

# BEYOND SOLVING MATH QUIZ: EVALUATING ABILITIES OF LRMS TO *Ask for Information*

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

The recent development of Large Reasoning Models (LRMs) has demonstrated remarkable problem-solving abilities in mathematics, as evaluated by existing benchmarks exclusively on well-defined problems. However, such evaluation setup constitutes a critical gap, since a genuine intelligent agent should not only know how to solve problems (being a math quiz solver), but also know to *ask for information* when the problems lack sufficient information, enabling proactivity in responding users' requests. To bridge such a gap, we propose a novel dataset consisting of two types of incomplete problems with diverse contexts. Based on the dataset, our systematical evaluation of LRMs reveals their inability in proactively asking for information. In addition, we uncover the behaviors related to overthinking and hallucination of LRMs, and highlight the potential and challenges of supervised fine-tuning in learning such ability. We hope to provide new insights in developing LRMs with genuine intelligence, rather than just solving problems.

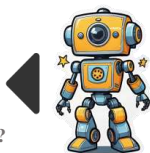
### Incomplete Problem: The shortest distance between two paths in a tree

#### *Ask for Information*

What is the definition of distance?

...

Can you provide more information?



**Genuine Intelligence**



**Math Quiz Solver**

#### *"Solve" the Problem*

Maybe the distance is the minimal ...

...

The shortest distance in a tree is 0.

## 1 INTRODUCTION

Mathematical reasoning, a sequence of steps to draw conclusions from given premises (van Eemeren et al., 2001; Angeles, 1981), presents an interesting challenge to artificial intelligence (AI), driving numerous studies and benchmarks (Gao et al., 2025; Lightman et al., 2024; Face, 2025). Recently, Large Reasoning Models (LRMs) (Jaech et al., 2024; Guo et al., 2025) represent remarkable mathematical reasoning abilities in solving competitive problems (Online, 2025), bolstering the conviction regarding the imminent arrival of genuine AI models. However, the creation of genuine AI should not be exclusively evaluated by solving well-defined mathematical problems in existing benchmarks (Lightman et al., 2024; Gao et al., 2025; Face, 2025; Online, 2025). From the standpoint of AI, John McCarthy said that an agent is intelligent "if it can get additional information from the external world when required" (McCarthy & Hayes, 1981). In addition, real-world problems are inherently characterized by incompleteness that fails to provide all necessary information (Zhang et al., 2024; Deng et al., 2023a; Belinda Z. Li, 2025). Such incompleteness makes proactive information-seeking critical for AI assistants to provide truly helpful responses. For example, a user may ask "My living room is 6 meters long. How many tiles (with a side length of 60 centimeter) do I need in total?", and any specific answer based on AI-made assumptions is meaningless as we do not know the width of the user's living room and need ask for this premise.<sup>1</sup> Therefore, being genuine intelligence in mathematics, LRMs should not only solve problems, but also acquire the ability to *ask for information* on incomplete problems. Otherwise, an impractical math quiz solver is what we will ultimately develop.

<sup>1</sup>Responses of different LRMs are presented in Appendix A.

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

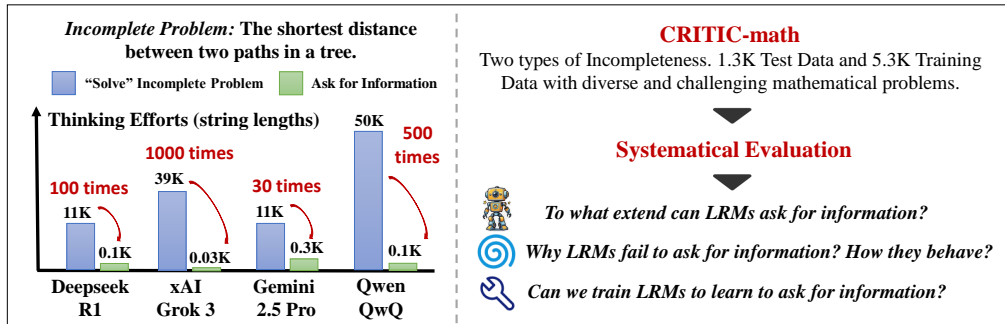


Figure 1: On an incomplete problem (i.e., lacking a definition of "distance"), "solving" the problem exhibits overthinking, manifesting as protracted thinking (measured by thoughts' string lengths), while asking for information reduces thinking efforts. Beyond this example, we provide systematical evaluation of: to what extent can LLMs ask for information, why they fail, and how to improve them.

Unfortunately, the ad-hoc example in Figure 1 (left) alerts us current LLMs being far from genuinely intelligent. Confronted with a problem (proposed by Jiayu Yao (2025)) missing a precise definition, LLMs overthink to infer an answer (which is critically impossible because of lacking premises), resulting in significant thinking efforts and delay in responding. This example reveals the limitations of current LLMs in proactively identifying incomplete problems and asking for information. Beyond this example, **to what extent can LLMs ask for information on incomplete problems? Why LLMs fail to ask questions? And can we train LLMs to know how to ask for information?** These questions remain to be systematically evaluated and play vital roles in developing AI models.

Driven by this concern, in this paper, we introduce CRITIC-math (CRITICThinking of Completeness on mathematical problems, Figure 1 right) to provide systematical evaluation. CRITIC-math contains two types of incomplete problems through rewriting well-defined problems in open-source datasets (Lightman et al., 2024; Gao et al., 2025; Face, 2025). In total, CRITIC-math generates 1.3K test and 5.3K training data, and undergoes manual verification to ensure its quality. Based on CRITIC-math, we reveal the inability of current LLMs in asking for information, and uncover the underlying causes regarding overthinking and hallucination within LLMs' thinking process. We also demonstrate the potential of supervised fine-tuning (SFT) (Ouyang et al., 2022; Muennighoff et al., 2025; Labs, 2025) in learning such ability. In addition, during SFT, we find a dilemma between problem-solving and asking for information, where the current mode of deep-thinking could weaken the ability to ask for information. These results suggest that the existing approach of developing LLMs is biased to only solve math quiz, overlooking the other aspect of intelligence of proactive information-seeking.<sup>2</sup>

In summary, this work makes the following three-fold contributions and offer new insights to develop genuine AI in mathematics that can not only solve problems but also ask for information when needed:

- Propose a new dataset consisting of two types of incompleteness and large-scale problems.
- Based on the dataset, we provide systematical evaluation of state-of-the-art LLMs, uncovering their inability to ask for information and understand how LLMs behave when they fail to ask questions.
- Demonstrate the potential and challenges (a dilemma between the current mode of deep-thinking in solving problems and asking for information) in training LLMs to learn to ask for information.

## 2 CRITIC-MATH

To provide systematical analysis, we introduce CRITIC-math, a new benchmark consisting of two categories of incomplete problems: missing goal and missing premises. We construct CRITIC-math by transforming well-defined problems from open-source datasets into incomplete ones, followed by manual verification to ensure quality. Table 1 and 2 provide illustrative examples and data statistics. In the following, we adopt Deepseek R1 to synthesize data. To avoid model bias, we use Gemini-2.5 Pro with the same process, and evaluate LLMs performances on data synthesized by both models.

<sup>2</sup>We open-source the datasets under Apache 2.0 License using anonymous accounts on HuggingFace. CRITIC-math: <https://huggingface.co/datasets/anonymsub25/CRITIC-math>, CRITIC-math-sft: <https://huggingface.co/datasets/anonymsub25/CRITIC-math-sft>.

108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161

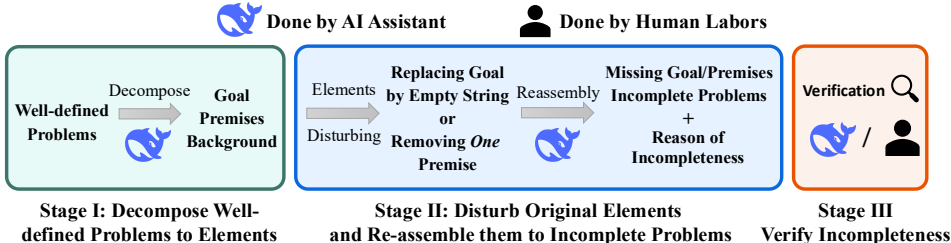


Figure 2: Construction pipeline of CRITIC-math. We rewrite well-defined problems in open-source datasets to incomplete ones in 3 steps: Decomposing, Incomplete Problem Constructing, Verifying.

### 2.1 TYPES OF INCOMPLETE MATHEMATICAL PROBLEMS THAT NEED CLARIFICATION

Reasoning involves a sequence of steps to derive conclusions from premises, and we focus on problems whose goals are assigning a value to a variable (Belinda Z. Li, 2025; Lightman et al., 2024; Gao et al., 2025; Cobbe et al., 2021). Formally, we define reasoning as an assignment function  $\mathcal{A}(y|p_1, p_2, \dots, p_n)$ , where  $y$  is the target variable and  $p$  are premises. A problem is information-complete if  $y$  is well-defined and  $\mathcal{A}$  is injective, i.e., only one value can be assigned given the premises. Based on this, as shown in Table 1, we introduce two types of incomplete mathematical problems:

- **Missing Goal.** When the goal is missing, we cannot know the target space of  $\mathcal{A}$ . Taking the example from Table 1, possible goals can be “how many minutes it takes to cover 1 mile?” or “how many feet can be covered in 1 hour?” Different goals have different values to assign.
- **Missing Premise(s).** When premises are missing, none or multiple values can be assigned to  $y$ . Taking the example from Table 1, the bicycle speeds can vary in  $[1, 10, 20, \dots]$  feet, different premises can result in different conclusions, and no specific value can be assigned.

Table 1: Two types of incomplete problems and a corresponding example.

<b>Well-defined Problem from MATH 500</b> (Lightman et al., 2024):	
A bicycle is traveling at 20 feet per minute. ( <b>Sufficient Premises</b> )	
What is the bicycle’s speed expressed in inches per second? ( <b>Clear Goal</b> )	
<b>Type of Incompleteness</b>	<b>Incomplete Mathematical Problems</b>
Missing Goal	A bicycle is traveling at 20 feet per minute.
Missing Premises	A bicycle is traveling. What is the bicycle’s speed expressed in inches per second?

### 2.2 CONSTRUCTION OF INCOMPLETE MATHEMATICAL PROBLEMS

**Overview.** CRITIC-math is constructed by rewriting the well-defined problems from open-source datasets. This construction involves three key stages: I) Parsing each problem into its constituent elements (i.e., goal, premises, and Background); II) Disturbing elements and constructing incomplete problems by reassembling these disturbed elements; and III) Verifying the resulting incompleteness. Each stage is elaborated in the following sections, and the overall pipeline is illustrated in Figure 2.

**Data Sources.** For our test set, we adopt problems from test sets of Omni-MATH (Cobbe et al., 2021) and MATH 500 (Lightman et al., 2024). For training set, we use problems from OpenR1-Math (Face, 2025). Given the challenging nature of Omni-MATH, where OpenAI’s o1-mini (Jaech et al., 2024) achieves an accuracy of 60% (compared to its 90% accuracy on MATH 500 (Guo et al., 2025)), we limited our selection to Omni-MATH problems with a difficulty level of 4 or lower<sup>3</sup>.

**Stage I – Parsing Problems into Goal and Premises.** We prompt Deepseek R1<sup>4</sup> to decompose each problem from the Data Sources into three elements: Goal, Premises, and Background. The Goal represents the result that the problem asks for. Premises are essential information required to achieve the Goal. Background encompasses any remaining information, such as descriptive text and unnecessary examples. All prompts used in the data construction are provided in Appendix B.1.

<sup>3</sup>MATH 500 is under MIT License, and Omni-MATH and OpenR1-Math are under Apache 2.0 License.

<sup>4</sup>We select Deepseek R1 for its high performance and suitability for our budget

**Stage II – Constructing Incomplete Problems.** We introduce disturbances to each parsed problem. "Missing Goal" are created by replacing the original goal with empty string. "Missing Premises" are generated by randomly removing *one* entry from original premises list (create up to two variants if the original list contains multiple entries). After disturbances, we prompt DeepSeek R1 to reassemble modified elements and untouched elements into free-form problems, mimics real-world scenarios where elements are not explicitly separated. During this process, Deepseek R1 compares the incomplete problem to its original counterpart and generates the reason for incompleteness.

**Stage III – Human-LLM Quality Verification.** We involve a multi-stage verification process. Specifically, we first prompt DeepSeek R1 to analyze the incomplete problems alongside their reasons for incompleteness (generated in Stage II). If Deepseek R1 judges a problem to be genuinely incomplete, we request it to provide at least two different understanding and possible solutions of the problem. Then, human evaluators provide a final confirmation by reviewing sampled rewritten problems and their AI’s judgments. Details on human evaluation and results are in Appendix B.2.

Table 2: Dataset Statistics of CRITIC-math. MP is "Missing Premises", MG is "Missing Goal", and Raw is well-defined problems.

Source	MP	MG	Raw
CRITIC-math Test			
MATH 500	380	234	243
Omni-Math	443	254	266
CRITIC-math Training			
OpenR1-math	3197	2153	5487

**Data Format, Statistics, and Features of CRITIC-math.** Each sample in CRITIC-math is a tuple  $(p, s, r, l_d)$ , where:  $p$  represents the incomplete problem;  $s$  represents the solution to the original problem where  $p$  is rewritten;  $r$  denotes the reason of incompleteness; and  $l_d$  indicates the difficulty of the original problem ( $l_d$  is not applicable in training set, as OpenR1-Math dataset lacks this rating). In addition to the incomplete problems, CRITIC-math also includes the original problems, structured as  $(p, s, \text{None}, l_d)$ , where "None" indicates the absence of incompleteness. Table 2 presents the dataset statistics for CRITIC-math, showing a total of 1311 incomplete problems in the test set and 5350 in the training set<sup>5</sup>. For problems’

difficulty, MATH 500 provides five levels (1 to 5), with the corresponding data ratios for "Missing Premises" (MP) being [10.3%, 17.6%, 16.8%, 25.5%, 29.7%] and for "Missing Goal" (MG) being [9.8%, 17.1%, 17.9%, 25.6%, 29.5%]. Omni-Math provides fine-grained levels [1, 1.5, 2, 2.25, 2.5, 3, 3.5, 4], with some levels have little samples. To ensure sufficient samples per level, we organize Omni-Math levels as follows: (1, 1.5) to 1, (2, 2.25) to 2, (2.5) to 3, (3, 3.5) to 4, and (4) to 5. This results in data ratios: MP [24.2%, 16.3%, 10.6%, 10.8%, 38.1%] and MG [25.2%, 15.4%, 11.8%, 11.42%, 36.2%]. In summary, CRITIC-math features two types of incompleteness, more challenging and diverse mathematical problems, a realistic task formulation (i.e., free-form problems), and a training set that allows us to examine how fine-tuning affects the ability of asking for information.

### 3 EVALUATE *Asking for Information* BASED ON CRITIC-MATH

We evaluate LRMs abilities to ask for information on incomplete mathematical problems. Recognizing that Supervised Fine-Tuning (SFT) can adapt LRMs for specific tasks (Muennighoff et al., 2025; Team, 2025a; Xu et al., 2025; Muennighoff et al., 2025), we study two research questions (RQs):

- **RQ1 (Evaluate LRMs):** To what extend can existing LRMs identify incomplete mathematical problems and raise questions to ask for information?
- **RQ2 (Evaluate SFT Effectiveness):** Can SFT effectively train LLMs to acquire the ability of *asking for information*? Furthermore, does incorporating the deep thinking from powerful LRMs during SFT enhance LLMs’ ability to ask for information?

#### 3.1 EVALUATE LRMs (R1)

##### 3.1.1 EVALUATION SETUPS

To conduct a comprehensive evaluation, we introduce two levels of analysis: a coarse level, focused on assessing LRMs’ overall performance, and a fine-grained level, designed to understand the thinking characteristics of LRMs and reveal underlying phenomena that why LRMs fail to ask for information.

<sup>5</sup>The training problems have been filtered for sampling responses to construct SFT data, cf. Section 3.2.1.

**Prompt Setups:** To examine how LRMs raise clarification questions, we prompt them using two prompts: "Implicit prompt" requires greater proactivity from LRMs in raising questions, and are more closely resemble realistic scenarios where only the problem is provided. In contrast, "explicit prompt" explicitly instructs LRMs to ask for information if needed. The prompts are detailed below.

Implicit Prompt:	Explicit Prompt:
{problem}	# Task to Solve {query}
	# Instruction If you need ask for information, please raise clarification question and start your response STRICTLY with: "Clarification Question" followed by your questions. Otherwise, please provide your answer in <code>\boxed{}</code> .

**Models:** We use Deepseek-R1 (Guo et al., 2025), Qwen3-plus thinking (Team, 2025b), Claude 3.7 thinking (Antropic, 2025), OpenAI o3-mini (Jaech et al., 2024), and Grok-3-mini (xAI, 2025). These models are the state-of-the-art from various institutions, renowned for advanced reasoning skills.

**Evaluation Metrics:** For the coarse level evaluation of the overall performance regarding raising questions to ask for information and the corresponding thinking efforts, we consider the following metrics, with all evaluations are conducted on both well-defined and incomplete problems.

- Clarification Ratio (CR) is the percentage of responses that raise questions. For "implicit prompts", we utilized LLM-as-a-Judge (Gu et al., 2024) (Deepseek R1) to determine whether a response raised questions. For "explicit prompt", since it requests raising questions by starting with "clarification", we simply check for the presence of the string "clarification".
- Thoughts Lengths in Clarifying (TLC), and Thoughts Lengths in No Clarifying (TLNC) are average thoughts lengths generated by LRMs when raising and not raising questions.
- Clarification Accuracy (ACC) is the overall percentage of responses that raise clarification questions on incomplete problems and not raise clarification questions on well-defined problems.

Fine-grained level evaluates thoughts when fail to ask for information.<sup>6</sup> We split thoughts by `\n\n` Zhang et al. (2025) and use following metrics. ROR and CNR are evaluated by LLM-as-a-Judge<sup>7</sup>. To avoid model bias, we adopt GPT-4.1 and Claude-3.7 as additional judge LLMs (Refer Appendix D.1.2 for results). We observe no significant differences among using different judge LLMs.

- Reflection Step (RS) is the number of steps that contains reflection keywords. Following Hu et al. (2025), we select reflection keywords to include 'alternative', 'wait', 'but', and 'check'.
- Reflection on Incompleteness Ratio (ROR) is the ratio of thoughts whose Reflection Steps specifically aimed at addressing the incompleteness of the problem. In specific, this includes making assumptions about the missing premises and imagining a goal when the goal is missing.
- Clarification Noticing Ratio (CNR) is the ratio of thoughts that identified the need for clarification.

**Implementation Details.** All tested LRMs are accessed via their respective APIs. To ensure deterministic output, we set the temperature to 0 and generate a single response per problem. Appendix D explores experimental results with temperature greater than 0 and multiple responses. Our analysis demonstrates there is no significant differences between sampling one and multiple responses. In addition, Appendix C details the implementation of our LLM-as-a-Judges. Finally, when calculating the thoughts length, we adopt different methods depending on the return format of different API's. For Deepseek R1, we measure the string lengths of the thoughts returned by APIs. For other LRMs, we directly use the number of "reasoning tokens" reported in the API's usage data.

### 3.1.2 RESULTS & DISCUSSION

Table 7 presents the results of our coarse-grained evaluation, while Table 9 details the fine-grained evaluation results<sup>8</sup>. An integrated analysis on both tables yields the following observations:

<sup>6</sup>This is only applicable on Deepseek-R1 and Qwen3 as other APIs do not return the thoughts contents.

<sup>7</sup>For longer thoughts, only the first ten steps are evaluated.

<sup>8</sup>In the main page, we report results on dataset synthesized by Deepseek R1. Please refer Appendix D.1.1 for results on dataset synthesized by Gemini-2.5 Pro. We observe no significant differences on these two datasets.

Table 3: Coarse-grained assessment of overall performance. For readability, TLCs/TLNCs are reported as the multiples of the TLNC on well-defined problems, with the absolute lengths shown in parentheses. Multiples greater than 1 are marked in **red**; otherwise, in **green**. Results obtained using "implicit prompt" are indicated with **gray lines**; otherwise, with white.

Models	Missing Premises			Missing Goal			Well-Defined			ACC
	CR	TLC	TLNC	CR	TLC	TLNC	CR	TLC	TLNC	
CRITIC-math (from Math 500 Data Source)										
Deepseek R1	48.68%	0.61	2.11	48.72%	0.49	1.05	1.23%	3.08	1 (6387)	62.89%
	21.05%	2.35	2.15	4.27%	1.73	1.34	0.41%	4.24	1 (7973)	38.74%
Qwen3 Plus	51.58%	0.40	1.78	58.12%	0.41	0.99	0.00%	nan	1 (3578)	67.09%
	21.58%	1.81	1.49	35.48%	1.56	1.07	0.00%	nan	1 (4087)	39.91%
o3 mini	50.70%	1.00	2.07	35.04%	1.34	1.09	0.41%	5.90	1 (694)	60.33%
	27.37%	1.43	2.63	19.23%	0.93	1.23	0.00%	nan	1 (738)	45.74%
Grok 3 mini	55.70%	0.49	1.52	48.72%	0.48	0.97	0.00%	nan	1 (1525)	66.39%
	34.21%	1.14	1.27	15.81%	0.68	1.06	0.00%	nan	1 (1587)	47.84%
Claude 3.7	53.42%	0.42	2.07	44.87%	0.16	0.93	2.47%	0.02	1 (4818)	63.59%
	33.68%	0.66	1.42	11.97%	0.56	0.94	0.41%	6.01	1 (3882)	46.44%
CRITIC-math (from Omni-Math Data Source)										
Deepseek R1	44.02%	0.34	1.27	53.54%	0.27	0.83	3.76%	0.43	1 (14177)	60.96%
	18.74%	1.71	1.41	3.15%	1.70	1.01	3.38%	2.53	1 (15583)	36.14%
Qwen3 Plus	46.73%	0.30	1.34	57.87%	0.24	1.01	5.64%	0.54	1 (6284)	62.82%
	27.65%	1.47	1.15	9.45%	1.98	0.95	4.14%	1.54	1 (6740)	37.80%
o3 mini	48.31%	0.63	1.37	40.16%	0.45	0.68	6.39%	0.38	1 (2085)	58.67%
	20.32%	0.86	1.65	20.87%	0.33	0.88	2.63%	0.50	1 (2143)	41.74%
Grok 3 mini	53.05%	0.36	1.18	55.38%	0.38	0.89	8.27%	0.37	1 (2433)	64.28%
	31.83%	1.00	1.07	20.47%	0.52	0.90	3.38%	1.55	1 (2379)	46.73%
Claude 3.7	50.34%	0.31	1.89	46.46%	0.11	1.10	7.89%	0.02	1 (7094)	60.85%
	23.48%	1.34	1.97	14.57%	0.58	1.26	2.26%	2.64	1 (4642)	41.64%

Table 4: Fine-grained analysis of LRMs *when fail to ask for information*. We report RS on well-defined problems for comparison. Results obtained using "implicit prompt" are indicated in **gray**; otherwise, in white. We average results of two data sources considering limited pages.

Model	Missing Premises			Missing Goal			Well-Defined
	RS	ROR	CNR	RS	ROR	CNR	RS
Deepseek R1	36.58	74.04%	21.22%	17.21	97.68%	3.78%	23.39
	50.38	77.27%	24.33%	29.55	94.77%	1.06%	26.41
Qwen3 Plus	79.79	76.43%	18.57%	50.37	98.45%	4.88%	47.79
	84.51	81.40%	28.35%	59.28	94.60%	3.80%	87.39

**Overall, LRMs lack the ability to Ask for Information.** In Table 7, LRMs achieve significantly low clarification ratios (CRs, around 25%) and accuracies (ACCs, around 40%) when prompted with only problems ("implicit prompt"), and explicitly instructing LRMs to ask for information ("explicit prompt") can improve CRs from 25% to 50% and ACCs from 40% to 65%. These results indicate that LRMs lack the ability to proactively ask for information. In most cases, LRMs act as math quiz solvers to "solve" problems regardless whether the problem is complete or not. Even given "explicit prompt", CRs are around 50%, indicating that LRMs struggle to effectively ask for information. These phenomena are even more pronounced on difficult problems. As detailed in Appendix D.1.3, we observe a negative correlation between problem difficulty and CRs: CRs decrease on harder problems. Unexpectedly, in Appendix D.1.3, we find that when LRMs ask for information, their questions are rather accurate that target at the incompleteness. For such, we hypothesize that LRMs actually lack the ability to ask rather than to notice the incompleteness. LRMs tend to "solve" a problem even know it is incomplete. Our fine-grained analysis further support this hypothesis.

**Fine-grained Evaluation** assesses LRMs' behavior modes when LRMs fail to ask for information. In total, we identify two major failure modes (*thoughts-to-answer unfaithfulness* and *overthinking*) on missing premises problems and one dominate failure mode (*hallucination*) on missing goal problems.



324 • **Thoughts-to-Answer Unfaithfulness When Missing Premises.** CNR is the ratio of thoughts that  
 325 have recognized the need to ask. The results in Table 9 suggest that, confronted with missing premises  
 326 problems, in approximately 20% of failure cases, LRMs recognize the need to ask for information,  
 327 but ultimately fail to do so. We provide a case for illustration below, where the thought recognizes  
 328 the necessity ("*the best approach*") and provides a candidate question ("*Are there any specific ...*").  
 329 However, the LRM fails to follow its thought, instead generating an answer. This exemplifies a type of  
 330 thoughts-to-answer unfaithfulness in LRMs, and further support our hypothesis that LRMs know the  
 331 incompleteness but tend to be math quiz solvers rather than be proactive to ask for information.

332 **Thoughts:** If I don't ask, I might give the wrong answer. So the best approach here is to request more  
 333 details. Clarification Question: Are there any specific conditions given for ...

334 **Answers:** To determine ... **\*\*Final Answer\*\*:** If the equation has equal roots, then  $\boxed{\pm 2}$ .

336 • **Overthinking when Missing Premises.** As Table 7 shows, confronted with missing premises prob-  
 337 lems, TLNCs (thoughts lengths in no clarifying) significantly increase. The results in Table 9 further  
 338 reveal that when LRMs fail to ask for information, their thoughts exhibit higher RSs (reflection steps).  
 339 These thoughts, as measured by RORs (Reflection on Incompleteness Ratio), contains reflection  
 340 steps that focus on the incompleteness in roughly 75% cases. In together, if problems lack premises,  
 341 LRMs tend to address such deficiency through their internal thinking, failing to ask for information,  
 342 leading to significant delays. In contrast, TLCs (thoughts lengths in clarifying) are much shorter,  
 343 demonstrating the efficiency of asking questions. In addition, we find the overthinking to be more  
 344 pronounced on easy problems (detailed in Appendix D), aligning with previous findings (Chen et al.,  
 345 2024b). Such overthinking also supports our hypothesis, as overthinking on incompleteness indeed  
 346 indicates that LRMs have noticed the incompleteness. However, LRMs more likely to rely on internal  
 347 thinking to address the incompleteness and "solve" the problem, rather than asking for information.

348 • **Hallucinations when Missing Goals.** As shown in Table 9, in more than 90% cases (RORs), the  
 349 thoughts address the incompleteness by imagining a goal. We consider this behavior as instruction  
 350 inconsistency (Huang et al., 2025b), where the output deviates from users' directive. Interestingly,  
 351 TLNCs are shorter compared to well-defined problems, suggesting that LRMs tend to imagine goals  
 352 that can be quickly addressed. For illustration, refer to Appendix D for case studies.

353 Unfaithfulness and overthinking explain 80% failures in missing premises, and hallucination explains  
 354 95% failures in missing goals. We leave analyzing other minor cases of behaviors to future works.  
 355 These results help us understand how and why LRMs fail to ask questions: current LRMs are more like  
 356 math quiz solvers, tending to "solve" incomplete problems by three major kinds of crooked behaviors  
 357 instead of asking for information even explicitly prompted and recognized the incompleteness.

## 359 3.2 EVALUATE SFT EFFECTIVENESS (RQ2)

### 360 3.2.1 EVALUATION SETUPS

362 **SFT Training Data.** Following the method of distilling LLMs (Hsieh et al., 2023; Hinton et al.,  
 363 2015), we use Deepseek R1 with "explicit prompt" to generate data for SFT. Specifically, for each  
 364 problem in the training set of CRITIC-math: (1) For well-defined problems, we store answers and  
 365 thoughts only if the answers are correct (determined by Math-Verify (Kydliček)). (2) For incomplete  
 366 problems, we store answers and thoughts if the answers raise clarification questions (determined  
 367 by string-matching on "clarification"). The training set was filtered to include only well-defined  
 368 problems with correct answers and incomplete problems with answers raising clarification questions.  
 369 This process results in a total of 10.8K SFT samples.

370 **Implementation Details:** Using Qwen3-8B-Base, we train two models and their variants: **CRITIC-**  
 371 **Qwen**, using answers only, and **CRITIC-Qwen-thinking**, using both thoughts and answers. For  
 372 ablation studies, we train **CRITIC-Qwen<sub>WI</sub>** and **CRITIC-Qwen-thinking<sub>WI</sub>** to analyze the impact  
 373 of problem types, where "W" indicates trained on well-defined problems and "I" indicates trained on  
 374 incomplete problems. We use OpenRLHF (Hu et al., 2024) to SFT for 1 epoch, with a learning rate  
 375 of  $1e-5$  and batch size of 64.<sup>9</sup> Training CRITIC-Qwen took 10 minutes and CRITIC-Qwen-thinking  
 376 took 60 minutes on 8 NVIDIA A800s. We use vLLM (Kwon et al., 2023) to infer responses. We

377 <sup>9</sup>Other hyperparameters, like the learning rate schedule and warmup steps, are kept at their default values.

conduct additional experiments using Llama-3.1-8B-Instruct. Please refer Appendix D.2 for detail results. We observe no significant differences in the results between the choices of backbone LLMs.

**Evaluation Metrics:** We follow the metrics in Section 3.1.1. Notably, according to the training data-format, we prompt SFT models by "explicit prompt" according to training data-format. To further demonstrate the effectiveness of SFT in improving mathematical reasoning, we introduce Solved Ratio (SR), defined as the percentage of answers that correctly solve the well-defined problems.

### 3.2.2 RESULTS & DISCUSSION

Table 5: Coarse level evaluation of SFT models. *CRITIC-Qwen T* denotes CRITIC-Qwen-thinking. *Qwen3-8B T* denotes the official LRM (also fine-tuned from Qwen3-8B-Base) with thinking enabled.

Model	Missing Premises			Missing Goal			Well-Defined				ACC
	CR	TLC	TLNC	CR	TLC	TLNC	CR	TLC	TLNC	SR	
Math 500 Subset											
CRITIC-Qwen	78.42%	/	/	94.87%	/	/	4.12%	/	/	73.39%	87.86%
CRITIC-Qwen T	57.37%	0.22	2.21	62.82%	0.25	0.93	1.23%	1.59	1 (4554)	80.83%	70.60%
Qwen3-8B T	51.58%	0.39	1.71	50.43%	0.41	1.05	0.00%	nan	1 (3853)	97.53%	64.99%
Omni-Math Subset											
CRITIC-Qwen	77.88%	/	/	97.64%	/	/	13.16%	/	/	41.13%	85.57%
CRITIC-Qwen T	56.66%	0.14	1.51	69.29%	0.14	1.09	7.52%	0.32	1 (8895)	54.88%	69.89%
Qwen3-8B T	45.82%	0.29	1.32	54.72%	0.30	0.90	5.26%	0.32	1 (7027)	83.74%	61.68%

Table 6: Fine-grained analysis of SFT model *when fails to ask for information*. Only CRITIC-Qwen-thinking can be evaluated. we average results of two data sources considering limited pages.

Model	Missing Premises			Missing Goal			Well-Defined
	RS	ROR	CNR	RS	ROR	CNR	Rs
CRITIC-Qwen T	82.84	74.01%	21.75%	44.10	92.12%	6.06%	41.30

Table 13 presents the results of our coarse-grained evaluation, while Table 6 details the fine-grained evaluation results. An integrated analysis on both tables yields the following observations:

**SFT Improves the Ability to Ask for Information.** CRITIC-Qwen achieves higher CRs on incomplete problems. Although more false positives on well-defined problems, the accuracies are 87.86% and 85.57%, surpassing the highest accuracy of close-source LRMs (67.09% and 64.28% achieved by Grok 3 mini). CRITIC-Qwen-thinking also surpasses close-source LRMs. In addition, both SFT models surpass Qwen3-8B T, which fine-tuned from the same backbone but not tailored for asking for information. These results demonstrate the effectiveness of SFT in learning to ask for information.

**A Dilemma between Deep Thinking and Asking for Information.** As shown in Table 13 and 6, when SFT with thinking process, solved ratios on well-defined problems increase, demonstrating improved mathematical reasoning. However, CRs on incomplete problems significantly decrease, mirroring the overthinking and hallucinations observed in the closed-source LRMs discussed in Section 3.1.2: large TLNCs/RSs and high RORs indicate that CRITIC-Qwen-thinking overthinks on missing premises. Additionally, the barely changed TLNCs and high RORs on missing goal problems suggest that CRITIC-Qwen-thinking imagines and accomplishes a goal not specified by the input.

These results demonstrate a dilemma between the current mode of deep-thinking, which encourages self-reflections in solving problems (Muennighoff et al., 2025; Guo et al., 2025), and asking for information. Given high RSs and RORs in Table 9 and 6, we argue that one root cause is that the current deep-thinking makes LRMs treat missing information as an internal issue (poor understanding of input) and stimulating self-reflections. This represents a flaw in current research: the development of LRMs is significantly biased towards problem-solving, which overemphasizes extensive internal thinking to address any issue (including the incompleteness) and "solve" problems. Ablation studies, detailed in Appendix D, further support our analysis, where CRs of CRITIC-Qwen-thinking increase as its self-reflections are less strengthened by learning on well-defined problems. In addition, there is no significant differences between CRITIC-Qwen-thinking<sub>w</sub> and CRITIC-Qwen-thinking on well-defined problems, indicating that asking for information has limited side-effects on solving problems.



## 4 RELATED WORKS

**The Ability to Identify Incomplete Problem.** In psychology, a child’s ability to identify missing information in mathematical problems is considered an important indicator of their intelligence (Dempster & Corkill, 1999; Edens & Potter, 2008), reflecting their use of schematic knowledge (understanding problem structures) (Low & Over, 1989) and metacognition (supervising one’s thoughts) (Lai, 2011; Medina et al., 2017). Current LRMs demonstrate advanced reasoning abilities, inspiring recent exploration into their capabilities from a cognitive perspective (Gandhi et al., 2025). Our work examines the behaviors and limitations of LRMs in identifying incomplete mathematical problems, revealing challenges in developing AI with human-level cognition. In addition, proactively identifying and gathering missing information is a fundamental philosophical problem in AI (McCarthy & Hayes, 1981), constituting central topics in various topics (Ren et al., 2021; Gal et al., 2017; Sutton et al., 1998; Curtis et al., 2024; Piquepal & Toussaint, 2019). Following this tradition, our work suggests that LRMs currently over-rely on internal thinking and struggle to effectively ask for information yet.

**Clarifying User Requests.** Incompleteness frequently appears in natural language due to various factors (Piantadosi et al., 2012; Wasow et al., 2005; Degani & Tokowicz, 2010), and recognizing and resolving it has posed long-standing challenges for developing AI models (Dreyfus, 1972; Anwar et al., 2024; Huang et al., 2025a). In the era of LLMs, extensive research has explored ambiguity in various scenarios, including conversation (Zhang et al., 2024), semantics Kuhn et al. (2023), question-answering (Min et al., 2020), and chatbots Chen et al. (2024a). To improve abilities of LLMs to handle ambiguity, recent approaches utilize Chain-of-Thought, few-shot prompting, or training adapters (Chen et al., 2024c; Deng et al., 2023b; Kuhn et al., 2022; Cole et al., 2023) to develop assistants that proactively understand users’ intents. In this paper, given the current trending research of LRMs, we extend the research scope of clarifying questions from general domain to mathematics, and argue that asking for information on incomplete problems stands for a critical aspect to develop genuine intelligence. We also explore the potential and challenges in learning such ability, hoping to providing insights to develop LRMs that can ask for information.

**Benchmarks in Asking for Information.** Existing benchmarks primarily focus on users requests in general domains like task-oriented dialogue (Budzianowski et al., 2018; Rastogi et al., 2020; Zhang et al., 2024; Basile et al., 2021; Wan et al., 2023). Limited efforts are dedicated to analyzing LRMs in handling incomplete math problems. Given the increasing prevalence of LRMs and current exclusive evaluation on well-defined problems, we propose a new dataset to systematically evaluate LRMs in asking for information on incomplete problems. We acknowledge a concurrent work, QuestBench (Belinda Z. Li, 2025) (released March 28, 2025 on Arxiv), that examines LLMs’ information-gathering in reasoning tasks. However, QuestBench only evaluates can LLMs pick up the correct question among several candidate choices. Such setting fails to evaluate the proactivity of asking for information, and does not discuss the relationship between problem-solving and asking questions, where a dilemma is found in this work. CRITIC-math aims to provide more comprehensive evaluation of LRMs in proactively asking for information, extending current scope of solving problems.

## 5 CONCLUSIONS

Existing evaluation of LRMs exclusively focuses on solving well-defined mathematical problems, ignoring another critic aspect of genuine intelligence regarding asking for information on incomplete problems. To bridge such gap, we introduce a new benchmark, CRITIC-math, provide systematical evaluation, and find LRMs’ inability to identify incompleteness and proactively seek clarification. Our observed dilemma between the current mode of deep thinking and asking for information also demands a fundamental shift in how we develop LRMs. We hope to offer guidance and insights for developing LRMs that transcend solely being math quiz solver, evolving into genuine intelligent agents capable requiting information, navigating uncertainties inherent in real-world scenarios.

Moving forward, we intend to broaden the scope of our research to new domains, such as computer programming. This expansion, however, necessitates overcoming several challenges. Chief among them are the increased contextual diversity—spanning multiple programming languages and frameworks—and the pronounced scarcity of datasets structured suitably for evaluating information-seeking tasks. Such datasets, which are essential for our methodology, are not yet readily available. Nevertheless, we are committed to advancing this line of inquiry as a vital progression of our work.

## 6 ETHICS STATEMENT

This paper focuses on mathematical reasoning and our major argument is that a genuine intelligence should not know solving problems, but also asking for information. We affirm that this study fully complies with the ICLR Code of Ethics. All data are sourced from public, open-source repositories, obviating the use of human participants or sensitive personal data. All procedures, including any human-in-the-loop verification performed under controlled conditions.

## 7 REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we publicly release the full dataset. CRITIC-math: <https://huggingface.co/datasets/anonysub25/CRITIC-math>, CRITIC-math-sft: <https://huggingface.co/datasets/anonysub25/CRITIC-math-sft> (both under Apache 2.0 License). We have also provided the detail experimental setups in the this paper. In addition, the corresponding code are provided in the supplementary files and will be open-sourced.

## REFERENCES

- Peter Adam Angeles. Dictionary of philosophy. 1981.
- Anthropic. Claude 3.7 sonnet and claude code. <https://www.anthropic.com/news/claude-3-7-sonnet>, 2025. Accessed: 2025-02-25.
- Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We need to consider disagreement in evaluation. *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, 2021. URL <https://api.semanticscholar.org/CorpusID:236486317>.
- Zi Wang Belinda Z. Li, Been Kim. Questbench: Can llms ask the right question to acquire information in reasoning tasks? *arXiv preprint arXiv:2503.22674*, 2025.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, 2018.
- Sanxing Chen, Sam Wiseman, and Bhuwan Dhingra. Chatshop: Interactive information seeking with language agents. *arXiv preprint arXiv:2404.09911*, 2024a.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for  $2+3=?$  on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024b.
- Yue Chen, Chen Huang, Yang Deng, Wenqiang Lei, Dingnan Jin, Jia Liu, and Tat-Seng Chua. Style: Improving domain transferability of asking clarification questions in large language model powered conversational agents. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10633–10649, 2024c.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 530–543, 2023.

- 540 Aidan Curtis, George Matheos, Nishad Gothoskar, Vikash Mansinghka, Joshua Tenenbaum, Tomás  
541 Lozano-Pérez, and Leslie Pack Kaelbling. Partially observable task and motion planning with  
542 uncertainty and risk awareness. *arXiv preprint arXiv:2403.10454*, 2024.
- 543 Tamar Degani and Natasha Tokowicz. Semantic ambiguity within and across languages: An integra-  
544 tive review. *Quarterly Journal of Experimental Psychology*, 63(7):1266–1303, 2010.
- 545 Frank N Dempster and Alice J Corkill. Interference and inhibition in cognition and behavior: Unifying  
546 themes for educational psychology. *Educational Psychology Review*, 11:1–88, 1999.
- 547 Yang Deng, Wenqiang Lei, Minlie Huang, and Tat-Seng Chua. Rethinking conversational agents in  
548 the era of llms: Proactivity, non-collaborativity, and beyond. In *SIGIR-AP*, pp. 298–301, 2023a.  
549 URL <https://doi.org/10.1145/3624918.3629548>.
- 550 Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting  
551 and evaluating large language models for proactive dialogues: Clarification, target-guided, and  
552 non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*,  
553 pp. 10602–10621, 2023b.
- 554 Hubert L Dreyfus. What computers can’t do: The limits of artificial intelligence. 1972.
- 555 Kellah Edens and Ellen Potter. How students “unpack” the structure of a word problem: Graphic  
556 representations and problem solving. *School Science and Mathematics*, 108(5):184–196, 2008.
- 557 Hugging Face. Open rl: A fully open reproduction of deepseek-rl, January 2025. URL <https://github.com/huggingface/open-rl>.
- 558 Yarín Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data.  
559 In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.
- 560 Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive  
561 behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv*  
562 *preprint arXiv:2503.01307*, 2025.
- 563 Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao  
564 Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran  
565 Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang.  
566 Omni-MATH: A universal olympiad level mathematic benchmark for large language models.  
567 In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=yaqPf0KA1N>.
- 568 Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen,  
569 Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*,  
570 2024.
- 571 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
572 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-rl: Incentivizing reasoning capability in llms  
573 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 574 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*  
575 *preprint arXiv:1503.02531*, 2015.
- 576 Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner,  
577 Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger  
578 language models with less training data and smaller model sizes. In *Findings of the Association*  
579 *for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.
- 580 Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An  
581 easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*,  
582 2024.
- 583 Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum.  
584 Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base  
585 model. *arXiv preprint arXiv:2503.24290*, 2025.

- 594 Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Xiangji Huang.  
595 How to enable effective cooperation between humans and nlp models: A survey of principles,  
596 formalizations, and beyond. *arXiv preprint arXiv:2501.05714*, 2025a.
- 597  
598 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong  
599 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language  
600 models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information  
601 Systems*, 43(2):1–55, 2025b.
- 602 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec  
603 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint  
604 arXiv:2412.16720*, 2024.
- 605  
606 Kunpeng Ning Jiayu Yao. Reasoning attack: Inducing llm to never-end thinking, 2025.  
607 <https://github.com/PKU-YuanGroup/Reasoning-Attack/tree/master>.
- 608 Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Clam: Selective clarification for ambiguous  
609 questions with generative language models. *arXiv preprint arXiv:2212.07769*, 2022.
- 610  
611 Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances  
612 for uncertainty estimation in natural language generation. In *The Eleventh International Confer-  
613 ence on Learning Representations*, 2023. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=VD-AYtP0dve)  
614 [VD-AYtP0dve](https://openreview.net/forum?id=VD-AYtP0dve).
- 615  
616 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.  
617 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model  
618 serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating  
619 Systems Principles*, 2023.
- 620  
621 Hynek Kydlíček. Math-Verify: Math Verification Library. URL [https://github.com/  
huggingface/math-verify](https://github.com/huggingface/math-verify).
- 622  
623 B. Bespoke-stratos Labs. The unreasonable effectiveness of reasoning distillation.  
624 <https://hf.co/bespokelabs/Bespoke-Stratos-32B>, 2025. Accessed: 2025-01-22.
- 625  
626 Emily R Lai. Metacognition: A literature review. 2011.
- 627  
628 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan  
629 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth  
630 International Conference on Learning Representations*, 2024. URL [https://openreview.  
net/forum?id=v8L0pN6EOi](https://openreview.net/forum?id=v8L0pN6EOi).
- 631  
632 Renae Low and R. Over. Detection of missing and irrelevant information within algebraic story  
633 problems. *British Journal of Educational Psychology*, 59:296–305, 1989. URL [https://api.  
semanticscholar.org/CorpusID:123199718](https://api.semanticscholar.org/CorpusID:123199718).
- 634  
635 J. McCarthy and P.J. Hayes. Some philosophical problems from the standpoint of artificial  
636 intelligence. In Bonnie Lynn Webber and Nils J. Nilsson (eds.), *Readings in Artificial In-  
637 telligence*, pp. 431–450. Morgan Kaufmann, 1981. ISBN 978-0-934613-03-3. doi: <https://doi.org/10.1016/B978-0-934613-03-3.50033-7>. URL [https://www.sciencedirect.  
638 com/science/article/pii/B9780934613033500337](https://www.sciencedirect.com/science/article/pii/B9780934613033500337).
- 639  
640 Melissa S Medina, Ashley N Castleberry, and Adam M Persky. Strategies for improving learner  
641 metacognition in health professional education. *American journal of pharmaceutical education*, 81  
642 (4):78, 2017.
- 643  
644 Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering  
645 ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.
- 646  
647 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke  
Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time  
scaling. *arXiv preprint arXiv:2501.19393*, 2025.

- 648 AoPS Online. 2024 aime i. URL: [https://artofproblemsolving.com/wiki/index.php/2024\\_AIME\\_I](https://artofproblemsolving.com/wiki/index.php/2024_AIME_I),  
649 2025.
- 650
- 651 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
652 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow  
653 instructions with human feedback. *Advances in neural information processing systems*, 35:27730–  
654 27744, 2022.
- 655 Camille Phiquepal and Marc Toussaint. Combined task and motion planning under partial observ-  
656 ability: An optimization-based approach. In *2019 International Conference on Robotics and  
657 Automation (ICRA)*, pp. 9000–9006. IEEE, 2019.
- 658
- 659 Steven T Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in  
660 language. *Cognition*, 122(3):280–291, 2012.
- 661 Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards  
662 scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings  
663 of the AAAI conference on artificial intelligence*, volume 34, pp. 8689–8696, 2020.
- 664 Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen,  
665 and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40,  
666 2021.
- 667
- 668 Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT  
669 press Cambridge, 1998.
- 670 NovaSky Team. Sky-t1: Train your own o1 preview model within \$450. [https://novasky-  
671 ai.github.io/posts/sky-t1](https://novasky-ai.github.io/posts/sky-t1), 2025a. Accessed: 2025-01-09.
- 672
- 673 Qwen Team. Qwen3: Think deeper, act faster. <https://qwenlm.github.io/blog/qwen3/>, 2025b.  
674 Accessed: 2025-04-29.
- 675 Frans H. van Eemeren, Rob Grootendorst, Sally Jackson, Scott Jacobs, Agnes van Rees, Fran-  
676 cisca Snoeck Henkemans, Eveline T. Feteris, Erik C. W. Krabbe, Agnes Verbiest, Charles Arthur  
677 Willard, and John Hayden Woods. What is reasoning? what is an argument?\*. 2001. URL  
678 <https://api.semanticscholar.org/CorpusID:15543854>.
- 679
- 680 Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone’s voice matters: Quantifying annotation  
681 disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial  
682 Intelligence*, volume 37, pp. 14523–14530, 2023.
- 683 Thomas Wasow, Amy Perfors, and David Beaver. The puzzle of ambiguity. *Morphology and the web  
684 of grammar: Essays in memory of Steven G. Lapointe*, pp. 265–282, 2005.
- 685
- 686 xAI. Grok 3 beta — the age of reasoning agents. URL: <https://x.ai/blog/grok-3>, 2025.
- 687 Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang,  
688 Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning  
689 systems? *arXiv preprint arXiv:2501.11284*, 2025.
- 690
- 691 Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru  
692 Liang, and Tat-Seng Chua. Clamber: A benchmark of identifying and clarifying ambiguous  
693 information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the  
694 Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, 2024.
- 695 Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu,  
696 Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical  
697 reasoning. *arXiv preprint arXiv:2501.07301*, 2025.
- 698
- 699
- 700
- 701

## A RESPONSES OF DIFFERENT LRMS

In this section, we present the responses of different LRMs in responding the question “My living room is 6 meters long. How many tiles (with a side length of 60 centimeter) do I need in total?” Among LRMs, only o3-mini asks for information. The others make assumptions to solve the problem.

**o3-mini:** To figure out the total number of tiles you’ll need, I also need the width of your living room (in meters). With both length and width, we can compute the total area and then divide by the area of one tile (0.6 m × 0.6 m). What’s the width of the room?

**Deepseek R1:** The user’s living room is 6 meters long, and the tiles have a side length of 60 centimeters (0.6 meters). To determine how many tiles are needed in total:

1. **Convert the length of the living room to centimeters**:

6 meters = 600 centimeters.

2. **Calculate the number of tiles along the length**:

Each tile is 60 centimeters long, so the number of tiles needed along the length is:

$$\frac{600 \text{ cm}}{60 \text{ cm/tile}} = 10 \text{ tiles}$$

3. **Consider the total number of tiles**:

Since the width of the living room is not provided, the problem is ambiguous. However, if we assume the question is asking how many tiles fit along the 6-meter length, the answer is 10 tiles.

Given the ambiguity but following the straightforward calculation for the length provided, the final answer is:

10

**Gemini 2.5 pro:** To figure out the total number of tiles, we need both the length and the width of your living room. You’ve given the length as 6 meters.

First, let’s make sure our units are the same:

Room length: 6 meters = 600 centimeters

Tile side length: 60 centimeters

Tiles along the length: Number of tiles = Room length / Tile side length  
Number of tiles = 600 cm / 60 cm = 10 tiles

So, you would need 10 tiles to go along the 6-meter length of your room.

To find the total number of tiles, you’ll also need the width of your room.

Let’s say, for example, your room is also 6 meters wide (making it a square room):

Room width: 6 meters = 600 centimeters

Tiles along the width: 600 cm / 60 cm = 10 tiles

Total tiles for a 6m x 6m room: 10 tiles (length) \* 10 tiles (width) = 100 tiles

If your room has a different width, please provide that dimension, and I can calculate the exact total for you. For instance, if your room was 3 meters (300 cm) wide:

Tiles along the width: 300 cm / 60 cm = 5 tiles

Total tiles: 10 tiles (length) \* 5 tiles (width) = 50 tiles

In summary: You need 10 tiles to span the 6-meter length. Please provide the width of your living room to calculate the total number of tiles needed.



756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

**Qwen3:** To determine how many tiles are needed to cover the floor of a living room that is 6 meters long, we need to calculate the total area of the room and divide it by the area of a single tile.

### Step 1: Convert Units for Consistency

Each tile has a side length of 60 cm, which is equivalent to:

$$60 \text{ cm} = 0.6 \text{ m}$$

So, the area of one tile is:

$$0.6 \text{ m} \times 0.6 \text{ m} = 0.36 \text{ m}^2$$

### Step 2: Determine the Area of the Room

The problem only gives the length of the room as 6 meters. To compute the area, we also need the width. However, since the width is not provided, and the problem asks for the total number of tiles, it is reasonable to assume that the room is square (i.e., length = width = 6 meters), which is a common assumption in such problems when only one dimension is given.

So, the area of the room is:

$$6 \text{ m} \times 6 \text{ m} = 36 \text{ m}^2$$

### Step 3: Calculate the Number of Tiles

Now, divide the total area of the room by the area of one tile:

$$\frac{36 \text{ m}^2}{0.36 \text{ m}^2/\text{tile}} = 100 \text{ tiles}$$

### Final Answer

100

**Claude 3.7:** To solve this problem, I need to find how many 60 cm tiles are needed to cover a length of 6 meters.

First, I'll convert the measurements to the same unit:

- Living room length = 6 meters
- Tile side length = 60 centimeters = 0.6 meters

To find the number of tiles needed, I'll divide the room's length by the tile's side length:

$$\begin{aligned} \text{Number of tiles} &= \text{Length of room} \div \text{Side length of tile} \\ \text{Number of tiles} &= 6 \text{ m} \div 0.6 \text{ m} = 10 \text{ tiles} \end{aligned}$$

Therefore, you need 10 tiles to cover the 6-meter length of your living room.

## B CRITIC-MATH CONSTRUCTION AND CASE STUDIES

### B.1 PROMPTS UTILIZED IN CONSTRUCTING CRITIC-MATH

CRITIC-math is constructed by rewriting the well-defined problems from open-source datasets. This construction involves three key stages: I) Parsing each problem into its constituent elements (i.e., goal, premises, and Background); II) Disturbing elements and constructing incomplete problems by reassembling these disturbed elements; and III) Verifying the resulting incompleteness.

This appendix details prompts used for each stage of our process (DeepSeek R1 was used). Specifically: Prompt in Figure 3 is used for parsing problems into goals, premises and background;. Stage II – Constructing Incomplete Problems – is mainly completed automatically and Prompts in Figure 4 and B.1 is used for reassembly and generating reason of incompleteness. Finally, Prompt in Figure 6 is used for verify the incompleteness.

Given a TEXT, its "Goal", "Necessary Information", and "Background" are as follows:

- "Goal" refers to the task that the TEXT requires to solve. "Goal" should be short and concise.
- "Necessary Information" and "Background" refer to everything apart from the Goal, including data, facts, examples, etc. Among them:
  - "Necessary Information" refers to the contents that are strictly needed to accomplish the Goal.
  - "Background" refers to the other contents that can be ignored, such as examples and descriptions.

Based on the above definitions, please SPLIT the following TEXT into "Goal", "Necessary Information", and "Background".

\*\*\*

TEXT:

problem

\*\*\*

If Background is None, leave Background empty.

Split long Necessary Information to short items.

For "Necessary Information", provide a numbered list.

**\*\*Requirements\*\***

1. **\*\*You should EXACTLY COPY contents from the TEXT to Goal, Necessary Information, and Background!!!\*\***
2. **\*\*DO NOT ADD, REWRITE, REPHRASE, ANY CONTENT WHEN COPYING TEXT to YOUR OUTPUTS!!!\*\***
3. **\*\*KEEP ALL SYMBOLS, such as EMPTY LINES, SPACES, MATH SYMBOLS, and FORMATTING SYMBOLS WHEN COPYING TEXT to YOUR OUTPUTS!!!\*\***
4. **\*\*Goal, Necessary Information, and Background SHOULD NOT OVERLAP!!!\*\***
5. **\*\*Goal, Necessary Information, and Background SHOULD COVER ALL contents of the TEXT!!!\*\***

Figure 3: The Prompt to Parse Problem into Goal and Premises.

864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

```

# Goal
{goal}

# Information
{information}

# Background
{background}

# Instruction

DO NOT consider rationality, assemble the Goal, Information, and Background
into a coherent text to get an Unclear Question.
Because the Unclear Question contains Missing Information compared with the Original
Question shown below, the Unclear Question should not be able to be answered directly.

***
Original Question:
{original_question}
***

Provide the reasons that why the Unclear Question is Unclear (i.e, can not be an-
swered directly).
Generate your output STRICTLY in the following format.
Unclear Question:
Reason of Unclearness:

# Requirements

1. **The Unclear Question SHOULD NOT CONTAIN the title of Goal, Informa-
tion, and Background.**
2. **The Unclear Question SHOULD CONTAIN ALL contents in the Goal, Information,
and Background (if the Information and Background are not empty).**
3. **DO NOT ADD ANY content that is not included in the Goal, Information and
Background from the Original Question to the Unclear Question.**

```

Figure 4: The Prompt to Construct Missing Premises Problems and Reason of Incompleteness

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

```
# Information
{information}

# Background
{background}

# Instruction
Assemble the Information and Background into a coherent text.

Generate your output STRICTLY in the following format.
Assembled Text:

# Requirements

1. **The Assembled Text SHOULD NOT CONTAIN the title of Information and Background.**
1. **The Assembled Text SHOULD CONTAIN ALL contents in the Information and Background (if the Background is not None).**
2. **DO NOT ADD ANY other content that is not included in the Information and Background to the Assembled Text.**
```

Figure 5: The Prompt to Construct Missing Goal Problems and Reason of Incompleteness

```
# Task
{query}

# Reason of Unclearness
{reason_of_unclearness}

# Instruction
According to the reason of unclearness, the task is classified to be unclear.
Please act as an impartial judge to evaluate whether the task is truly unclear and whether the reason of unclearness is correct.
If you agree that task is truly unclear, provide at least two different understandings of the unclear task corresponding to the unclearness as the verification of your judgement.

Output your final verdict by strictly following this format:
Analysis: [Your analysis about the task and the reason of unclearness]
Judge: "[[Yes]]" if you agree with the reason and that the task is truly unclear, "[[No]]" if you do not agree with the reason and that the task is clear.
Verification: [At least two different understandings of the unclear task corresponding to the unclearness]
```

Figure 6: The Prompt to verify the incompleteness

## B.2 HUMAN EVALUATION & QUALITY CONTROL

To verify the incomplete problems, we recruit three PhD students as human evaluators, two of them major in computer science and the other majors in mathematics. Since the incomplete problems have already been verified by Deepseek R1, we sample 10% of the problems and ask human evaluators to provide confirmation. Specifically, we ask human evaluators to provide their decision that whether the problem is incomplete given the incomplete problem, the counterpart well-defined problem, the reason of incompleteness, and R1’s verification. We type the above information into Microsoft Word and present to human evaluators for them to verify. Totally, 128 incomplete problems are evaluated. All evaluators are paid at least the minimum wage in their country.

The evaluation results show that human agree with the R1’s verification in 95% cases (i.e., human disagrees in only 5 cases). We present two cases that we provide human evaluators to make decisions. We highlight the missing elements in red.

**Potential risks of Human Evaluation:** Our human evaluation involves verification specific characteristic (i.e. the incompleteness) of mathematical problems. After a thorough examination, we do not notice any potential risk of human subjects. We have resorted to the Institutional Review Board and got the approval regarding our human evaluation. If we notice or receive any risk regarding the human evaluation, we will update this section and provide comprehensive discussion.

### Incomplete Problem (Missing Premises)

A robot moving forward at a constant speed takes 2.5 hours to travel 1 kilometer. How many meters long is the hallway?

### Well-defined Problem

a robot moving forward at a constant speed takes 2.5 hours to travel 1 kilometer. **moving forward at this same constant speed, it takes the robot 90 seconds to travel the length of a particular hallway.** how many meters long is the hallway?

### Reason of Incompleteness

The question does not provide the time it takes for the robot to travel the hallway. While the robot’s speed can be calculated from the given information (2.5 hours for 1 kilometer), the hallway’s length cannot be determined without knowing how long the robot took to traverse the hallway. The missing information (e.g., 90 seconds in the Original Question) is critical for calculating the distance.

### R1 Verification

Analysis: The task states the robot’s speed (1 km in 2.5 hours) but does not specify the time it took to traverse the hallway. To calculate the hallway’s length, both speed and time are required. The absence of the traversal time makes the problem unsolvable as posed. The reason for unclearness correctly identifies the missing critical information (time to traverse the hallway), rendering the task ambiguous.

Judge: [[Yes]]

Verification:

- Misinterpretation 1**: Assume the 2.5 hours mentioned is the time taken to traverse the hallway. This would incorrectly equate the hallway’s length to 1 km (1000 meters), conflating the speed-calibration distance with the hallway’s actual length.
- Misinterpretation 2**: Assume the problem implicitly asks for the robot’s speed in meters per second (e.g., converting 1 km/2.5 hours to  $\approx 0.111$  m/s) and answers with the speed instead of the hallway length, misaligning with the question’s intent due to confusion over missing variables.

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

### Incomplete Problem (Missing Goals)

In a school of 360 students, 15 are enrolled in calculus, physics, and chemistry simultaneously, while 15 students do not take any of these three subjects. A total of 180 students take calculus. The number of students taking chemistry is twice the number taking physics. There are 75 students enrolled in both calculus and chemistry, and another 75 in both physics and chemistry. However, only 30 students take both physics and calculus.

### Well-defined Problem

there are 360 people in my school. 15 take calculus, physics, and chemistry, and 15 don't take any of them. 180 take calculus. twice as many students take chemistry as take physics. 75 take both calculus and chemistry, and 75 take both physics and chemistry. only 30 take both physics and calculus. **how many students take physics?**

### Reason of Incompleteness

The problem lacks a specified goal to answer.

### R1 Verification

Analysis: The task provides data about student enrollments in three subjects but does not specify a question to answer. The reason for unclearness—lack of a goal—is valid. Without a defined objective (e.g., finding the number of students in a specific subject or verifying consistency), the problem is ambiguous.

Judge: [[Yes]]

Verification:

- Understanding 1**: The task aims to find the total number of students enrolled in physics. Using the given overlaps and relationships (chemistry =  $2 \times$  physics), one might solve for physics using set theory formulas.
- Understanding 2**: The task requires checking if the provided numbers are mathematically consistent (e.g., ensuring overlaps do not exceed subject totals). For instance, verifying if 75 students in both physics and chemistry (with 15 in all three) do not exceed the total physics or chemistry enrollments.



## C IMPLEMENTATION OF LLM-AS-A-JUDGE FOR EVALUATION

In this section, we detail the prompts we used to conduct LLM-as-a-Judge in Evaluation.

### **Prompts to Evaluate that Whether a response Raises Clarification Questions when using "Implicit Prompt"**

---

# Question

{problem}

# Response

{response}

# Instruction

Please act as an impartial judge to evaluate whether the response has raised questions to ask additional information for answering the Question.

Output your final verdict by strictly following this format:

Analysis: [Your analysis about the response]

Judge: "[[Yes]]" if the response has raised questions to ask additional information, and "[[No]]" if the response directly provide an answer without raising clarification.

### **Prompts to Evaluate CNR, i.e. whether the thoughts have noticed the need to raise clarification questions.**

---

# Incomplete Question

{incomplete\_task}

# Reason of Incompleteness

{reason\_of\_incompleteness}

# Output

{output}

# Instruction

Please act as an impartial judge to evaluate whether the output has noticed the need to ask clarification questions to address the incompleteness.

Your assessment should be grounded on the Incomplete Question and the Reason of Incompleteness.

Output your final verdict by strictly following this format:

Analysis: [Your analysis about the Output]

Judge: "[[Yes]]" if the output has noticed the need to ask clarification questions, otherwise "[[No]]"

1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187

**Prompts to Evaluate ROR on missing premises problems, i.e. whether the reflection thoughts have focused on addressing the incompleteness.**

---

# Incomplete Question

{incomplete\_task}

# Reason of Incompleteness

{reason\_of\_incompleteness}

# Output

{output}

# Instruction

Please act as an impartial judge to evaluate whether the output has tried to address the incompleteness.  
For example, the output tried different possibilities about the incompleteness or guessing how the incompleteness arises. Your assessment should be grounded on the Incomplete Question and the Reason of Incompleteness.

Output your final verdict by strictly following this format:  
Analysis: [Your analysis about the Output]  
Judge: "[[Yes]]" if the output has tried to address the incompleteness, otherwise "[[No]]"

**Prompts to Evaluate ROR on missing goal problems, i.e. whether the thoughts imagine a goal by themselves.**

---

# Incomplete Question

{unclear\_task}

# Reason of Incompleteness

{reason\_of\_unclearness}

# Output

{output}

# Instruction

According to the reason of incompleteness, the question lacks a specific goal.  
Please act as an impartial judge to evaluate whether the output has tried to address the incompleteness by imagining a goal itselfes.

Output your final verdict by strictly following this format: Analysis: [Your analysis about the Output]  
Judge: "[[Yes]]" if the output has tried to address the incompleteness, otherwise "[[No]]"

## D ADDITIONAL EXPERIMENTAL ANALYSIS

### D.1 ADDITIONAL ANALYSIS FOR RQ1

#### D.1.1 ADDITIONAL RESULTS ON DATASET SYNTHESIZED BY GEMINI-2.5 PRO

Table 7: Coarse-grained evaluation of LRMs on datasets synthesized by Gemini-2.5 Pro.

Models	Missing Premises			Missing Goal		
	CR	TLC	TLNC	CR	TLC	TLNC
CRITIC-math (from Math 500 Data Source)						
Deepseek R1	57.00%	0.61	4.03	52.00%	0.88	2.56
Qwen3 Plus	53.00%	0.33	2.38	55.98%	0.27	0.85
o3 mini	55.00%	1.50	2.12	34.00%	1.42	1.18
Grok 3 mini	58.00%	0.50	1.43	47.00%	0.47	1.16
Claude 3.7	48.00%	0.76	2.28	40.00%	0.66	2.06
CRITIC-math (from Omni-Math Data Source)						
Deepseek R1	56.00%	0.34	2.38	55.00%	0.40	1.34
Qwen3 Plus	45.00%	0.63	1.54	57.09%	0.41	1.12
o3 mini	47.00%	0.52	1.19	45.00%	0.45	0.88
Grok 3 mini	52.00%	0.33	1.00	52.00%	0.37	0.96
Claude 3.7	44.00%	0.92	2.14	35.00%	0.33	1.73

To avoid model bias in data synthesis, we additionally use Google’s Gemini 2.5 Pro to generate supplementary data and perform corresponding evaluation. As detailed in the table, our core findings consistently hold across data synthesized by different models. These include:

- LRMs perform poor to ask for information on incomplete problems (low CR-Clarification Ratio).
- LRMs tend to overthink on incomplete problems (large TLNC-Thinking Length Not Clarifying).
- Raising clarification questions on incomplete problems allow LRMs to reduce thinking effort (low TLC, Thinking Length when Clarifying).

#### D.1.2 ADDITIONAL RESULTS USING GPT-4.1 AS THE JUDGE LLM

Table 8: Fine-grained analysis of LRMs with GPT-4.1 as the Judge LLM.

Model	Missing Premises		Missing Goal	
	ROR	CNR	ROR	CNR
Deepseek R1	75.62%	40.63%	84.45%	16.39%
Qwen3 Plus	77.38%	45.71%	80.19%	16.51%

Table 9: Fine-grained analysis of LRMs with Claude-3.7 Thinking as the Judge LLM.

Model	Missing Premises		Missing Goal	
	ROR	CNR	ROR	CNR
Deepseek R1	79.91%	58.69%	98.74%	35.29%
Qwen3 Plus	79.29%	61.90%	97.64%	4.81%

We provide additional results where we use other two LLMs to judge the thoughts of Deepseek R1 and Qwen3 Plus. As detailed in the table, our main findings remain consistent with those in our main page, even when evaluated by different LLMs. Specifically, the high ratio of overthinking and high CNR (Clarification Noticing Ratio) in scenarios with missing information are consistently observed across these different evaluators, although the specific evaluation scores may vary between models.

### D.1.3 THE RELATIONSHIP BETWEEN CRS AND PROBLEM DIFFICULTIES

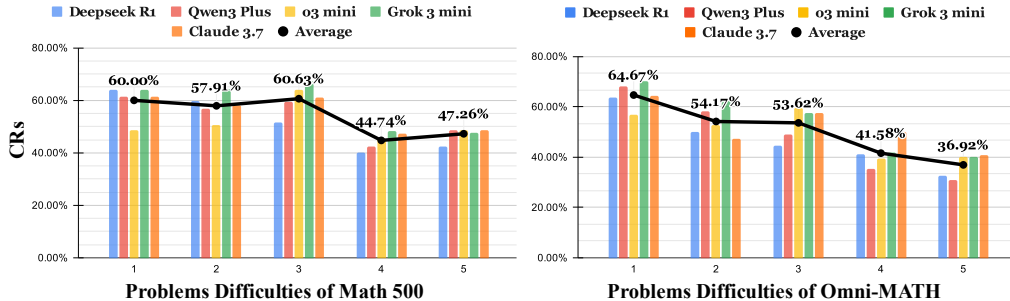


Figure 7: CRs of different LRMs on missing premises problems with different difficulties.

Figure 7 show the clarification ratios (CRs) of different LRMs on missing premises problems. We can observe a negative correlation between the problem difficulty and the CRs, which indicate that the ability to ask for information is unrobust to the problem difficulty.

### D.1.4 QUESTIONS ACCURACIES IN ADDRESSING THE INCOMPLETENESS

In this section, we provide the evaluation results regarding the accuracy of asking for information, which is measured by the ratio of questions asked by LRMs that can appropriately address the incompleteness. We use LLM-as-a-Judge (Deepseek R1) as the evaluation, where the prompt are detailed in Figure 8. From the results, we can observe that all LRMs, when they successfully raise questions, can achieve rather high accuracy in asking for information.

Table 10: The accuracy of asking for information, which is measured by the ratio of questions asked by LRMs that can appropriately address the incompleteness.

Model	Missing Premises	Missing Goal
CRITIC-math (from Math 500 Data Source)		
Deepseek R1	91.25%	98.25%
Qwen3 Plus	95.41%	99.26%
o3 mini	89.16%	90.48%
Grok 3 mini	92.23%	100.00%
Claude 3.7	87.26%	99.12%
CRITIC-math (from Omni-MATH Data Source)		
Deepseek R1	87.18%	95.59%
Qwen3 Plus	89.86%	95.92%
o3 mini	87.00%	91.53%
Grok 3 mini	85.51%	98.04%
Claude 3.7	85.53%	95.71%

### D.1.5 SAMPLING MULTIPLE RESPONSES

. In the evaluation in Section 3, we generate one response with the temperature set to 0 from close-source LRMs. To analyze that how the temperature affects the results and will the results significantly vary in different samples on the same problem, we conduct additional evaluation. Considering our budget, we conduct this evaluation only using the "explicit prompt" on randomly sampled 50 problems from each category of missing premises, missing goal, and well-defined in each data source.

1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318

```

# Incomplete Task
{incomplete_task}

# Reason of Incompleteness
{reason_of_incompleteness}

# Clarification Question
{clarification_question}

# Instruction

Please act as an impartial judge to evaluate whether the clarification question can
appropriately address the incompleteness of the incomplete question, given the content of
the incomplete question and the reason of incompleteness.

Output your final verdict by strictly following this format:
Analysis: [Your analysis about the clarification question]
Judge: "[[Yes]]" if the task clarification question is appropriate and sufficient, otherwise
"[[No]]"

```

Figure 8: Prompts to evaluate the accuracy of asking for information.

1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332

For each problem, we generate one response with temperature set to 0 (denotes as S) and 5 responses with temperature set to 0.6 (denoted as M).

Table 11 display the results. From the results, we can observe that sampling multiple responses will slightly influence the CRs, TLCs, and TLNCs. However, the influence varies in different LRMs and different types of incompleteness. Generally speaking, the results of S mostly fall into the 95% confidential interval of M. Also, the main conclusion remains the same, where the CRs are around 50%, TLNCs are greater than 1 on missing premises problems, barely changed TLNCs on missing goal problems, and lower TLCs on both cases. These additional results demonstrate that our analysis and results in the main text are robust to the choice of temperature and different sampled responses.

#### TLNCs on Problems of Different Difficulties.

1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349

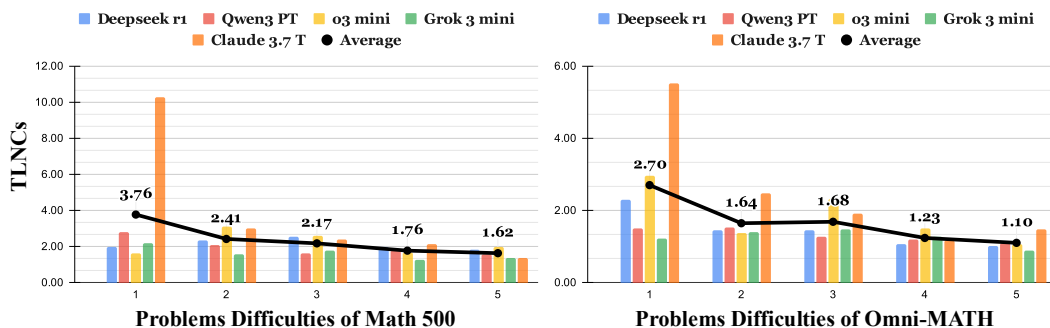


Figure 9: TLNCs of different LRMs on incomplete problems with different difficulties.

The results are plotted in Figure 9. From the results, we can observe a negative correlation between the problem difficulty and the TLNCs, where the overthink are more significant on easier problems. For example, on the difficulty level of 1 in Math 500, the thoughts lengths are in average 3.76 times longer thoughts compared with the lengths on well-defined problems. In addition, the thoughts lengths are

Table 11: The results of sampling multiple responses. S denotes generating one response with temperature set to 0, and M denotes sampling 5 responses with temperature set to 0.6. For M, 95% confidential intervals are reported. We omit the results of well-defined problems for better readability.

Model	Type	Missing Premises			Missing Goal		
		CR	TLC	TLNC	CR	TLC	TLNC
CRITIC-math (Subset from Math 500 Data Source)							
Deepseek R1	S	55.56%	0.53	2.25	46.00%	0.41	1.03
	M	37.60%±3.24%	0.62±0.32	2.38±0.22	48.80%±4.84%	0.49±0.13	1.29±0.22
Qwen3 PT	S	44.00%	0.50	2.45	64.00%	0.73	1.23
	M	44.00%±3.51%	0.55±0.01	2.60±0.13	60.00%±2.51%	0.47±0.01	1.31±0.26
o3 mini	S	46.00%	1.62	3.20	36.00%	1.54	1.33
	M	43.20%±1.36%	1.78±0.37	2.98±0.65	38.80%±0.02	1.88±0.33	1.32±0.58
Grok 3 mini	S	46.00%	0.57	1.61	54.00%	0.64	0.86
	M	48.00%±3.04%	0.70±0.19	1.72±0.20	49.20%±2.83%	0.47±0.04	1.15±0.09
Claude 3.7	S	28.00%	0.30	2.56	46.00%	0.16	1.59
	M	38.40%±17.86%	0.29±0.24	1.76±1.12	48.00%±8.42%	0.28±0.27	1.51±0.74
CRITIC-math (Subset from Omni-Math Data Source)							
Deepseek R1	S	46%	0.28	1.25	59.00%	0.24	1.16
	M	50.00%±3.51%	0.39±0.03	1.28±0.08	50.80%±0.42	0.26±0.12	1.33±0.09
Qwen3 PT	S	48.00%	0.23	1.24	58.00%	0.27	1.26
	M	51.20%±2.22%	0.33±0.03	1.21±0.08	58.00%±4.65%	0.33±0.07	1.21±0.14
o3 mini	S	56.00%	0.42	1.71	40.00%	0.72	1.08
	M	55.60%±6.43%	0.57±0.39	1.48±0.14	39.20%±0.03	0.66±0.13	0.87±0.13
Grok 3 mini	S	56.00%	0.44	1.07	52.00%	0.45	1.10
	M	61.60%±5.38%	0.40±0.05	1.06±0.09	53.60%±3.24%	0.41±0.04	1.07±0.10
Claude 3.7	S	42.00%	0.45	1.38	38.00%	0.36	1.16
	M	42.00%±1.76%	0.45±0.14	1.45±0.10	36.40%±4.44%	0.21±0.11	1.27±0.14

increased more significant on Math 500 compared with Omni-MATH, given that Omni-MATH is a dataset containing more challenging problems.

#### D.1.6 CASES STUDIES ABOUT OVERTHINKING WHEN MISSING PREMISES

We provide 3 case studies of Deepseek R1 with "explicit prompt" related to the overthinking when missing premises. The case studies are provided in Figure 10, 11, and 12. From the cases, we can observe that LRMs try to address the incompleteness by themselves (marked as **red**) rather than asking for information. Such self-reflections on incompleteness suggests that the current mode of deep-thinking may overemphasize the internal thinking while ignoring the external interactions. These phenomena also support our analysis in Section D.2 about the dilemma between learning the current mode of deep-thinking and asking for information.

#### D.1.7 CASES STUDIES ABOUT HALLUCINATIONS WHEN MISSING GOALS

We provide 6 case studies of Deepseek R1 (3 with "implicit prompt" and 3 with "explicit prompt") about the hallucinations when missing goals. The results are shown in Figure 13, 14, 15, 16, 17, and 18. We highlight the goal that not specified in the input but imagined by Deepseek R1 in **red**. In those cases, we can observe that Deepseek R1 immediately generates a goal in their first paragraph of thinking, and dedicates to solve the imagined goal and finally provide a corresponding answer.



1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457

# Problem to Solve

The greek army contained two types of soldiers: the upper class and the lower class soldiers. If there were a total of 5 upper class soldiers and 10 lower class soldiers in a certain part of athens, and the battle of thermopylae demands a force of 4 upper class soldiers, how many different battalions can be sent?

# Instruction

If you need ask information, please raise clarification question and start your response STRICTLY with: "Clarification Question" followed by your questions. Otherwise, please reason step by step, and put your final answer within  $\{\}$ .

---

...

**But wait, maybe there's more to it. The problem mentions "battalions," which usually consist of multiple soldiers.** However, since the requirement is specifically for 4 upper class soldiers, perhaps lower class soldiers are not part of this battalion. Alternatively, maybe the problem expects that the battalion includes both upper and lower class soldiers, but with exactly 4 upper class ones. But the problem doesn't specify a total number of soldiers needed for the battalion. So that's a bit confusing.

...

**Alternatively, maybe the problem is just asking for the number of ways to choose the 4 upper class soldiers from the available 5,** regardless of the lower class soldiers. Because if the battalion is required to have 4 upper class soldiers, and the number of lower class soldiers isn't specified,

...

**But I need to make sure. Let me think again. The problem states "a force of 4 upper class soldiers."** If the battalion is just those 4, then yes, 5 choose 4 is 5.

...

Figure 10: Case Studies of Overthinking when Missing Premises.

1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

# Problem to Solve

find the remainder when  $(n + 2)(n + 4)(n + 6)$  is divided by 7.

# Instruction

If you need ask information, please raise clarification question and start your response STRICTLY with: "Clarification Question" followed by your questions. Otherwise, please reason step by step, and put your final answer within  $\{\}$ .

---

...

**Case 1:  $n = 0 \pmod 7$**

Then:

$n + 2 = 2 \pmod 7$

$n + 4 = 4 \pmod 7$

$n + 6 = 6 \pmod 7$

Product:  $2 * 4 * 6 = 48$ . 48 divided by 7 is  $6 * 7 = 42$ , remainder 6. So  $48 \pmod 7 = 6$ .

...

**Case 2:  $n = 1 \pmod 7$**

Then:

$n + 2 = 3 \pmod 7$

$n + 4 = 5 \pmod 7$

$n + 6 = 7 = 0 \pmod 7$

Product:  $3 * 5 * 0 = 0$ . So remainder 0.

...

But wait, the problem is asking for "the remainder" in general. But since the remainder depends on  $n$ , unless there's more information about  $n$ , we can't give a unique answer. **Wait, but maybe the problem is expecting an answer that works for any  $n$ , or perhaps there's a pattern that can be generalized.**

...

Figure 11: Case Studies of Overthinking when Missing Premises.

1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565

# Problem to Solve

a pentagon is rotated counterclockwise. what is the number of degrees it must be rotated until it coincides with its original position?

# Instruction

If you need ask information, please raise clarification question and start your response STRICTLY with: "Clarification Question" followed by your questions. Otherwise, please reason step by step, and put your final answer within  $\{\}$ .

---

...  
The exact wording is: "what is the number of degrees it must be rotated until it coincides with its original position?" So, does that mean the total rotation needed to get back to the original position? For example, if you rotate 72 degrees each time, after five such rotations, you would have  $5 \times 72 = 360$  degrees, which brings it back. **But is the question asking for the smallest such angle?** Because 360 degrees is a full rotation, which would obviously bring it back. But maybe the question is trickier than that  
But I need to make sure. Let me think. For example, a square has rotational symmetry every 90 degrees. So, rotating it 90 degrees four times brings it back to the original position. So, each 90 degrees is a step. So, the minimal angle is 90 degrees. But the total rotation required to get back to original is 360 degrees, **but the question is probably asking for the minimal angle that when you rotate by that angle, the figure coincides with itself.**  
...

Figure 12: Case Studies of Overthinking when Missing Premises.

A proper divisor of an integer  $n$  is a positive divisor of  $n$  that is less than  $n$ . For example, the proper divisors of 12 are 1, 2, 3, 4, and 6.

---

Okay, **so I need to figure out the proper divisors of an integer.** Let's start by recalling what a proper divisor is. The problem statement says a proper divisor of an integer  $n$  is a positive divisor that is less than  $n$ . For example, for 12, the proper divisors are 1, 2, 3, 4, and 6. That makes sense because those numbers divide 12 without leaving a remainder and are all less than 12.  
...  
**The proper divisors of 12 are**  $\{1, 2, 3, 4, 6\}$ .

Figure 13: Case Studies of Hallucinations when Missing Goals.

$\overline{bc}$  is parallel to the segment through  $a$ , and  $ab = bc$ . The diagram shows two horizontal lines connected by segments from  $a$  to  $b$  and  $a$  to  $c$ . The angle at  $a$  is labeled  $124^\circ$ , and an angle labeled  $x^\circ$  is marked on the upper horizontal line near  $(4.5, 3)$ .

---

Okay, let's see. **I need to figure out the value of angle  $x$  in this diagram.** Hmm. The problem says that segment  $bc$  is parallel to the segment through  $a$ , and that  $ab$  equals  $bc$ . There are two horizontal lines connected by segments from  $a$  to  $b$  and  $a$  to  $c$ . The angle at  $a$  is 124 degrees, and angle  $x$  is on the upper horizontal line near the point  $(4.5, 3)$ .  
...  
**\*\*Final Answer\*\***  $\{28\}$

Figure 14: Case Studies of Hallucinations when Missing Goals.

1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

The problem must be solved without the use of a calculator.

---

Okay, so I need to solve this problem without using a calculator. **Let me see... The problem is about a triangle with vertices at points A, B, and C.** The coordinates given are A(1, 2), B(4, 6), and C(6, 3). The question is asking for the area of this triangle. Hmm, right. I remember there are a few ways to calculate the area of a triangle when you have the coordinates of the vertices. One method is using the shoelace formula, another might be using vectors or determinants. Let me recall the shoelace formula because that seems straightforward for coordinates.

...

**The area of the triangle is  $\boxed{\frac{17}{2}}$ .**

Figure 15: Case Studies of Hallucinations when Missing Goals.

# Problem to Solve

A regular hexagon can be divided into six equilateral triangles. The perimeter of one of the triangles is 21 inches.

# Instruction

If you need ask information, please raise clarification question and start your response STRICTLY with: "Clarification Question" followed by your questions. Otherwise, please reason step by step, and put your final answer within  $\boxed{\quad}$ .

---

Okay, let's see. The problem says a regular hexagon can be divided into six equilateral triangles. The perimeter of one of those triangles is 21 inches. I need to find something, but wait, the original problem statement isn't here. Wait, looking back. **Oh, the user said "If you need to ask information, please raise clarification question..." So maybe the problem is to find the perimeter of the hexagon? Because that would make sense.**

...

So yes, the side of each triangle is equal to the side of the hexagon. **Therefore, perimeter of hexagon is  $6 * 7 = 42$ . So I think that's correct**

Figure 16: Case Studies of Hallucinations when Missing Goals.

1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673

# Problem to Solve

Let  $z = 2 + \sqrt{2} - (3 + 3\sqrt{2})i$  and  $c = 2 - 3i$ . The point  $w$  is obtained by rotating  $z$  around  $c$  by  $\frac{\pi}{4}$  radians counter-clockwise. An Asymptote diagram illustrates this configuration:  $z$  is plotted at the coordinate  $(2 + \sqrt{2}, -3 - 3\sqrt{2})$ ,  $c$  at  $(2, -3)$ , and  $w$  is the result of rotating  $z$  by  $45^\circ$  around  $c$ . The figure includes lines connecting  $z$ ,  $c$ , and  $w$ , with labeled points and an angle label  $\frac{\pi}{4}$  near  $c$ .

# Instruction

If you need ask information, please raise clarification question and start your response STRICTLY with: "Clarification Question" followed by your questions. Otherwise, please reason step by step, and put your final answer within `\boxed{ }`.

---

Okay, **so I need to find the point w**, which is the result of rotating  $z$  around  $c$  by  $\pi/4$  radians counter-clockwise. Let me start by recalling how rotation works in the complex plane.

...

Therefore, **the complex number w is 6 - 5i**. So I think that's the correct answer. I don't see any mistakes in the steps. So the final answer is `\boxed{6 - 5i}`.

Figure 17: Case Studies of Hallucinations when Missing Goals.

# Problem to Solve

Denali and Nate work for a dog walking business and are paid for each dog they walk. Denali is responsible for 16 dogs, and Nate is responsible for 12 dogs. Under the company's new policy, they will be assigned or unassigned new dogs in groups of  $x$  dogs. The ratio of Denali's pay to Nate's pay would be the same if Denali started walking  $4x$  more dogs and Nate stayed at 12 dogs or if  $x$  of Nate's dogs were reassigned to Denali.

# Instruction

If you need ask information, please raise clarification question and start your response STRICTLY with: "Clarification Question" followed by your questions. Otherwise, please reason step by step, and put your final answer within `\boxed{ }`.

---

Okay, let's see. So the problem is about Denali and Nate who work for a dog walking business. They get paid per dog they walk. Currently, Denali has 16 dogs and Nate has 12. The company is changing their policy so that they'll be assigned or unassigned dogs in groups of  $x$  dogs. The key point is that the ratio of their pay remains the same under two different scenarios: either Denali gets  $4x$  more dogs and Nate stays at 12, or  $x$  of Nate's dogs are reassigned to Denali. **I need to find x**. ... Yes, both equal 3. So the answer is 5. So  $x=5$ . **Therefore, boxed answer is `\boxed{5}`**

Figure 18: Case Studies of Hallucinations when Missing Goals.

## D.2 ADDITIONAL ANALYSIS FOR RQ2

We conduct ablation studies to analyze the effect of different types of problems in SFT models. The results are provided in Table 12, where the subscript W indicates trained solely using well-defined problems, and the subscript I indicates trained solely using incomplete problems. From the results, we have the following observations:

**Learning to ask for information can benefit the ability of solving problems.** Comparing CRITIC-Qwen<sub>W</sub> with CRITIC-Qwen and CRITIC-Qwen-thinking<sub>W</sub> with CRITIC-Qwen-thinking, we can observe that SRs on well-defined problems even decreased when solely learning to solve well-defined problems (W). Therefore, learning to ask for information on incomplete problems is not contradict with, even benefit, learning the ability to solve well-defined problems

**The Dilemma between Deep Thinking and Asking for Information.** When trained solely on the answers of incomplete problems, CRITIC-Qwen<sub>I</sub> achieves 100% CR on both types of problems, which makes sense since the SFT model will overfit to raising clarification questions. However, when trained with the thoughts (CRITIC-Qwen-thinking<sub>I</sub>), the CRs significantly decrease, and comparable SRs are achieved on well-defined problems where SFT models do not raise questions. This suggests that the current mode of thoughts overemphasize the contents regarding how to solve problems, even when thinking to raise clarification questions. Trained on such kind of deep thinking process, the model will more likely to just learn how to solve problems rather than how to interact with users and ask for information. In addition, we can observe that CRITIC-Qwen-thinking<sub>I</sub> achieve higher CRs compared with CRITIC-Qwen-thinking where their deep thinking is less strengthened by learning to solve well-defined problems, therefore being more willing to ask for information.

Table 12: Ablation studies of SFT models trained on different types of problems.

Model	Missing Premises	Missing Goal	Well-defined	
	CR	CR	CR	SR
CRITIC-math (from Math 500 Data Source)				
CRITIC-Qwen	78.42%	94.87%	4.12%	73.39%
CRITIC-Qwen <sub>W</sub>	0.53%	0.00%	0.00%	73.25%
CRITIC-Qwen <sub>I</sub>	100.00%	100.00%	100.00%	0.00%
CRITIC-Qwen-thinking	57.37%	62.82%	1.23%	80.83%
CRITIC-Qwen-thinking <sub>W</sub>	34.21%	28.63%	0.00%	80.25%
CRITIC-Qwen-thinking <sub>I</sub>	63.42%	73.93%	7.00%	79.20%
CRITIC-math (from Omni-MATH Data Source)				
CRITIC-Qwen	77.88%	97.64%	13.16%	41.13%
CRITIC-Qwen <sub>W</sub>	0.23%	0.00%	0.00%	39.85%
CRITIC-Qwen <sub>I</sub>	100.00%	100.00%	100.00%	0.00%
CRITIC-Qwen-thinking	56.66%	69.29%	7.52%	54.88%
CRITIC-Qwen-thinking <sub>W</sub>	24.60%	37.40%	3.38%	51.75%
CRITIC-Qwen-thinking <sub>I</sub>	57.34%	69.53%	12.41%	52.79%

In Figure 19 we provide a case study about the training samples. We can observe that, even the final answer raises clarification questions and LRMs quickly identify the missing information (marked in **blue**), the thinking process still tries hard to address the incompleteness (marked in **red**). Such thinking process strengthen the ability of solving problems (i.e., trying different ways to solve a problem) while weaken the ability of asking for information.

### D.2.1 ADDITIONAL RESULTS TRAINED USING LLAMA-3.1-8B-INSTRUCT

We have further fine-tune Llama 3.1 8B Instruct to analyze whether "deep-thinking dilemma" exhibits across different models. The results are shown in the above table. As detailed in the table, our core findings regarding "deep-thinking dilemma" consistently hold when fine-tuning Llama 3.1 8B Instruct, an instruction-tuned model as you requested. These include:



1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781

# Problem to Solve

a group of boys and girls have eaten at a restaurant where only pizzas cut into 12 slices are served. each boy ate 6 or 7 slices and each girl 2 or 3 slices. calculate the number of boys and girls in the group.

# Instruction

If you need ask information, please raise clarification question and start your response STRICTLY with: Clarification Question followed by your questions. Otherwise, please reason step by step, and put your final answer within `\boxed{ }`.

---

<think>

Okay, let's try to figure out this problem. So, there are some boys and girls who ate pizza at a restaurant. The pizzas are each cut into 12 slices. Each boy ate either 6 or 7 slices, and each girl ate 2 or 3 slices. We need to find out how many boys and girls are in the group.

Hmm, the problem says "calculate the number of boys and girls," but I don't see any specific numbers given here. **Maybe there's some missing information? Wait, maybe the total number of slices or the number of pizzas ordered is implied? Let me check the problem again.**

Wait, the problem states that "a group of boys and girls have eaten at a restaurant where only pizzas cut into 12 slices are served." So the total number of slices must be a multiple of 12 because they can only order whole pizzas. But the problem doesn't mention how many pizzas they ordered. That's probably the key here. So the total slices consumed by the group must be a multiple of 12.

**Let me restate the problem.** Let's denote the number of boys as B and girls as G. Each boy ate 6 or 7 slices, so total slices eaten by boys would be 6B or 7B, but actually, each individual boy can eat either 6 or 7. So the total for boys is 6B + some number depending on how many ate 7 instead of 6. Similarly, each girl ate 2 or 3, so total for girls is 2G + some number depending on how many ate 3. But this seems complicated because there are multiple variables and possible combinations.

**Wait, maybe there's more constraints.** The problem might be expecting a unique answer, which would mean that the total number of slices is a multiple of 12, and with the number of boys and girls such that the possible combinations of 6/7 and 2/3 sum up to that multiple. But without knowing the total slices or the number of pizzas, how can we find a unique answer?

...

So I need to ask for clarification. The problem must have had more data, but maybe it's missing in translation. Therefore, the correct approach here is to ask for the total number of pizzas or total slices consumed.

Figure 19: Case Studies of One Training Sample.

Table 13: Coarse level evaluation of SFT models. *CRITIC-Qwen T* denotes CRITIC-Qwen-thinking. *Qwen3-8B T* denotes the official LRM (also fine-tuned from Qwen3-8B-Base) with thinking enabled.

Model	Missing Premises			Missing Goal			Well-Defined				ACC
	CR	TLC	TLNC	CR	TLC	TLNC	CR	TLC	TLNC	SR	
CRITIC-Llama	88.42%	/	/	90.15%	/	/	27.12%	/	/	32.10%	88.81%
CRITIC-Llama T	66.32%	0.11	1.50	73.93%	0.06	1.02	7.82%	0.34	1 (4554)	43.62%	77.48%
Llama-3.1-8B-Instruct	51.30%	/	/	49.57%	/	/	41.98%	/	/	6.17%	64.47%

- 1782 • Fine-tuning on training data of CRITIC helps to improve the ability of asking for information of  
1783 Llama, suggested by the improved clarification accuracy (ACC: the ratio of clarifying incomplete  
1784 questions but not clarifying well-defined questions) of CRITIC-Llama compared with Llama 3.1  
1785 8B Instruct.
- 1786 • Incorporating deep-thinking, the problem-solving ability of fine-tuned model is improved, suggested  
1787 by the higher SR (solving ratio of well-defined problems) of CRITIC-Llama T(hinking) compared  
1788 with CRITIC-Llama.
- 1789 • Incorporating deep-thinking, the asking for information ability of fine-tuned model is reduced,  
1790 suggested by the lower CR (clarification ratio) and ACC (clarification accuracy) of CRITIC-Llama  
1791 T compared with CRITIC-Llama.

1792  
1793 The above results exactly demonstrate the deep-thinking dilemma, where incorporating current style  
1794 of deep-thinking improves the problem-solving ability while reduces the ability to ask for information.  
1795 Such results alert the community that the current type of thinking may overlook another ability of  
1796 genuine intelligence.

## 1797 1798 E LLMs USAGE STATEMENT

1799  
1800 LLMs/LRMs become powerful tools for efficiently processing data and human-like evaluation in  
1801 many recent works. Following these common practices, in this paper, we use LLMs to synthesize  
1802 large-scale incomplete problems and conduct LLM-as-a-Judge evaluation. However, all conceptual  
1803 work and method design in this work are conducted independently by the authors. LLMs are used to  
1804 polish the writing, help to synthesize data, and evaluate results, which do not contribute to the dataset  
1805 design or the shape of other core ideas of this paper. Any part steps that involve LLMs to synthesize  
1806 data and evaluate results are guided by carefully designed frameworks proposed by the authors.

1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835