



Motivation

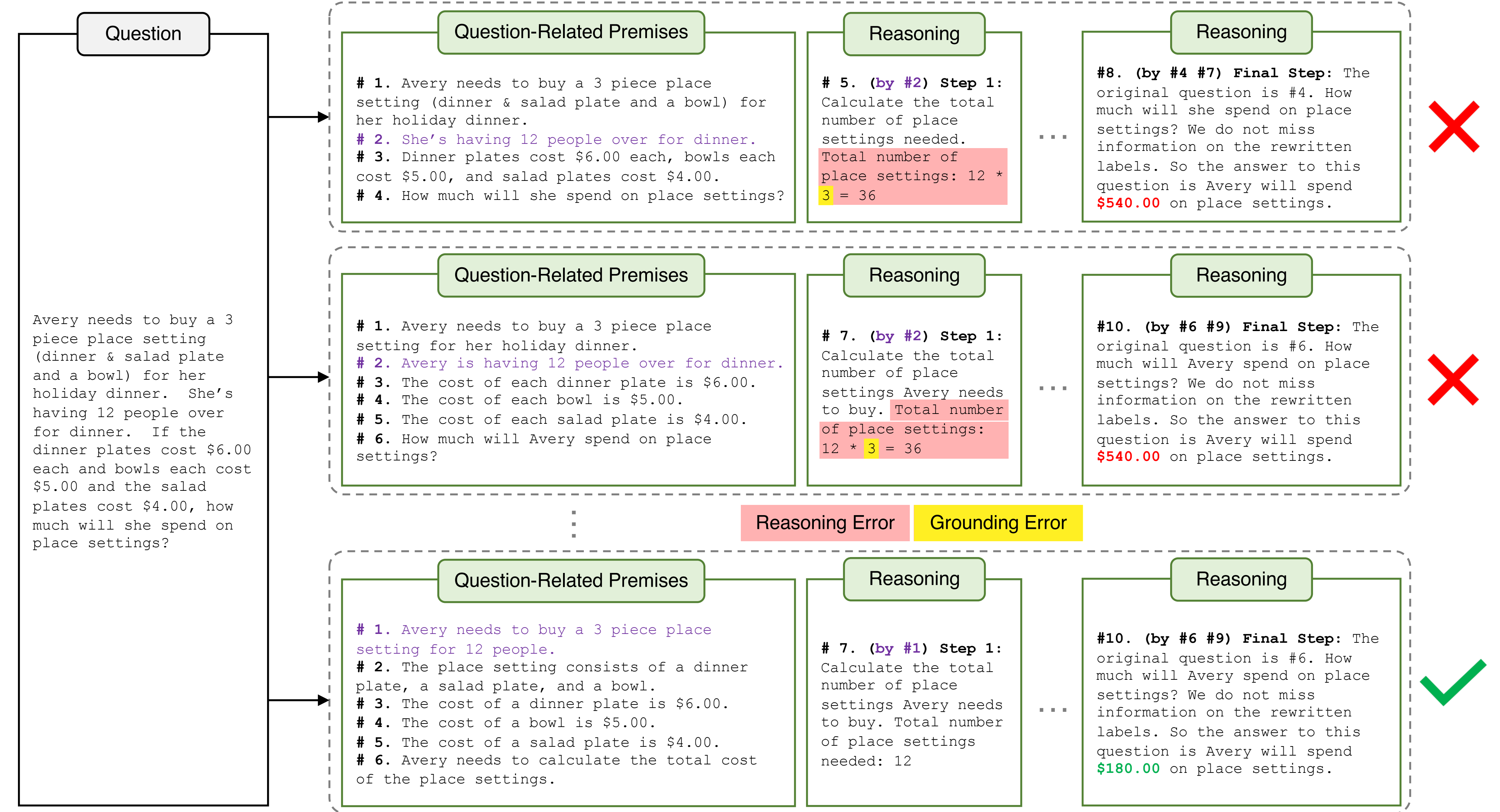
- Chain-of-thought reasoning introduces **hallucinations** and **accumulated errors** which is limiting models' ability to solve complex reasoning tasks.
- However, humans can utilize careful **deductive logical reasoning processes** to solve tasks and ensure the **trustworthiness** of their reasoning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

Question: 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?
Ground Truth: 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
Output: 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.

Prompting	Reasoning Correctness	GSM8K	AQuA	MATH	AddSub	Date	Last Letters
Zero-shot	Correct	98%	96%	100%	98%	98%	100%
	Incorrect (Average)	4%	6%	4%	2%	4%	4%
Two-shot	Correct	98%	96%	100%	92%	100%	96%
	Incorrect (Average)	2%	4%	0%	6%	26%	6%

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

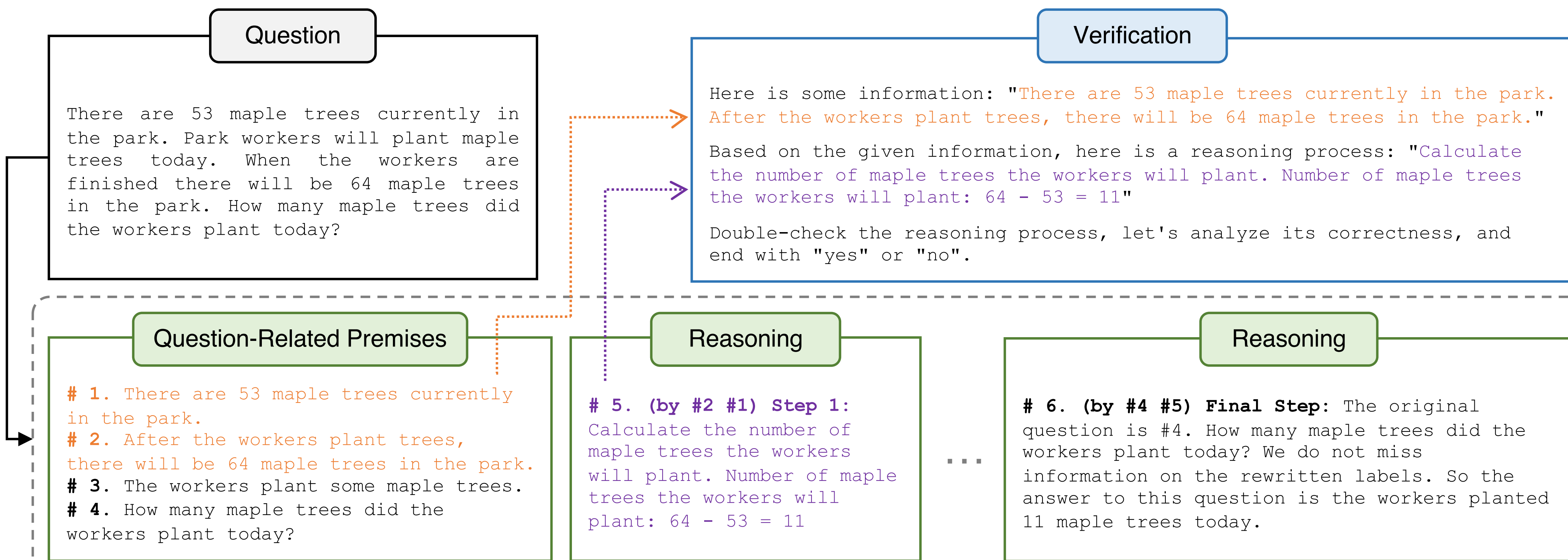
A flawed reasoning chain with correct answer.



Unanimity-Plurality Voting

Deductive Verification with Natural Program

- Natural Program:** a deductive CoT reasoning format for LLMs.
- Deductive verification of a Natural Program reasoning chain:**
 - Decompose a long reasoning chain into a series of reasoning steps.
 - Verify step-by-step.**
 - A reasoning chain is valid if and only if every reasoning step is valid.



Results and Limitations

Verification Method	Reasoning Correctness	GSM8k	AQuA	MATH	AddSub	Date	Last Letters	Overall
CoT Two-shot	Correct	98%	96%	100%	92%	100%	96%	97%
	Incorrect	2%	4%	0%	6%	26%	6%	7%
	(Average)	50%	50%	50%	49%	63%	51%	52%
Natural Program One-shot	Correct	84%	72%	70%	95%	90%	96%	85%
	Incorrect	84%	62%	76%	40%	56%	6%	54%
	(Average)	84%	67%	73%	68%	73%	51%	69%

Methods	Arithmetic				Commonsense	
	GSM8K	AQuA	MATH*	AddSub	Date	Last Letters
CoT + Voting	87.62%	70.18%	35.93%	92.36%	69.97%	81.60%
Faithful CoT + Voting	75.80%	61.80%	31.78% ¹	88.35% ¹	73.50%	-
Ours (Natural Program (NP), No Verification)	87.05%	70.34%	36.75%	93.67%	72.49%	92.98%
Ours (NP + Deductive Verification + UPV)	86.01%	69.49%	36.48%	93.54%	71.45%	92.60%

Question: 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?
Ground Truth: It rained $0.81 - 0.35 = 0.46$ inches.

GPT-3.5-turbo (ChatGPT):

Question-related premises:

- It rained 0.81 inches in March.
- It rained 0.35 inches less in April than in March.
- 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