# UC San Diego Qualcom

Al research

## Motivation

- Chain-of-thought reasoning introduces hallucinations and accumulated errors which is limiting models' ability to solve complex reasoning tasks.
- However, humans can utilize careful deductive logical reasoning processes to solve tasks and ensure the trustworthiness of their reasoning process through **self-verification**.
- Verifying the validity of an entire vanilla CoT reasoning chain is challenging, even with advanced models like ChatGPT.

<b>Question</b> : Jerry is rolling a six-sided die. How much more likely is it (expressed as a percentage) that he rolls a number greater than \$3 than that he rolls two even numbers in a row?	Pro						
<b>Ground Truth:</b> There are 3 numbers greater than 3 on the dice, so the chances of rolling one of them are $3/6 = 50\%$ . The chance of rolling one even number is 50%, so the chance of rolling two in a row is $50\% * 50\% = 25\%$ . The difference between these two probabilities is $50\% - 25\%$ = $25\%$ . #### 25							
<b>Output</b> : There are two possible outcomes for rolling a number greater than 3 out of 6 total outcomes, so the probability is $2/6$ or $1/3$ . For rolling two even numbers in a row, there are three possible outcomes (2-4, 4-6, 6-2) out of 36 total outcomes, so the probability is $3/36$ or							
1/12. To find the difference in probability, we subtract the probability of rolling two even numbers in a row from the probability of rolling a number greater than 3: $1/3 - 1/12 = 1/4$ . This means that it is 25% more likely to roll a number greater than 3 than to roll two even numbers in a row.							
A flawed reasoning chain with correct answer.							

	<b>Deductive Verif</b>	ficati	ion v	
• Na	atural Program: a ded	luctive	CoT	
• De	eductive verification of	of a N	atura	
• [	)ecompose a long reasc	ning c	hain ir	
• V	erify step-by-step.			
• Δ	reasoning chain is valid	d if and	d only	
	Question There are 53 maple trees currently in the park. Park workers will plant maple trees today. When the workers are finished there will be 64 maple trees in the park. How many maple trees did the workers plant today?		Here is so After the Based on the the number the worker Double-che end with	
	<pre> Question-Related Premises # 1. There are 53 maple trees currently in the park. # 2. After the workers plant trees, there will be 64 maple trees in the park. # 3. The workers plant some maple trees. # 4. How many maple trees did the </pre>	Reasoning # 5. (by #2 #1) Ster Calculate the numbe maple trees the wor will plant. Number trees the workers w		
	<pre># 1. There are 53 maple trees currently in the park. # 2. After the workers plant trees, there will be 64 maple trees in the park. # 3. The workers plant some maple trees. # 4. How many maple trees did the workers plant today?</pre>			

## **Deductive Verification of Chain-of-Thought Reasoning**

Zhan Ling<sup>1</sup>\*, Yunhao Fang<sup>1</sup>\*, Xuanlin Li<sup>1</sup>, Zhiao Huang<sup>1</sup>, Mingu Lee<sup>2</sup>, Roland Memisevic<sup>2</sup>, Hao Su<sup>1</sup> <sup>1</sup>University of California, San Diego, <sup>2</sup>Qualcomm AI Research

ompting	Reasoning Correctness	GSM8K	AQuA	MATH	AddSub	Date	Last Letters
ero-shot	Correct	98%	96%	100%	98%	98%	100%
	Incorrect	4%	6%	4%	2%	4%	4%
	(Average)	51%	51%	52%	50%	51%	52%
wo-shot	Correct	98%	96%	100%	92%	100%	96%
	Incorrect	2%	4%	0%	6%	26%	6%
	(Average)	50%	50%	50%	49%	63%	51%

Zero-shot and two-shot verification accuracy of GPT-3.5-turbo for 100 generated CoT reasoning hains: 50 valid and 50 with mistakes.

### with Natural Program

reasoning format for LLMs.

**Program reasoning chain:** 

nto a series of reasoning steps.

### if every reasoning step is valid.

Verification

ome information: "There are 53 maple trees currently in the park. workers plant trees, there will be 64 maple trees in the park." the given information, here is a reasoning process: "Calculate of maple trees the workers will plant. Number of maple trees rs will plant: 64 - 53 = 11"

eck the reasoning process, let's analyze its correctness, and "yes" or "no".





## **Results and Limitations**

Verification Method	Reasoning Correctness	GSM8k	AQuA	MATH	AddSub	Date	Last Letters	Overall
СоТ	Correct	98% 2%	96% 4%	100% 0%	92% 6%	100% 26%	96% 6%	97%
Two-shot	(Average)	50%	50%	50%	49%	63%	51%	52%
Natural Dragram	Correct	84%	72%	70%	95%	90%	96%	85%
One shot	Incorrect	84%	62%	76%	40%	56%	6%	54%
One-shot	(Average)	84%	67%	73%	68%	73%	51%	69%

		Arith	Commonsense			
Methods	GSM8K	AQuA	MATH*	AddSub	Date	Last Letters
CoT + Voting Faithful CoT + Voting	<b>87.62%</b> 75.80%	70.18% 61.80%	35.93% 31.78% <sup>1</sup>	92.36% 88.35% <sup>1</sup>	69.97% <b>73.50%</b>	81.60% -
Ours (Natural Program (NP), No Verification) Ours (NP + Deductive Verification + UPV)	87.05% 86.01%	<b>70.34%</b> 69.49%	<b>36.75%</b> 36.48%	<b>93.67%</b> 93.54%	72.49% 71.45%	<b>92.98%</b> 92.60%
<b>Question</b> : In March it rained 0.81 inches. It rained 0.35 inches less in April than in March. How much did it rain in April? <b>Ground Truth:</b> It rained 0.81 - 0.35 = 0.46 inches.						
GPT-3.5-turbo (ChatGPT): Question-related premises: #1. It rained 0.81 inches in March.						

#2. It rained 0.35 inches less in April than in March. #3. How much did it rain in April? The first reasoning step: #4. (by #2) Step 1: Calculate how much it rained in April by subtracting the difference of 0.35 inches from the amount in March. Amount of rain in April: 0.81 - 0.35 = 0.46 inches