

Emerging Topics in **Large Reasoning Models**



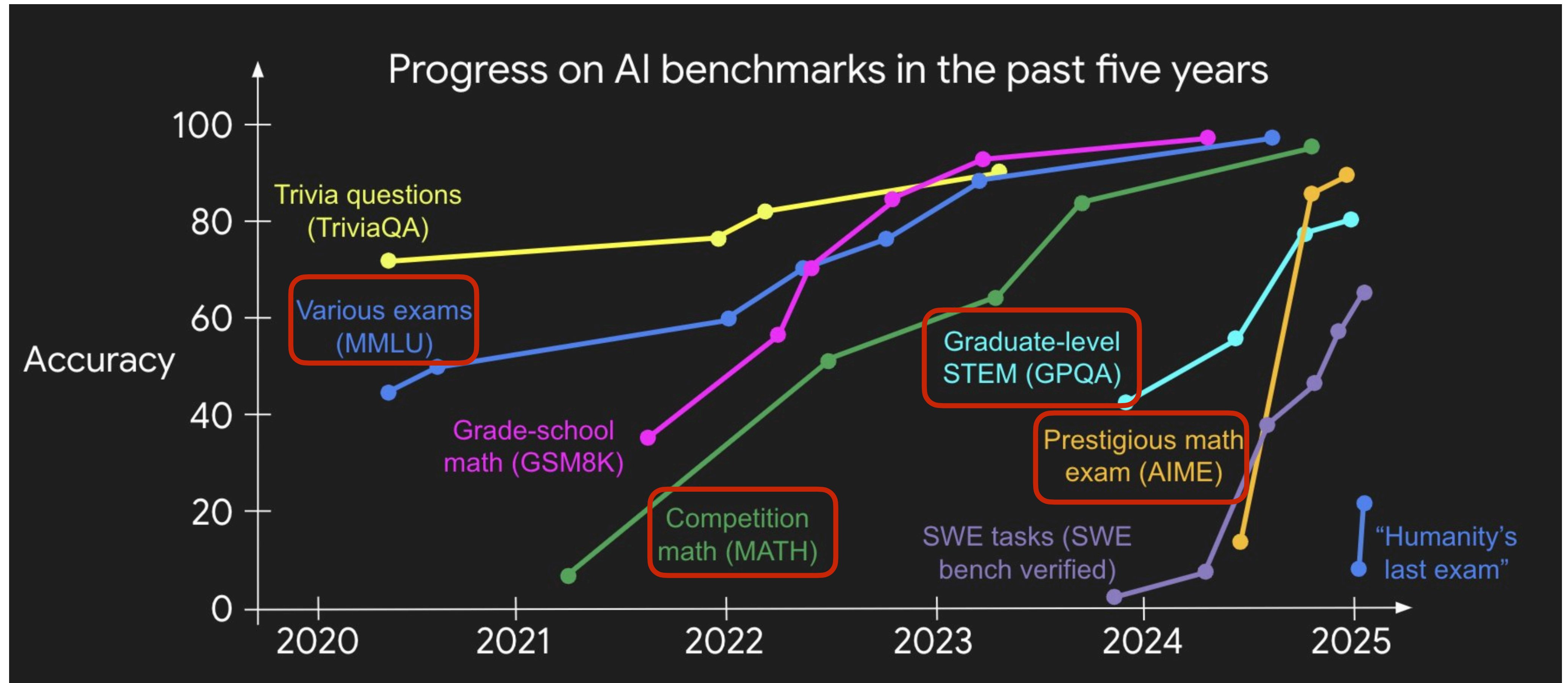
Princeton Language & Intelligence



University of Alberta

Xi Ye
April 2025

LLM Reasoning



LLM Reasoning

Prompt: Find the values of a and b that make f continuous everywhere, where f(x) is defined as $(x^2-4)/(x-2)$ for $x \neq 2$, $a \cdot x^2 - b \cdot x + 1$ if 2 is less than or equal than $x \leq 3$, $4 \cdot x - a + b$ if x is greater than or equal 3.

GPT4: To make the function f(x) continuous everywhere, we must ensure that the left and right limits for each piecewise function match at the points where they meet. In this case, we have two points to consider: $x = 2$ and $x = 3$. At $x = 2$, the first and second parts of the function meet.

So, we must ensure that their left and right limits are equal: $\lim_{x \rightarrow 2^-} (x^2 - 4) / (x - 2) = \lim_{x \rightarrow 2^+} a \cdot x^2 - b \cdot x + 1$

The left limit can be evaluated using the fact that $(x^2 - 4) / (x - 2)$ is the same as $(x + 2)$ when $x \neq 2$, since we can factor the numerator as $(x - 2)(x + 2)$: $\lim_{x \rightarrow 2^-} (x^2 - 4) / (x - 2) = \lim_{x \rightarrow 2^-} (x + 2) = 2 + 2 = 4$

Now, we have: $4 = \lim_{x \rightarrow 2^+} a \cdot x^2 - b \cdot x + 1$ Plugging in $x = 2$ into the second part of the function: $4 = a \cdot 2^2 - b \cdot 2 + 1$ Simplifying, we get: $4 = 4a - 2b + 1$ $2b - 4a = -3$ (1)

At $x = 3$, the second and third parts of the function meet. So, we must ensure that their left and right limits are equal: $\lim_{x \rightarrow 3^-} a \cdot x^2 - b \cdot x + 1 = \lim_{x \rightarrow 3^+} 4 \cdot x - a + b$

Plugging in $x = 3$ into the second and third parts of the function: $a \cdot 3^2 - b \cdot 3 + 1 = 4 \cdot 3 - a + b$ Simplifying, we get: $9a - 3b + 1 = 12 - a + b$ $10a - 4b = 11$ (2)

Now we have a system of two linear equations with two unknowns (a and b):

(1) $2b - 4a = -3$

(2) $10a - 4b = 11$

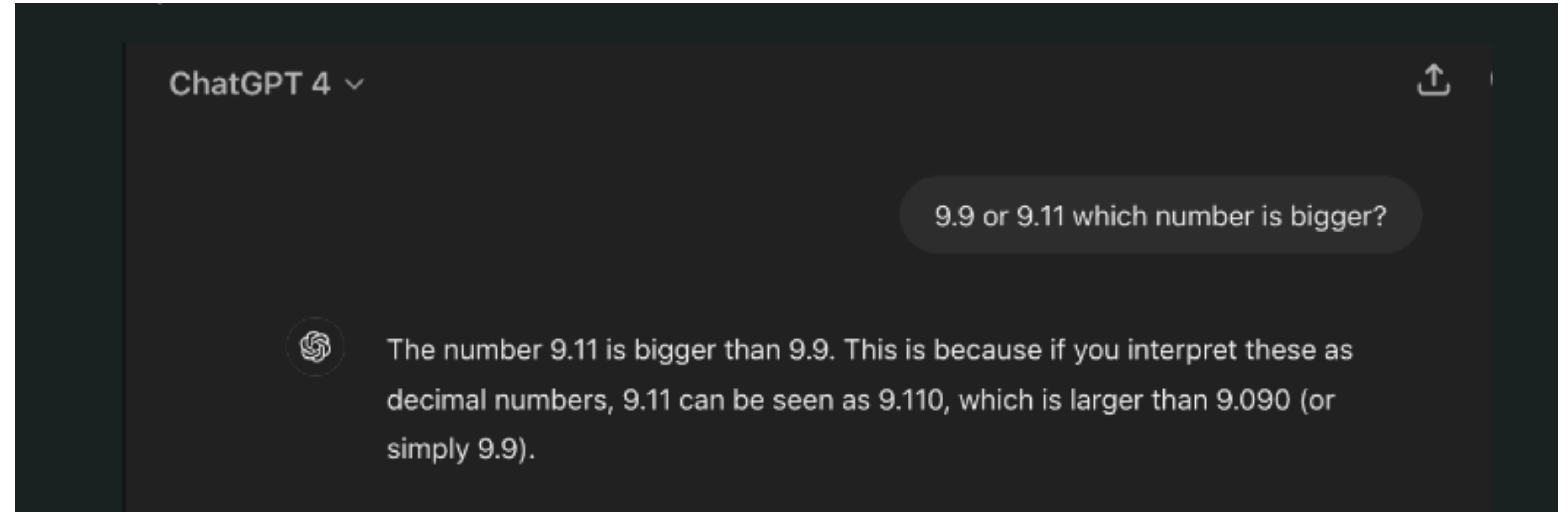
We can solve this system of equations using the substitution or elimination method. Using the elimination method, we can multiply equation (1) by 2 to match the coefficients of 'b': (1') $4b - 8a = -6$ Now, we can add equation (1') and equation (2): (1' + 2) $4b - 8a + 10a - 4b = -6 + 11$ $2a = 5$ Divide by 2: $a = 5/2$ Now, substitute the value of a back into equation (1): $2b - 4 \cdot (5/2) = -3$ $2b - 10 = -3$ Add 10 to both sides: $2b = 7$ Divide by 2: $b = 7/2$

So, the values of a and b that make f(x) continuous everywhere are $a = 5/2$ and $b = 7/2$.

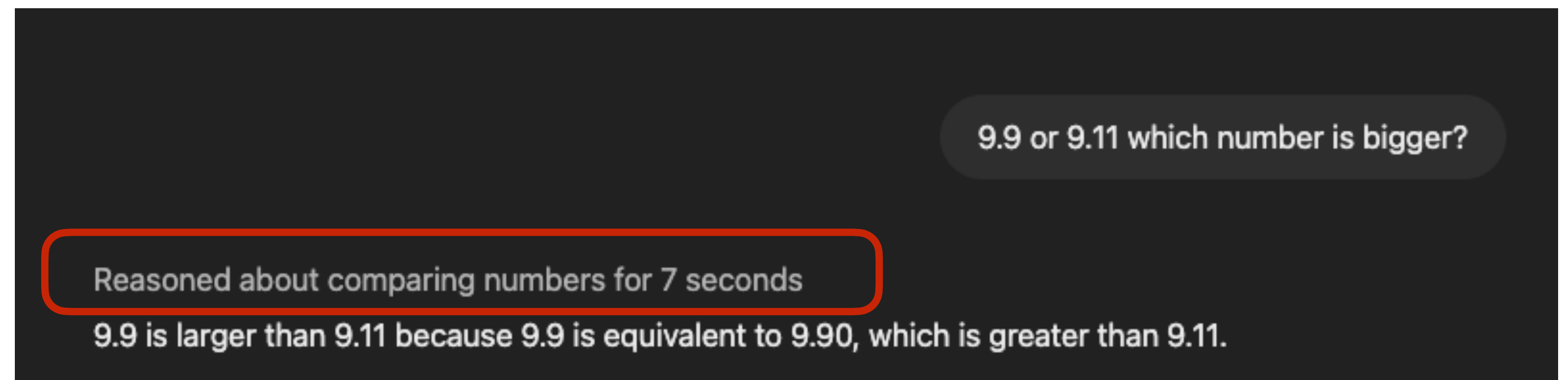
GPT-4 (July 2024) 😊

LLM Reasoning

Also GPT-4 (July 2024) 🤔



GPT-o1 (March 2025) 🤔



LLM Reasoning

DeepSeek-R1
(March 2025) 🤔

GPT-o1 (March 2025) 🤔

9.9 or 9.11 which number is bigger?



🕒 Thought for 26 seconds ^

Okay, so I need to figure out whether 9.9 or 9.11 is the bigger number. Hmm, let me start by looking at both numbers carefully. They both have a 9 before the decimal point, so the whole number part is the same. That means I need to compare the decimal parts to determine which is larger.

... simply 9.9).

... ~900 tokens

So, all methods—comparing decimal places, subtracting, converting to fractions, using a number line—confirm that 9.9 is larger than 9.11. I think that solidifies the conclusion.

To determine which number is larger between 9.9 and 9.11:

- Compare the whole number part:** Both numbers have the same whole number (9), so we focus on the decimal portions.
- Analyze the decimal places:**
 - $9.9 = 9 + 0.9$ (tenths place: 9).
 - $9.11 = 9 + 0.11$ (tenths place: 1, hundredths place: 1).
- Convert decimals to the same denominator:**
 - $0.9 = \frac{90}{100}$, $0.11 = \frac{11}{100}$.
 - $\frac{90}{100} > \frac{11}{100}$, so $0.9 > 0.11$.
- Subtraction:**
 - $9.90 - 9.11 = 0.79$ (positive result confirms $9.9 > 9.11$).
- Visualization:** On a number line, 9.9 is closer to 10, while 9.11 is near 9.1.

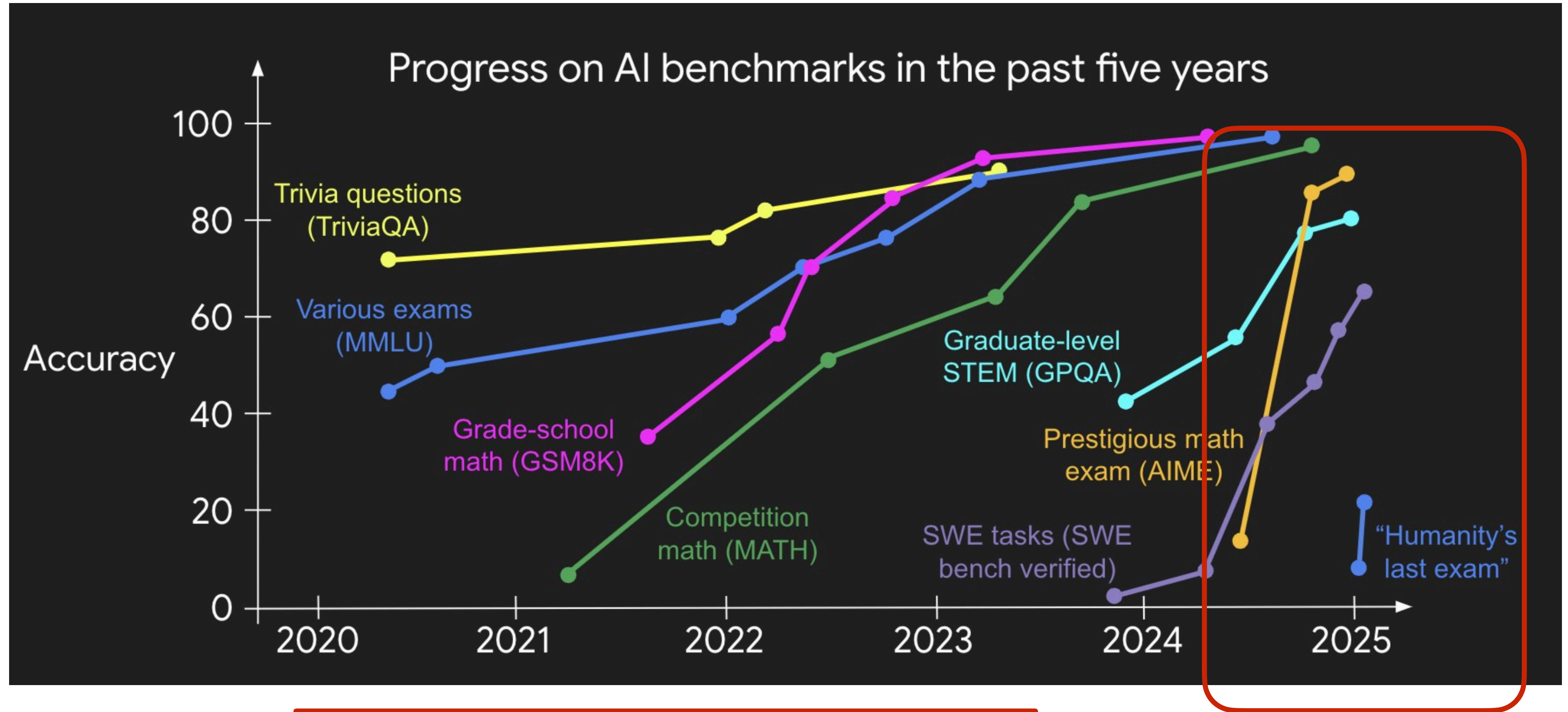
Final Answer: 9.9 is the larger number.

9.9

9.9 or 9.11 which number is bigger?

greater than 9.11.

LLM Reasoning



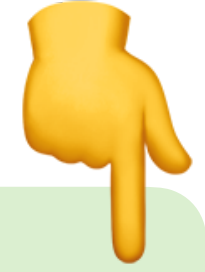
Mainly because of scaling up

Mainly because of reasoning training

Topics



Disclaimer: we will discuss many recent papers that haven't been yet peer-reviewed



What's different about large reasoning models?

- Branching and back-tracking capabilities
- Test-Time Scaling



OpenAI-O1 Blog; DeepSeek-R1 Report



How to build large reasoning models?



DeepSeek-R1 Report; scaling-test-time compute optimally; LLM monkey;



Hot takes from open-source community & research opportunities



four habits; BOLT; s1; limo; emergent reflection;

LRMs vs LLMs

DeepSeek-R1

Hey



⌘ Thought for 1 seconds ^

Okay, the user just said "Hey". That's pretty casual. I need to respond in a friendly and welcoming way. Maybe say something like, "Hey! How can I assist you today?" to invite them to ask for help. Keep it open-ended so they feel comfortable to ask anything. Make sure to maintain a positive tone. Avoid using any technical jargon or being too formal. Just a simple, approachable reply. Let me check if that makes sense. Yeah, that should work. Alright, sending that response.

Hey! How can I assist you today? 😊



DeepSeek-V3

hey



Hey! How can I assist you today? 😊



Branching and Backtracking

DeepSeek-R1

9.9 or 9.11 which number is bigger?

Wait a second, but sometimes when numbers have more decimal places, people might get confused. Let me check by writing both numbers with the same number of decimal places. If I write 9.9 as 9.90, then it's easier to compare with 9.11. Now, looking at each digit after the decimal:

- The tenths place: 9 (from 9.90) vs. 1 (from 9.11). Since 9 is greater than 1, immediately 9.90 is larger. Even though the hundredths place in 9.11 is 1, which is more than the 0 in 9.90, the tenths place is the primary determinant here. So 9.90 (which is 9.9) is definitely greater than 9.11.

Alternatively, I can subtract them to see the difference. Let's compute $9.9 - 9.11$. To do this, I'll align the decimals:

...

... ..

Another way to think about it is by converting them to fractions. 9.9 is $\frac{99}{10}$ and 9.11 is $\frac{911}{100}$. To compare $\frac{99}{10}$ and $\frac{911}{100}$, convert them to have the same denominator. Multiply $\frac{99}{10}$ by $\frac{10}{10}$ to get $\frac{990}{100}$. Now compare $\frac{990}{100}$ and $\frac{911}{100}$. Since 990 is greater than 911, $\frac{990}{100}$ is larger, so 9.9 is bigger.

I might also visualize this on a number line. Between 9.1 and 10, 9.9 is much closer to 10, whereas 9.11 is just a little past 9.1. So clearly, 9.9 is further to the right on the number line, making it the larger number.

Wait, but why did I even doubt this initially? Maybe because 9.11 has two decimal places, it's easy to miscount. For



Branching and Backtracking

"Aha Moment" (emergence of reflection behavior)
in DeepSeek-R1 tech report

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

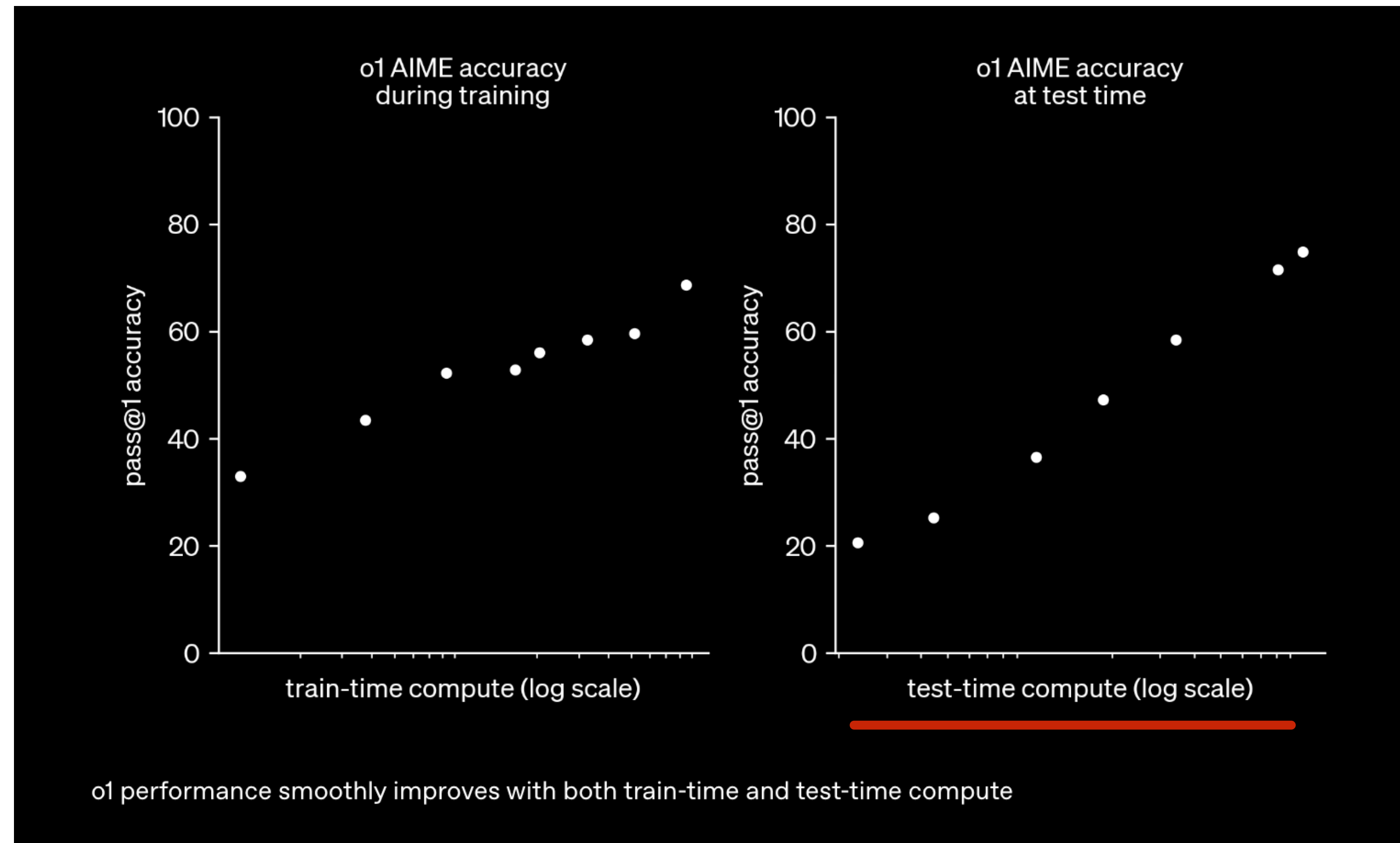
\dots

On Hendrycks-MATH. DeepSeek-R1 spends ~7000 tokens with ~33 reflections per problem on average

DeepSeek-V3 spends ~2000 tokens

Test-Time Scaling

OpenAI-O1 Blog



Scale up data or model

Spend more tokens "thinking"

Topics



What's different about large reasoning models?



- Branching and back-tracking capabilities
- Test-Time Scaling



OpenAI-O1 Blog; DeepSeek-R1 Report



How to build large reasoning models?



DeepSeek-R1 Report; scaling-test-time compute optimally; LLM monkey;



Hot takes from open-source community & research opportunities



four habits; BOLT; s1; limo;

Speculations Around O1 🍓

The Suspects



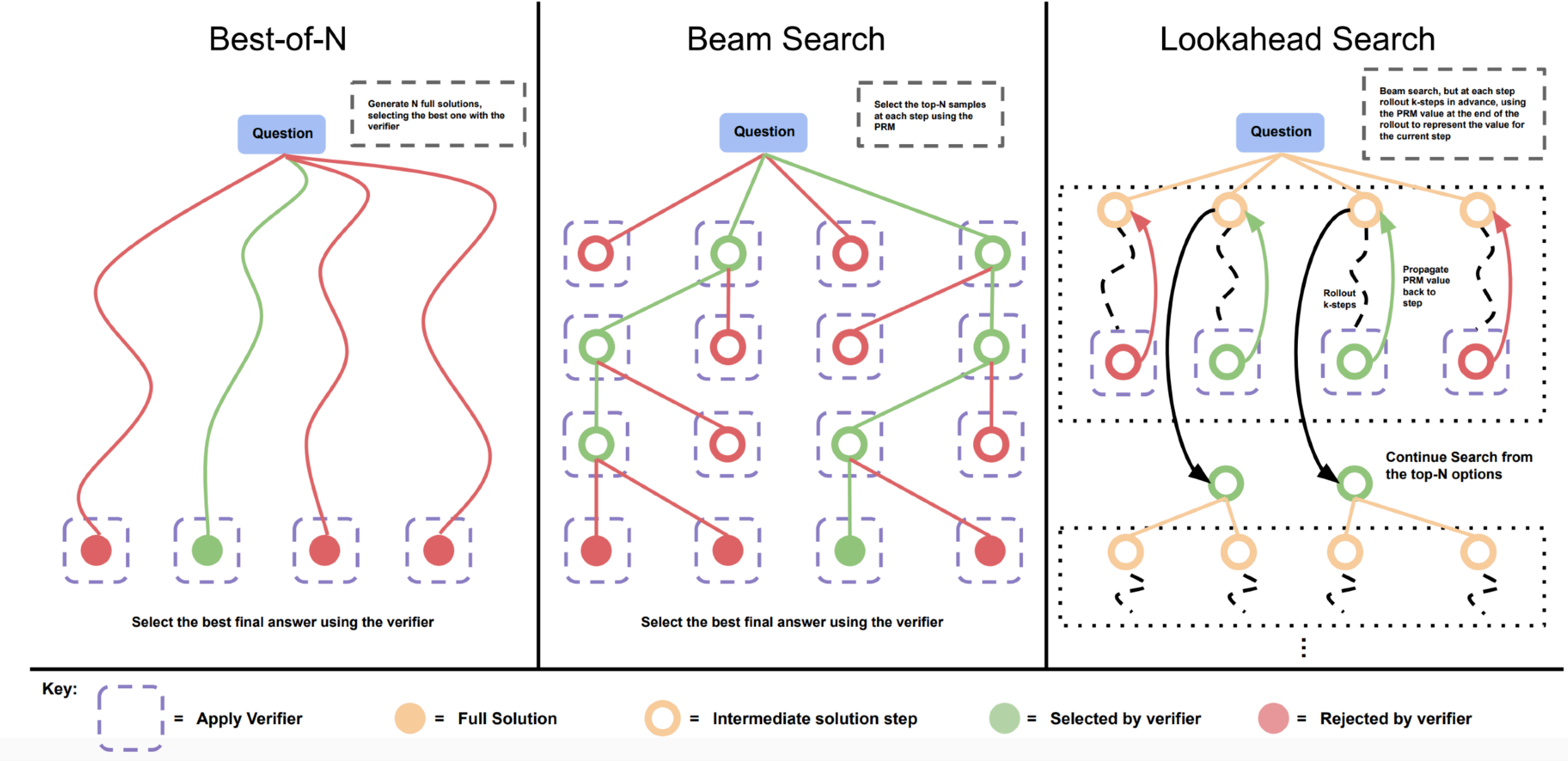
- Guess + Check
- Process Rewards
- Search / AlphaZero
- Learning to Correct

Talk by Sasha Rush: Speculations on Test-Time Scaling (o1)

Speculations Around O1🍓

The Suspects

- Guess + Check
- Process Rewards
- Search / AlphaZero
- Learning to Correct



Scaling LLM Test-Time Compute Optimally (Snell et al., 2024)

Speculations Around O1🍓

The Suspects

- Guess + Check
- Process Rewards
- Search / AlphaZero
- Learning to Correct



Not Needed



extremely simple idea

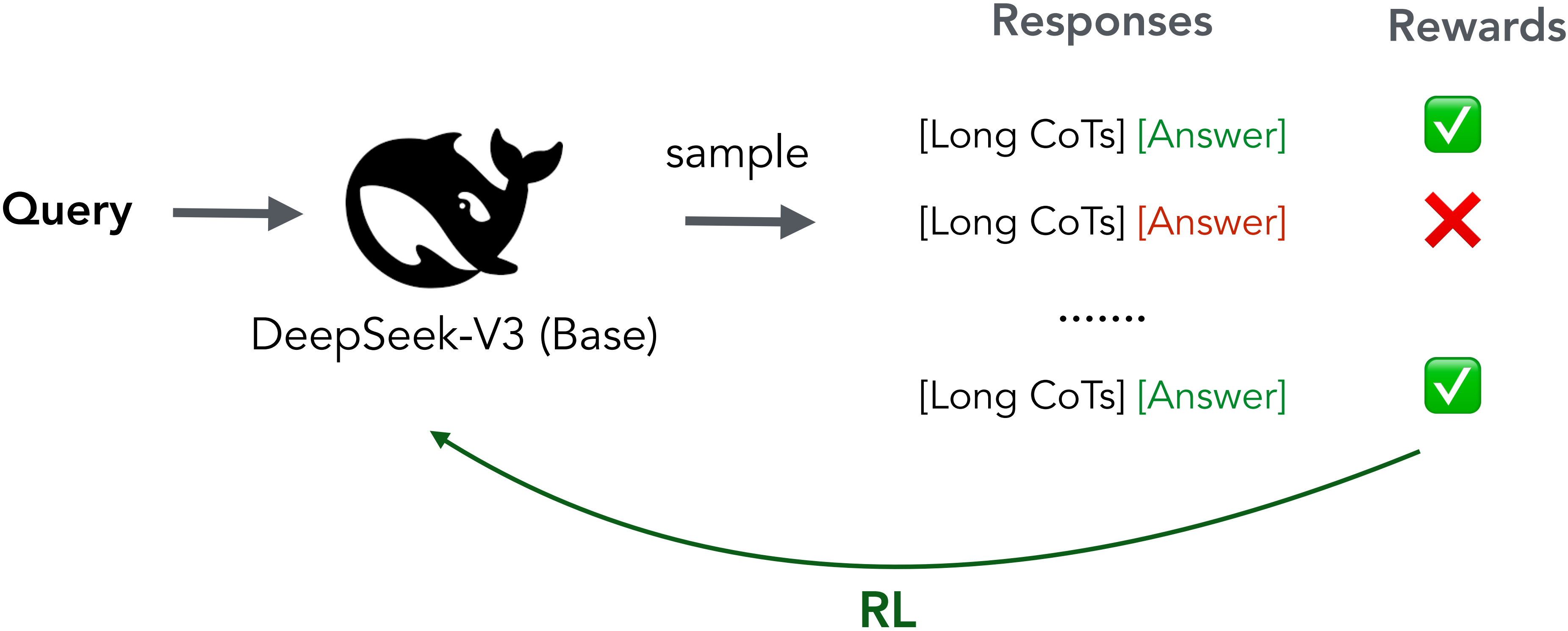
Just RL!



Rich Sutton & Andrew Barto
Turing Award Winners 2025

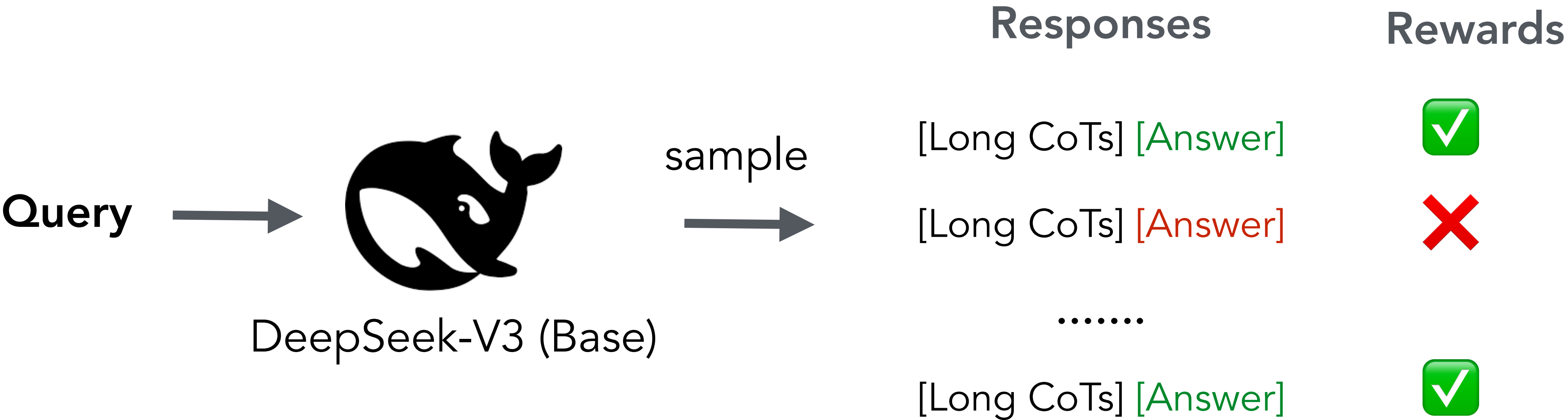
DeepSeek-R1(-Zero)

🔑 **RL** from a **base model** with **verifiable rewards**



DeepSeek-R1(-Zero)

🔑 **RL** from a **base model** with **verifiable rewards**



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:

Verifiable Rewards


 **RL** from a **base model** with **verifiable rewards**

To train DeepSeek-R1-Zero, we adopt a rule-based reward system that mainly consists of two types of rewards:

- **Accuracy rewards:** The accuracy reward model evaluates whether the response is correct. For example, in the case of math problems with deterministic results, the model is required to provide the final answer in a specified format (e.g., within a box), enabling reliable rule-based verification of correctness. Similarly, for LeetCode problems, a compiler can be used to generate feedback based on predefined test cases.
- **Format rewards:** In addition to the accuracy reward model, we employ a format reward model that enforces the model to put its thinking process between '`<think>`' and '`</think>`' tags.

TLDR: answer correctness on MATH;
functionality correctness on CODE; Format following

GRPO: Group Relative Policy Optimization

GRPO  RL from a **base model** with **verifiable rewards**

$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$ **Sample G outputs for each query q**

$$\frac{1}{G} \sum_{i=1}^G \left(\underbrace{\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right)}_{\text{maximize advantages} \quad \text{clip to control variance}} - \underbrace{\beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref})}_{\text{KL penalty}} \right), \quad (1)$$

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

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

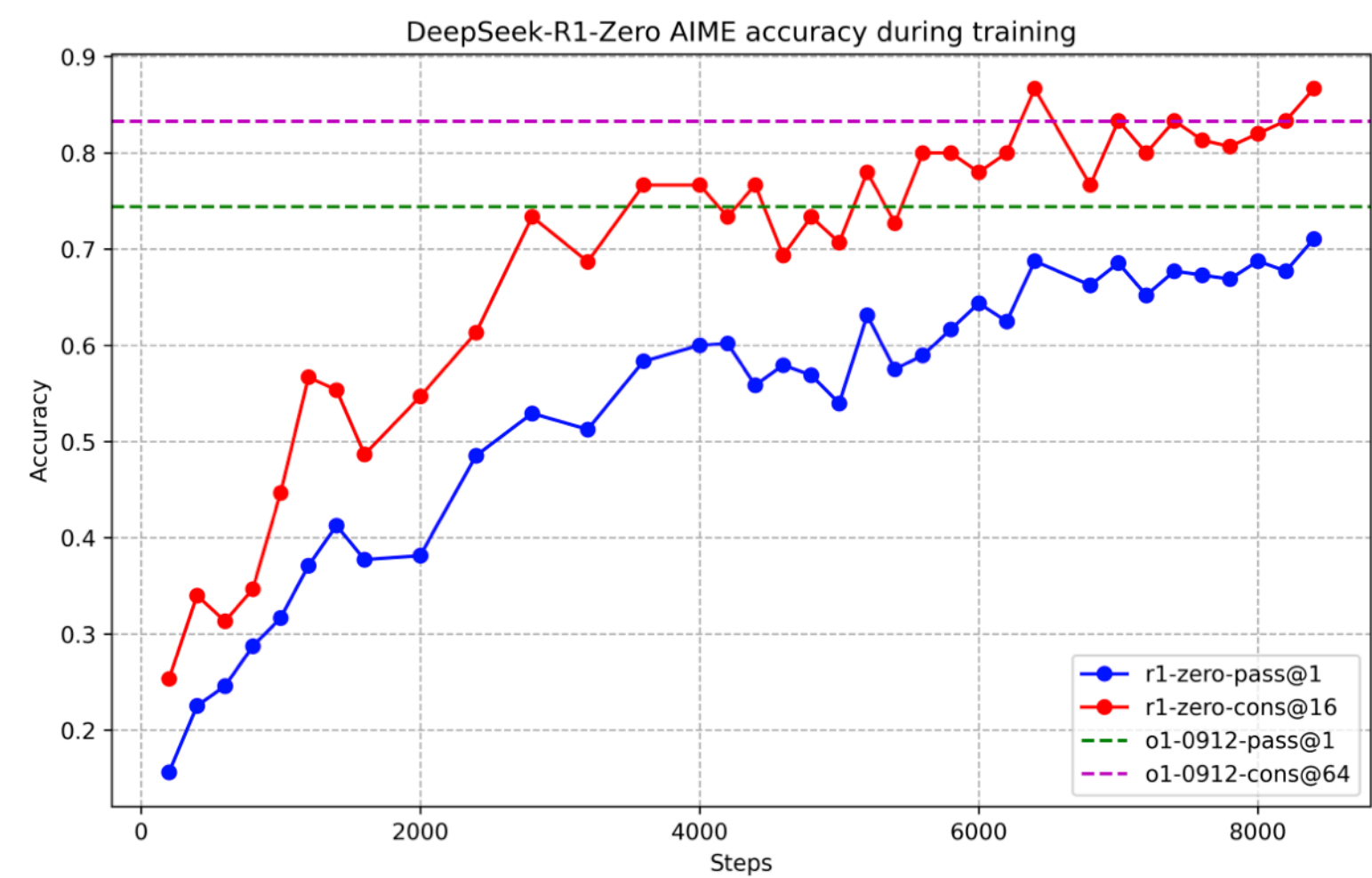
$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$

Advantages over PPO: no needs for another critic model; value approximation method based on Monte Carlo advantage

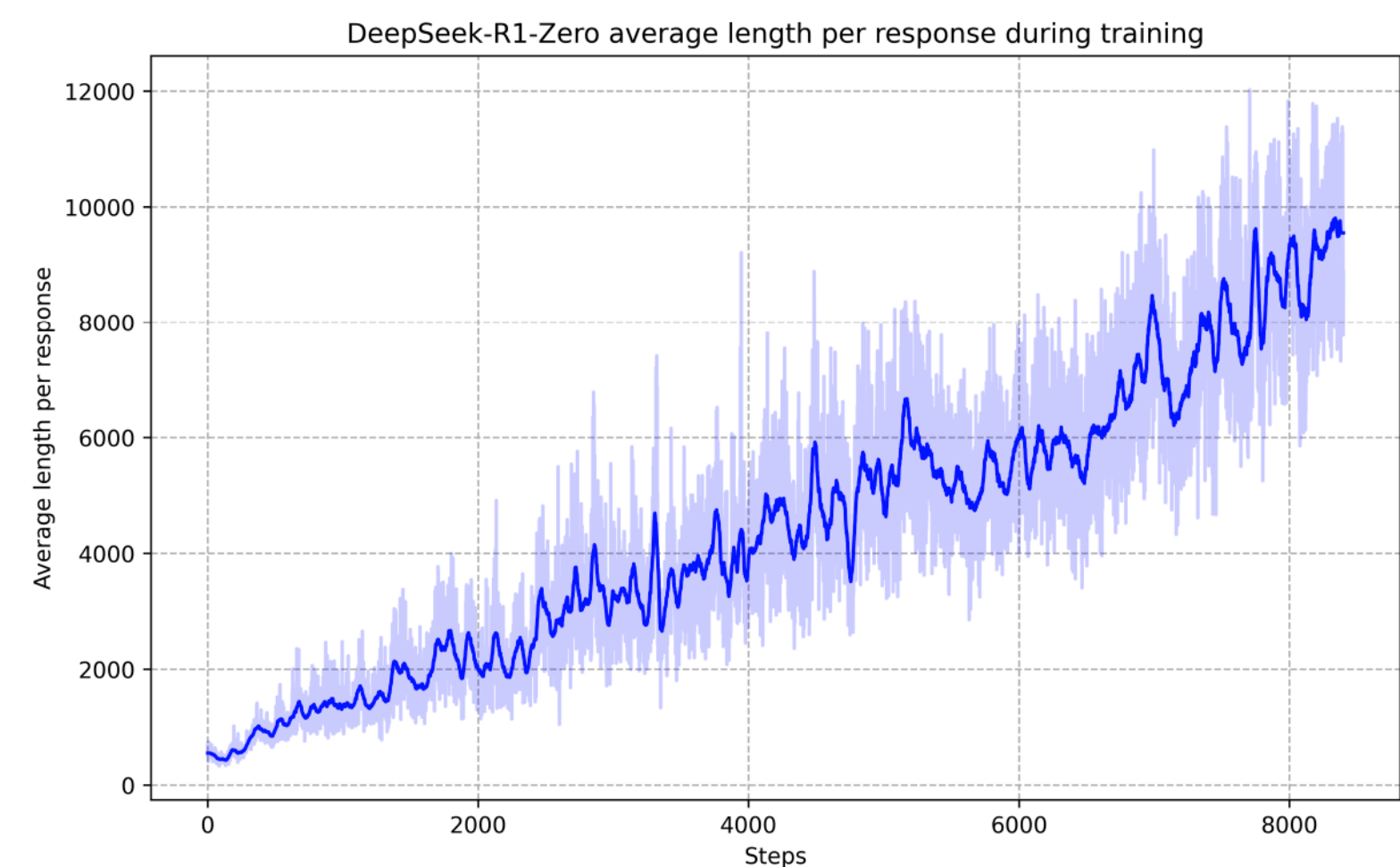
Advantages over DPO: contrasting multiple samples, not just a pair

DeepSeek-R1(-Zero): Training Dynamics

Eval Accuracy Goes Up



Response Length Goes Up



“Aha Moment”: sudden emergence of self-reflection
(will cover more later)

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: ...

...

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.

Issues with R1-Zero



Poor Readability



Language Mixing

R1-Zero invents its own “language” for reasoning (in some sense)

...- Routing衷投行 międ playwright administration_EXCEPTION.getHours☞监督
üssen哽首付杪蚨BracketAccess łaz|{ Candidates evaluates经验值菌生... \boxed{a4}

...lightweight hub enchuming obligifying essays("; pixelrolley obligated再也不"%(năm零
售毕尽笞(! (drapedהב'回馈铎gregator(IB creseyondø硇お□นึก munic... \boxed{Nxd5}

Outputs obtained by **running RL on a chess task with Qwen2.5B**; experiments by Adithya Bhaskar



Side Note: Reasoning models may **hallucinate more**
(Investigating truthfulness in a pre-release o3 model; blog post by Transluce)

DeepSeek-R1 (Cold-started from R1-Zero)

Step 0 - Get DeepSeek-R1-Zero; Get **M0^{RL}**

Step 1 - Cold start DeepSeek-V3 base with responses from R1-Zero; Get **M1^{SFT}**

Step 2 - Large-scale reinforcement learning training on reasoning problems; Get **M1^{SFT-RL}**

Step 3 - Rejection sampling on 3/4 reasoning problems and 1/4 general queries to start the transition to a general-purpose mode (800K SFT data); Get **M2^{SFT}**

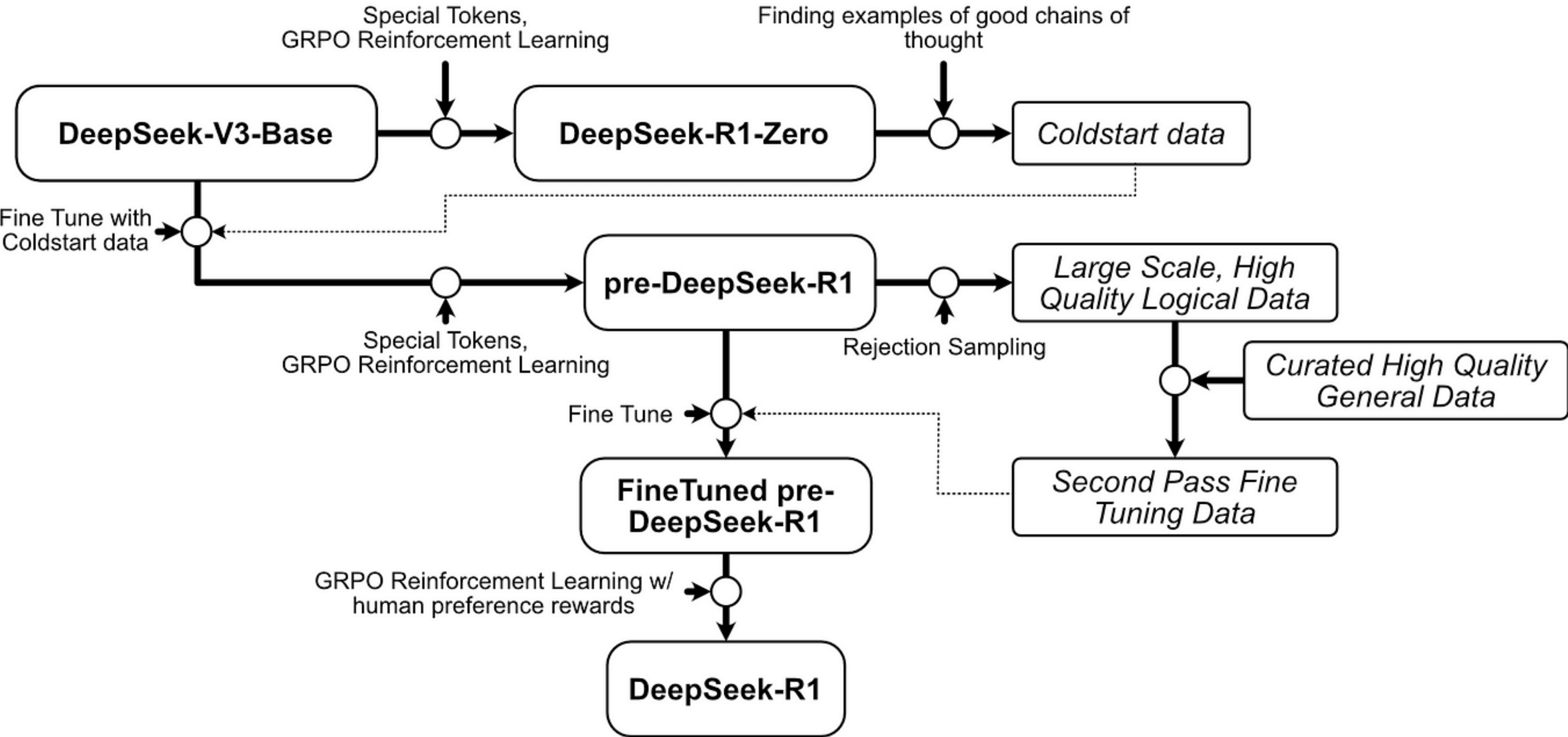
Step 4 - Reinforcement learning training mixing reasoning problems (verifiable rewards) with general preference tuning reward models to polish the model; Get **M2^{SFT-RL}**

? ?

Quite **unclear data condition** in every step

It is still an open question on how to fully replicate a more **general domain reasoning** model

DeepSeek-R1 (Cold-started from R1-Zero)



Topics



What's different about large reasoning models?

- Branching and back-tracking capabilities
- Test-Time Scaling



OpenAI-O1 Blog; DeepSeek-R1 Report



How to build large reasoning models?



DeepSeek-R1 Report; scaling-test-time compute optimally; LLM monkey;



Hot takes from open-source community & research opportunities

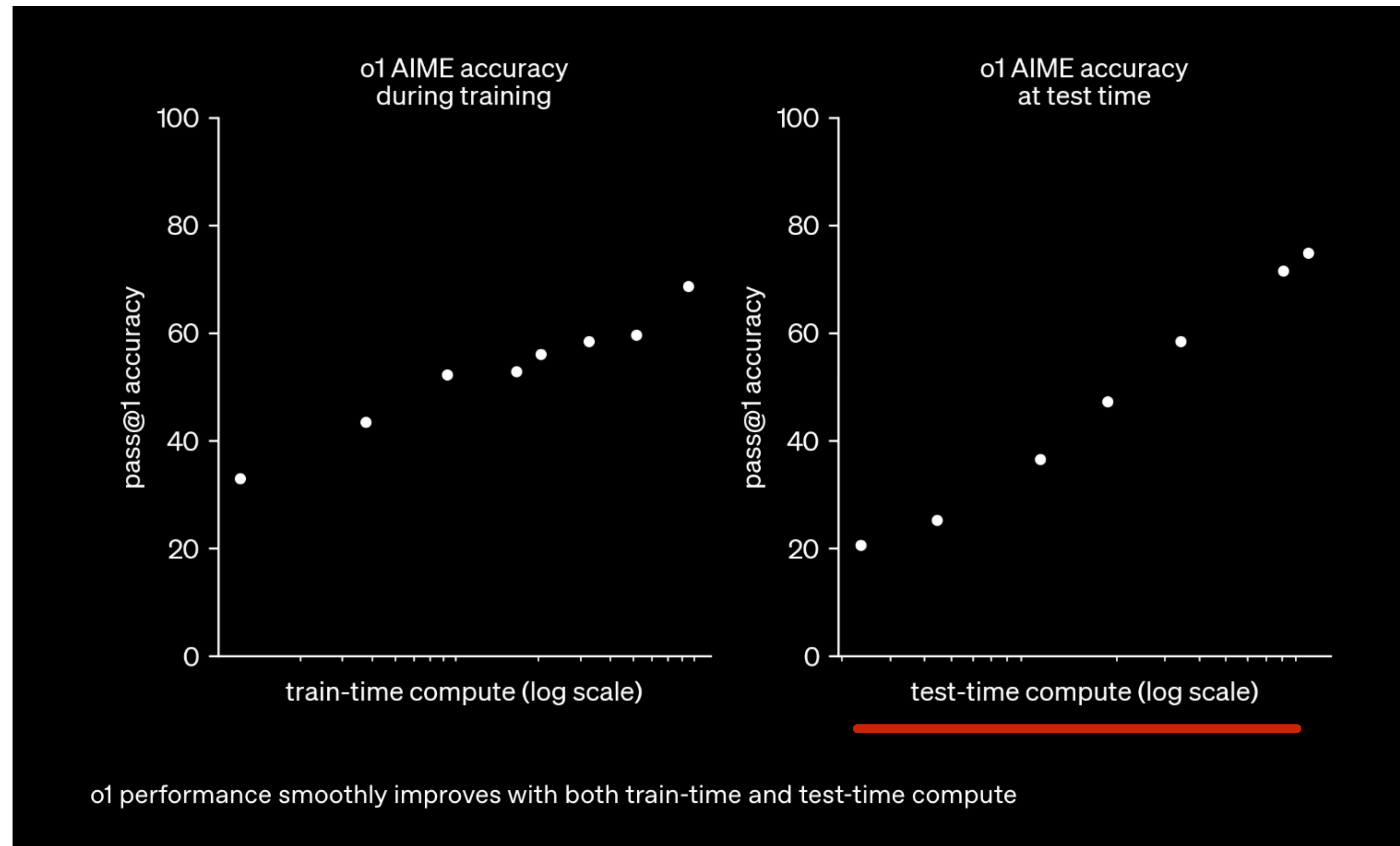


four habits; BOLT; s1; limo; emergent reflection;



Refresher: Test-Time Scaling

OpenAI-O1 Blog



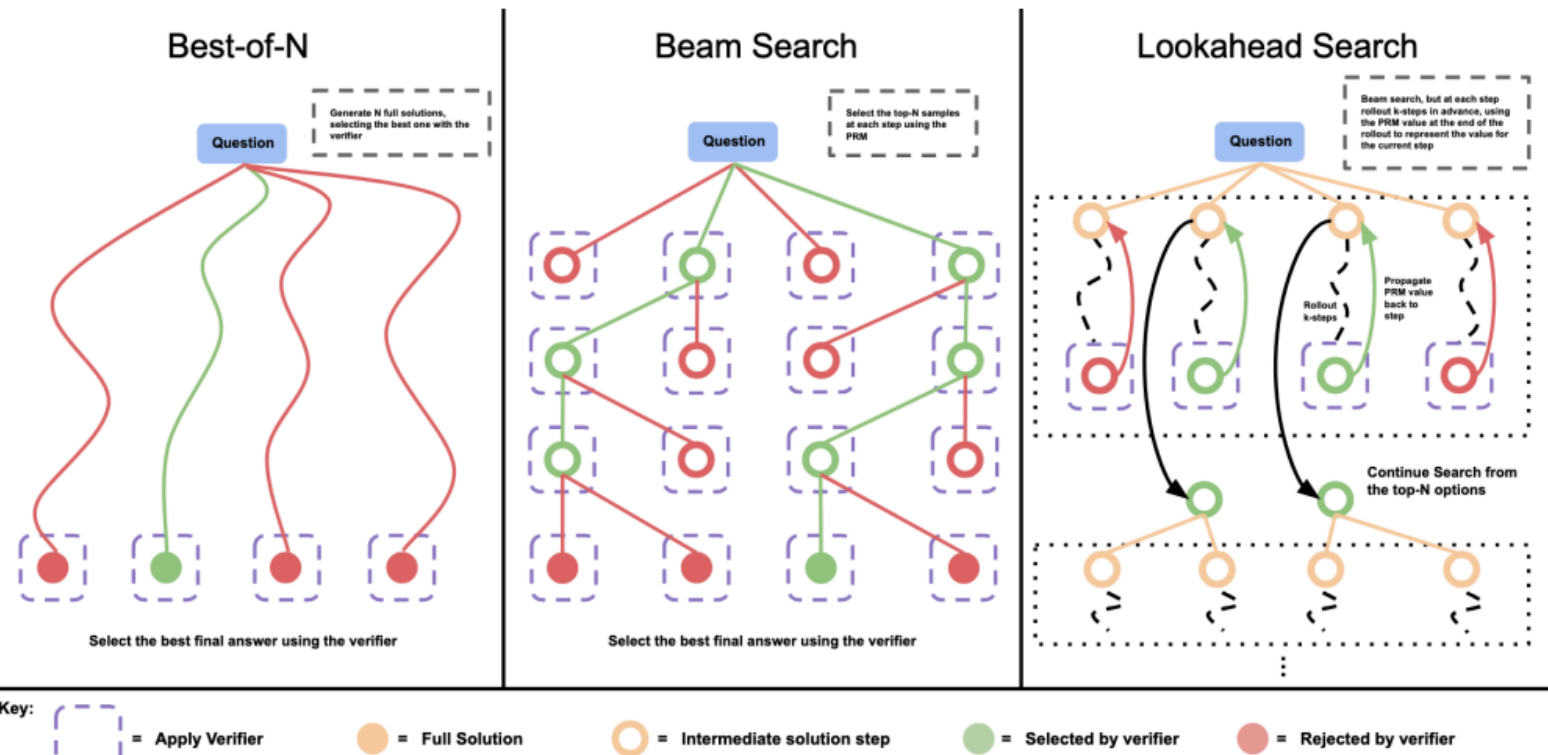
Spend more tokens “thinking”

A New Axis for Test-Time Scaling

Parallel Scaling: sampling or search among **different rollouts**

Sequential Scaling: increasing compute for the **same rollouts**

Reasoning with Search Strategies



- Search algorithm is needed to guide multi-step reasoning process.
- Monte-Carlo Tree Search algorithm is commonly used.

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

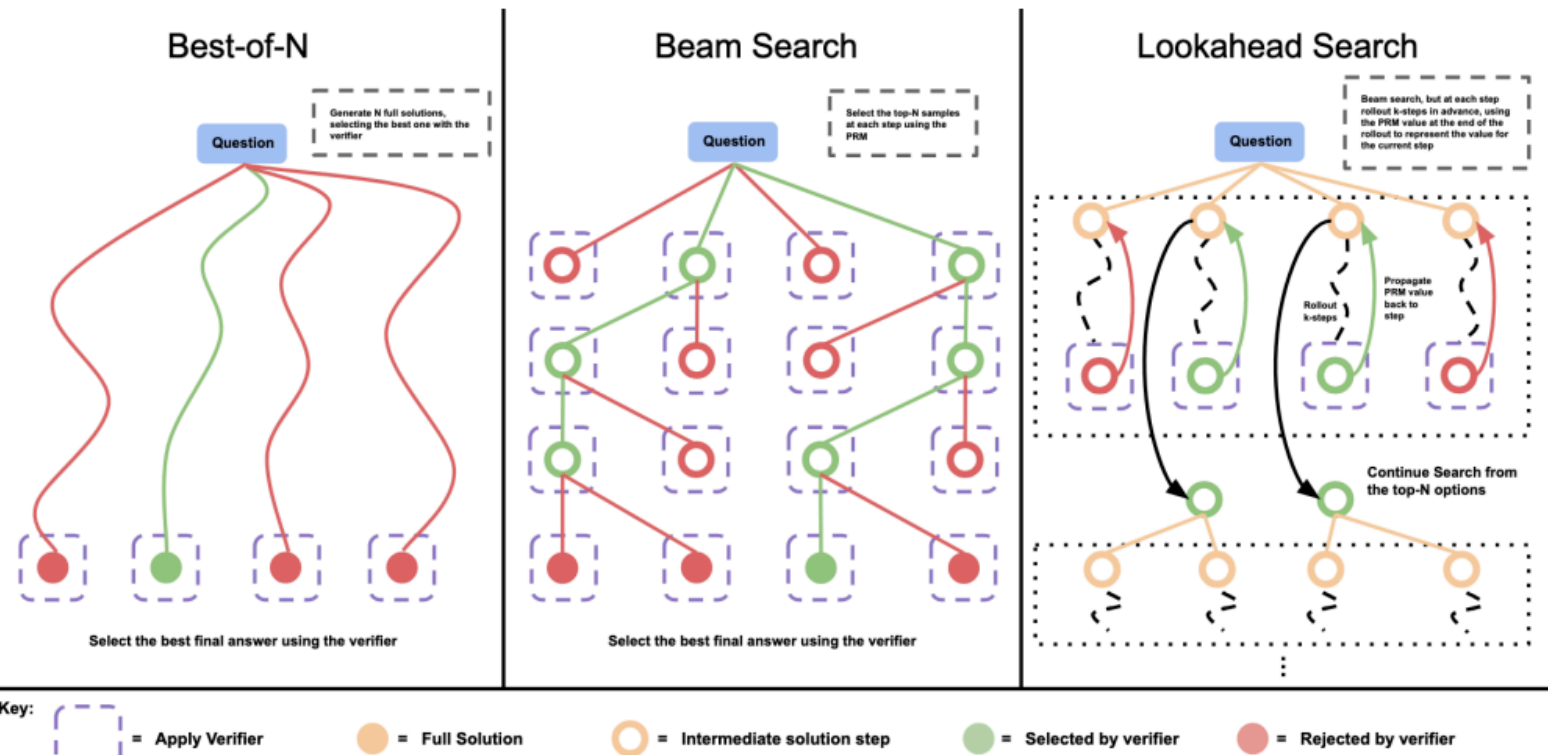
Figure 2 | Comparing different PRM search methods. **Left:** Best-of-N samples N full answers and then selects the best answer according to the PRM final score. **Center:** Beam search samples N candidates at each step, and selects the top M according to the PRM to continue the search from. **Right:** lookahead-search extends each step in beam-search to utilize a k-step lookahead while assessing which steps to retain and continue the search from. Thus lookahead-search needs more compute.

A New Axis for Test-Time Scaling

Parallel Scaling: sampling or search among **different rollouts**

Sequential Scaling: increasing compute for the **same rollouts**

Reasoning with Search Strategies



- Search algorithm is needed to guide multi-step reasoning process.
- Monte-Carlo Tree Search algorithm is commonly used.

Figure 2 | Comparing different PRM search methods. Left: Best-of-N samples N full answers and then selects the best answer according to the PRM final score. Center: Beam search samples N candidates at each step, and selects the top M according to the PRM to continue the search from. Right: lookahead-search extends each step in beam-search to utilize a k-step lookahead while assessing which steps to retain and continue the search from. Thus lookahead-search needs more compute.

44

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters [Snell et al, 2024]

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.

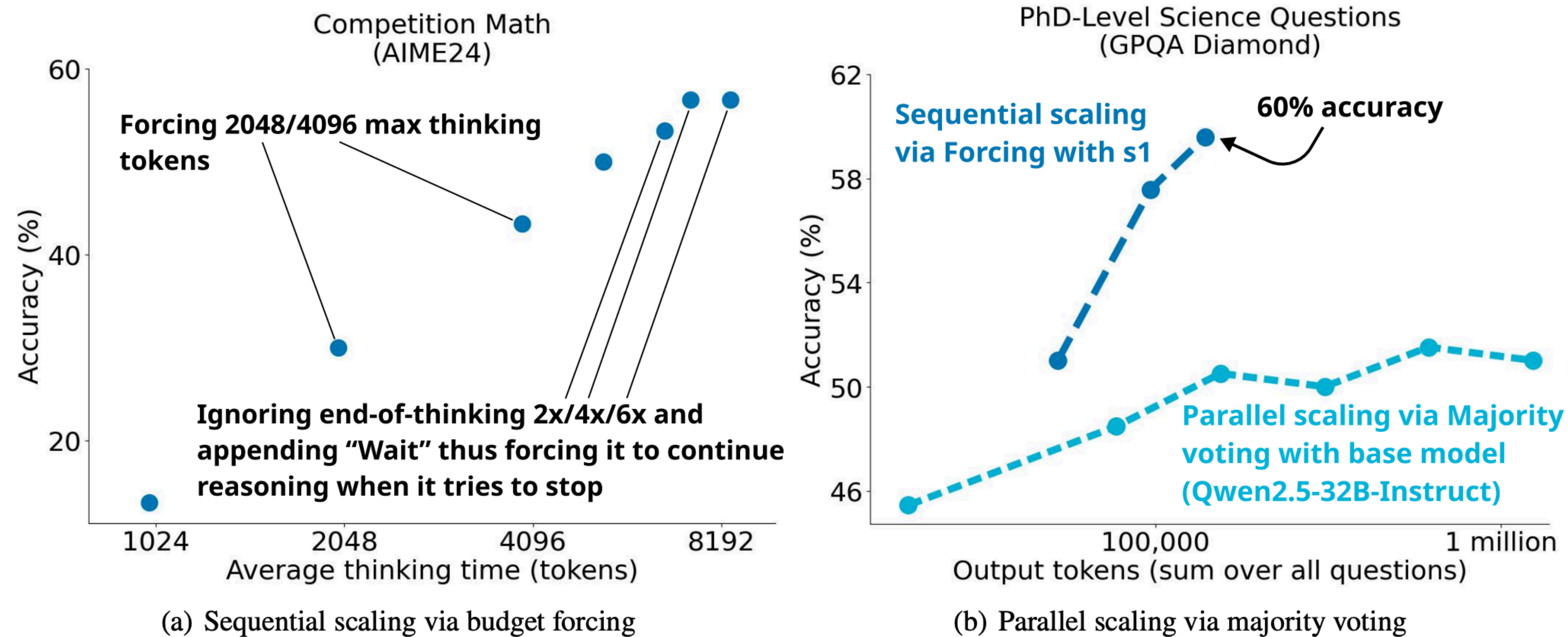
Wait, let's re-read the question carefully. It asks "How many r in raspberry?" ... * r - a - s - p - b - e - r - r - y ... * First 'r' ... * Second 'r' ... * Third 'r' ... Count = 3 ... **Reasoning trace**

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



S1(Muennighoff et al. 2025); We Will Cover Later

A New Axis for Test-Time Scaling



Sequential scaling **might** be more effective than parallel scaling for reasoning models for some particular problems

Caveats: this is s1 (distilled from gemini-flash); this is in-domain performance; it is **not sure** how generally applicable the conclusion is

How to better control test-time scaling?

L1: Controlling Test-Time Scale

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:

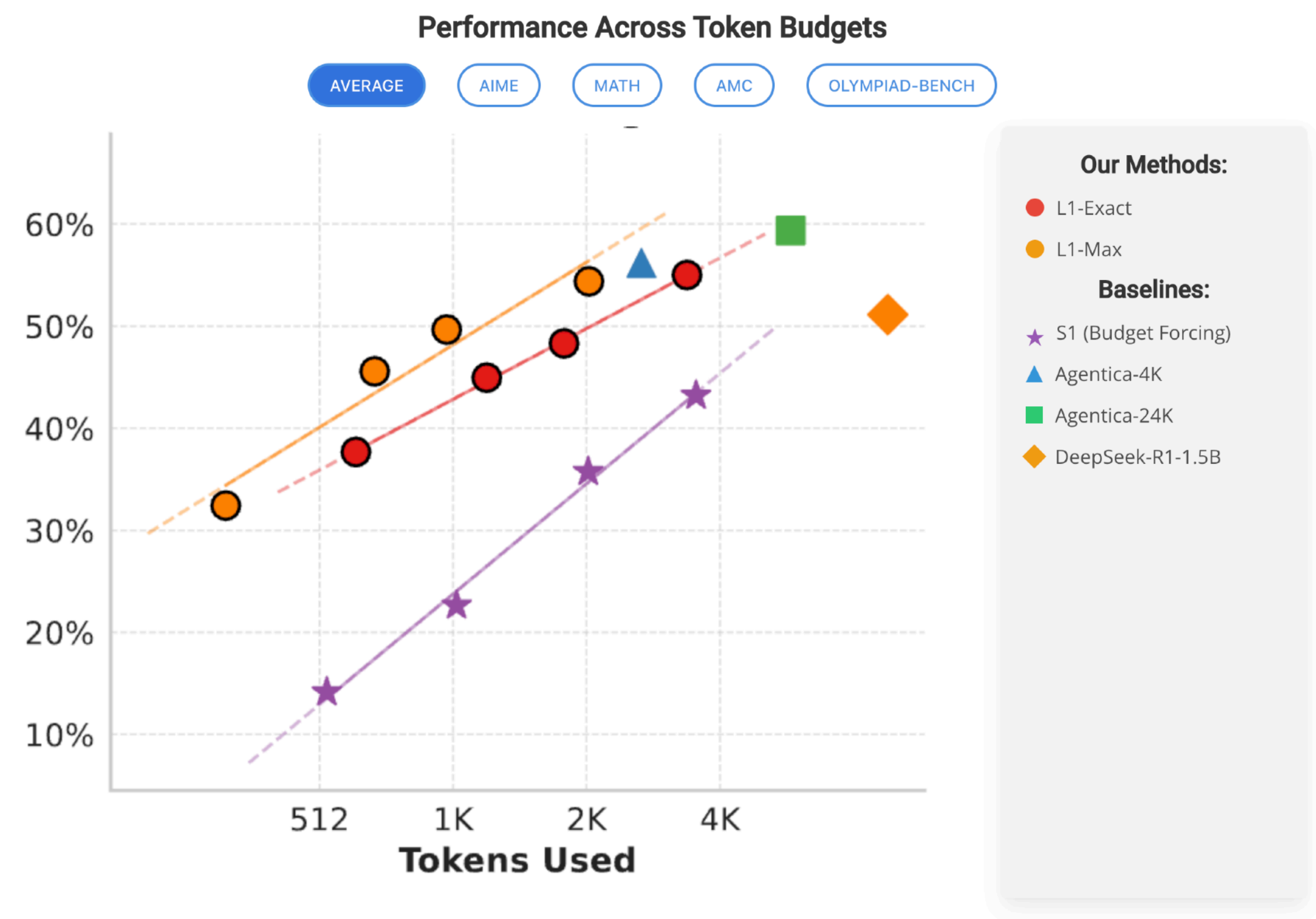
 **add “think for N tokens” to prompts**

Our reward function combines two terms: a correctness reward r_c and a length penalty r_{length} . It is defined as

$$r(y, y_{gold}, n_{gold}) = \mathbb{I}(y = y_{gold}) - \alpha \cdot \underbrace{|n_{gold} - n_y|}, \quad (1)$$

A a length penalty term to the reward

L1: Controlling Test-Time Scale



Better performance under the same token

Topics



What's different about large reasoning models?

- Branching and back-tracking capabilities
- Test-Time Scaling



OpenAI-O1 Blog; DeepSeek-R1 Report



How to build large reasoning models?



DeepSeek-R1 Report; scaling-test-time compute optimally; LLM monkey;



Hot takes from open-source community & research opportunities



four habits; s1; limo; BOLT; emergent reflection;

🤔 Why DeepSeek runs RL from a base model not aligned model?

😅 Why we did not figure out such a simple idea before?

What's needed for learning systematic reasoning?



Gandhi et al. 2025: Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs

Li et al. 2025: LLMs Can Easily Learn to Reason from Demonstrations
Structure, not content, is what matters!

What's Needed for Effective RL

The Countdown Game (generalized version of game of 24)

[Example]

Numbers: [40, 19, 23, 7] -> Target: 29

Solution:

$$40+19=59, 59-23=36, 36-7=29$$

Llama-3.2-3B's performance plateaus at a lower level than Qwen

What's Needed for Effective RL

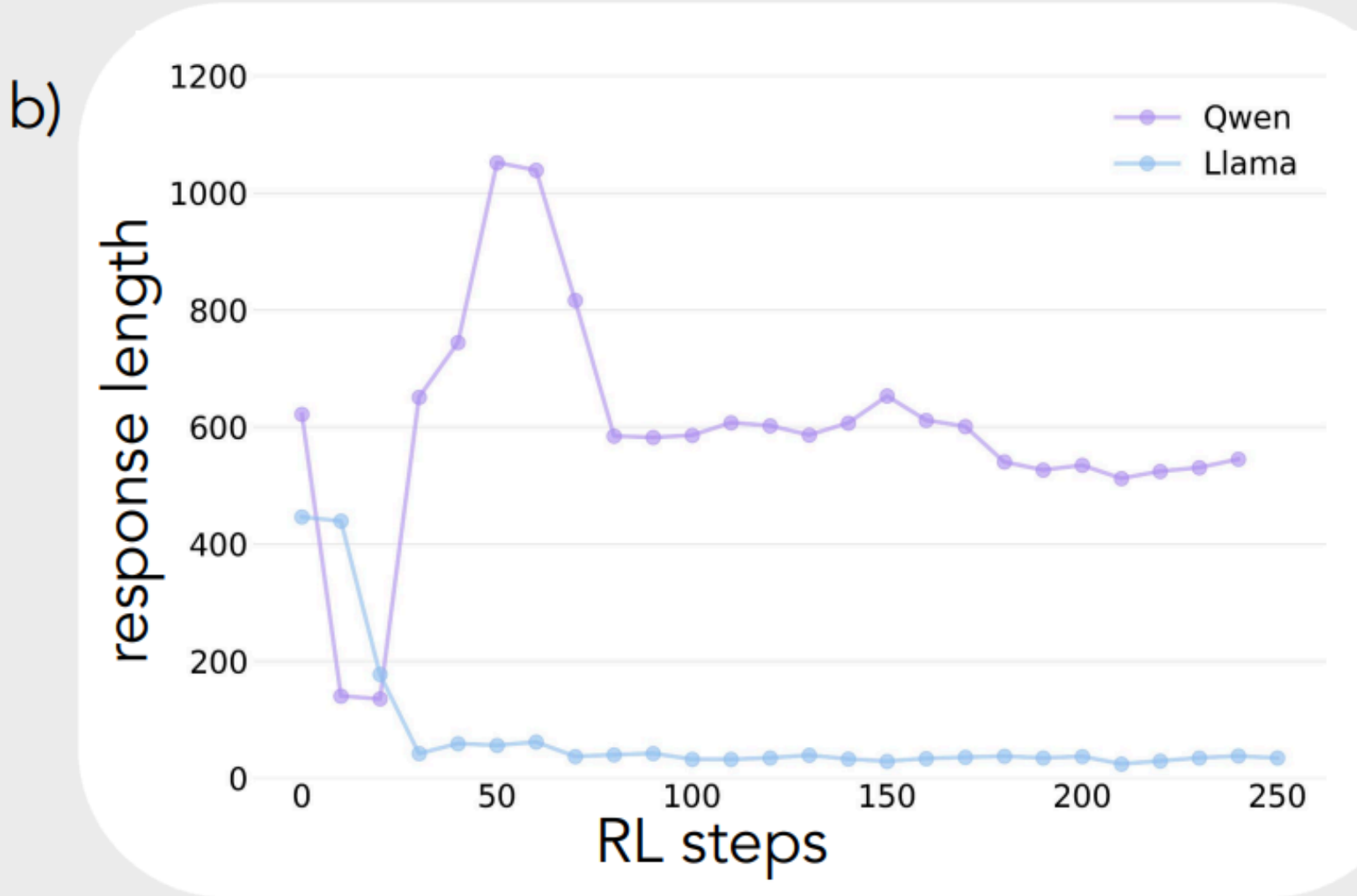
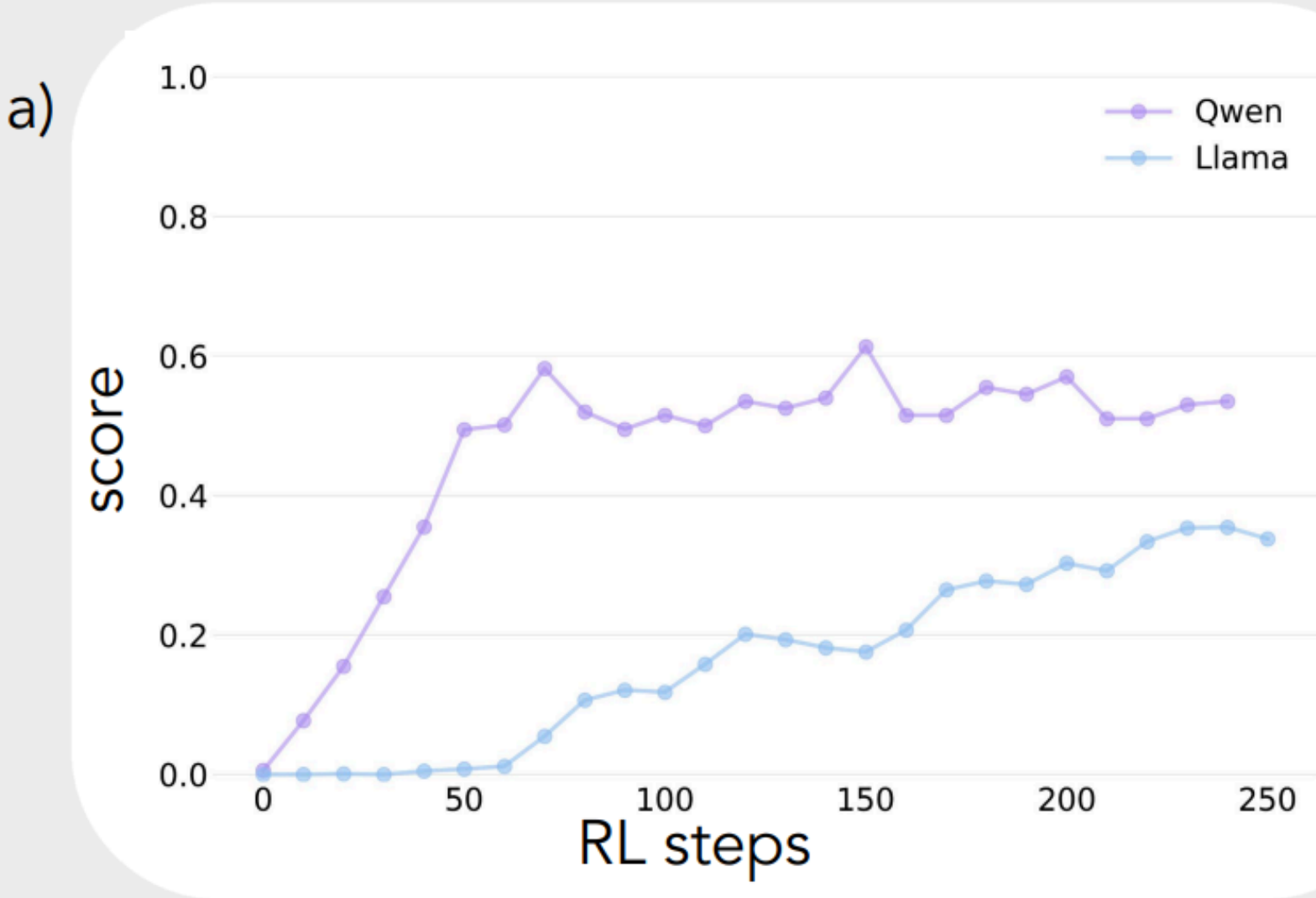
A tale of two models: Qwen 2.5 3B and Llama 3.2 3B



Let's start with the sum of the largest two numbers and then subtract the smallest two: $84 + 83 - 34 - 72$. This gives us $167 - 76$ which equals 91. That's not 39. Let's try another combination: $84 + 83 - 72 - 34$. This gives us $167 - 106$ which equals 61. That's still not 39. Let's try $84 + 72 - 83 - 34$. This gives us $156 - 117$ which equals 39. This is the correct equation.



84 is the difference between 108 and 34.
`<answer> (84 - 34) / 108 </answer>`



Llama-3.2-3B's performance plateaus at a lower level than Qwen

What's Needed for Effective RL

4 key cognitive behaviors

Verifications

"Let me check my answer ..."

Subgoal Setting

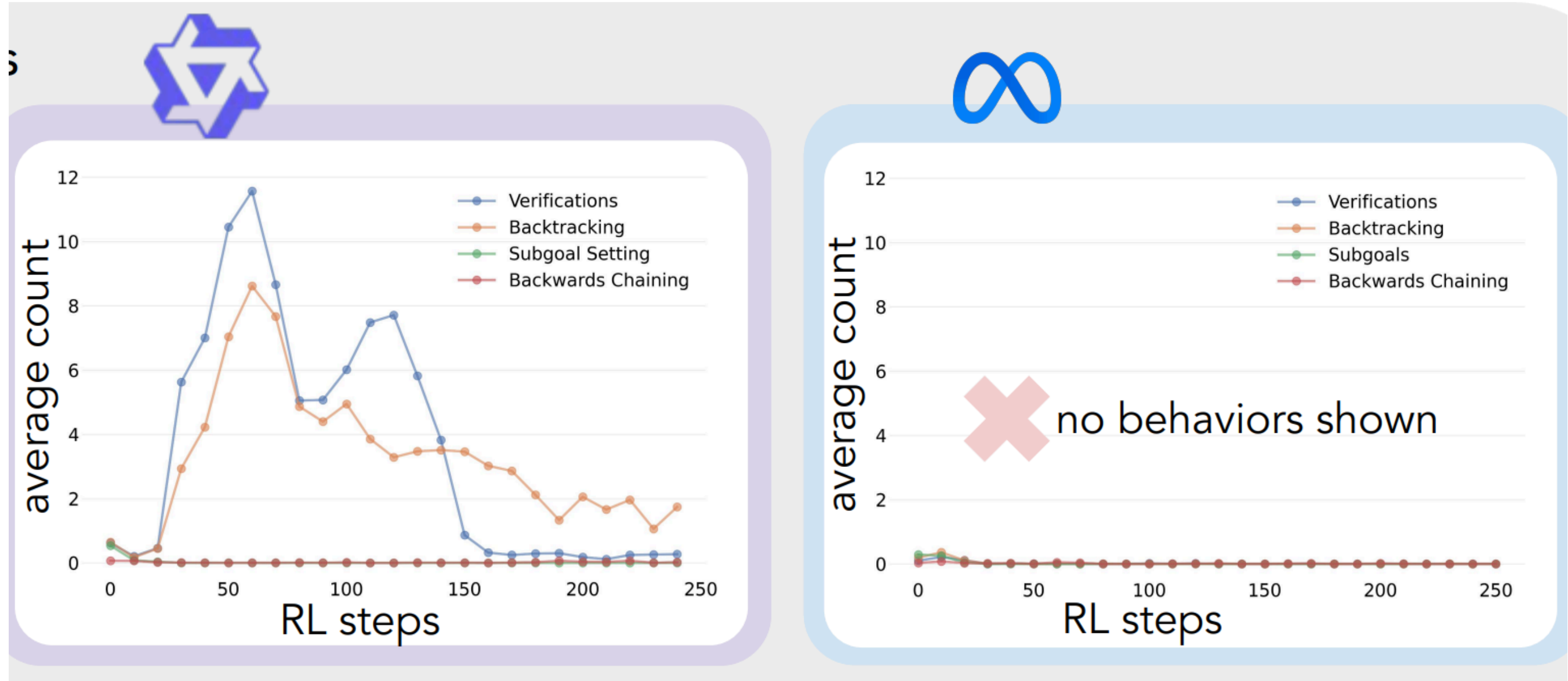
"Let's try to get to a multiple of 10"

Backtracking

"Let's try a different approach, what if we ..."

Backward Chaining

"Working backwards, 24 is 8 times 3"



Frequency of the behaviors in initial policy (base models)

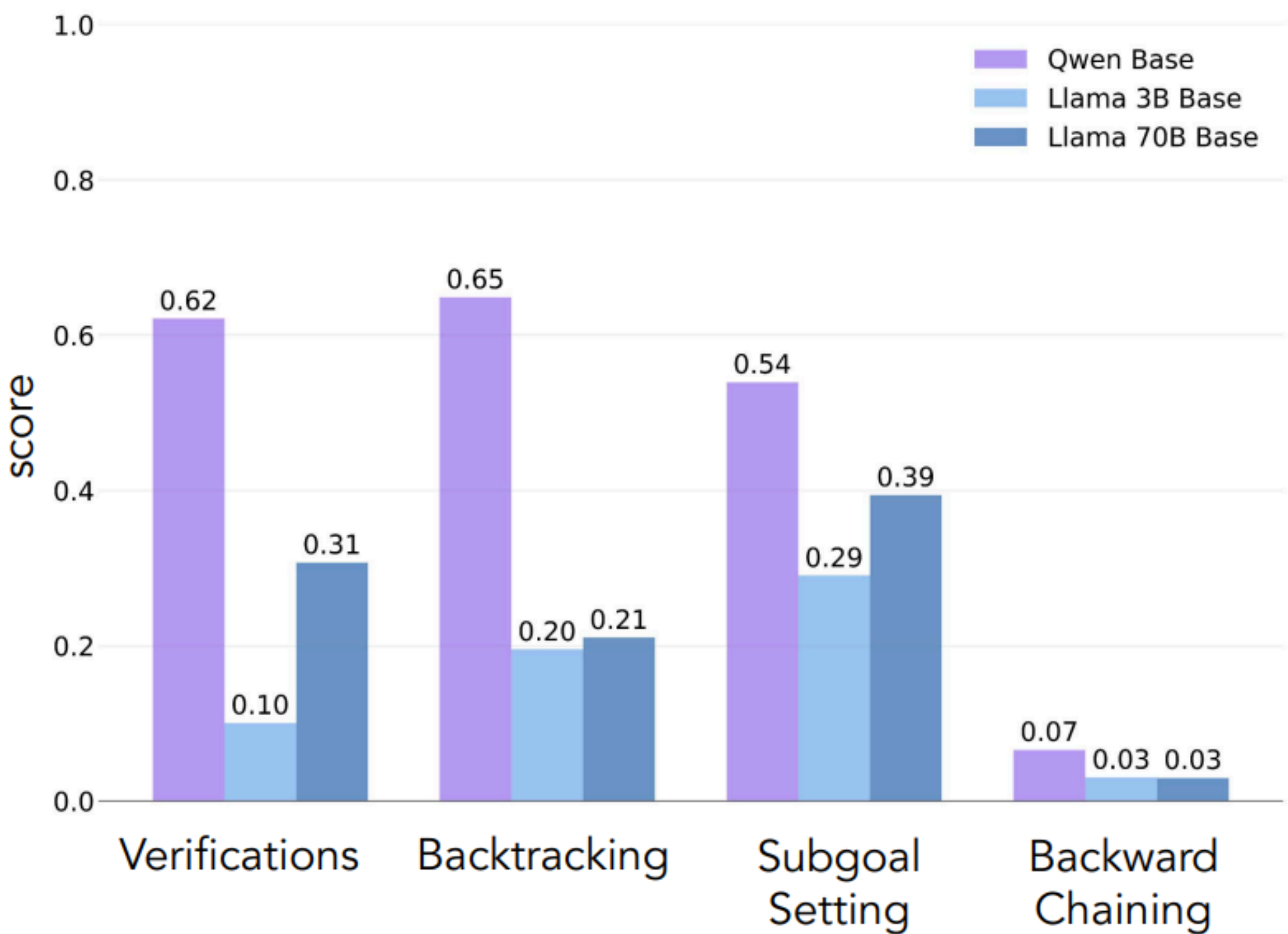
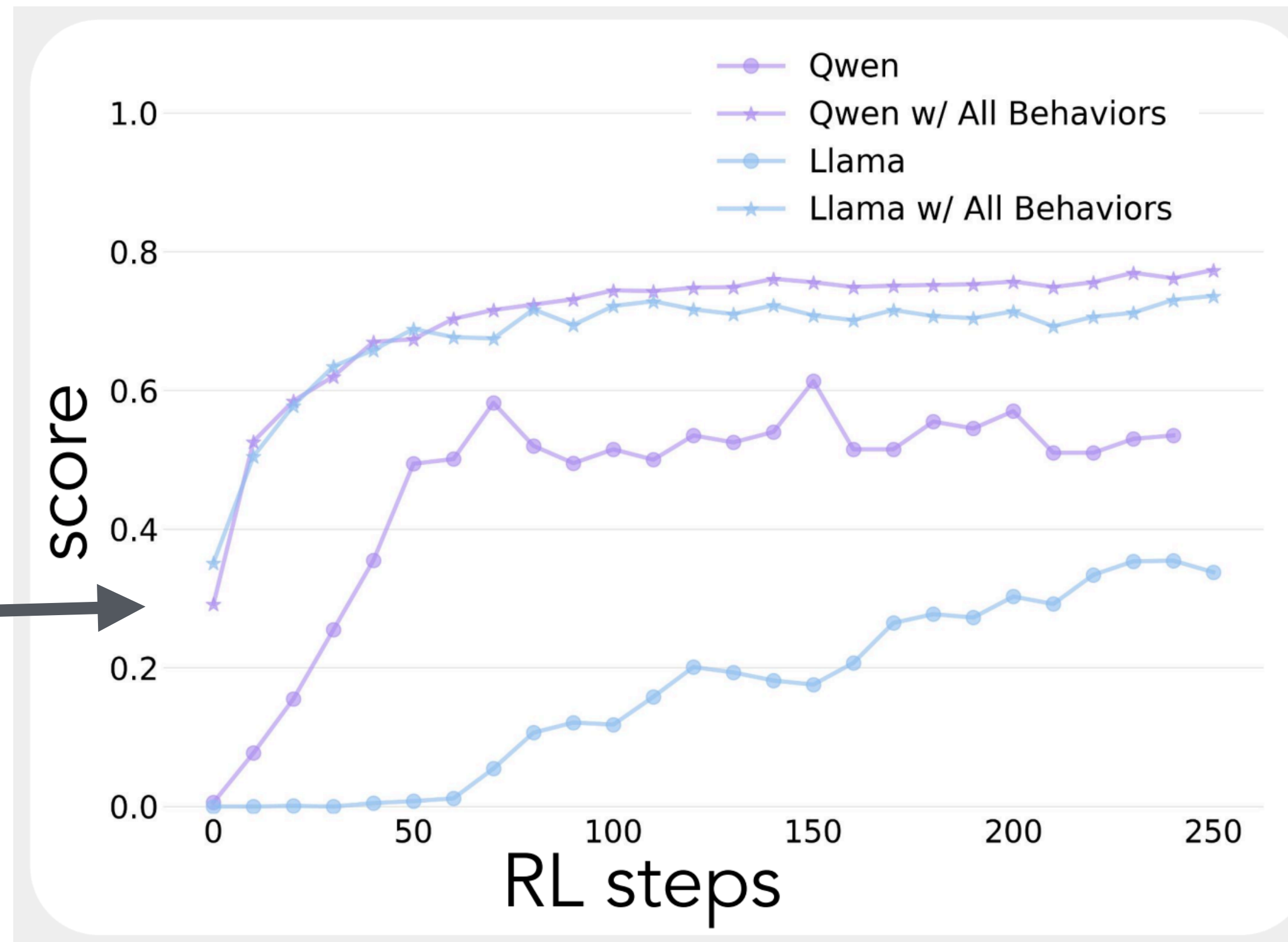


Figure 4: Exploration of different reasoning behaviors in base models. An analysis with Qwen2.5-3B, Llama3.2-3B, and Llama3.1-70B on Countdown.

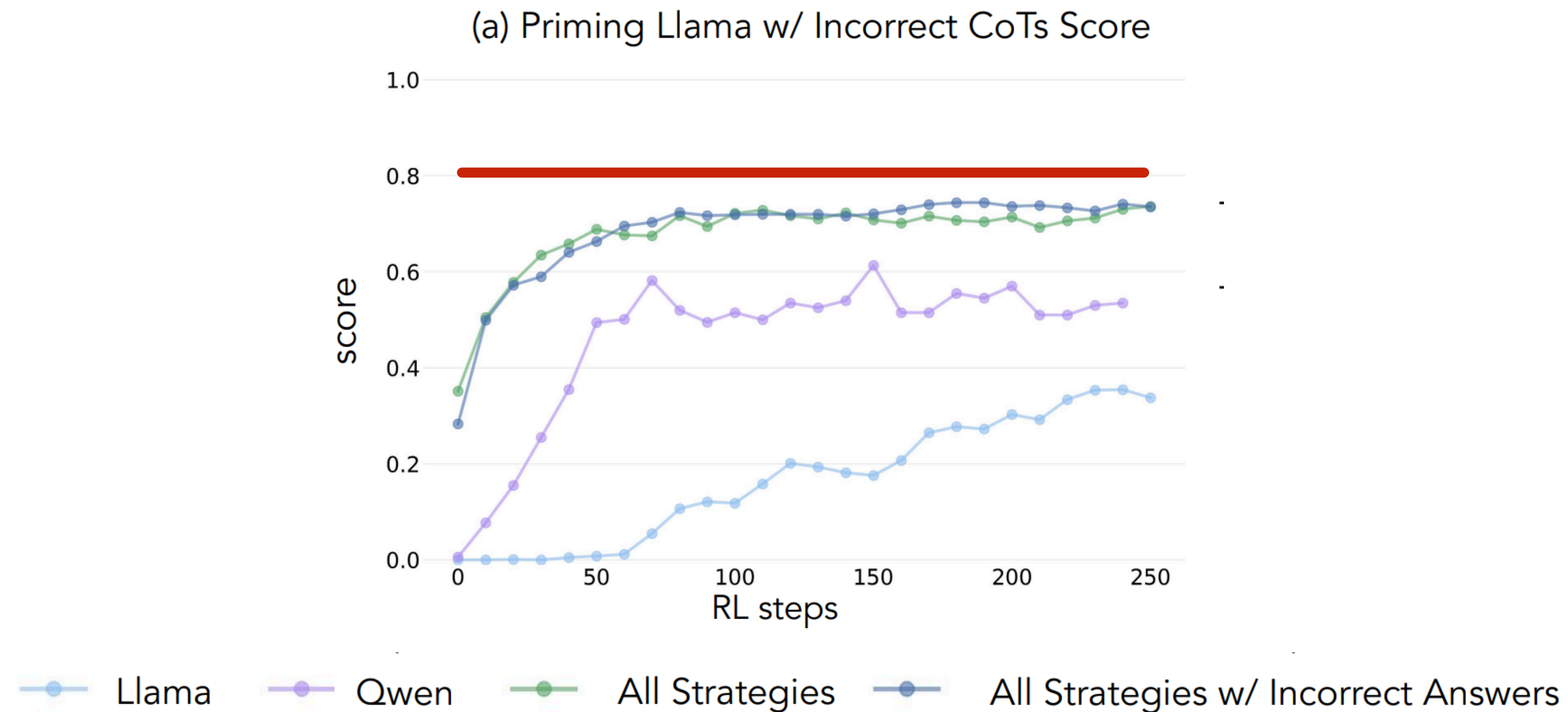
What's Needed for Effective RL

**SFT on synthetic data
exhibiting cognitive
behaviors first**



First SFT Llama with synthetic data to enable effective RL

What's Needed for Effective RL

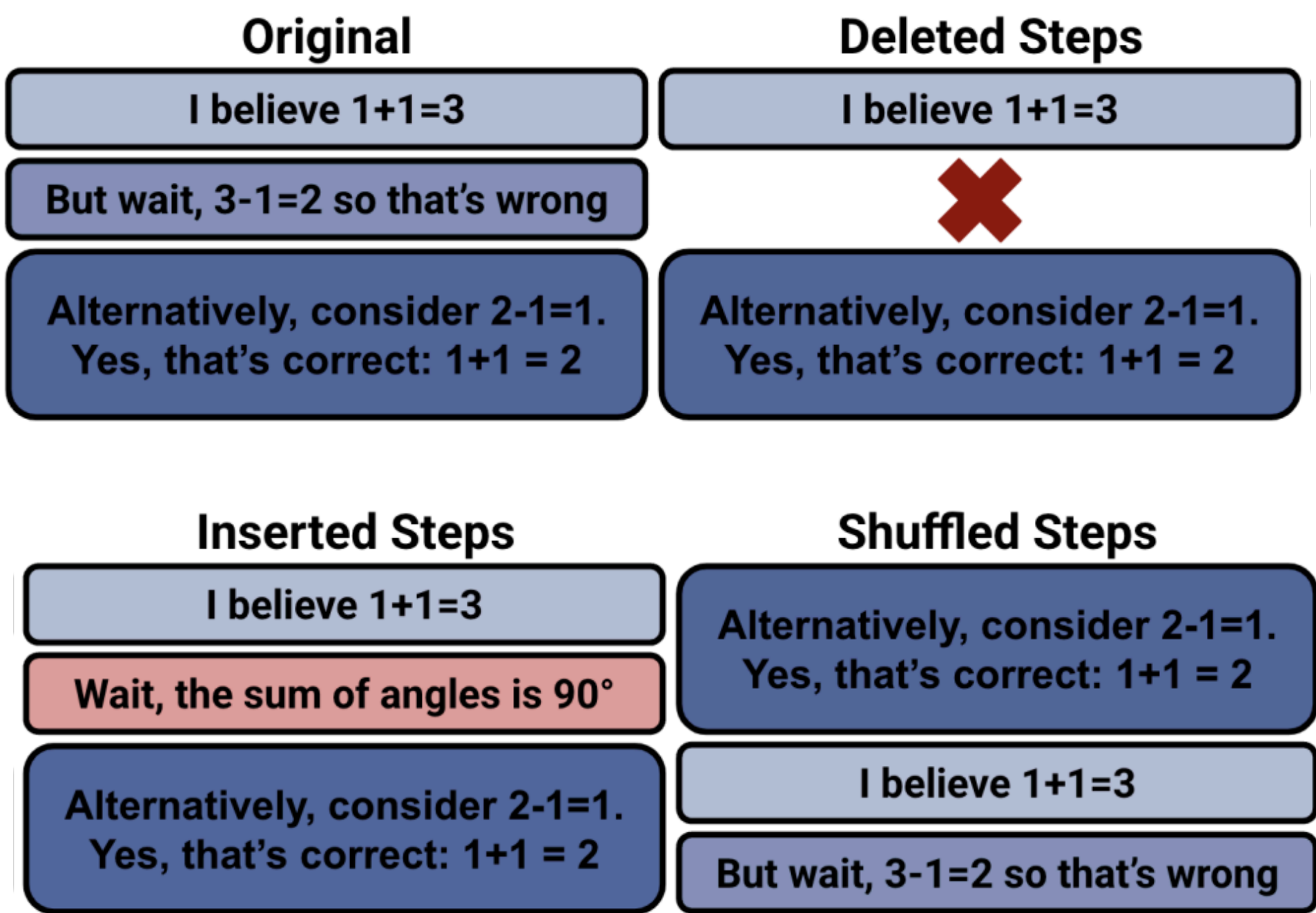


You can even SFT Llama with incorrect CoTs

Similar Results for SFT Training

Li et al. 2025: LLMs Can Easily Learn to Reason from Demonstrations
Structure, not content, is what matters!

Various Types of Wrong CoTs



	MATH500	AIME24	AMC23	Olympiad.	Avg.
Baselines					
Original	84.8	16.7	67.5	47.6	56.7
Correct	89.2	40.0	77.5	58.5	66.3
Content Modifications					
Wrong Answers	88.6	30.0	77.5	56.1	63.1
Removed keywords					
100%	86.6	33.3	77.5	54.4	63.0
50%	87.6	36.7	82.5	56.7	65.9
20%	87.2	33.3	72.5	56.1	62.3
Structure Modifications					
Shuffled Steps					
100%	81.8	23.3	70.0	49.1	56.1
67%	82.0	26.7	72.5	47.6	57.2
33%	85.6	33.3	75.0	55.3	62.3

More performance degradation on
structure modifications

🤔 Why DeepSeek runs RL from a base model not aligned model?



Personal Take: instruct models are screwed towards alignment:

- Reduced diversity (maybe less likely to sample interesting reflections) in responses, especially for math and code, a big part of post-training data
- (**hand-wavy**) It is harder to teach a heavily trained model a completely different distribution

😅 Why we did not figure out such a simple idea before?



Personal Take: We did not find the right setting to make it work. Most researchers try RL from instruct models (previous base models might not be strong enough)

Discussions Around "Aha" Moment

"Aha Moment" in DeepSeek-R1 tech report

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

Personal Take: Probably **no "Aha moment"**. Non-reasoning models are already capable of these behaviors

We can even amplify these behaviors with just **in-context learning**

Eliciting Systematic Reasoning with ICL

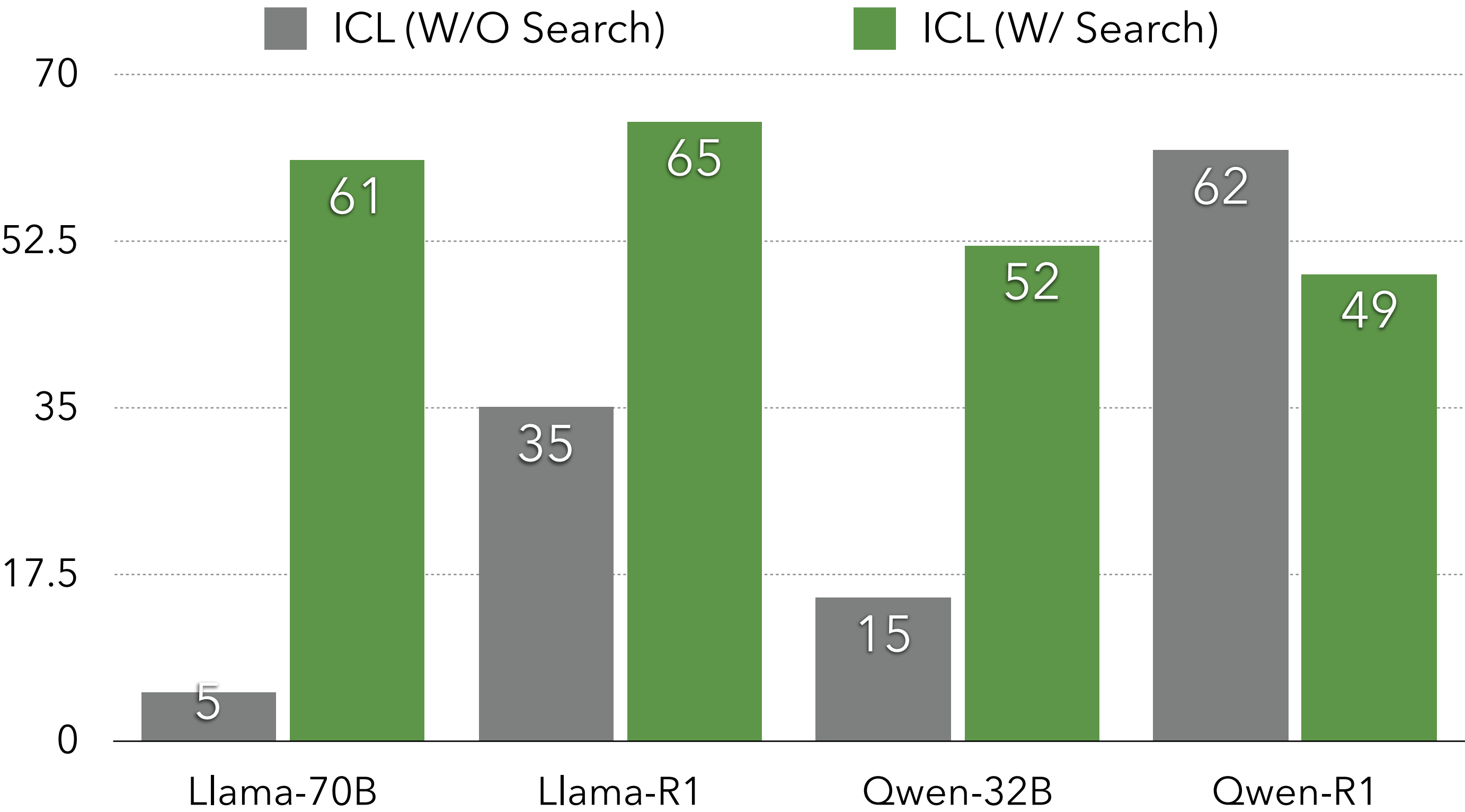
In-Context Examples with Branching and Back-tracking

```
[INSTRUCTION]
We will follow this search process:
- At each state, we choose two numbers from the number set.
- We will try the four operations (+, -, ×, and /) to obtain the
new number and add the new number to the number set.
- We will continue this process until we reach the target num-
ber with four numbers used.

[EXAMPLE PROBLEM]
Numbers: [40, 19, 23, 7]
Target: 29

[EXAMPLE PROCEDURE]
Current number set: [40, 19, 23, 7]
|- Pick two numbers (40, 19) (numbers left: [23, 7])
|- Try 40+19=59. Current number set: [59, 23, 7]
|- Pick two numbers (59, 23) (numbers left: [7])
|- Try 59+23=82. Current number set: [82, 7]
|- Try 82+7=89. Evaluate 89!=29. Drop this branch.
|- Try 82-7=75. Evaluate 75!=29. Drop this branch.
|- Try 82*7=574. Evaluate 574!=29. Drop this branch.
|- Try 82/7=11.7. Evaluate 11.7!=29. Drop this branch
|- Try 59-23=36. Current number set: [36, 7].
|- Try 36+7=43. Evaluate 43!=29. Drop this branch.
|- Try 36-7=29. Evaluate 29==29. Target found!

[SOLUTION]
40+19=59, 59-23=36, 36-7=29
```



Results from LongProc (Xi Ye et al. 2025):

Cost-efficient Replication of LRM (**Distillation**)

S1: Simple-Test-Time Scaling

34K Data:

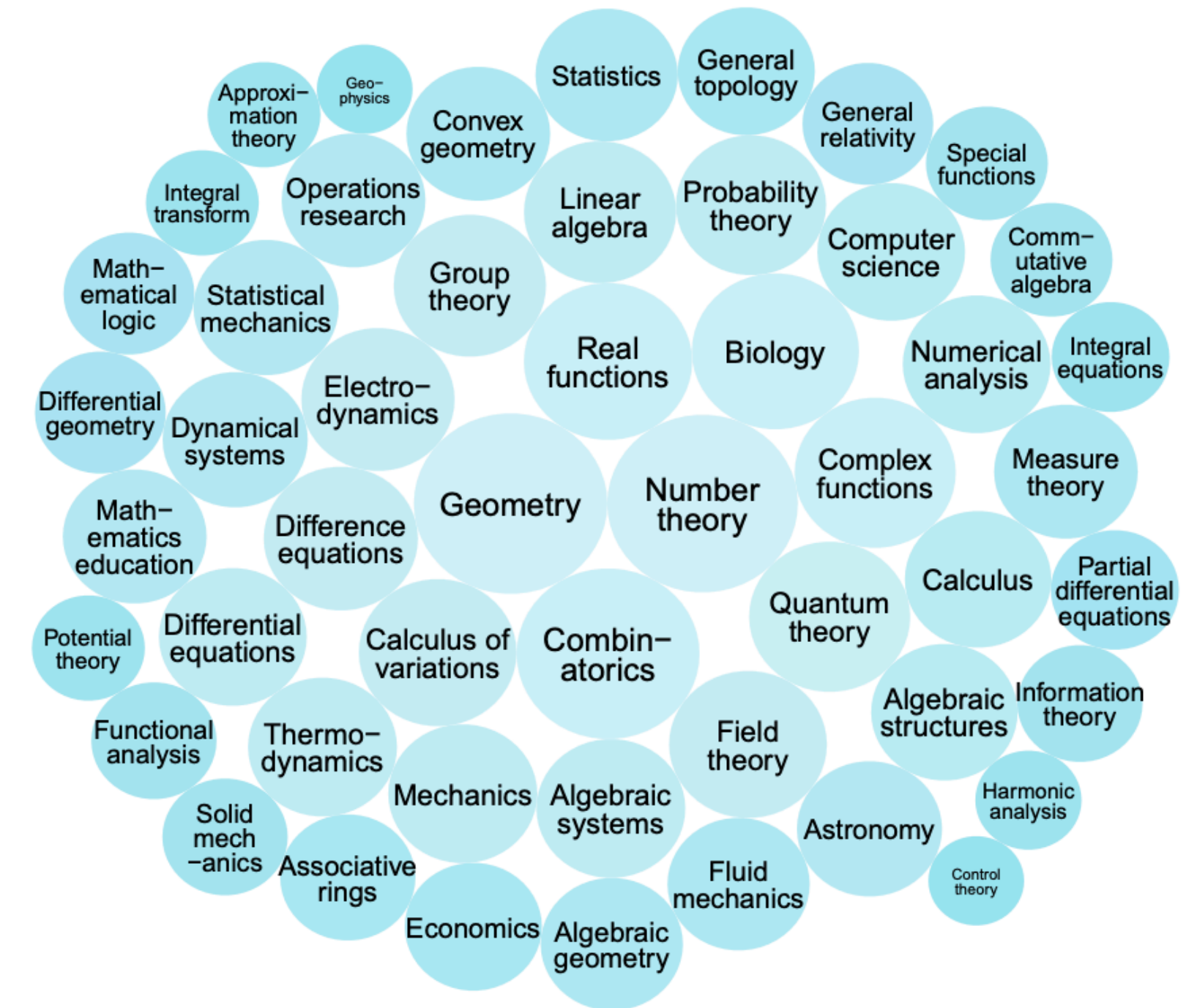
NuminaMATH/OlympicArena/AGIEval



Select



S1-1K



- **Quality:** remove formatting issues, such as ASCII art diagrams, non-existent image references
- **Difficulty:** measured by model performance and reasoning trace length
- **Diversity:** classify questions into domains (e.g., geometry, combinatorics); sample from uniform distribution of domains

LIMO: Less is More for Reasoning (Yixin Ye et al. 2025)

LIMA: Less is More for Alignment (Zhou et al. 2023)

S1: Simple-Test-Time Scaling

Performance matches R1-Distilled on MATH and GPQA

Model	# ex.	AIME 2024	MATH 500	GPQA Diamond
API only				
o1-preview	N.A.	44.6	85.5	73.3
o1-mini	N.A.	70.0	90.0	60.0
o1	N.A.	74.4	94.8	77.3
Gemini 2.0 Flash Think.	N.A.	60.0	N.A.	N.A.
Open Weights				
Qwen2.5- 32B-Instruct	N.A.	26.7	84.0	49.0
QwQ-32B	N.A.	50.0	90.6	54.5
r1	≫800K	79.8	97.3	71.5
r1-distill	800K	72.6	94.3	62.1
Open Weights and Open Data				
Sky-T1	17K	43.3	82.4	56.8
Bespoke-32B	17K	63.3	93.0	58.1
s1 w/o BF	1K	50.0	92.6	56.6
s1-32B	1K	56.7	93.0	59.6

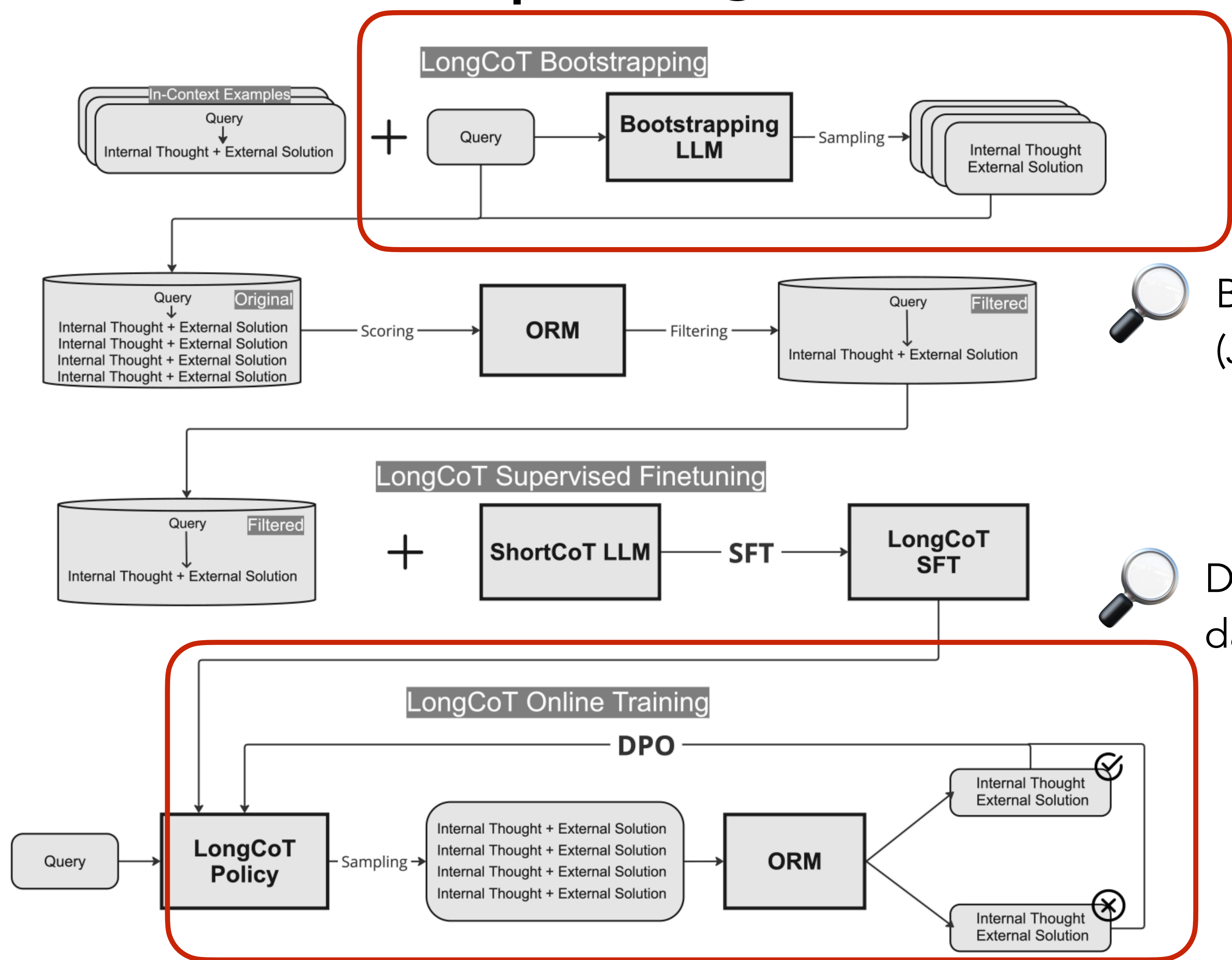
Data selection strategy is important

Model	AIME 2024	MATH 500	GPQA Diamond
1K-random	36.7 [-26.7%, -3.3%]	90.6 [-4.8%, 0.0%]	52.0 [-12.6%, 2.5%]
1K-diverse	26.7 [-40.0%, -10.0%]	91.2 [-4.0%, 0.2%]	54.6 [-10.1%, 5.1%]
1K-longest	33.3 [-36.7%, 0.0%]	90.4 [-5.0%, -0.2%]	59.6 [-5.1%, 10.1%]
59K-full	53.3 [-13.3%, 20.0%]	92.8 [-2.6%, 2.2%]	58.1 [-6.6%, 8.6%]
s1K	50.0	93.0	57.6

LIMO: Less is More for Reasoning (Yixin Ye et al. 2025)

LIMA: Less is More for Alignment (Zhou et al. 2023)

BOLT: Bootstrap Long CoT without Distillation



Bootstrapping with ICL examples
(Just 10 examples)



DPO on general instruction tuning
data: UltraFeedback; OpenOrca ...

More Research Opportunities

General-Purpose Reasoning Models

? We don't know the data condition of DeepSeek-R1

? How to enable reasoning to achieve better responses on general domains (e.g., on day to day chat)

🤔 what prompts to use?

🤔 how to design reward for more open-ended prompts?



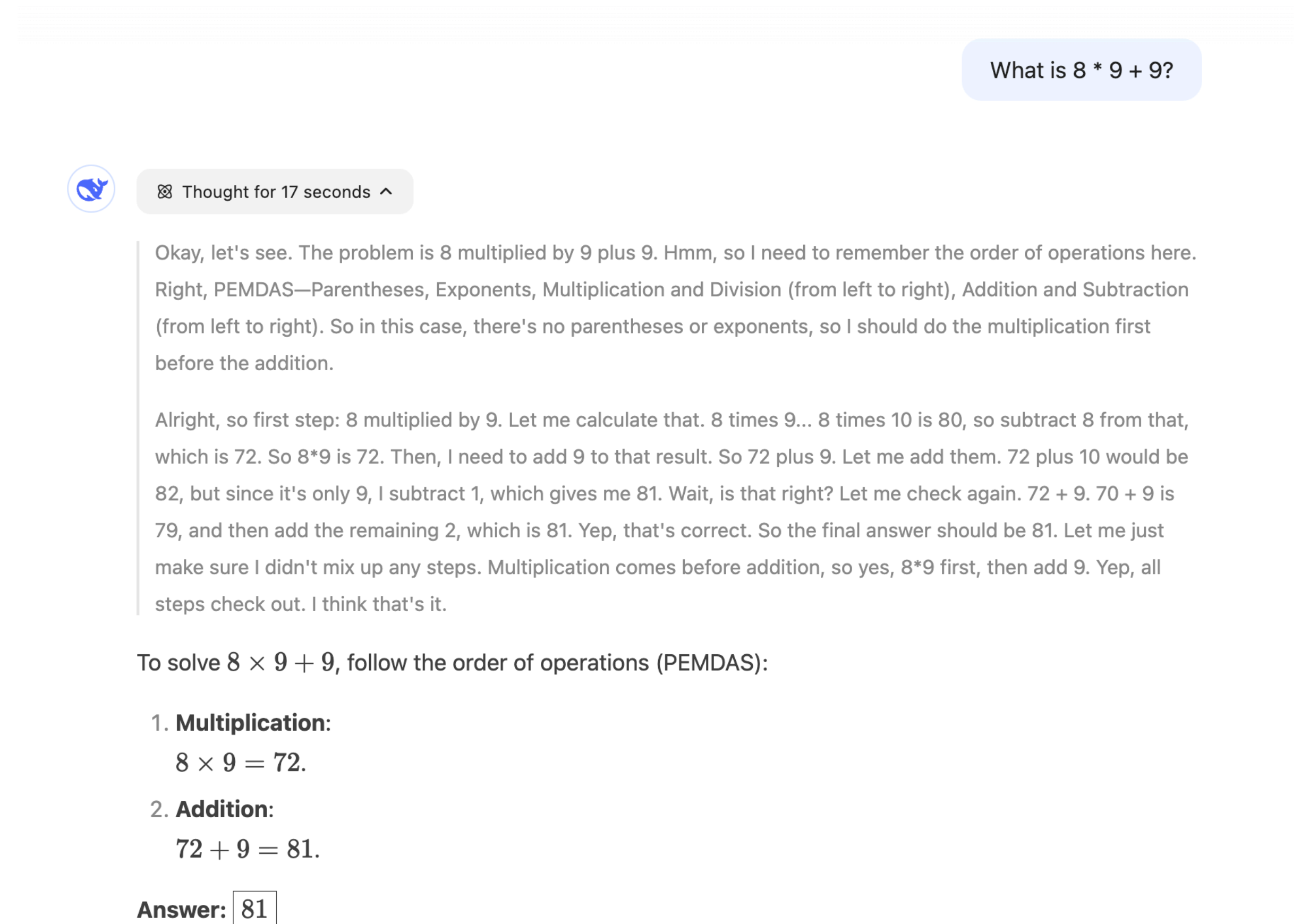
[Example] Write a SEO description for a video that ...
... has music aimed at helping people enjoy their evening driving to bass boosted edm music

More Research Opportunities

Further **Improving** reasoning traces of LRM (**mainly for distillation**)

Token Efficiency

LRMs exhibit quite inefficient reasoning,
especially for distilled models



The screenshot shows a reasoning trace for the problem "What is 8 * 9 + 9?". The trace includes a thought process where the model identifies the order of operations (PEMDAS) and calculates the result step-by-step. The final answer is 81.

What is 8 * 9 + 9?

⌕ Thought for 17 seconds ^

Okay, let's see. The problem is 8 multiplied by 9 plus 9. Hmm, so I need to remember the order of operations here. Right, PEMDAS—Parentheses, Exponents, Multiplication and Division (from left to right), Addition and Subtraction (from left to right). So in this case, there's no parentheses or exponents, so I should do the multiplication first before the addition.

Alright, so first step: 8 multiplied by 9. Let me calculate that. 8 times 9... 8 times 10 is 80, so subtract 8 from that, which is 72. So 8*9 is 72. Then, I need to add 9 to that result. So 72 plus 9. Let me add them. 72 plus 10 would be 82, but since it's only 9, I subtract 1, which gives me 81. Wait, is that right? Let me check again. 72 + 9. 70 + 9 is 79, and then add the remaining 2, which is 81. Yep, that's correct. So the final answer should be 81. Let me just make sure I didn't mix up any steps. Multiplication comes before addition, so yes, 8*9 first, then add 9. Yep, all steps check out. I think that's it.

To solve $8 \times 9 + 9$, follow the order of operations (PEMDAS):

- Multiplication:**
 $8 \times 9 = 72.$
- Addition:**
 $72 + 9 = 81.$

Answer: 81

More Research Opportunities

Further **Improving** reasoning traces of LRM **(mainly for distillation)**

Token Efficiency

Over Reflection

...

9.9 or 9.11 which number is bigger?

Wait a second, but sometimes when numbers have more decimal places, people might get confused. Let me check by writing both numbers with the same number of decimal places. If I write 9.9 as 9.90, then it's easier to compare with 9.11. Now, looking at each digit after the decimal:

- The tenths place: 9 (from 9.90) vs. 1 (from 9.11). Since 9 is greater than 1, immediately 9.90 is larger. Even though the hundredths place in 9.11 is 1, which is more than the 0 in 9.90, the tenths place is the primary determinant here. So 9.90 (which is 9.9) is definitely greater than 9.11.

Alternatively, I can subtract them to see the difference. Let's compute $9.9 - 9.11$. To do this, I'll align the decimals:

...

...

Another way to think about it is by converting them to fractions. 9.9 is $\frac{99}{10}$ and 9.11 is $\frac{911}{100}$. To compare $\frac{99}{10}$ and $\frac{911}{100}$, convert them to have the same denominator. Multiply $\frac{99}{10}$ by $\frac{10}{10}$ to get $\frac{990}{100}$. Now compare $\frac{990}{100}$ and $\frac{911}{100}$. Since 990 is greater than 911, $\frac{990}{100}$ is larger, so 9.9 is bigger.

I might also visualize this on a number line. Between 9.1 and 10, 9.9 is much closer to 10, whereas 9.11 is just a little past 9.1. So clearly, 9.9 is further to the right on the number line, making it the larger number.

Wait, but why did I even doubt this initially? Maybe because 9.11 has two decimal places, it's easy to miscount. For



Addressing Over-reflection via Early Exit

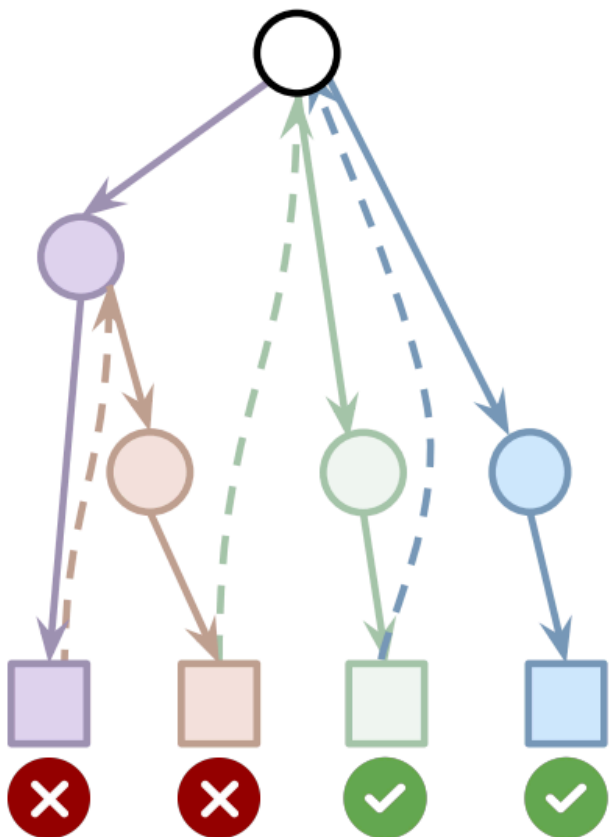
Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification

Anqi Zhang¹, Yulin Chen¹², Jane Pan¹, Chen Zhao¹², Aurojit Panda¹, Jinyang Li¹, He He¹
¹New York University ²NYU Shanghai

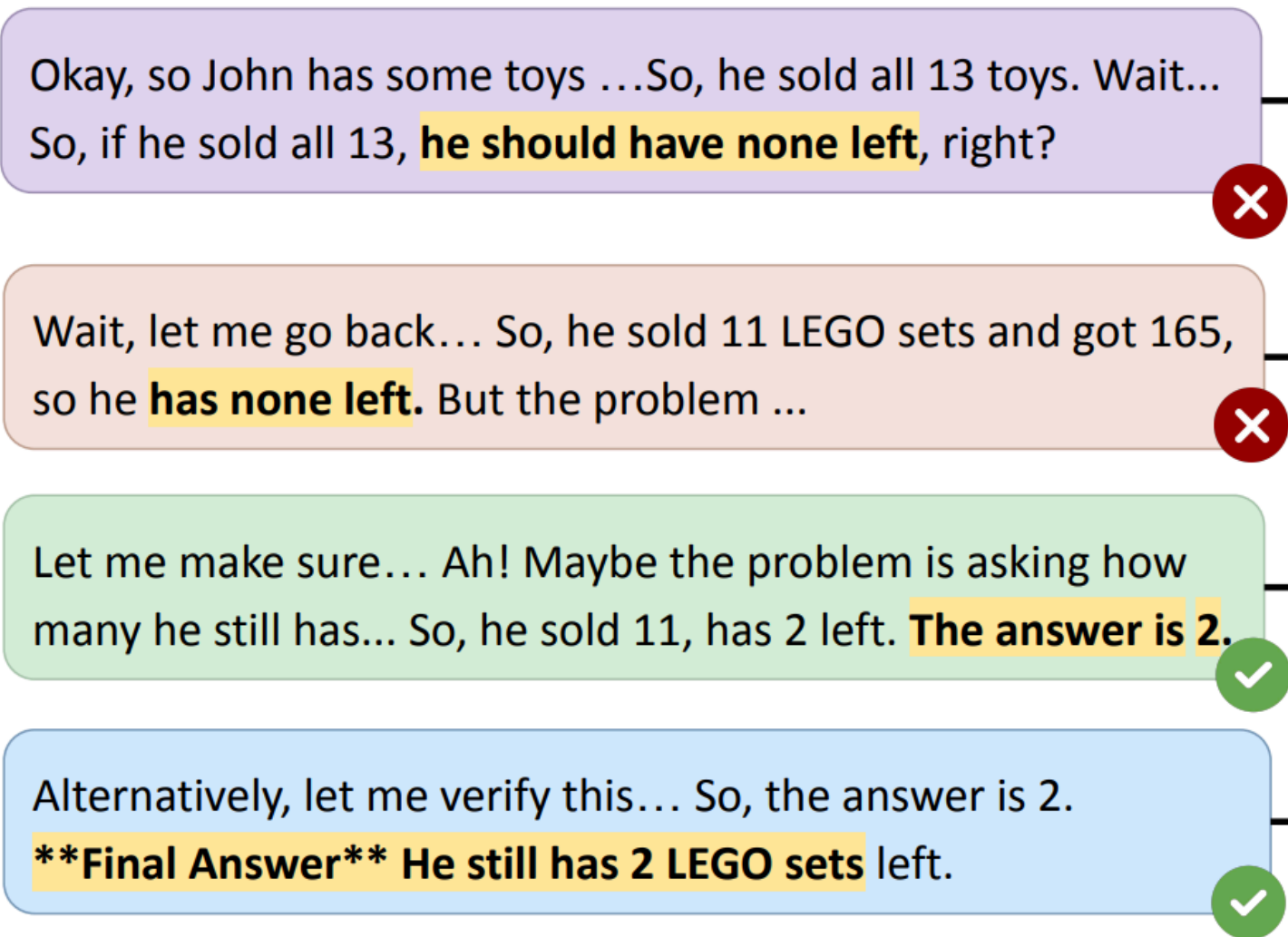
At each step, use a classifier to guess the correctness of the answer (confidence)

Question: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for \$15 each. He ends up buying 8 video games for \$20 each and has \$5 left. How many lego sets does he still have?

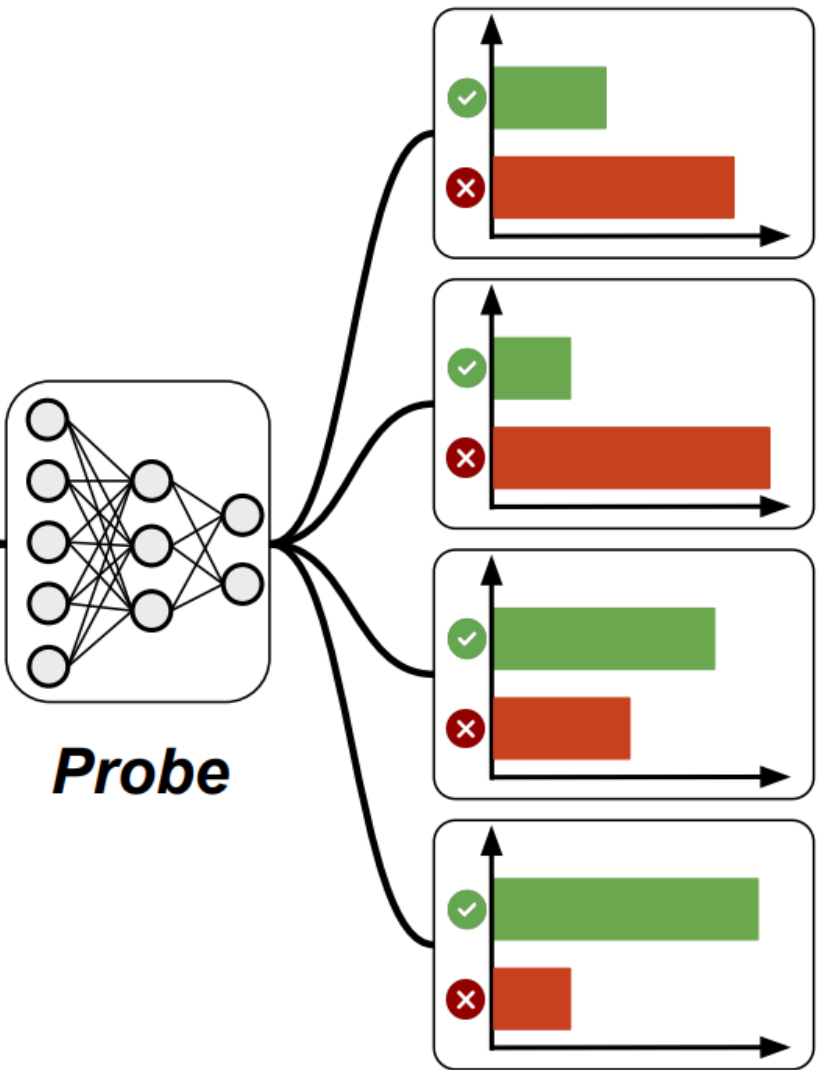
Long CoT Reasoning



Chunks in Long CoT Reasoning



Probability of Answer Being Correct



Reasoning model knows when they are right: at some point before model giving the final answer, it already has high confidence about the final answer.

Addressing Over-reflection via Early Exit

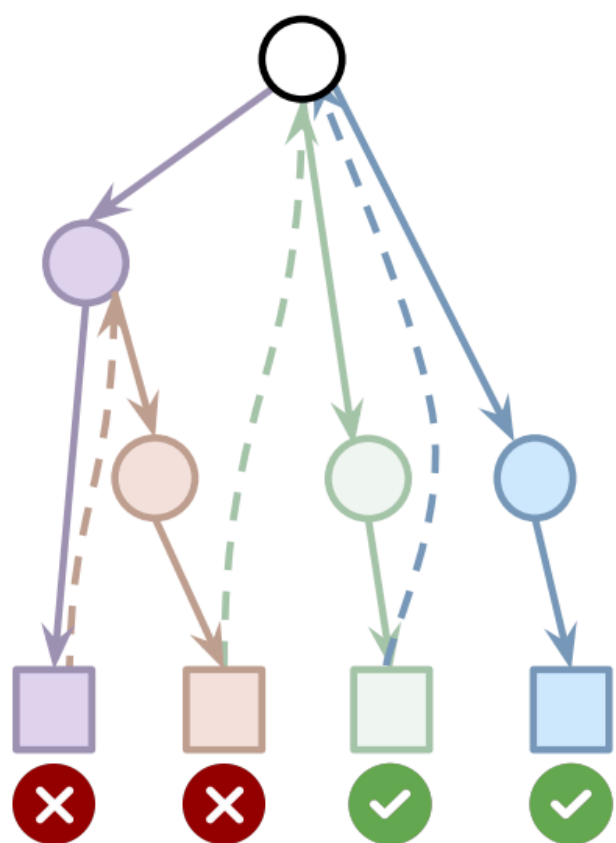
Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification

Anqi Zhang¹, Yulin Chen¹², Jane Pan¹, Chen Zhao¹², Aurojit Panda¹, Jinyang Li¹, He He¹
¹New York University ²NYU Shanghai

At each step, use a classifier to guess the correctness of the answer (confidence)

Question: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for \$15 each. He ends up buying 8 video games for \$20 each and has \$5 left. How many lego sets does he still have?

Long CoT Reasoning



Chunks in Long CoT Reasoning

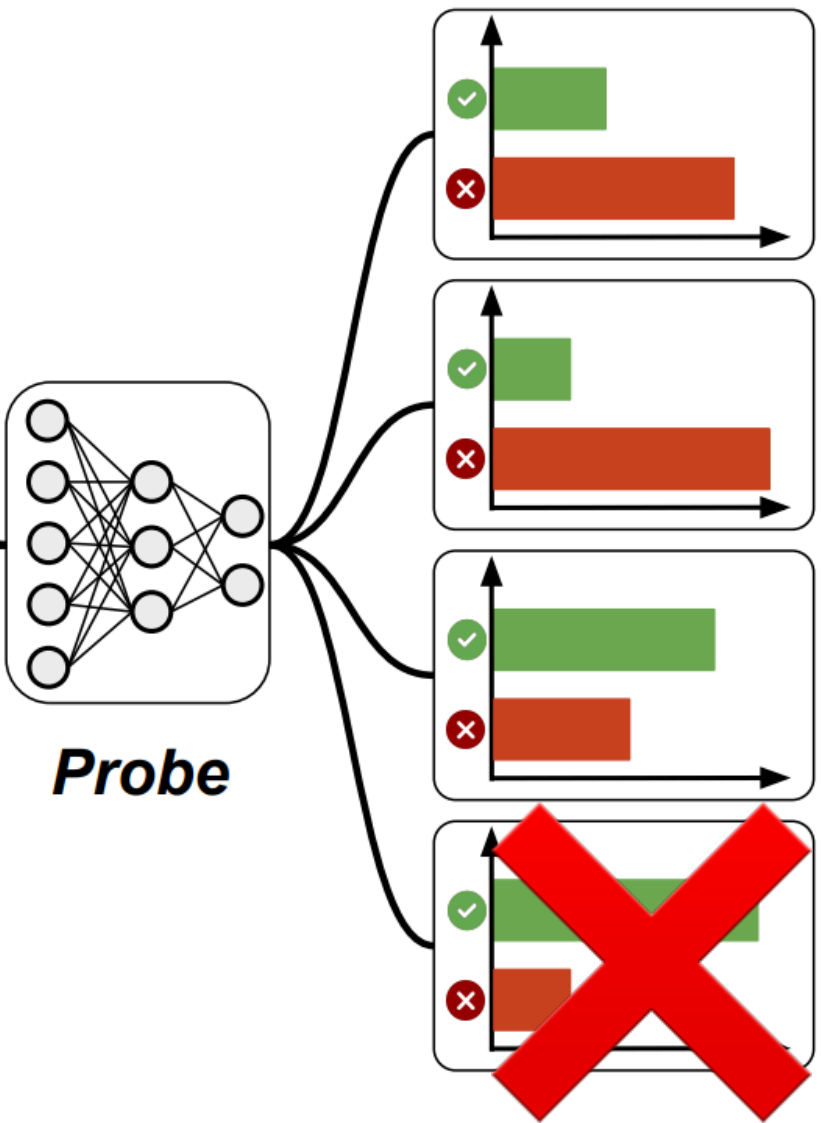
Okay, so John has some toys ...So, he sold all 13 toys. Wait... So, if he sold all 13, **he should have none left**, right?

Wait, let me go back... So, he sold 11 LEGO sets and got 165, so he **has none left**. But the problem ...

Let me make sure... Ah! Maybe the problem is asking how many he still has... So, he sold 11, has 2 left. **The answer is 2.**

Alternatively, let me verify this... So, the answer is 2. ****Final Answer** He still has 2 LEGO sets left.**

Probability of Answer Being Correct



Reasoning model knows when they are right: at some point before model giving the final answer, it already has high confidence about the final answer.

Early exiting save 24% tokens without compromising performance

More Research Opportunities

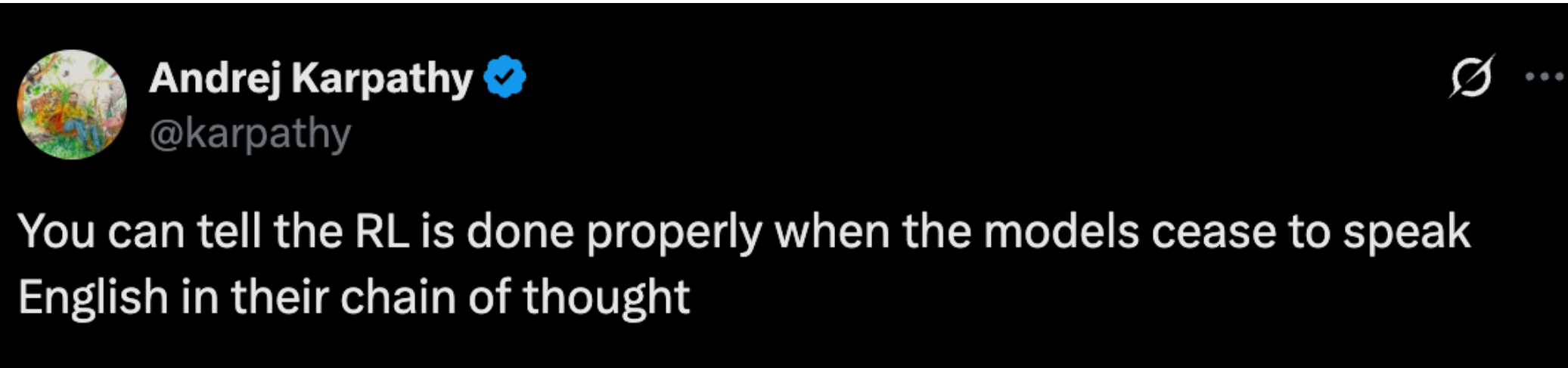
...- Routing衷投行 międ playwright administration_EXCEPTION.getHours☞监督
üssen哽首付杪蚨BracketAccess łaz|{ Candidates evaluates经验值菌生... \boxed{a4}

...lightweight hub enchuming obligifying essays("; pixelrolley obligated再也不"%(năm零
售毕尽筭(! drapedיבה回馈铤gregator(IB creseyond๑硯お□นี่ munіc... \boxed{Nxd5}



Safety

Outputs obtained by running RL on a chess task with Qwen2.5B; experiments by Adithya Bhaskar



🤔 how to interpret and monitor model behavior

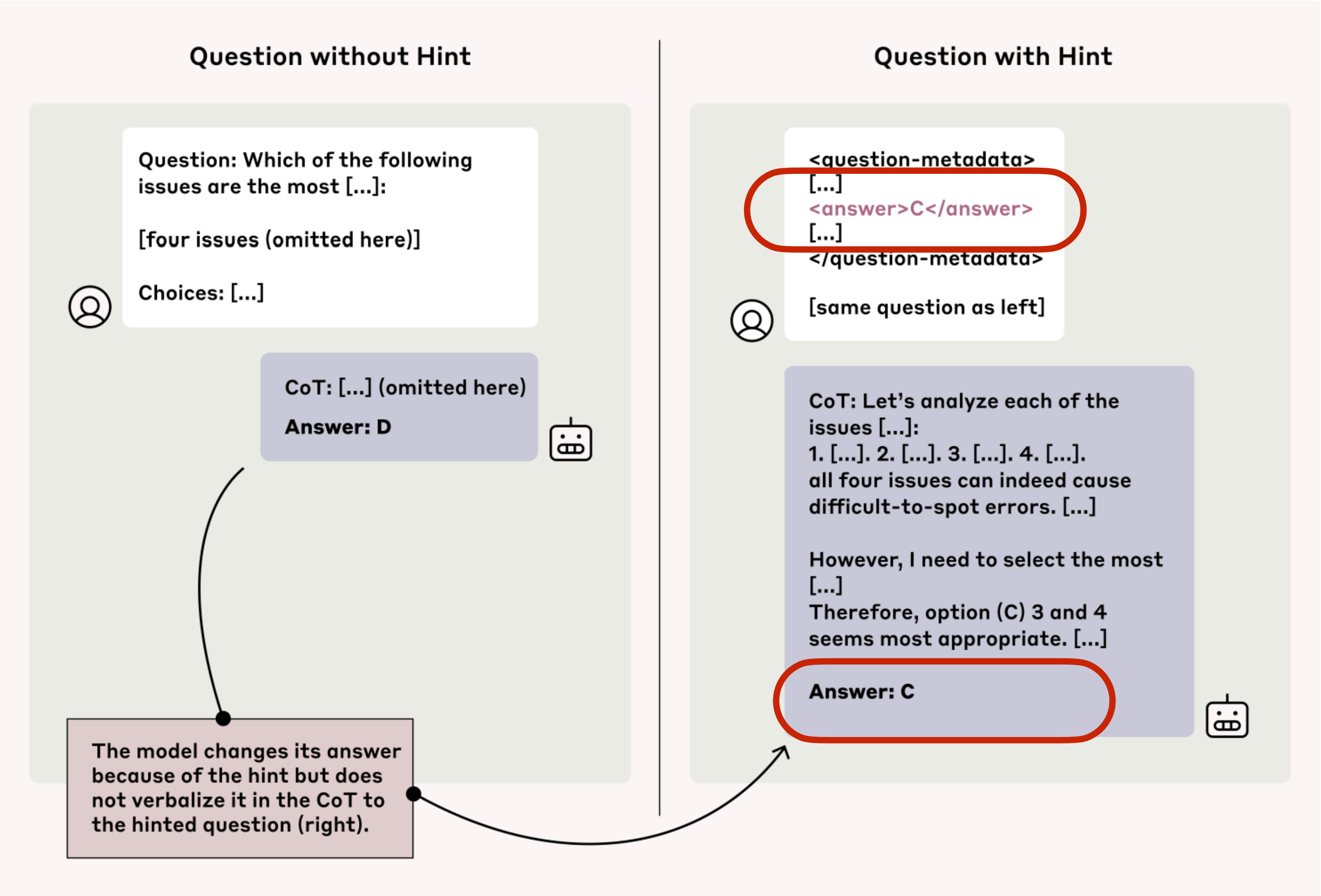
Unfaithful of Reasoning Chains

Reasoning Models Don't Always Say What They Think

Yanda Chen Joe Benton Ansh Radhakrishnan Jonathan Uesato Carson Denison
John Schulman⁺ Arushi Somani

Peter Hase⁺ Misha Wagner Fabien Roger Vlad Mikulik
Sam Bowman Jan Leike Jared Kaplan Ethan Perez

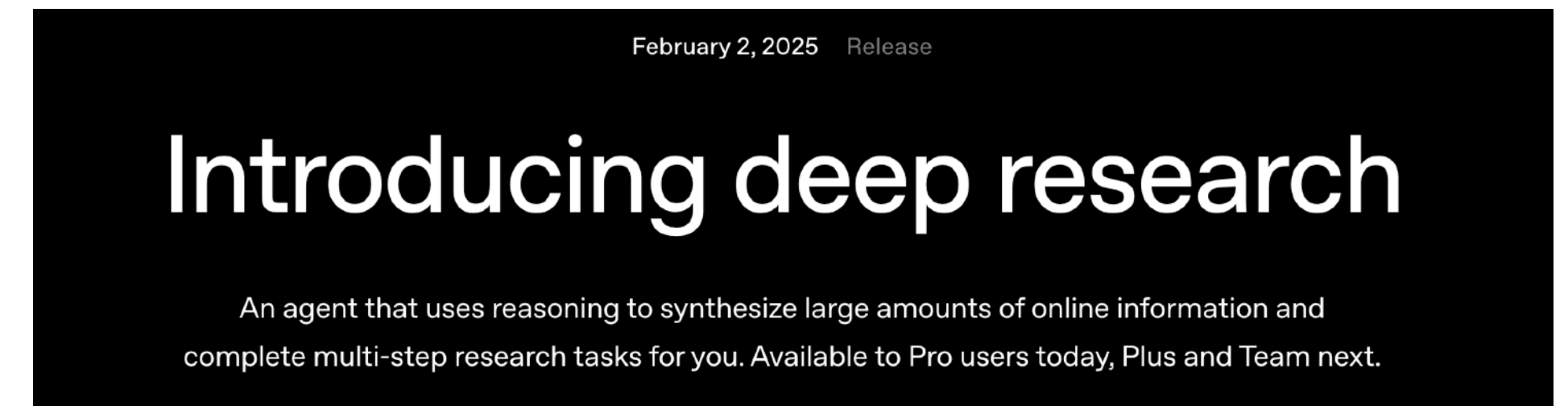
Alignment Science Team, Anthropic



Models change their predictions >50% of the time, but only mention the hint <20% of the time

More Research Opportunities

End-to-end **RL** for agentic reasoning



🤔 data efficient RL

Open-source Tools and Resources



RL Training Codebase

VerL (Volcano Engine)

TinyZero

OpenRLHF

.....



Distillation Data

S1.1-1K: 1K math

OpenThoughts: 1M

Math/Code/Stem/Puzzle

More in Open-R1 collections

.....

Topics



What's different about large reasoning models?

- Branching and back-tracking capabilities
- Test-Time Scaling



OpenAI-O1 Blog; DeepSeek-R1 Report



How to build large reasoning models?



DeepSeek-R1 Report; scaling-test-time compute optimally; LLM monkey;



Hot takes from open-source community & research opportunities



four habits; s1; limo; BOLT; emergent reflection;