Answering the Unanswerable Is to Err Knowingly: Analyzing and Mitigating Abstention Failures in Large Reasoning Models

Yi Liu[1], **Xiangyu Liu**[1], **Zequn Sun**[1], **Wei Hu**[1,2,*]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] National Institute of Healthcare Data Science, Nanjing University, China
{yiliu07, xyl}.nju@gmail.com, {sunzq, whu}@nju.edu.cn

Abstract

Large reasoning models (LRMs) have shown remarkable progress on complex reasoning tasks. However, some questions posed to LRMs are inherently unanswerable, such as math problems lacking sufficient conditions. We find that LRMs continually fail to provide appropriate abstentions when confronted with these unanswerable questions. In this paper, we systematically analyze, investigate, and resolve this issue for trustworthy AI. We first conduct a detailed analysis of the distinct response behaviors of LRMs when facing unanswerable questions. Then, we show that LRMs possess sufficient cognitive capabilities to recognize the flaws in these questions. However, they fail to exhibit appropriate abstention behavior, revealing a misalignment between their internal cognition and external response. Finally, to resolve this issue, we propose a lightweight, two-stage method that combines cognitive monitoring with inference-time intervention. Experimental results demonstrate that our method significantly improves the abstention rate while maintaining the reasoning performance.

1 Introduction

Large reasoning models (LRMs), such as GPT-o1 (Jaech et al. 2024) and DeepSeek-R1 (DeepSeek-AI et al. 2025), have demonstrated strong performance on complex reasoning tasks (Xu et al. 2025). By introducing the concept of "thought" and generating longer chains of thought (CoT), LRMs are able to explore diverse reasoning paths while spontaneously reflecting and correcting errors. This enables LRMs to tackle complex tasks with greater depth and flexibility, which is particularly valuable in high-stakes reasoning scenarios.

Most current research on LRMs aims to improve the reasoning process for solving complex problems (Xu et al. 2025), such as reducing redundant reasoning steps to alleviate overthinking (Yang et al. 2025b) and refining the reasoning process to further enhance reasoning performance (Chen et al. 2025; Liao et al. 2025). Limited work has considered that not all complex questions are answerable (Kirichenko et al. 2025; Song, Shi, and Zhao 2025). When confronted with noisy and diverse user queries, there will always be cases where LRMs are unable to produce reliable responses, such as the math problems missing sufficient conditions.

*Corresponding author

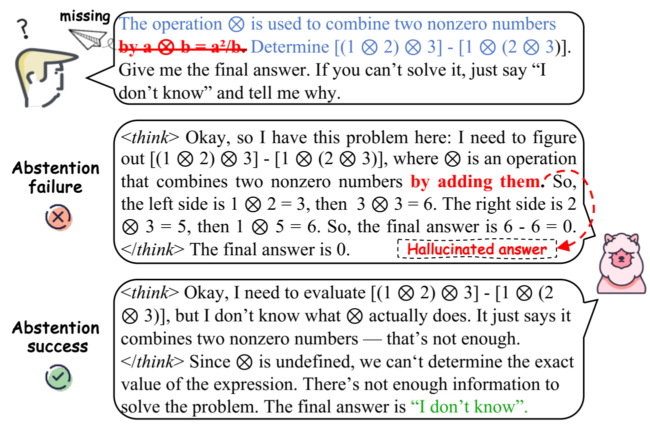

Figure 1: Examples of abstention failure and success for LRMs on unanswerable questions.

In Figure 1, when a question is unanswerable due to missing necessary conditions, we prefer LRMs to reason about why it cannot be answered and have the ability to abstain by responding with "I don't know" (i.e., abstention success), rather than generating a reasoning process with hallucination and arriving at an incorrect answer (i.e., abstention failure). Since model reliability is foundational to user trust, LRMs need to possess both strong reasoning abilities and the capacity to abstain from answering unanswerable questions (Kirichenko et al. 2025; Song, Shi, and Zhao 2025). We first validate the phenomenon that LRMs often struggle to abstain. In Figure 2, we evaluate several LRMs on SUM (Song, Shi, and Zhao 2025), which contains mathematical unanswerable questions. Our results show that most LRMs fail to abstain on more than half of the examples. To address the above issue, we systematically analyze, investigate, and propose solutions to improve the abstention behavior for LRMs.

First, to understand how LRMs fail to abstain, we analyze three types of responses generated by LRMs when faced with unanswerable questions. We find that LRMs exhibit two types of behavior when they fail to abstain. Further, we explore their awareness of unanswerable questions. We conduct analysis on LRMs at both external level (intermediate responses during reasoning) and internal level (latent representations). We find that LRMs possess sufficient cognitive capabilities

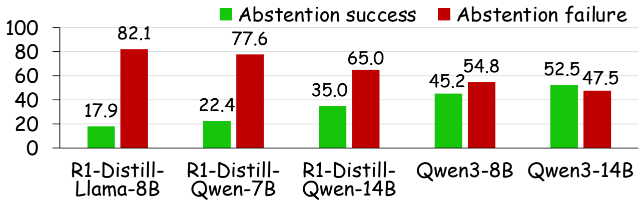

Figure 2: Abstention performance comparison of different LRMs for unanswerable questions on the SUM dataset.

Prompt

{Question} Let's think step by step and output the final answer within \boxed{}. Please solve the problem strictly based on the information provided. Do not introduce any additional assumptions. If you believe the problem lacks sufficient information or is unsolvable, first reply with \boxed{I don't know.}, and then provide your corresponding reason in the format: Reason {your explanation here}.

Figure 3: Prompt used for math problems.

to recognize flaws in such questions. This reveals a misalignment between internal cognition and external output: a LRM internally realizes that a question is unanswerable, yet still fails to act on this realization and abstain accordingly. Answering the unanswerable is to err knowingly.

Second, we seek to improve the abstention ability of LRMs for unanswerable questions. Our further analysis shows that although LRMs may internally exhibit a tendency to abstain during reasoning, such signals are generally not strong enough to interrupt reasoning and result in abstention. Based on this insight, we propose a method that combines cognitive monitoring with inference-time intervention, aiming to improve LRMs' ability to abstain from unanswerable questions while preserving their reasoning abilities on answerable ones.

Finally, we conduct extensive experiments with two datasets using LRMs from various model families and scales. Our method enhances LRMs' ability to abstain from answering unanswerable questions without degrading their reasoning on answerable ones. Furthermore, our experiments reveal that different types of abstention failures benefit from different intervention strategies. Our method achieves significant improvements across all failure types. We release our code at https://github.com/nju-websoft/AbstentionReasoning.

2 Related Work

LLM Abstention on Unanswerable Questions. As the demand for more reliable language models grows, the ability of LLMs to abstain from answering unanswerable questions has become an important evaluation criterion (Yin et al. 2023; Amayuelas et al. 2024; Madhusudhan et al. 2025; Sun et al. 2024; Tomani et al. 2024). LRMs have attracted attention for their strong performance on reasoning tasks (Yang et al. 2025b; Fu et al. 2025). However, their behavior on unanswerable questions has been less studied. Prior work (Kirichenko et al. 2025; Song, Shi, and Zhao 2025) finds that LRMs show weaker abstention ability. We further provide a more detailed analysis of the different types of outputs produced by LRMs when faced with unanswerable questions. We show that LRMs have an internal understanding of a problem's solvability but fail to express it through explicit abstention, and propose a method to improve their abstention behavior.

Inference-Time Improvements for LRMs. While LRMs benefit from richer chain-of-thought (CoT) reasoning and achieve strong performance on complex tasks, recent efforts have have begun to analyze (Dang, Huang, and Chen 2025; Zhu et al. 2025; Liu et al. 2025) and implement (Chen et al. 2025) inference-time interventions to further enhance their reasoning accuracy and efficiency. For example, (Fu et al. 2025) and (Li et al. 2024) monitor intermediate consistency between reasoning steps to decide when to output the answer. (Yang et al. 2025b) estimates confidence in intermediate steps to decide whether the model has reached a sufficiently certain conclusion. (Liao et al. 2025) uses process-level reward models to guide the model toward better reasoning trajectories. However, the above methods are designed for answerable questions. In our work, we evaluate their effectiveness on unanswerable questions and propose a new inference-time intervention to improve the abstention capability for LRMs.

3 Analysis of Abstention Failure

We first analyze how LRMs respond to unanswerable math problems. Then, we investigate LRMs' awareness of unanswerable questions by examining whether they possess the ability to recognize the unanswerability of such questions, from both internal and external perspectives.

LRMs. We evaluate five LRMs across diverse model families and scales, including R1-Distill-Llama-8B, R1-Distill-Qwen-7B, R1-Distill-Qwen-14B (DeepSeek-AI et al. 2025), Qwen3-8B, and Qwen3-14B (Yang et al. 2025a).

Datasets. Following previous work (Song, Shi, and Zhao 2025; Ouyang 2025), we focus on unanswerable math problems, i.e., ill-posed problems that are challenging for LRMs and offer objectively defined criteria for unanswerability. Specifically, we use the Synthetic Unanswerable Math (SUM) (Song, Shi, and Zhao 2025) dataset, which includes diverse problems from AIME (1984–2023), AMC (pre-2023), Omni-MATH (Gao et al. 2025), and Still (Team 2025). The unanswerable problems in SUM are generated based on five criteria: (1) key information deletion, (2) ambiguous key information, (3) unrealistic conditions, (4) unrelated objects, and (5) question deletion. We randomly sample 1,000 problems for our analysis. For each unanswerable case, we use GPT-4o to generate a brief explanation as the ground-truth rationale.

Objective and Prompt. Given a question, we encourage the LRM to perform normal reasoning when the question is answerable, and to respond with "I don't know." along with a corresponding explanation when it is not (i.e., correct abstention). The prompt used in this task is shown in Figure 3.

3.1 Reactions to Unanswerable Questions

We evaluate the reactions of LRMs to unanswerable questions using the prompts in Figure 3 with a maximum token budget of 10,000 per response. We identify three distinct response types, as shown in Figures 4:

Correct abstention

<think>\nOkay, so I have this problem here: I need to figure out [(1 ⊗ 2) ⊗ 3] minus [1 ⊗ (2 ⊗ 3)], where ⊗ is some operation that combines two nonzero numbers. Wait, the problem is asking me to compute this expression, but **I don't know** what the operation ⊗ actually does…
</think>\n\nThe problem requires determining the value of [(1 ⊗ 2) ⊗ 3] - [1 ⊗ (2 ⊗ 3)] where ⊗ is an undefined operation. **Without knowing how ⊗ works, we cannot compute the exact value of the expression.** Therefore, the problem lacks sufficient information to solve it.\n\n\\boxed{**I don't know.**}

Hallucinated answer

<think> First, I need to determine the total cost of purchasing the appliance through the television advertisement. The advertisement offers three payments of $29.98 each, plus **a one-time shipping and handling charge of $9.99.**
Next…
</think>\n\nTo determine the savings by purchasing the appliance through the television advertisement…
\n\n**Final Answer:**\n\\[\n**boxed{6}**\n\\]

Cognitive fixation

<think>\nOkay, so I have this problem here: I need to figure out what [(1 ⊗ 2) ⊗ 3] minus [1 ⊗ (2 ⊗ 3)] is, where the operation ⊗ is some custom operation used to combine two nonzero numbers…
Wait, but the problem says \"the operation ⊗ is used to combine two nonzero numbers,\" so maybe it's expecting me to realize that…
Wait, but the problem says \"the operation ⊗ is used to combine two nonzero numbers,\" so maybe the problem lacks sufficient information …
Wait, but the problem says \"the operation ⊗ is used to combine two nonzero numbers,\" so maybe the operation is associative?…
… … …

Figure 4: Example of different types of response outcomes.

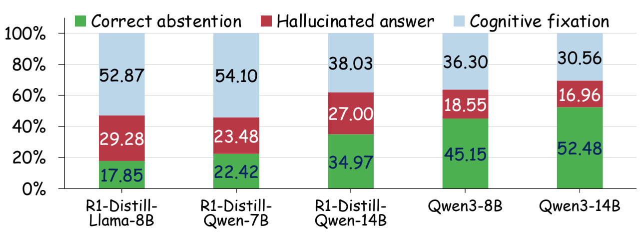

Figure 5: Distribution of the response types of LRMs on unanswerable math problems.

- **Correct abstention:** The LRM identifies the question as unanswerable with the given information and explicitly abstains from answering, typically responding with statements such as *"I don't know"*.
- **Hallucinated answer:** The LRM produces a complete solution by assuming or fabricating missing details not present in the question. As illustrated in Figure 4, the LRM infers a $9.99 handling charge that is never mentioned in the input in order to compute a final answer.
- **Cognitive fixation:** The LRM fails to reach a conclusion within the token limit. It often enters a prolonged reasoning process, stubbornly reformulating or pursuing invalid solution paths without terminating the response, even after recognizing the question is unanswerable.

Figure 5 shows the response type distribution. As model capacity increases, the proportion of correct abstention tends to rise, while those of hallucinated answer and cognitive fixation decrease. A substantial portion of unanswerable questions do not receive correct abstentions. Overall, the results reveal a

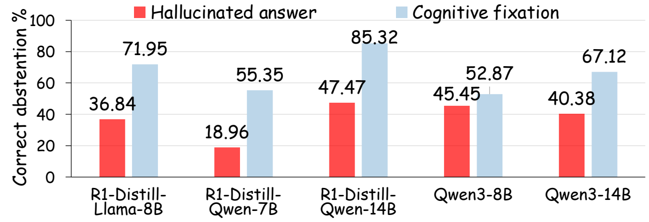

Figure 6: The proportion of questions that can respond with a correct abstention in stopping points during the reasoning in the types of "hallucinated answer" and "cognitive fixation".

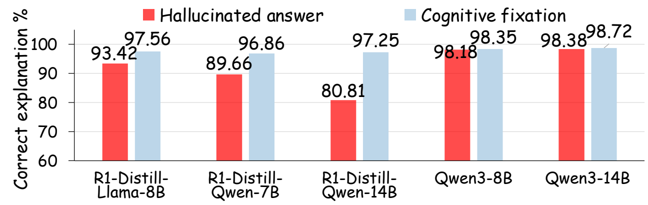

Figure 7: The proportion of questions that can provide a correct explanation in stopping points during the reasoning.

core limitation: LRMs often fail to abstain from answering unanswerable problems, despite increased model capacity.

3.2 Awareness of Unanswerable Questions

We investigate whether LRMs can recognize unanswerable math problems, probing both their external behavior and internal cognition. Our two-layered analysis assesses (1) behavioral signals: whether LRMs outwardly indicate question answerability, and (2) latent signals: whether answerability is internally encoded during reasoning.

Behavioral Signals of Question Answerability. Inspired by prior work (Yang et al. 2025b; Chen et al. 2025; Wang et al. 2025) showing that LRMs form and revise intermediate conclusions during reasoning, we insert stopping points into the reasoning trajectory. At each stopping point, we prompt the LRM to (1) directly provide an answer, and (2) to explain why the question is unanswerable (details in the appendix).

We apply this intervention to the two failure types. For "cognitive fixation", we use the keyword "wait" as a stopping point and prompt the model to directly output its answer. For "hallucinated answer", which typically have shorter reasoning trajectories, we use "\n\n" as the stopping point. We then compute the proportion of questions in each category where the model can respond with a correct abstention at stopping points. As shown in Figure 6, for cognitive fixation, more than half of the cases can result in correct abstentions. A notable portion of hallucinated answer cases receive correct abstentions, and the rate improves as model size increases.

Since "I don't know" is a simple response and may be produced randomly, we additionally prompt the LRMs at each stopping point to explain why the question is unanswerable, to better assess their awareness of unanswerability. We measure the percentage of questions in each category with correct explanations. Figure 7 shows that both failure types

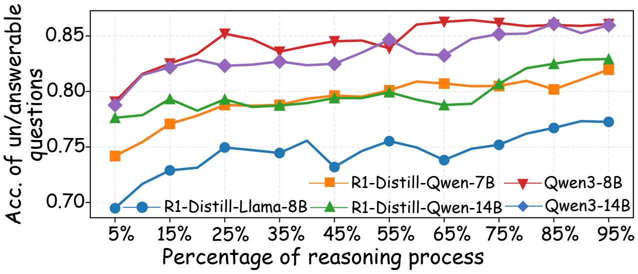

Figure 8: Classification accuracy of answerable and unanswerable questions at varying stages of the reasoning process.

yield high percentages of correct explanations.

Overall, even when an LRM fails to abstain, it may still recognize unanswerability during reasoning. Its ability to assess answerability exists but is underused in final decisions.

Latent Signals of Question Answerability. We conduct a probing-based analysis on the latent representations during reasoning. The linear representation hypothesis posits that high-level concepts, such as language, gender, and truthfulness, are linearly embedded in the latent space of LLMs (Park, Choe, and Veitch 2024). We hypothesize that question answerability may also be linearly represented during the reasoning process. Inspired by prior work on representation probing (Li et al. 2023; Orgad et al. 2025; Marks and Tegmark 2023; Slobodkin et al. 2023), we train lightweight linear classifiers (probes) on hidden activations from the LRMs' reasoning trajectory. The goal is to assess whether answerable and unanswerable questions can be distinguished from latent representations during the reasoning process, thereby revealing the presence of answerability-related signals in internal state.

We use the output of the multi-head attention before the residual connection as the input to the probe (Li et al. 2023):

$$x_l^c = \sum_{h=1}^{H} Q_l^h \, \mathrm{Att}_l^h(P_l^h x_l), \tag{1}$$

where x_l is the input of layer l, P_l^h projects the input into a head-specific subspace, Q_l^h maps it back, Att is the attention operator. The probe is defined as a simple linear classifier: $p_\theta(x_l^c) = \sigma(\langle \theta, x_l^c \rangle)$, where θ is the trainable weight and σ denotes the sigmoid function. One probe is trained per layer.

We sample 2,200 pairs of answerable and unanswerable questions from the SUM dataset (2,000 pairs for training and 200 for validation). For each question, we randomly sample 1,000 token-level activations x_l^c from the reasoning trajectory, and construct a dataset $\left\{(x_l^c, y)_i\right\}_{i=1}^N$, where $y \in \{0, 1\}$ indicates question answerability. At inference time, we aggregate the prediction probabilities across all tokens up to the current reasoning step and use the average as the overall answerability prediction. We select the optimal probing layer for each model based on the validation set, and evaluate on the test dataset used for analysis in the previous section.

The results are shown in Figure 8, where we plot the probe's classification accuracy at various percentages of the reasoning process, using a 0.5 threshold to distinguish unanswerable (prediction probability of probe > 0.5) from answerable cases. For all LRMs, we observe that the probe's classification accuracy increases steadily as reasoning progresses, with most accuracies exceeding 0.8 by the end of the trajectory. These results suggest that signals related to question answerability are indeed encoded in the representations of LRMs during reasoning. Although these signals may not always be reflected in the LRMs' final output behavior, they are implicitly present in the internal computation process.

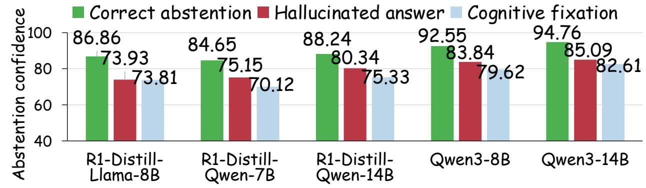

Figure 9: The confidence of abstention answer: "I don't know" in different types of response.

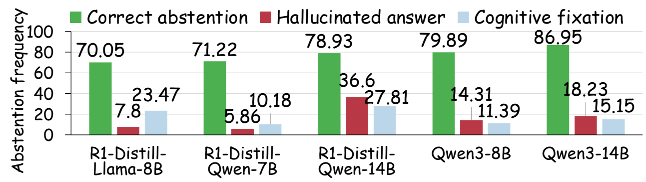

Figure 10: The frequency of abstention responses across stopping points in different types of response outcomes.

4 Mitigation of Abstention Failure

We seek to mitigate abstention failures of LRMs when faced with unanswerable questions. Although LRMs show signs of recognizing unanswerable questions during reasoning, they often fail to abstain in their final answers. We aim to identify the causes of this misalignment and explore ways to fix it.

First, we examine the LRMs' confidence (Kuhn, Gal, and Farquhar 2023) in abstaining at the stopping point across three output types. We sample an equal number of questions from each output type and compute the average confidence in generating "I don't know" at the stopping points. Following (Yang et al. 2025b), the confidence score C is computed as

$$p(a_t) = P(a_t \,|\, H, I, a_{<t}), C = \left(\sum_{i=1}^{n} \max_{a_t \in V} p(a_t)\right)^{\frac{1}{n}}, \tag{2}$$

where $p(a_t)$ is the model's predicted probability of answer token a_t, H includes the input prompt and the generated thoughts, I is the prompt to elicit the answer, n is the number of decoding steps, and V is the vocabulary. In Figure 9, hallucinated answer and cognitive fixation both exhibit lower confidence in abstention compared to correct abstention.

Next, we analyze the average frequency of "I don't know" outputs by the LRMs across all stopping points in the reasoning trajectory. In Figure 10, the hallucinated answer and cognitive fixation types consistently show lower abstention frequencies than correct abstention.

Method	SUM					UMWP				
	Unanswerable			Answerable		Unanswerable			Answerable	
	Abstention ↑	Reason Acc ↑	Token ↓	Answer Acc ↑	Token ↓	Abstention ↑	Reason Acc ↑	Token ↓	Answer Acc ↑	Token ↓
R1-Distill-Llama-8B										
Vanilla	16.90%	14.44	5088	61.97	3446	30.67%	26.00	1829	77.67	613
Dynasor-CoT	35.92%	27.82	3084	55.28	**2619**	39.33%	30.67	**958**	73.33	**495**
DEER	24.30%	19.79	4339	**60.92**	2675	34.11%	28.67	1616	**77.67**	637
Ours	**60.92%**	**53.17**	**2419**	**60.92**	3151	**54.67%**	**44.00**	1246	77.33	574
R1-Distill-Qwen-7B										
Vanilla	21.13%	19.37	4878	69.72	3169	47.67%	43.67	1935	90.30	597
Dynasor-CoT	56.34%	45.89	**1869**	62.68	**2074**	64.33%	53.00	**763**	88.67	**486**
DEER	28.87%	25.70	3747	63.73	2191	54.33%	49.33	1335	**91.67**	590
Ours	**73.94%**	**61.86**	2247	**67.25**	3001	**77.33%**	**64.33**	1256	90	569
R1-Distill-Qwen-14B										
Vanilla	36.97%	33.45	3820	70.42	2671	49.67%	45.00	539	90.00	384
Dynasor-CoT	51.17%	45.18	2622	66.20	**2029**	50.67%	46.00	**504**	90.00	**384**
DEER	39.79%	36.97	3417	63.75	2303	50.33%	46.67	644	**90.33**	532
Ours	**74.30%**	**62.68**	**2621**	**67.96**	2541	**60.33%**	**54.33**	606	89.67	388
Qwen3-8B										
Vanilla	47.18%	41.90	4411	60.92	4245	80.00%	72.33	1906	94.33	1317
Dynasor-CoT	65.61%	58.21	**1710**	60.27	2190	83.67%	75.00	736	92.00	803
DEER	63.38%	51.17	1902	**63.77**	**1993**	80.00%	70.33	**479**	**94.33**	**474**
Ours	**75.27%**	**64.44**	2912	61.62	3875	**87.33%**	**79.67**	980	93.67	1252
Qwen3-14B										
Vanilla	54.22%	48.24	3713	66.55	3768	82.33%	76.67	1279	94.33	877
Dynasor-CoT	66.18%	56.62	**1375**	63.38	**1862**	84.00%	76.67	752	**92.67**	662
DEER	63.90%	56.91	1749	**68.18**	2192	83.33%	75.67	**408**	93.00	**447**
Ours	**78.17%**	**69.01**	2311	65.03	3528	**92.67%**	**82.67**	959	92.48	848

Table 1: Performance of methods on (un)answerable questions across LRMs. Best scores are marked in **bold**.

All the findings suggest that while the LRM may recognize a question as unanswerable, it often lacks sufficient confidence to act upon it. The gap between internal awareness and output behavior reveals a key misalignment in LRMs: though they may be aware of unanswerability, their decision process remains biased toward answering rather than abstaining.

Motivated by the above findings, we propose a method to help LRMs improve their abstention capability. It consists of two key components: cognitive monitoring and inference-time intervention. The goal is to monitor the LRM's evolving recognition of question unanswerability during reasoning, and to intervene when necessary to guide the LRM toward making an abstention by encouraging abstention behaviors.

Cognitive Monitoring. The first step of our method aims to identify when the model internally recognizes that a question may be unanswerable. We track the token-level hidden states generated during inference and segment the reasoning process into semantically coherent units (e.g., clauses or transitions marked by discourse cues such as "wait"). At the end of each segment, we apply a lightweight linear probe, which is trained from our analysis of latent signals of question answerability, to estimate the probability that the current question is unanswerable. If the predicted probability exceeds a threshold, the model takes the intervention strategy.

Inference-Time Intervention. Once the LRM exhibits sufficient internal evidence that a question is unanswerable, we intervene the reasoning process to reinforce this recognition and increase the likelihood of a correct abstention. The intervention is implemented via an instructional guidance prompt

Intervention Prompt

Instruction:\n You are not permitted to make assumptions that are not explicitly stated in the question. There are signs that this question may lack sufficient information to answer definitively. If you find that any part of your reasoning depends on undefined operations, missing values, or unspecified conditions, you must immediately stop and output: \boxed{I don't know.} Do not attempt to guess, infer, or continue reasoning with incomplete information. This is a strict constraint! \n</think>\n\n

Figure 11: The prompt used for inference-time intervention.

(see Figure 11). We append a prompt that restates the possibility that the question may be unanswerable, helping the LRM overcome cognitive fixation and avoid speculative guesses based on missing or unstated information. Inspired by prior work (Yang et al. 2025b; Fu et al. 2025), we also incorporate an early exit strategy to prevent unnecessary continuation of reasoning. It is designed to be minimally intrusive yet semantically salient, encouraging the model to consider abstention as a valid and even preferred option under certain conditions.

5 Experiments

5.1 Setup

Datasets. We use SUM (Song, Shi, and Zhao 2025) and UMWP (Sun et al. 2024). For SUM, we used its test set, which contains 284 unanswerable and 284 answerable questions manually verified. UMWP is derived from four widely used math word problem datasets: SVAMP (Patel, Bhattamishra, and Goyal 2021), MultiArith (Koncel-Kedziorski et al. 2016), GSM8K (Cobbe et al. 2021), and ASDiv (Miao,

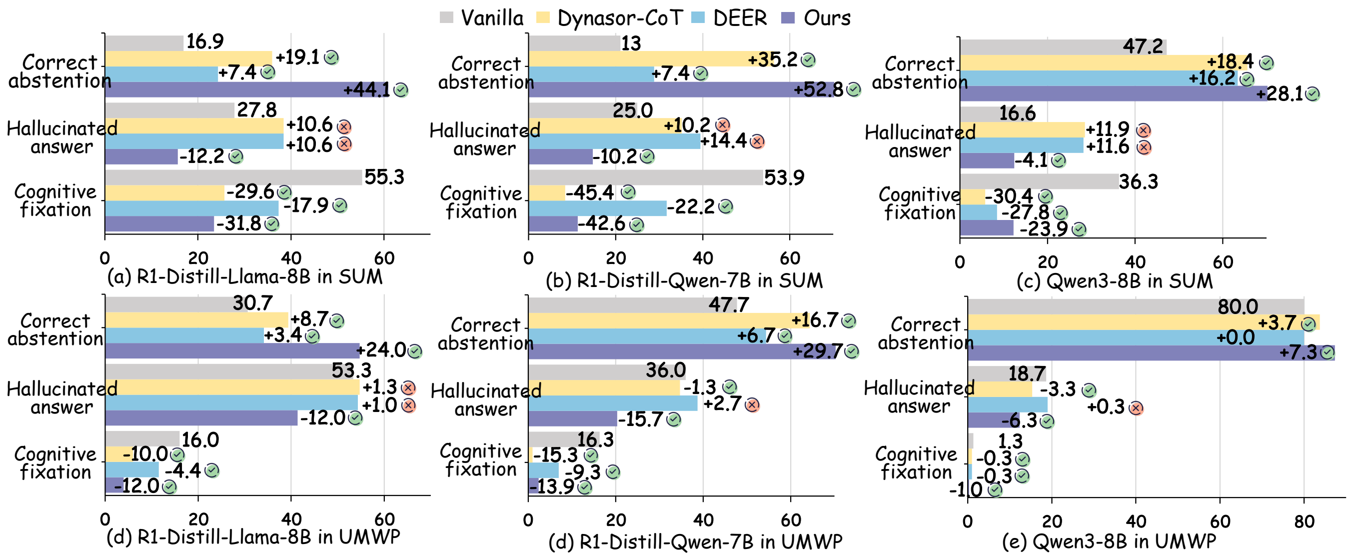

Figure 12: Comparison of response type distributions across different methods on unanswerable questions. Proportions are reported across models and datasets, with numbers indicating absolute changes from the vanilla model.

Liang, and Su 2020). The full dataset consists of 5,200 instances. We sample 600 questions (300 unanswerable and 300 answerable) to form the test set. For the unanswerable questions in UMWP, we also use GPT-4o to generate the reasoning behind their unanswerability as ground truth.

Metrics. Following (Song, Shi, and Zhao 2025; Kirichenko et al. 2025), we evaluate model performance with (1) *Abstention Rate*: The proportion of unanswerable questions for which the model correctly abstains from answering by outputting "I don't know" in accordance with the instruction. (2) *Reason Accuracy*: The percentage of unanswerable questions for which the model provides the correct rationale for unanswerability. (3) *Token Usage*: The number of tokens generated by the model during the reasoning process. (4) *Answer Accuracy*: The proportion of answerable questions for which the model produces the correct final answer.

Baselines and LRMs. Our method is an inference-time intervention for LRMs without training. There are no directly comparable baselines. We select two baselines based on output consistency (Dynasor-CoT) and confidence (DEER), assuming that correct abstention is the correct answer for unanswerable questions. Dynasor-CoT (Fu et al. 2025) prompts intermediate answers and halts when the same answer appears three times consecutively. DEER (Yang et al. 2025b) monitors sentence-level confidence and exits early once a threshold is met. Besides, the Vanilla method denotes unaltered LRM outputs. For LRMs, we choose R1-Distill-Llama-8B, R1-Distill-Qwen-7B, R1-Distill-Qwen-14B, Qwen3-8B, and Qwen3-14B. More details are in the appendix.

5.2 Main Results

Table 1 shows the main results. We have the following observations: (1) Our method achieves the highest Abstention Rate and Reason Accuracy on unanswerable questions, demonstrating its strong ability to guide LRMs in recognizing and abstaining from unanswerable inputs. (2) Our method maintains comparable Answer Accuracy on answerable questions. In most settings, accuracy remains close to the vanilla model and sometimes improves slightly, indicating minimal impact on solving answerable tasks. (3) Our method can reduce token usage on unanswerable questions, with an average reduction of 30–50% across different LRMs compared to the vanilla model. It also slightly decreases token consumption on answerable questions, indicating improved reasoning efficiency. (4) We observe a positive correlation between Abstention Rate and Reason Accuracy: as the LRM becomes better at abstaining from unanswerable questions, it also produces more accurate explanations. This suggests that the intervention not only changes the final output but also improves intermediate reasoning quality. (5) Qwen3 models generally outperform other models in abstention-related metrics. Larger LRMs tend to exhibit stronger abstention capabilities. This trend indicates that both model scale and architecture contribute to reliable unanswerability detection.

5.3 Further Discussions

Impact on Response Type Distribution. We analyze the changes in the proportions of different response types across methods and datasets. Figure 12 shows that our method consistently reduces the hallucinated answers and cognitive fixation outputs, which contribute to a substantial increase in the rate of correct abstentions. While Dynasor-CoT and DEER employ early exits to mitigate cognitive fixation, they lead to a higher proportion of hallucinated answers. Early exits without appropriate guidance may cause the model to make up assumptions or imagined scenarios for giving a definite answer, rather than acknowledging uncertainty. This highlights the importance of combining monitoring with guided intervention to steer LRMs toward proper behavior.

Method	SUM		UMWP	
	Abst. conf.	Abst. rate	Abst. conf.	Abst. rate
R1-Distill-Llama-8B				
Pre-Interv.	79.7	30.1%	84.1	41.5%
Post-Interv.	87.3 (↑9.4%)	78.1% (×2.6)	90.0 (↑7.0%)	81.4% (×1.9)
R1-Distill-Qwen-7B				
Pre-Interv.	77.1	24.5%	87.1	50.8%
Post-Interv.	86.8 (↑12.6%)	80.6% (×3.3)	92.1 (↑5.7%)	71.5% (×1.4)
Qwen3-8B				
Pre-Interv.	90.9	48.3%	90.6	75.9%
Post-Interv.	98.9 (↑8.7%)	74.9% (×1.5)	98.1 (↑8.2%)	93.1% (×1.2)

Table 2: Results of intervention effects. "Abst. conf." denotes the average abstention confidence when getting the answer "I don't know". "Interv." is the inference-time intervention.

Intervention Effects. We assess how our intervention influences the LRM's confidence in abstention and its actual abstention behavior, to evaluate whether our method helps bridge the gap between cognitive awareness and abstention behavior. At the intervention point identified by our method, we prompt the LRM to generate intermediate outputs before and after the intervention. We measure two key indicators: the confidence when producing "I don't know" responses, and the proportion of questions for which the model outputs "I don't know". Table 2 shows that our method consistently enhances the confidence in generating abstention responses. The abstention rate also shows corresponding improvements.

Further Analysis of Cognitive Monitoring. We analyze the cognitive monitoring component and compare our default monitoring strategy based on latent representations with alternative strategies relying on behavioral signals. The behavioral signal approach monitors the LRMs' intermediate outputs at the end of the paragraph generation phase (e.g., when it reaches a "wait" token), and uses these outputs to determine whether to trigger an intervention. We investigate three variants: The Direct Behavior strategy checks whether the model's intermediate output is "I don't know" and triggers an intervention immediately if so. The Consistency strategy triggers intervention only if the model produces "I don't know" in three consecutive intermediate outputs (inspired by Dynasor-CoT). The Confidence Score strategy triggers intervention when the model outputs "I don't know" with a confidence score exceeding a predefined threshold (inspired by DEER). In Table 3, all monitoring strategies contribute to the improvement in abstention behavior, showing that cognitive monitoring is generally effective. The strategy based on latent representation signals achieves the best and most consistent performance across models and datasets. The Direct Behavior method is simple and works well, but it may be too aggressive and hurt performance on answerable questions.

Ablation Study. We evaluate two aspects of inference-time intervention: the instructional guidance prompt and the early exit strategy. We analyze how each aspect affects abstention behavior and answer quality. The results are shown in Table 4. For correct abstention, the impact of instructional guidance is greater than that of early exit. The early exit strategy helps reduce the number of cognitive fixation cases. Without instructional guidance, the proportion of hallucinated answers

Monitoring Signal	Unanswerable			Ans.
	Correct abstention ↑	Hallucinated answer ↓	Cognitive fixation ↓	Acc ↑
R1-Distill-Llama-8B				
Vanilla	16.9	27.8	55.3	61.9
Latent Representation	**60.9**	**15.7**	**23.4**	60.9
Direct Behavior	53.1	20.4	26.5	58.1
Consistency	47.9	23.6	28.5	59.8
Confidence Score	37.0	24.7	38.4	**61.3**
R1-Distill-Qwen-7B				
Vanilla	21.1	25.0	53.9	69.7
Latent Representation	**73.9**	**14.8**	**11.3**	67.3
Direct Behavior	41.6	22.6	35.9	66.8
Consistency	35.7	23.1	41.2	69.3
Confidence Score	31.2	23.9	44.9	**69.4**
Qwen3-8B				
Vanilla	47.2	16.6	36.3	60.9
Latent Representation	**75.3**	12.4	**12.3**	**61.6**
Direct Behavior	67.3	10.9	21.8	60.2
Consistency	61.6	13.0	25.4	61.3
Confidence Score	64.3	**10.6**	25.2	61.3

Table 3: Comparison of cognitive monitoring strategies.

Variant	Unanswerable			Ans.
	Correct abstention ↑	Hallucinated answer ↓	Cognitive fixation ↓	Acc ↑
R1-Distill-Llama-8B				
Vanilla	16.9	27.8	55.3	61.9
Ours	60.9	15.7	23.4	60.9
w/o Early Exit	43.3 (↑**26.4**)	18.3 (↓**9.5**)	38.4 (↓**16.9**)	61.9
w/o Instr. Guidance	26.1 (↑9.2)	33.4 (↑5.6)	40.5 (↓14.8)	**62.3**
R1-Distill-Qwen-7B				
Vanilla	21.1	25.0	53.9	69.7
Ours	73.9	14.8	11.3	67.3
w/o Early Exit	49.7 (↑**28.6**)	16.2 (↓**8.8**)	34.2 (↓19.7)	69.0
w/o Instr. Guidance	43.5 (↑22.4)	35.2 (↑10.2)	21.3 (↓**32.6**)	**70.0**
Qwen3-8B				
Vanilla	47.2	16.6	36.3	60.9
Ours	75.3	12.4	12.3	61.6
w/o Early Exit	73.6 (↑**26.4**)	13.7 (↓**2.9**)	12.7 (↓23.6)	**62.7**
w/o Instr. Guidance	59.2 (↑12.0)	30.3 (↑13.7)	10.6 (↓**25.7**)	**62.7**

Table 4: Ablation results of intervention components across different LRMs on SUM. We report the effect of removing either Instructional Guidance or Early Exit component on three types of responses for unanswerable questions, as well as the accuracy on answerable questions.

increases. This again shows that without proper guidance, the model tends to make up conditions and generate unsupported answers. Instructional guidance also has a slight impact on the performance of answerable questions.

6 Conclusion and Future Work

We investigate the failure of LRMs to abstain from answering unanswerable questions, despite having the cognitive ability to detect the unanswerability. We identify a misalignment between the model's internal cognition and its external response behavior. We propose a lightweight, two-stage method that significantly improves abstention behavior without harming reasoning performance. Future work aims to explore training-time alignment strategies to improve abstention fidelity.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 62272219 and 62406136).

References

Amayuelas, A.; Wong, K.; Pan, L.; Chen, W.; and Wang, W. Y. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. In *ACL (Findings)*, 6416–6432. Association for Computational Linguistics.

Chen, R.; Zhang, Z.; Hong, J.; Kundu, S.; and Wang, Z. 2025. SEAL: Steerable Reasoning Calibration of Large Language Models for Free. *arXiv preprint arXiv:2504.07986*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Dang, R.; Huang, S.; and Chen, J. 2025. Internal Bias in Reasoning Models leads to Overthinking. *arXiv preprint arXiv:2505.16448*.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; and Li, S. S. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Fu, Y.; Chen, J.; Zhuang, Y.; Fu, Z.; Stoica, I.; and Zhang, H. 2025. Reasoning without self-doubt: More efficient chain-of-thought through certainty probing. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Gao, B.; Song, F.; Yang, Z.; Cai, Z.; Miao, Y.; Dong, Q.; Li, L.; Ma, C.; Chen, L.; Xu, R.; Tang, Z.; Wang, B.; Zan, D.; Quan, S.; Zhang, G.; Sha, L.; Zhang, Y.; Ren, X.; Liu, T.; and Chang, B. 2025. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark for Large Language Models. In *ICLR*. OpenReview.net.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; Iftimie, A.; Karpenko, A.; Passos, A. T.; Neitz, A.; Prokofiev, A.; Wei, A.; Tam, A.; Bennett, A.; Kumar, A.; Saraiva, A.; Vallone, A.; Duberstein, A.; Kondrich, A.; Mishchenko, A.; Applebaum, A.; Jiang, A.; Nair, A.; Zoph, B.; Ghorbani, B.; Rossen, B.; Sokolowsky, B.; Barak, B.; McGrew, B.; Minaiev, B.; Hao, B.; Baker, B.; Houghton, B.; McKinzie, B.; Eastman, B.; Lugaresi, C.; Bassin, C.; Hudson, C.; Li, C. M.; de Bourcy, C.; Voss, C.; Shen, C.; Zhang, C.; Koch, C.; Orsinger, C.; Hesse, C.; Fischer, C.; Chan, C.; Roberts, D.; Kappler, D.; Levy, D.; Selsam, D.; Dohan, D.; Farhi, D.; Mely, D.; Robinson, D.; Tsipras, D.; Li, D.; Oprica, D.; Freeman, E.; Zhang, E.; Wong, E.; Proehl, E.; Cheung, E.; Mitchell, E.; Wallace, E.; Ritter, E.; Mays, E.; Wang, F.; Such, F. P.; Raso, F.; Leoni, F.; Tsimpourlas, F.; Song, F.; von Lohmann, F.; Sulit, F.; Salmon, G.; Parascandolo, G.; Chabot, G.; Zhao, G.; Brockman, G.; Leclerc, G.; Salman, H.; Bao, H.; Sheng, H.; Andrin, H.; Bagherinezhad, H.; Ren, H.; Lightman, H.; Chung, H. W.; Kivlichan, I.; O'Connell, I.; Osband, I.; Gilaberte, I. C.; and Akkaya, I. 2024. OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720*.

Kirichenko, P.; Ibrahim, M.; Chaudhuri, K.; and Bell, S. J. 2025. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. *arXiv preprint arXiv:2506.09038*.

Koncel-Kedziorski, R.; Roy, S.; Amini, A.; Kushman, N.; and Hajishirzi, H. 2016. MAWPS: A Math Word Problem Repository. In *HLT-NAACL*, 1152–1157. The Association for Computational Linguistics.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *ICLR*. OpenReview.net.

Li, K.; Patel, O.; Viégas, F. B.; Pfister, H.; and Wattenberg, M. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *NeurIPS*.

Li, Y.; Yuan, P.; Feng, S.; Pan, B.; Wang, X.; Sun, B.; Wang, H.; and Li, K. 2024. Escape Sky-high Cost: Early-stopping Self-Consistency for Multi-step Reasoning. In *ICLR*. OpenReview.net.

Liao, B.; Xu, Y.; Dong, H.; Li, J.; Monz, C.; Savarese, S.; Sahoo, D.; and Xiong, C. 2025. Reward-Guided Speculative Decoding for Efficient LLM Reasoning. *arXiv preprint arXiv:2501.19324*.

Liu, Y.; Sun, Z.; Shao, Z.; Cui, Y.; and Hu, W. 2025. Are LLMs Really Knowledgeable for Knowledge Graph Completion? In *International Semantic Web Conference*, 77–95.

Madhusudhan, N.; Madhusudhan, S. T.; Yadav, V.; and Hashemi, M. 2025. Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models. In *COLING*, 9329–9345. Association for Computational Linguistics.

Marks, S.; and Tegmark, M. 2023. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv preprint arXiv:2310.06824*.

Miao, S.; Liang, C.; and Su, K. 2020. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *ACL*, 975–984. Association for Computational Linguistics.

Orgad, H.; Toker, M.; Gekhman, Z.; Reichart, R.; Szpektor, I.; Kotek, H.; and Belinkov, Y. 2025. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In *ICLR*. OpenReview.net.

Ouyang, J. 2025. TreeCut: A Synthetic Unanswerable Math Word Problem Dataset for LLM Hallucination Evaluation. *arXiv preprint arXiv:2502.13442*.

Park, K.; Choe, Y. J.; and Veitch, V. 2024. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *ICML*. OpenReview.net.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *NAACL-HLT*, 2080–2094. Association for Computational Linguistics.

Slobodkin, A.; Goldman, O.; Caciularu, A.; Dagan, I.; and Ravfogel, S. 2023. The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models. In *EMNLP*, 3607–3625. Association for Computational Linguistics.

Song, L.; Shi, T.; and Zhao, J. 2025. The Hallucination Tax of Reinforcement Finetuning. *arXiv preprint arXiv:2505.13988*.

Sun, Y.; Yin, Z.; Guo, Q.; Wu, J.; Qiu, X.; and Zhao, H. 2024. Benchmarking Hallucination in Large Language Models Based on Unanswerable Math Word Problem. In *LREC/-COLING*, 2178–2188. ELRA and ICCL.

Team, Q. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Tomani, C.; Chaudhuri, K.; Evtimov, I.; Cremers, D.; and Ibrahim, M. 2024. Uncertainty-Based Abstention in LLMs Improves Safety and Reduces Hallucinations. *arXiv preprint arXiv:2404.10960*.

Wang, Y.; Liu, Q.; Xu, J.; Liang, T.; Chen, X.; He, Z.; Song, L.; Yu, D.; Li, J.; Zhang, Z.; Wang, R.; Tu, Z.; Mi, H.; and Yu, D. 2025. Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs. *arXiv preprint arXiv:2501.18585*.

Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; Shao, C.; Yan, Y.; Yang, Q.; Song, Y.; Ren, S.; Hu, X.; Li, Y.; Feng, J.; Gao, C.; and Li, Y. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025a. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, C.; Si, Q.; Duan, Y.; Zhu, Z.; Zhu, C.; Lin, Z.; Cao, L.; and Wang, W. 2025b. Dynamic Early Exit in Reasoning Models. *arXiv preprint arXiv:2504.15895*.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? In *ACL (Findings)*, 8653–8665. Association for Computational Linguistics.

Zhu, R.; Liu, Y.; Sun, Z.; Wang, Y.; and Hu, W. 2025. When Can Large Reasoning Models Save Thinking? Mechanistic Analysis of Behavioral Divergence in Reasoning. *arXiv preprint arXiv:2505.15276*.