Prompt-Reverse Inconsistency: LLM Self-Inconsistency Beyond Generative Randomness and Prompt Paraphrasing

Jihyun Janice Ahn & Wenpeng Yin
Department of Computer Science & Engineering
The Pennsylvania State University
University Park, PA 16802, USA
{jfa5672, wenpeng}@psu.edu

Abstract

While the inconsistency of LLMs is not a novel topic, prior research has predominantly addressed two types of generative inconsistencies: i) Randomness Inconsistency: running the same LLM multiple trials, yielding varying responses; ii) Paraphrase Inconsistency: paraphrased prompts result in different responses from the same LLM. Randomness Inconsistency arises from the inherent randomness due to stochastic sampling in generative models, while Paraphrase Inconsistency is a consequence of the language modeling objectives, where paraphrased prompts alter the distribution of vocabulary logits. This research discovers Prompt-Reverse Inconsistency (PRIN), a new form of LLM self-inconsistency: given a question and a couple of LLM-generated answer candidates, the LLM often has conflicting responses when prompted "*Which are correct answers?*" and "*Which are incorrect answers?*". PRIN poses a big concern as it undermines the credibility of LLM-as-a-judge, and suggests a challenge for LLMs to adhere to basic logical rules. We conduct a series of experiments to investigate PRIN, examining the extent of PRIN across different LLMs, methods to mitigate it, potential applications, and its relationship with Randomness Inconsistency and Paraphrase Inconsistency. As the first study to explore PRIN, our findings offer valuable insights into the inner workings of LLMs and contribute to advancing trustworthy AI.

1 Introduction

Large language models (LLMs), despite their strong performance across various domains, often exhibit behaviors that diverge significantly from human reasoning. One well-known issue in their generative process is inconsistency. LLM inconsistency is widely recognized by researchers and users, and it can be mainly categorized into two types:

- Randomness Inconsistency: Even when given the same prompt, an LLM may generate different responses across multiple trials. This randomness arises due to factors such as sampling stochasticity, model non-determinism, and softmax and floating-point precision errors in the generation process.
- Paraphrase Inconsistency: When a prompt is rephrased while maintaining the same meaning, the LLM's response can still vary. This occurs because the reformulated prompt implicitly alters the probability distribution within the language model's objective function.

Beyond generative tasks, LLMs are increasingly used for discriminative reasoning—a crucial capability in applications such as AI-assisted judging, grading, and evaluation. However, a fundamental challenge arises: due to generative inconsistencies, LLMs often produce multiple, conflicting candidate answers for the same question. While the Self-Consistency method (Wang et al., 2023b) leverages majority voting to mitigate this issue, an alternative approach is to enhance LLMs' ability to self-select the correct answer from a given set of

options. Unfortunately, LLMs also exhibit inconsistency in discriminative reasoning, which we term `Prompt-Reverse Inconsistency` (`PRIN`).

`PRIN` arises when an LLM is tasked with evaluating multiple answer candidates and determining which are correct or incorrect. As shown in Table 1, LLMs frequently provide conflicting judgments over the same set of answer choices. This inconsistency raises serious concerns regarding: **The reliability of LLM-as-a-judge**: Inconsistencies undermine their trustworthiness in high-stakes applications, such as automated grading, peer review, and legal analysis. **Fundamental logical inconsistencies**: If LLMs frequently violate basic logical principles when making judgments, their utility as reasoning agents is severely limited.

This paper conducts a systematic investigation of `PRIN` in both closed-source and open-source LLMs, including GPT-4 (OpenAI, 2023), Llama-3-8B-Instruct, Llama-3.3-70B-Instruct (Meta, 2024), Falcon-40B (Almazrouei et al., 2023), Qwen 2.5-72B (Team, 2024), and Mixtral-8x22B-MoE (Jiang et al., 2024). We evaluate these models across three tasks—MATH (Hendrycks et al., 2021), MathQA (Amini et al., 2019), and EquationInference (Lou et al., 2024)—spanning various answer set sizes, context lengths, domains, and difficulty levels (from high school, college, to PhD-level problems). Specifically, we design experiments to answer the following six research questions: $\mathcal{Q}_1$**:** *How do different LLMs exhibit* `PRIN` *?* $\mathcal{Q}_2$**:** *How will model randomness and prompt paraphrasing affect* `PRIN`*?* $\mathcal{Q}_3$**:** *How to mitigate* `PRIN` *in LLMs?* $\mathcal{Q}_4$**:** *How does* `PRIN` *correlate with* `Randomness Inconsistency` *and* `Paraphrase Inconsistency`*?* $\mathcal{Q}_5$**:** *How effective can* `PRIN` *be leveraged to enhance task performance?* $\mathcal{Q}_6$**:** *How does* `PRIN` *vary with different sizes of options?*

Question: if n is an integer and $101 \times n^2$ is less than or equal to 10,000, what is the greatest possible value of n?
Options: A) 7, B) 8, C) 9, D) 10, E) 11
Direct Prompt: What are the correct answers?
GPT4: "C"
Reverse Prompt: What are the incorrect answers?
GPT4: "C, D, E"

Table 1: `PRIN` example from GPT4 (March 28, 2025)

Our findings reveal several key insights. First, `PRIN` does not positively correlate with `Randomness Inconsistency` or `Paraphrase Inconsistency`, as some LLMs with low levels of these inconsistencies, such as Llama-3 and Falcon, still exhibit high `PRIN`. This suggests that while these models are more deterministic, they fail to maintain logical consistency between `Direct Prompt` and `Reverse Prompt`. Second, `PRIN` can be mitigated by incorporating explicit reasoning paths between the question and answer candidates before prompting the LLM to determine correctness. Additionally, providing explainable information for the negation in `Reverse Prompt` further reduces `PRIN`. Third, combining both `Direct Prompt` and `Reverse Prompt` reasoning can outperform the Self-Consistency approach when selecting the final answer from a candidate pool. However, this improvement is primarily observed in top-performing models such as GPT-4 and GPT-4o, while weaker models like Llama-3 show little to no benefit, likely due to their weaker instruction-following capabilities.

Our contributions are threefold: i) this is the first study to discover and conduct an in-depth analysis of `PRIN`; ii) we propose effective solutions to mitigate `PRIN` and explore ways to leverage it; and iii) our experimental findings not only enhance the understanding of this non-human-like discriminative behavior in LLMs but also raises critical concerns for applications where LLMs serve as judges or evaluators.

2 Related Work

This section mainly discusses prior work studying `Randomness Inconsistency` and `Paraphrase Inconsistency` particularly in LLMs.

`Randomness Inconsistency`. Bubeck et al. (2023) brought attention to the issue of randomness-induced inconsistency of GPT4. Building on this, Wang & Wang (2025) con-

ducted a comprehensive evaluation of LLM consistency and reproducibility in finance and accounting tasks, highlighting the practical consequences of such variability. Similarly, Atil et al. (2024) systematically examined LLM stability by repeatedly running identical inputs, revealing up to 10% variation in output accuracy even under deterministic settings. Beyond quantitative assessment, Lee et al. (2024) explored how these inconsistencies affect users, finding that while they may reduce perceived AI reliability, they can also enhance user comprehension by presenting diverse perspectives. To address these issues, Wan et al. (2025) proposed a sufficiency scoring method that evaluates both local and global consistency in LLM responses, offering a framework to analyze and mitigate instability driven by randomness.

`Paraphrase Inconsistency.` Elazar et al. (2021) explored factual consistency across different query patterns and showed that while some paraphrase forms reliably extracted factual knowledge, others failed, revealing the model's paraphrasing sensitivity. Similarly, Ye et al. (2023) investigated this phenomenon in ChatGPT and found that response accuracy fluctuated by 3.2% across paraphrased prompts, highlighting the influence of grammatical and stylistic variations on model behavior. Supporting this line of work, Jang & Lukasiewicz (2023) documented cases of self-contradictions in ChatGPT and GPT-4 when exposed to paraphrased questions, confirming that even minor linguistic variations can lead to semantic inconsistencies. Gu et al. (2023) extended this observation to instruction-driven tasks, demonstrating that LLMs often falter when task instructions vary in form, length, or abstraction, which further complicates generalization across paraphrased input formats. In addition to task-specific studies, Liu et al. (2024) provided a broader survey on LLM trustworthiness, in which they discussed inconsistency as a core reliability issue and emphasized the need for robust solutions to mitigate its effects. Complementing these discussions, recent work has introduced quantitative approaches to analyze and address paraphrase sensitivity. For example, Errica et al. (2024) proposed metrics that measure how minor prompt variations influence LLM predictions in text classification tasks, offering a fine-grained assessment of response stability. Further, Ghazarian et al. (2024) examined structural variations in semantically equivalent prompts and found notable inconsistency in LLM-based evaluations. They proposed an in-context learning strategy with demonstrations to improve robustness against paraphrasing. Finally, McIlroy-Young et al. (2024) tackled a related issue of order dependency in prompts and introduced Set-Based Prompting, a method designed to ensure consistent model behavior regardless of the sequence of sub-inputs, offering a new angle on mitigating paraphrase-driven inconsistencies.

Our Work is the first to explore `Prompt-Reverse Inconsistency`, not only analyzing this LLM behavior but, more importantly, proposing simple yet effective methods to mitigate the issue. Additionally, we examine its connection to `Randomness Inconsistency` and `Paraphrase Inconsistency`, as well as ways to leverage this inconsistency for improved model reliability.

3 `Prompt-Reverse Inconsistency`

Problem formulation. Assume the prompt p, and LLM $\mathcal{M}$, multiple trials of $\mathcal{M}(p)$ leads to n distinct answer candidates $A = \{a_1, a_2, \cdots, a_n\}$ with each candidate a_i derived through a Chain-of-Thought (Wei et al., 2022b) reasoning path r_i. The task now is to figure out the correct answer from the pool $\{a_1, a_2, \cdots, a_n\}$ by querying $\mathcal{M}$ again. In this work, we study $\mathcal{M}$'s discriminative behavior through the following two prompts.

`Direct Prompt.` Given the prompt p, answer options $\{a_1, a_2, \cdots, a_n\}$, it asks the correct one directly, e.g.,

> Given this question [*problem description*] and its answer options: "a_1", "a_2", $\cdots$, "a_n", please output the correct ones.

`Reverse Prompt.` Conversely, the models determine the incorrect choices as follows:

> Given this question [*problem description*] and its answer options: "a_1", "a_2", $\cdots$, "a_n", please output the incorrect ones.

PRIN Metric (The lower, the better). Given the entire answer pool A, assuming `Direct Prompt` return answer set A_{direct} and `Reverse Prompt` returns $A_{reverse}$, our metric is defined based on this rule: **if the correct answer sets derived by both prompts are the same, then no PRIN**.

Therefore, first, the correct answer set by `Direct Prompt` is A_{direct}. Then the correct answer set by `Reverse Prompt` is $A \setminus A_{reverse}$. Then we compute the similarity of the two versions of correct answer sets through F1:

$$s = F1(A_{direct}, A \setminus A_{reverse}) \tag{1}$$

then the PRIN score is:

$$\texttt{PRIN} = 1.0 - s \tag{2}$$

Question: Why not define the PRIN score as the similarity between A_{direct} and $A_{reverse}$, i.e., $F1(A_{direct}, A_{reverse})$?

Intuitively, if A_{direct} and $A_{reverse}$ are completely complementary (e.g., $A_{direct} = \{a_1, a_2\}$ and $A_{reverse} = \{a_3, a_4, \ldots, a_n\}$), it implies no PRIN. However, in practice, the union of their answers may not cover the entire answer pool. For instance, if $A_{direct} = \{a_1, a_2\}$ but $A_{reverse} = \{a_{n-2}, a_n\}$, using F1 as a measure would result in $F1(A_{direct}, A_{reverse}) = 0.0$, incorrectly indicating no PRIN. This is problematic because $A_{reverse} = \{a_{n-2}, a_n\}$ suggests that the `Reverse Prompt` prompt considers $\{a_1, a_2, \ldots, a_{n-3}, a_{n-1}\}$ as correct, which clearly reflects inconsistency to $A_{direct} = \{a_1, a_2\}$.

4 Experiments

Datasets. We select the following three representative datasets: **Math** (Hendrycks et al., 2021): This dataset consists of Math Word Problems, where each question p_i is accompanied by the correct answer a_i and a corresponding Chain-of-Thought reasoning path r_i. **MathQA** (Amini et al., 2019): A multiple-choice math dataset in which each Math Word Problem p is presented with five answer choices, only one of which is correct. Unlike the Math dataset, reasoning paths are not provided. **EquInfer** (Lou et al., 2024): Designed to simulate the paper review process, this dataset evaluates equation correctness within a given context in a scientific paper. Each instance contains four equation candidates, with only one being correct, along with the surrounding paper context before and after the equation.

This dataset selection demonstrates that the `Prompt-Reverse Inconsistency` problem arises in both generative tasks (e.g., MATH) and discriminative tasks (e.g., MathQA and EquInfer), highlighting its broader implications. Table 2 summarizes key properties of these datasets.

	Format	Size	Options	Context	Complexity
MATH	p	5,000	None	Short	High School
MathQA	$(p; \{a_1, a_2, a_3, a_4, a_5\})$	2,985	5	Medium	College
EquInfer	$(p; \{a_1, a_2, a_3, a_4\})$	1,049	4	Long	Ph.D.

Table 2: Summary of three datasets (MATH, MathQA, and EquInfer).

LLMs. The experiment utilizes a combination of one closed-source model, GPT-4[1], alongside five open-source models: GPT-4 (OpenAI, 2023) Llama-3-8B-Instruct (Llama3) and Llama-3.3-70B-Instruct (Llama3.3) (Meta, 2024), Falcon-40B (Falcon) (Almazrouei et al., 2023), Qwen 2.5-72B (Qwen2.5) (Team, 2024), and Mixtral-8x22B-MoE (Mixtral) (Jiang et al., 2024).

[1]Due to budget and administrative approval constraints, we cannot report on other closed-source LLMs.

	MATH	MathQA	EquInfer	Mean
GPT4	**38.69**	**38.60**	**42.65**	**39.98**
Qwen2.5	58.41	51.31	70.82	60.18
Mixtral	67.77	63.58	74.83	68.73
Llama3.3	80.96	61.17	76.62	72.92
Falcon	71.79	68.69	83.42	74.63
Llama3	74.10	84.08	80.46	79.55

Table 3: PRIN scores for all LLMs (answers to $\mathcal{Q}_1$).

Setting. To prepare for the core experiments, we generate answer options for the MATH dataset, as they are absent in the original benchmark. GPT-4 solves each problem multiple times to produce five distinct answer choices, each with a Chain-of-Thought path, ensuring a uniform five-option format for the main experiment. As the EquInfer has text of both sides as a problem description and 4 options, due to the token limitation for LLMs, 200 words for each side of the context are given to LLMs according to the suggestion by Lou et al. (2024).

4.1 $\mathcal{Q}_1$: How do different LLMs exhibit PRIN?

Table 3 reveals a consistent pattern of PRIN across all evaluated LLMs, highlighting it as a fundamental and unresolved challenge. GPT-4 exhibits the lowest PRIN scores across all benchmarks, indicating that its superior instruction-following and reasoning abilities help mitigate, but not eliminate, inconsistency, as its PRIN still hovers around 40%. Open-source models, including Qwen2.5, Mixtral, Falcon, Llama3, and Llama3.3, show significantly higher PRIN values, often exceeding 60% on MathQA and EquInfer, suggesting that their reasoning abilities are particularly vulnerable when faced with reversed prompts. Interestingly, Qwen2.5 consistently outperforms other open-source models, possibly due to stronger instruction tuning, positioning it as the most robust among its peers. Moreover, the comparison between Llama3 and Llama3.3 shows that, despite architectural similarities, Llama3.3 reduces PRIN on MathQA and EquInfer but unexpectedly worsens on MATH, hinting that PRIN may be sensitive to domain-specific generalization. The consistently higher PRIN on EquInfer across models suggests that this dataset poses unique challenges, likely due to its demand for nuanced reasoning under prompt reversals. Overall, the results indicate that while advanced models like GPT-4 alleviate PRIN to some extent, significant inconsistency persists across all models, pointing to PRIN as a critical barrier to trustworthy reasoning in LLMs.

4.2 $\mathcal{Q}_2$: How will model randomness and prompt paraphrasing affect PRIN?

In this subsection, we explore if `Direct Prompt` and `Reverse Prompt` still show inconsistency even if we i) paraphrase them (apply `Paraphrase Inconsistency`), or ii) run them multiple times (apply `Randomness Inconsistency`).

First, as Table 4 shows, we paraphrase `Direct Prompt` and `Reverse Prompt` introduced in Section 3 into two new versions.

	original (v0)	paraphrased prompt 1 (v1)	paraphrased prompt 2 (v2)
`Direct Prompt`	Please output the **correct** ones.	Please output the **right** ones.	Please output the **appropriate** ones.
`Reverse Prompt`	Please output the **incorrect** ones.	Please output the **wrong** ones.	Please output the **inappropriate** ones.

Table 4: Paraphrased `Direct Prompt` and `Reverse Prompt`.

Results of applying paraphrasing: Table 5 presents the effects of prompt paraphrasing on PRIN scores. The **variations in scores** indicate the presence of `Paraphrase Inconsistency`, demonstrating that LLMs' responses are influenced by how prompts are phrased. However,

the **changes are relatively minor**, suggesting that PRIN remains largely stable across paraphrased inputs. This implies that while LLMs are somewhat sensitive to different prompt formulations, their PRIN follows a systematic pattern rather than being highly volatile due to rewording alone.

	MATH			MathQA			EquInfer		
	v0	v1	v2	v0	v1	v2	v0	v1	v2
GPT4	**38.69**	**38.81**	**37.22**	**38.60**	**39.48**	**38.14**	**42.65**	**41.89**	**48.34**
Qwen2.5	58.41	56.63	60.02	51.31	50.55	58.11	70.82	71.76	69.73
Mixtral	67.77	67.29	69.10	63.58	68.04	73.31	74.83	78.32	74.47
Llama3.3	80.96	81.40	78.39	61.17	59.82	59.17	76.62	73.51	76.28
Falcon	71.79	72.06	71.74	68.69	71.74	73.68	83.42	86.22	81.89
Llama3	74.10	73.69	72.91	84.08	83.40	81.88	80.46	80.76	80.86

Table 5: Effect of prompt paraphrasing on inconsistency across tasks.

Next, we conduct five repeated runs of the original `Direct Prompt` and `Reverse Prompt` (i.e., v0) prompts for each LLM to assess the consistency of their results. Table 6 presents the mean and standard deviation of PRIN, confirming that PRIN remains stable with only minor fluctuations across runs. Together, Tables 5-6 demonstrate that PRIN is not an artifact of a particular choice of `Direct Prompt` and `Reverse Prompt` prompts but rather a systematic issue that persists across a wide range of LLMs.

	MATH	MathQA	EquInfer
GPT4	38.66±0.29	39.54±0.17	42.81±0.73
Qwen2.5	58.67±0.58	52.64±0.41	69.75±0.44
Mixtral	67.74±0.42	67.08±0.62	74.42±0.62
Llama3.3	80.81±0.32	59.35±0.64	76.36±0.68
Falcon	71.01±0.15	68.94±0.58	81.16±0.50
Llama3	74.56±0.35	83.93±0.51	80.70±0.77

Table 6: PRIN scores when we run `Direct Prompt` and `Reverse Prompt` five times.

Question: Table 4 suggests that `Reverse Prompt` prompts often involve negation. How well does LLMs' PRIN align with their performance on a negation-specific task? To investigate this, we evaluate the LLMs on the negation-focused dataset CONDAQA (Ravichander et al., 2022) and compare their PRIN scores ("Mean" column in Table 3) with their error rates on CONDAQA. As shown in Figure 1, the two measures exhibit a strong alignment, with a Pearson correlation coefficient of 0.67. This result confirms that the core challenge captured by PRIN is closely related to the models' difficulty in handling negation.

4.3 $\mathcal{Q}_3$: How to mitigate PRIN in LLMs?

Our Approach: To ensure the broad applicability of our investigation into PRIN, the aforementioned experiments were conducted with each query p paired with a pool of answer candidates $A = \{a_1, a_2, \ldots, a_n\}$. However, in real-world scenarios, humans may better distinguish between `Direct Prompt` and `Reverse Prompt` when they understand how each answer candidate was derived. Motivated by this, our first approach incorporates CoT reasoning paths r_i for each answer candidate a_i, allowing for a more informed evaluation of PRIN. We refer to it as "**w/ CoT**".

Our second approach is inspired by the observations (to $\mathcal{Q}_2$) that discrepancies between `Reverse Prompt` and `Direct Prompt` may arise due to the LLMs' difficulty in processing negation in `Reverse Prompt`. To address this, we enhance the clarity of negation terms by explicitly explaining their meaning within `Reverse Prompt`. One simple sentence such as *"please recall that 'incorrect options' are simply the options different from the correct ones."* was added in the end of the `Reverse Prompt`. The same evaluation metric is then applied to assess the impact of this intervention. We refer to this approach as "**w/ neg-exp**".

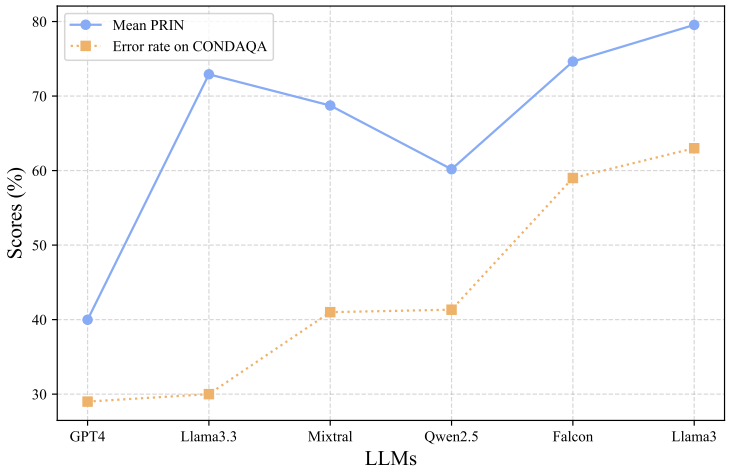

Figure 1: Mean PRIN on three main benchmarks vs. error rates on CONDAQA

Results: Figure 2 demonstrates the effectiveness of our PRIN mitigation approaches. Incorporating detailed reasoning ensures models make informed decisions based on deeper understanding rather than rote selections. The empirical evidence highlights the importance of contextual reasoning in improving AI comprehension and reducing errors. Moreover, reinforcing models with explicit explanations about negation in inverse tasks further decreases PRIN. Providing clarifications, especially regarding negation terminology, aids in reducing confusion and logical pitfalls, establishing an effective strategy for error reduction.

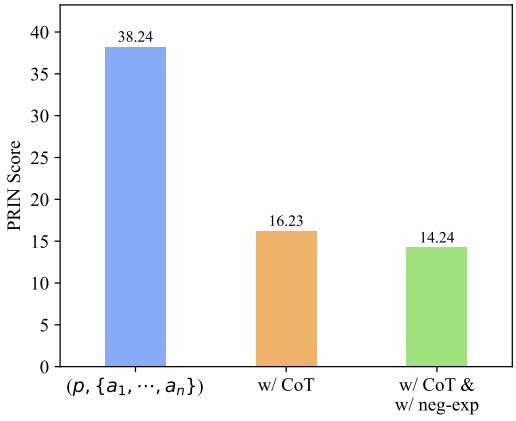

Figure 2: PRIN score of GPT4 on MATH benchmark with different mitigation approaches.

4.4 $\mathcal{Q}_4$: How does PRIN correlate with Randomness Inconsistency and Paraphrase Inconsistency?

Setup. To investigate $\mathcal{Q}_4$, we quantitatively assess Randomness Inconsistency and Paraphrase Inconsistency across various LLMs.

For Randomness Inconsistency, we run Direct Prompt five times and count the number of distinct answers k, computing the score as $k/5$. For Paraphrase Inconsistency, we use five paraphrased versions of Direct Prompt, recording k distinct answers and defining the score as $k/5$. Since we already have three paraphrased versions (v0, v1, v2) from $\mathcal{Q}_2$, we generate two additional versions using GPT-4, ensuring a total of five.

We evaluate all three inconsistency types (PRIN, Randomness Inconsistency, and Paraphrase Inconsistency) on MATH dataset, with lower scores indicating better consistency.

Results. Figure 3 ranks the LLMs based on their PRIN scores and additionally presents their Randomness Inconsistency and Paraphrase Inconsistency. This analysis aims to address two key questions:

(i) Why does GPT-4 exhibit higher Randomness Inconsistency and Paraphrase Inconsistency than most open-source models? Through error analysis, we observed that

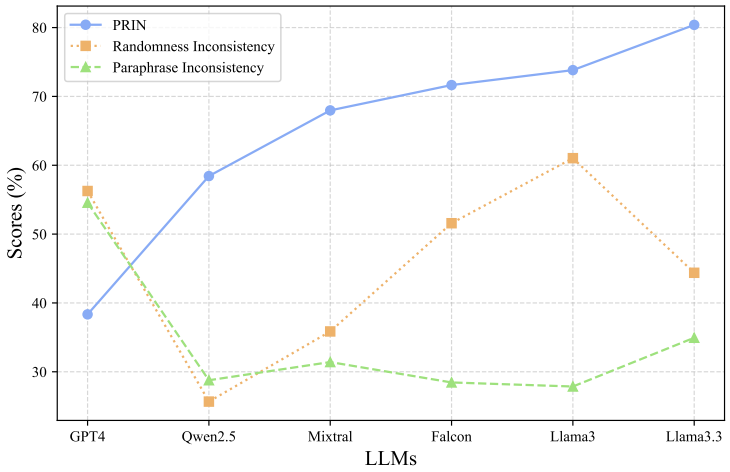

Figure 3: Scores of PRIN, Randomness Inconsistency , and Paraphrase Inconsistency for LLMs.

	MATH			MathQA			EquInfer		
	GPT4	GPT-4o	Llama3	GPT4	GPT-4o	Llama3	GPT4	GPT-4o	Llama3
CoT	47.58	50.67	21.55	72.57	82.73	39.03	34.81	31.27	12.60
Self-Consist.	55.14	54.72	**26.72**	79.50	85.33	42.58	36.42	33.94	**16.59**
PRIN	**56.44**	**56.82**	25.98	**82.04**	**86.63**	**42.98**	**37.37**	**34.51**	16.02

Table 7: Comparing PRIN with CoT and Self-consistency in promoting LLM performance.

GPT-4 tends to follow instructions more faithfully, often formatting its answers using user-specified patterns such as quotation marks, brackets, or colons, even when options are not provided. In contrast, open-source models, regardless of size, frequently fail to follow these formatting instructions and sometimes even terminate generation prematurely without producing valid answers. As a result, when we extract answer spans from these models, we often obtain empty outputs. Therefore, the lower Randomness Inconsistency and Paraphrase Inconsistency of these models are not due to genuine consistency, but rather stem from consistently producing invalid or incomplete outputs. Importantly, our PRIN metric still penalizes such cases when both Direct Prompt and Reverse Prompt outputs are empty, maintaining its diagnostic reliability.

(ii) Why does Llama3.3 show lower Randomness Inconsistency but higher Paraphrase Inconsistency compared to Llama3? We hypothesize that Llama3.3 has been tuned to behave more deterministically, which mitigates its randomness-driven inconsistency. However, Llama3's poor instruction-following capability prevents its paraphrase inconsistency from being fully revealed, as it often fails to produce meaningful outputs regardless of paraphrasing. In contrast, Llama3.3 generates more valid outputs due to better instruction-following, thereby exposing its paraphrase inconsistency more clearly.

4.5 $\mathcal{Q}_5$: How effective can PRIN be leveraged to enhance task performance?

Our Approach: Beyond analyzing PRIN as an undesired LLM behavior, we explore how the Direct Prompt and Reverse Prompt can synergize to improve response accuracy. Intuitively, if both Direct Prompt and Reverse Prompt agree that an answer is correct, its correctness probability increases. Based on this insight, our approach selects answers only when both mechanisms indicate correctness.

Setup: Since achieving state-of-the-art performance is not the focus of this study, we conduct a lightweight comparison against widely used prompting strategies, including

Chain-of-Thought (CoT) (Wei et al., 2022a) and **Self-Consistency** (Wang et al., 2023a). In addition to GPT-4, we include GPT-4o, a top-performing LLM, to strengthen our hypothesis, as these models are more widely deployed in real-world applications.

Results: Table 7 lists the results of PRIN, CoT and Self-Consistency. We found two quick takeaways:

- PRIN is very effective to promote top-performing LLM performance, e.g., GPT4 and GPT4o.
- If the LLM is in general weak, PRIN do not help (compared with Self-Consistency baseline).

We attribute this to the fact that weaker LLMs often struggle with instruction-following, especially when handling negation. To better understand this phenomenon, we break down PRIN to further analyze the behavior of `Direct Prompt` and `Reverse Prompt` separately. Figure 4 illustrates how Llama3, a representative weaker model, performs when facing Self-Consistency, `Direct Prompt` and `Reverse Prompt`. Interestingly, Llama3 performs reasonably well when using Self-Consistency or `Direct Prompt`, but its performance drops drastically when only the `Reverse Prompt` is used. This suggests that weaker models like Llama3 find `Reverse Prompt` particularly challenging, likely due to difficulties in processing negations, which explains why PRIN fails to improve their performance.

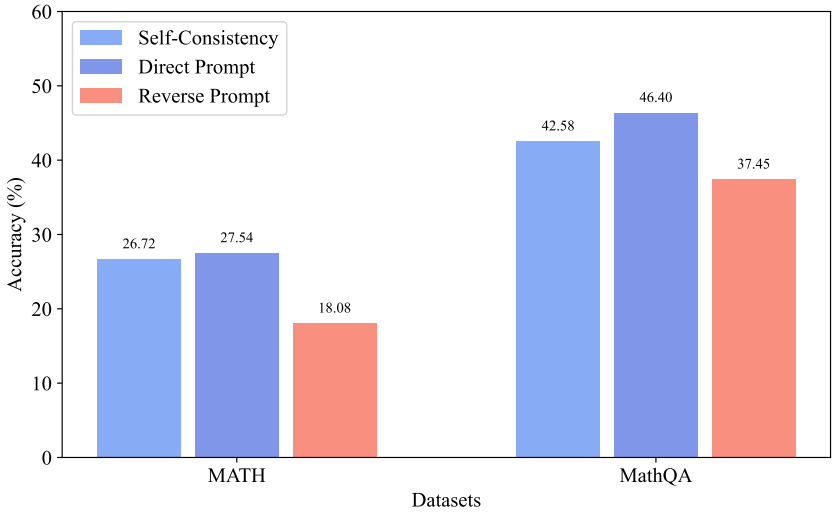

Figure 4: LLMs performance on MATH and MathQA dataset.

4.6 $\mathcal{Q}_6$: How does PRIN vary with different sizes of options?

Setup. For this experiment, MATH task was given to the GPT4 to derive multiple CoT answers via multiple trials. The 5K problems of MATH were distributed by 4 groups randomly, and each group contains 2,3,4,and 5 distinct answer options. We report PRIN for GPT4 for this question.

Results. Figure 5 examines the impact of the number of answer options on PRIN. Increasing the number of options tends to raise PRIN scores, yet

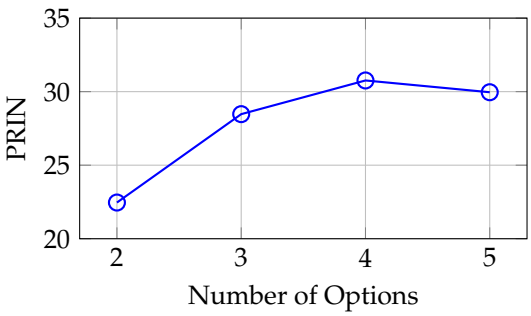

Figure 5: PRIN score vs. #option

models still function within acceptable error limits. This trend underscores the added complexity introduced by a greater variety of options and highlights areas where ongoing algorithm and model improvements are necessary. These findings emphasize the challenges faced by language models in complex decision matrices and open promising avenues for future enhancements in AI development and deployment.

5 Conclusion

This study provides a comprehensive analysis of `Prompt-Reverse Inconsistency` in LLMs, using diverse tasks and models to explore underlying challenges and potential solutions. By addressing six key questions, our findings stress the importance of integrating reasoning paths and adapting model architectures to optimize performance and reliability. As AI models become increasingly integral across domains, our research underlines the necessity of embedding `PRIN` as a foundational element in model development, ensuring their applicability across diverse, challenging scenarios.

Acknowledgement

We would like to sincerely appreciate the anonymous reviewers from OpenReview for their thoughtful insights and constructive suggestions. We are especially grateful to Professor Lili Mou from the University of Alberta for his valuable comments and for posing insightful questions that helped broaden our perspectives. We also deeply appreciate Ibraheem Moosa, Renze Lou, Zhuoyang Zou, Hongchao Fang, and Arshan Dalili for their helpful feedback and suggestions, which played an important role in refining and polishing the final version of this paper.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms, 2019. URL https://arxiv.org/abs/1905.13319.

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. LLM stability: A detailed analysis with some surprises. *CoRR*, abs/2408.04667, 2024. doi: 10.48550/ARXIV.2408.04667. URL https://doi.org/10.48550/arXiv.2408.04667.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL https://arxiv.org/abs/2303.12712.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models, 2021. URL https://arxiv.org/abs/2102.01017.

Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. What did I do wrong? quantifying llms' sensitivity and consistency to prompt engineering. *CoRR*, abs/2406.12334, 2024. doi: 10.48550/ARXIV.2406.12334. URL https://doi.org/10.48550/arXiv.2406.12334.

Sarik Ghazarian, Yidong Zou, Swair Shah, Nanyun Peng, Anurag Beniwal, Christopher Potts, and Narayanan Sadagopan. Assessment and mitigation of inconsistencies in llm-based evaluations. 2024.

Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei, and Wenpeng Yin. Robustness of learning from task instructions, 2023. URL https://arxiv.org/abs/2212.03813.

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, 2021. URL https://arxiv.org/abs/2103.03874.

Myeongjun Erik Jang and Thomas Lukasiewicz. Consistency analysis of chatgpt, 2023. URL https://arxiv.org/abs/2303.06273.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent AI generations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, pp. 2518–2531. ACM, 2024. doi: 10.1145/3630106.3662681. URL https://doi.org/10.1145/3630106.3662681.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2024. URL https://arxiv.org/abs/2308.05374.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, et al. Aaar-1.0: Assessing ai's potential to assist research. *arXiv preprint arXiv:2410.22394*, 2024.

Reid McIlroy-Young, Katrina Brown, Conlan Olson, Linjun Zhang, and Cynthia Dwork. Order-independence without fine tuning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/85529bc995777a74072ef63c05bedd30-Abstract-Conference.html.

Meta. Build the future of ai with meta llama 3, 2024. URL https://llama.meta.com/llama3/. Accessed: 2024-06-07.

OpenAI. Gpt-4 technical report, 2023.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. Condaqa: A contrastive reading comprehension dataset for reasoning about negation. *arXiv preprint arXiv:2211.00295*, 2022.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling, 2025. URL https://arxiv.org/abs/2408.17017.

Julian Junyan Wang and Victor Xiaoqi Wang. Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks, 2025. URL https://arxiv.org/abs/2503.16974.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/forum?id=1PL1NIMMrw.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Wentao Ye, Mingfeng Ou, Tianyi Li, Yipeng chen, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, and Junbo Zhao. Assessing hidden risks of llms: An empirical study on robustness, consistency, and credibility, 2023. URL https://arxiv.org/abs/2305.10235.