



# Benchmarking Hallucination in Large Language Models based on Unanswerable Math Word Problem

Yuhong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, Hui Zhao

# Introduction: Hallucination towards Math Word Problems (MWP)



Q: Tom had a total of 50 salty cookies and sweet cookies combined. He ate 14 sweet cookies and 9 salty cookies. How many salty cookies did Tom have left?



GPT-3.5

A: Tom had 41 salty cookies left.



GPT-4

A: The problem doesn't provide information on how many salty cookies and sweet cookies Tom had at the beginning, so we can't definitively answer how many salty cookies Tom has left.





# Why choosing Math Word Problems(MWP) as benchmark?

---

## □ Focus on reasoning:

- Compared with general questions, MWP is challenging to mitigate hallucination through additional text retrieval.
- Answering MWP heavily relies on the LLM's intrinsic abilities, including comprehension, reasoning, and computation abilities.

## □ Easy to check:

- The answer to MWP is often unique and represented as a numerical value or variable expression.
- In determining whether a model is prone to hallucination, the MWP-based method only involves evaluating the correctness of a numerical or variable expression output.

# Dataset Construction: Modification Strategy

---

Original Question: The sum of three consecutive even numbers is 246. What is the number?



Contradiction

Modified Question: The sum of one consecutive even number is 247. What is the number?

# Dataset Construction: Modification Strategy

Strategy	Example	Original
Key information Deletion	Suzanne wants to raise money for charity by <u>running a race</u> . Her parents have pledged to donate \$10 for her first kilometer and double the donation for every successive kilometer. If Suzanne finishes the race, how much money will her parents donate?	running a 5-kilometer race
Range	Nadine collected different colored pebbles. She has <u>more than 20</u> white pebbles and half as many red pebbles. How many pebbles does she have in all?	20
Contradiction	The sum of <u>one consecutive even number is 247</u> . What is the number?	three consecutive even numbers is 246
Negation	There were 8 friends playing a video game online when 2 more players joined the game. If each player <u>had not 6 lives</u> , how many lives did they have in total?	had 6 lives
Summation	Baker made 61 pastries and 167 cakes. If he sold <u>totally 108 cakes and pastries altogether</u> . How many cakes would baker still have?	108 cakes and 44 pastries
Unrealism	Sue works in a factory and <u>every 0 minutes</u> , a machine she oversees produces 30 cans of soda. How many cans of soda can one machine produce in 8 hours?	every 30 minutes
Subject Substitution	Brittany, Alex, and Jamy all share 600 marbles divided between them in the ratio 3:5:7. If Brittany gives Alex half of her marbles, what's the total number of marbles that <u>Johnson has?</u>	Alex
Question Deletion	Jennifer will be 30 years old in ten years. At that time, her sister Jordana will be three times as old Jennifer. <u>How ?</u>	How old is Jennifer's sister now?

# Dataset Construction: Unanswerable Question

Type	Example	Percentage
Key Information Missing	Samanta has 8 more points than Mark, and Mark has 50% more points than <u>Eric</u> . How many points do Samanta, Mark, and <u>Eric</u> have in total?	32%
Ambiguous Key Information	Jack received <u>some</u> emails in the morning, 5 emails in the afternoon, and 8 emails in the evening. How many more emails did Jack receive in the afternoon and evening than in the morning?	49%
Unrealistic Conditions	How many <u>triangles with a height of 0 inches and a width of 0 inches</u> could fit inside a square with 2-inch sides?	11%
Unrelated Object	Joshua bought 25 <u>oranges</u> for \$12.50. He sells each one for 60c, how much profit in cents will he make on each <u>apple</u> ?	4%
Question Missing	Baker made 13 cakes. He sold 91 of them and bought 154 new cakes. <u>How many?</u>	5%

Table 1: Unanswerable questions in the UMWP dataset that span across multiple categories.



# Dataset Construction: Answerable Question

---

Source	Total	Percentage	Average Length
SVAMP	500	19.2%	30.38
MultiArith	300	11.5%	31.76
<b>GSM8K</b>	<b>1700</b>	<b>65.4%</b>	<b>45.38</b>
ASDiv	100	3.8%	28.37

Table 2: Statistics of answerable questions.

# Evaluation Method

---

## Algorithm 1 Answerability Evaluation

---

```
1: Input: Generated text  $v$  of a question by LLM
2: Output: Answerable or not
3:  $S \leftarrow f_{\text{sim}}(v, u_i)$ 
4: if  $\max(S) \geq \mathcal{T}$  then
5:   return False
6: end if
7:  $T \leftarrow \text{TokenizeText}(v)$ 
8:  $T' \leftarrow \text{RemoveCommonVocabulary}(T)$ 
9:  $v' \leftarrow \text{RemoveWhitespace}(T')$ 
10: if  $\text{ContainsExpression}(v')$  then
11:   return False
12: end if
13: return True
```

---

 $f_{\text{sim}}$ 

SimCSE

 $\mathcal{T}$ 

Threshold

 $U = \{u_1, u_2, \dots, u_i\}$ 

Unanswerable template sentences

e.g. "There is no definitive answer."

# Evaluation Method

---

We adopt the F1 score as the metric for evaluating LLMs' degree of hallucination.

Positive Case: Unanswerable Question

Negative Case: Answerable Question

$$\textit{precision} = \frac{TP}{TP + FP}$$

$$\textit{recall} = \frac{TP}{TP + FN}$$

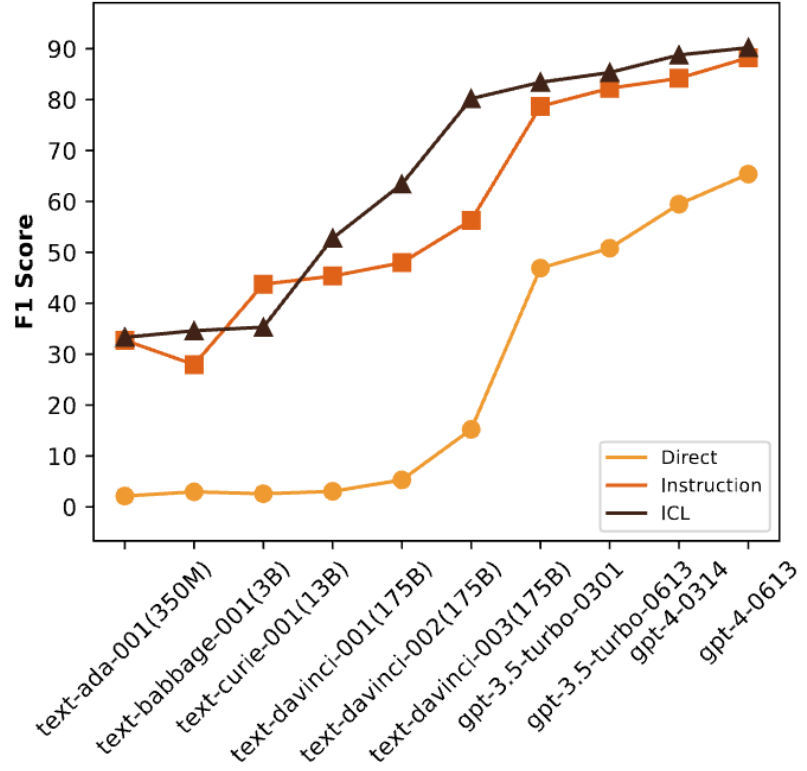
$$F_1 = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

# Experiment: Main Result

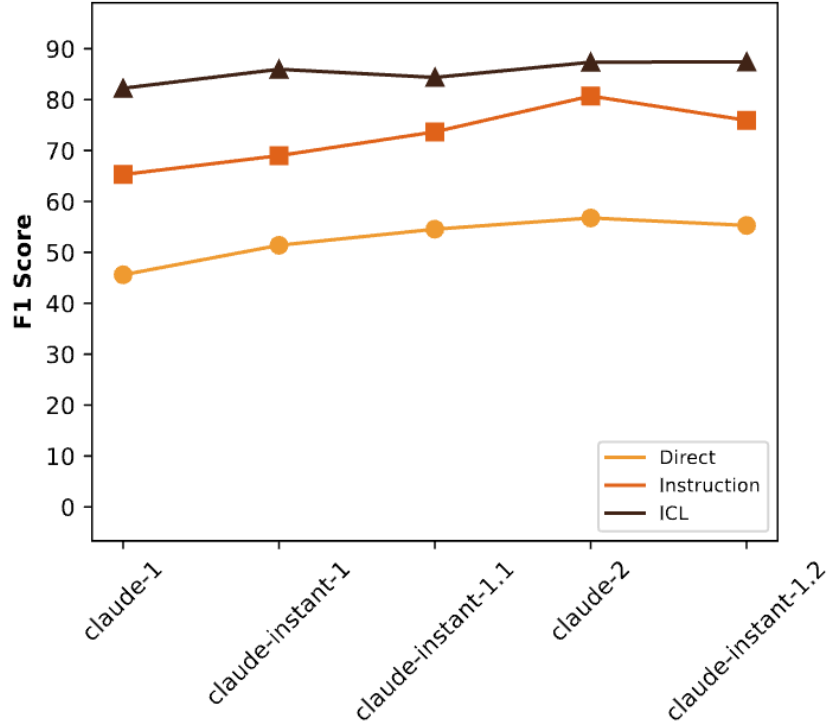
3 Input Forms: Direct, Instruction, ICL

31 LLMs (Some are not shown in the Figure)

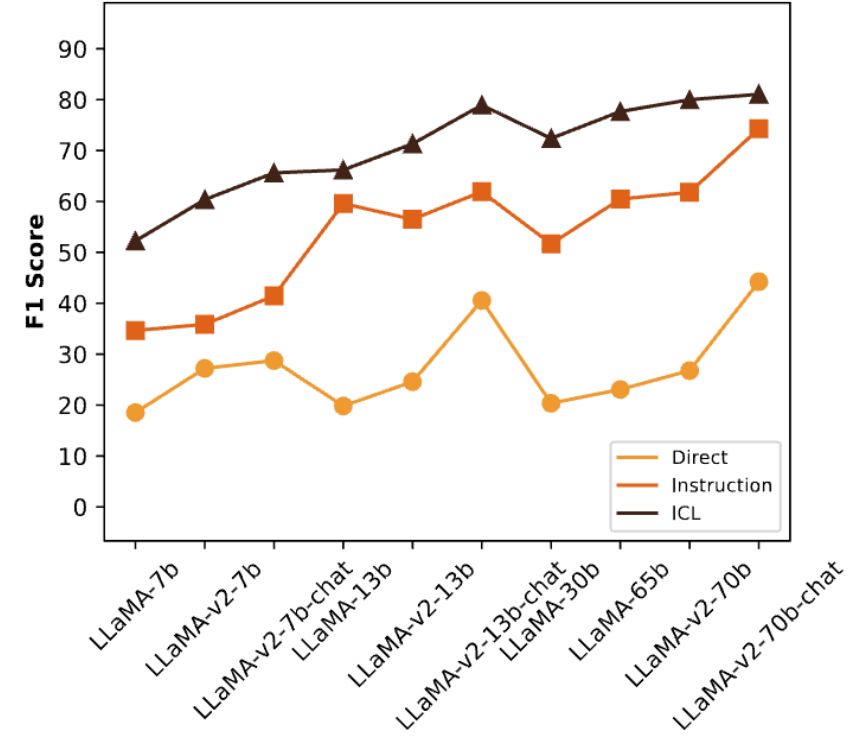
**InstructGPT Series**



**Claude Series**



**LLaMA Series**



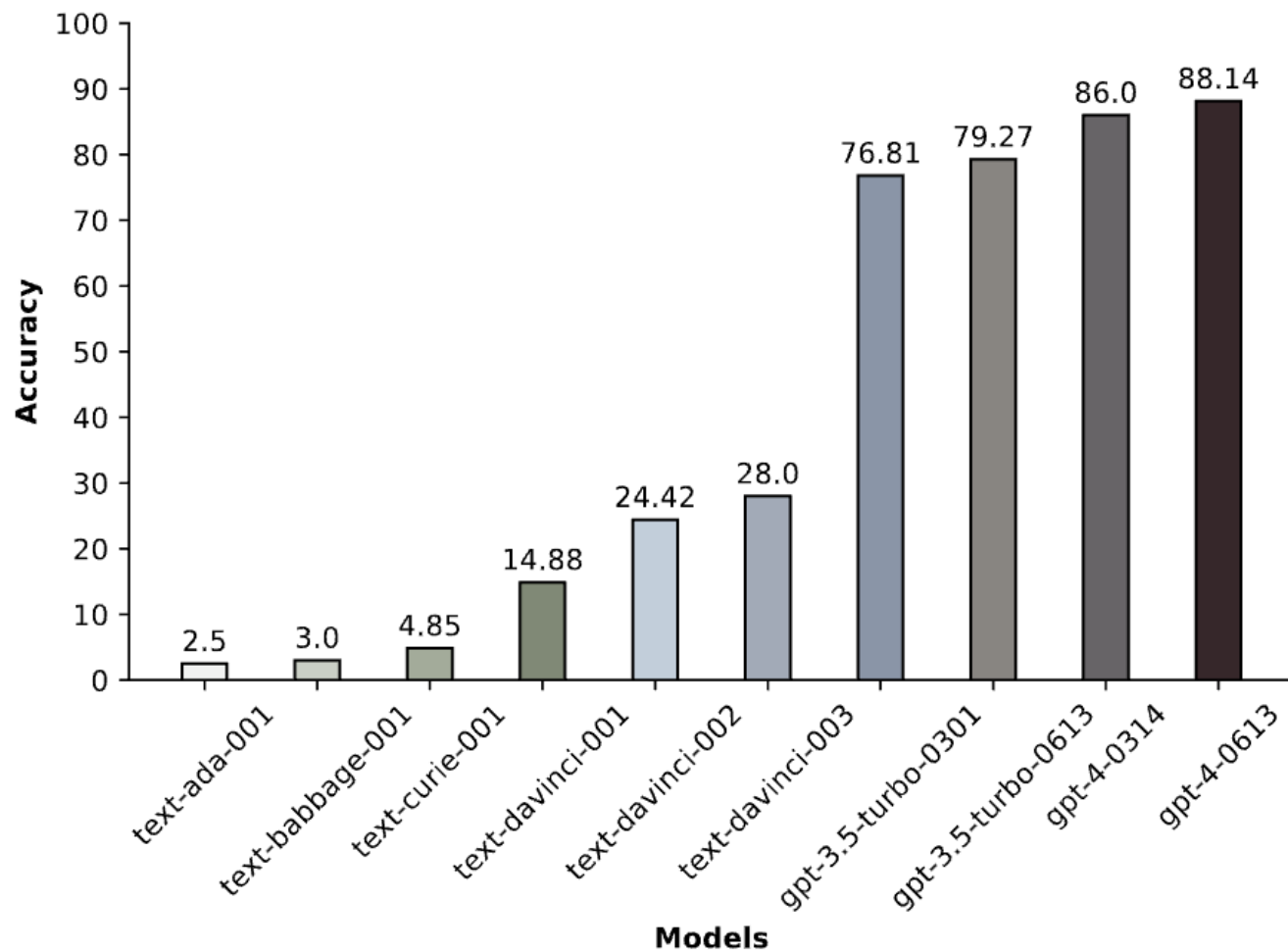
# Experiment: Compared with human judgment

Table 3 shows that using a template-based approach combined with mathematical expression detection can improve the consistency with human judgment. (520 samples 10% of UMWP dataset)

Model	Template	TemplateRule
text-davinci-003	0.732	0.804(+0.072)
claude-1	0.744	0.791(+0.047)
Llama-7b	0.702	0.757(+0.055)
gpt-3.5	0.753	0.802(+0.049)
gpt-4	0.864	0.891(+0.027)

Table 3: Cohen's Kappa comparison between two evaluation methods in the direct input form.

# Supplementary Experiment: Answerable Question Accuracy



## Reference:

<https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k>

# Case Study

---

## Case 1: Making arbitrary assumptions

**Q:** Zoe had 42 bottles of water in her fridge. If **she drank some of them** and then bought 30 more. How many bottles would she have?

**Llama-v2-7b-chat:** Zoe had 42 bottles of water in her fridge to start with. If she drank some of them, then she had fewer bottles of water. **Let's say she drank half of them**, so she had  $42 / 2 = 21$  bottles left. Then, she bought 30 more bottles of water, so she had  $21 + 30 = 51$  bottles of water in her fridge.

## Case 2: Ignoring scope condition

**Q:** The Razorback t-shirt shop sells each t-shirt for \$ 51 dollars. During the Arkansas and Texas tech game they offered a discount of **more than \$ 8** per t-shirt and sold roughly 130 t-shirts. How much money did they make from selling the t-shirts?

**gpt-4-0613:** **The t-shirts were sold for  $\$51 - \$8 = \$43$**  each after the discount. So, the t-shirt shop made approximately  $\$43 * 130 = \$5,590$  from selling the t-shirts.

# Case Study

---

## Case 3: Ignoring unreasonable condition

**Q:** Faye was placing her pencils into **-14 rows** with 11 pencils in each row. How many pencils does she have?

**claude-instant-1.2:** Faye was placing her pencils into **-14 rows**. Each row has 11 pencils. To calculate the total number of pencils, we multiply the number of rows by the number of pencils in each row. So the total number of pencils is: Number of rows = -14. Pencils per row = 11. So total pencils =  $-14 * 11 = -154$ . Therefore, the total number of pencils Faye has is -154.

## Case 4: Misclassifying answerable question

**Q:** Jack received 3 emails in the afternoon, 6 emails in the morning and some more in the evening. If he received a total of 10 emails in the day. How many emails did jack receive in the evening?

**Llama-v2-70b-chat:** The problem states that Jack received some more emails in the evening, but it doesn't provide any information about the number of emails he received in the evening. **Without this information, we cannot determine the total number of emails Jack received in the evening.**