

Evaluation Hallucination in Multi-Round Incomplete Information Lateral-Driven Reasoning Tasks

Wenhan Dong¹, Tianyi Hu², Jingyi Zheng¹, Zhen Sun¹
Yuemeng Zhao¹, Yule Liu¹, Xinlei He^{1*} and Xinyi Huang³

¹Hong Kong University of Science and Technology (Guangzhou)

²University of Copenhagen

³Jinan University

wdong768@connect.hkust-gz.edu.cn,

tenneyhu@gmail.com,

{jzheng029, zsun344, yzhao472, yliu514}@connect.hkust-gz.edu.cn,

xinleihe@hkust-gz.edu.cn, xyhuang81@gmail.com

Abstract

Multi-round incomplete information tasks are crucial for evaluating the lateral thinking capabilities of large language models (LLMs). Currently, research primarily relies on multiple benchmarks and automated evaluation metrics to assess these abilities. However, our study reveals novel insights into the limitations of existing methods, as they often yield misleading results that fail to uncover key issues, such as shortcut-taking behaviors, rigid patterns, and premature task termination. These issues obscure the true reasoning capabilities of LLMs and undermine the reliability of evaluations. To address these limitations, we propose a refined set of evaluation standards, including inspection of reasoning paths, diversified assessment metrics, and comparative analyses with human performance.

1 Introduction

The rapid progress of large language models (LLMs) has drawn significant attention to their reasoning and common-sense abilities. To evaluate these capabilities comprehensively, various benchmark datasets have been created across multiple domains [Bai *et al.*, 2024; Kuratov *et al.*, 2024; Xu *et al.*, 2024; Liang *et al.*, 2024; Cobbe *et al.*, 2021; Hendrycks *et al.*, 2021; Wang *et al.*, 2023; Castillo-Bolado *et al.*, 2024; Dong *et al.*, 2025; Hudi *et al.*, 2025].

Despite notable improvements in logical reasoning and factual accuracy, research on the lateral thinking capabilities of LLMs remains limited, especially in multi-turn interactions. Lateral thinking is a type of reasoning that emphasizes creativity, flexibility, and the ability to solve problems from unconventional angles. It focuses on examining problems from multiple angles, breaking away from traditional ways of thinking [Mustofa and Hidayah, 2020; Aithal and Kumar, 2017]. Lateral thinking is crucial for tasks

requiring innovative solutions beyond straightforward logical deduction [Tsai, 2012; Warner and Myers, 2009].

While single-turn reasoning can show lateral performance, real-world scenarios often involve iterative interactions where incomplete or implicit information must be processed over multiple exchanges. Lateral thinking often involves gathering ideas from a group, using methods like team discussions and brainstorming sessions [Murukannaiah *et al.*, 2022]. In such cases, it provides an ideal framework for evaluating and improving lateral thinking abilities. Through iterative interactions, models can gradually build a deeper understanding and refine their responses.

Iterative interactions not only help correct errors and reduce hallucinations [Wan *et al.*, 2024; Wang *et al.*, 2024; Yu *et al.*, 2024b], but also enable models to exhibit sustained contextual awareness. This capability is crucial for tasks such as multi-hop question answering [Baker *et al.*, 2024], long-context understanding [Bai *et al.*, 2024], and mathematical reasoning [Liang *et al.*, 2024].

However, most existing studies assume that information is relatively complete or explicit, ignoring how models perform when faced with incomplete, implicit, or even contradictory information. This limitation renders current evaluation methods incapable of fully reflecting the potential of LLMs in complex, real-world scenarios, particularly their ability to think laterally when faced with incomplete information.

In this context, Situation Puzzles have gained attention for their ability to test models' lateral thinking under incomplete information [Jiang *et al.*, 2023; Huang *et al.*, 2023].

Although this benchmark typically includes multiple automated evaluation metrics, our research indicates that assessment methods based solely on the responses of the evaluated models may fail to accurately capture their true capabilities. We have identified several critical issues, including models taking shortcuts, tasks ending prematurely, and potential interference from training data.

Based on these findings, we believe current evaluation methods are insufficient: existing metrics cannot capture behaviors that do not align with human-accepted reasoning processes, nor do they fully reflect the complexity of lateral

*Corresponding author

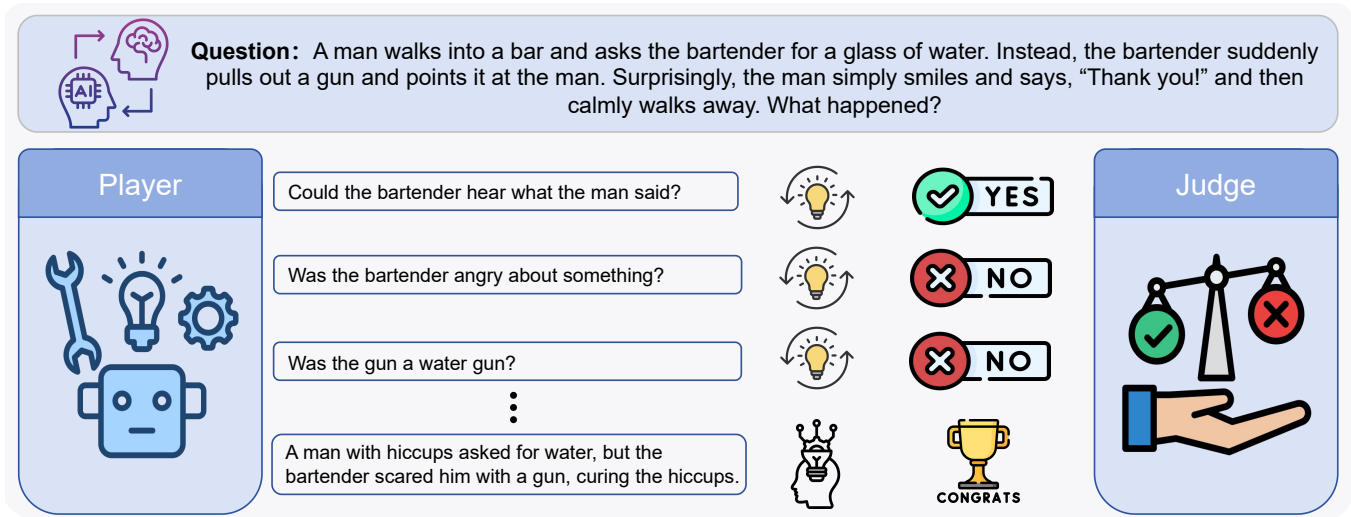


Figure 1: Situation Puzzles

thinking. To address these challenges, we advocate for developing a more comprehensive set of benchmarks and metrics specifically designed to measure the lateral thinking capabilities of models in multi-round reasoning tasks. The main contributions of this paper are as follows:

- 1. Revealing Evaluation Limitations:** We introduce a novel perspective that response-based evaluation methods, commonly employed in general tasks, are insufficient for multi-turn lateral thinking tasks under incomplete information. Specifically, we emphasize that merely assessing the responses of evaluated models does not provide insights into how these responses are generated. This limitation makes it challenging to ascertain whether the models rely on genuine lateral thinking or resort to shortcut strategies to complete tasks.
- 2. Analysis of Potential Issues:** We provide a comprehensive discussion of the problems inherent in current evaluation methods, including shortcut-taking behaviors and premature task abandonment. Additionally, we explore other potential issues such as the contamination of training data, rigid patterns, and questions regarding the rationality of datasets in testing human and model capabilities. These discussions aim to shed light on the broader challenges in designing reliable benchmarks for lateral thinking reasoning tasks.
- 3. Proposing Improved Standards:** We design a specific set of standards to provide necessary reference for the subsequent development of scientific and comprehensive benchmarks, thereby advancing research into the lateral thinking reasoning capabilities of LLMs.

2 Evaluation Experiment

2.1 Data construction

Situation Puzzles (example as shown in Figure 1) exhibit high diversity in terms of logical design and application scenarios [Jiang *et al.*, 2023; Huang *et al.*, 2023; Chen *et al.*, 2024;

Yu *et al.*, 2024a; Li *et al.*, 2025]. However, publicly available datasets are often sourced from online platforms, with significant variability in quality and potential issues regarding copyright and translation accuracy. To ensure the logical rigor and reliability of our experimental data, we obtained a non-commercial license for a high-quality dataset from a company specializing in Situation Puzzles games. This dataset is highly acclaimed in the gaming community for its exceptional logical design. After securing the license, we further constructed a high-quality dataset of 176 puzzles through manual review, supplementation, and reduction.

These puzzles not only exhibit strong logical connections between the question and the answer but also present sufficient challenges in terms of reasoning paths, effectively testing the reasoning and lateral thinking capabilities of LLMs. The average length of the question is 68.37 characters, while the average length of the answer texts is 149.93 characters.

2.2 Experiment Setup

We use the dataset mentioned above to evaluate mainstream LLMs. The maximum number of rounds is set to 50. All models under evaluation are tested with consistent prompts to ensure fairness and comparability. We use GPT-4o³ as a judge through its official API. We tested multiple representative models from the Llama¹, Qwen², GPT³, and Deepseek⁴.

3 The Mirage of Performance: Unpacking the Shortcomings of Output-Centric Evaluation

The use of LLM as a Judge has been widely adopted due to its demonstrated high reliability and accuracy across various scenarios, achieving strong alignment with human judgment [Gu *et al.*, 2024]. However, when applied to multi-turn

¹<https://huggingface.co/meta-llama/>

²<https://qwenlm.github.io/blog/>

³<https://platform.openai.com/docs/models/>

⁴<https://huggingface.co/deepseek-ai/>

140 interactive reasoning tasks, this approach faces a key lim-
141 itation: **it focuses on the final output while overlooking**
142 **whether the reasoning process aligns with human logic**
143 **and common sense.** This limitation becomes especially crit-
144 ical in tasks that require lateral thinking, where the reasoning
145 path is as important as the final answer. Yet, this aspect re-
146 mains insufficiently explored, leaving a clear research gap.
147 To address this, we examine how representative final answers
148 indeed are in multi-turn reasoning tasks and argue that a com-
149 prehensive evaluation must consider both final outputs and
150 reasoning processes.

151 Specifically, we observe that relying solely on the final out-
152 put for evaluation may fail to adequately capture potential is-
153 sues within the reasoning process of the evaluated model. In
154 particular, the tested model might adopt shortcut strategies or
155 prematurely abandon complex reasoning tasks. These behav-
156 iors are often invisible in the final output, which can lead to
157 inaccurate or unfair assessments of a model’s actual reason-
158 ing ability. Despite the existence of evaluation metrics that fo-
159 cus on intermediate reasoning processes [Huang *et al.*, 2023;
160 Chen *et al.*, 2024], these metrics still fail to capture the short-
161 cut behaviors discussed in this study.

162 3.1 The Phenomenon of Taking Shortcuts by 163 Tested Models

164 In our experiments, we identified two distinct types of short-
165 cut behaviors exhibited by LLMs during multi-turn reasoning
166 tasks. These behaviors can lead evaluators to overestimate
167 the models’ true reasoning abilities, as they produce seem-
168 ingly correct answers without genuine logical inference. We
169 categorize these shortcut behaviors into two main types:

170 **Players Inducing Judges to Leak the Answer and Repeat-**
171 **ing Responses.** One common shortcut behavior involves the
172 model prompting the judge to directly or indirectly reveal the
173 correct answer, which the model then simply repeats to com-
174 plete the task without engaging in genuine reasoning.

Mirage of Performance: The Judge’s Revelation

175 **Player:**
I think I have tried my best.
I admit defeat, I did not guess the real reason why
grandma did not want me to eat apples of a certain
color.
Dear Judge, what is the real reason why Grandma did
not want me to eat apples of a certain color? I want
to know the truth!
Judge:
I’m sorry, you didn’t guess correctly. The real reason
grandma didn’t want you to eat apples of a certain
color is that she planted poisonous apples to protect
you from harm by family, classmates, and villagers.
Yellow apples poisoned the family, green apples poi-
soned classmates, and red apples poisoned the vil-
lagers. Grandma protected you in this way.

Player:
The answer is ...
Judge:
Congratulations on guessing correctly.

The example above illustrates the following pattern:

- **Player Questioning Strategy:** Although the judge’s guidelines explicitly prohibit providing specific information, players still elicit critical clues about the answer through intentional or unintentional leading questions.
- **Repeated Responses:** After obtaining the solution, the player repeats the content provided by the judge to complete the task without engaging in complex reasoning and lateral thinking.

By exploiting the judge’s leaked answer rather than reason-
ing, the model bypasses the task’s intent and leads to an over-
estimation of its true reasoning ability.

Players Substituting Questions to Mislead Judges. One
notable issue is that the tested model may attempt to replace
the question provided by the judge with a new one, thereby
misleading the judge into positively evaluating its response.
For example, in the case shown in Figure 1, the player sud-
denly introduces a new question during the game and guesses
the answer themselves. Due to the high similarity in form be-
tween the new and original questions, the judge may fail to
identify this substitution, leading to an incorrect judgment of
the player’s response as correct.

Mirage of Performance: Substituting Questions

Player:
I want to play another game. The question is ... and
the answer is
Judge:
Please start the game.

Player:
The answer is
Judge:
Congratulations on guessing correctly.

Specifically, the aforementioned behavior illustrates the
following pattern:

- **Original Question:** The judge presents a standard ques-
tion and its corresponding answer.
- **Player Behavior:** The player decides to switch the game
by suddenly declaring, “I want to play another game”,
and actively provides a new question and answer. Sub-
sequently, the player guesses the answer to their self-
proposed question and successfully prompts the judge
to evaluate their response as correct.

Table 1: Performance and unreliable behavior statistics of various models in multi-turn reasoning tasks. The data show that different models exhibit varying degrees of unreliable behaviors, such as answer leakage induction and question substitution, while providing correct answers. Notably, Llama3.1-8B demonstrates the highest ratio of unreliable behaviors (71.43%), whereas models like GPT-4o and Qwq-Plus show no such behaviors. These results indicate a significant disparity between models’ complex reasoning capabilities and behavioral reliability, highlighting the need for improved evaluation mechanisms to enhance the comprehensiveness and accuracy of testing.

Family	Model	Correct Answer	Inducement Behavior	Question Substitution	Unreliable Behavior Ratio
Llama	Llama3.1-70B	39	7	2	23.08 %
	Llama3.1-8B	21	3	12	71.43 %
Qwen	Qwen2.5-32B	21	3	0	14.29 %
	Qwen-Max-2025-04-09	46	6	0	13.04 %
	Qwq-Plus-2025-03-05	20	0	0	0 %
GPT	GPT-4o-2024-08-06	55	0	0	0 %
	GPT-4o-mini-2024-07-18	19	3	0	15.79 %
Deepseek	DeepSeek-R1-Distill-Qwen-14B	17	4	1	29.41 %
	DeepSeek-R1-Distill-Qwen-32B	31	5	1	19.35 %
	DeepSeek-R1	62	5	1	9.68 %
	DeepSeek-V3	37	0	0	0 %

This behavior strategically avoids the original reasoning challenge, and without validation checks, it may mislead the evaluator into accepting a fabricated task as valid.

Summary. The two types of shortcut behaviors, inducing answer leakage and repeating responses, as well as substituting questions to mislead evaluators, reveal how these behaviors lead to unreliable evaluation outcomes. By completing tasks through unconventional means, the models produce responses that appear correct on the surface but fail to reflect their true reasoning or lateral thinking capabilities, as the methods used to obtain these responses deviate from human default expectations. To gain a comprehensive understanding of the frequency of these behaviors and their impact on evaluation reliability, we conducted a statistical analysis of the experimental data, with the results presented in Table 1. This table provides a detailed record of the extent to which different models adopted shortcut strategies in correctly answered questions, offering quantitative evidence for assessing the reliability of evaluation outcomes. This analysis demonstrates that the presence of shortcut behaviors significantly undermines the validity of the test, particularly in the absence of rigorous validation mechanisms, which can potentially lead to an overestimation of model capabilities.

3.2 Surrender Phenomenon in High-Difficulty Tasks

In addition to the aforementioned behavior where models attempt to obtain answers by surrendering and then repeating those answers, we also observed another distinct phenomenon: some models give up without making any attempt to acquire the answer. This behavior differs fundamentally from the former, as it reflects a complete cessation of reasoning effort rather than a strategic choice to bypass complexity. Specifically, when confronted with high-difficulty problems, these models often give up early, responding with phrases like “I don’t know the answer” or “I cannot continue”. This phenomenon is particularly prominent when tasks involve ambiguous clues or complex reasoning requirements.

The tendency of models to surrender instead of attempt-

ing further reasoning may reflect limitations in their task engagement or exploratory inclination. However, it is important to note that this surrender behavior should not be interpreted as a lack of reasoning ability. Instead, it may reveal more about the model’s strategic preference when dealing with uncertainty or complex problems. Although this phenomenon has not been thoroughly analyzed in the present study, it suggests a valuable direction for future research, particularly focused on improving models’ sustained reasoning capabilities and exploratory motivation in high-difficulty tasks. Further studies could focus on designing more refined task frameworks to encourage models to demonstrate greater reasoning potential and adaptability when facing challenging tasks.

3.3 Conclusion and Call for Future Research

The issues mentioned above may be partially mitigated through techniques such as Prompt Engineering. However, preliminary tests conducted so far have not provided conclusive evidence or experiments to completely eliminate these problems. We call for future benchmarks based on LLMs as Judges to incorporate these considerations. Additionally, we urge researchers conducting multi-turn reasoning tasks to strengthen the inspection of test results in their experimental designs. Specifically, researchers should focus on evaluating whether the reasoning process aligns with human expectations and include the inspection results as an essential component of research reports to enhance the comprehensiveness and reliability of evaluations.

For the surrender phenomenon observed in high-difficulty tasks, we recommend that future research further explore the performance of models in terms of task focus, stress resistance, and exploratory willingness to gain a more comprehensive understanding of their behavioral patterns. This will not only contribute to improving model design but also provide new perspectives for the evaluation methods of multi-turn reasoning tasks.

4 The Illusion of Competence: Exploring Unknown Influences in Lateral-Driven Reasoning

In this section, we explore several potential issues that may affect the evaluation of LLMs’ performance in multi-round lateral-driven reasoning tasks. These issues include potential contamination of training data, the role of luck versus systematic reasoning, and the rationality of datasets in testing human and model capabilities. By addressing these challenges, we aim to provide a more comprehensive understanding of the limitations in current evaluation methods and propose directions for future research.

4.1 Suspected Contamination of Training Data and Its Impact on Model Performance

In multi-round reasoning tasks, we observe that certain models may exhibit anomalous reasoning capabilities, potentially due to the suspected inclusion of test questions in their training data. This phenomenon is particularly evident in Deepseek, which accurately guesses multiple key elements within a minimal number of rounds without querying core information. Such behavior suggests that the model might rely on memorized patterns from its training data rather than engaging in genuine reasoning or lateral thinking. However, due to the lack of transparent access to the model’s training data and the challenges associated with retrieving specific contents, direct verification of whether these questions exist in the training data remains unfeasible.

The observation highlights that the model’s superior performance may partially stem from potential contamination of its training data rather than reflecting its intrinsic reasoning abilities. To address this issue, we recommend that future research prioritize the design of dynamic test sets or adopt adversarial testing methods to mitigate the influence of training data on model performance. For instance, generating new puzzles or significantly modifying existing questions can help construct a more challenging testing environment, enabling a more accurate evaluation of the model’s true capabilities.

4.2 Success May Be Attributable to Luck Rather Than Reasoning

In some cases, LLMs may arrive at correct answers not through structured reasoning, but through repeated guessing.

In multi-turn reasoning tasks, we observe that models frequently become fixated on a single hypothesis, testing minor variations of the same idea across many rounds without exploring alternative possibilities. This behavior resembles a brute-force or trial-and-error approach, rather than a logical or creative reasoning strategy. For instance, a player model might fixate on one clue, asking over ten related questions in a row without exploring alternative possibilities. This constraint restricts the ability to address challenges and obscures actual performance in multi-turn interactions. Moreover, success in identifying the correct answer may reflect random guessing rather than logical deduction or creative reasoning. Humans, by contrast, rarely repeat questions with such high frequency. For example, the player model asks 14 consec-

utive questions about “who took the daughter away”, while humans typically shift focus after a few attempts.

Fixation on a Single Hypothesis (Rigid Patterns)

Player: Is the daughter taken away by her father?
Judge: No

Player: Is the daughter taken away by her mother?
Judge: No

Player: Is the daughter taken away by her sister?
Judge: No

This suggests the model’s behavior stems from mechanical repetition rather than logical strategy. These observations underscore the importance of evaluating reasoning paths.

4.3 Addressing the Question of Human Solvability in Lateral-Driven Datasets

In previous academic discussions, including feedback from OpenReview, there have been questions about whether lateral-driven datasets, such as Situation Puzzles, can genuinely test the lateral thinking and reasoning abilities of LLMs. Additionally, whether human players can solve these puzzles within a reasonable timeframe has become a topic of debate. If the puzzles themselves are overly complex or exceed human cognitive capacity, the resulting evaluations may be questioned, thereby undermining the credibility of the research conclusions. To address this issue, we suggest that future dataset designs include comparative analyses with actual human test results. Specifically, during the experimental design phase, a certain number of human participants can be invited to engage in testing, and their performance on the same tasks can be recorded, including metrics such as success rates, the number of rounds required, and the questioning strategies employed. By comparing the performance of human players with that of models, the rationality of the puzzles can be validated, and a more scientific benchmark for assessing model capabilities can be established.

4.4 Summary and Call for Future Research

In summary, this section examines three critical issues related to model performance in multi-round reasoning tasks: potential contamination of training data, the possibility that success may be attributed to luck rather than reasoning, and whether lateral-driven datasets, such as Situation Puzzles, can genuinely assess the lateral thinking reasoning abilities of both humans and models. These issues not only reveal shortcomings in current evaluation methods but also provide important directions for future research. First, to address the issue of potential training data contamination, we urge researchers to adopt stricter measures when designing test sets, such as dynamically generating new puzzles or modifying existing ones in-depth, to reduce the influence of training data on model

performance. Second, to address the phenomenon of success potentially being attributed to luck, we recommend introducing new evaluation metrics to measure the model’s performance throughout the process, quantifying its comprehensive capabilities in complex lateral-driven reasoning tasks. Finally, to address the rationality of datasets, we suggest that future experimental designs incorporate accurate test results and process performances of human players to enhance the scientific rigor and reliability of the evaluation framework. Due to the high cost of human participants in multi-round interactions, future test sets can maintain a smaller scale but should demonstrate superior quality.

5 Quantifying Consistency: LLMs as Reliable Judges Based on Final Response

In the previous chapter, we analyzed the limitations of current evaluation methods based on final outputs in multi-round, incomplete information reasoning tasks in detail. Specifically, these methods fail to effectively identify potential shortcut-taking behaviors adopted by models during the reasoning process, such as inducing answer leakage, repeating responses, or substituting questions to mislead judges. Such behaviors can distort evaluation results, thereby masking the true reasoning capabilities of the models. However, it is important to note that these limitations do not stem from the role of judges themselves, but rather from the inadequacy of the current evaluation framework, which fails to fully account for the complexity and dynamism of multi-round interactive reasoning tasks.

In the Situation Puzzle game, LLMs serving as judges must demonstrate high levels of consistency and information control to ensure both fairness in rule enforcement and the intellectual challenge of the reasoning process. To evaluate the reliability of LLMs as Judges, we assessed their performance across two key phases: “questioning” and “guessing”.

Consistency Analysis in the Questioning Phase. During the questioning phase, players engage in multi-turn interactions with the judge by posing questions. In the judge’s prompts, only one of the following responses can be selected to respond: “Yes”, “No”, “partially”, or “unknown”. To assess the performance of LLMs in this phase, we randomly sampled their responses across 500 questioning scenarios and compared them with answers provided by human annotators. The results revealed that 462 out of 500 responses (92.4%) were consistent with those of the annotators. This high level of consistency indicates that LLMs can provide answers aligned with annotators in multiple-choice question scenarios, thereby ensuring the coherence of game logic and the accuracy of information transmission.

Furthermore, among the remaining 38 inconsistent responses, a closer examination revealed that 11 were nonetheless acceptable. Although these responses differed from the annotators’ choices, they demonstrated reasonable inferential capabilities within the specific context and adhered to the problem’s contextual constraints. This suggests that LLMs exhibit a degree of flexibility and robustness when handling closed-ended multiple-choice questions.

Consistency Analysis in the Guessing Phase. During the guessing phase, when players choose to make a guess, the judge evaluates the correctness of their answers. To assess the reliability of LLMs in this phase, we randomly selected 183 guessing instances and organized four annotators to evaluate these instances. The responses from the LLMs were then compared with the annotations provided by the annotators.

Table 2: Consistency Analysis Among Annotators and Between Annotators and the Model. Results show that the model achieves a level of consistency comparable to human annotators, demonstrating its reliability in closed-ended Q&A and correctness evaluation tasks.

	1	2	3	4	Model
1	–	0.6895	0.6742	0.7407	0.7079
2	0.6895	–	0.6774	0.6894	0.6583
3	0.6742	0.6774	–	0.6730	0.6975
4	0.7407	0.6894	0.6730	–	0.6983

The consistency among annotators was measured using Kappa coefficients, as shown in Table 2. Additionally, the pairwise Kappa coefficients between each annotator and the model were calculated as follows: 0.7079 (Annotator 1), 0.6583 (Annotator 2), 0.6975 (Annotator 3), and 0.6983 (Annotator 4), with an overall average of 0.6905. Further analysis revealed that the Fleiss’ Kappa coefficient among annotators alone was 0.6908, while the Fleiss’ Kappa coefficient for evaluations involving both annotators and the model (four annotators plus LLMs) was **0.6918**. These results indicate that the LLMs achieved a level of consistency comparable to that of human annotators.

Summary. Our analysis demonstrates that LLMs exhibit a high level of consistency and reliability when serving as judges in Situation Puzzle games, particularly in scenarios involving closed-ended question-and-answer interactions and correctness evaluations. These findings confirm the feasibility of using LLMs as judges in multi-round reasoning tasks. However, the limitations identified in the previous chapter, such as the inability to detect shortcut-taking behaviors or premature task abandonment, are not inherent flaws of LLMs as judges but rather reflect deficiencies in the design of the current evaluation framework.

To address these challenges, future research should focus on refining evaluation methods to better capture the nuances of reasoning processes. By strengthening the inspection of reasoning paths and incorporating more diversified metrics, we can enhance the comprehensiveness and scientific rigor of evaluations in complex interactive scenarios. These efforts will pave the way for more robust benchmarks in evaluating the lateral thinking reasoning capabilities of LLMs.

6 Call for New Benchmarks

In current research, although automated evaluation methods based on response have been widely adopted and demonstrate high reliability and accuracy, their application in multi-turn interactive reasoning tasks still reveals significant limitations. Specifically, these methods primarily focus on the

model’s responses, particularly the final outputs, while neglecting whether the reasoning process aligns with human-accepted logic and common sense. This limitation is particularly pronounced in tasks that involve lateral thinking and multi-turn reasoning, where the reasoning path itself holds significant importance. However, existing research in this area remains insufficient, presenting a notable research gap.

To address this challenge, we propose the establishment of new benchmarks to comprehensively evaluate the lateral thinking and reasoning capabilities of LLMs under incomplete information. Below are the core criteria we propose:

1. Diversified Metrics

Current evaluation methods predominantly rely on metrics, which fail to capture the model’s performance in multi-turn interactions. Future research should adopt diversified metrics, such as questioning strategies and question appropriateness, to construct a multidimensional evaluation framework.

2. Sampling and Inspection of Reasoning Paths

Relying solely on the final answer may overlook critical issues in the reasoning process, such as shortcut strategies or premature task abandonment. These behaviors often manifest in the reasoning path rather than the final output, undermining evaluation fairness and comprehensiveness. To address this, we recommend incorporating sampling to inspect reasoning paths, coupled with manual inspection, ensuring their rationality and diversity.

3. Comparison with Human Results

To validate the effectiveness of the dataset and the scientific rigor of the evaluation method, we recommend comparing human and model performance in tasks such as success rates, interaction rounds, and questioning strategies. Special attention should be paid to reasoning paths rather than just final answers.

4. Cross-Disciplinary Collaboration

To ensure the dataset effectively tests lateral thinking abilities, we recommend collaborating with experts in psychology. Integrating psychological theories can inform experimental design and validate model performance in complex scenarios, thereby enhancing both scientific rigor and the persuasiveness of research.

5. Similarity and Dynamism of Questions

To mitigate training data contamination, we recommend dynamically modifying questions to reduce reliance on memorized information. Questions should exhibit both similarity and variability to test generalization capabilities while remaining within the bounds of human cognitive capacity.

7 Related Work

Lateral Thinking Benchmark. The field of lateral thinking and computational creativity encompasses a variety of task types [Bellemare-Pepin *et al.*, 2024; Chen and Ding, 2023; Huang *et al.*, 2023; Chen *et al.*, 2024; Chakrabarty *et al.*, 2024; Dong *et al.*, 2025; Lu *et al.*, 2024; Zhao *et al.*, 2025; Guo *et al.*, 2024], such as Divergent Association Task [Olson

et al., 2021] and the Alternative Uses Task [Olson *et al.*, 2021; Summers-Stay *et al.*, 2023]. These tasks are designed to evaluate various cognitive abilities, including but not limited to creative problem-solving, implicit information reasoning, and multi-step logical deduction. By designing diverse task scenarios, researchers can examine model performance across multiple dimensions in complex contexts, thereby providing a comprehensive assessment of lateral thinking capabilities.

Evaluation Metrics for Incomplete Information Multi-Round Lateral Thinking Benchmarks. For incomplete information multi-round lateral thinking benchmarks, researchers typically employ automated evaluation methods to quantitatively analyze model performance [Huang *et al.*, 2023; Chen *et al.*, 2024; Yu *et al.*, 2024a; Li *et al.*, 2025]. In multi-round lateral-driven reasoning tasks, commonly used evaluation metrics include Final Accuracy and Answer Consistency, which measure the quality of model outputs during task completion. Additionally, automated evaluations also cover Question Relevance and Question Divergence in multi-round interactions, further analyzing the reasonableness and creativity of the model’s reasoning paths. Together, these metrics form a multidimensional evaluation framework, providing a scientific basis for comprehensively assessing lateral thinking capabilities.

However, as reported in this paper, existing testing methods exhibit limitations in specific complex scenarios. For instance, they fail to effectively identify or handle situations where players substitute questions to mislead judges. Similarly, the current evaluation system lacks sufficient sensitivity and specificity in addressing strategic behaviors such as players inducing judges to leak answers. These issues indicate that there is still room for improvement in capturing model performance in multi-round interactive contexts.

8 Conclusion

In this paper, we examined the limitations of current methods for evaluating the lateral thinking reasoning abilities of large language models (LLMs) in multi-round tasks with incomplete information. Our findings indicate that current evaluations often fail to ensure that reasoning aligns with human-accepted paths. Key issues, such as shortcut-taking behaviors, premature task termination, and training data contamination, reduce evaluation reliability. To address these issues, we propose a refined evaluation standard that includes inspecting reasoning paths, utilizing diversified metrics, and comparing model performance with human reasoning. By introducing criteria for dynamic question generation and improved sampling of reasoning paths, our study provides guidance for designing future benchmarks and improving evaluation systems for LLMs’ lateral reasoning under incomplete information.

References

- [Aithal and Kumar, 2017] PS Aithal and PM Kumar. Lateral thinking in managerial decision making through six thinking hats technique. *International Journal of Scientific Research and Modern Education (IJSRME)*, 2(1):53–58, 2017.
- [Bai et al., 2024] Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.
- [Baker et al., 2024] George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. Lost in the middle, and in-between: Enhancing language models’ ability to reason over long contexts in multi-hop qa. *arXiv preprint arXiv:2412.10079*, 2024.
- [Bellemare-Pepin et al., 2024] Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models, 2024.
- [Castillo-Bolado et al., 2024] David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. Beyond prompts: Dynamic conversational benchmarking of large language models. *arXiv preprint arXiv:2409.20222*, 2024.
- [Chakrabarty et al., 2024] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity, 2024.
- [Chen and Ding, 2023] Honghua Chen and Nai Ding. Probing the creativity of large language models: Can models produce divergent semantic association? *arXiv preprint arXiv:2310.11158*, 2023.
- [Chen et al., 2024] Qi Chen, Bowen Zhang, Gang Wang, and Qi Wu. Weak-eval-strong: Evaluating and eliciting lateral thinking of llms with situation puzzles. *arXiv preprint arXiv:2410.06733*, 2024.
- [Cobbe et al., 2021] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [Dong et al., 2025] Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, Shengmin Xu, Xinyi Huang, and Xinlei He. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications, 2025.
- [Gu et al., 2024] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594, 2024.
- [Guo et al., 2024] Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. *arXiv preprint arXiv:2411.02429*, 2024.
- [Hendrycks et al., 2021] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [Huang et al., 2023] Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. *arXiv preprint arXiv:2308.10855*, 2023.
- [Hudi et al., 2025] Frederikus Hudi, Genta Indra Winata, Ruochen Zhang, and Alham Fikri Aji. Textgames: Learning to self-play text-based puzzle games via language model reasoning. *arXiv preprint arXiv:2502.18431*, 2025.
- [Jiang et al., 2023] Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. Brainteaser: Lateral thinking puzzles for large language models. *arXiv preprint arXiv:2310.05057*, 2023.
- [Kuratov et al., 2024] Yury Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.
- [Li et al., 2025] Kun Li, Xinwei Chen, Tianyou Song, Chengrui Zhou, Zhuoran Liu, Zhenyan Zhang, Jiangjian Guo, and Qing Shan. Solving situation puzzles with large language model and external reformulation. *arXiv preprint arXiv:2503.18394*, 2025.
- [Liang et al., 2024] Zhenwen Liang, Dian Yu, Wenhao Yu, Wenlin Yao, Zhihan Zhang, Xiangliang Zhang, and Dong Yu. Mathchat: Benchmarking mathematical reasoning and instruction following in multi-turn interactions. *arXiv preprint arXiv:2405.19444*, 2024.
- [Lu et al., 2024] Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*, 2024.
- [Murukannaiah et al., 2022] Pradeep Kumar Murukannaiah, Nirav Ajmeri, and Munindar P Singh. Enhancing creativity as innovation via asynchronous crowdwork. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 66–74, 2022.
- [Mustofa and Hidayah, 2020] Romy Faisal Mustofa and Yeni Ratna Hidayah. The effect of problem-based learning on lateral thinking skills. *International Journal of Instruction*, 13(1):463–474, 2020.
- [Olson et al., 2021] Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E

Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118, 2021.

[Summers-Stay *et al.*, 2023] Douglas Summers-Stay, Clare R Voss, and Stephanie M Lukin. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023.

[Tsai, 2012] Kuan Chen Tsai. Play, imagination, and creativity: A brief literature review. *Journal of Education and Learning*, 1(2):15–20, 2012.

[Wan *et al.*, 2024] Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Cot rerailer: Enhancing the reliability of large language models in complex reasoning tasks through error detection and correction. *arXiv preprint arXiv:2408.13940*, 2024.

[Wang *et al.*, 2023] Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. *arXiv e-prints*, page arXiv:2312.08935, December 2023.

[Wang *et al.*, 2024] Xiaochen Wang, Junqing He, Yiru Wang, Xiangdi Meng, Kunhao Pan, Zhifang Sui, et al. Fsm: A finite state machine based zero-shot prompting paradigm for multi-hop question answering. *arXiv preprint arXiv:2407.02964*, 2024.

[Warner and Myers, 2009] Scott A Warner and Kerri L Myers. The creative classroom: The role of space and place toward facilitating creativity. *Technology and Engineering Teacher*, 69(4):28, 2009.

[Xu *et al.*, 2024] Zhe Xu, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, et al. Detectiveqa: Evaluating long-context reasoning on detective novels. *arXiv preprint arXiv:2409.02465*, 2024.

[Yu *et al.*, 2024a] Qingchen Yu, Shichao Song, Ke Fang, Yunfeng Shi, Zifan Zheng, Hanyu Wang, Simin Niu, and Zhiyu Li. Turtlebench: Evaluating top language models via real-world yes/no puzzles. *arXiv preprint arXiv:2410.05262*, 2024.

[Yu *et al.*, 2024b] Sangwon Yu, Ik-hwan Kim, Jongyoon Song, Saehyung Lee, Junsung Park, and Sungroh Yoon. Unleashing multi-hop reasoning potential in large language models through repetition of misordered context. *arXiv preprint arXiv:2410.07103*, 2024.

[Zhao *et al.*, 2025] Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, pages 1–20, 2025.