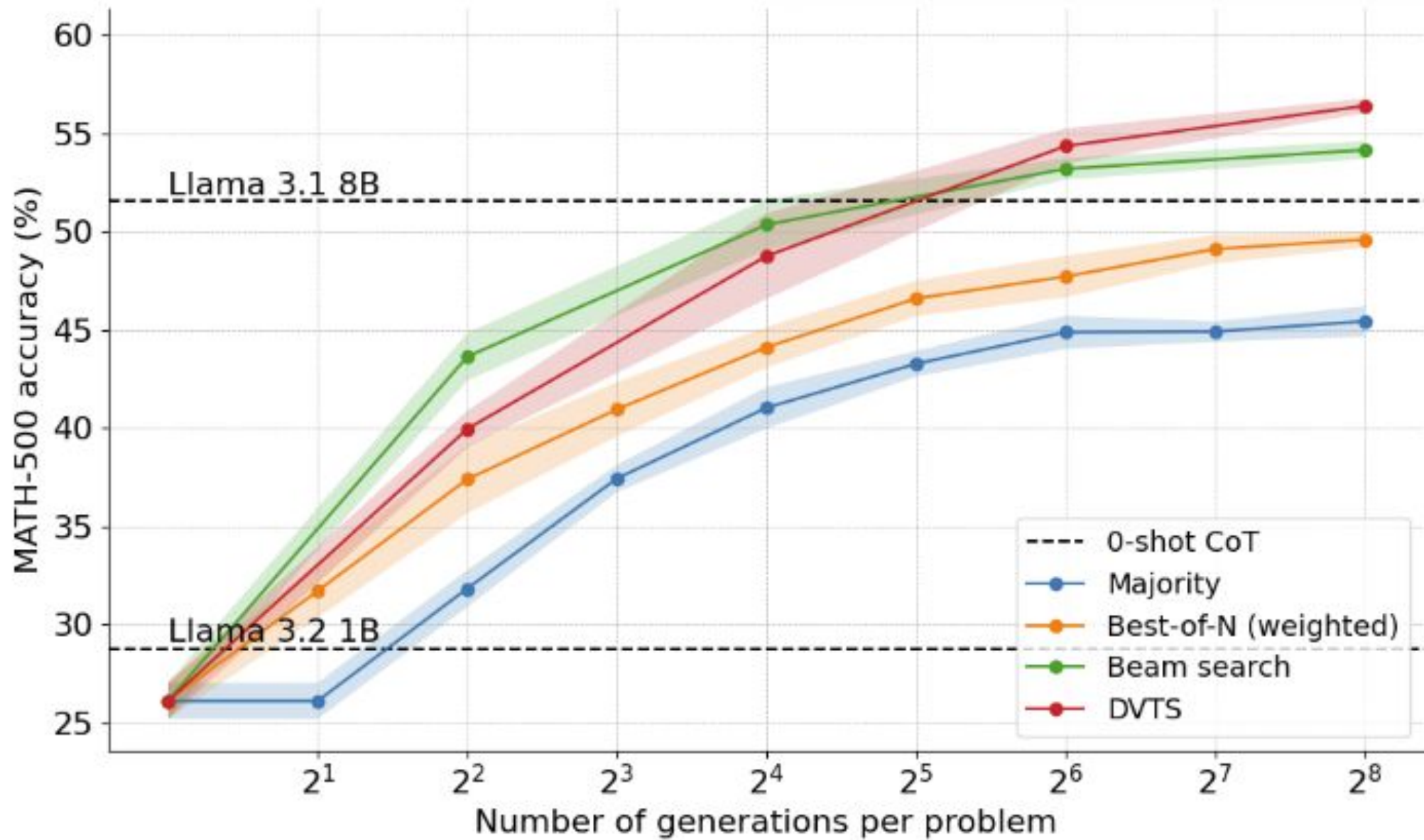


Beam Search

<https://arxiv.org/abs/2305.00633>

<https://arxiv.org/abs/2401.17686>

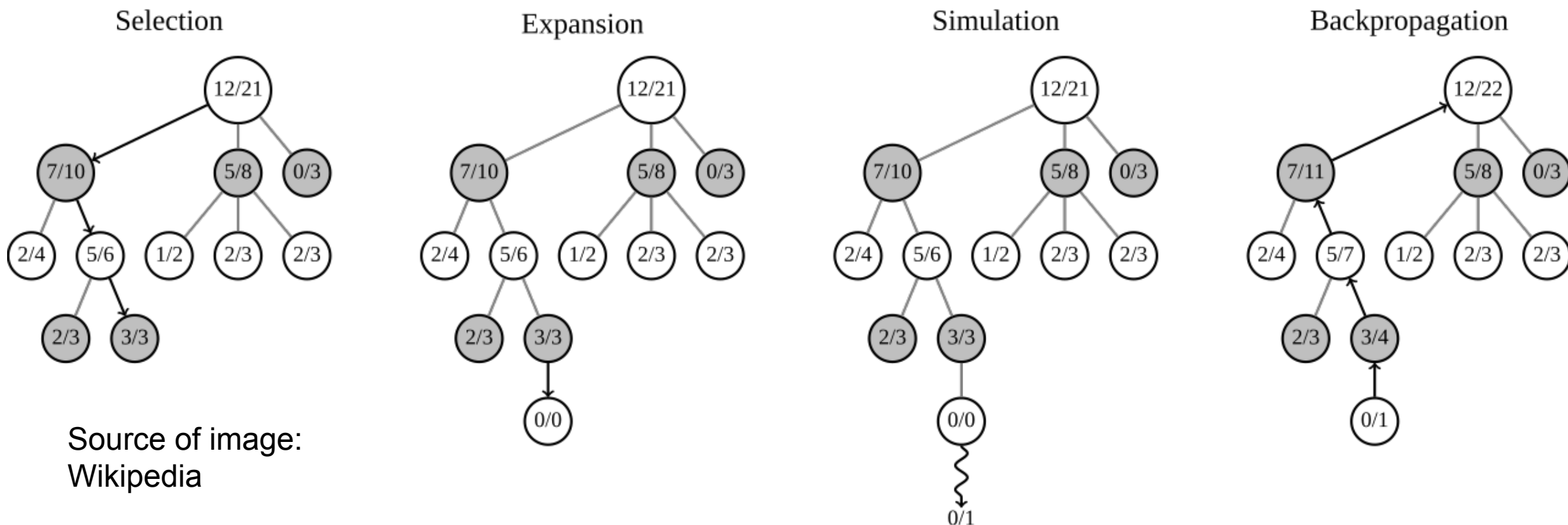




<https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute>

Heuristic Search Algorithm

e.g. Monte Carlo
Tree Search (MCTS)



Source of image:
Wikipedia

Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning<https://arxiv.org/abs/2405.00451>

ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search<https://arxiv.org/abs/2406.03816>

Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solvers<https://arxiv.org/abs/2408.06195>

打造「推理」語言模型的方法



教模型推理過程 (Imitation Learning)

以結果為導向學習推理 (Reinforcement Learning, RL)

需要微調參數

打造「推理」語言模型的方法

更強的思維鏈 (Chain-of-Thought, CoT)

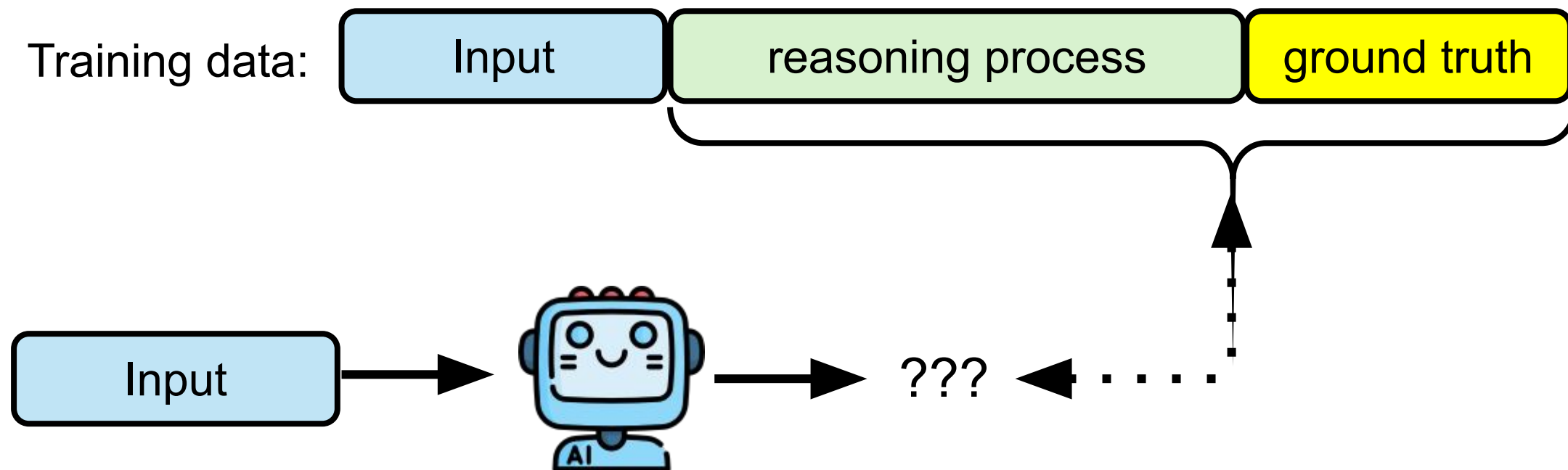
給模型推論工作流程

教模型推理過程 (Imitation Learning)

以結果為導向學習推理 (Reinforcement Learning, RL)

教模型推理過程

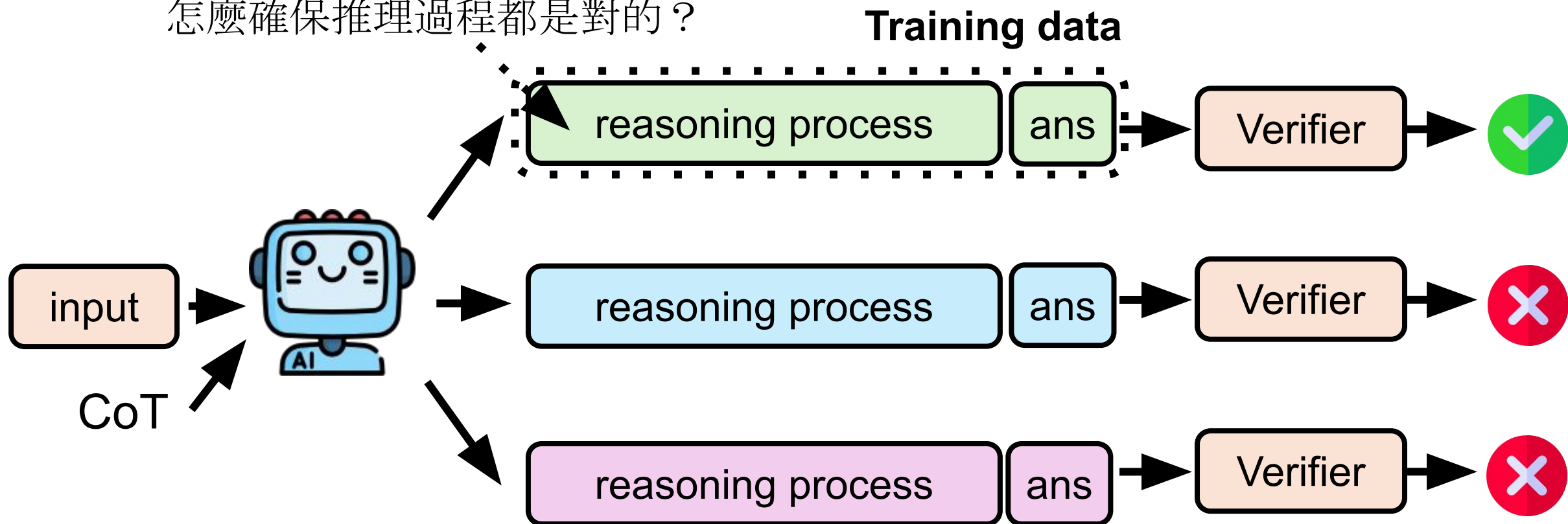
哪裡來？



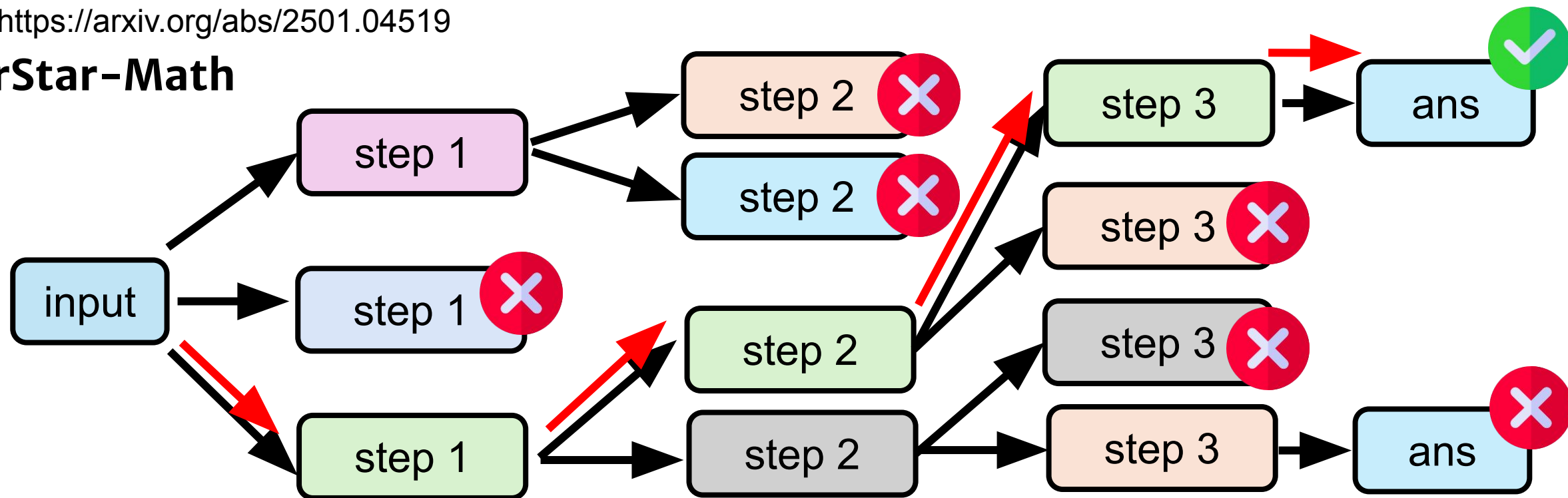
想辦法生成推論過程的訓練資料

Training Data: input ground truth

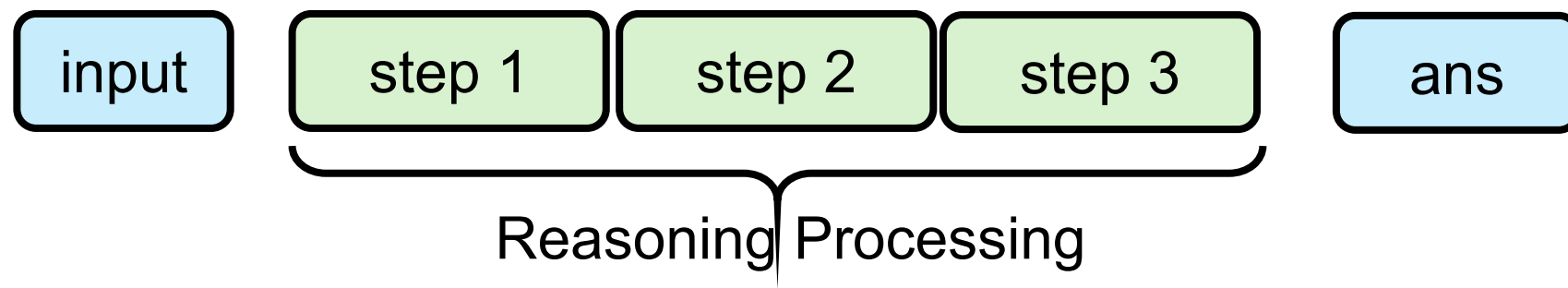
怎麼確保推理過程都是對的？



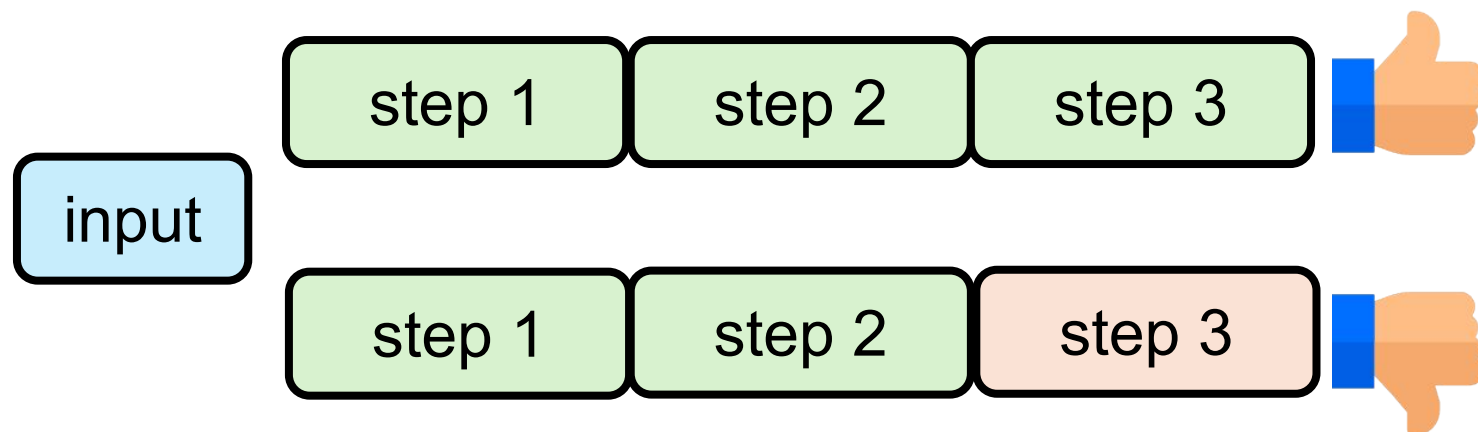
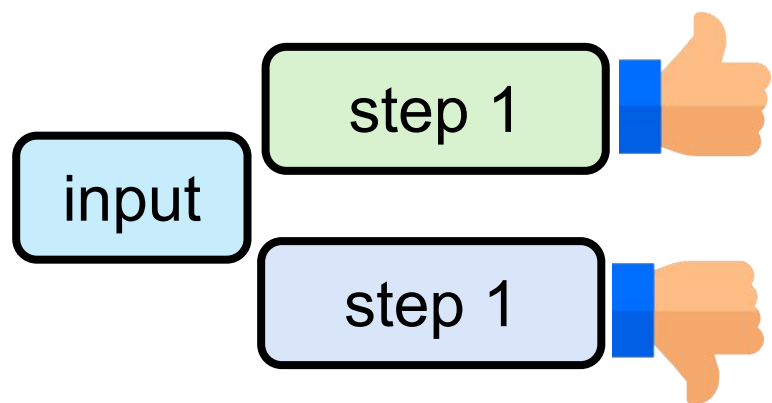
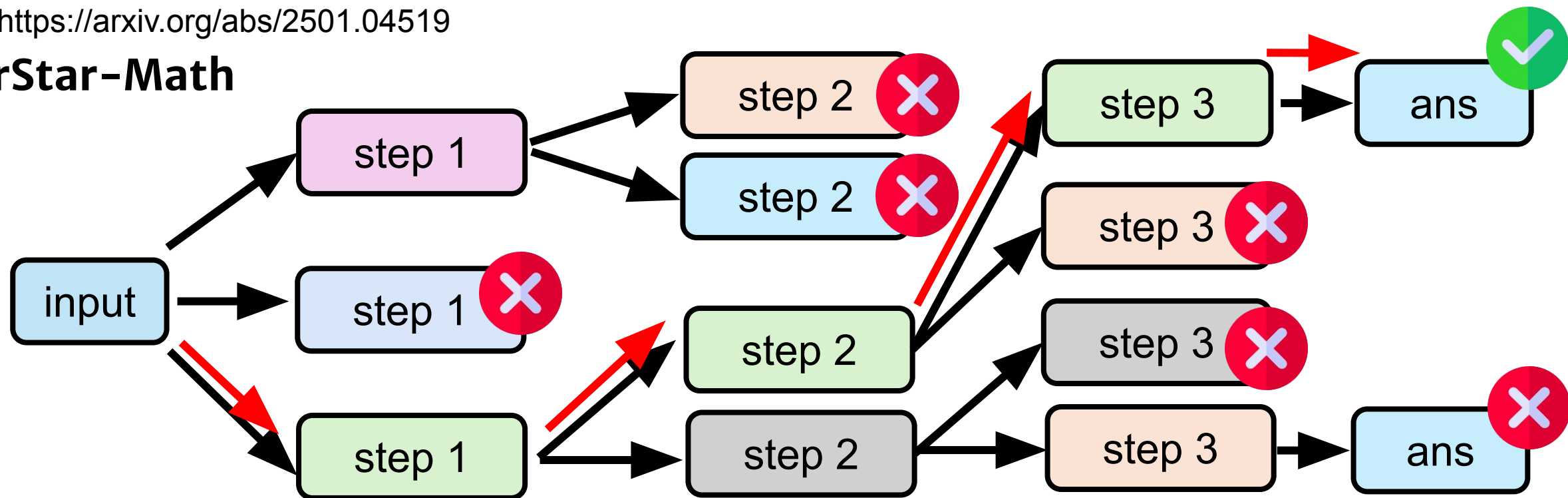
rStar-Math



Training data:



rStar-Math



推理過程需要每一步都是對的嗎？

s..t.rraw.b.e.r..ry 有幾個 "."

推理關於字串中的點數花了 26 秒 ✓

(正確答案為 9)

找找自己的問題！

只要最後答案正確就好

The user provided a query asking how many periods are in "s..t.rraw.b.e.r..ry".

Let's break it down: I see the string "s..t.rraw.b.e.r..ry".

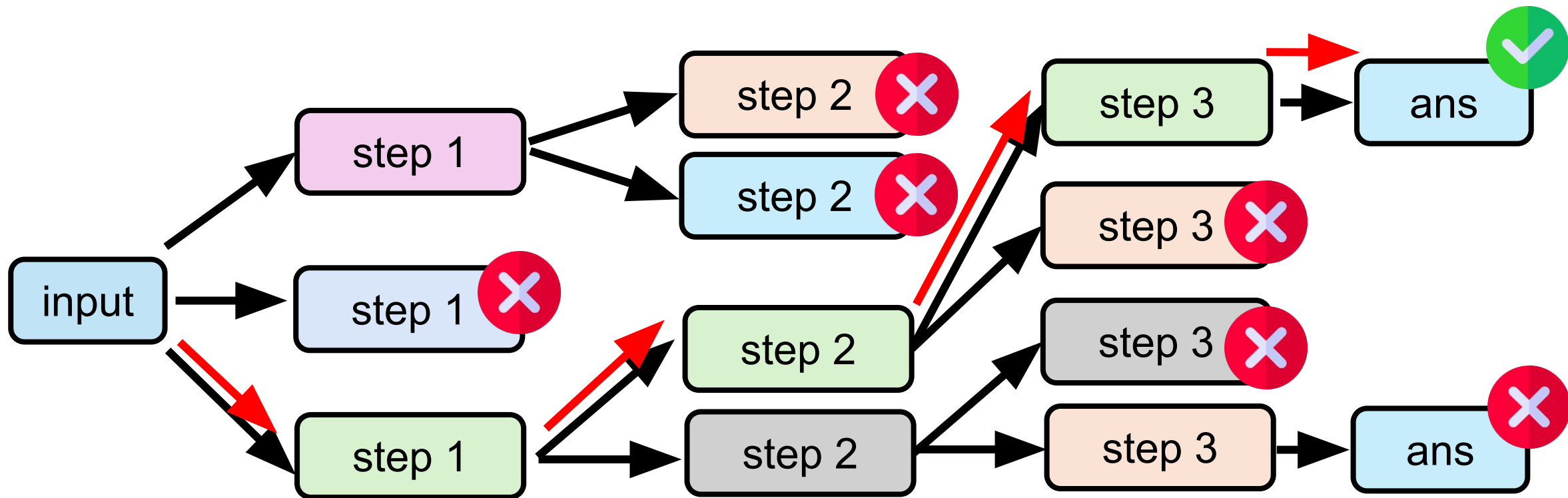
The periods are:

So, there are 8 periods in total based on my count.

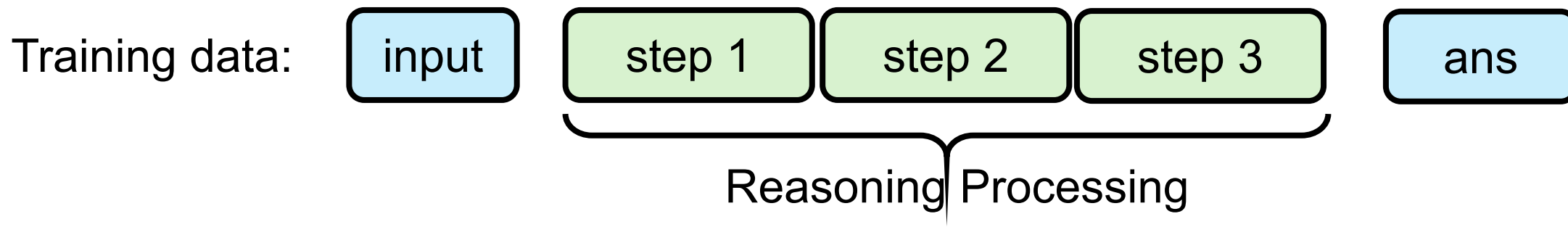
That's the answer to the question: There are **8** periods in the string.

Okay, let's carefully count the dots in the string "s..t.rraw.b.e.r..ry".

First, I see "s.." gives 2 periods. Then "t." adds 1 more for a total of 3. After "rraw" there's only 1 dot, bringing the total to 4. Then "b." adds 1 more, totaling 5, and "e." gives 6. The ".." after "r" adds 2, bringing a running total of 8. The final "r." adds 1 more, giving us a total of 9. So, in total, there are **9 periods** in "s..t.rraw.b.e.r..ry".

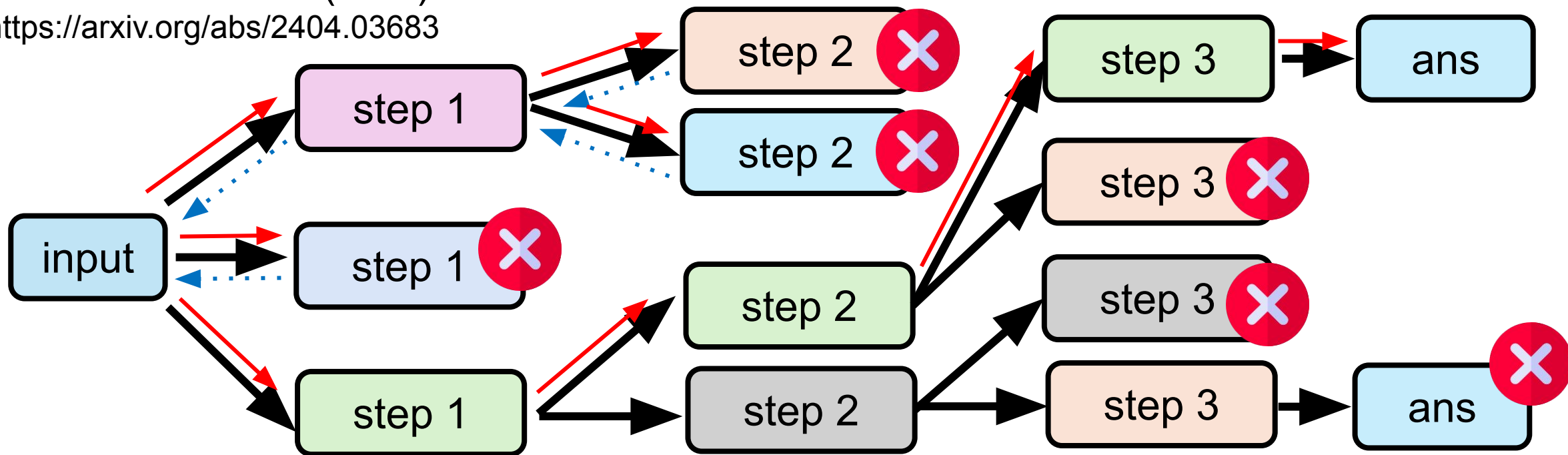


推理過程全是正確的，不知道找找自己的問題！

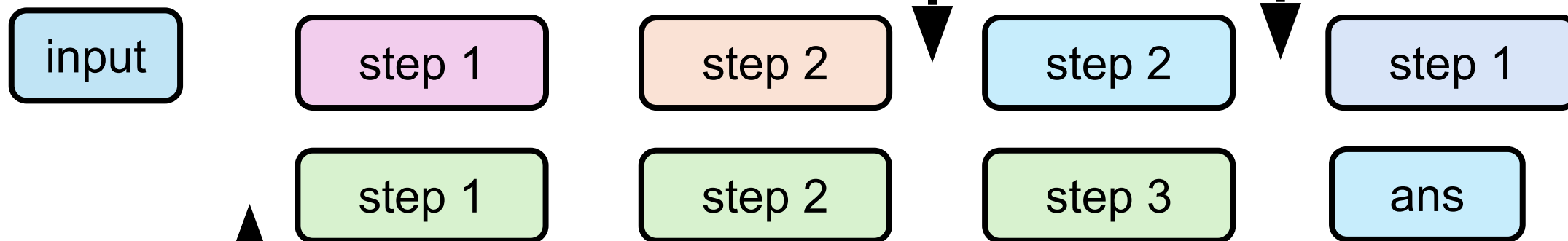


Stream of search (SoS)

<https://arxiv.org/abs/2404.03683>

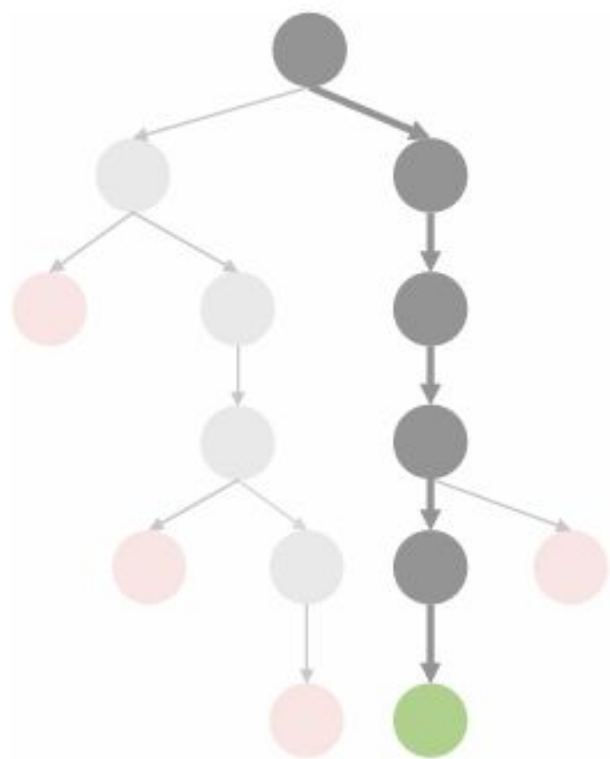


[插入 Verifier 的回饋] [插入 Verifier 的回饋]‘重新來過’

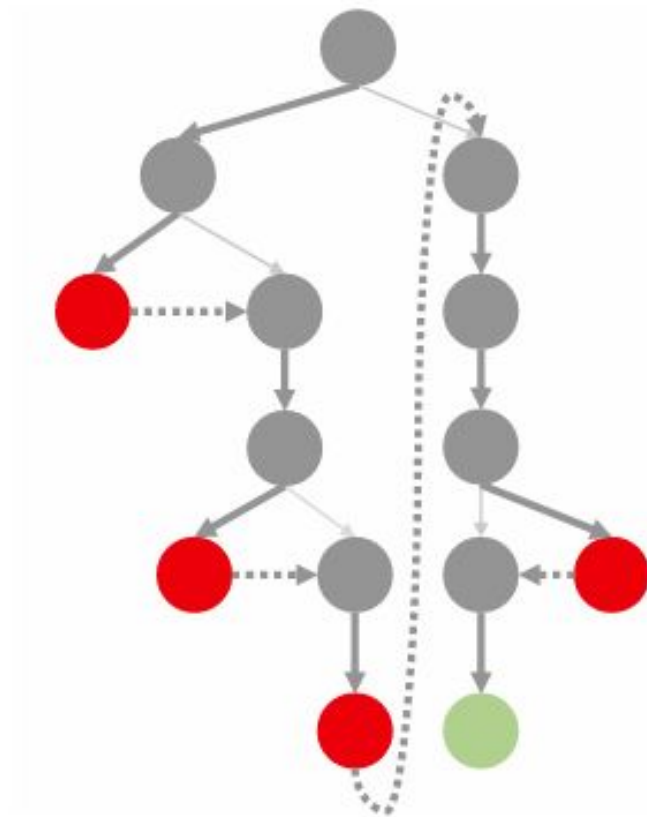


[插入 Verifier 的回饋]

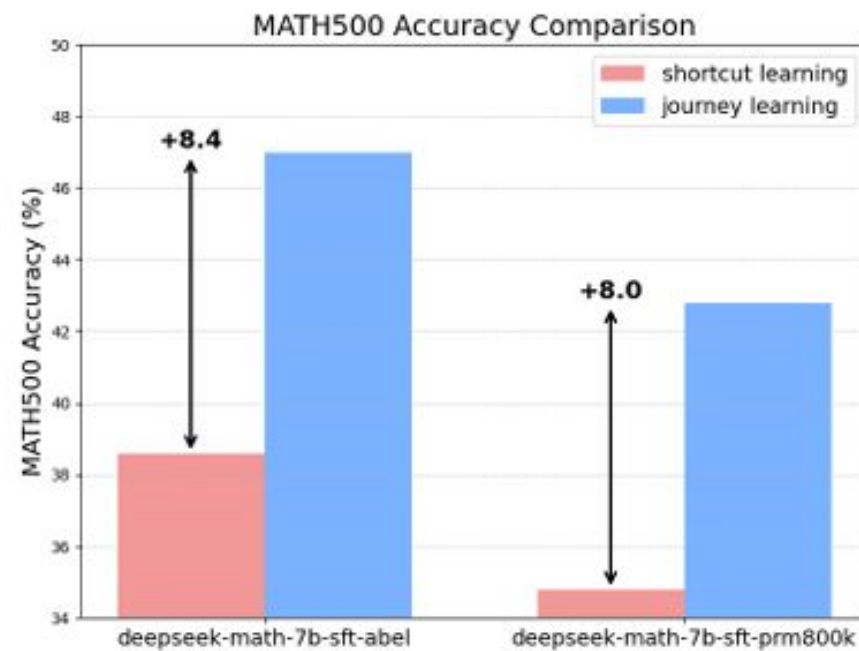
<https://arxiv.org/abs/2410.18982>



(a) Shortcut learning.

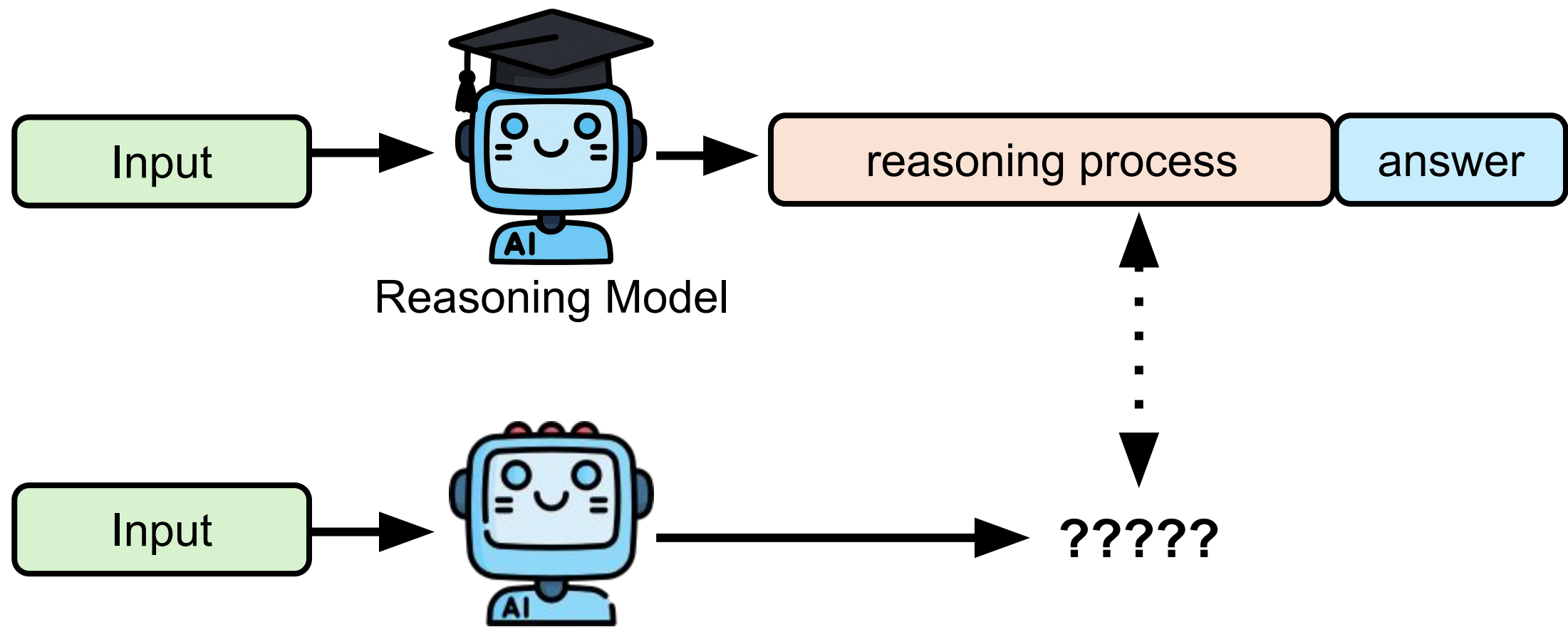


(b) Journey learning



(c) Performance Comparison

現在可以做 Knowledge Distillation



Sky-T1: <https://novasky-ai.github.io/posts/sky-t1/>
s1: <https://arxiv.org/abs/2501.19393>

現在可以做 Knowledge Distillation

<https://arxiv.org/abs/2501.12948>

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

Foundation Model

打造「推理」語言模型的方法

更強的思維鏈 (Chain-of-Thought, CoT)

給模型推論工作流程

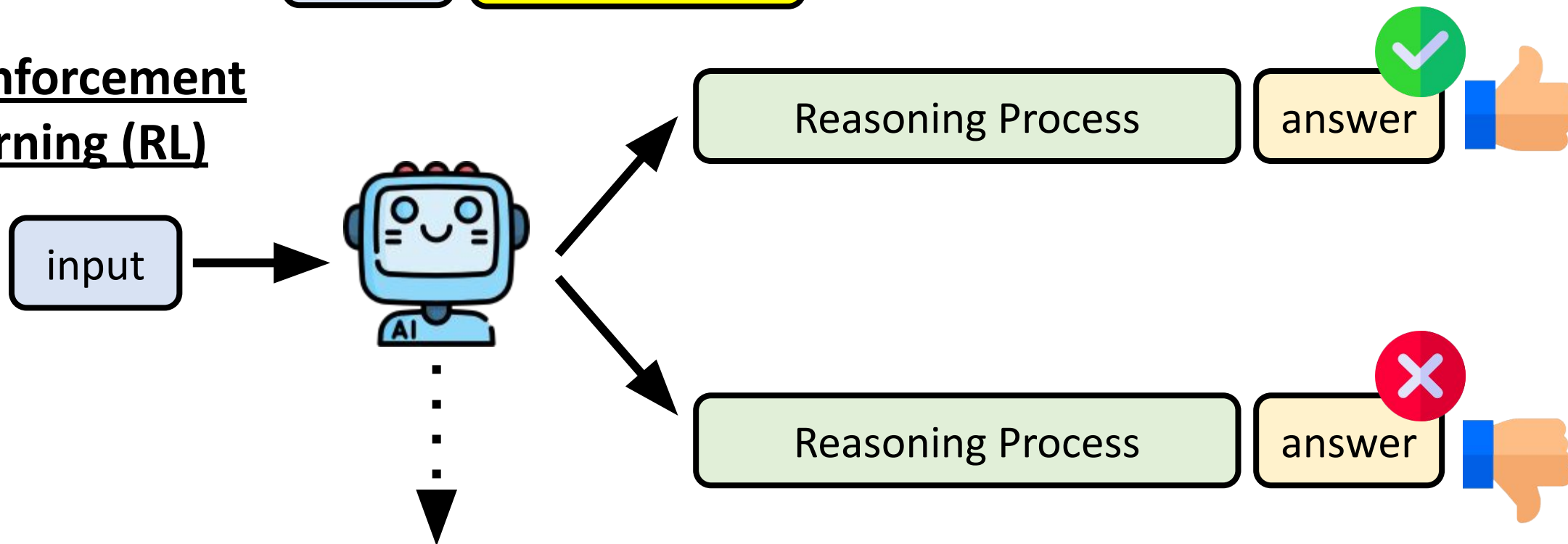
教模型推理過程 (Imitation Learning)

以結果為導向學習推理 (Reinforcement Learning, RL)

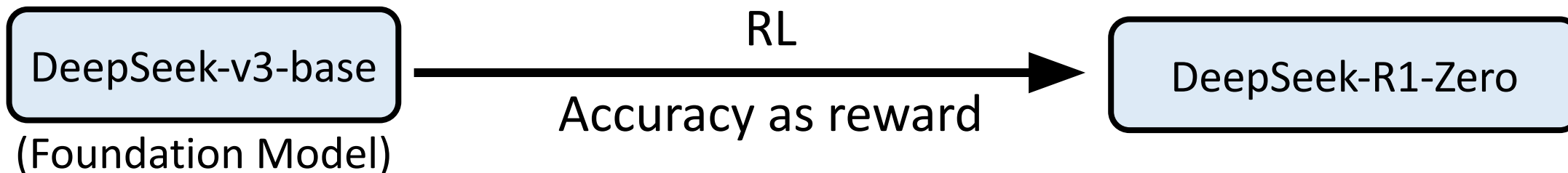
DeepSeek-R1 系列的作法

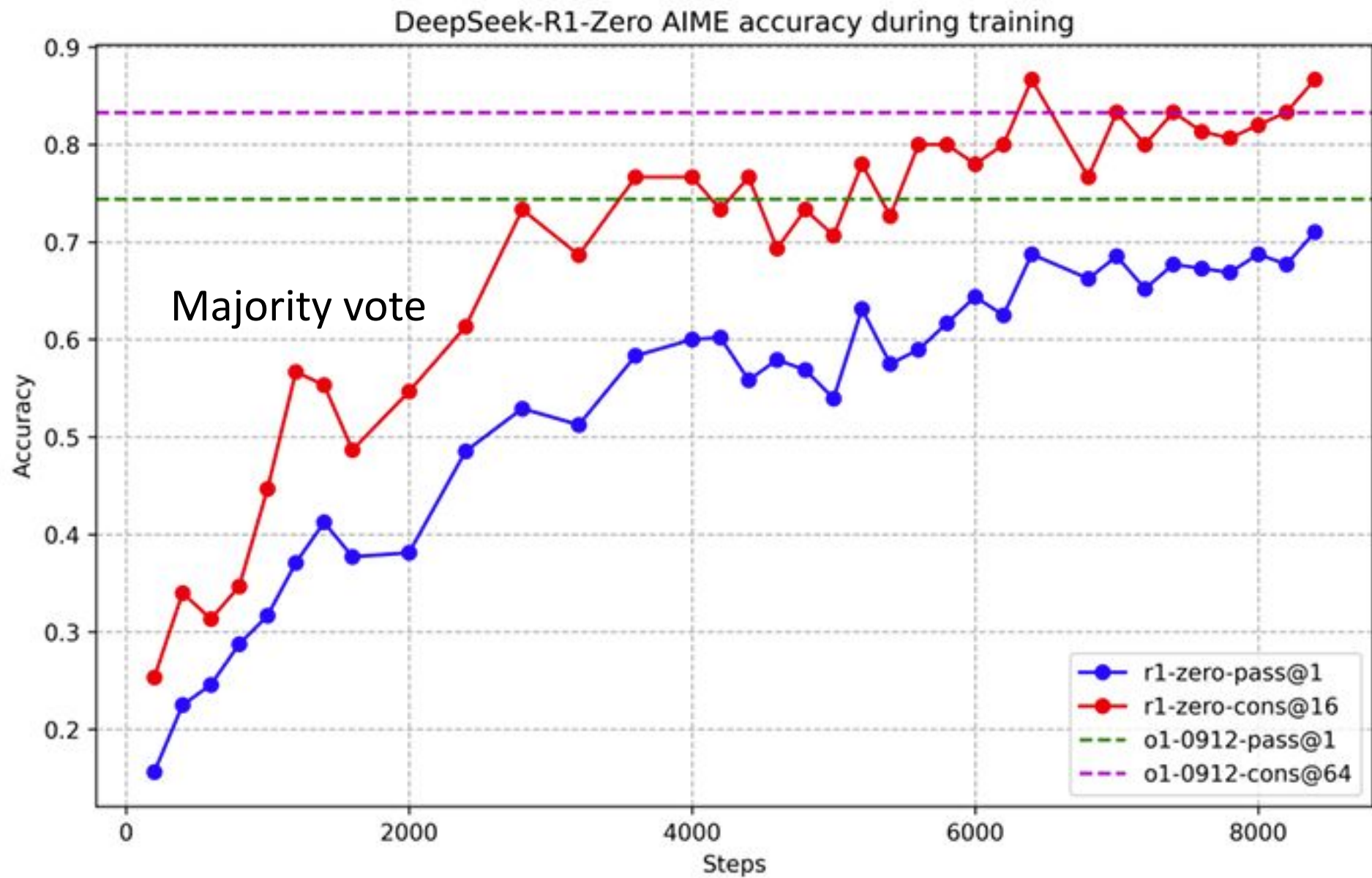
Training Data: input ground truth

Reinforcement Learning (RL)



推理過程內容不重要，最後答案是對的就好





Source of image: <https://arxiv.org/abs/2501.12948>

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 ...

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

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:

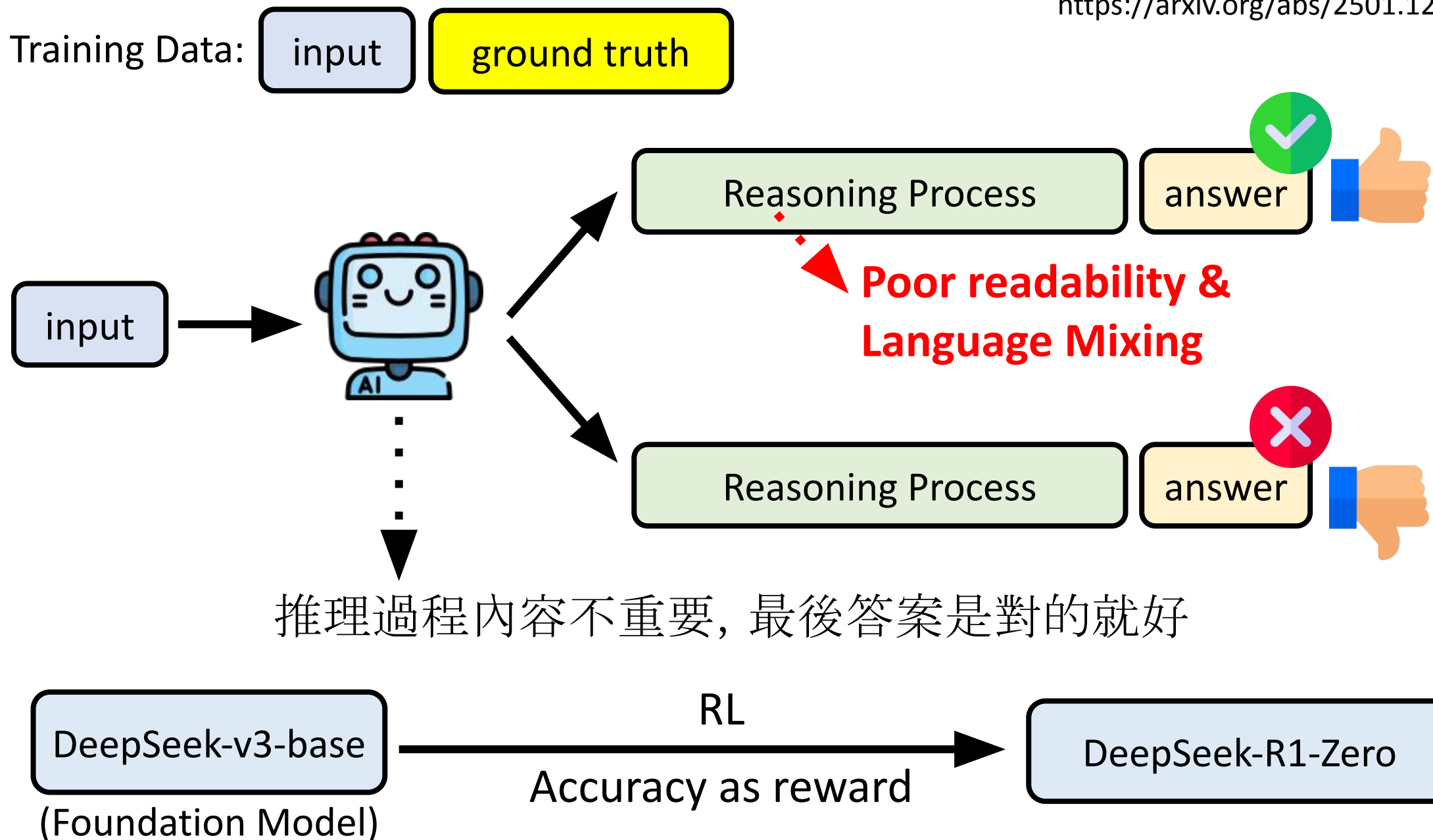
$$\sqrt{a - \sqrt{a + x}} = x$$

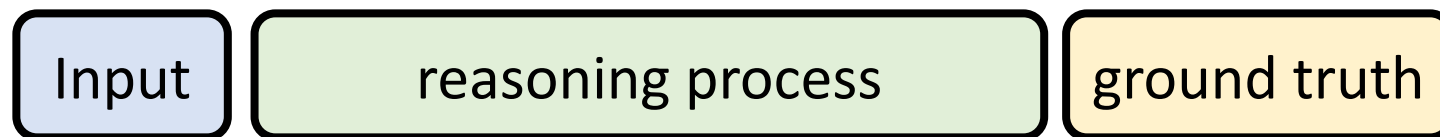
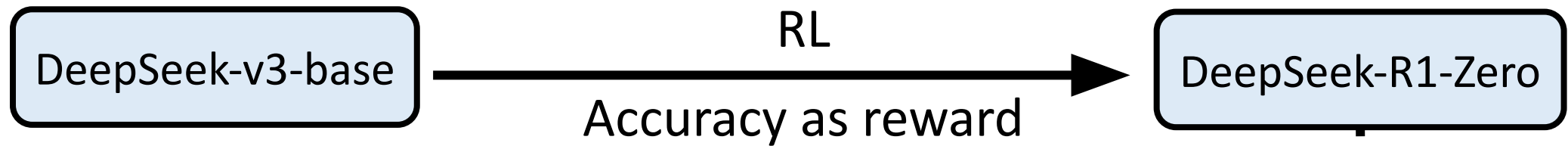
First, let's square both sides:

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

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

...





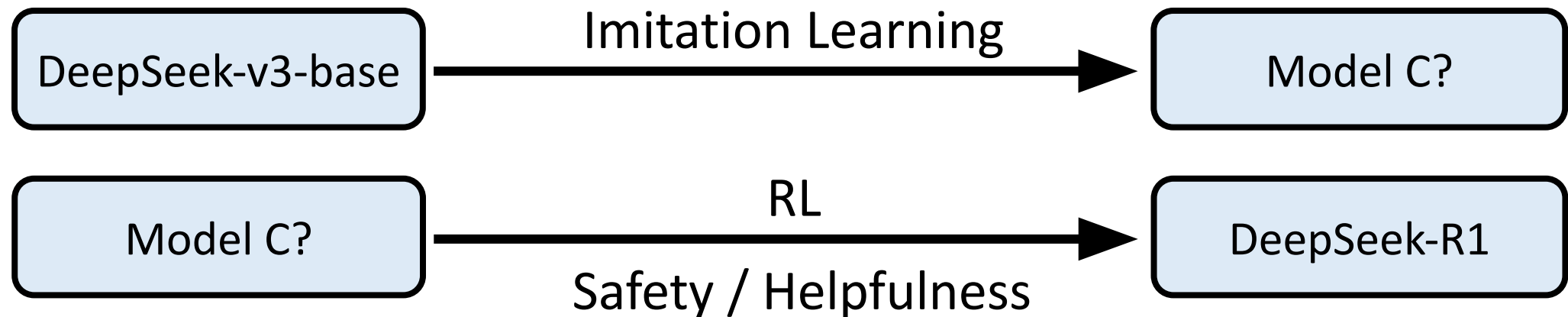
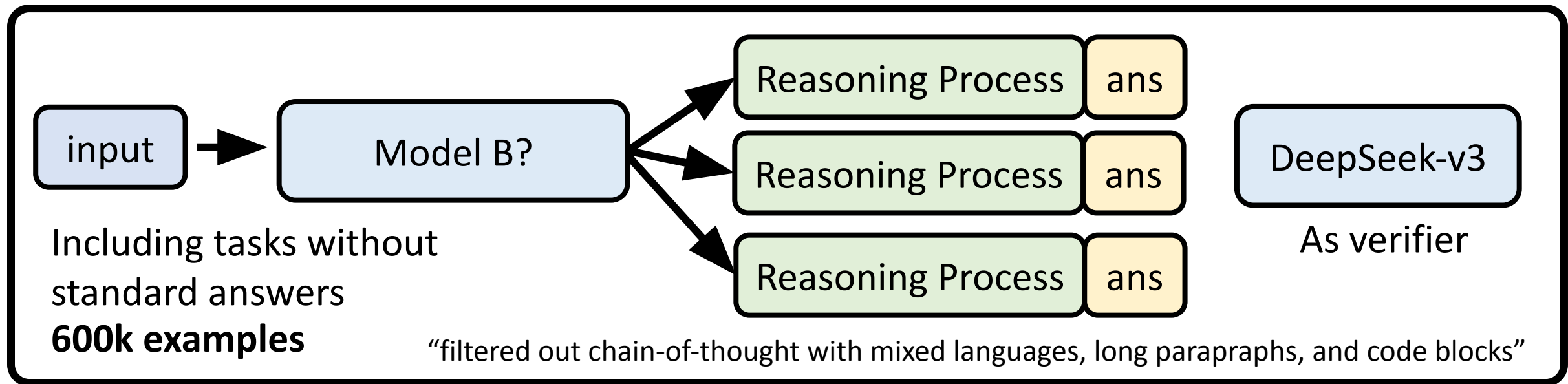
- using few-shot prompting with a long CoT as an example
- directly prompting models to generate detailed answers with reflection and verification



(Thousands of examples)

Generated data
+ **human annotation**





Based on the Deepseek-R1 paper, both the process verifier and MTCS were tried but ultimately not used.

strawberry 有幾個 r

所以r出现在第3、4、9、10的位置，总共有四个r。

不过这可能吗？因为正确的拼写是strawberry（不过正确的拼写其实是strawberry，中间一个r，后面两个r，所以正确的拼写有三个r？或者不，strawberry的正确拼写是s-t-r-a-w-b-e-r-r-y，中间一个r在straw，然后两个r在berry部分，所以总共有三个r？或者straw中的r是第一个，berry中的两个r，所以总共三个r？

不过用户输入的是strawberry，即前面有两个r，后面可能有两个r，所以总共有四个r？

或者可能用户输入的是strawberry，其中字母顺序可能有误，比如中间多了一个r，导致前面有两个r，后面两个r，总共四个？

所以根据用户提供的错误拼写“strawberry”，其中的r的数量是四个？

Foundation Model 很重要

以 Qwen-32B-Base 作為 Foundation Model

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

- 透過 RL 無法大幅強化 Qwen-32B-Base 的能力
- Imitation Learning (向 DeepSeek-R1學習) 對 Qwen-32B-Base 反而比較有效

RL 是強化模型原有的能力

Foundation Model
本來就會 Aha

Question(a): Find the least positive integer n such that $1/\sin 45^\circ \sin 46^\circ + 1/\sin 47^\circ \sin 48^\circ + \cdots + 1/\sin 133^\circ \sin 134^\circ = 1/\sin n^\circ$.

Response
First, I notice that the left-hand side of the equation has the form $\sum_{k=1}^n \frac{1}{\sin a_k \sin a_{k+1}}$, where $a_1 = 45^\circ, a_2 = 47^\circ, \dots, a_{90} = 133^\circ$. I also notice that $\sin a_{k+1} = \sin(a_k + 2^\circ)$.
...
Simplifying, I get $\sin(3a) = 3 \sin a - 4 \sin^3 a$.
Aha! I can use this to get $\sin(a + 2b) = \sin((a + b) + b) = 3 \sin(a + b) \dots$
This looks messy, but I notice that it has the form $\sin a \cos b + \cos a \sin b$ again.
This gives me an idea. Maybe I can use this identity repeatedly to simplify the left-hand side of the equation.
...

Question(b): An element is randomly chosen from among the first 15 rows ...

Response
In Pascal's Triangle, every row starts and ends with 1, ...
...
This can be calculated as: *awkward silence* **Wait, I'm overthinking. Let's try again.** The number of elements in the first n rows of Pascal's Triangle is $(n + 1)!/[2! * (n - 1)!]$.
...

Question(c): Denali and Nate work for a dog walking business and are paid ...

Response
To solve this problem, let's start by defining the variables...
...
Again, since $x \neq 0$, this equation doesn't give us a valid solution either. **Let's check if we made an error. We should verify the problem conditions again.** The correct approach is to ...
...

打造「推理」語言模型的方法

不用微調參數

更強的思維鏈 (Chain-of-Thought, CoT)

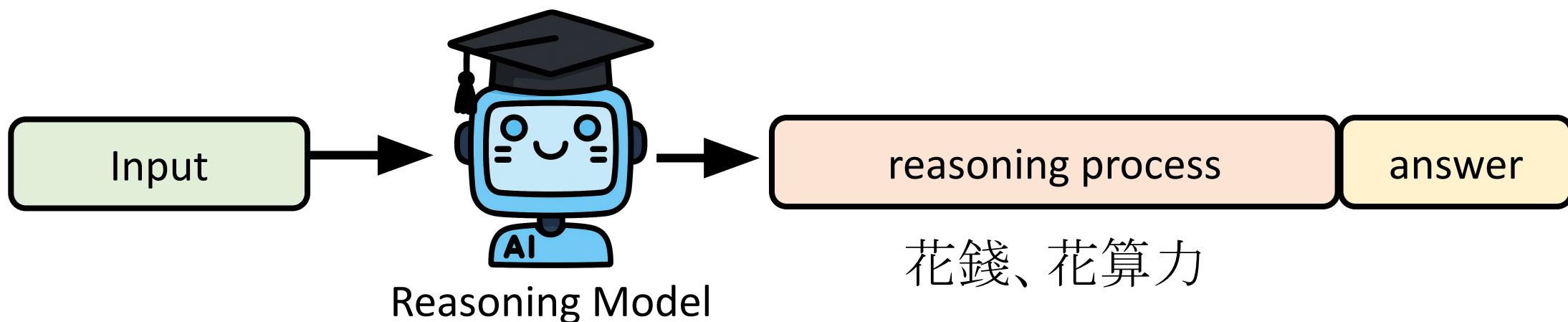
給模型推論工作流程

教模型推理過程 (Imitation Learning)

以結果為導向學習推理 (Reinforcement Learning, RL)

需要微調參數

下回預告：推理模型的挑戰與未來發展



$$123 \times 456 = ?$$



已深度思考 (用时 59 秒) ^

planning

Okay, so I need to multiply 123 by 456. Hmm, let's see. I remember that multiplying larger numbers can be broken down into smaller parts, right? Maybe I can use the distributive property here. Let me think.

First, I should probably write it out step by step. 123 times 456. Let me break down 456 into 400, 50, and 6.

That way, I can multiply 123 by each of those parts separately and then add them up. So, 123×400 , 123×50 , and 123×6 . Then add all those results together. That should work.

Verification (for a step)

Starting with 123×400 . Well, 123 times 4 is 492, right? Because 100×4 is 400, 20×4 is 80, and 3×4 is 12.

Adding those together: $400 + 80 = 480$, plus 12 is 492. So, 123×400 would be 492 with two zeros added at the end, which makes 49,200. Got that part.

(sub)planning

Next up is 123×50 . Hmm, multiplying by 50 is the same as multiplying by 5 and then adding a zero. So, 123×5 . Let me calculate that. 100×5 is 500, 20×5 is 100, and 3×5 is 15. Adding those together: $500 + 100 = 600$, plus 15 is 615. Then add the zero to make it 50, so 6150. So, 123×50 equals 6,150. Wait, no, adding a zero would make it 6150, right? Yeah, that's 6,150. Okay.

Now the last part is 123×6 . Let's do that. 100×6 is 600, 20×6 is 120, and 3×6 is 18. Adding those: $600 + 120 = 720$, plus 18 is 738. So, 123×6 is 738.

Now I need to add up all these three results: $49,200 + 6,150 + 738$. Let's add them step by step. First, $49,200 + 6,150$. Let's see, $49,200 + 6,000$ is 55,200. Then add the remaining 150: $55,200 + 150 = 55,350$. Then add the last part, 738. So, $55,350 + 700$ is 56,050, and then +38 more is 56,088. Wait, is that right? Let me check again.

verification

Wait, $55,350 + 738$. Let me break down 738 into $700 + 38$. So, $55,350 + 700$ is 56,050. Then add 38: $56,050 + 38$. $56,050 + 30$ is 56,080, then +8 is 56,088. Okay, so total is 56,088. So, 123×456 should be 56,088. Hi



verification + explore

But let me verify that again because it's easy to make a mistake in these calculations. Maybe another way to check. Let's try multiplying 123 by 456 using the standard algorithm. Let me write it out:

123

x 456

First, multiply 123 by 6: $123 \times 6 = 738$. That's the same as before.

Then, multiply 123 by 50, which is 123×5 (615) and add a zero, making 6,150. That's the second line, shifted one position to the left because it's the tens place.

Next, multiply 123 by 400, which is 123×4 (492) and add two zeros, making 49,200. That's the third line, shifted two positions to the left because it's the hundreds place.



Now, add them all together:

738

+6150

+49200

56,088

verification + explore

Yep, same result. So, that seems correct. But just to be absolutely sure, maybe I can use another method or check with estimation. Let's see.

Estimation: 123 is approximately 120, and 456 is approximately 450. So, $120 \times 450 = 54,000$. But our answer is 56,088, which is a bit higher. That makes sense because we rounded both numbers down. 123 is 3 more than 120, and 456 is 6 more than 450. The exact product should be higher than 54,000, which matches our result of 56,088. So, the estimation gives a ballpark figure that seems reasonable.

Alternatively, maybe I can use the vertical multiplication method again step by step to confirm. Let's do that:

verification + explore

Multiplying 456 by 123:

Start with $456 \times 3 = 1,368$

Then $456 \times 20 = 9,120$

Then $456 \times 100 = 45,600$

Add them together: $1,368 + 9,120 = 10,488$; $10,488 + 45,600 = 56,088$.



verification

Wait, that's another way to break it down. Instead of breaking down the first number, break down the second. So, 123 is $100 + 20 + 3$. So, $456 \times 100 = 45,600$; $456 \times 20 = 9,120$; $456 \times 3 = 1,368$. Adding those gives $45,600 + 9,120 = 54,720$; $54,720 + 1,368 = 56,088$. Same answer. So, that's reassuring.

Another cross-check: use calculator logic. Let me compute 123×456 step by step mentally. $123 \times 400 = 49,200$; $123 \times 50 = 6,150$; $123 \times 6 = 738$. Adding those gives $49,200 + 6,150 = 55,350$; $55,350 + 738 = 56,088$. All right, consistent.

verification

Alternatively, maybe use the formula $(a + b)(c + d)$ but that might complicate. Alternatively, use the difference of squares? Probably not necessary here. Alternatively, use the fact that $123 \times 456 = (100 + 20 + 3)(400 + 50 + 6)$. Expanding term by term:

verification + explore

$$100 \times 400 = 40,000$$

$$100 \times 50 = 5,000$$

$$100 \times 6 = 600$$

$$20 \times 400 = 8,000$$

$$20 \times 50 = 1,000$$

$$20 \times 6 = 120$$

⋮

$$54,720 + 1,200 = 55,920$$

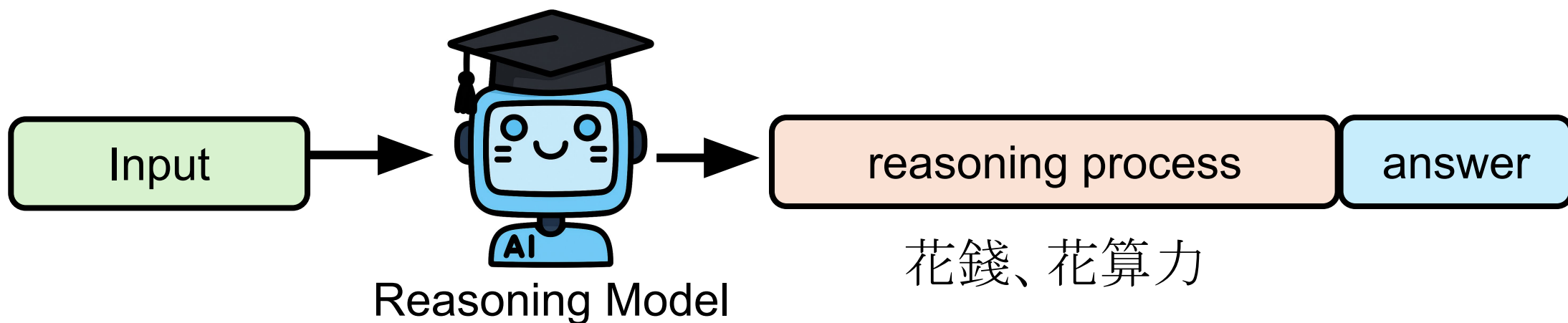
$$55,920 + 150 = 56,070$$

$$56,070 + 18 = 56,088.$$

Wow, that works too. So, breaking down each digit and multiplying each pair, then adding gives the same result. So, all methods confirm that $123 \times 456 = 56,088$.

Therefore, I'm pretty confident the answer is 56,088. No steps seem to have errors, cross-check to the same result. So, I think that's it.

下回預告：推理模型的挑戰與未來發展



真的需要產生這麼長的推論過程嗎？
有沒有辦法縮短推論過程