

# CSCSE 638 Natural Language Processing Foundation and Techniques

## Lecture 14: Test-Time Scaling, Large Reasoning Models

Kuan-Hao Huang

Spring 2026



(Some slides adapted from Hung-Yi Lee)

# Project Proposal

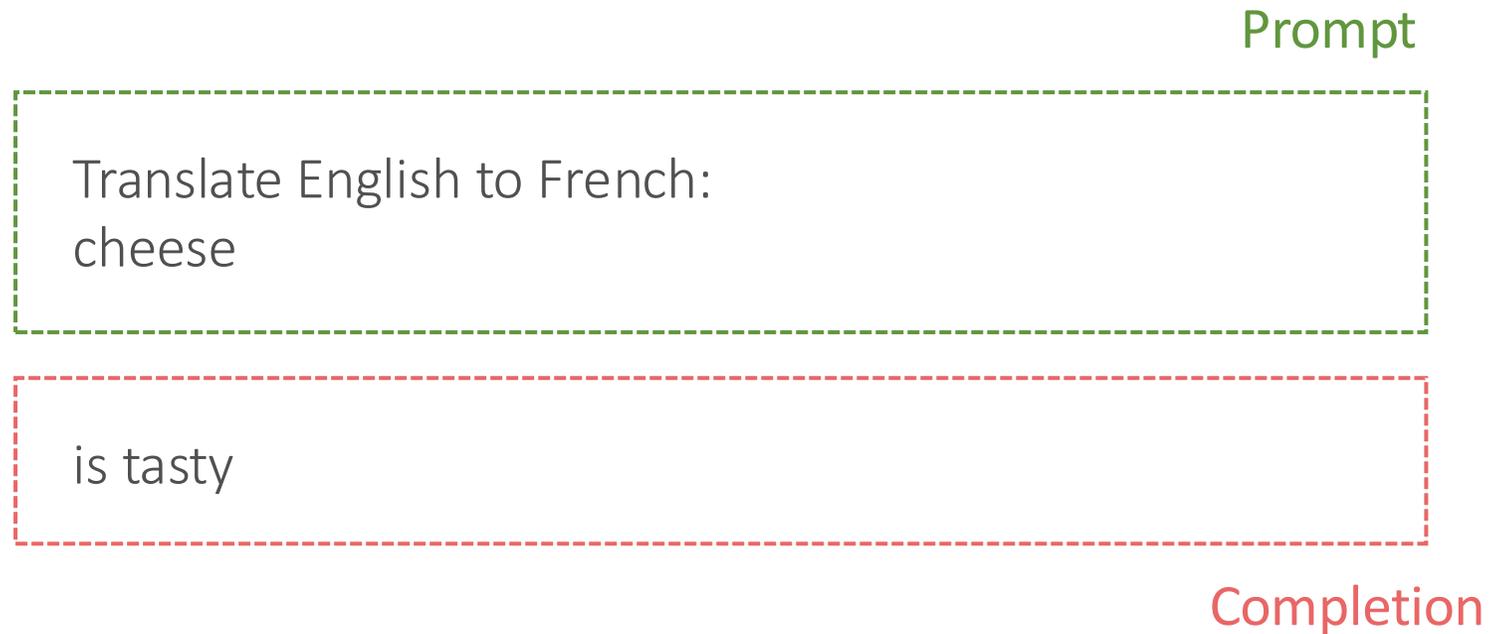
- Due: Mar 6
- Page limit: 2 pages (excluding reference)
- Format: ACL style

# Team Sign-Up

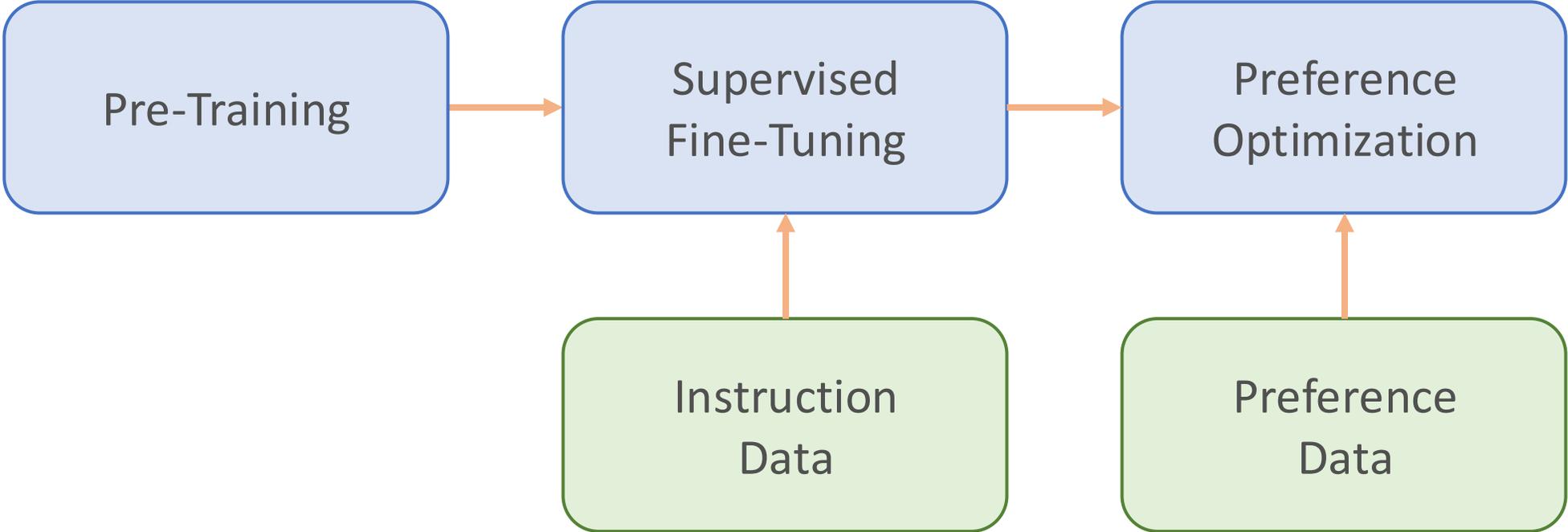
- <https://docs.google.com/spreadsheets/d/1qUZPFI4wciToJsXye8-WN4L7xVG38IWdS2GCCzmu84A/edit?usp=sharing>

# Recap: Alignment and Post-Training

- Prompt → Completion
  - Continue writing



# Recap: Post-Training Pipeline



# Recap: Instruction Tuning

- LLMs have knowledge, but don't always generate the outputs we want
- Training LLMs to following **human instructions**
  - Convert existing tasks to (instruction, input, output) format
  - Create many prompts and collect human answers

Annotated task definitions	Category	Description
<p>You will be given two pieces of text... <u>One of them is simpler ...</u> You are expected to output 'Text one' if the first sentence is simpler. Otherwise output 'Text two'.</p>	<u>Input Content</u> <u>Additional Input Content</u>	Primary description of the task input Additional details on task input
<p>Given a sentence with a missing word, <u>pick the answer option that best fills out</u> the missing word in the sentence. <u>Indicate each answer with its index ('a', 'b', 'c', 'd')</u>.</p>	<u>Action Content</u> <u>Input Mention</u> <u>Output Content</u>	Action to perform for task Mentions of input within action content Primary description of task output
<p>Given a document, <u>generate</u> a short title of the document. <u>The title should convey the main idea/event/topic about which the document is being written.</u></p>	<u>Additional Output Content</u> <u>Label List</u> <u>Label Definition</u>	Additional details on task output Task output labels (classification only) Task Label definitions (classification only)

# Recap: Reinforcement Learning with Human Feedback

- Human reward

SAN FRANCISCO,  
California (CNN) --  
A magnitude 4.2  
earthquake shook the  
San Francisco

...  
overturn unstable  
objects.

An earthquake hit  
San Francisco.  
There was minor  
property damage,  
but no injuries.

$$R(s_1) = 8.0$$

The Bay Area has  
good weather but is  
prone to  
earthquakes and  
wildfires.

$$R(s_2) = 1.2$$

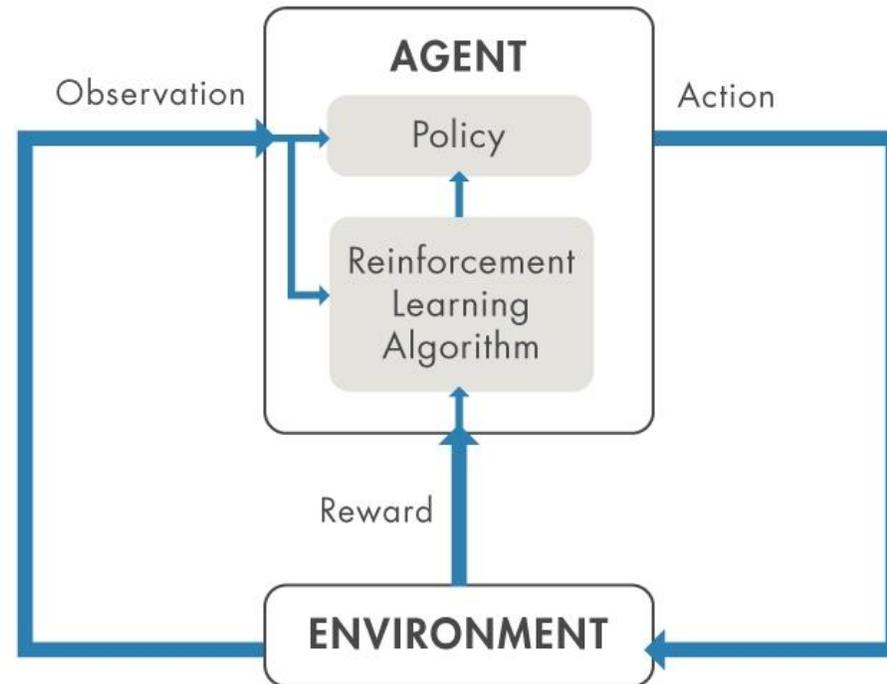
Goal: maximize the expected reward of samples from our LM

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})]$$

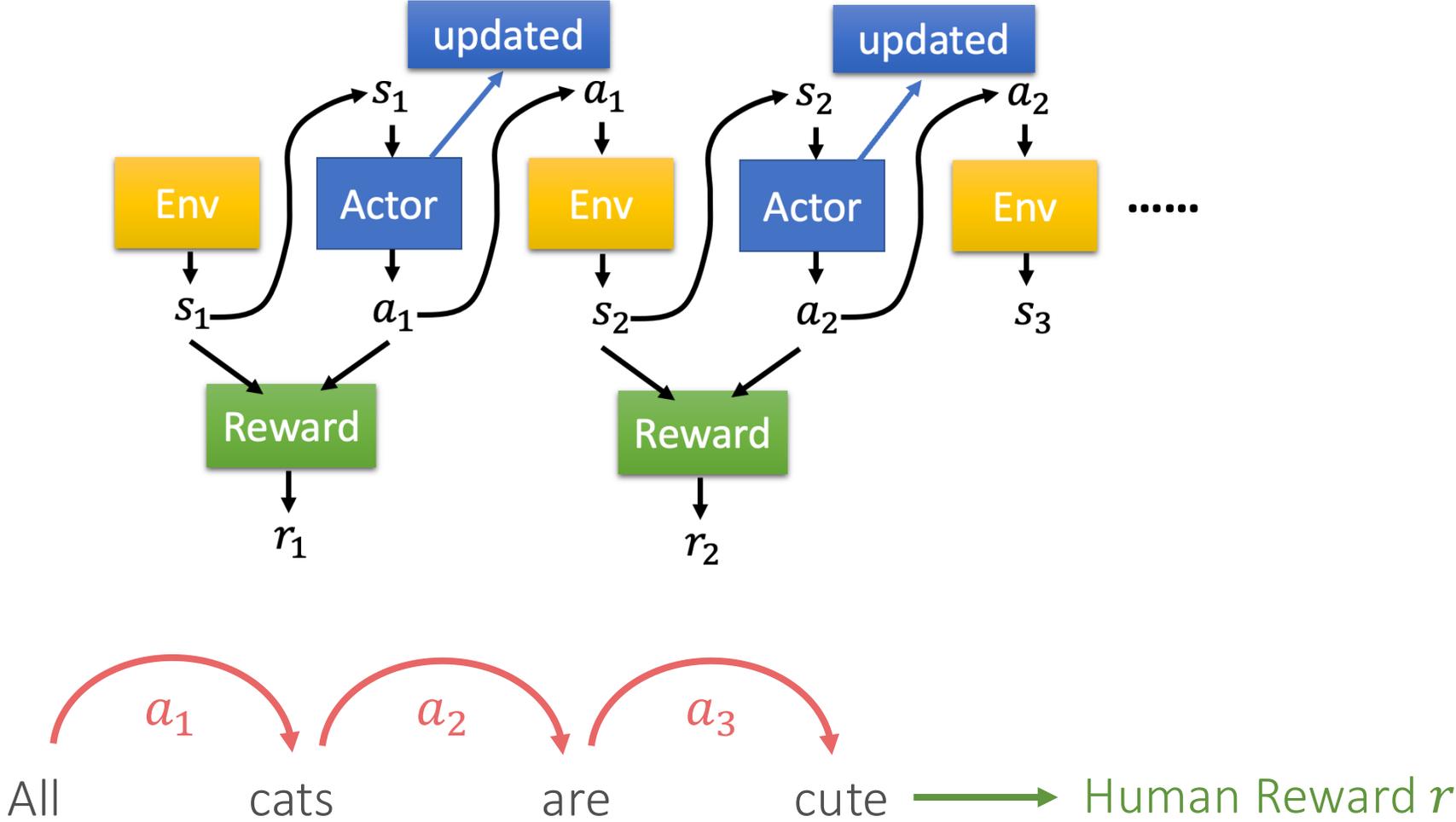
# Recap: Reinforcement Learning with Human Feedback

How do we change the LM parameters  $\theta$  to maximize this?

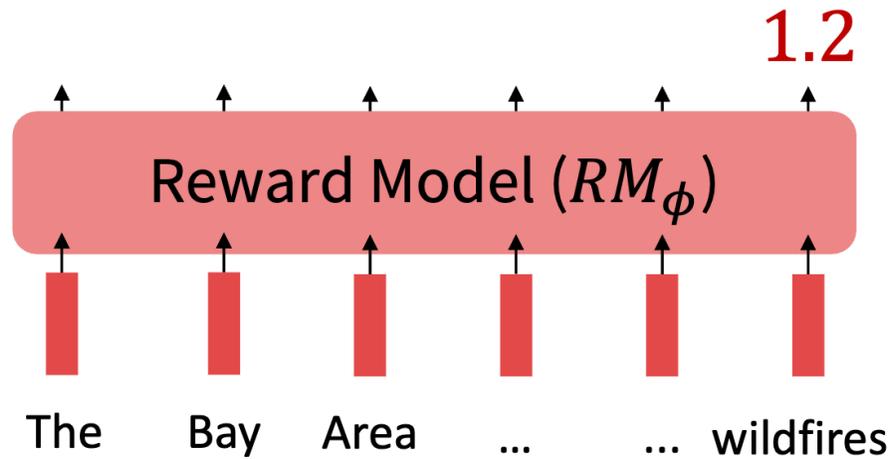
$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})]$$



# Recap: Reinforcement Learning with Human Feedback



# Recap: Reward Model



Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l) \sim D} [\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))]$$

“winning”  
sample

“losing”  
sample

$s^w$  should score  
higher than  $s^l$

# Recap: RLHF/PPO

- We want to optimize:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y} | x)} [RM_{\phi}(x, \hat{y})]$$

- Do you see any problems?
  - Learned rewards are imperfect; this quantity can be imperfectly optimized
- Add a penalty for drifting too far from the initialization:

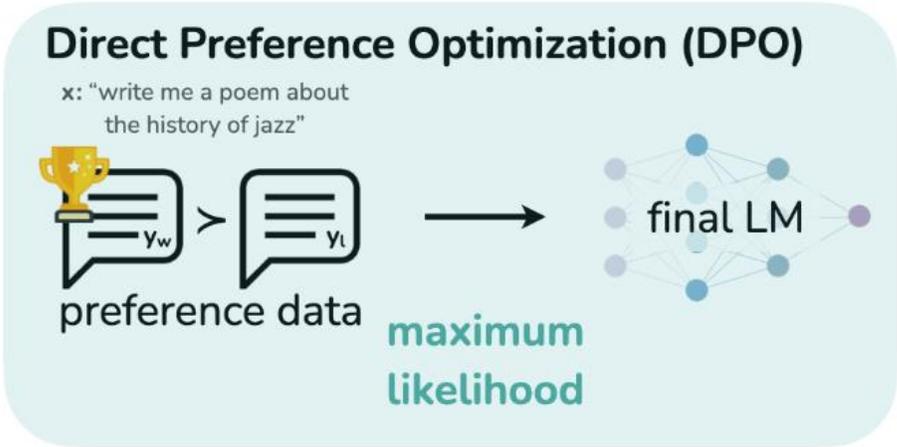
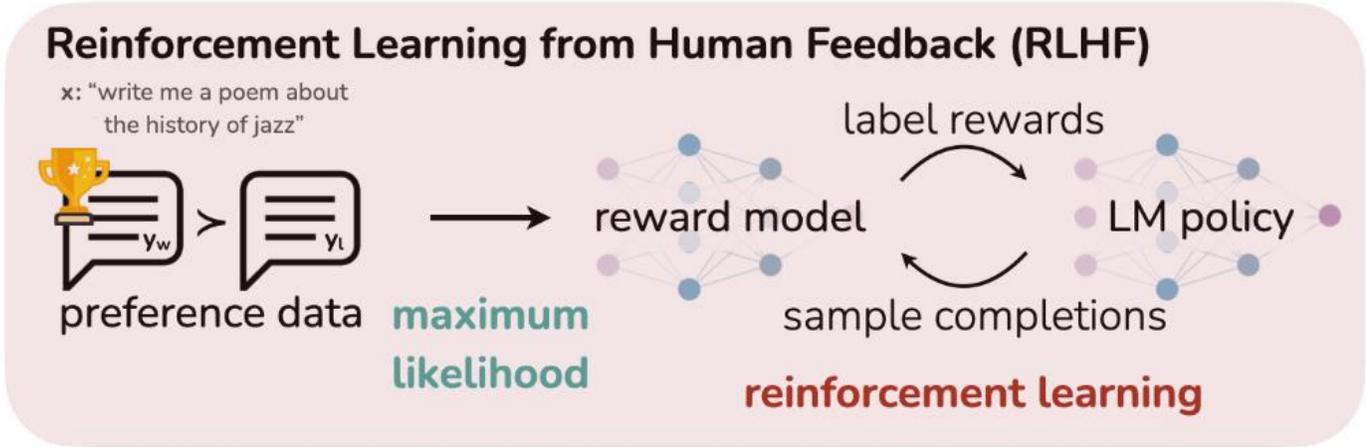
$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y} | x)} [RM_{\phi}(x, \hat{y}) - \underbrace{\beta \log \left( \frac{p_{\theta}^{RL}(\hat{y} | x)}{p^{PT}(\hat{y} | x)} \right)}_{\text{penalty}}]$$

Pay a price when

$$p_{\theta}^{RL}(\hat{y} | x) > p^{PT}(\hat{y} | x)$$

This penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler (KL)** divergence between  $p_{\theta}^{RL}(\hat{y} | x)$  and  $p^{PT}(\hat{y} | x)$ .

# Recap: Direct Preference Optimization (DPO)



# Direct Preference Optimization (DPO)

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r(x, y_w) - r(x, y_l))]$$

**A loss function on reward functions**



**A transformation between reward functions and policies**

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)} + \beta \log Z(x)$$



**A loss function on policies**

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]$$

Reward of preferred response

Reward of dispreferred response

# Lecture Plan

- Test-Time Scaling
- Training Reasoning Models
  - Imitation Learning
  - Reinforcement Learning

# Recap: Chain-of-Thought Prompting

(a) Few-shot

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: 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 answer is 8.* ✘

(b) Few-shot-CoT

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.* ✔

(c) Zero-shot

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: The answer (arabic numerals) is

(Output) *8* ✘

(d) Zero-shot-CoT (Ours)

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.* ✔

# Model Reasoning is The Key

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. ✓*

How to encourage model to generate better reasoning chain?

# Reasoning Models

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: ...

...

GPT-5.2	GPT-4.1	o4-mini
The best model for coding and agentic tasks across industries	Smartest non-reasoning model	Fast, cost-efficient reasoning model, succeeded by GPT-5 mini
<a href="#">Learn more</a>	<a href="#">Learn more</a>	<a href="#">Learn more</a>
<a href="#">Playground</a>	<a href="#">Playground</a>	<a href="#">Playground</a>
Reasoning: ●●●●●	Intelligence: ●●●●●	Reasoning: ●●●●●
Speed: ⚡⚡⚡	Speed: ⚡⚡⚡	Speed: ⚡⚡⚡
Input: [🗨️] [📄] [📁] [📧]	Input: [🗨️] [📄] [📁] [📧]	Input: [🗨️] [📄] [📁] [📧]
Output: [🗨️] [📄] [📁] [📧]	Output: [🗨️] [📄] [📁] [📧]	Output: [🗨️] [📄] [📁] [📧]
Reasoning tokens: ✓	Reasoning tokens: ☒	Reasoning tokens: ✓

# Reasoning Models

ChatGPT o1-preview



Prove that every bounded sequence in  $\mathbb{R}$  has a convergent subsequence. Provide a detailed explanation and proof.

Thought for 9 seconds ^

## Understanding the principle

I'm exploring the Bolzano-Weierstrass theorem, which states every bounded sequence in  $\mathbb{R}$  has a convergent subsequence. This involves constructing such subsequences and utilizing nested intervals or the completeness of  $\mathbb{R}$ .

## Dividing intervals

I'm working through the proof of constructing a convergent subsequence by dividing a closed interval into two equal parts, focusing on the sequence's bounded nature and the midpoint.

## Chewing through intervals

I'm picking infinite-term intervals and splitting them equally. Each step, I identify which interval holds the next infinite terms, repeating until all terms are included.

I'm thinking through the process of creating a nested sequence of closed intervals that shrink in length, aiming to pinpoint a single intersection point  $L$ .

## Proof of the Bolzano-Weierstrass Theorem:

**Statement:** Every bounded sequence in  $\mathbb{R}$  has a convergent subsequence.

## Let's begin by understanding the problem:

Given a bounded sequence  $\{x_n\}$  in  $\mathbb{R}$ , we need to show that there exists a subsequence  $\{x_{n_k}\}$

Message ChatGPT

ChatGPT can make mistakes. Check important info.

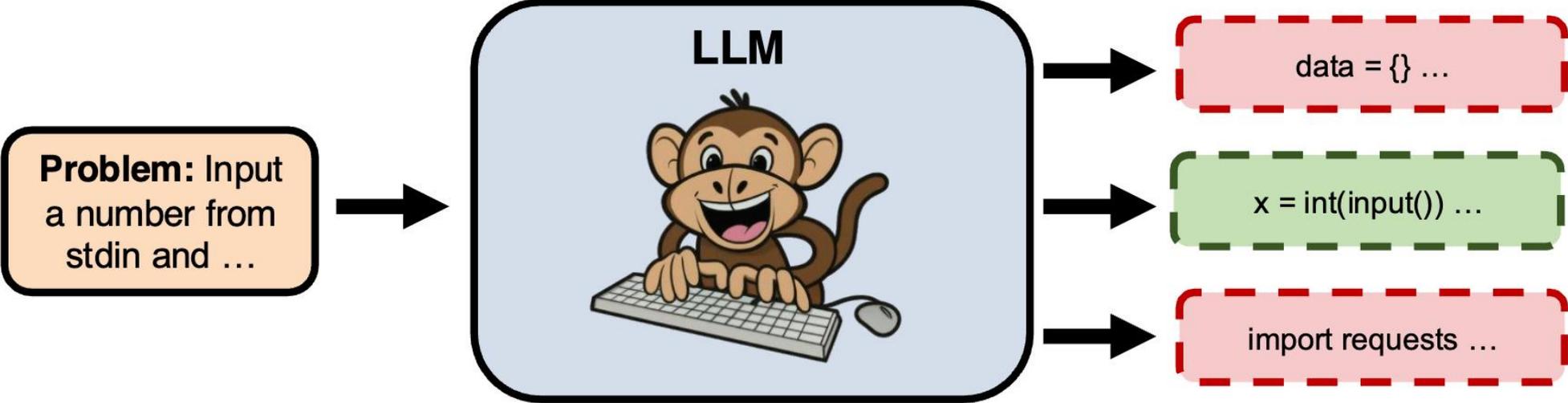
# Test-Time Scaling

- Improve model performance at **inference time**
- Allocate more **computation during testing**

# Repeated Sampling



# Repeated Sampling

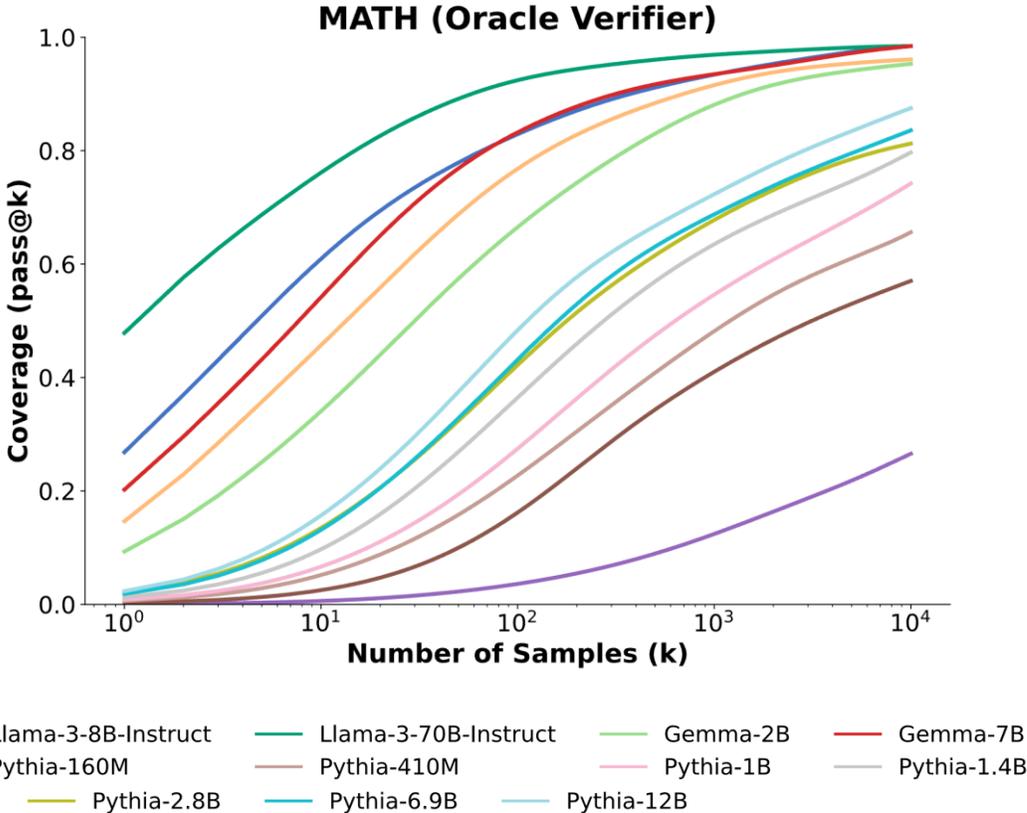


# Infinite Monkey Theorem: Large Language Monkeys

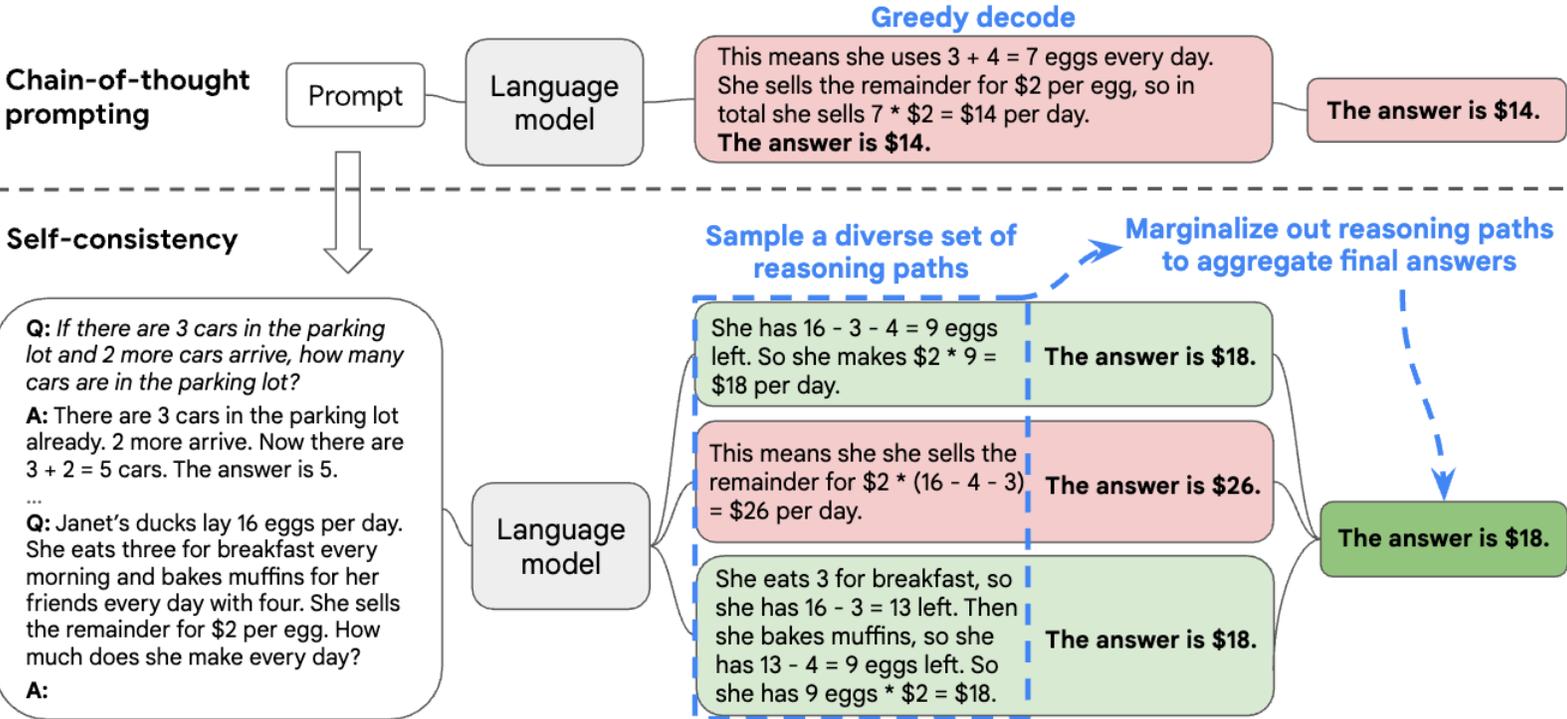
- A monkey randomly hitting keys on a typewriter for an infinite amount of time will almost surely produce any given finite text (e.g., the complete works of Shakespeare)

pass@k

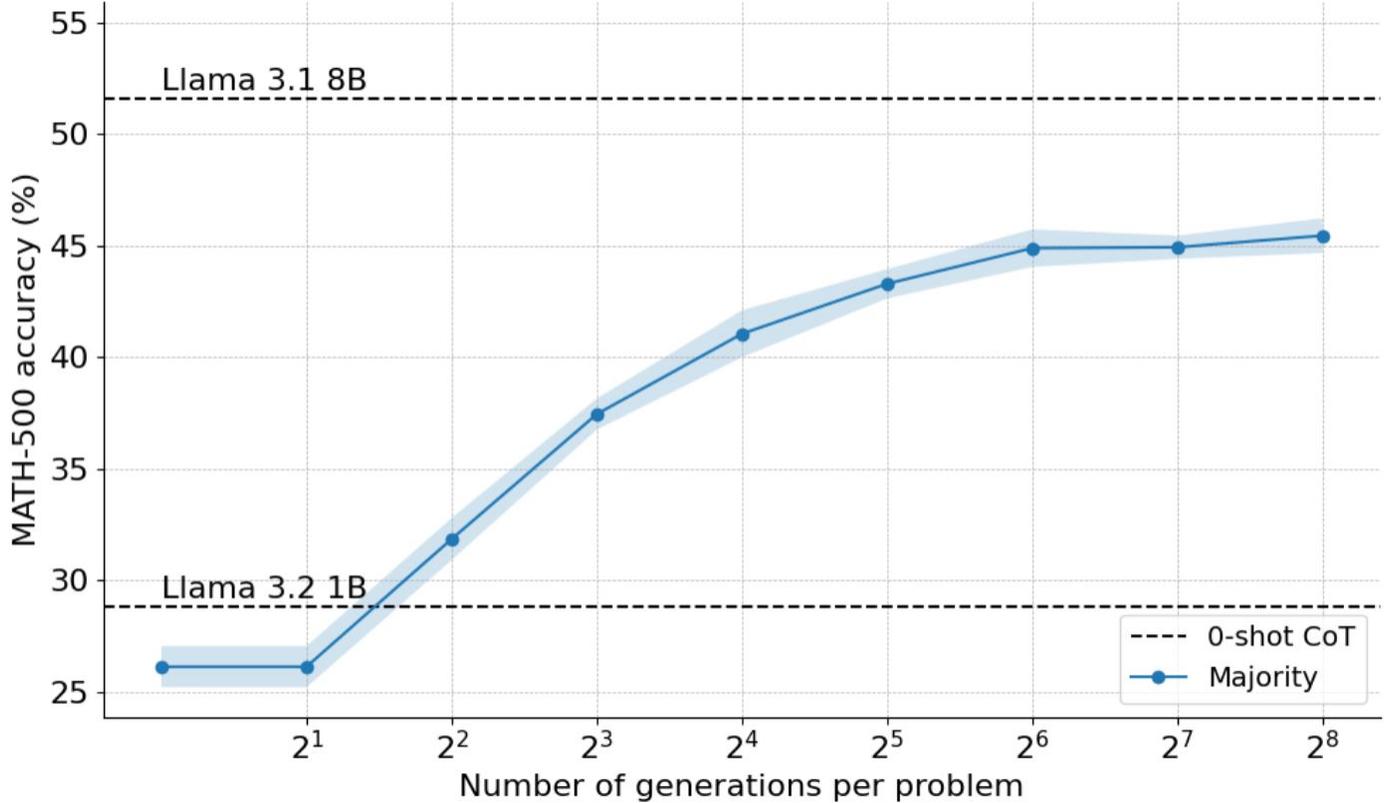
- For the same input, sample k outputs
- If 1 out of k outputs is correct → pass



# Choosing Final Answers: Majority Vote (Self-Consistency)

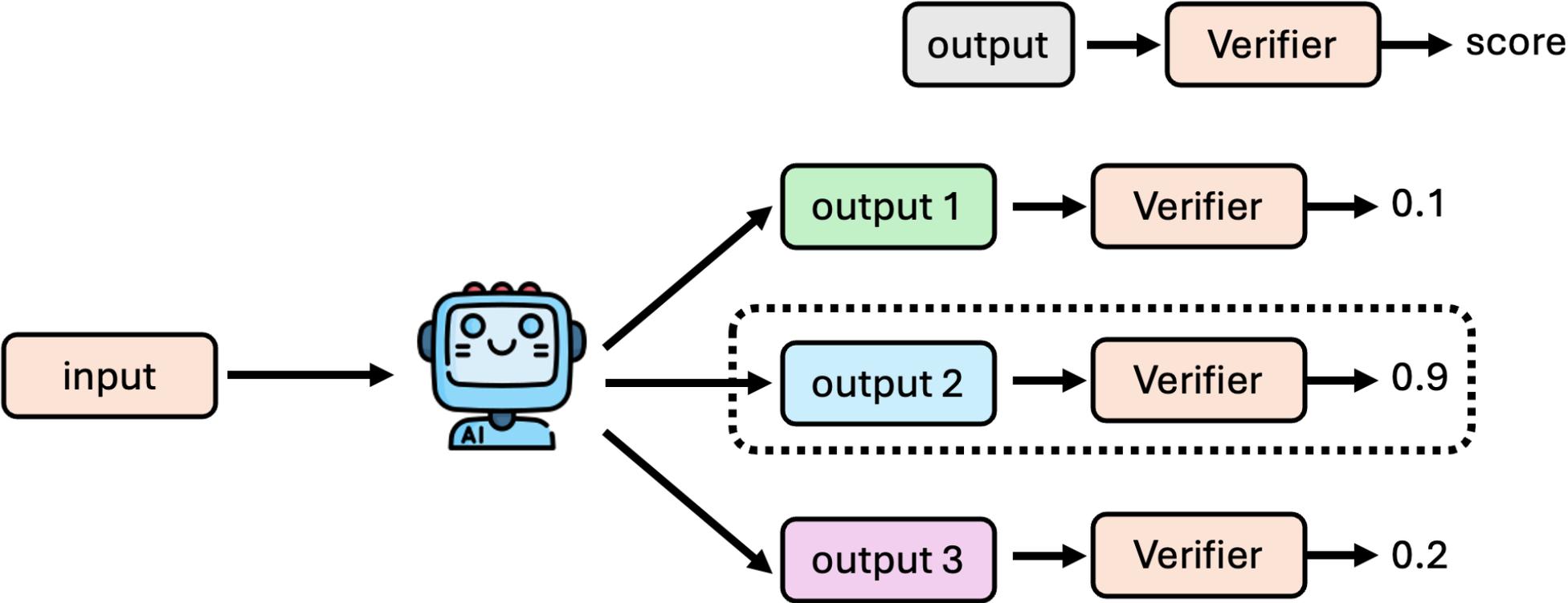


# Choosing Final Answers: Majority Vote (Self-Consistency)

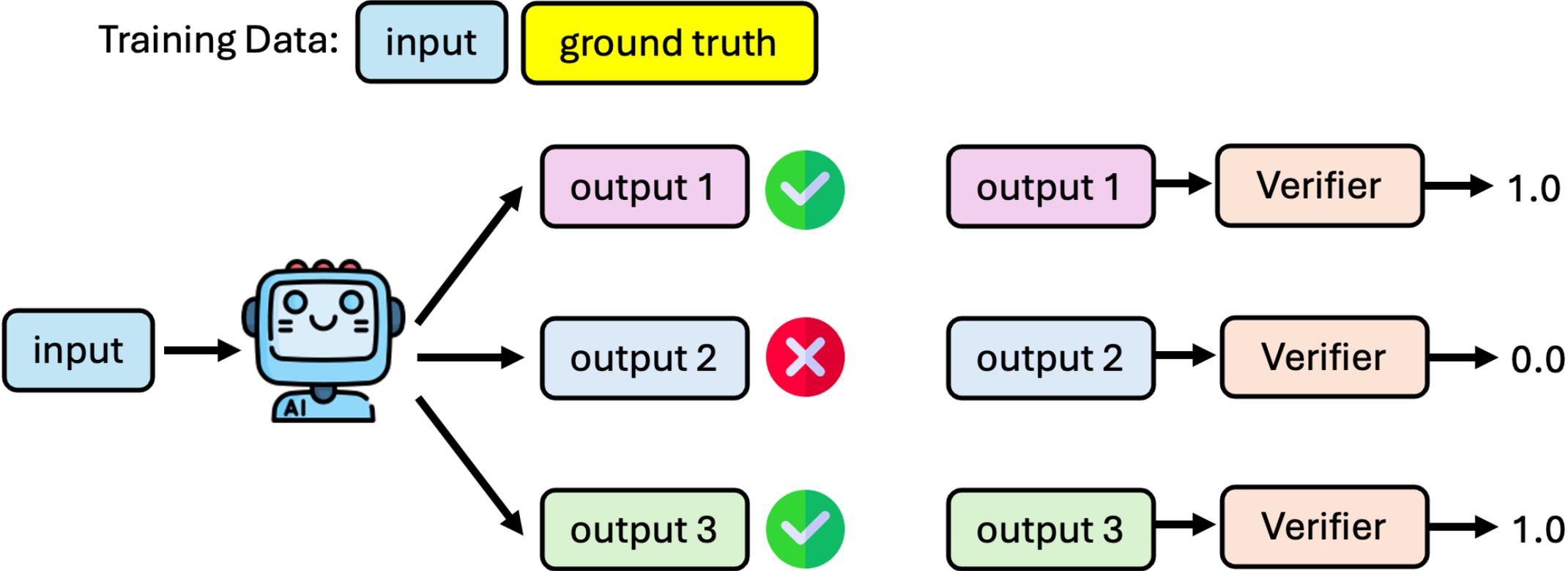


Spending more compute at testing time → Better performance

# Choosing Final Answers: Best-of-N with Verifier

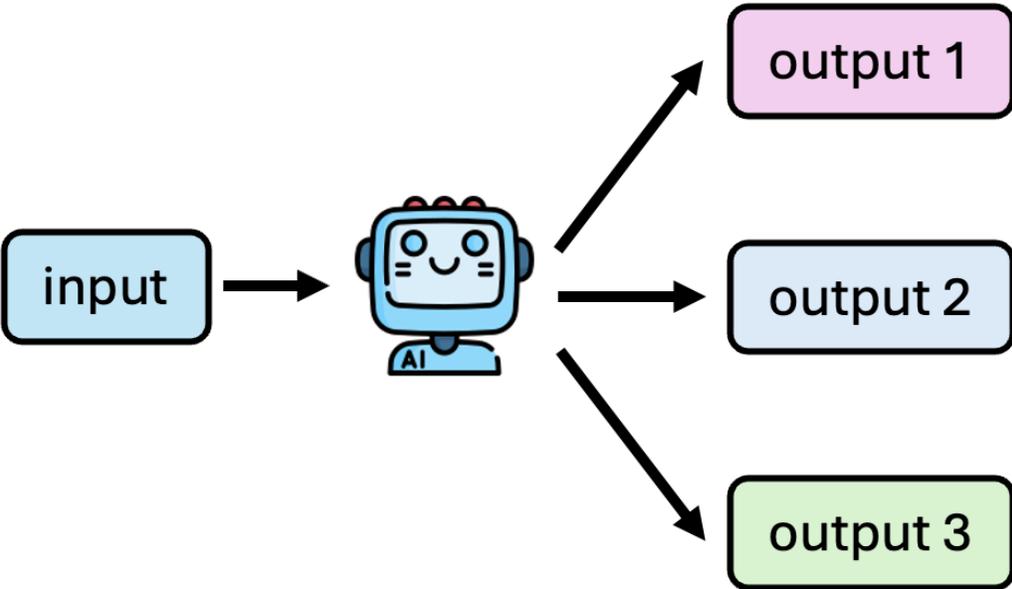


# How to Train A Verifier?

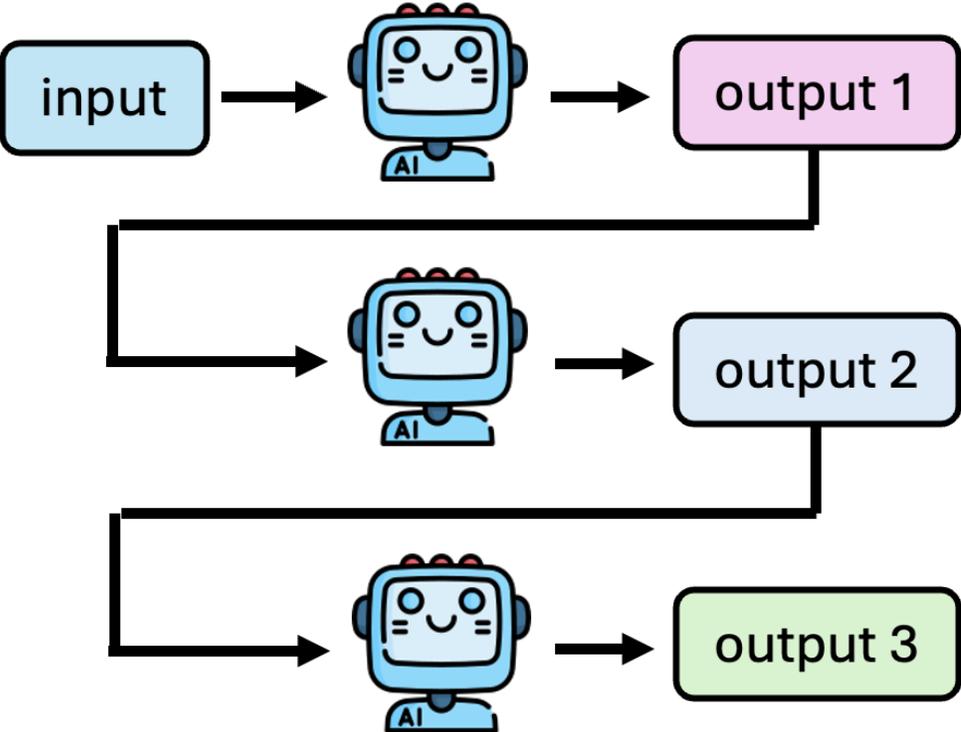


# Parallel Scaling vs. Sequential Scaling

## Parallel



## Sequential



# Sequential Scaling

- Longer reasoning chains → Better performance?
  - How to control the length of reasoning chains?

# Budget Forcing

How many r in raspberry?

**Question**

Let's break down the process of counting the letter 'r' in the word "raspberry" ...

- \* First letter: 'r' - This is an 'r', count = 1.
- \* Second letter: 'a' - Not an 'r', count remains 1 ...
- \* Sixth letter: 'e' - Not an 'r', count remains 1.
- \* Seventh letter: 'r' - This is an 'r', count = 2.
- \* Eighth letter: 'y' - Not an 'r', count remains 2 ...

The number of 'r's in "raspberry" is 2.

**Wait**, let's re-read the question carefully. It asks "How many r in raspberry?" ... \* r - a - s - p - b - e - r - r - y ... \* First 'r' ... \*

Second 'r' ... \* Third 'r' ... Count = 3 ...

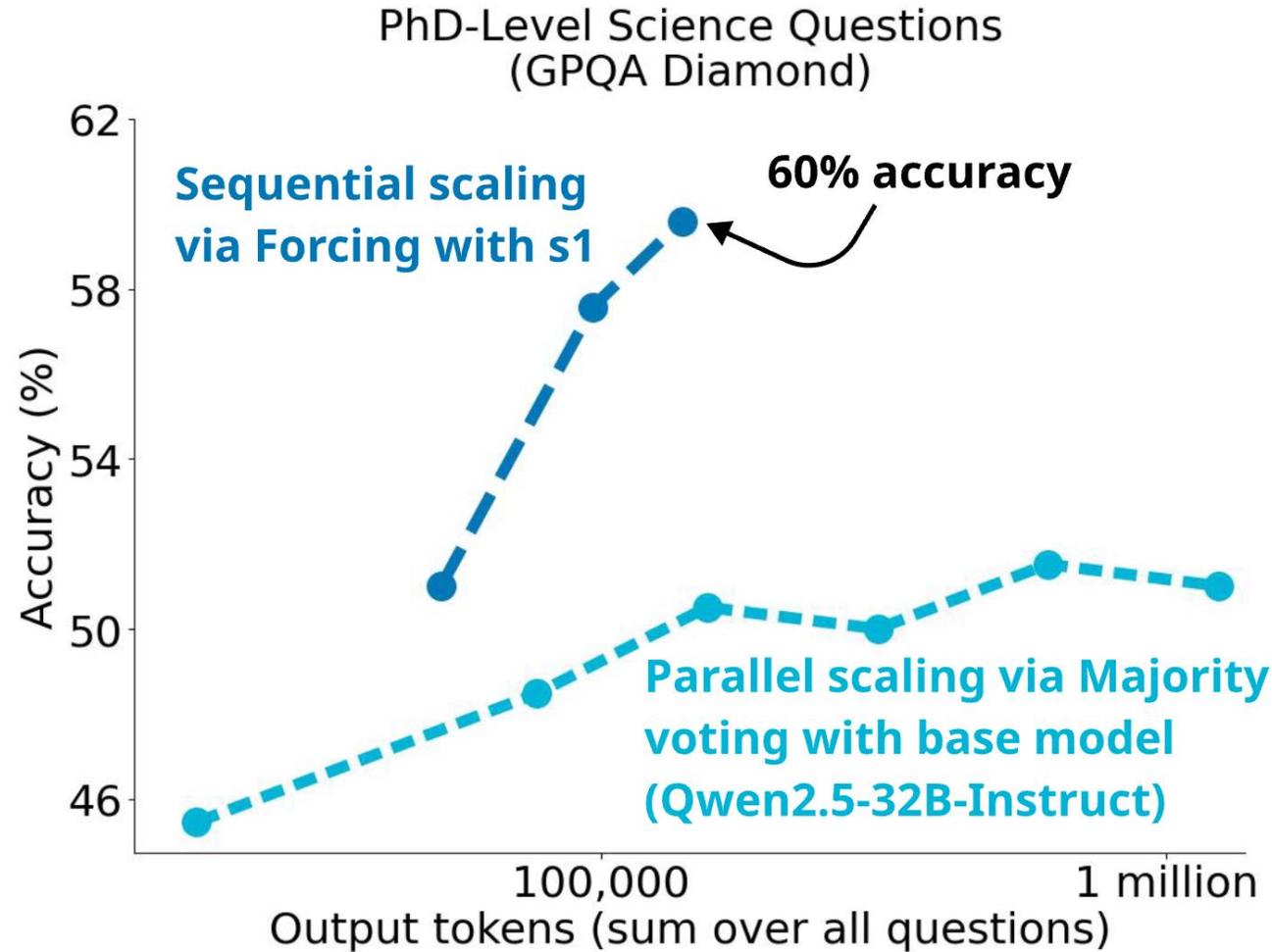
**Reasoning trace**

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is **3** **Response**

Force model to stop

Force model to keep generating

# Sequential Scaling



Spending more compute at testing time → Better performance

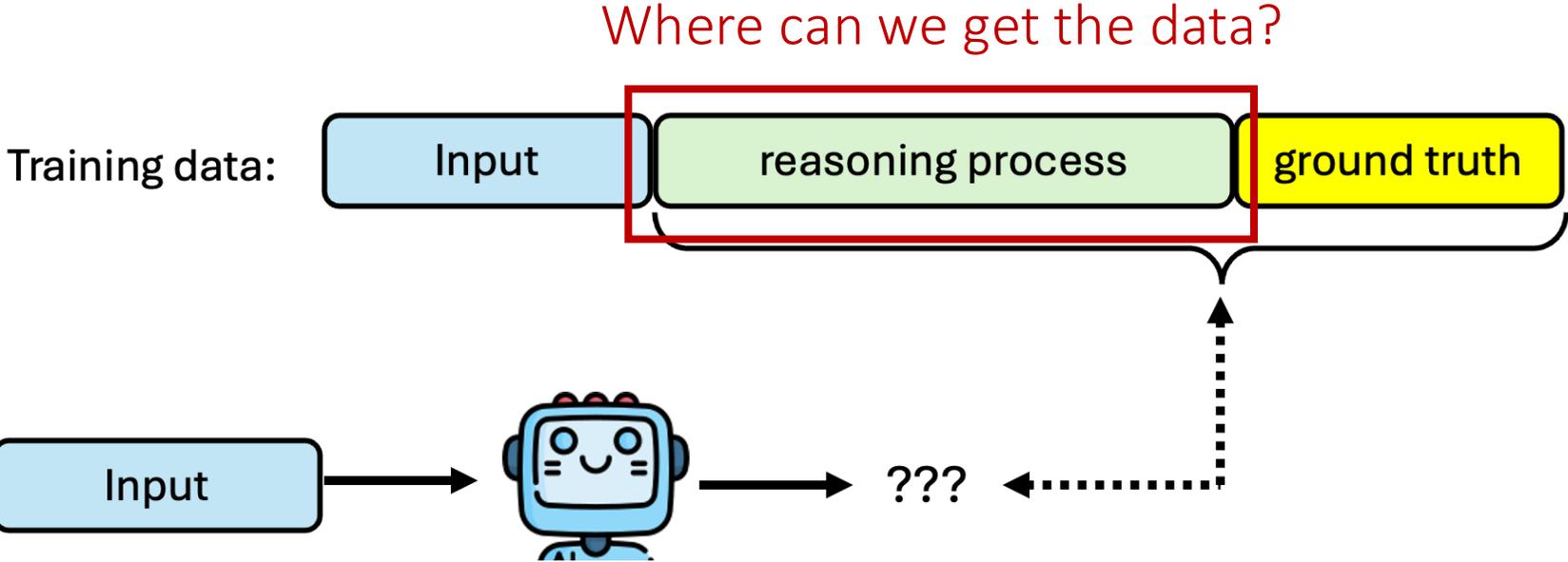
# Test-Time Scaling

- Improve model performance at **inference time**
- Allocate more **computation during testing**

# Lecture Plan

- Test-Time Scaling
- Training Reasoning Models
  - Imitation Learning
  - Reinforcement Learning

# Imitation Learning: Teach Models to Reason



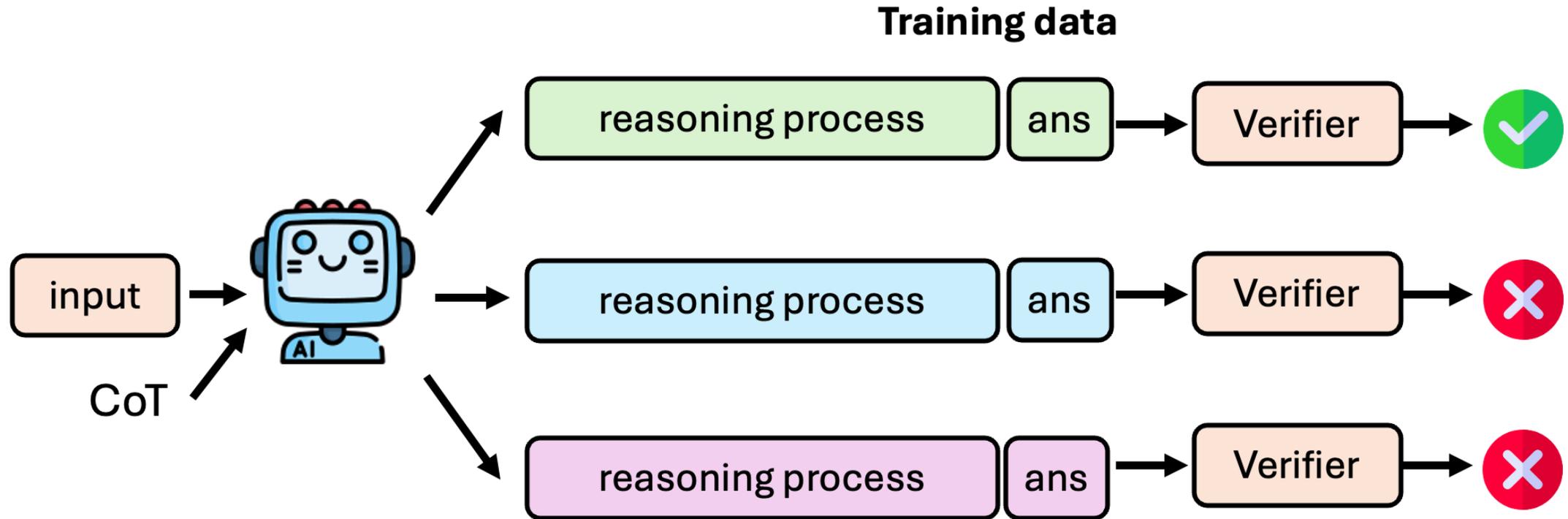
Previous full o-series reasoning model

[Learn more](#)

[Playground](#)

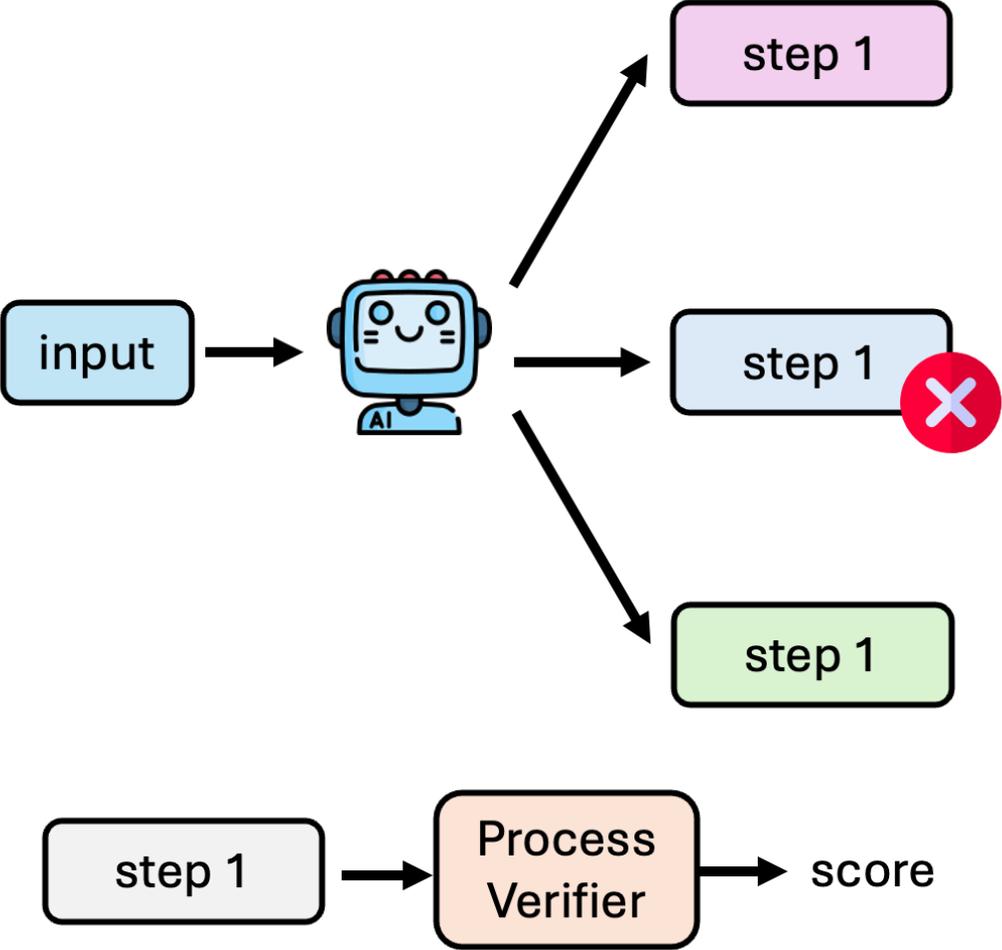
- Reasoning
- Speed
- Input
- Output
- Reasoning tokens

# How to Collect Reasoning Data?

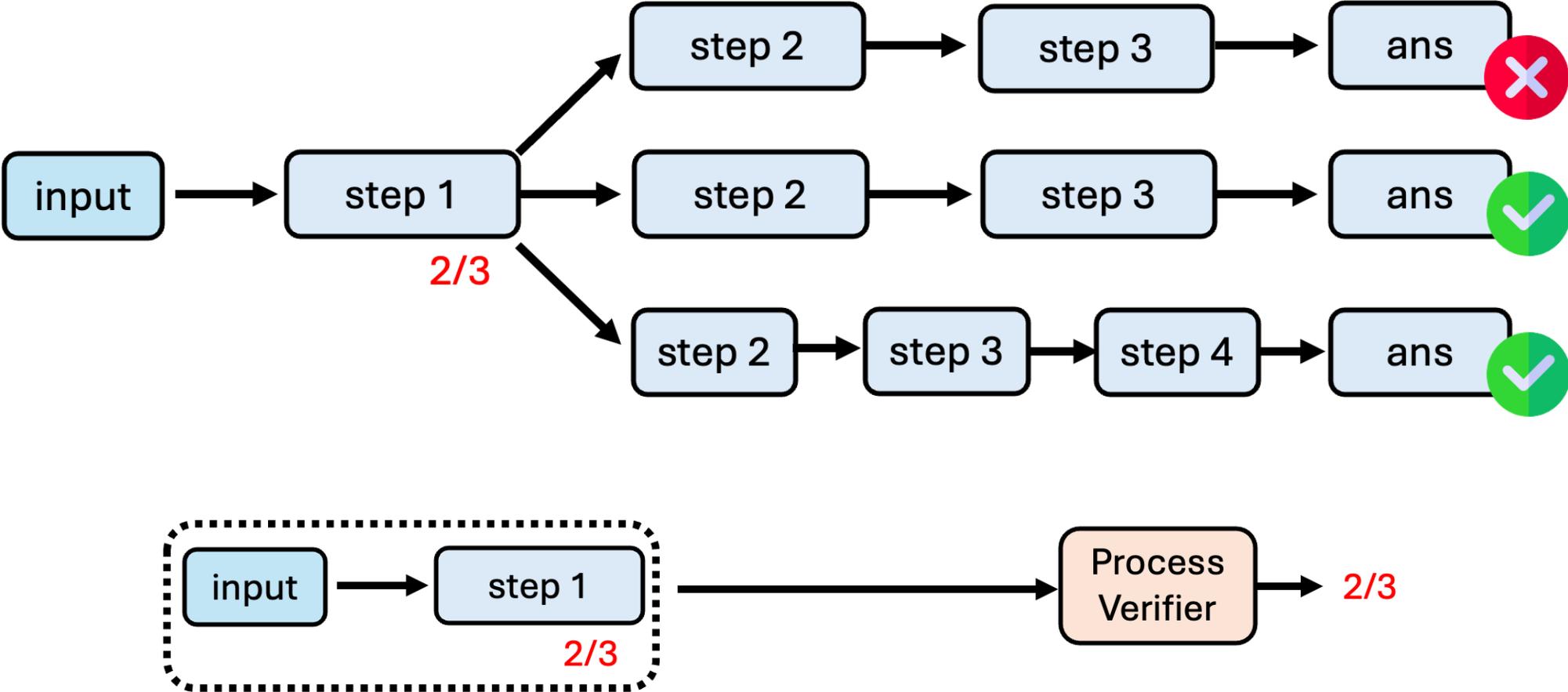


If the model is not strong enough, we have to sample many times  
Part of reasoning chain can be correct even if the final answer is wrong  
Reasoning chain can be wrong even if the final answer is correct

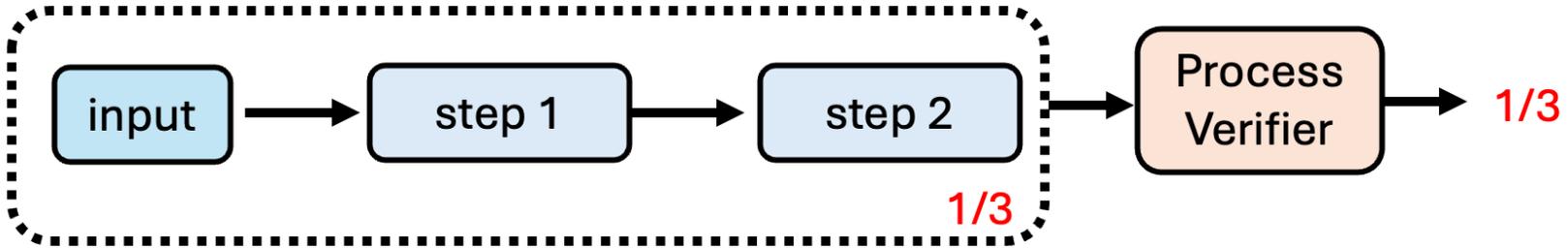
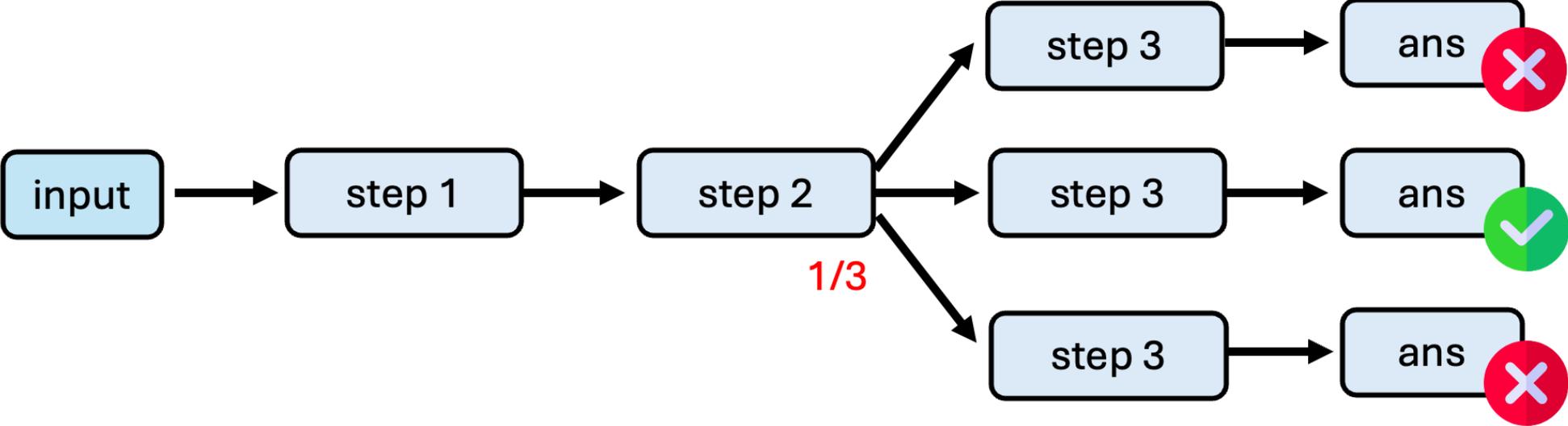
# Process Verifier



# How to Collect Process-Level Data?



# How to Collect Process-Level Data?



# Imitation Learning (Supervised Fine-Tuning)

- Simple
- Require high quality reasoning process data

Can we let the model to learn to reasoning by itself?

# Lecture Plan

- Test-Time Scaling
- Training Reasoning Models
  - Imitation Learning
  - Reinforcement Learning

# DeepSeek-R1

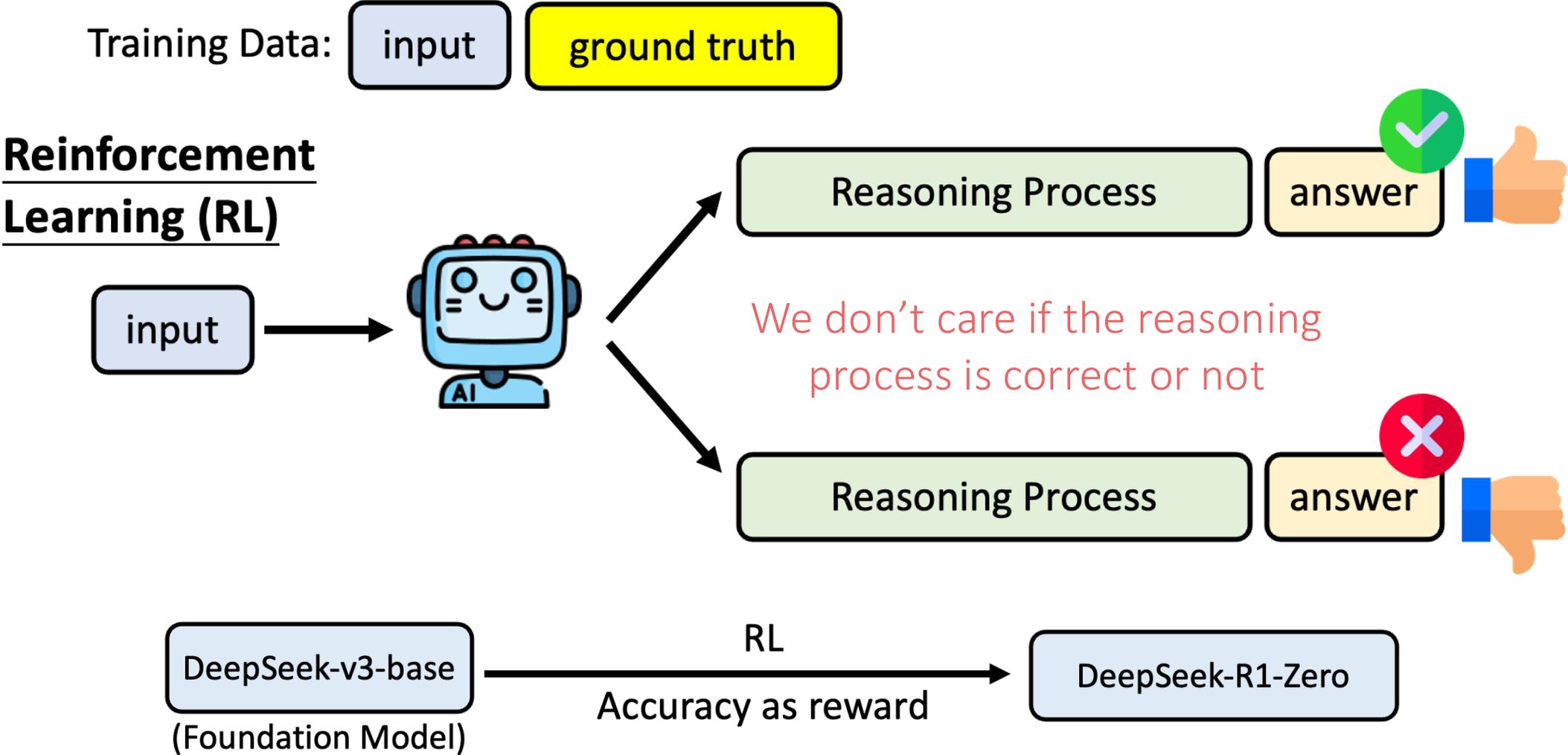


## **DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning**

DeepSeek-AI

`research@deepseek.com`

# Learning to Reason with Reinforcement Learning



# Learning to Reason with Reinforcement Learning

- Format reward for thinking token `<think>` and `</think>`

---

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:

---

```
>>> What is 5 + 5 * 10
<think>
I need to solve the expression  $(5 + 5 \times 10)$ . According to
mathematical rules, multiplication should be performed before addition.

First, I'll calculate  $(5 \times 10)$ , which equals 50.

Then, I'll add this result to 5:  $(5 + 50 = 55)$ .

So, the final answer is 55.
</think>
```

To solve the expression  $(5 + 5 \times 10)$ , we need to follow the **order of operations**, often remembered by the acronym **PEMDAS** (Parentheses, Exponents, Multiplication and Division, Addition and Subtraction).

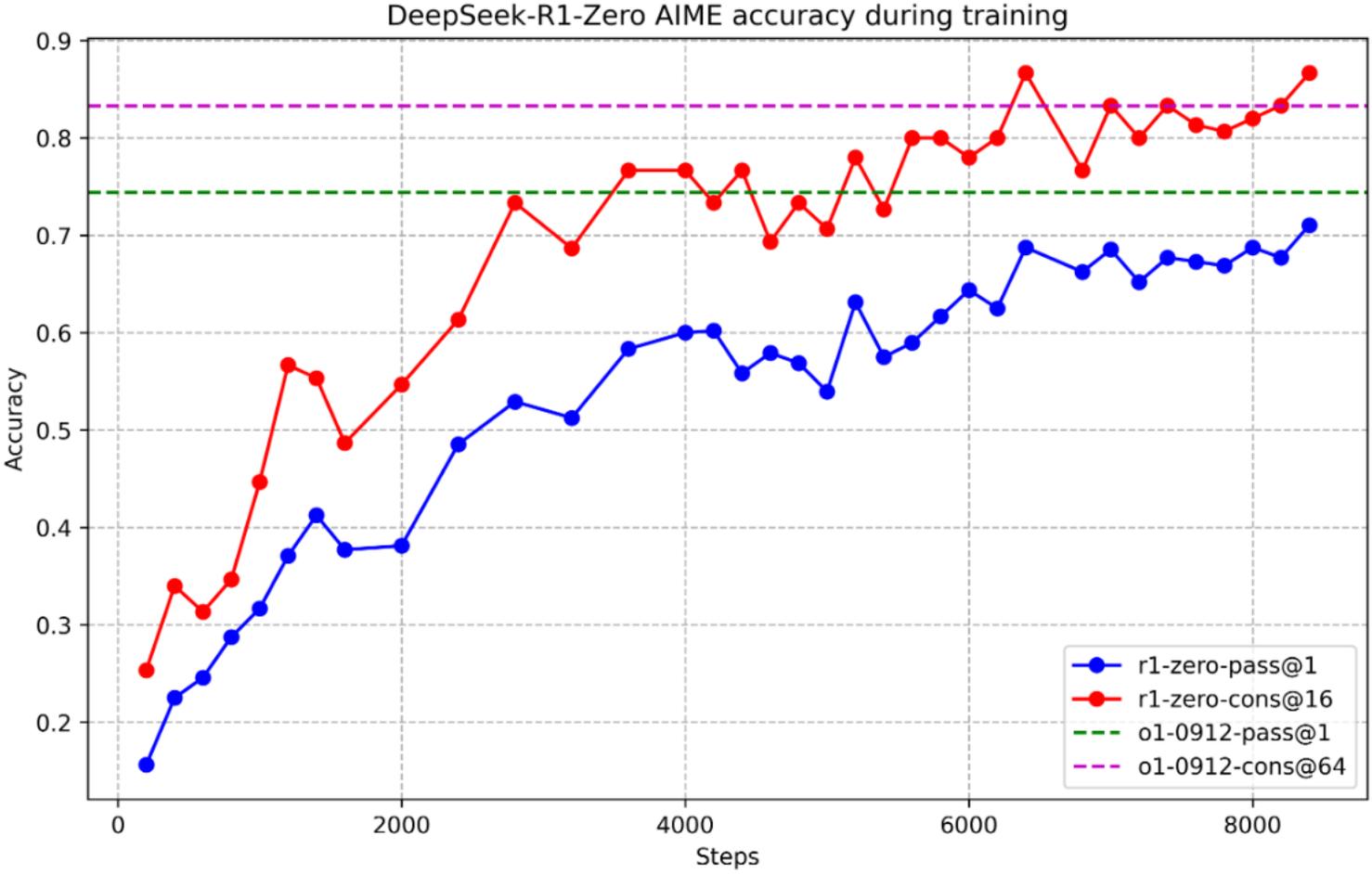
Here's a step-by-step breakdown:

- Multiplication:**  
Perform the multiplication operation first.  
 $5 \times 10 = 50$
- Addition:**  
Now, add the result of the multiplication to 5.  
 $5 + 50 = 55$

So, the final answer is:  
 $\boxed{55}$

```
>>> █ end a message (? for help)
```

# Does Reinforcement Learning Work?



# 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  $\dots$

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

Rearrange to isolate the inner square root term:

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

$\dots$

**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  $\dots$

We started with the equation:

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

First, let's square both sides:

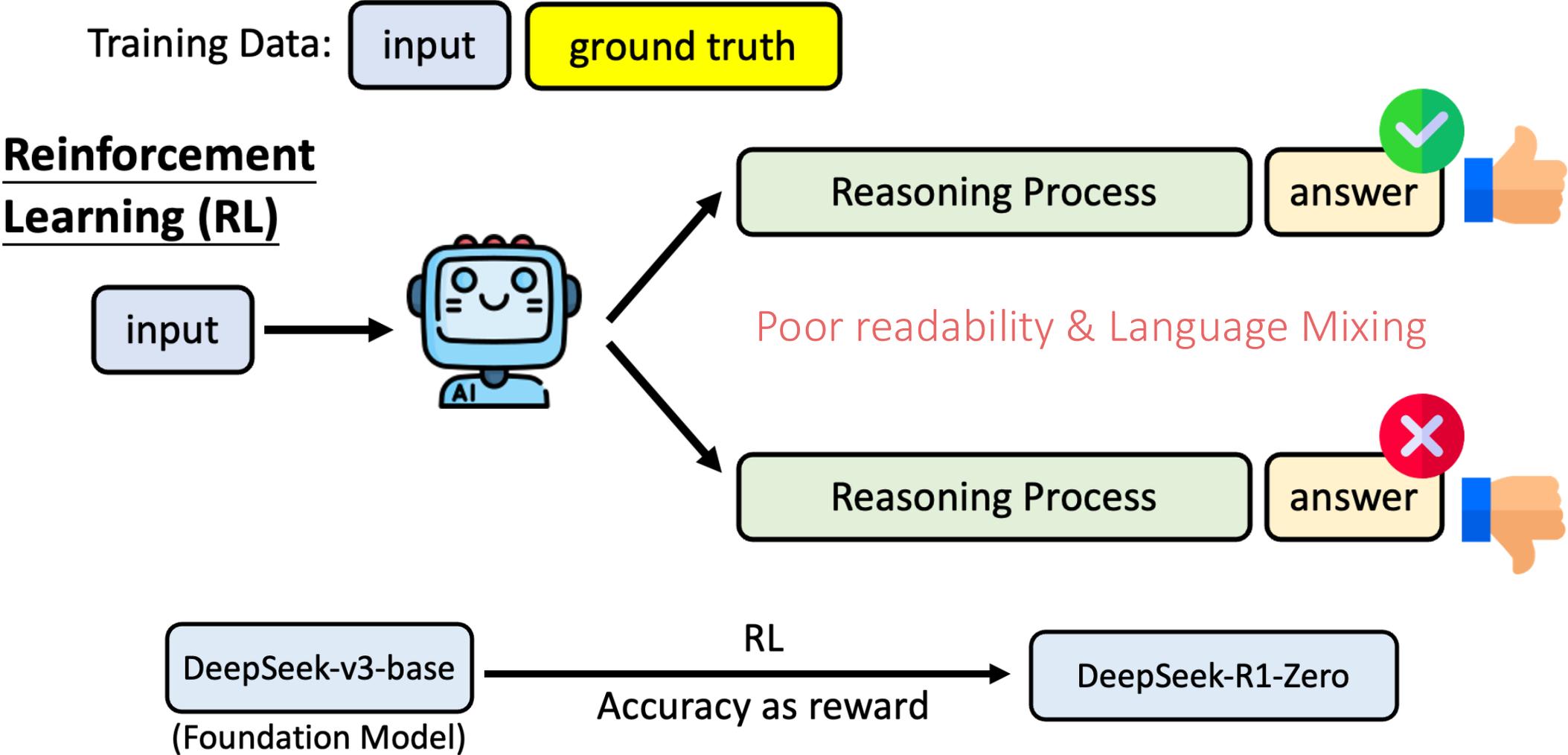
$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation:  $\dots$

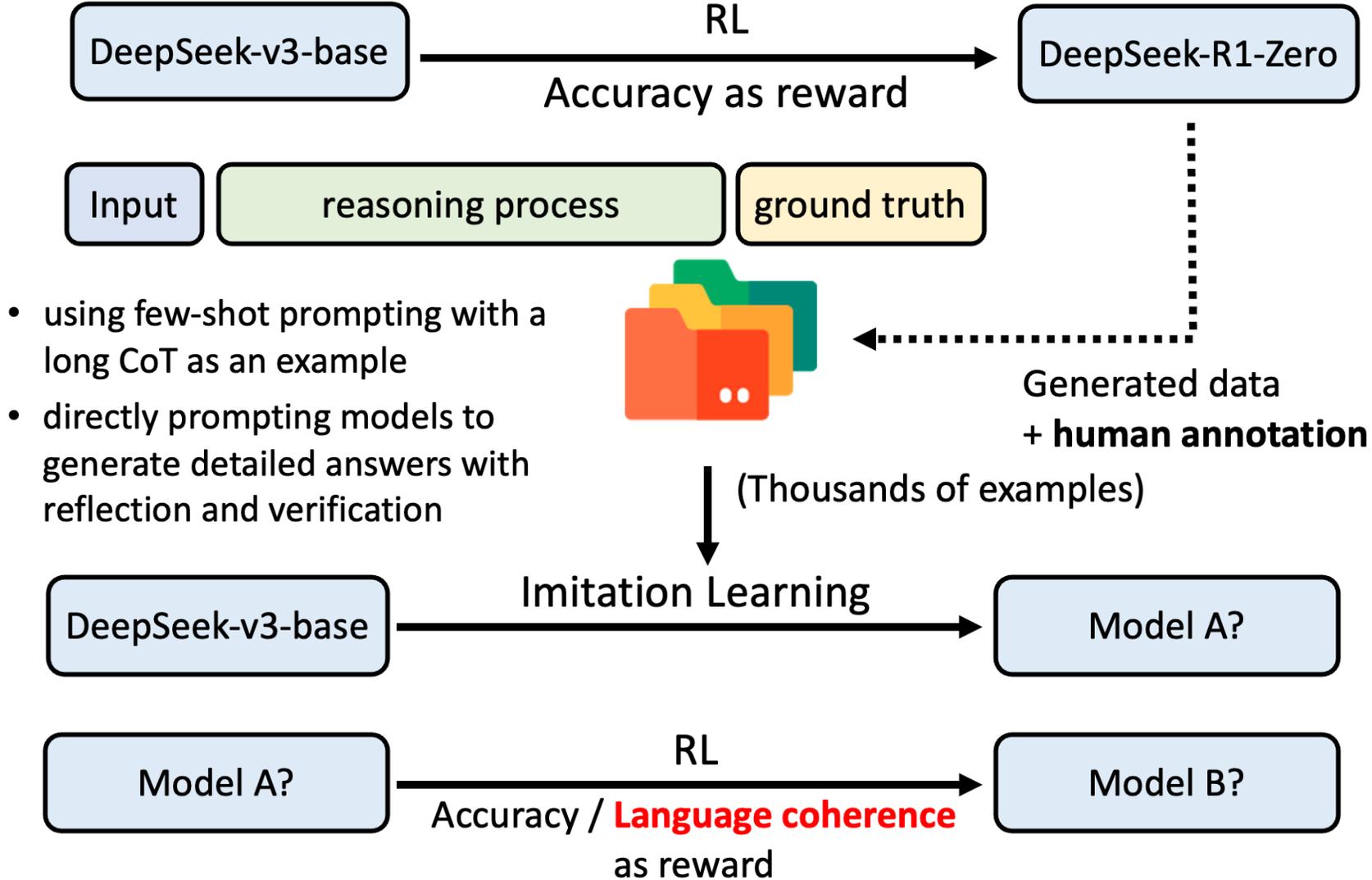
$\dots$

---

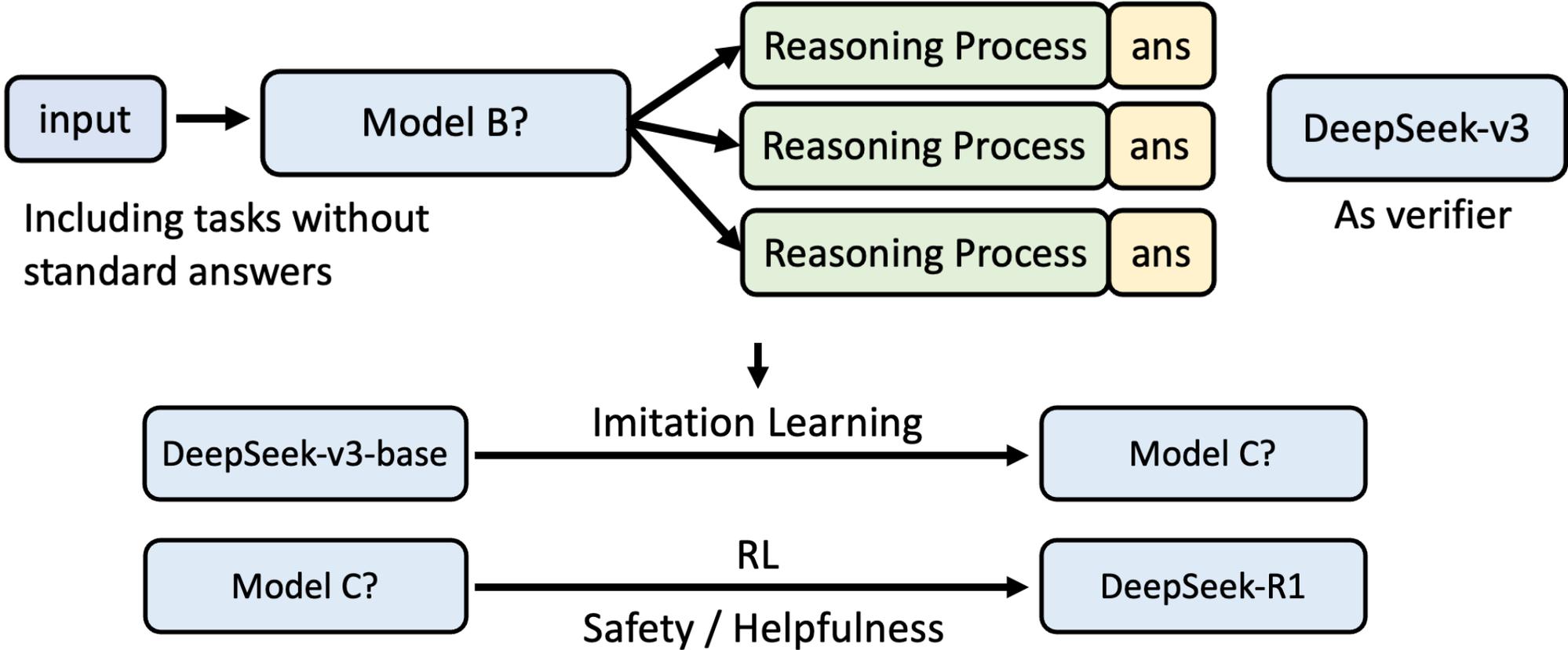
# DeepSeek-R1-Zero Can't Be Used Directly



# From DeepSeek-R1-Zero to DeepSeek-R1



# From DeepSeek-R1-Zero to DeepSeek-R1



# Real Example

solve  $x^3 + 2x^2 + 5x - 6 = 0$



Alright, let's solve the cubic equation step by step:

We are solving

$$x^3 + 2x^2 + 5x - 6 = 0$$

## 1. Check for possible rational roots

Rational root theorem: possible roots are factors of constant term  $-6$  over factors of leading coefficient  $1$ .

Possible rational roots:  $\pm 1, \pm 2, \pm 3, \pm 6$

Test  $x = 1$ :

$$1 + 2 + 5 - 6 = 2 \neq 0$$

Test  $x = -1$ :

$$-1 + 2 - 5 - 6 = -10 \neq 0$$

Test  $x = 2$ :

$$8 + 8 + 10 - 6 = 20 \neq 0$$

Test  $x = -2$ :

$$-8 + 8 - 10 - 6 = -16 \neq 0$$

Test  $x = 3$ :

$$27 + 18 + 15 - 6 = 54 \neq 0$$

Test  $x = -3$ :

$$-27 + 18 - 15 - 6 = -30 \neq 0$$

Test  $x = 6$ : huge  $\rightarrow$  skip

Test  $x = -6$ :

$$-216 + 72 - 30 - 6 = -180 \neq 0$$

No rational roots.



# Real Example

## 2. Try grouping or special factorizations

We can attempt to factor as  $(x^2 + ax + b)(x + c)$  type:

$$x^3 + 2x^2 + 5x - 6 = (x^2 + ax + b)(x + c)$$

Expand:

$$x^3 + (a + c)x^2 + (ac + b)x + bc$$

Match coefficients:

$$a + c = 2$$

$$ac + b = 5$$

$$bc = -6$$

From  $bc = -6$ , possibilities:  $b = 2, c = -3$  or  $b = -2, c = 3$ , etc.

Try  $b = 2, c = -3$ :

$$a + c = a - 3 = 2 \Rightarrow a = 5$$

$$ac + b = 5 \cdot (-3) + 2 = -15 + 2 = -13 \text{ (not 5)} \rightarrow \text{no.}$$

Try  $b = -2, c = 3$ :

$$a + 3 = 2 \Rightarrow a = -1$$

$$ac + b = (-1) \cdot 3 + (-2) = -3 - 2 = -5 \text{ (not 5)} \rightarrow \text{no.}$$

Try  $b = 3, c = -2$ :

$$a - 2 = 2 \Rightarrow a = 4$$

$$ac + b = 4 \cdot (-2) + 3 = -8 + 3 = -5 \text{ (no).}$$

Try  $b = -3, c = 2$ :

$$a + 2 = 2 \Rightarrow a = 0$$

$$ac + b = 0 + (-3) = -3 \text{ (no).}$$

Try  $b = 1, c = -6$ :

$$a - 6 = 2 \Rightarrow a = 8$$

$$ac + b = 8 \cdot (-6) + 1 = -48 + 1 = -47 \text{ (no).}$$

Try  $b = -1, c = 6$ :

$$a + 6 = 2 \Rightarrow a = -4$$

$$ac + b = (-4)(6) + (-1) = -24 - 1 = -25 \text{ (no).}$$

So no nice integer factorization.

# Real Example

## 3. Use cubic formula approach

General depressed cubic: try  $x = t - \frac{2}{3}$  since coefficient of  $x^2$  is 2,  
 $x^3 + 2x^2 + 5x - 6 = 0$ .

Let  $x = y - \frac{2}{3}$  to remove quadratic term.

---

$$x^2 = y^2 - \frac{4}{3}y + \frac{4}{9}$$
$$x^3 = y^3 - 2y^2 + \frac{4}{3}y - \frac{8}{27}$$

Now substitute into original:

$$x^3 \text{ term: } y^3 - 2y^2 + \frac{4}{3}y - \frac{8}{27}$$
$$+ 2x^2: + 2y^2 - \frac{8}{3}y + \frac{8}{9}$$
$$+ 5x: + 5y - \frac{10}{3}$$
$$- 6 \text{ constant.}$$

Sum  $y^3$  terms:  $y^3$

Sum  $y^2$  terms:  $-2y^2 + 2y^2 = 0$  ✓

Sum  $y$  terms:  $\frac{4}{3}y - \frac{8}{3}y + 5y = (4 - 8)/3 + 5 = -4/3 + 5 = (-4 + 15)/3 = 11/3$

Constants:  $-\frac{8}{27} + \frac{8}{9} - \frac{10}{3} - 6$

Get common denom 27:  $-8 + 24 - 90 - 162 = -236$

So constants =  $-\frac{236}{27}$ .

Equation:

$$y^3 + \frac{11}{3}y - \frac{236}{27} = 0$$

Multiply by 27:

$$27y^3 + 99y - 236 = 0$$

---

## 4. Now use depressed cubic solution: $t^3 + pt + q = 0$ form.

Here  $p = \frac{99}{27} = \frac{11}{3}$  (already in orig?), Wait, careful:

We have  $y^3 + \frac{11}{3}y - \frac{236}{27} = 0$ .

So  $p = 11/3$ ,  $q = -236/27$ .

---

Discriminant  $\Delta = (q/2)^2 + (p/3)^3$

$$q/2 = -\frac{236}{54} = -\frac{118}{27}$$
$$(q/2)^2 = \frac{13924}{729}$$

$$p/3 = \frac{11}{9}$$
$$(p/3)^3 = \frac{1331}{729}$$

$$\text{Sum: } \frac{13924+1331}{729} = \frac{15255}{729}$$



# Real Example

$$\frac{15255}{729} = \frac{1695}{81}?? \text{ Check: } 15255/729 = 20.928 \text{ — hmm.}$$

But  $15255 / 9 = 1695$ , /9 again = 188.333... not nice. Let's check if 1695 divisible by 3:  $1695/3 = 565$ .

$565/5=113$ . So  $15255 = 3^2 \cdot 5 \cdot 113 \cdot 9??$  better:  $15255/729 =$  divide 15255 by 9 = 1695,  $1695/81?$

$81 \cdot 20 = 1620$ , remainder 75, so  $1695/81 = 20 + 75/81 = 20 + 25/27 = (540+25)/27 = 565/27$ .

$$\text{So indeed } 15255/729 = \frac{565}{27}.$$

So discriminant  $\Delta = \frac{565}{27} > 0$  so one real root, two complex.

## 5. Real root formula

$$\text{Let } u = \sqrt[3]{-\frac{q}{2} + \sqrt{\Delta}}$$

$$v = \sqrt[3]{-\frac{q}{2} - \sqrt{\Delta}}$$

$$\text{Here } -\frac{q}{2} = \frac{118}{27}$$

$$\sqrt{\Delta} = \sqrt{\frac{565}{27}} = \frac{\sqrt{565}}{3\sqrt{3}}, \text{ better to keep symbolic.}$$

$$u^3 = \frac{118}{27} + \frac{\sqrt{565}}{3\sqrt{3}}$$

$$v^3 = \frac{118}{27} - \frac{\sqrt{565}}{3\sqrt{3}}$$

Then  $y = u + v$  is the real root.

This  $u + v$  is messy, so I'll leave the exact form as:

$$y = \sqrt[3]{\frac{118}{27} + \frac{\sqrt{565}}{3\sqrt{3}}} + \sqrt[3]{\frac{118}{27} - \frac{\sqrt{565}}{3\sqrt{3}}}$$

$$\text{Then } x = y - \frac{2}{3}.$$

But since the problem likely expects an approximate numerical answer or factor form (irrational)

with one real root,

let's approximate:

Numerically:

$$118/27 \approx 4.37037$$

$$\frac{\sqrt{565}}{3\sqrt{3}} \approx \frac{23.7707}{5.19615} \approx 4.574$$

$$\text{So } u^3 \approx 8.94437 \rightarrow u \approx 2.075$$

$$v^3 \approx -0.20363 \rightarrow v \approx -0.588$$

$$\text{So } y \approx 1.487$$

$$\text{Then } x \approx 1.487 - 0.6667 \approx 0.8203.$$

# Real Example

Check  $x = 0.82$  in original eq:

$$0.82^3 \approx 0.551,$$

$$2x^2 \approx 2(0.6724) = 1.3448,$$

$$5x \approx 4.1, \text{ sum} = 0.551 + 1.3448 + 4.1 - 6 = (6 - 6?) \text{ Actually } 0.551 + 1.3448 = 1.8958 + 4.1 = 5.9958 - 6 \approx -0.0042 \approx 0 \text{ (good).}$$

So one real root  $x \approx 0.820$ , other two complex.

---

**Final answer:**

$$x \approx 0.820$$

Exact form:

$$x = \sqrt[3]{\frac{118}{27} + \frac{\sqrt{565}}{3\sqrt{3}}} + \sqrt[3]{\frac{118}{27} - \frac{\sqrt{565}}{3\sqrt{3}}} - \frac{2}{3}$$

# Real Example

$$123 \times 456 = ?$$

## 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,  $123 \times 400$ ,  $123 \times 50$ , and  $123 \times 6$ . Then add all those results together. That should work.

## Verification (for a step)

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

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

# Real Example

## (sub)planning

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

Now the last part is  $123 \times 6$ . Let's do that.  $100 \times 6$  is 600,  $20 \times 6$  is 120, and  $3 \times 6$  is 18. Adding those:  $600 + 120 = 720$ , plus 18 is 738. So,  $123 \times 6$  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,  $123 \times 456$  should be 56,088. Hi 

# Real Example

## 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  $123 \times 5$  (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  $123 \times 4$  (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.



# Real Example

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.

# Real Example

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



# Real Example

## 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  $123 \times 456$  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

# Real Example

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**

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

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

$$100 \times 6 = 600$$

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

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

$$20 \times 6 = 120$$

⋮



# Real Example

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

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

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

Wow, that works too. So, breaking down each digit and multiplying each pair, then adding all together. That 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, and multiple methods cross-check to the same result. So, I think that's it.

# Foundation Models are Important

Model	AIME 2024		MATH	GPQA Diamond	LiveCode Bench
	pass@1	cons@64	pass@1	pass@1	pass@1
<b>QwQ-32B-Preview</b>	50.0	60.0	90.6	54.5	41.9
<b>Qwen2.5-32B-Zero</b>	47.0	60.0	91.6	55.0	40.2
<b>DeepSeek-R1-Distill-Qwen-32B</b>	<b>72.6</b>	<b>83.3</b>	<b>94.3</b>	<b>62.1</b>	<b>57.2</b>

<p><b>Question(a):</b> An element is randomly chosen from among the first 15 rows ...</p> <hr/> <p><b>Response</b>            In Pascal’s Triangle, every row starts and ends with 1, ...            ...            This can be calculated as: *awkward silence* <b>Wait, I’m overthinking. Let’s try again.</b> The number of elements in the first n rows of Pascal’s Triangle is <math>(n + 1)! / [2! * (n - 1)!]</math>.            ...</p>	<p><b>Question(b):</b> Find the least positive integer <math>n</math> such that <math>1/\sin 45^\circ \sin 46^\circ + \dots</math></p> <hr/> <p><b>Response</b>            First, I notice that the left-hand side of the equation has the form ...            Simplifying, I get ...  <b>Aha! I can use this to get</b> <math>\sin(a + 2b) = \sin((a + b) + b) = 3 \sin(a + b) \dots</math>            This looks messy, but I notice that it has the form <math>\sin a \cos b + \cos a \sin b</math> again.            This gives me an idea. Maybe I can ...            ...</p>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 13: Cases showing that DeepSeek-V3-Base already exhibits “Aha moment” even before RL tuning.

Reinforcement learning enhances the ability of foundation models, not teaching them new knowledge

# Group Relative Policy Optimization (GRPO)

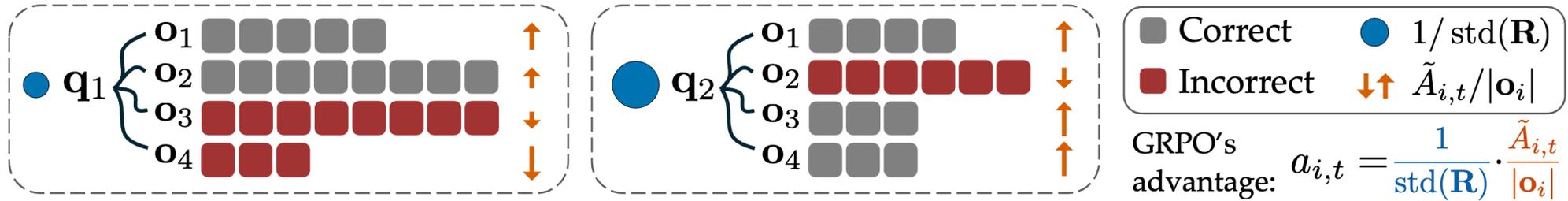
## GRPO

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[ \frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})} \hat{A}_{i,t}, \text{clip} \left( \frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] \right\},$$

$$\text{where } \hat{A}_{i,t} = \frac{R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}{\text{std}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}.$$

- Sample multiple responses for the same prompt
- Score them with a reward model
- Compute relative advantages within the group
- Update the policy to increase probability of better responses

# Dr. GRPO



## GRPO

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[ \frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})} \hat{A}_{i,t}, \text{clip} \left( \frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] \right\},$$

$$\text{where } \hat{A}_{i,t} = \frac{R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}{\text{std}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}.$$

## Dr. GRPO

GRPO Done Right (without bias)

$$\frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[ \frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})} \hat{A}_{i,t}, \text{clip} \left( \frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] \right\},$$

$$\text{where } \hat{A}_{i,t} = R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\}).$$

# Lecture Plan

- Test-Time Scaling
- Training Reasoning Models
  - Imitation Learning
  - Reinforcement Learning