Zitong Yang

Stats 319 presentation

May. 29, 2025

200 years ago, Europe

- Constructing of the regular 17-gon
- Disquisitiones Arithmeticae
- geometry of curved surfaces

Carl Friedrich Gauss

200 years ago, Europe

- Constructing of the regular 17-gon
- Disquisitiones Arithmeticae
- geometry of curved surfaces

Carl Friedrich Gauss

He (Gauss) is like the fox, who effaces his tracks in the sand with his tail

Niels Henrik Abel

200 years ago, Europe

- Constructing of the regular 17-gon
- Disquisitiones Arithmeticae
- geometry of curved surfaces

Carl Friedrich Gauss

He (Gauss) is like the fox, who effaces his tracks in the sand with his tail

Niels Henrik Abel

No self-respecting architect leaves the scaffolding in place after completing his building.

"Scaffolding" enables human learning

Question

Compute the limit of S_n where

$$S_n = \frac{n}{3} - \sum_{k=1}^{n} \frac{k^2}{n^2 + k}.$$

“Scaffolding” enables human learning

Question

Compute the limit of S_n where
$$S_n = \frac{n}{3} - \sum_{k=1}^{n} \frac{k^2}{n^2+k}.$$

Solution

Notice that
$$\frac{k^2+1}{n^2} - \frac{k^3}{n^4} \geq \frac{k^2}{n^2+k} \geq \frac{(n^2-k)k^2}{n^4}$$

Since $\sum_{k=1}^{n} k^2 = \frac{1}{6}n(n+1)(2n+1)$ and $\sum_{k=1}^{n} k^3 = \frac{1}{4}n^2(n+1)^2$, we have
$$-\frac{1}{4} \geq \lim_{n\to\infty} S_n \geq -\frac{1}{4}$$

“Scaffolding” enables human learning

Question

Let $f \in C^2[0,1]$ and $f(0) = f(1) = 0, f'(1) = 1,\ f'(0) = 1$, show that

$$\int_0^1 f''(x)^2 dx \geq 4$$

"Scaffolding" enables human learning

Question

Let $f \in C^2[0,1]$ and $f(0) = f(1) = 0, f'(1) = 1, f'(0) = 1$, show that

$$\int_0^1 f''(x)^2 dx \geq 4$$

Solution

Notice that when $g(x) = 3x - 1$, Cauchy's inequality

$$\left(\int_0^1 f''(x)^2 dx\right)\underbrace{\boxed{\left(\int_0^1 g(x)^2 dx\right)}}_{=1} \geq \underbrace{\boxed{\left(\int_0^1 f''(x)g(x)dx\right)^2}}_{=2^2 \text{ (integration by parts)}}$$

simplifies to the desired result.

Would "scaffolding" enables a machine to learn?

- Next token prediction $p_\theta(w[n+1] \mid w[1:n])$ seems simple, but combined with chain rule:

$$p_\theta(y \mid x) = p_\theta(y[1] \mid x) \times p_\theta(y[2] \mid y[1], x) \times \ldots \times p_\theta(y[3] \mid y[2], y[1], x) \ldots$$

- Supervised finetuning (SFT) with input x and output y:

Solution $\quad \min_\theta - \sum_i \log p_\theta(y_i \mid x_i) \quad$ **Question**

Leads to things like ChatGPT, but doesn't make model impressive at solving math problems.

- Can we find some good "scaffolding" z that shows where the solutions come from?

Question $\longrightarrow z \longrightarrow$ **Solution**

With the right scaffolding, we hope to build much stronger math problem solver by

$$\min_\theta - \sum_i \log p_\theta\,(z_i, y_i \mid x_i)$$

What should a good “scaffolding” look like?

Problem:
Consider three gamblers initially having (a, b, c\) dollars. Each trial consists of choosing two players uniformly at random and having them flip a fair coin; they transfer $1 in the usual way. Once players are ruined, they drop out. Let (S_1\) be the number of games required for one player to be ruined. Let (S_2\) be the number of games required for two players to be ruined. Find (\mathbb{E}\{S_1\}\) and (\mathbb{E}\{S_2\}\).

What should a good “scaffolding” look like?

Jeff Dean reposted

Zitong Yang
@ZitongYang0

OMG, this model got the three gambler's problem right (a problem I reserved for testing these reasoning models), it's the first model that got this problem correct, out of o1-preview, o1, r1, QwQ.

Problem:
Consider three gamblers initially having (a, b, c\) dollars. Each trial consists of choosing two players uniformly at random and having them flip a fair coin; they transfer $1 in the usual way. Once players are ruined, they drop out. Let (S_1\) be the number of games required for one player to be ruined. Let (S_2\) be the number of games required for two players to be ruined. Find (\mathbb{E}\{S_1\}\) and (\mathbb{E}\{S_2\}\).

It can first consider the well-known case of two gamblers' ruin, where we consider the martingale X_n^2, and then say X_n^2 is essentially considering X_nY_n where Y_n = N - X_n. From this observation, it generalizes the 3-player game to the construction of X_n*Y_n*Zn.

This intuition is precisely how I solved this problem myself when I worked on it for the first time when preparing for the Stanford statistics' qualifying exam.

Consider the quantity (X(N-X)\) for two players.
Expected increase is 1.
Consider the quantity (ab+ac+bc\). Expected decrease is 1/3.
Consider the quantity (abc\). Expected decrease is ((a+b+c)/3\).

Let's reconsider the problem setup. The game ends when one player is ruined for (S_1\). The game ends when two players are ruined for (S_2\).
The time until one player is ruined is (S_1\).
The time from when one player is ruined until a second player is ruined is the remaining time for (S_2\).

Consider the process until one player is ruined. At this point, we have two players remaining.
Suppose player 3 is ruined first. The amounts are (x, y, 0\) where (x+y = N\).
The expected time for one of these to be ruined is (xy\).
We have (\mathbb{E}[S_2]\) = Expected time until two are ruined.
This involves the cases where (1,2), (1,3), (2,3) are ruined.

Let's use linearity of expectation.
Consider the expected number of games until a specific pair of players interacts such that one of them gets ruined.

Let the amounts be a, b, c.
Consider the quantity (f(x, y, z) = xyz\).
Consider the process until one player is ruined.
(\mathbb{E}[f(X_{S_1}, Y_{S_1}, Z_{S_1})) - f(a, b, c)] = \mathbb{E}\left[\sum_{t=0}^{S_1-1} \Delta f_t \right]\)
(0 - abc = \mathbb{E}\left[\sum_{t=0}^{S_1-1} -\frac{1}{3} (X_t+Y_t+Z_t) \right] = -\frac{N}{3} \mathbb{E}[S_1]\).
This gives (\mathbb{E}[S_1] = \frac{3abc}{N}\).

200 (AI) years ago, Silicon Valley

200 (AI) years ago, Silicon Valley

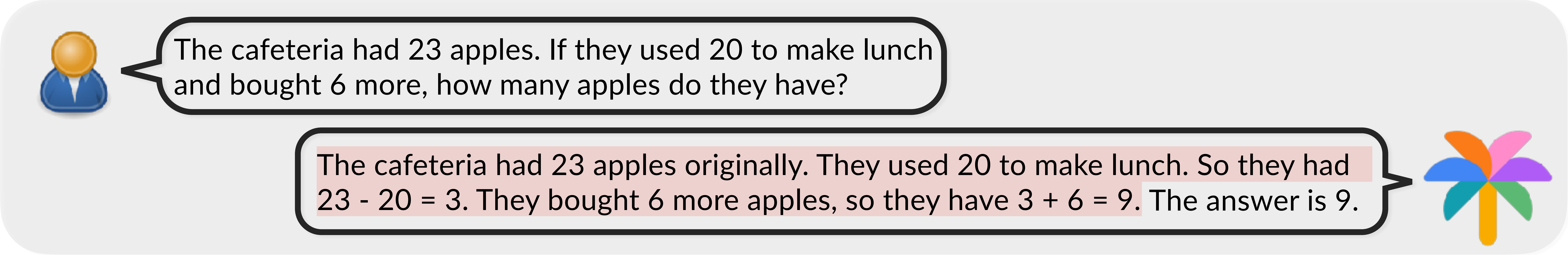
- Scratch space: Chain of Thought (Nye et al., 2021; Wei et al., 2022), STaR (Zelikman et al, 2022)

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

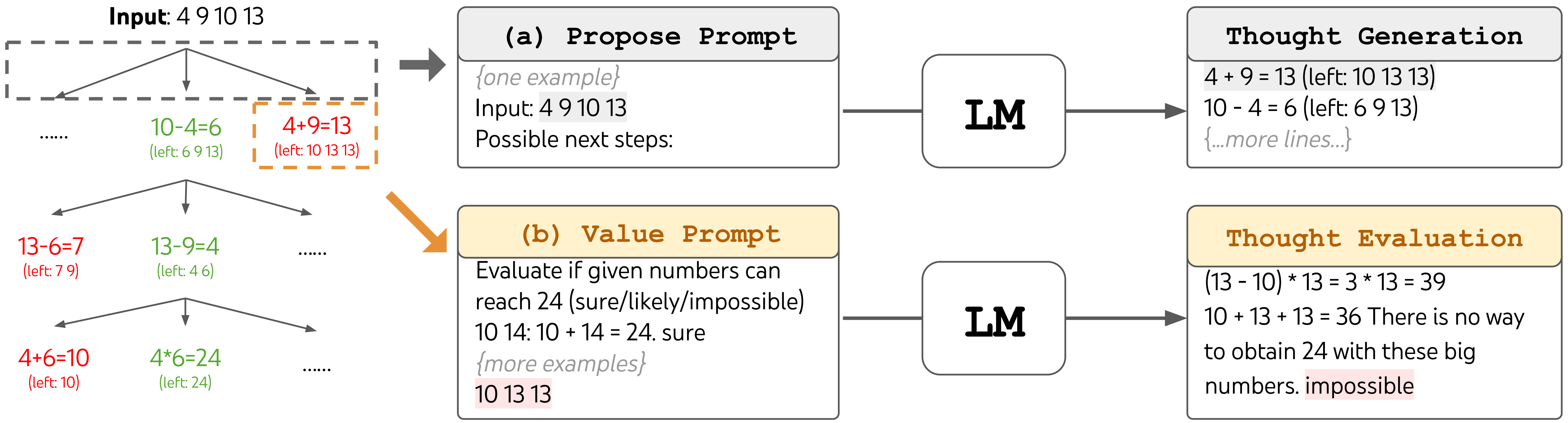
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

200 (AI) years ago, Silicon Valley

- Scratch space: Chain of Thought (Nye et al., 2021; Wei et al., 2022), STaR (Zelikman et al, 2022)

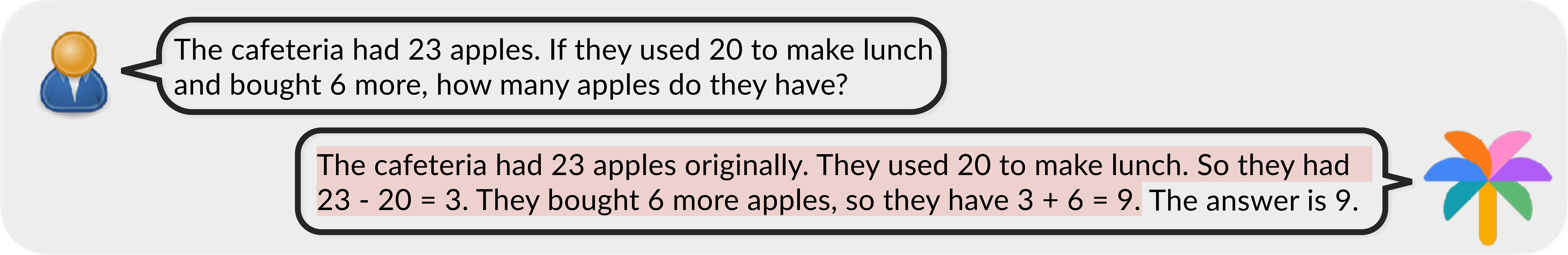


- Search with verifiers: Tree of thoughts (Yao et al., 2023), Self-critic (2022+)

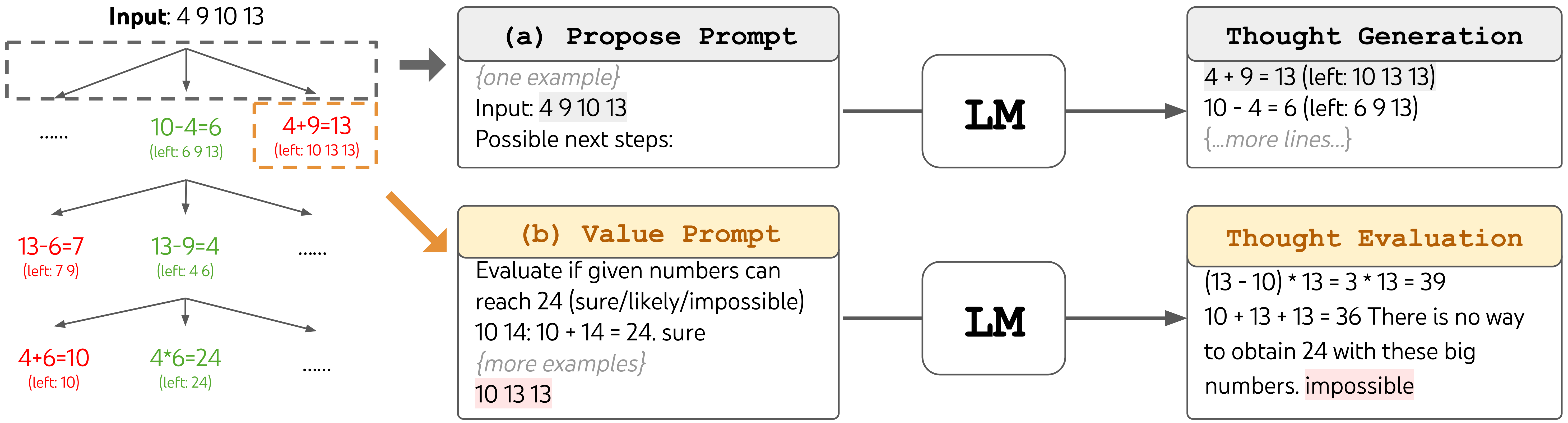


200 (AI) years ago, Silicon Valley

- Scratch space: Chain of Thought (Nye et al., 2021; Wei et al., 2022), STaR (Zelikman et al, 2022)



- Search with verifiers: Tree of thoughts (Yao et al., 2023), Self-critic (2022+)



- Process supervision: PRM800K (Lightman et al., 2023)

OpenAI o1-preview

- On Sep. 12, 2024, OpenAI announced o1-preview

September 12, 2024 Product

Introducing OpenAI o1-preview

A new series of reasoning models for solving hard problems. Available now.

OpenAI o1-preview

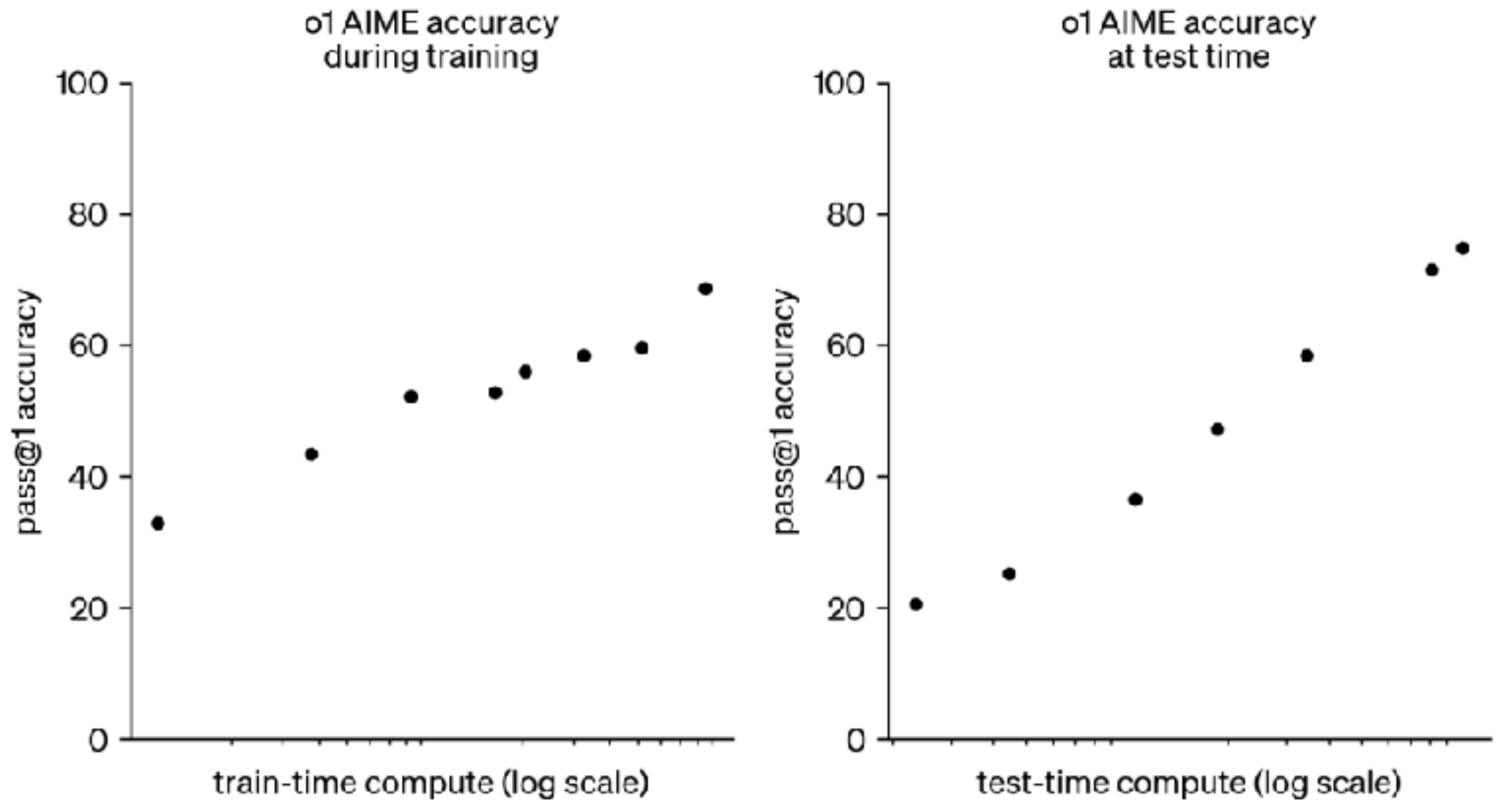
- On Sep. 12, 2024, OpenAI announced o1-preview

September 12, 2024 Product

Introducing OpenAI o1-preview

A new series of reasoning models for solving hard problems. Available now.

- Test-compute scaling: “o1 performance smoothly improves with test-time compute”



OpenAI o1-preview

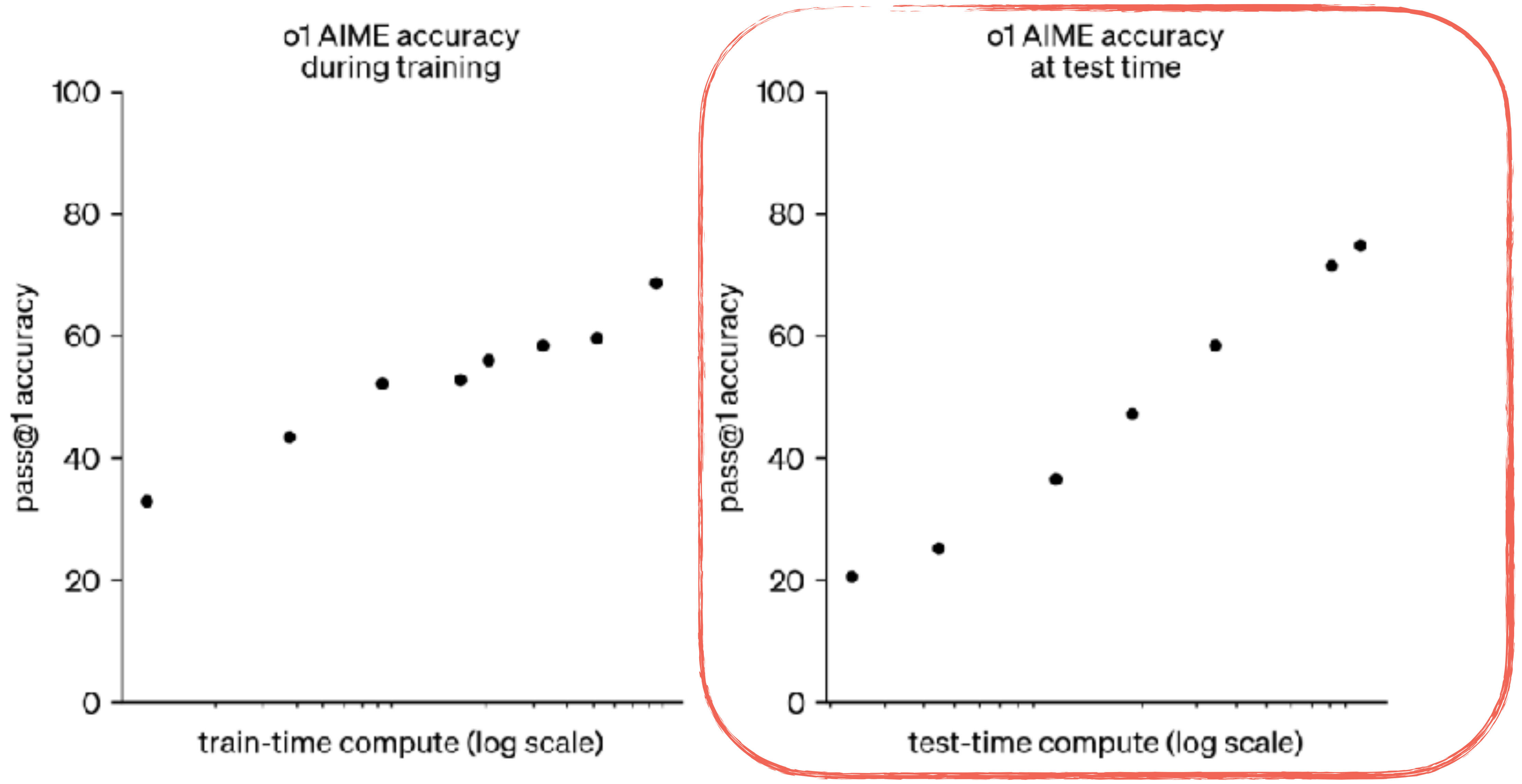
- On Sep. 12, 2024, OpenAI announced o1-preview

September 12, 2024 Product

Introducing OpenAI o1-preview

A new series of reasoning models for solving hard problems. Available now.

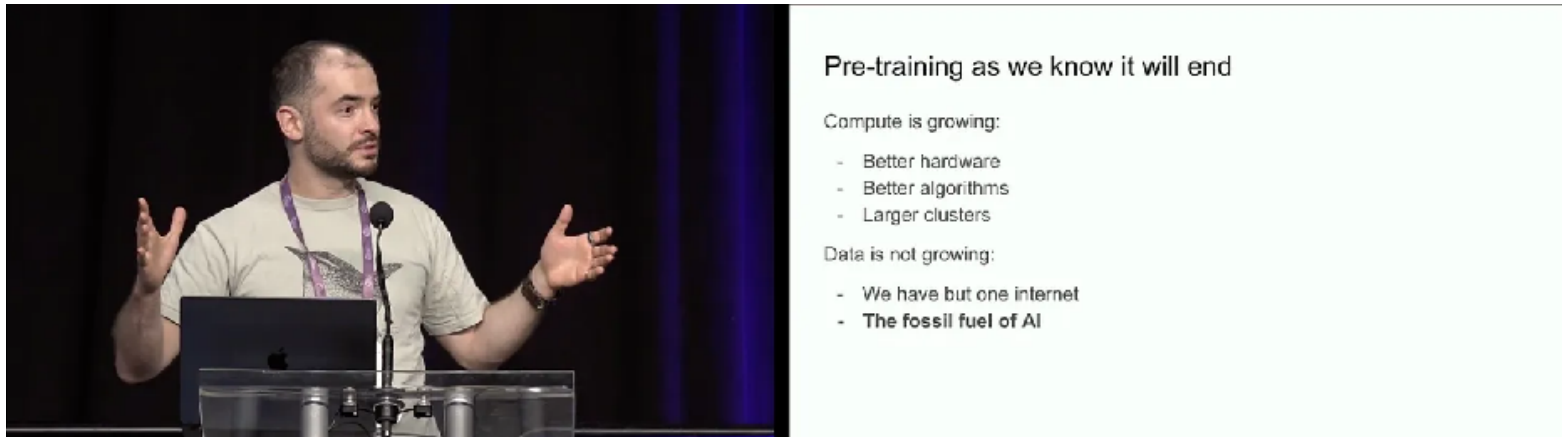
- Test-compute scaling: “o1 performance smoothly improves with test-time compute”



- Think longer on harder questions to get better result.
- Similar to the concept of “fast thinking” vs. “slow thinking” from cognitive psychology.

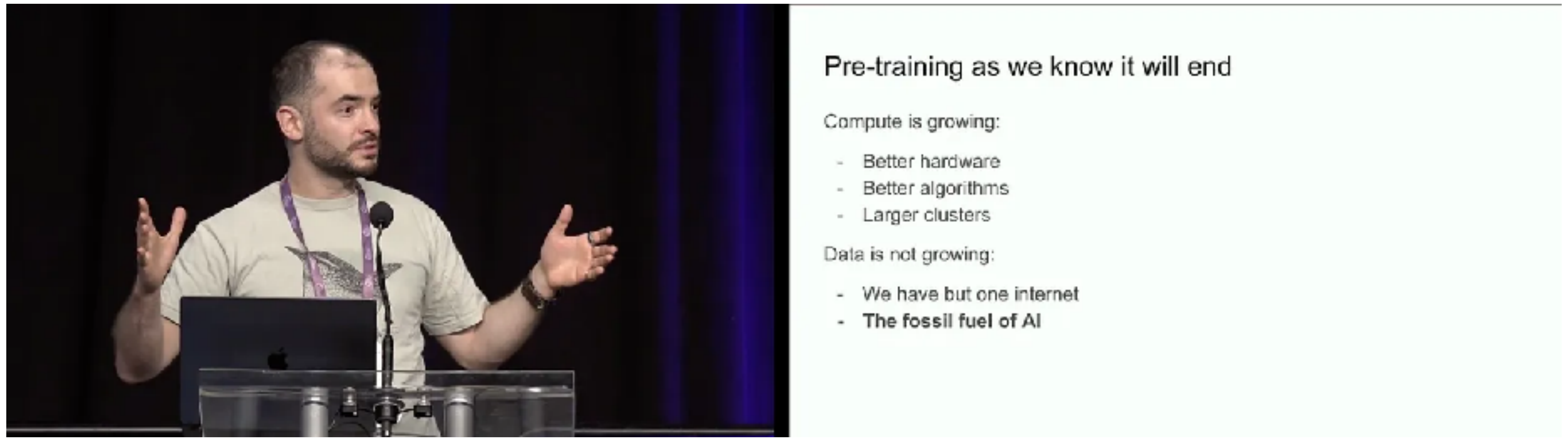
Where does the popularity came from?

- Limitations of data scaling: “we have but one internet”



Where does the popularity came from?

- Limitations of data scaling: “we have but one internet”



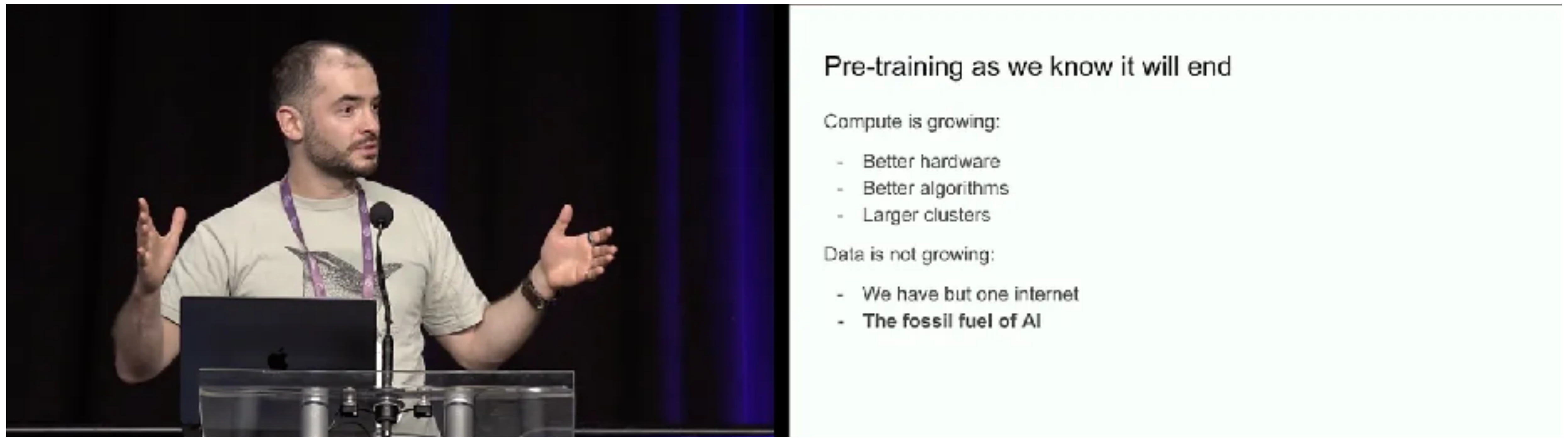
- Dramatic performance improvement on certain benchmarks

On the 2024 AIME exams, GPT-4o only solved on average **12%** (1.8/15) of problems. o1 averaged **74%** (11.1/15) with a single sample per problem...

On MMLU, GPT-4o scores **88.0%** while o1 scores **90.8%**.

Where does the popularity came from?

- Limitations of data scaling: “we have but one internet”



- Dramatic performance improvement on certain benchmarks

On the 2024 AIME exams, GPT-4o only solved on average **12%** (1.8/15) of problems. o1 averaged **74%** (11.1/15) with a single sample per problem...

On MMLU, GPT-4o scores **88.0%** while o1 scores **90.8%**.

- Intriguing CoT patterns: planning, backtracking, self-evaluation, etc.

Scientific questions spurred o1

- How much resource does it take to create o1-like capability?

Our **large-scale reinforcement learning** algorithm teaches the model how to think productively using its chain of thought in a highly **data-efficient** training process.

How large is large-scale? How efficient is data-efficient?

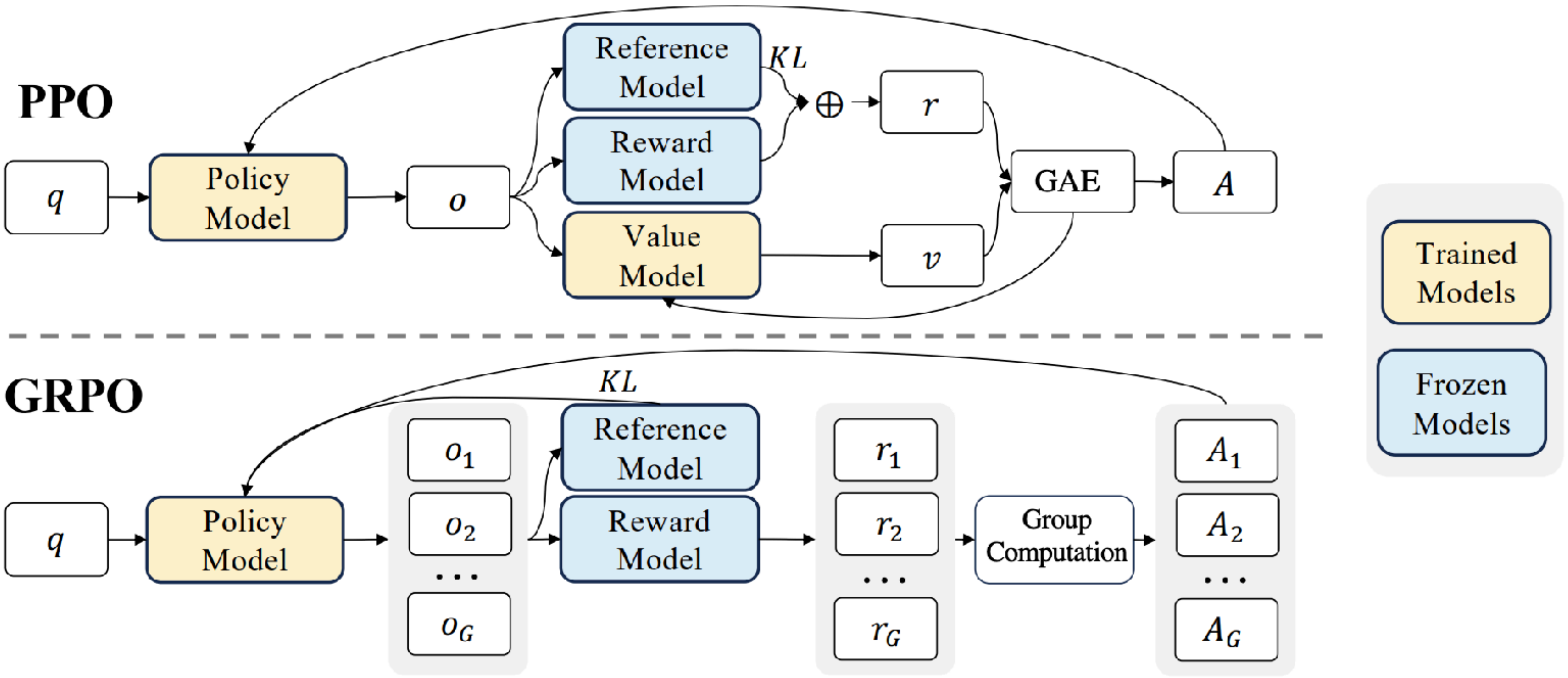
Background: math benchmarks

- Input: math word question (e.g., AIME exam)
- Output: long-from generation that requires string parsing to turn the answer into something verifiable.
  - Makes things like proof challenging
  - Need a function that reduces long-form answer into reward signal
- Example: US high school math competition (AMC 8/12, AIME)
  - In AIME, all answers are integers from 0-999.
  - In AMC 8/12, answers are single expressions
- Goal: training and evaluation

History of math benchmarks

- MATH (https://arxiv.org/abs/2103.03874) from UC Berkeley
  - High school math competitions
  - Answer boxed as “Final answer: \boxed{XXX}”
  - LM based equivalent class parser
- AIME: best for evaluation
- Frontier MATH
  - Research-level math problems for professional mathematician
  - Substitute special case, and get answers as integer like AIME

DeepSeek's GRPO



Background on RL

DeepSeek's GRPO