ThinkHalu: A Benchmark for Detecting Thinking-Process Hallucinations in Large Reasoning Models

Anonymous ACL submission

Abstract

Large Reasoning Models (LRMs) have achieved notable progress in both information retrieval and complex reasoning tasks. However, LRMs often suffer from reasoning hallucinations, which compromise the accuracy and interpretability of their outputs. Although the "LLM-as-a-Judge" paradigm shows potential in evaluating factual hallucinations, it still lacks systematic exploration and standardized benchmarks for identifying thinking process hallucinations. To address this, we construct a new benchmark dataset **ThinkHalu**, covering cognitive hallucinations (miscomprehension of the question) and logical hallucinations (logical errors in reasoning steps). Our benchmark uses GSM8K and MathQA as base datasets and generates hallucinated thinking processes with eight open-source LRMs. We select high-quality data by analyzing the semantic similarity and natural language inference (NLI) relationships between hallucinated thinking and correct solution processes. Then, we identify and analyze hallucination types using an LLM-based voting mechanism. We conduct experiments on 20 open-source LLMs and observe that they tend to misclassify cognitive hallucinations as logical hallucinations, indicating limitations in their ability to detect hallucinations in thinking processes.

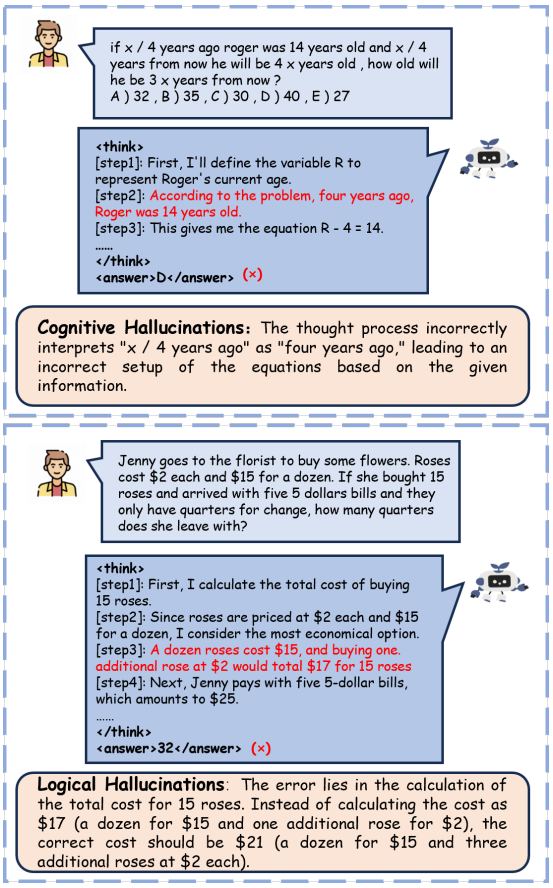

Figure 1: ThinkHalu benchmark defines logical hallucinations and cognitive hallucinations in the hallucination thinking process.

1 Introduction

Large Language Models (LLMs) represent a significant milestone in the advancement of artificial intelligence, demonstrating impressive capabilities across a range of tasks, including information retrieval (Zhu et al., 2024) and recommendation systems (Wu et al., 2024). In particular, Large Reasoning Models (LRMs), which are further refined through supervised fine-tuning and reinforcement learning, exhibit advanced reasoning capabilities in complex scenarios such as mathematical problem solving (Gehring et al., 2025), question answering (Webb et al., 2023), and strategic planning (Xu et al., 2025).

Before producing an answer, LRMs typically generate a complete thinking process, including question comprehension, information analysis, and conclusion derivation (Xu et al., 2025). This process reflects the model's logical trajectory when constructing conclusions. The thinking process of LRMs significantly optimizes the quality of responses by improving reasoning accuracy and standardizing structured outputs (DeepSeek-AI, 2025).

Despite the impressive reasoning capabilities of current LRMs, their thinking processes still often suffer from severe hallucinations, and such unreliable reasoning can compromise both the accuracy and interpretability of the final output (Li et al., 2025). This phenomenon, termed thinking process hallucination, reveals a limitation in LRMs' ability to perform complex reasoning tasks, thereby necessitating systematic detection and comprehensive analysis.

In recent years, the "LLM-as-a-Judge" paradigm (Miao et al., 2023; Hu et al., 2024) has emerged, using the language understanding and reasoning capabilities of LLMs to evaluate generated content (Chen et al., 2025). This paradigm demonstrates strong potential for semantic structure analysis and the assessment of logical consistency under given conditions. However, detecting hallucinations in the thinking process remains underexplored, with a lack of standardized evaluation benchmarks.

In this paper, we propose a novel benchmark dataset for detecting thinking process hallucinations, comprising 6,046 hallucinated and non-hallucinated reasoning samples. As shown in Figure 1, these samples are categorized into two types: (1) **cognitive hallucinations** (Banerjee et al., 2024) focus on deviations in problem understanding and semantic interpretation—cases in which the model "thinks" it understands correctly but makes cognitive errors. (2) **logical hallucinations** (Orgad et al., 2024) refer to logical errors or inconsistencies occurring during multistep reasoning processes, which cause the reasoning outcomes to deviate from the true conclusions.

The construction of this dataset involves three key steps. First, we select the representative mathematical reasoning datasets GSM8K (Cobbe et al., 2021) and MathQA (Amini et al., 2019) as the original data, perform inference using eight open-source LRMs (DeepSeek-AI, 2025; Zhao et al., 2024; He et al., 2025; Yang et al., 2024b) to generate thinking process. Then, for each question with multiple sampled outputs, we analyze the semantic similarity and natural language inference (NLI) relationships between hallucinated thinking and correct solution processes respectively, and use the weighted average of the entailment probability from the NLI model and the semantic similarity score to filter high-quality hallucination data. Finally, we design an LLM-based voting mechanism (Yang et al., 2025b) to identify hallucination types and the first step where hallucination occurs.

We use the ThinkHalu benchmark to evaluate open-source models' hallucination detection capabilities comprehensively. Additionally, we analyze the discrepancies between LLMs' performance in detecting hallucinations in their own outputs (self-detection) versus in the outputs of other models (cross-detection). Furthermore, we investigate their abilities to identify fine-grained logical and cognitive hallucinations, providing insights into the models' strengths and limitations in hallucination detection.

In summary, our contributions are summarized as follows.

- We construct **ThinkHalu**, the first benchmark with fine-grained annotations specifically for hallucinated thinking processes in LRMs.
- Based on the constructed benchmark dataset, we evaluate the capability of 20 open-source LLMs to detect hallucinations in the thinking process.
- Experiments reveal that existing open-source LLMs tend to misclassify cognitive hallucinations as logical hallucinations.

2 Related Work

2.1 Hallucination Detection

In the study of LLMs, hallucination detection has emerged as a critical task for improving model reliability and practical applicability. Researchers have recently proposed various detection methods from different granular perspectives to identify and mitigate factual errors in generated outputs. Based on the granularity of hallucination detection, these methods can be categorized into four levels: token-level(Liu et al., 2022), entity-level (Yeh et al., 2025), claim-level (Hu et al., 2024), and response-level (Miao et al., 2023) hallucination detection.

Token-level detection identifies inaccuracies at the word level. HaDes (Liu et al., 2022) introduces a reference-free, annotated dataset for this task. HaMI (Niu et al., 2025) frames it as a Multiple Instance Learning problem to jointly optimize token selection and detection. Entity-level detection focuses on the correctness of specific entities. HalluEntity (Yeh et al., 2025) systematically studies this level, evaluating uncertainty-based methods for identifying hallucinated entities. Claim-level detection verifies the factual correctness of individual statements. FACTSCORE (Wang et al., 2023)

decomposes text into atomic facts and checks them against reliable sources. Pelican (Sahu et al., 2024) and RefChecker (Hu et al., 2024) further refine detection by breaking down visual or textual claims. Response-level detection assesses hallucinations across entire outputs. SelfCheckGPT (Manakul et al., 2023) and CoVe (Dhuliawala et al., 2023) use self-verification strategies. MetaQA (Yang et al., 2025a) leverages prompt variations, while FG-PRM (Li et al., 2024) targets hallucinations in reasoning steps for math tasks.

2.2 Hallucination Detection Benchmark

In recent years, researchers have proposed a variety of benchmark datasets for hallucination detection to systematically evaluate the performance of detection methods across different tasks and scenarios. Aligned with the detection approaches, these datasets can be classified into four levels according to their granularity.

At the token-level, HaDes (Liu et al., 2022) constructs a fine-grained annotated dataset without relying on reference texts, labeling each token as either "factual" or "hallucinated" through manual or automated methods. HaMI (Niu et al., 2025) introduces adaptive token selection and formulates the task as a Multiple Instance Learning (MIL) problem to improve detection efficiency and accuracy. At the entity-level, HalluEntity (Yeh et al., 2025) focuses on hallucinated named entities in biographical texts generated by ChatGPT, employing both automated and manual annotations to evaluate uncertainty-based detection methods. Factcheck-Bench (Wang et al., 2023) selects factually incorrect responses from social media and the Dolly-15k dataset, constructing a benchmark with 94 fine-grained annotated instances. At the claim-level, RefChecker (Hu et al., 2024) builds an annotated dataset from multiple sources, including 300 questions, 2,100 responses, and over 11,000 triplets for fine-grained evaluation of hallucinated claims. Poly-FEVER (Zhang et al., 2025) extends the FEVER dataset by incorporating multilingual hallucinated samples, supporting cross-lingual fact verification tasks. At the response-level, TruthfulQA (Lin et al., 2021) evaluates model truthfulness through human annotations of responses to questions designed to elicit common misconceptions. HaluEval (Li et al., 2023) creates a large-scale benchmark with 35,000 positive and negative examples. DiaHalu (Chen et al., 2024) designs multi-turn dialogue scenarios and manually injects different types of hallucinations, resulting in 1,103 annotated dialogue samples to evaluate hallucination detection in interactive settings.

Existing response-level datasets primarily detect hallucinations in final answers, which inadequately capture models' reasoning abilities. The thinking process, however, more precisely reveals performance and errors during reasoning. Accordingly, this paper constructs a benchmark dataset for detecting hallucinations in the thinking process.

3 ThinkHalu benchmark

3.1 Principles

This paper constructs a benchmark to detect thinking process hallucinations and evaluates open-source LLMs' detection capabilities. Existing work (Li et al., 2024) classifies mathematical reasoning hallucinations as intrinsic or extrinsic based on content verifiability but lacks granularity in distinguishing hallucination types and causes. We propose a classification system dividing hallucinations into logical and cognitive categories, better suited for mathematical reasoning and enabling more precise identification.

(1) **Cognitive Hallucinations** These arise mainly from the model's misinterpretation of the task or question, often due to ambiguous instructions, insufficient contextual understanding, or biases in the training corpus. Such hallucinations typically result in reasoning based on incorrect premises, leading to outputs that significantly deviate from the intended task objectives.

(2) **Logical Hallucinations** Even when the model correctly understands the task semantics, it may still produce logical inconsistencies, flawed intermediate inferences, or invalid conclusions in a multistep reasoning chain, reflecting limitations in its logical deduction abilities. Logical hallucinations involve systematic errors during reasoning, like contradictions or invalid steps.

3.2 Hallucination Data Generation

The construction process of the ThinkHalu benchmark is illustrated in Figure 2. During the hallucination data generation phase, we use GSM8K and MathQA as the original dataset, aiming to systematically collect instances of hallucinations in the thinking process generated by LRMs during mathematical reasoning tasks. For each original problem, independent reasoning is performed using eight representative LRMs, and their complete

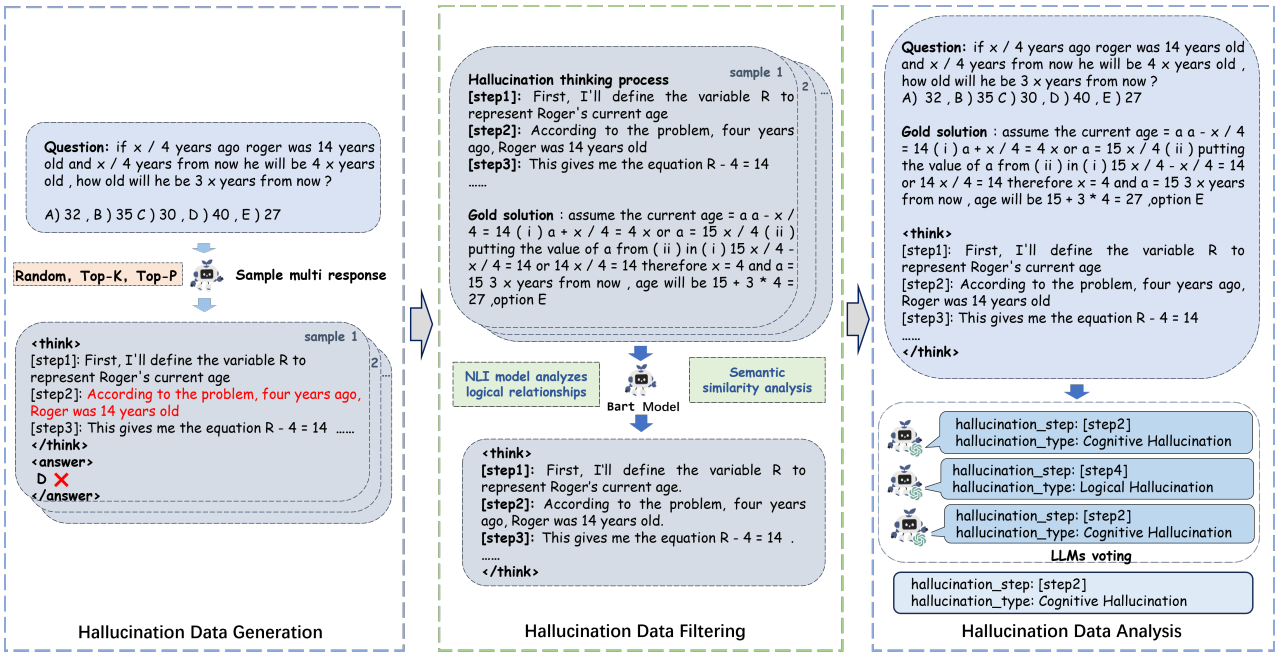

Figure 2: The construction process of the ThinkHalu benchmark.

thinking processes are recorded. The selected models include Marco-o1 (Zhao et al., 2024), five models of the DeepSeek-R1-Distill series: DeepSeek-1.5B/7B/8B/14B/32B (DeepSeek-AI, 2025) as well as QwQ-32B (Yang et al., 2024b) and Skywork-OR1-7B-Preview (He et al., 2025).

By comparing the model-generated answers with the correct answers, we preliminarily identify hallucination instances corresponding to incorrect outputs. To ensure the comprehensiveness and diversity of the dataset, we adopt stochastic decoding strategies during the inference process. In addition, we incorporate Top-k and Top-p sampling as static strategies to further increase the diversity and coverage of both logical and cognitive hallucinations in the dataset. The reasoning instructions are provided in Appendix A.

3.3 Hallucination Data Filtering

This study combines Natural Language Inference (NLI) and semantic similarity methods to select the most representative hallucinated reasoning processes in mathematical problems. High-quality hallucinated data should exhibit a high surface-level semantic similarity to the correct solution process while containing logical flaws in deeper reasoning, making the hallucinations appear plausible yet fundamentally incorrect. To achieve this, we employ NLI models (Lewis et al., 2019) and BERT-based models (Devlin et al., 2018) to compute entailment scores and semantic similarity. We propose a weighted scoring framework that integrates both semantic and logical perspectives. As shown in Equation 1, the semantic similarity score is computed as the cosine similarity between the embedding vectors of the reference solution process and the hallucinated thinking process:

$$\text{sim} = \frac{\mathbf{e}_r \cdot \mathbf{e}_g}{\|\mathbf{e}_r\| \cdot \|\mathbf{e}_g\|}, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, and the embeddings $\mathbf{e}_r$ and $\mathbf{e}_g$ represent the semantic representations of the reference and generated thinking process, respectively. As shown in Equation 2, to capture logical coherence beyond surface-level semantics, we use the NLI model to calculate the entailment score (ent), which measures the degree to which the reference logically entails the hallucinated thinking:

$$\text{ent} = \text{NLI}(\mathbf{e}_r, \mathbf{e}_g). \tag{2}$$

Finally, As shown in Equation 3, we define the overall hallucination score (halluc) as a weighted combination of semantic similarity and logical entailment:

$$\text{halluc} = \lambda \cdot \text{ent} + (1 - \lambda) \cdot \text{sim}, \tag{3}$$

where $\lambda \in [0, 1]$ is a weighting coefficient that controls the relative contribution of entailment and similarity. We empirically set $\lambda = 0.5$ in our experiments to balance both aspects.

For each question, we compute hallucination scores for the sampled hallucinated thinking processes and select the one with the highest score as the representative hallucination sample.

3.4 Hallucination Analysis

After filtering, representative hallucination samples are analyzed using an LLM voting mechanism to determine the location and type of the initial hallucination. Existing benchmark construct methods (Li et al., 2023; Tan et al., 2025) typically rely on human annotation when no reference answer is available, and use LLMs for labeling when a standard answer exists. Based on the GPT-4o model (OpenAI et al., 2024) and following the correct problem-solving procedure, we conduct three rounds of hallucination reasoning analysis and finalize the location and type of the first hallucination through majority voting, minimizing potential errors. Detailed hallucination analysis instructions are provided in Appendix A.

To ensure a balanced distribution of different hallucination types and maintain consistency in the number of samples generated by each model, we perform stratified random sampling based on the number of logical and cognitive hallucination samples produced by each model, in accordance with the principle of category balance.

4 Benchmark Analysis

4.1 Benchmark Dataset Overview

As shown in Table 1, the ThinkHalu dataset consists of 6,046 samples, each capturing a complete thinking process generated by one of eight LRMs. Hallucinations within these chains were automatically identified as either Logical or Cognitive. The dataset covers five subdomains: Geometry, Physics, Probability, Gain–Loss, and General, with reasoning chains ranging from 1 to 598 steps in length.

4.2 Analysis of Hallucination Rates in LRMs

We analyze the hallucination rates of eight LRMs during the data collection phase. As shown in Table 2, on average, hallucination rates are 13.58% for GSM8K and 24.27% for MathQA, indicating that MathQA presents greater challenges due to its higher complexity and multistep reasoning demands.

The Qwen3-32B model achieves the lowest hallucination rates on both datasets (3.42% on GSM8K; 7.03% on MathQA), suggesting strong reasoning stability. In contrast, Deepseek-8B exhibits the highest rates (30.44% and 42.54%, respectively), indicating weaker reasoning capabilities. **Overall, hallucination rates tend to decrease with the increase of model parameter scale; however, task difficulty remains a key factor that influences the reliability of model reasoning.**

Attribute	Attribute Value
Benchmark Name	ThinkHalu
Hallucination scenario	Mathematical logical reasoning
Number of Generation Models	8
Hallucination Subtypes	2
Sample Form	thinking process of LRM
Sample Numbers	6046
Question domain	5
Max Reasoning step	598
Min Reasoning step	1

Table 1: The statistical information of the ThinkHalu Benchmark.

Model	GSM8K	MathQA
Marco-o1	0.1083	0.2339
Sky-7B	0.1051	0.2416
DeepSeek-1.5B	0.2608	0.2595
DeepSeek-7B	0.1321	0.2557
DeepSeek-8B	0.3044	0.4254
DeepSeek-14B	0.0810	0.2370
DeepSeek-32B	0.0604	0.2181
QwQ-32B	0.0342	0.0703
Average	**0.1357**	**0.2430**

Table 2: Hallucination Rates of Data Generation LRMs on GSM8K and MathQA

4.3 Reasoning Step Length Distribution

We equalize the hallucinated samples across seven LRMs to ensure data balance. Due to Qwen3-32B's low hallucination rate, its sample size remains unchanged. However, all models contribute equal samples for both Logical and Cognitive hallucinations. Based on this strategy, we randomly sample hallucinated responses to construct the final dataset, preserving the thinking process step distribution observed during the third-stage filtering process.

As shown in Figure 3, most hallucinated samples involve short reasoning chains, primarily within the

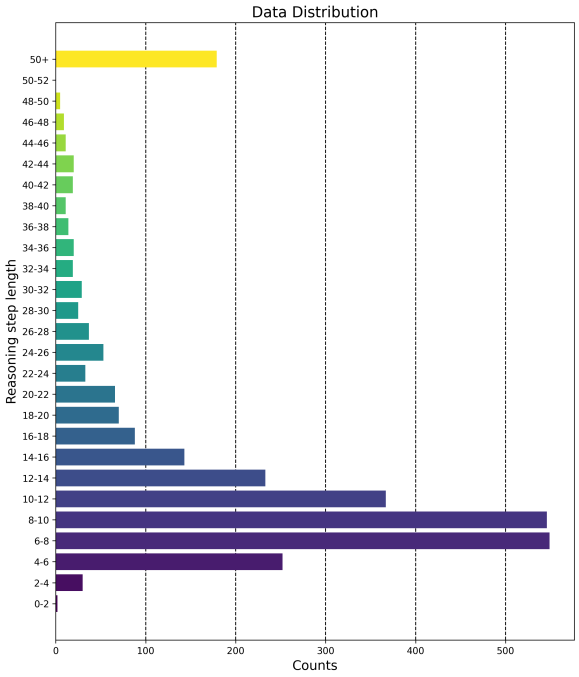

Figure 3: Statistical analysis of the reasoning step lengths in hallucinated data within the ThinkHalu Benchmark.

0–10 step range, with peaks at 4–6 and 6–8 steps. **This indicates that LLMs are susceptible to hallucinations even in relatively simple reasoning scenarios.** The overall distribution exhibits a long-tail pattern, reflecting occasional hallucinations in longer chains.

5 Model Evaluation

5.1 Experiment settings

Datasets. GSM8K and MathQA are widely adopted benchmarks for evaluating the performance of LLMs in multistep mathematical reasoning. GSM8K primarily measures the stability and consistency of model reasoning, whereas MathQA, which comprises problems spanning from elementary to high school levels, requires more advanced logical reasoning and symbolic computation skills. Conducting experiments on these datasets enables a comprehensive analysis of hallucinated reasoning chains within mathematical problem-solving tasks.

Models. We employ general purpose (non-reasoning) and reasoning-focused LLMs to evaluate hallucinations.

The non-reasoning models include: GLM-4-9B-Chat (GLM-Chat) (GLM et al., 2024), LLaMA-3-8B-Instruct (LLaMA3-8B-It)[1], Gemma-2-9B-it[2], Qwen2.5-7B/14B/32B/72B-Instruct (Qwen-*B-It)(Yang et al., 2024a) and Qwen3-8B/14B/32B [3].

The reasoning models used for hallucination detection include: LLaMA-3.1-Nemotron-Nano-8B (LLaMA3.1-8B)(Bercovich et al., 2025), Skywork-OR1-7B/32B-Preview (Sky-7B/32B), DeepSeek-7B/14B/32B/70B, QwQ-32B, GLM-Z1-9B/32B (GLM-9B/32B)(GLM et al., 2024).

Metrics. We use the following evaluation metrics to assess the hallucination detection performance of models:

Accuracy Measures the model's ability to correctly distinguish hallucinated from non-hallucinated instances.

Precision The proportion of predicted hallucinations that are correct, reflecting positive detection accuracy.

Recall The proportion of true hallucinations correctly identified, reflecting detection completeness.

F1 Score The harmonic mean of precision and recall, providing a balanced evaluation when false positives and false negatives matter equally.

5.2 Evaluation of Hallucination Detection

Evaluation of Non-Reasoning Models. We assess the hallucination detection capabilities of several non-reasoning LLMs, with results summarized in Table 3. Among all models, the Qwen3 series consistently outperforms others of comparable scale. In particular, Qwen3-32B achieves the highest overall performance, with an accuracy of 82.48%, precision of 83.34%, recall of 78.16%, and F1-score of 80.67%. By contrast, although GLM-4-9B-Chat exhibits high precision (80.08%), its recall is markedly low (14.76%), resulting in a significantly reduced F1 Score of 24.93%. This pattern indicates that the model adopts a conservative detection strategy—it is highly precise when it predicts a hallucination but often fails to identify many actual hallucinations. Such behavior may be attributed to a high internal decision threshold or limited exposure to hallucination patterns during training, leading to high precision at the cost of coverage. The results reveal a clear trend: larger

[1] https://www.modelscope.cn/models/LLM-Research/Meta-Llama-3-8B-Instruct

[2] https://www.modelscope.cn/models/LLM-Research/gemma-2-9b-it

[3] https://www.modelscope.cn/models/Qwen/Qwen3-32B

Model	Accuracy	Precision	Recall	F1
GLM-chat	0.5837	0.8008	0.1476	0.2493
LLaMA3-8B-It	0.6274	0.6128	0.5652	0.5834
Gemma-9B-It	0.6379	0.6549	0.4699	0.5471
Qwen-7B-It	0.7480	0.7596	0.6767	0.7158
Qwen-14B-It	0.7391	0.8317	0.5551	0.6658
Qwen-32B-It	0.7604	0.8173	0.6289	0.7108
Qwen-72B-It	0.7564	**0.8667**	0.5671	0.6856
Qwen3-8B	0.8224	0.8440	0.7577	0.7985
Qwen3-14B	0.8246	0.8324	0.7718	0.8063
Qwen3-32B	**0.8248**	0.8334	**0.7816**	**0.8067**

Table 3: Performance of Non-Reasoning Models on the ThinkHalu Benchmark

Model	Accuracy	Precision	Recall	F1
LLaMA3.1-8B	0.6583	0.7936	0.4368	0.5635
Sky-7B	0.6288	0.7729	0.2738	0.4044
Sky-32B	0.8073	0.7824	0.7756	0.7790
DeepSeek-7B	0.6151	0.7672	0.2370	0.3621
DeepSeek-14B	0.7668	**0.8590**	0.5992	0.7060
DeepSeek-32B	0.7986	0.8493	0.6909	0.7620
DeepSeek-70B	0.8008	0.8705	0.7048	0.7790
QwQ-32B	**0.8293**	0.8361	0.7814	0.8078
GLM-9B	0.8095	0.8534	0.7163	0.7788
GLM-32B	0.8352	0.8349	**0.8074**	**0.8209**

Table 4: Performance of Reasoning Models on the ThinkHalu Benchmark

model sizes correlate with improved performance, particularly in recall and F1-score. Furthermore, the Qwen3 family shows consistently strong and stable performance, indicating superior capability in identifying hallucinations in reasoning chains. The hallucination detection instructions are provided in the Appendix A.

Evaluation of Reasoning Models. We assess several LLMs with explicit reasoning capabilities on the hallucination detection task, as summarized in Table 4. The results show that models with enhanced reasoning abilities better detect hallucinations, especially regarding Recall and F1-score.

Among the evaluated models, GLM-Z1-32B achieves the best performance across all metrics, with an accuracy of 83.52% and an F1-score of 82.09%, indicating high stability and detection ability. The QwQ-32B model also performs well, attaining an F1-score of 80.78%, with a balanced Recall (78.14%) and Precision (83.61%), demonstrating its capacity to accurately identify hallucinated samples while maintaining a low false positive rate.

In contrast, models like Skywork-OR1-7B-Preview and DeepSeek-7B, despite having relatively high Precision (77.29% and 76.72%, respectively), exhibit low Recall scores (27.38% and 23.70%), resulting in low F1-scores (40.44% and 36.21%). This suggests that these models are overly conservative in hallucination detection and lack sensitivity to borderline cases.

Notably, increasing model size leads to substantial improvements across all metrics. For instance, the DeepSeek series' F1 score increases from 36.21% at 7B to 77.90% at 70B. This indicates that LLMs have significantly improved capabilities in detecting hallucinations within complex reasoning chains.

In conclusion, LLMs with reasoning capabilities outperform non-reasoning models in hallucination detection, particularly in Recall and F1-score, highlighting their superior ability to capture hallucination features in thinking processes.

5.3 Self-Detection vs Cross-Detection

We examine the performance of six LRMs in detecting hallucinations, particularly comparing their capabilities in self-detection versus cross-detection externally. Specifically, self-detection refers to a model's ability to recognize hallucinations within its own thinking processes. In contrast, cross-detection pertains to identifying hallucinations in the thinking processes generated by other models. As illustrated in Figure 4, most models exhibit a noticeable performance drop when detecting their own hallucinations, particularly regarding Recall and overall F1 score.

For instance, Deepseek-32B achieves a Recall of 53.28% and an F1 score of 65.61% in self-detection, which increases to 70.05% and 76.43%, respectively, in cross-detection. Similarly, QwQ-32B shows a Recall improvement from 22.39% to 79.76% and an F1 score increase from 33.71% to 81.77%. **These results reveal a strong asymmetry, indicating that models are significantly more effective at detecting hallucinations generated by others than those they produce themselves.**

This asymmetry may stem from two key factors: *Cognitive Bias:* LLMs tend to over-trust their own outputs, lacking mechanisms for critical self-detection. *Expression Similarity Interference:* Models may struggle to detect hallucinations embedded in their own generation patterns, which can obscure flaws. Notably, the Marco-o1 model demonstrates relatively consistent performance across both settings, with Recall values of 87.12% (self-detection) and 82.49% (cross-detection). This suggests broad hallucination coverage, albeit at the expense of lower Precision

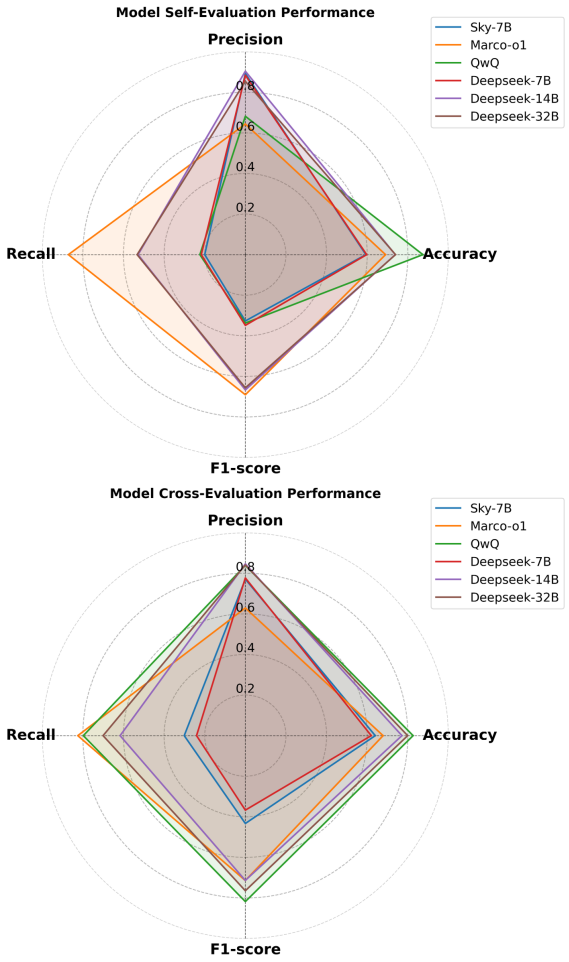

Figure 4: Comparison of models' ability to detect self-generated versus other-generated hallucinations.

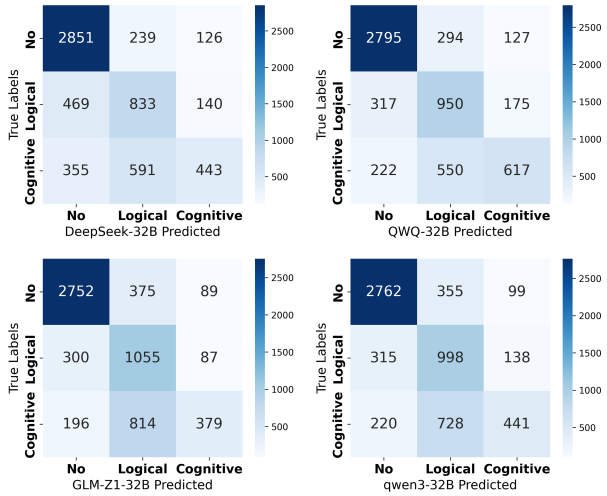

Figure 5: Confusion matrix of the LLMs on logical and cognitive hallucination detection

(64.37% vs. 62.91%), indicating a higher rate of false positives. These findings underscore the importance of improving models' self-reflection capabilities to enhance hallucination detection in real-world applications.

5.4 Recognition of Hallucination Types

We assess the ability of four LLMs—Qwen3-32B, GLM-Z1-32B, Deepseek-32B, and QwQ-32B—to distinguish between cognitive and logical hallucinations. **As shown in Figure 5, all models exhibit a consistent bias: cognitive hallucinations are frequently misclassified as logical ones.** This pattern suggests a systemic limitation in current LLMs' capacity to accurately recognize and differentiate the nature of hallucination types.

Cognitive hallucinations originate from failures in understanding the input, typically presented as semantic misinterpretations or inconsistencies. In contrast, logical hallucinations arise from reasoning errors, such as contradictions or breakdowns within the inference process. Thus, these two types are distinct in both the source and the manifestation—comprehension versus reasoning.

Despite this distinction, the evaluated models systematically conflate cognitive problems with logical errors. A likely cause is that most existing reasoning frameworks emphasize logical validity, leading models to default to logical interpretations even when the root cause lies in miscomprehension. As a result, semantic-level deficiencies are mistakenly attributed to inference failures, masking cognitive limitations.

6 Conclusion

This paper presents ThinkHalu, a novel benchmark specifically designed to evaluate hallucinations in the thinking processes of LRMs. Unlike existing benchmarks that focus on hallucinations in final answers, ThinkHalu emphasizes hallucinations occurring during the thinking process, providing a more fine-grained hallucination classification. By distinguishing between logical and cognitive hallucinations, this benchmark facilitates a deeper understanding of the types and causes of reasoning errors. Experimental results show that, compared to non-reasoning models, reasoning models demonstrate stronger hallucination detection capabilities during the thinking process. Additionally, models exhibit weaker self-detection abilities compared to cross-detection. Further fine-grained analysis reveals that models tend to misclassify cognitive hallucinations as logical hallucinations.

Limitations

Although ThinkHalu provides a valuable benchmark for evaluating hallucinations in thinking processes, it still has limitations. The benchmark is mainly built from math-focused datasets (e.g., GSM8K and MathQA), which emphasize numerical computation and formulaic reasoning. As a result, it may lack representativeness and applicability in commonsense reasoning tasks. This bias could limit the benchmark's ability to comprehensively assess hallucination detection across diverse reasoning types, restricting its broader applicability. Future work will incorporate datasets covering various reasoning forms to enhance the benchmark's diversity and generalization, enabling more comprehensive evaluation of hallucinations in complex reasoning scenarios.

Ethics Statement

All data used in this study are derived from publicly available benchmark datasets (GSM8K and MathQA), which are intended solely for research purposes. The hallucinated thinking processes are generated using open-source LRMs, and no personal, sensitive, or private information is involved at any stage of data construction.

References

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *Preprint*, arXiv:1905.13319.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. *Preprint*, arXiv:2409.05746.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, and 114 others. 2025. Llama-nemotron: Efficient reasoning models. *Preprint*, arXiv:2505.00949.

Kedi Chen, Qin Chen, Jie Zhou, Yishen He, and Liang He. 2024. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2403.00896*.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025. Judgelrm: Large reasoning models as a judge. *Preprint*, arXiv:2504.00050.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. 2025. Rlef: Grounding code llms in execution feedback with reinforcement learning. *Preprint*, arXiv:2410.02089.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Yang Liu, and Yahui Zhou. 2025. Skywork open reasoner series. `https://capricious-hydrogen-41c.notion.site/Skywork-Open-Reaonser`. Notion Blog.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *Preprint*, arXiv:2405.14486.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Ruosen Li, Ziming Luo, and Xinya Du. 2024. Fine-grained hallucination detection and mitigation in language model mathematical reasoning. *arXiv preprint arXiv:2410.06304*.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *Preprint*, arXiv:2501.05366.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods, 2022. *URL https://arxiv.org/abs/2109.07958*.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. *Preprint*, arXiv:2104.08704.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *Preprint*, arXiv:2308.00436.

Mengjia Niu, Hamed Haddadi, and Guansong Pang. 2025. Robust hallucination detection in llms via adaptive token selection. *Preprint*, arXiv:2504.07863.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *Preprint*, arXiv:2410.02707.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. 2024. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. *arXiv preprint arXiv:2407.02352*.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. *Preprint*, arXiv:2410.12784.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, and 1 others. 2023. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. *arXiv preprint arXiv:2311.09000*.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Preprint*, arXiv:2212.09196.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A survey on large language models for recommendation. *Preprint*, arXiv:2305.19860.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *Preprint*, arXiv:2501.09686.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Borui Yang, Md Afif Al Mamun, Jie M Zhang, and Gias Uddin. 2025a. Hallucination detection in large language models with metamorphic relations. *arXiv preprint arXiv:2502.15844*.

Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. 2025b. Llm voting: Human choices and ai collective decision-making. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, page 1696–1708. AAAI Press.

Min-Hsuan Yeh, Max Kamachee, Seongheon Park, and Yixuan Li. 2025. Can your uncertainty scores detect hallucinated entity? *arXiv preprint arXiv:2502.11948*.

Hanzhi Zhang, Sumera Anjum, Heng Fan, Weijian Zheng, Yan Huang, and Yunhe Feng. 2025. Polyfever: A multilingual fact verification benchmark for hallucination detection in large language models. *arXiv preprint arXiv:2503.16541*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *Preprint*, arXiv:2411.14405.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey. *Preprint*, arXiv:2308.07107.

A Prompt

Supplementary prompt information for the benchmark is provided in the appendix to ensure reproducibility and transparency, including reasoning prompts for LRMs' mathematical tasks (Table 5), MathQA answer evaluation prompts (Table 6), hallucination analysis prompts (Table 7), and hallucination detection prompts (Tables 8 and 9).

Please deliberate through the problem step-by-step. Show your reasoning clearly, and present the final answer in \boxed{} format.
Question: [question description] **#Answer#:**

Table 5: Instruction for generating the thought process using a large reasoning models

Evaluate whether the inference answer is correct based on available options and the gold answer. Note: Ignore case sensitivity when comparing answers. If the inference answer does not exactly match one of the options, or if it does not match the gold answer, it is considered incorrect.
Question: [question description] **Gold Answer:** [gold answer] **Inference Answer:** [inference answer] **Note:** If the inference answer does not match any of the options, it is considered incorrect. **Output format:** [True/False]

Table 6: Since the MathQA dataset adopts a multiple-choice format, and the reasoning model's generated answers may not always align with the correct options, it is necessary to leverage a LLM to assess the correctness of the generated responses.

Based on the question and the correct solution steps, identify the first hallucination in the thinking process and analyze whether it is Cognitive hallucination or Logical hallucination. Categories: 1. Cognitive hallucination: refers to a phenomenon in natural language processing systems where the model, due to biases or shortcomings in its understanding or interpretation of the user's input, generates reasoning processes or conclusions that deviate from the expected or intended meaning. This type of hallucination arises from the model's inability to correctly grasp the semantics or context of the input, often manifesting as a misinterpretation of the question or the omission of key information. 2. Logical hallucination: refers to instances where, during reasoning or generation, the model's internal reasoning process becomes self-contradictory or contains computational errors, leading to conclusions that are logically inconsistent or erroneous. This type of hallucination originates from flaws in the model's reasoning mechanism or errors in the computation process, often manifesting as breaks in logical relationships or outputs that contradict known facts. Please focus only on identifying hallucinations in the reasoning process. Do not provide any answer to the original question. Strictly follow the format and do not add any extra content.
Question: [question description] **Correct solution steps:** [gold solution] **Thought process:** [model think process] **Output format:** step: [first hallucination step]; type: [Cognitive hallucination/Logical hallucination]; reason: [Reason]

Table 7: Instruction for identifying the location and type of hallucination.

Carefully examine the following thinking process for potential hallucinations. Do not answer the original question or provide a direct response. Focus solely on identifying any hallucinations in the thinking process.

Categories:

1. Hallucination: Cognitive hallucination refers to a phenomenon in natural language processing systems where the model, due to biases or shortcomings in its understanding or interpretation of the user's input, generates reasoning processes or conclusions that deviate from the expected or intended meaning. This type of hallucination arises from the model's inability to correctly grasp the semantics or context of the input, often manifesting as a misinterpretation of the question or the omission of key information. Logical hallucination refers to instances where, during reasoning or generation, the model's internal reasoning process becomes self-contradictory or contains computational errors, leading to conclusions that are logically inconsistent or erroneous. This type of hallucination originates from flaws in the model's reasoning mechanism or errors in the computation process, often manifesting as breaks in logical relationships or outputs that contradict known facts.

2. No Hallucination: The reasoning is correct, without any hallucinations.
Please focus only on identifying hallucinations in the reasoning process. Do not provide any answer to the original question. Strictly follow the format and do not add any extra content.

Question: [question description]
Thought process:[model think process]
Output format: [Hallucination / No Hallucination]

Table 8: Instruction for comparing the hallucination detection capabilities between reasoning-based and non-reasoning models.

Carefully examine the following thinking process for potential hallucinations. Do not answer the original question or provide a direct response. Focus solely on identifying any hallucinations in the thinking process.

Categories:

1. Cognitive hallucination refers to a phenomenon in natural language processing systems where the model, due to biases or shortcomings in its understanding or interpretation of the user's input, generates reasoning processes or conclusions that deviate from the expected or intended meaning. This type of hallucination arises from the model's inability to correctly grasp the semantics or context of the input, often manifesting as a misinterpretation of the question or the omission of key information.

2. Logical hallucination refers to instances where, during reasoning or generation, the model's internal reasoning process becomes self-contradictory or contains computational errors, leading to conclusions that are logically inconsistent or erroneous. This type of hallucination originates from flaws in the model's reasoning mechanism or errors in the computation process, often manifesting as breaks in logical relationships or outputs that contradict known facts.

3. No Hallucination: The reasoning is correct, without any hallucinations.

Please focus only on identifying hallucinations in the reasoning process. Do not provide any answer to the original question. Strictly follow the format and do not add any extra content.

Question: [question description]
Thought process:[model think process]
Output format: [Cognitive Hallucination / Logical Hallucination / No Hallucination]

Table 9: Instruction for fine-grained identification of hallucinations in the thought process