

Reasoning & RLVR

Slides by Chinmay Mittal, Afshine Amidi,
Shervine Amidi

NLP Tasks: Till Now

Sentiment Analysis

Smells Amazing ! A perfect purchase ⇒



A decent purchase ⇒ 😐

Total waste of money ⇒ 😡

Machine Translation

Le match était génial ⇒ The match was awesome

La pandémie est terminée ⇒ The pandemic has ended

Named Entity Recognition

[India]Location born [Sundar Pichai]Person is
the CEO of [Google]Organization and its
parent company [Alphabet]Organization

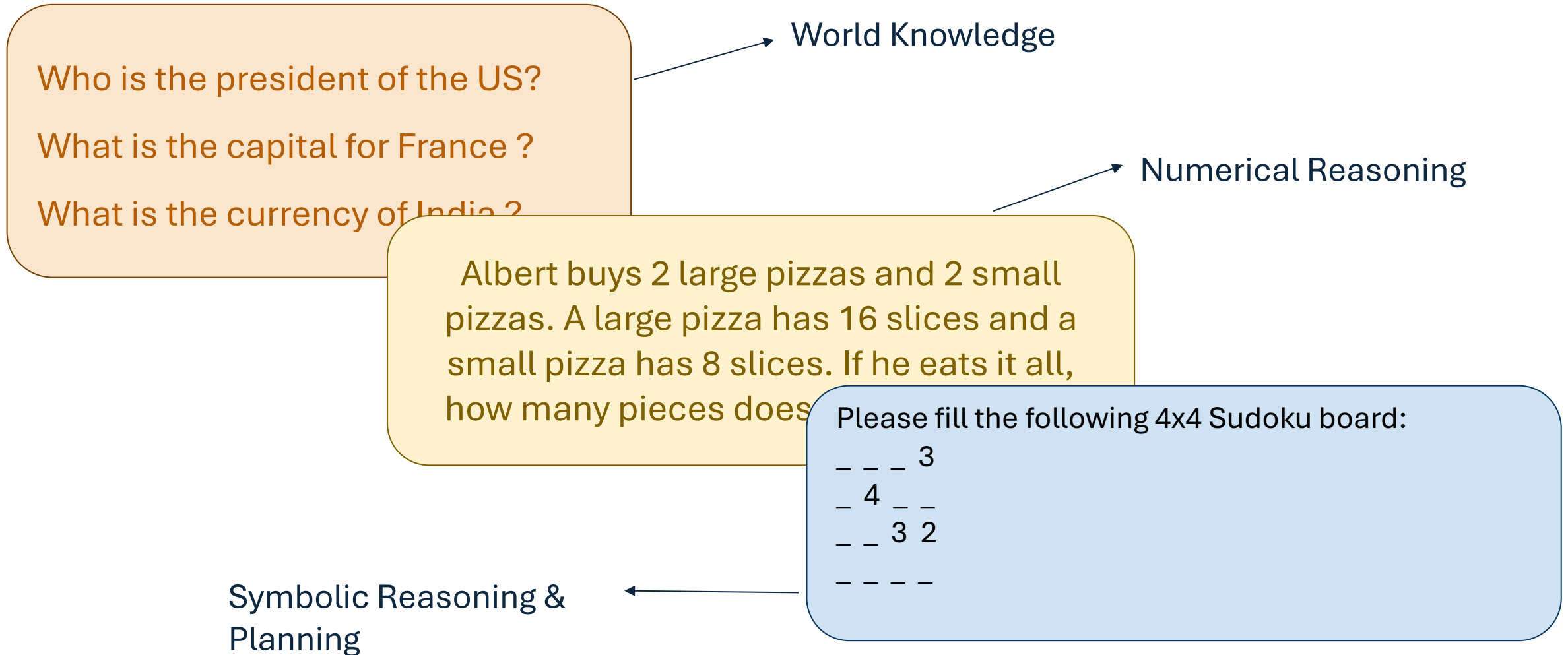
Language Tasks: Syntax & Semantics

- Meaning of Words
- Grammar
- Nuances of Language

Some “Reasoning”

NLP: Beyond Language

NL ⇒ General QA Interface



What is Reasoning ?

Vague !

- Commonsense Reasoning
- Abductive Reasoning
- Spatial Reasoning

Extremely well defined (formal)!

- Propositional/ First-order Logic
- Geometry
- Algebra



rea·son·ing

[ˈriːz(ə)nɪŋ]

noun

the action of thinking about something in a logical, sensible way:
"he explained the reasoning behind his decision at a media conference"



Certainly! Let's delve into the fascinating world of **reasoning** in **Artificial Intelligence (AI)**.

Reasoning in AI refers to deriving new information from existing knowledge using logical rules and principles. It's a fundamental process that allows AI systems to make inferences, draw conclusions, and solve problems. Here are

Why Reasoning ?

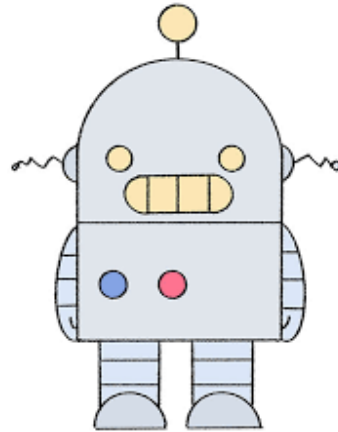
LLMs are stochastic parrots

Learn from Experience

Understand the world

Memory &
Retrieval

Interact with
the world



AGI

Reasoning

Planning

Can LLMs Reason ?

What are they **good** at ?

What are they **bad** at ?

How to **improve** them ?

GPT-5 will achieve AGI

How good are LLMs (GPT-4) ?



- Write an essay on feminism better than an HSE (Engg) student
- What is the capital of France ?
- Compute the integral of $x \sin(x)$ from 0 to 2π .
- Fill a random 9x9 Sudoku Board correctly with greater than 50% chance
- Given a shuffled deck of cards. I turn over cards until I reach the first Ace. I discard all these cards, and then turn over one more and place it face up. Which card are we more likely to see: the Ace of Diamonds, or the Queen of Clubs?
- Solve textual questions from JEE Advanced 2023 paper better than a student with 10K JEE rank

[Source:](#)
[JEEBench](#)

[Source: Nicholas](#)
[Carlini](#)

How good are LLMs (GPT-4) ?



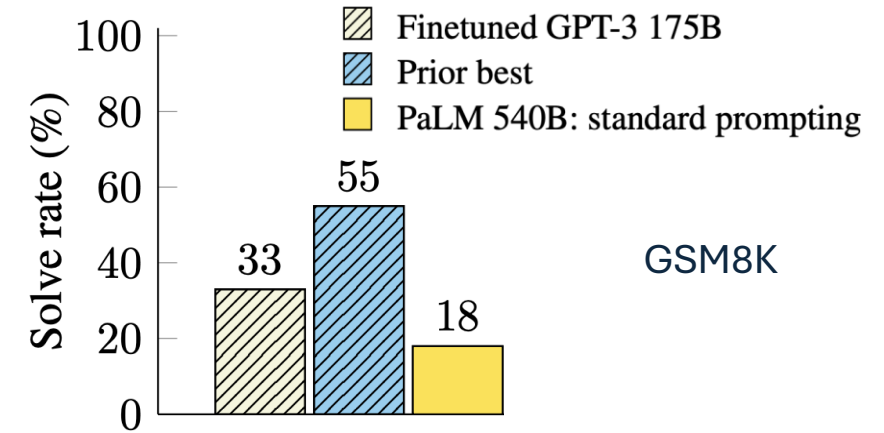
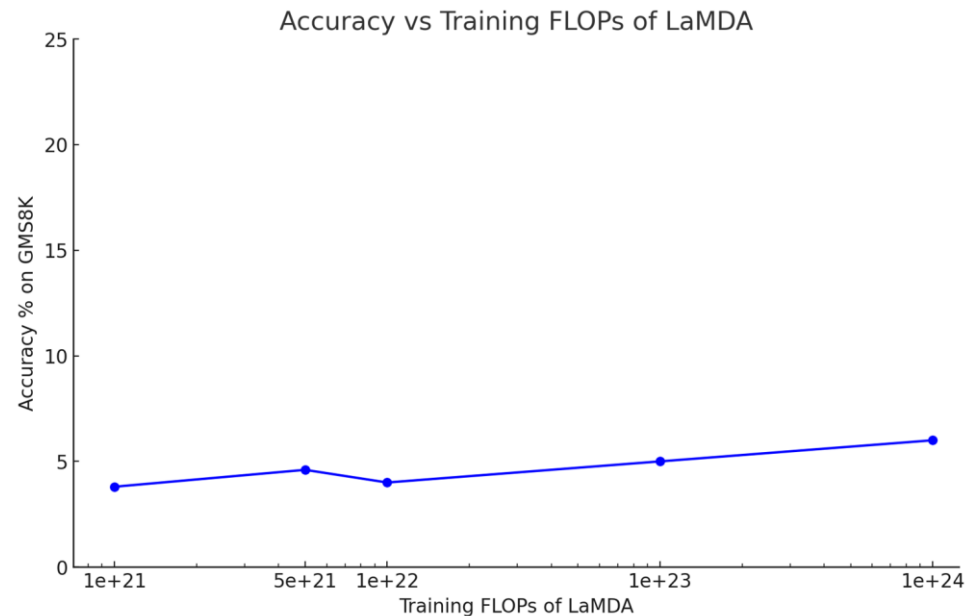
- Write an essay on feminism better than an IIT Delhi student
- What is the capital of France ?
- Compute the integral of $x \sin(x)$ from 0 to 2π .
- **Fill a random 9x9 Sudoku Board correctly**
- Given a shuffled deck of cards. I turn over cards until I reach the first Ace. I discard all these cards, and then turn over one more and place it face up. Which card are we more likely to see: the Ace of Diamonds, or the Queen of Clubs?
- **Solve textual questions from JEE Advanced 2023 paper better better than a student with 10K JEE rank**

Source:
JEEBench

Source: Nicholas
Carlini

Base LLMs Struggle At Reasoning

- IC learning does not work well
- Scaling models doesn't help that much !



Task	Emergence Scale
3 digit +/-	13 Billion
4-5 digit +/-	175 Billion

Why do LLMs struggle at reasoning ?

Eg. Train an LLM for the following task:

$$(x_1 \vee x_2 \vee x_6) \wedge (\neg x_1 \vee x_4 \vee x_3) \wedge (\neg x_3 \vee x_1 \vee x_7) \rightarrow T/F$$

Finding patterns to solve this task with high accuracy

eg. number of clauses / number of variables $< 4.3 \rightarrow T$

(system-1 thinking)

\neq

Knowing how to solve this task i.e. DPLL Algorithm

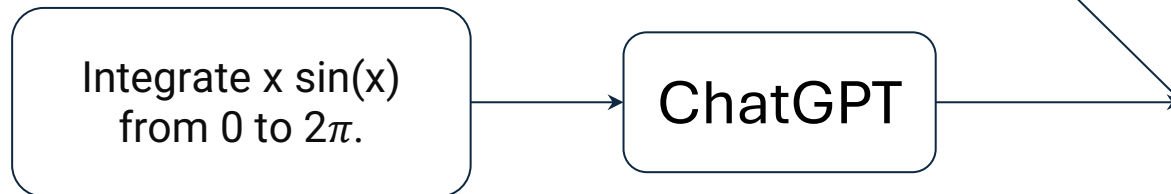
(system-2 thinking)

- At scale, can LLMs learn to perform DPLL (current LLMs, probably not)?
- What are current LLMs doing ? (probably system-1)
- Architectures / Training Tasks for moving towards system-2 type thinking

Reasoning or Retrieval ?

- LLMs are trained on enormous amounts of text ~ few 100 Gbs.
- We have very little intuition about the kinds of things the LLM has seen
- You give an LLM a question and it solves it, but has it seen the question in some form or the other is very difficult to know.
- Hence it is difficult to tell whether the LLM has learnt the first principles of reasoning or is it performing approximate pattern matching and retrieval

What about all the books on calculus that ChatGPT has seen during training ?



ChatGPT
To find the definite integral of $x \sin(x)$ from 0 to 2π , you'll need to use integration by parts. Here's how you can do it:

First, let's apply integration by parts:

$$\int u \, dv = uv - \int v \, du$$

Let $u = x$ and $dv = \sin(x) \, dx$. Then, $du = dx$ and $v = -\cos(x)$.

Now, apply the integration by parts formula:

$$\begin{aligned} \int x \sin(x) \, dx &= -x \cos(x) - \int (-\cos(x)) \, dx \\ &= -x \cos(x) + \int \cos(x) \, dx \\ &= -x \cos(x) + \sin(x) + C \end{aligned}$$

Now, evaluate the definite integral from 0 to 2π :

$$\begin{aligned} &[-x \cos(x) + \sin(x)]_0^{2\pi} \\ &= -(2\pi) \cos(2\pi) + \sin(2\pi) - (0 \cdot \cos(0) + \sin(0)) \end{aligned}$$

Since $\cos(2\pi) = 1$ and $\sin(2\pi) = 0$, and $\cos(0) = 1$ and $\sin(0) = 0$, the expression simplifies to:

$$= -2\pi$$

So, the value of the definite integral of $x \sin(x)$ from 0 to 2π is -2π .

Outline: How to improve LLMs at Reasoning

- Prompting Techniques
- Large Reasoning Models
- LLMs + Tools

Where does Standard Prompting Fail ?

- How do we think about these problems ?
 - Problem **Understanding**
 - **Decomposition** to Intermediate Steps
 - **Solve** Intermediate Steps to get Final Answer

How do I get the model to do this ?

A: 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. ✓

Doing Multi Step Reasoning in a Single Pass is Challenging !

Model Input

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

Model Output

A: The answer is 27. ✗

Chain of Thought Prompting

- Augment Exemplars with the Chain of Thought to get to the answer
- Chain of Thought is the sequence of intermediate reasoning steps
- LLMs learn to do Chain of Thought reasoning via in-context learning

Model Input

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

Model Output

A: 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. ✓

CoT Intuition

LLM pre training essentially stores training data as a compressed index of some sort in their weights

In-context examples guide the LLM to retrieve relevant pre-training data from the weights

In context exemplars guide the LLM to leverage pretraining data to produce CoT like output

Producing each token is much easier and can be done using pre-training data

LLM can produce more token for problems requiring more computation

Two tasks: Produce CoT like output & make sure the output is correct

Model Input

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

Model Output

A: 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. ✓



CoT Intuition

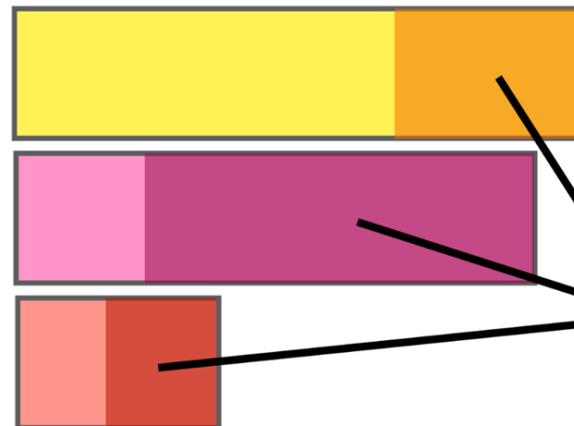
- Instead of solving the problem in a single pass, the model produces multiple tokens
- While producing an intermediate sequence of tokens the LLM is doing a particular task for eg. numerical reasoning
- With scale, LLMs get better at all of these tasks and hence better at the overall task reaching the correct answer

Types of errors made by a 62B language model:

Semantic understanding
(62B made 20 errors of this type,
540B fixes 6 of them)

One step missing
(62B made 18 errors of this type,
540B fixes 12 of them)

Other
(62B made 7 errors of this type,
540B fixes 4 of them)



Errors fixed by
scaling from
62B to 540B

CoT: Explainability

Question

Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?

62B Model Output

He travels from 1 to 4 PM. That means he travels for 3 hours. He travels back at 6 mph. That means he travels for $3 * 6 = 18$ miles. The answer is 18. **(Incorrect: semantic understanding error)**

540B Model Output

He travels at 10 mph from 1 to 4 PM. This means he travels 3 hours. 3 hours at 10 mph means he travels $3 * 10 = 30$ miles. He then travels back at 6 mph. This means he travels 6 miles per hour. He has to travel 30 miles, so it takes him $30 / 6 = 5$ hours. The answer is 5. **(Correct)**

CoT: Wide Applicability

Math Word Problems (free response)

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.

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

StrategyQA

Q: Yes or no: Would a pear sink in water?

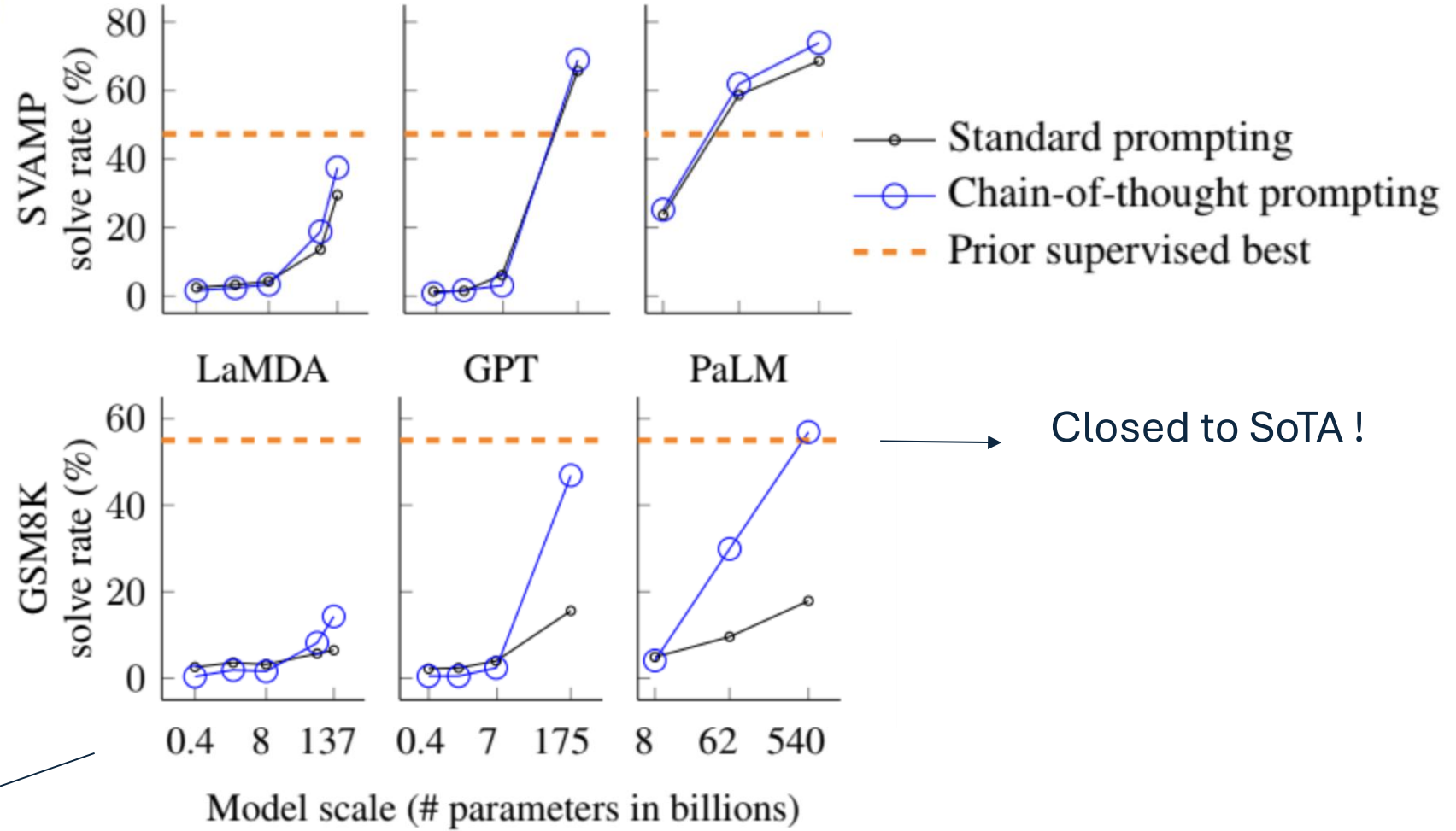
A: The density of a pear is about 0.6 g/cm^3 , which is less than water. Thus, a pear would float. So the answer is no.

Arithmetic Reasoning

Symbolic Reasoning

CommonSense Reasoning

CoT: Huge Performance Gains



Huge Gains on Challenging datasets !



Closed to SoTA !



Emergent Ability !



CoT: Pros ?

- Variable Computation
- Explainability
- Wide Applicability
- No Training
- Huge Performance Gains

Model Input

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

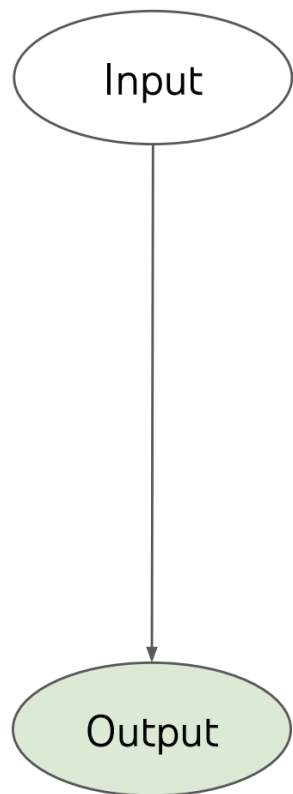
Model Output

A: 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. ✓

What have we achieved ?



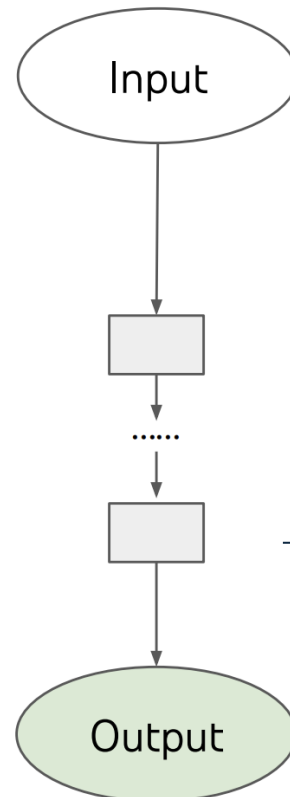
Often Multiple ways to reach the answer !



Standard Prompting

Guide the LLM through IC learning

to perform explicit reasoning



CoT prompting

What if LLM gets something wrong ?



Decoding with LLMs: Greedy Decoding



Question
+
Prompt

The probability distribution predicted by the LLM is a way to explore diverse reasoning paths !

model's vocabulary

0.298	0.213	0.125	0.098	0.087	0.056	0.013	0.011
the	their	another	a	an	victory	its	all

Select token w/
highest
probability

'the'



Self Consistency Decoding

Greedy Decoding

CoT

Prompt

LLM

This means she uses $3 + 4 = 7$ eggs every day. She sells the remainder for \$2 per egg, so in total she sells $7 * \$2 = \14 per day. The answer is \$14.

SC

Temperature Sampling to sample Diverse Reasoning Paths !

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?

LLM

She has $16 - 3 - 4 = 9$ eggs left. So she makes $\$2 * 9 = \18 per day.
The answer is 18

This means she she sells the remainder for $\$2 * (16 - 4 - 3) = \26 per day.
The answer is 26

She eats 3 for breakfast, so she has $16 - 3 = 13$ left. Then she bakes muffins, so she has $13 - 4 = 9$ eggs left. So she has $9 \text{ eggs} * \$2 = \18 .
The answer is 18

Aggregate Answers

The answer is 18

Marginalize out reasoning paths !

Self Consistency

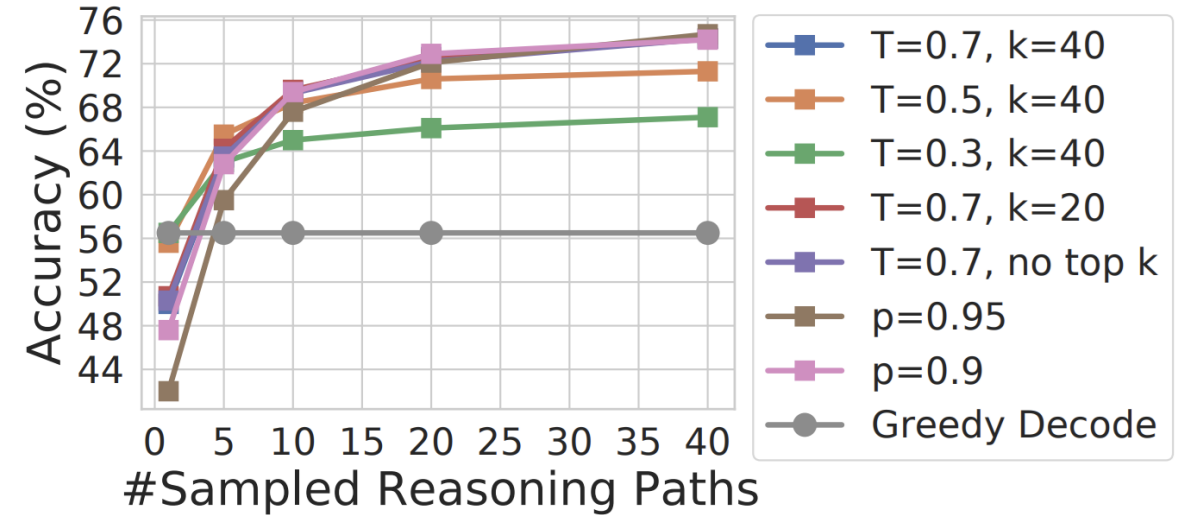
Pros

- Simple and Effective
- No Training / Fine Tuning
- Model Calibration / Confidence

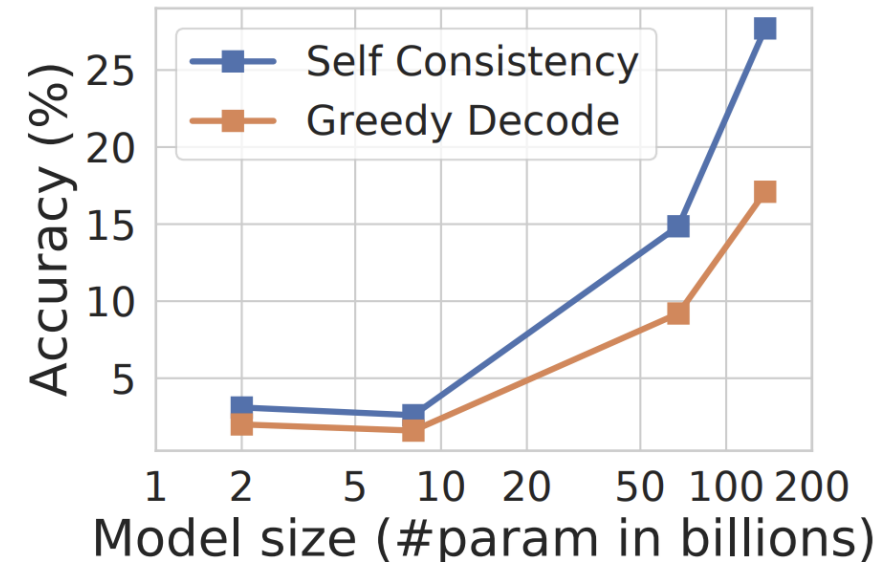
Cons

- Computation Cost
- Does not fix the inherent reasoning limitations of the LLM

GSM8K with PaLM- 540 B



GSM8K with LaMDA



Outline: How to improve LLMs at Reasoning

- Prompting Techniques
- Large Reasoning Models
- LLMs + Tools

Train an LLM that is a Natively Good Reasoner

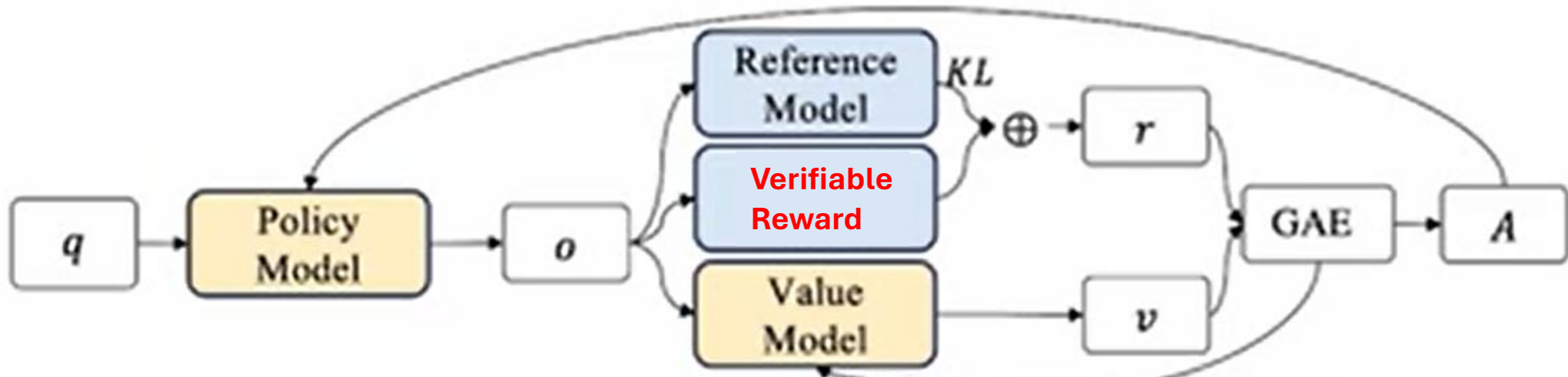
- Current Status
 - Train using LM objective
 - Use inference time compute (CoT) for reasoning
- Idea
 - Train for CoT
- Challenge
 - Writing CoT on large scale very expensive
 - Also human CoTs may not be ideal for a model to train on
- Idea
 - Use RL (natural reward: whether you got correct answer or not)

Reward

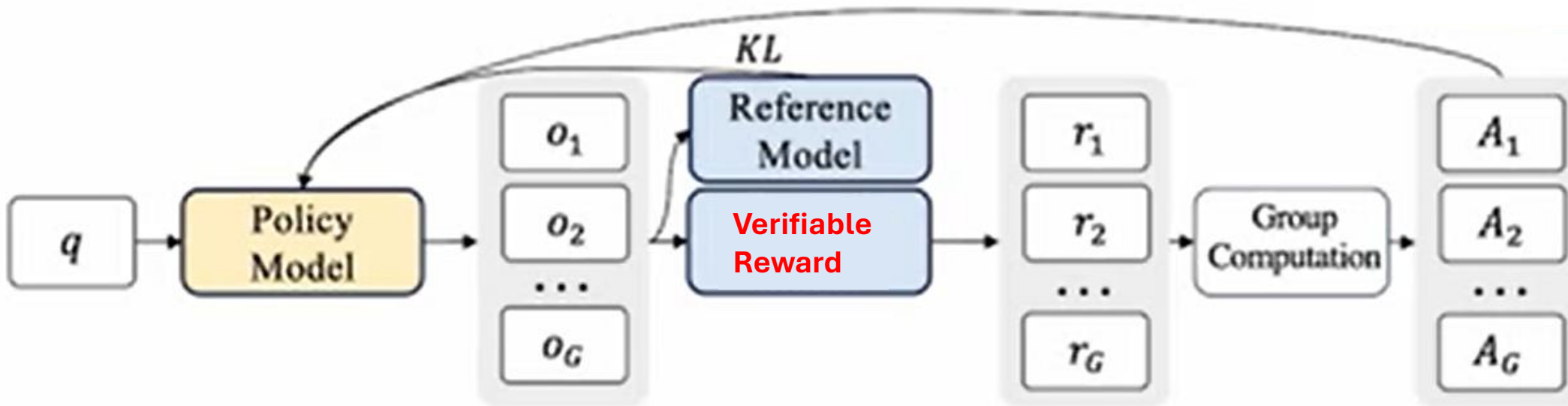
- Reward Model
 - Not needed
- Reward
 - Reward for generating CoT
 - (generally special tokens like <think>, </think>, <answer>, </answer>, etc.
 - Reward for getting correct answer

PPO → GRPO

PPO



GRPO



Mathematically

PPO

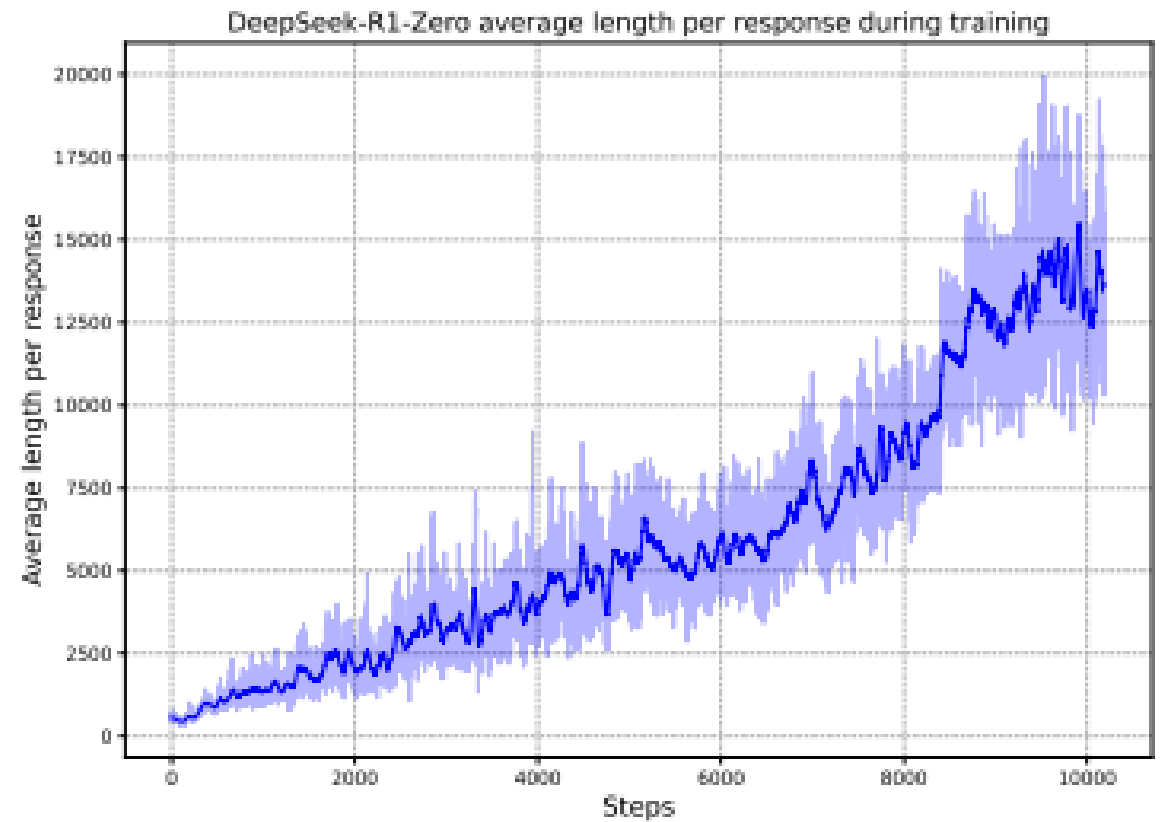
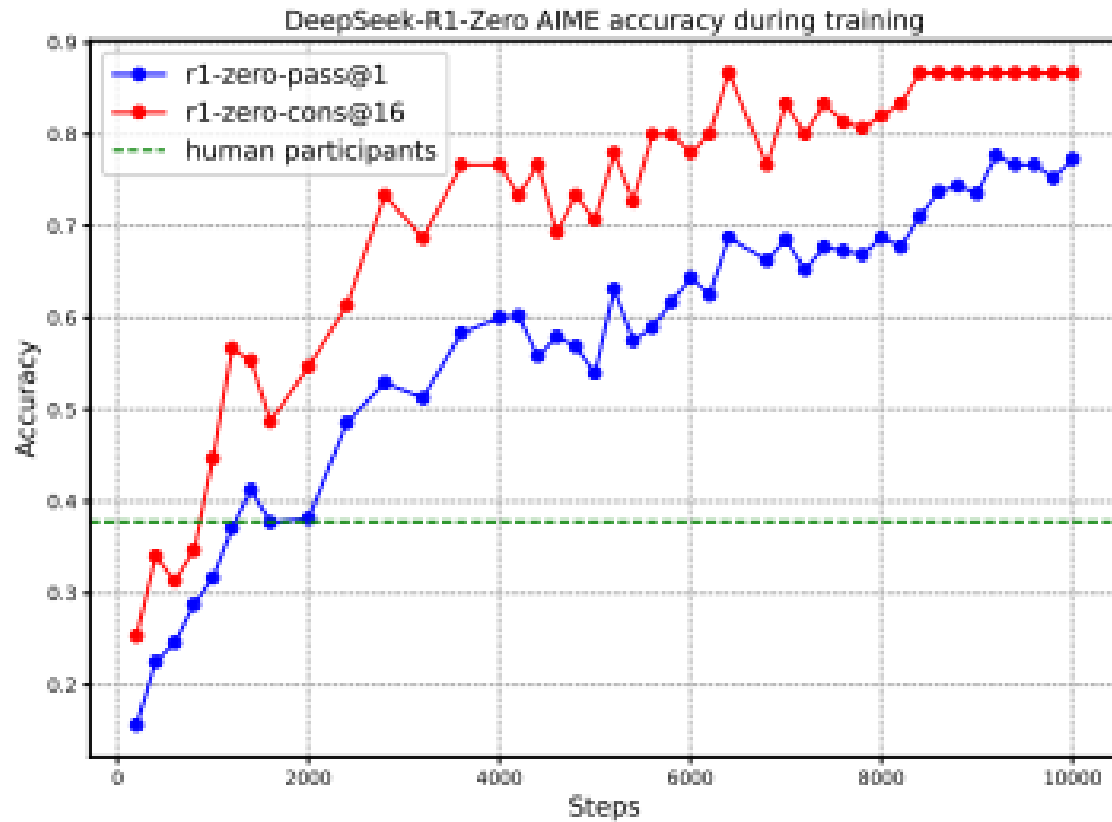
$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right] - \beta \mathbf{D}_{KL} [\pi_{\theta} || \pi_{ref}]$$

GRPO

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \\ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] - \beta \mathbf{D}_{KL} [\pi_{\theta} || \pi_{ref}] \right\}$$

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

DeepSeek R1-Zero Performance



CoT Length

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

$$\frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \boxed{\frac{1}{|o_i|}} \left\{ \min \left[\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{i,t} \right] - \beta \mathbf{D}_{KL} [\pi_{\theta} || \pi_{ref}] \right\}$$

Small $|o|$



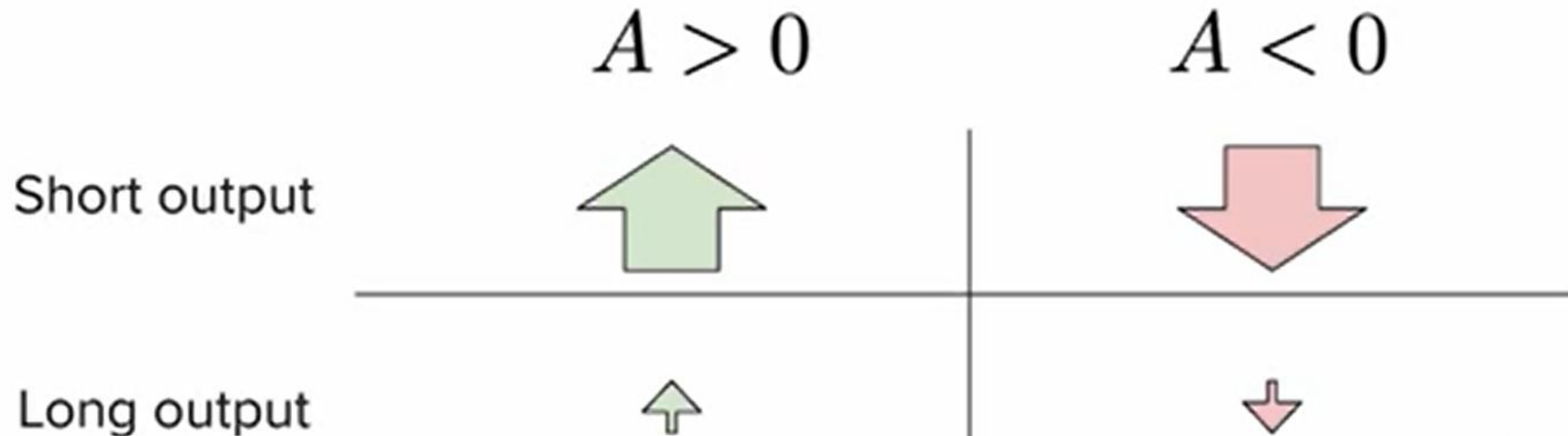
Big $|o|$



CoT Length

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

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



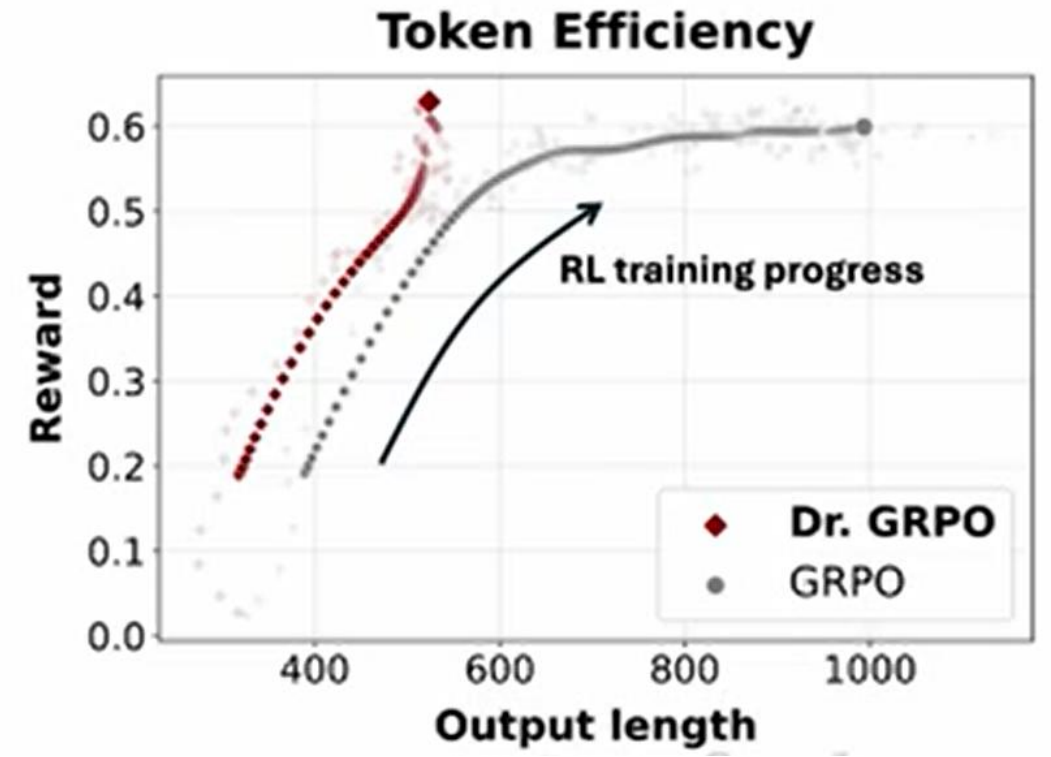
Solutions

Problem. $\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|}$

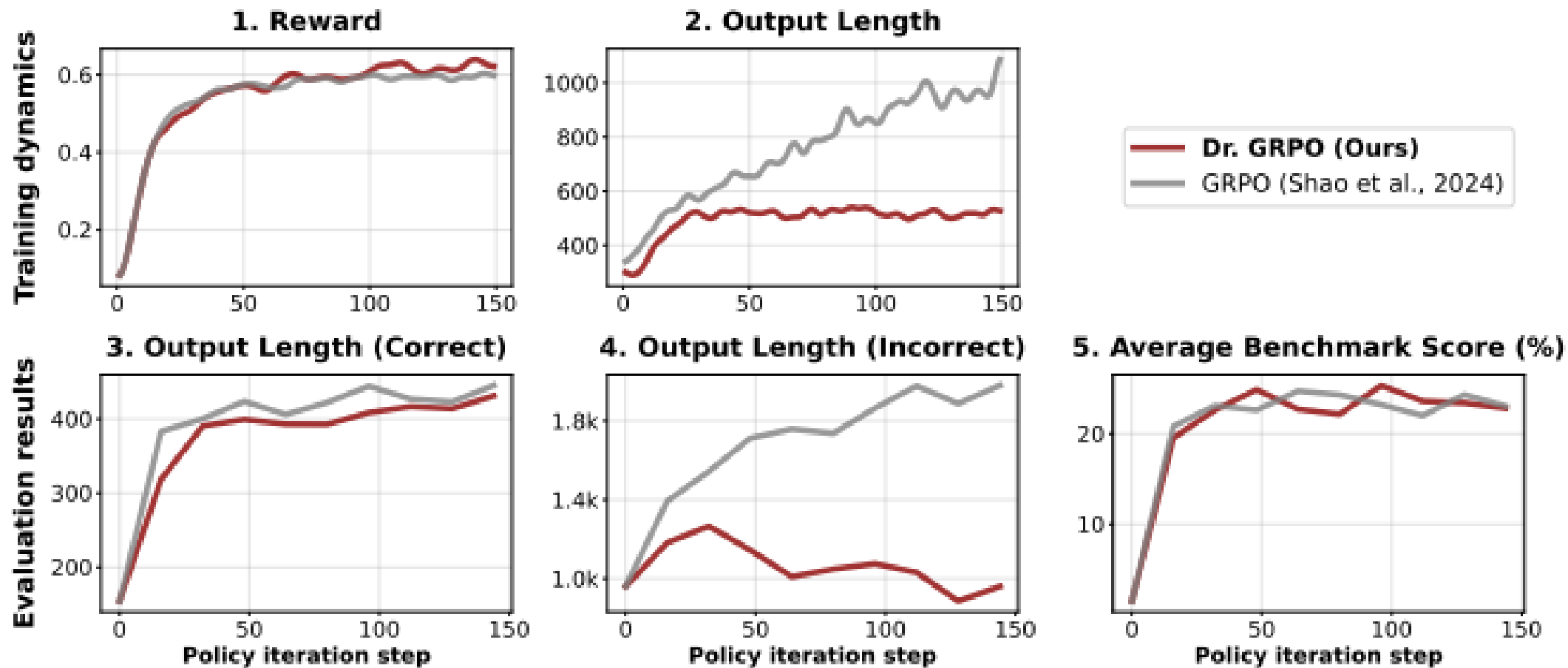
Remedy. Equalize token-level contributions

- DAPO $\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|}$

- Dr. GRPO $\frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|o_i|}$



Dr. GRPO Results



Other Modifications

Bias linked to **level of difficulty**

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

Encourage **diversity**

$$\text{clip}(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon) \longrightarrow \text{clip}(r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}})$$

DeepSeek R1 Recipe

- Train a Base Model: DeepSeek V3 Base
- Use RLVR with GRPO to get base reasoning model: DeepSeek R1-Zero
 - Issue: illegible/long CoTs, mixing language in CoT, syntax issues
- Generate CoT data using DeepSeek R1-Zero
 - Cold Start: Label / Rewrite CoT using human annotators (all CoT compliant)
- Start with DeepSeek V3 Base
- SFT on small-scale CoT data to train the model to generate good CoTs
- Use RLVR with GRPO until convergence to get good reasoning model
 - Reward = formatting reward + accuracy reward + language consistency reward
- Transition to general purpose model
- Large Scale SFT over
 - $\frac{3}{4}$ reasoning (600K, rejection sampling, V3 judge) and $\frac{1}{4}$ general (200K IFT) queries
- Combine RLHF and RLVR using GRPO to polish the model

Aha Moment in DeepSeek Training

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: ...

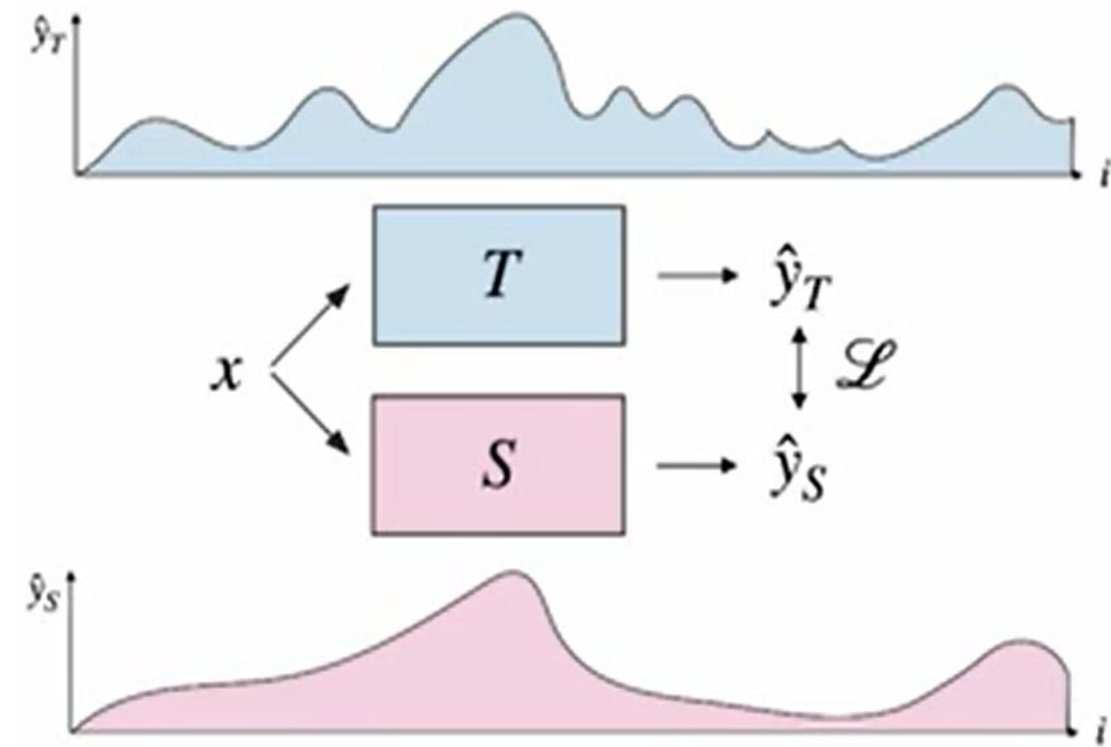
...

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.

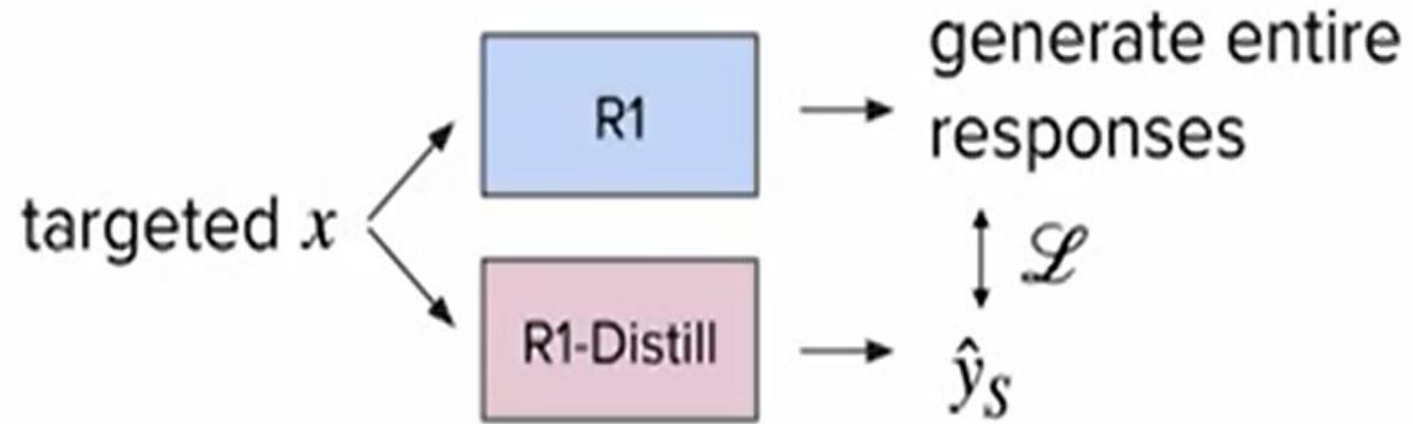
Are CoTs Faithful?

- Mostly No
- Some works say that random CoT also provides equally good results
- Gemini/OpenAI doesn't expose them
 - Can be long and uninterpretable
 - Only provides CoT summary
 - Continuous CoT (?)

DeepSeek R1 → DeepSeek-Distill

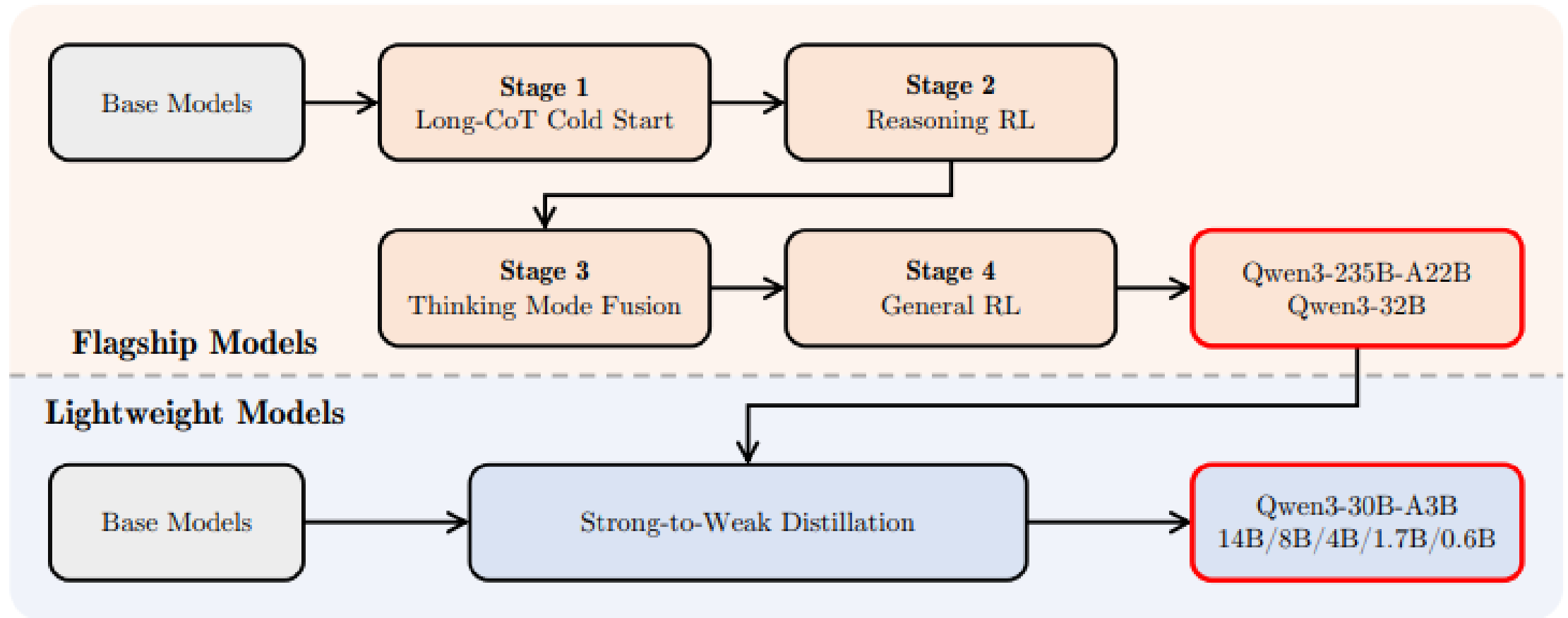


Goal: match next token distribution

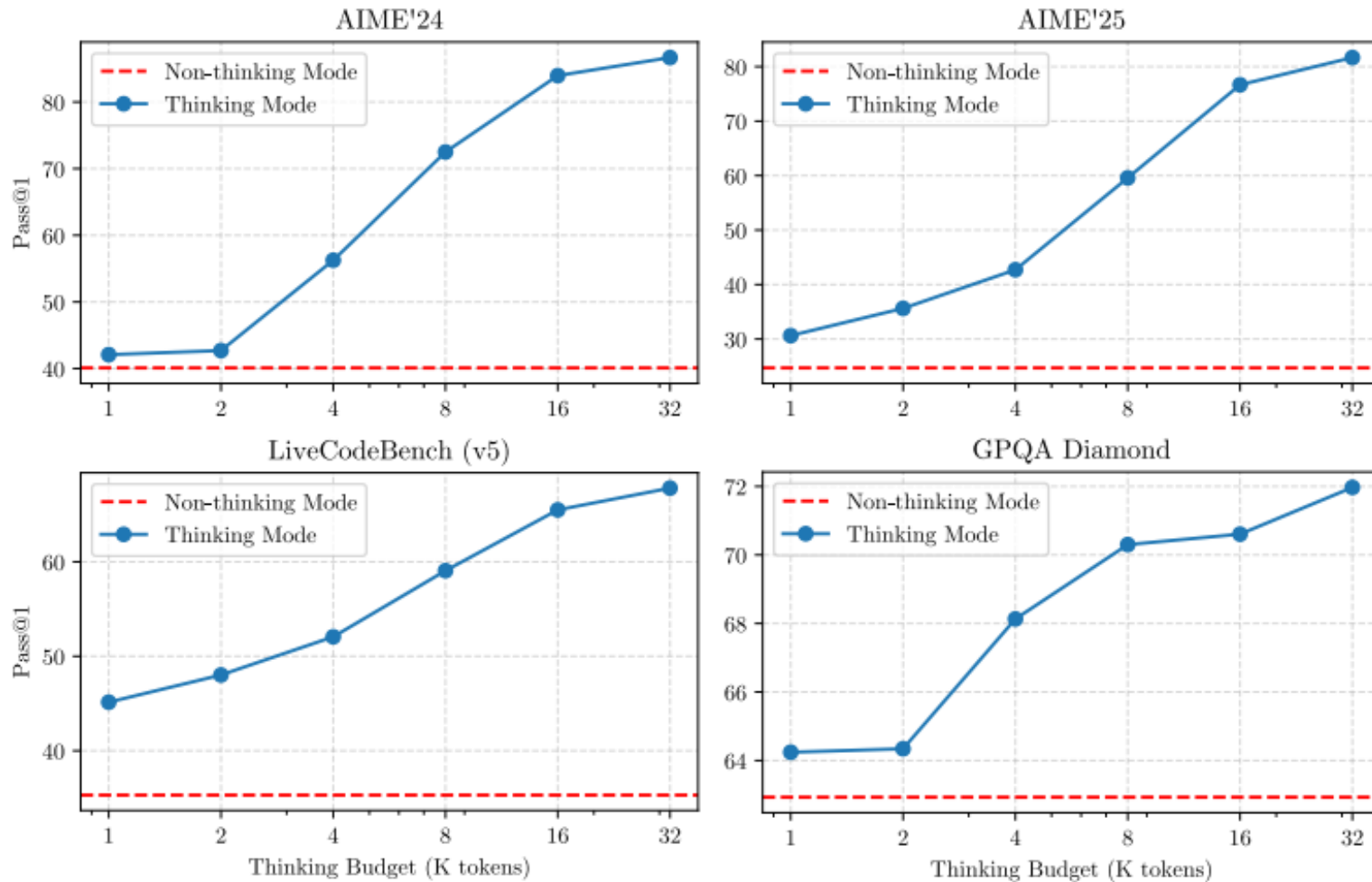


Goal: SFT on reasoning traces

Qwen3 Post-Training Pipeline



Thinking vs Non-Thinking: Graceful Degradation



Thinking Mode Fusion

Thinking Mode	Non-Thinking Mode
<pre>< im_start >user {query} /think< im_end > < im_start >assistant <think> {thinking_content} </think> {response}< im_end ></pre>	<pre>< im_start >user {query} /no_think< im_end > < im_start >assistant <think> </think> {response}< im_end ></pre>

Outline: How to improve LLMs at Reasoning

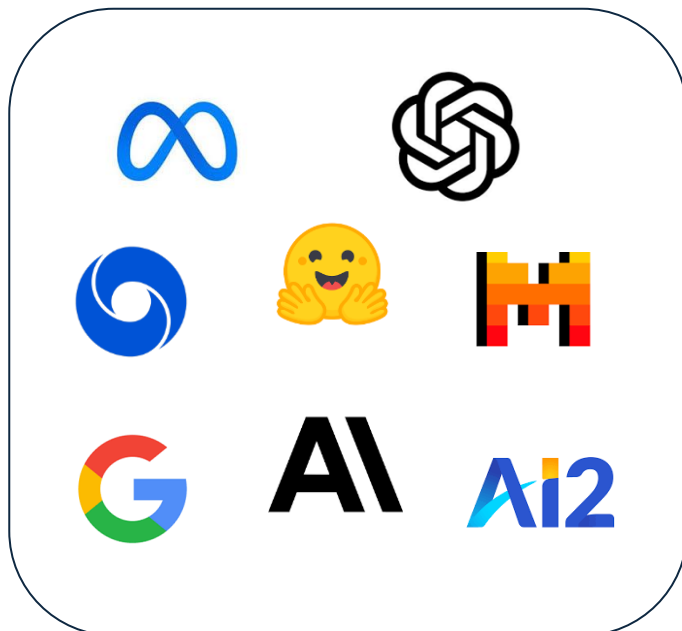
- Prompting Techniques
- Large Reasoning Models
- LLMs/LRMs + Tools

How to improve LLMs at Reasoning ?

LLM Modulo Frameworks



External Agent



Black Box LLM

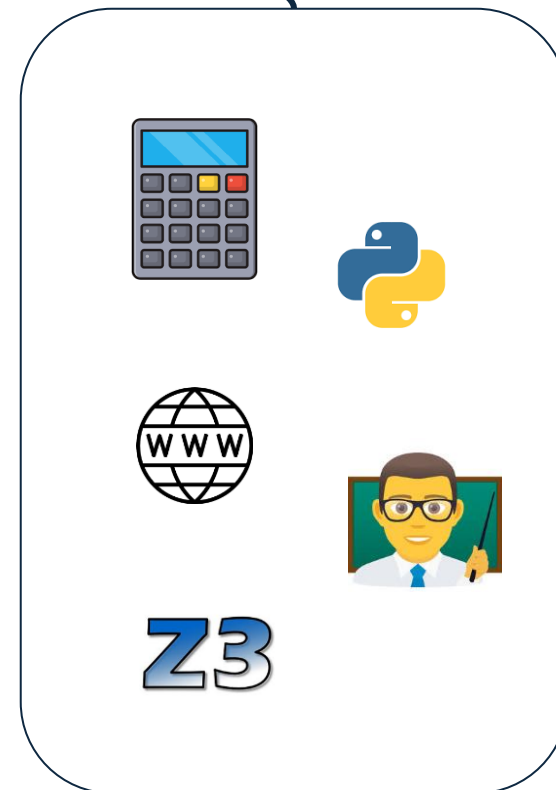
I am **GOOD** at



- Natural Language Understanding !
- Language Translation !
-

I am **BAD** at / costly for

- Maths
- Reasoning
-



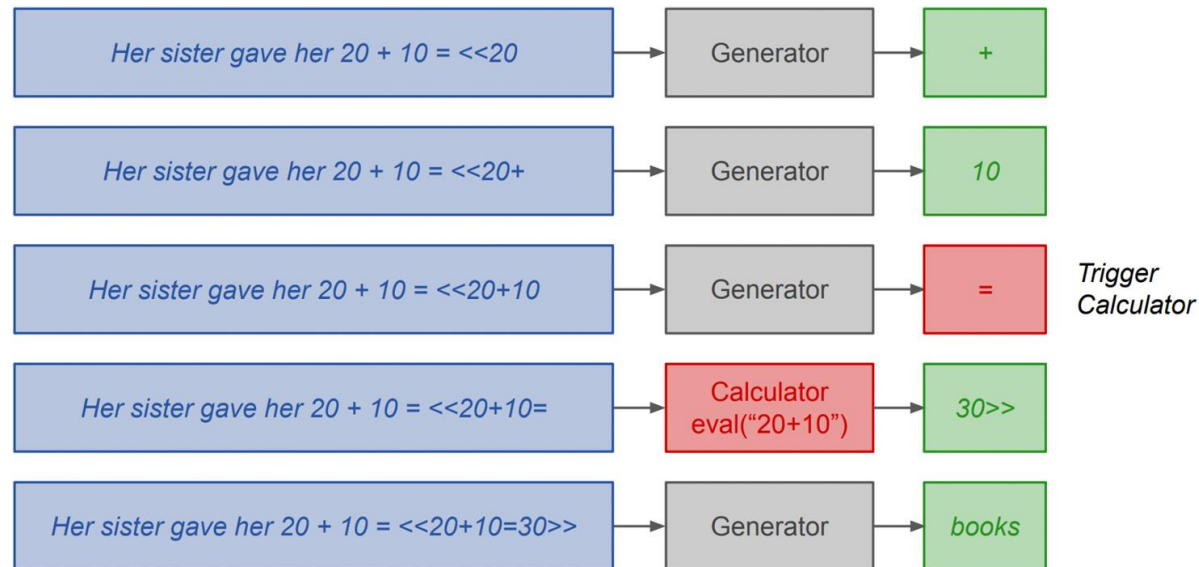
I do what I am good at, the agent does what it is good at

Fixing Numerical Reasoning Mistakes

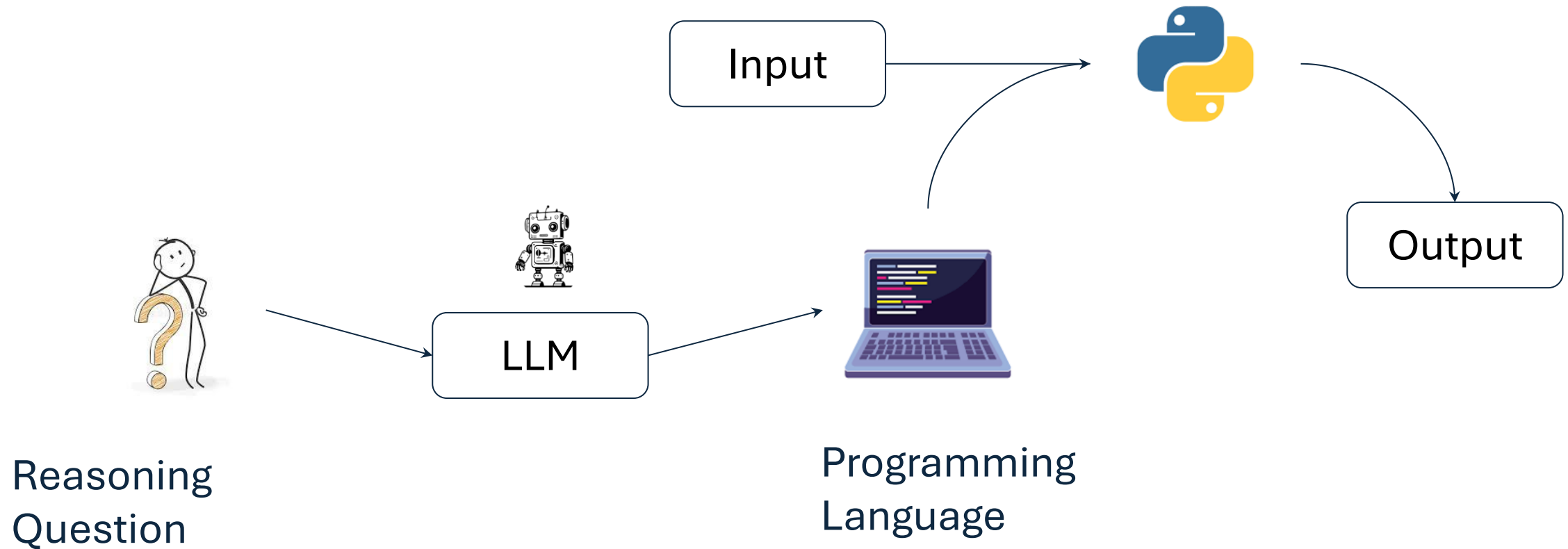
QUESTION: Dan plants 3 rose bushes. Each rose bush has 25 roses. Each rose has 8 thorns. How many thorns are there total?

MODEL ANSWER (INCORRECT; CALCULATOR ERROR ONLY): Dan plants 3 rose bushes. Each rose bush has 25 roses. Each rose has 8 thorns. So $3 \times 25 \times 8 = 300$. The answer is 300. ✗

Ask a calculator to do this ! But How ?



Program Interpreters as Calculators



Leverage LLMs Code Writing Abilities !



Program Aided Language Models



LLMs can Think in Code !

PAL

CoT

Input

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 tennis balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves. They sold 93 in the morning and 39 in the afternoon. So they sold $93 + 39 = 132$ loaves. The grocery store returned 6 loaves. So they had $200 - 132 - 6 = 62$ loaves left. The answer is 62.



Input

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 tennis balls.
`tennis_balls = 5`
2 cans of 3 tennis balls each is
`bought_balls = 2 * 3`
tennis balls. The answer is
`answer = tennis_balls + bought_balls`

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

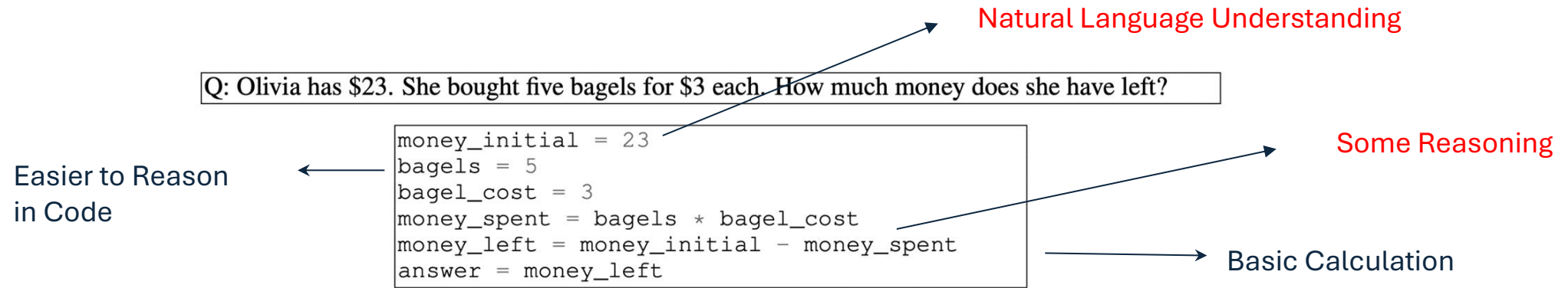
Model Output

A: The bakers started with 200 loaves
`loaves_baked = 200`
They sold 93 in the morning and 39 in the afternoon
`loaves_sold_morning = 93`
`loaves_sold_afternoon = 39`
The grocery store returned 6 loaves.
`loaves_returned = 6`
The answer is
`answer = loaves_baked - loaves_sold_morning`
`- loaves_sold_afternoon + loaves_returned`

```
>>> print(answer)  
74
```



Offloading Reasoning to the Interpreter



Q: On the table, you see a bunch of objects arranged in a row: a purple paperclip, a pink stress ball, a brown keychain, a green scrunchiephone charger, a mauve fidget spinner, and a burgundy pen. What is the color of the object directly to the right of the stress ball?

```
...
stress_ball_idx = None
for i, object in enumerate(objects):
    if object[0] == 'stress ball':
        stress_ball_idx = i
        break
# Find the directly right object
direct_right = objects[stress_ball_idx+1]
# Check the directly right object's color
answer = direct_right[1]
```

Looping & Conditionals

LLMs as Idea Generators !

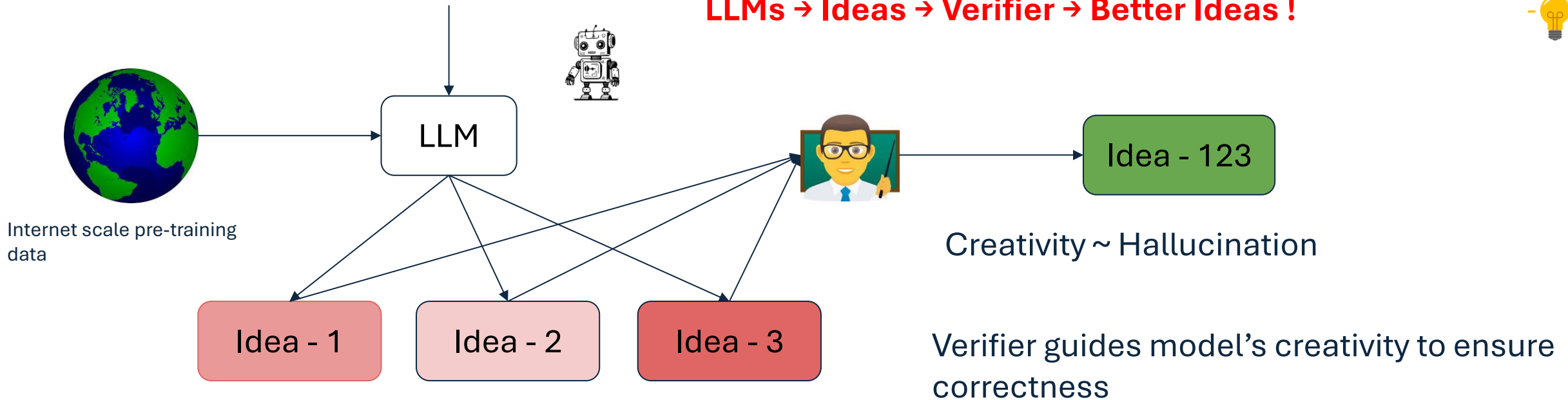
Reasoning requires being creative !

“The way to get good ideas is to get lots of ideas, and throw the bad ones away.”

- Linus Pauling



Coding Assignment
building a neural network for question table
answering



Bin Packing

- Given a set of blocks of different sizes and bins, decide which bin to place the block in
- Minimize the total number of bins used
- Practical Eg: Job Scheduling on clusters
- Combinatorial Optimization: NP Hard
- Online Setting → Heuristics



FunSearch Set up: Online Bin Packing as a Program

```
def main(problem):  
    """Runs `solve` on online 1d bin packing instance,  
    ↪ and evaluates the output."""  
    bins = problem.bins  
    # Packs `problem.items` into `bins` online.  
    for item in problem.items:  
        # Extract bins that have space to fit item.  
        valid_bin_indices =  
        ↪ utils_packing.get_valid_bin_indices(item,  
        ↪ bins)  
        best_index = solve(item,  
        ↪ bins[valid_bin_indices])  
        # Add item to the selected bin.  
        bins[valid_bin_indices[best_index]] -= item  
    return evaluate(bins, problem)
```

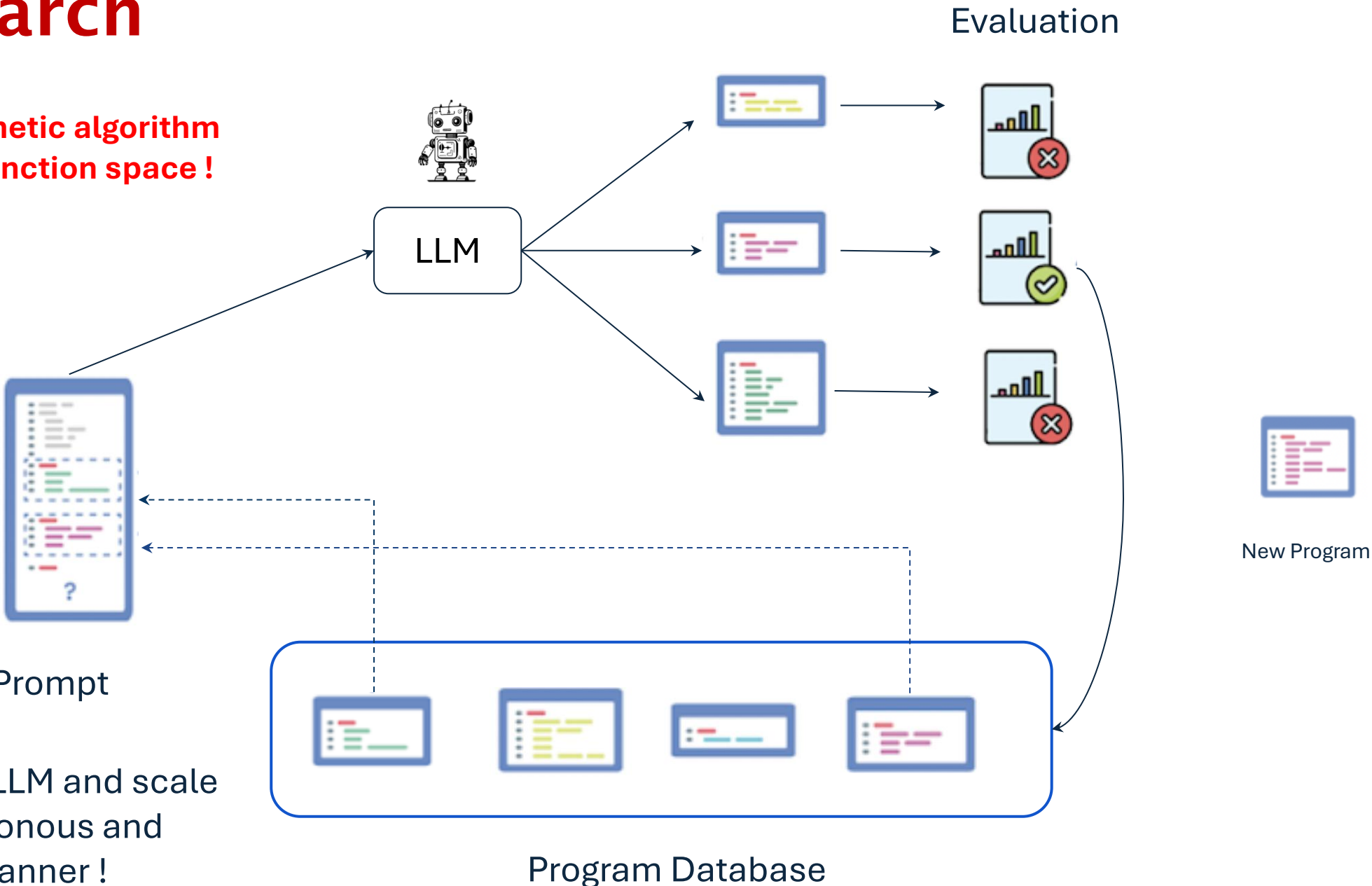
```
def heuristic(item, bins):  
    """Returns priority with which we want to add  
    ↪ `item` to each bin."""  
    return -(bins - item)
```

Use LLMs to discover new heuristics
and evaluate these !

```
def evaluate(bins, problem):  
    """Returns the negative of the number of bins  
    ↪ required to pack items in `problem`."""  
    if utils_packing.is_valid_packing(bins, problem):  
        return -utils_packing.count_used_bins(bins,  
        ↪ problem)  
    else:  
        return None
```

FunSearch

LLM guided genetic algorithm for search in function space !



Use a cheap LLM and scale in an asynchronous and distributed manner !

FunSearch Results

```
def heuristic(item: float, bins: np.ndarray) -> np.ndarray:
    """Online bin packing heuristic discovered with FunSearch."""
    score = 1000 * np.ones(bins.shape)
    # Penalize bins with large capacities.
    score -= bins * (bins-item)
    # Extract index of bin with best fit.
    index = np.argmin(bins)
    # Scale score of best fit bin by item size.
    score[index] *= item
    # Penalize best fit bin if fit is not tight.
    score[index] -= (bins[index] - item)**4
    return score
```



Best-fit heuristic



FunSearch

FunSearch Results

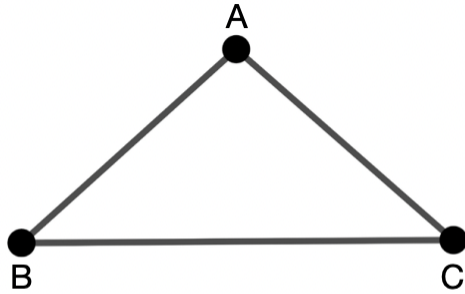
“

The solutions generated by FunSearch are far conceptually richer than a mere list of numbers. When I study them, I learn something.

JORDAN ELLENBERG, COLLABORATOR AND PROFESSOR OF MATHEMATICS AT THE UNIVERSITY OF WISCONSIN-MADISON

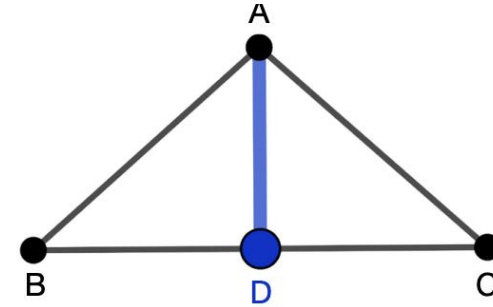
Geometry: Classical AI approach

LLMs can suggest constructions to guide search



“Let ABC be any triangle with $AB = AC$.
Prove that $\angle ABC = \angle BCA$.”

Construction lead to
large branching factors



► Construct D: midpoint BC,
 $AB=AC, BD = DC, AD=AD \Rightarrow \angle ABD=\angle DCA$ [1]
 [1], B C D collinear $\Rightarrow \angle ABC=\angle BCA$

Construction

X = angle bisector(A, B, C)
 X = angle mirror(A, B, C)
 X = circle(A, B, C)
 A, B, C, D = eq_quadrilateral()
 A, B, C, D = eq_trapezoid()
 X = eqtriangle(B, C)

Description

Construct a point X on the angle bisector of $\angle ABC$
 Construct a point X such that BC is the bisector of $\angle ABX$
 Construct point X as the circumcenter of A, B, C
 Construct quadrilateral ABCD with $AD = BC$
 Construct trapezoid ABCD with $AD = BC$
 Construct X such that XBC is an equilateral triangle

Perform Search

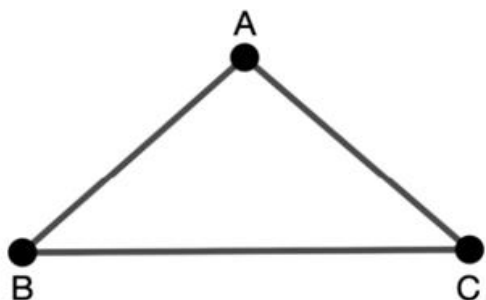
- cyclic(E,A,D,H) $\rightarrow \angle EAH = \angle EDH$
- X = eqtriangle(B, C) $\rightarrow \angle XBC = 60$

....

Represent problems
Symbolically

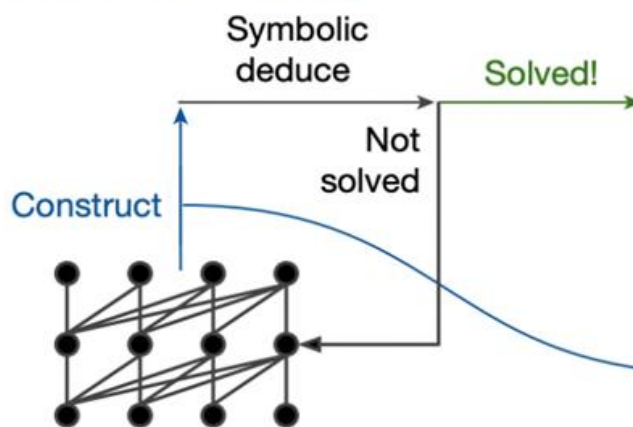
AlphaGeometry

a A simple problem



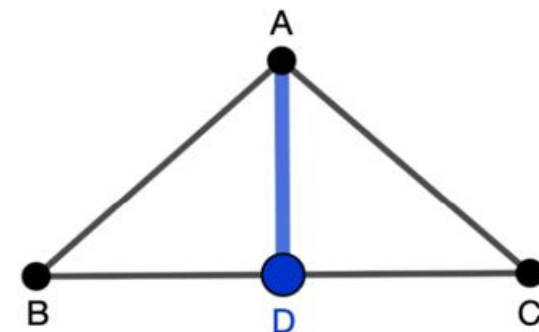
“Let ABC be any triangle with $AB = AC$.
Prove that $\angle ABC = \angle BCA$.”

b AlphaGeometry



c Language model

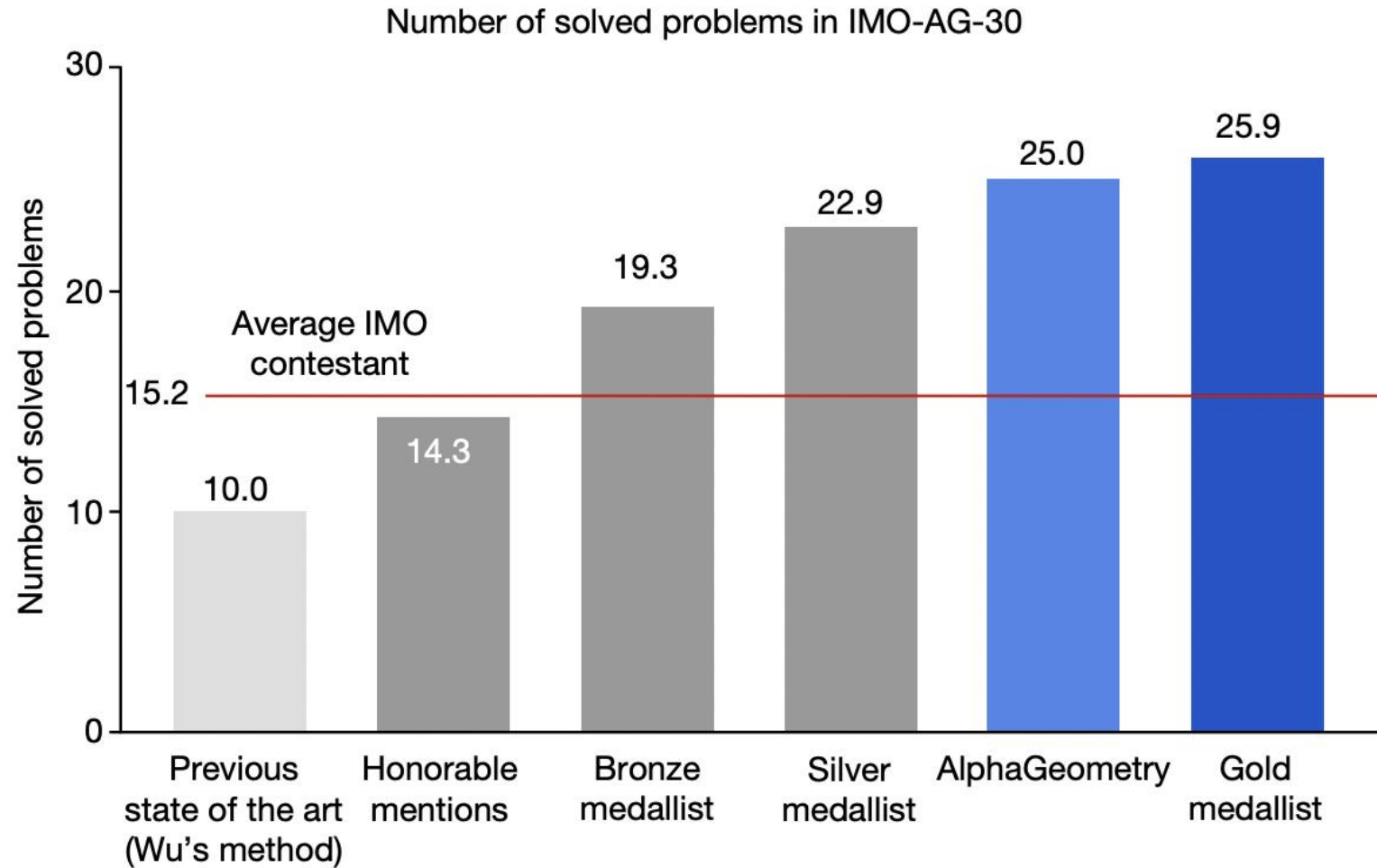
d Solution



Construct D: midpoint BC,
 $AB=AC, BD = DC, AD=AD \Rightarrow \angle ABD=\angle DCA$ [1]
[1], B C D collinear $\Rightarrow \angle ABC=\angle BCA$

AlphaGeometry: An Olympiad-level AI system for geometry

AlphaGeometry Results



Conclusion: LLMs for Reasoning

- Prompting Techniques (LLMs as Reasoners)
 - CoT
 - SC
 - There are many others
- Large Reasoning Models (LLMs as better reasoners)
 - RVLR based on CoT and RL
- LLMs/LRMs + Tools (LLMs not as reasoners)
 - PAL
 - LLMs as Idea Generators