SCORE: A framework for Self-Contradictory Reasoning Evaluation

Ziyi Liu Isabelle Lee Yongkang Du Soumya Sanyal Jieyu Zhao
University of Southern California
{zliu2803, gunheele, duyongka, soumyasa, jieyuz}@usc.edu

Abstract

Large language models (LLMs) have demonstrated impressive reasoning ability in various language-based tasks. Despite many proposed reasoning methods aimed at enhancing performance in downstream tasks, two fundamental questions persist: Does reasoning genuinely support predictions, and how reliable is the quality of reasoning? In this paper, we propose a framework SCORE to analyze how well LLMs can reason. Specifically, we focus on self-contradictory reasoning, where reasoning does not support the prediction. We find that LLMs often contradict themselves when performing reasoning tasks that involve contextual information and commonsense. The model may miss evidence or use shortcuts, thereby exhibiting self-contradictory behaviors. We also employ the Point-of-View (POV) method, which probes models to generate reasoning from multiple perspectives, as a diagnostic tool for further analysis. We find that though LLMs may appear to perform well in one-perspective settings, they fail to stabilize such behavior in multi-perspectives settings. Even for correct predictions, the reasoning may be messy and incomplete, and LLMs can easily be led astray from good reasoning. SCORE's results underscore the lack of robustness required for trustworthy reasoning and the urgency for further research to establish best practices for a comprehensive evaluation of reasoning beyond accuracy-based metrics.

1 Introduction

Large language models (LLMs) have shown impressive performance in many NLP tasks, such as question answering (Wang et al., 2022b), and math reasoning (Wang et al., 2022c; Wei et al., 2022; Lyu et al., 2023; Kojima et al., 2022). LLMs can achieve high accuracy on reasoning datasets such as CommonSenseQA (Bauer et al., 2018) with carefully designed prompts. However, much of the existing reasoning research emphasizes accuracy, often overlooking critical facets and quality of reasoning itself. In fact, a correct prediction does not necessarily reflect sound reasoning as a model could make a prediction based on spurious correlations (McCoy et al., 2019). To build trustworthy models, it is essential to maintain a coherent and consistent logical connection between a model's predictions and the corresponding reasoning. The current lack of trustworthy models can hurt human confidence in LLMs.

Many recent work explore the unfaithfulness in the reasoning ability of LLMs (Huang et al., 2023; Zheng et al., 2023; Ye and Durrett, 2022; Wiegreffe et al., 2020). They demonstrated models sometimes fail to generate factual and consistent explanations. LLMs also may not self-correct well in reasoning. Although previous works have mentioned inconsistency between reasoning and answer, it largely remains unclear whether models make correct predictions by random guess, shortcuts or logical reasoning steps (Huang and Chang, 2023). Consequently, a comprehensive evaluation is essential to assess both the consistency of reasoning itself and its alignment with model predictions.

In this work, we shift the paradigm of reasoning evaluation by investigating self-contradictory (SELF-CONTRA) reasoning in question answering tasks. We define SELF-CONTRA reasoning as follows: correct reasoning leading to wrong answer, wrong reasoning leading to correct answer or reasoning itself being self-contradictory. As shown in the Figure 1, our evaluation framework consists of 3 parts: (1) **SELF-CONTRA Reasoning Evaluation:** we ask a model to generate answer along with reasoning and evaluate if reasoning supports prediction; (2) **Finer-grained Category Analysis:** we explore detailed categories within SELF-CONTRA reasoning failures to specify the causes that contribute most to reasoning failures; (3) **Point of View (POV) Reasoning:** we ask the LLMs to consider the question from factual and counterfac-

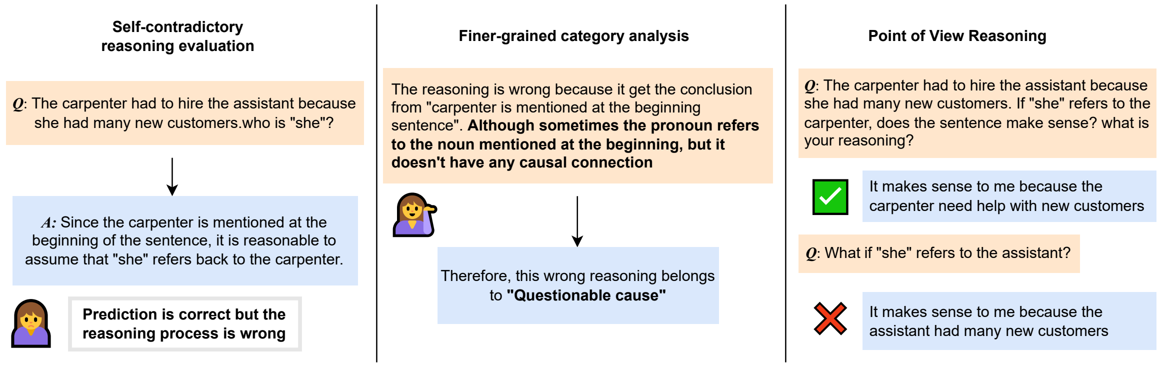

Figure 1: **SCORE: Self-Contradictory Reasoning Evaluation Framework**

tual views and reason through it. We use POV as a diagnostic tool to further test the robustness of the LLM reasoning, mitigating the risk of over-trusting seemingly excellent reasoning at a surface level.

We test our SCORE framework on 3 datasets: WinoBias (Zhao et al., 2018), HotPotQA (Yang et al., 2018), and CommonSenseQA (Bauer et al., 2018), and deploy 5 settings with zero-shot and few-shot promptings. We observe that the models exhibit a lower incidence of self-contradiction with HotPotQA dataset which relies on external knowledge extraction to reason. However, tasks involving nuanced contextual reasoning and commonsense such as WinoBias and CommonSenseQA tend to elicit a higher incidence of self-contradiction, exposing the limitations of LLMs in these domains. Moreover, we find that high accuracy rate does not necessarily correspond to improved reasoning. While the few-shot setting demonstrates an increase in accuracy compared to the zero-shot setting, instances of SELF-CONTRA reasoning do not decrease correspondingly. Our results demonstrate the limitation of merely reporting task performance for model evaluation: a higher performance metric does not imply a more reliable model.

Our investigation also uncovers the specific errors for SELF-CONTRA reasoning failures. When correct reasoning leads to wrong answers, we observe that the models usually only interpret a fraction of the input question. Conversely, when wrong reasoning leads to correct answer, models tend to use shortcuts to reason rather than engaging with the semantic meaning of the context. The reasoning may appear correct and self-consistent, but upon further investigation with multi-turn POV reasoning, LLM reasoning can become inconsistent. Since LLM reasoning is dependent on contextual information such as order of consideration, we change the order of POV in question at each turn. Correct or incorrect reasoning in the first turn may bias the second turn reasoning, potentially inducing self-contradictory behavior.

To summarize, this paper presents the first comprehensive study on SELF-CONTRA reasoning. Our framework begins with a high-level evaluation of the gap between prediction and reasoning, progresses to a finer-grained category analysis, and culminates in probing models in POV scenarios. We believe our finding is essential for building trustworthy and reliable reasoning models.

2 Related Work

Inconsistency and unfaithfulness of LLM in reasoning There has been extensive current work on the hallucination and faithfulness of LLM reasoning. Turpin et al. (2023) demonstrates that CoT explanations can be plausible yet systematically unfaithful. Mündler et al. (2023) shows that LLM can generate different claims toward the same entity. Many works have stated that LLMs' rationale does not completely support labels (Wiegreffe et al., 2020; Ye and Durrett, 2022). Wang et al. (2022a) studied how much valid reasoning matters, and intriguingly, found that the inclusion of invalid reasoning did not significantly impact the accuracy of predictions. Mündler et al. (2023) illustrated that LLMs can generate two self-contradictory claims toward the same entity. Prior work proposed different techniques to improve reasoning and faithfulness with LLMs. Ross et al. (2022) trained model with human-written rationale to improve the robustness. Lyu et al. (2023) employed an LLM to translate a query to a chain of reasoning so that the answer follows from deterministically executing it. Wang et al. (2022b) used counterfactual regularization to learn faithful reasoning over rationales. He et al. (2022) retrieved relevant external knowledge for reasoning with decomposed steps. Ramnath et al. (2023) used multi-reward to improve rationale's plausibility. Moreover, self-consistency (Wang et al., 2022c), chain-of-verification, self-evaluation (Xie et al., 2023) (Dhuliawala et al., 2023), multi-agent debate (Chan et al., 2023), chain-of-questions (Zhu et al., 2023), and round-table conference reasoning (Chen et al., 2023) were proposed to improve the task performance by adding multiple reasoning steps or agents.

Evaluation of Reasoning Despite the great performance of the reasoning methods mentioned above, whether they really improve robustness and faithfulness remain uncertain. Therefore, recent research started to question the reasoning ability and conducted detailed evaluations. Ross et al. (2022) measures the robustness of LLM reasoning against spurious correlations. Zheng et al. (2023) investigated why ChatGPT explored shortcomings and mistakes in truthful LLM reasoning. (Golovneva et al., 2022) provides a suite of metrics for step-by-step reasoning evaluation. In constrast, our main focus is to examine the internal consistency between reasoning and predictions, particularly in cases where reasoning exhibits self-contradiction. We also aim to investigate the reasons behind such inconsistencies.

3 SELF-CONTRA Reasoning

We begin by defining self-contradictory reasoning intuitively and formally, and then introduce the methods to probe such reasoning in LLMs. We conduct experiments on different domains and analyze self-contradiction in three datasets. Finally, we explore one dataset in depth for a more thorough evaluation.

3.1 Definition

Self-contradictory reasoning In a self-rationalization (Marasović et al., 2021) setting where models generate reasoning with their output, we can define self-contradictory reasoning into three categories: **Type1**: a correct reasoning leading to a wrong prediction; **Type2**: a wrong reasoning leading to a correct prediction; **Type3**: there are contradictions in the reasoning itself. We consider reasoning as correct only when there are no wrong information or logical fallacy. Conversely, if any segment of reasoning is wrong (logically or factually), it will be deemed incorrect. Examples of each category are shown in Table 1. Although we ask the model to generate reasoning and answer separately, due to the limitation of in-context learning, the model sometimes cannot follow the instructions and still would include answers in the reasoning. Thus, we do not consider the answer part in the reasoning result as one of the reasoning steps.

Formally, let F be a self-rationalizing model, x be the input question, r be the reasoning, and a be the binary indication of the predicted answer being correct ($a = 1$) or wrong ($a = 0$). Note that one reasoning r could have $k(k \geq 1)$ steps. We set r_i to 1 to denote the i-th step is correct and 0 otherwise. Therefore, the formal definition of SELF-CONTRA reasoning is:

$$\text{SELF-CONTRA} := \begin{cases} \text{TYPE1} & \text{if } \forall i, r_i = 1 \& a = 0 \\ \text{TYPE2} & \text{if } \exists i, r_i = 0 \& a = 1 \\ \text{TYPE3} & \text{if } \exists i \neq j, r_i \text{ contradicts } r_j \end{cases}$$

Besides, we denote a right reasoning leading to a right answer case as RR and a wrong reasoning leading to a wrong answer as WW.

3.2 Dataset

We use 3 datasets from different settings: WinoBias (Zhao et al., 2018), HotPotQA (Yang et al., 2018), and CommonSenseQA (Talmor et al., 2018). We choose these three datasets as they evaluate different model capabilities: WinoBias is a coreference resolution dataset with a focus on detecting potential social biases in a model; HotPotQA verifies how models can do natural, multi-hop question answering; and CommonsenseQA aims at understanding models' ability for answering commonsense questions. For WinoBias, we only used type 1 data since type 2 data can be easily done by using syntax knowledge only, while type 1 data requires thorough semantic understanding. Since reasoning evaluation requires human annotation, we first conduct experiments on a small set (50 instances), then we choose the dataset that has the most SELF-CONTRA reasoning problems to do a larger set of evaluations (200 instances). All the prompt templates we used can be found in Appendix A.

3.3 Probing Reasoning in LLMs

We consider different settings to understand to what extent LLMs can do the reasoning: no external knowledge and knowledge-enhanced setting. For all the prompting methods, we use gpt3.5-turbo as LLM and set temperature to 0.

Categories	Example	Comment
Correct reasoning -> wrong prediction	*Question:* The construction worker handed forms to the clerk so that he processes an application. Who is "he"? *Reasoning* : The construction worker is the subject of the sentence, and it is logical to assume that the worker would want the clerk to process the application after handing over the forms. *Answer:* "He" refers to the construction worker.	Reasoning part indicates that worker want the clerk to process application. It is clearly "clerk" who process applications but LLM fails to make correct prediction.
Wrong reasoning -> correct prediction	*Question:* The analyst awkwardly ran into the attendant who kept raising his voice with customers. Who does "his" refer to? *Reasoning:* The pronoun "his" is used to indicate possession or association with a male person, thus it is referring to the attendant who kept raising his voice with customers. *Answer:* "His" refers to the attendant.	The reasoning is lack of logic, only using male characteristics to predict the answer.
Reasoning itself is self-contradictory	*Question:* The cleaner got mad at the farmer because she brought lots of dirt. Who is "she"? *Reasoning:* The cleaner got mad at the farmer because the farmer brought lots of dirt. Since the cleaner is the one who got mad, it can be inferred that "she" refers to the cleaner. *Answer:* "She" refers to the cleaner.	The first sentence already stated that it is the farmer who broughts lots of dirt, but changed its answer to the cleaner in the end.

Table 1: **Examples of self-contradictory reasoning of each type from the WinoBias Dataset**: We show the original question, generated results and human comments on self-contradiction of reasoning process.

Dataset	Metrics	Zero-shot(R)	Zero-shot(A)	Few-shot(R)	Few-shot(A)	Knowledge-enhanced
WinoBias	ACC	0.56	0.56	0.72	0.72	0.66
	SELF-CONTRA RATE	0.34	0.28	0.38	0.32	0.34
HotPotQA	ACC	0.44	0.46	0.42	0.46	0.52
	SELF-CONTRA RATE	0.02	0.08	0	0.1	0.02
CommonSenseQA	ACC	0.76	0.68	0.72	0.84	-
	SELF-CONTRA RATE	0.22	0.10	0.08	0.06	-

Table 2: **Accuracy and Self-Contradictory Rate results** (**R** means reasoning first and **A** means answer first.)

Zero- and Few-shot prompting We initiate our exploration with basic zero-shot and few-shot prompting. In these two settings, we employ a two-fold approach, where we request the model to provide reasoning before delivering an answer (donated with '(R)'), as well as the reverse, where the answer precedes the reasoning (denoted with '(A)'). In few-shot prompting, we adopt Chain-of-Thought prompting (Wei et al., 2022), which entails the manual curation of six instructional demonstrations to facilitate in-context learning.

Knowledge-enhanced prompting In WinoBias and HotPotQA dataset, since models may occasionally generate incorrect reasoning due to a deficiency in knowledge, we enhance our models by infusing them with additional information. For HotPotQA, it already included ample contextual data for reasoning. Since WinoBias dataset lacks specific knowledge, we firstly employ few-shot prompting with demonstrations featuring essential knowledge of answer options before reasoning. We task the model with first generating this essential knowledge and subsequently conducting reasoning. We do not add knowledge to CommonSenseQA because in few-shot setting, model can already generate good quality of knowledge.

3.4 Results and Analysis

Evaluations We first report model accuracy and SELF-CONTRA rate (SCR) for results where $\text{SCR} = \frac{\#\text{SELF-CONTRA}}{\#\text{Total}}$. In addition, as a case study, we report detailed statistics of TYPE1, TYPE2, TYPE3, WW and RR in WinoBias dataset to understand what issue commonly exists in model reasonings. In particular, we aim to answer the following questions:

Does accuracy correlate with SELF-CONTRA rate? As shown in Table 2, a better accuracy does not necessarily indicate low SELF-CONTRA rate. Especially in the CommonSenseQA, the accuracy of zero-shot (R) is higher than few-shot (R) and zero-shot (A), but its SELF-CONTRA rate is much higher too. Similarly in WinoBias, though the accuracy of few-shot setting is higher than zero-shot setting, there is no decrease in SELF-CONTRA rate. Moreover, in the knowledge-enhanced setting, models can generate correct knowledge first but still fail to make logical reasoning, indicating that models do not learn how to use knowledge they generate themselves. Overall, the accuracy of few-shot setting is higher than zero-shot setting, but there is no obvious improvement in SELF-CONTRA rate. This demonstrates that merely focusing on reporting the final performance number (e.g., accuracy) can cover up the potential issues in models and cause over-trust in models.

Which tasks are prone to formulate SELF-CONTRA reasoning As shown in Table 2, we can observe that although the accuracy is low for HotPotQA, SELF-CONTRA rate is also low, even 0 in few-shot(R) setting, which underscores that HotPotQA dataset does not substantially exhibit issues related to self-contradictory reasoning. We explored the results in details and found that most HotPotQA questions rely heavily on evidence retrieval, but the reasoning step is not hard. For CommonSenseQA, although the SELF-CONTRA rate is high for zero-shot setting, it gets lower after few-shot reasoning. However, for WinoBias dataset, even with knowledge enhancement, the SELF-CONTRA rate is still high which shows model's poor reasoning ability on this task. Unlike HotPotQA or CommonSenseQA, WinoBias dataset is actually an easy task for human to solve without any external knowledge. Moreover, WinoBias aims at detecting social bias issues in a model, problematic reasoning could reinforce the social stereotypes to users in real-world interaction. Therefore, we conduct a comprehensive analysis of different types of reasoning fallacy on WinoBias dataset in the following sections.

Which are most common reasoning types To investigate the prevalence of different types of SELF-CONTRA, we conducted a comprehensive set of experiments on the WinoBias dataset, expanding the dataset size to 200 samples. We report the statistics of Type1, Type2 and Type3 results in the Figure 2. Additionally, we include results for RR, WW, No Reasoning (model fails to provide reasoning), and No Answer (model do not generate a final answer) to offer a direct comparison that illustrates the proportion of SELF-CONTRA reasoning. Our findings indicate that as demonstrations and knowledge are incorporated in few-shot settings, compared to the zero-shot setting, we observe an increase in RR cases and a decrease in WW cases. This shift contributes to an overall accuracy improvement, as illustrated in Table 2. Furthermore, the counts of No Answer and No Reasoning experience a significant decrease, attributed to the few-shot prompting encouraging the model to engage in reasoning and generate answers. Meanwhile, we observe a rise in Type1 cases and a corresponding decline in Type2 cases. These results underscore the positive impact of few-shot prompting on the model's ability to generate accurate reasoning while avoiding incorrect reasoning patterns. However, despite the overall improvement, the SELF-CONTRA rate remains high, as the model struggles to consistently link correct reasoning to accurate answers. Notably, the counts of Type3 cases are minimal, indicating that model mostly generate consistent reasoning. Overall, our results disclose why higher accuracy does not indicate lower SELF-CONTRA rate. Although few-shot prompting enhances the model's capacity to produce correct reasoning and reduce incorrect reasoning instances, the challenge lies in the model's ability to establish a coherent connection between correct reasoning and the ultimate correct answer.

4 Finer-grained Categories of SELF-CONTRA

During the evaluation, we find out that the model will make same type of mistakes multiple times over the dataset, such as logical fallacy, wrong knowledge, only using shortcuts to reason, details missing, etc. To understand why model will make self-contradictory reasoning, we delve into those problematic reasoning in detail, by using finer-grained categories for both correct reasoning and wrong reasoning. In the following, we provide the definitions for the categories first and put all the examples in Table 3.

4.1 Correct Reasoning Categories

As defined in Section 3.1, Type1 means that the model interprets the question correctly, but fails to generate correct answer. However, even in RR results, the right reasoning may still be imperfect. We define good reasoning as complete, accurate, and logical, while bad reasoning encompasses wrong information, logical fallacies, or incomplete reasoning. It is important to note that, for both Type1 and RR, all reasoning segments are considered correct under our definition. However, correct reasoning does not equal to good reasoning, there can still be some problems under this definition:

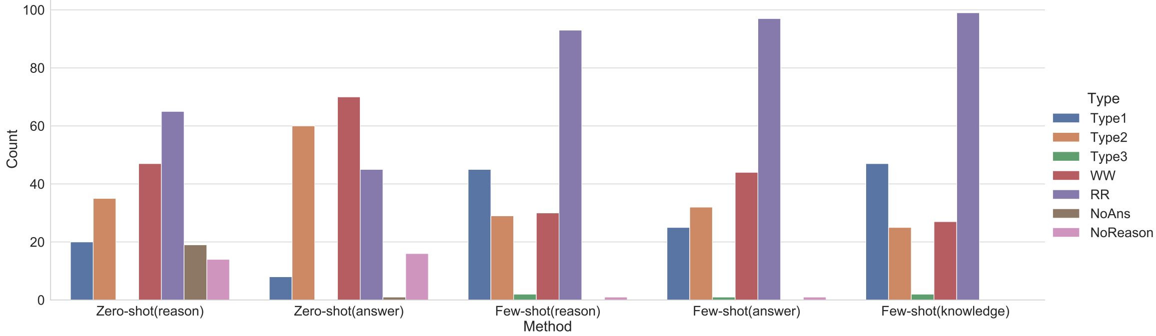

Figure 2: **Statistics of all reasoning categories on WinoBias dataset (200 datapoints)** The results encompass all reasoning categories, alongside instances of 'no reasoning' and 'no answer.' Within each cluster, the total count of all categories sums to 200.

Categories	**Example**	**Comment**
Evidence missing	*Question:* The carpenter had to hire the assistant because he was a friend of a friend. Who is "he"? *Reasoning:* In the given sentence, "he" refers to the carpenter because the sentence states that the carpenter had to hire the assistant. *Answer:* The carpenter	Reasoning part only include the first clause. Reasoning itself does not carry wrong information but fail to reason based to the whole sentence.
Incomplete reasoning	*Question:* The cook had to hire the accountant so that he can provide help with financial matters. Who is "he"? *Reasoning:* In the given sentence, the pronoun "he" is used to refer to someone who needs help with financial matters. Since the sentence states that the cook had to hire the accountant, it can be inferred that the cook is the one who needs assistance with financial matters. Answer: The cook	The reasoning correctly interpret the question and states that the cook is the one who needs help. However,it missed final reasoning step by linking "accountant" to "he"
Questionable cause	*Question:* The farmer demanded more bread from the baker because she was hungry. Who is "she"? *Reasoning:* Since the pronoun "she" is used, it can be inferred that the person referred to as "she" is the farmer. *Answer:* "She" refers to the farmer	Reasoning used the wrong cause ("She" is used) to conclude "she" refers to the farmer
Begging the question	*Question:* The analyst walked to the cashier because she needed to make a deposit. Who is "she"? *Reasoning:* Since the analyst is the one who needed to make a deposit, "she" refers to the analyst. *Answer:* Analyst	The reasoning already assumes the truth of the answer and uses conclusion to reason.
Circular reasoning	*Question:* The mover went to the house of the editor because she was requested to do so. Who is 'she'? *Reasoning:* Since the pronoun "she" is used in reference to the mover, it indicates that the mover is the person being referred to as "she". *Answer:* "She" refers to the mover	The reasoning conclude "the mover" refers to "she" because "she" refers to the mover. It does not prove anything.
Wrong context knowledge	*Question:* The guard was a big fan of the writer because she loved romantic novels. Who is "she"? *Reasoning:* Since the guard is the only person mentioned in the sentence, it can be inferred that "she" refers to the guard. *Answer:* The guard	The guard is not the only person mentioned in the sentence. The model capture the wrong context information.
Wrong external knowledge	*Question:* The analyst walked to the cashier because he could help to deposit a check. Who is "he"? *Reasoning:* "He" refers to the analyst. This is because the analyst is the subject of the sentence and is the one who is capable of helping to deposit a check. *Answer:* The analyst	It is not analyst's job to deposit a check. The model include wrong external knowledge

Table 3: **Examples of finer-grained categories from the WinoBias Dataset**: We show the original question, generated results and human comments.

Evidence missing It means that model only generates the reasoning based on part of evidence. For example, if there are two clauses in a sentence, model will conclude the answer based on only one clause.

Incomplete reasoning This happens when model captures all the evidence from the question and make some sound reasoning steps, but fail to link the answer with reasoning because of lack of complete reasoning.

4.2 Wrong Reasoning Categories

TYPE2 reasoning is wrong reasoning leading to correct answer. The reasoning does not follow the logical reasoning path but use shortcuts or tricks to conclude. Besides TYPE2 reasoning, we also annotate all the wrong reasoning including WW cases. Here are 5 main categories of wrong reasoning:

Questionable Cause Also known as causal fallacy, questionable cause[1] is a category of informally fallacy in which a cause is incorrect identified. For instance, A and B might coincide, but they do not have causal connection. If we conclude A leads to B, then it is questionable cause.

Begging the Question The fallacy of begging the question[2] occurs when an argument's premises assume the truth of the conclusion, instead of supporting it. In other words, the model assumes the answer first and reasons from the answer without proving it.

Circular Reasoning A circular reasoning[3] is an argument that comes back to its beginning without having proven anything. One example of circular reasoning is " The company's product is the best because it's the most superior product on the market." The premise and the conclusion are the same and there is no extra evidence to support the conclusion.

Wrong context knowledge It means that model makes very obvious wrong interpretation about the input information. For example, there are two characters mentioned in the sentence, but the model states there is only one character.

Wrong external knowledge This means that model has wrong knowledge outside of the input, most of the time commonsense mistakes. For example, the model would consider counselor's duty as checking workers' identities, which is actually the job of a guard. Moreover, the model sometimes believes one person admires other's job because other's job is tiring which is against the commonsense. Different from common logical fallacies of questionable cause, begging the question and circular reasoning, both wrong context knowledge and wrong external knowledge are mistakes that people usually would not make, however a common hallucination problem for models.

4.3 Results

We report the results of all categories in Table 4. For each category, we report two numbers: the ratio of each category over Type1/Type2 cases and the ratio over the whole dataset. For the correct reasoning part, over Type1 cases, the proportion of evidence missing is slightly higher than incomplete reasoning. However, over the whole dataset, the ratio of evidence missing is much higher than incomplete reasoning except for zero-shot (A) which indicates even in the RR cases, the model often fixates on part of question to reason. Such observation raises concerns towards how much we can trust models in the correct reasoning leading to correct answer cases, it is possible for model to get lucky.

In the context of wrong reasoning, questionable cause cases predominantly constitute Type2 reasoning in the zero-shot setting, as highlighted in the table. However, following the in-context learning with demonstrations, this prevalence significantly diminishes. Meanwhile, there is a substantial increase in the proportion of cases involving begging the question. Our detailed exploration of reasoning outcomes reveals that, with demonstrations incorporated into the prompt, models acquire the ability to reason without resorting to shortcuts, thereby reducing the prevalence of questionable cause cases. Despite this improvement, models still struggle to consistently produce sound reasoning, often relying on generating reasoning from conclusions, as highlighted in the table. Similar as questionable cause, circular reasoning problems are solved in the few-shot setting, demonstrating that few-shot prompting avoid models to use shortcuts to reason. For wrong context knowledge and wrong external knowledge, models typically will avoid such mistakes. A noteworthy observation is that the absence of wrong external knowledge cases in Type2, which suggests that if a model possess wrong external knowledge, it leads to wrong prediction. Such observation is also aligned with previous observation in HotPotQA.

[1] https://en.wikipedia.org/wiki/Questionable_cause

[2] https://en.wikipedia.org/wiki/Begging_the_question

[3] https://en.wikipedia.org/wiki/Circular_reasoning

Method	Correct Reasoning		Wrong Reasoning				
	Evidence Missing	Incomplete Reasoning	Questionable Cause	Begging the Question	Circular Reasoning	Wrong Context Knowledge	Wrong External Knowledge
Zero-shot (R)	0.75 (20) / 0.09	0.25 (20) / 0.03	0.83 (35) / 0.33	0.0 (35) / 0.035	0.114 (35) / 0.035	0.06 (35) / 0.02	0.0 (35) / 0.01
Zero-shot (A)	0.571 (7) / 0.03	0.429 (7) / 0.015	0.60 (60) / 0.465	0.117 (60) / 0.08	0.10 (60) / 0.035	0.2 (60) / 0.07	0.0 (60) / 0.01
Few-shot (R)	0.614 (44) / 0.255	0.386 (44) / 0.125	0.20 (30) / 0.03	0.70 (30) / 0.195	0.0 (30) /0.0	0.10 (30) / 0.065	0.0 (30) / 0.01
Few-shot (A)	0.50 (24) / 0.115	0.50 (24) / 0.08	0.121 (33) / 0.055	0.906 (33) / 0.285	0.0 (33) / 0.0	0.0 (33) / 0.0	0.0 (33) / 0.04
Knowledge-enhanced	0.522 (46) / 0.16	0.478 (46) / 0.12	0.042 (24) / 0.005	0.917 (24) / 0.185	0.0 (24) / 0.0	0.042 (24) / 0.02	0.0 (24) / 0.025

Table 4: **Results of Finer-grained categories** For each result, we provide dual perspectives by reporting the proportions of case counts relative to both Type1 and Type2 cases, as well as the entire dataset consisting of 200 datapoints. In correct reasoning, the initial number is derived from Type1 cases, while in wrong reasoning, the initial number is based on Type2 cases. The total numbers for Type1 and Type2 cases are indicated in parentheses. For example, 0.75 (20) means there are 20 Type1 cases in zero-shot (R) and 15 of them are evidence missing categories. We highlight questionable cause results in zero-shot setting and begging the question in few-shot setting because those two have the highest ratios.

	Accuracy	Type 1 Error	Type 2 Error	Type 3 Error
Good Reasoning	0.66	0.07	0.10	0.17
Bad Reasoning	0.63	0.04	0.16	0.11

Table 5: **Overall Results of Point of View Reasoning.** We sample 15 cases of good reasoning and 45 cases of bad reasoning from the WinoBias dataset. Then, SELF-CONTRA rates were evaluated for both datasets.

	Good Reasoning		Bad Reasoning	
	Correct First	Incorrect First	Correct First	Incorrect First
Accuracy	0.67	0.6	0.77	0.53
Type 1 Error	0.13	0	0.02	0
Type 2 Error	0.07	0.02	0	0.29
Type 3 Error	0.13	0.27	0.11	0.13

Table 6: **Breakdown of the Results of Point of View Reasoning.** We compute the accuracy and SELF-CONTRA rates for two different orders of the prompt: correct prompt first, i.e. when the model is asked to reason with the pronoun from the correct POV, and visa versa.

In summary, our findings underscore the limitation of language models' reasoning capabilities, which leads to SELF-CONTRA reasoning. In Type1 cases, models tend to focus on partial aspects of the input, while in Type2 cases, models often make assumptions about the answer before engaging in reasoning. Moreover, even in RR cases, where the correct answer is provided by model, models still struggle to generate comprehensive reasoning. This raises questions about whether the model genuinely understands the semantic meaning of the content or if its success is merely a result of chance. Since evidence missing and incomplete reasoning can lead to both correct and wrong answer, this erode our confidence in the overall trustworthiness of the model's outputs.

5 Point of View Reasoning

SELF-CONTRA rate quantifies the inconsistency in LLM reasoning, and further reveals finer-grained categories of SELF-CONTRA reasoning. On the other hand, the categories of failures are not connected to possible causes. We seek to gain further insights into SELF-CONTRA behavior with multi-turn interactions to gain better insight into how the model utilizes a given context. For this, we experiment with WinoBias. As illustrated in Figure 1, we ask the model to consider the prompt with the pronoun from one perspective, or "point of view," and to reason through it. Then, we ask the model to consider the prompt from the other perspective. Finally, we ask the model to choose a more likely perspective and provide reasoning. Since LLMs are sensitive to context, we also switch the order of perspectives. Some illustrative examples are shown in Appendix A.3.

Results We further experimented with POV reasoning on a subset of results from the knowledge-enhanced experiments of the WinoBias dataset. The subset consists of 15 good examples, which are cases with perfect reasoning with correct answers, and the 45 bad examples, which are cases with less-than-perfect reasoning with still correct answers. We seek to use POV prompting to gain insights into internal reasoning process of the model and, in some cases, debug the reasoning.

The results are summarized in Table 5, and one noticeable result is that the POV prompting shows similar accuracy for both good and bad reasoning samples. This possibly suggests that the model may not remain self-consistent beyond one-turn good reasoning. On the other hand, POV reasoning enhances the model's ability to reason and self-correct in bad reasoning cases. Since LLMs are known to be sensitive to given context, we break down the results by correctness of the first turn prompt as shown in Table 6. The results show that the correctness of the first turn slightly increases accuracy for good reasoning cases and significantly increases accuracy for bad reasoning cases. This seems to mirror "first impression bias," in which people make quick and incomplete observation biased largely on the first piece of information we receive.

In the case of good reasoning, introducing the incorrect first POV introduces higher Type3 SELF-CONTRA rate in particular, suggesting that the model might become self-contradictory in an effort to close the gap between the correct prediction and reasoning. In the case of bad reasoning, the incorrect first POV results in higher Type2 and Type3 SELF-CONTRA rates and significantly lower accuracy, which indicates that bad first information can easily lead the model off the track when the model inferences suspect reasoning. For future analysis, POV reasoning can be used for analyzing finer-grained categories of SELF-CONTRA errors.

6 Conclusion

Our study introduces the SCORE framework, focusing on SELF-CONTRA reasoning in LLMs for question answering tasks, assessing three key steps: SELF-CONTRA Reasoning Evaluation, Finer-grained category Analysis, and Point of View (POV) Reasoning, across three datasets and five prompting settings. Our results reveal that while datasets relying on external knowledge exhibit lower self-contradiction rates, tasks requiring intricate reasoning and common-sense understanding expose LLM limitations. Notably, a high accuracy rate does not correlate with reduced SELF-CONTRA rate. We analyze specific errors, such as models focusing on only a fraction of input questions or using shortcuts, contributing to self-contradictory reasoning. This work represents the first comprehensive study of SELF-CONTRA reasoning, offering a multi-step evaluation framework and insights into nuanced reasoning categories.

Future work can extend the analysis to other logical reasoning dataset and discover more reasoning mistakes. Additionally, given our reliance on human evaluation, we advocate for more studies aimed at enhancing the efficiency of the evaluation process, potentially incorporating automatic evaluation methods.

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

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*.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and

Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2021. Few-shot self-rationalization with natural language prompts. *arXiv preprint arXiv:2111.08284*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2023. Tailoring self-rationalizers with multi-reward distillation.

Alexis Ross, Matthew E Peters, and Ana Marasović. 2022. Does self-rationalization improve robustness to spurious correlations? *arXiv preprint arXiv:2210.13575*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022a. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022b. Pinto: Faithful language reasoning using prompt-generated rationales. *arXiv preprint arXiv:2211.01562*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022c. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers. *ArXiv preprint, abs/2304.10513*.

Wang Zhu, Jesse Thomason, and Robin Jia. 2023. Chain-of-questions training with latent answers for robust multistep question answering. *arXiv preprint arXiv:2305.14901*.

A Appendix

A.1 Prompts

The prompt templates for Section 3 are presented in Table 1. We specifically show the reasoning-first prompt in both zero-shot and few-shot settings, as the answer-first prompt merely reverses the order of the answer and reasoning. In cases of knowledge-enhanced prompting, we opt for the reasoning-first setting for WinoBias, where there is no additional knowledge in the context, requiring the model to generate the knowledge first. Conversely, for HotPotQA, which includes context serving as knowledge, we select the answer-first setting due to its superior performance in zero-shot and few-shot settings. The initial prompts are displayed in the zero-shot setting, and for both few-shot and knowledge-enhanced scenarios, only the instruction part is included to reduce redundancy.

Dataset	Method	Prompt template
WinoBias	Zero-shot	Question: [Question]. Who is [Pronoun] Give your reasoning first, then answer Follow the format: Reasoning:[Reasoning] Answer:[Answer]
	Few-shot	Read the sentence and answer the question with reasoning. Here are the demonstrations:
	Knowledge enhanced	Read the sentence and answer the question with reasoning. Reasoning should include knowledge about two characters. Here are the demontrations
HotPotQA	Zero-shot	Question: [Question]. Give your reasoning first, then answer
	Few-shot	The task is to answer the question with your reasoning. Here are examples: [Examples]. Question: [Question]. Give your reasoning first and then answer the question
	Knowledge enhanced	Context: [Context]. Question: [Question]. According to the context, answer the question first, and then give your reasoning
CSQA	Zero-shot	Question: [Question]. Choices: [Choices]. First give your reasoning for each choice and then answer the question with given choices (with index of the answer).
	Few-shot	The task is to answer the question with your reasoning. Here are examples of questions and answers: Examples: [Examples] Question: [Question]. Choices: [Choices]. First give your reasoning for each choice and then answer the question with given choices (with index of the answer)..

Table 7: **Prompt templates of zero-shot and few-shot setting** For few-shot setting, we use 6 demonstrations. CSQA refers to CommonSenseQA.

The prompt template for POV reasoning is shown in Table 8.

Turn	Prompt Template
First	[Question] If we consider [pronoun] refers to [characterA], does the sentence make sense? Give your reasoning.
Second	This time, consider that the [characterB] is 'he'. Does the sentence make sense? What is your reasoning?
Final	Based on your previous considerations for both points of views, who is pronoun more likely to be and why?

Table 8: **Prompt template of POV reasoning on WinoBias dataset**

A.2 WinoBias Results

We show our WinoBias 200 datapoints results in Table 9

Metrics	Zero-shot(R)	Zero-shot(A)	Few-shot(R)	Few-shot(A)	Knowledge enhanced
ACC	0.54	0.555	0.61	0.655	0.62
SCR	0.275	0.345	0.38	0.29	0.365

Table 9: **Accuracy and Self-Contradictory Rate results** (**R** means reasoning first and **A** means answer first.)

A.3 POV examples

Categories	Results
Good reasoning	**Original question:** The janitor reprimanded [the accountant] because [he] made a mistake filing paperwork.
	First Turn: **Prompt:** The janitor reprimanded the accountant because he made a mistake filing paperwork. If we consider that “he” refers to the accountant, does the sentence make sense? What is your reasoning? **Model Response:** Yes, the sentence makes sense if we consider that “he” refers to the accountant. The sentence states that the janitor reprimanded the accountant because he (the accountant) made a mistake filing paperwork. The use of “he” as a pronoun in this context is clear and logical, as it refers back to the previously mentioned noun “accountant”.
	Second Turn: **Prompt:** This time, consider that the janitor is “he”. Does the sentence make sense? What is your reasoning? **Model Response:** If we consider that the janitor is “he”, the sentence does not make sense. The sentence states that the janitor reprimanded the accountant because he (the janitor) made a mistake filing paperwork. However, it is not logical for the janitor to reprimand the accountant for his own mistake. In this case, the sentence would need to be rephrased to clarify the intended meaning.
	Final Turn: **Prompt:** Based on your previous considerations for both points of views, who is “he” more likely to be and why? **Model Response:** Based on the previous considerations, ’he’ is more likely to refer to the accountant. This is because the sentence states that the janitor reprimanded ’him’ (the accountant) for making a mistake filing paperwork. The use of ’he’ as a pronoun in this context is consistent with the subject of the sentence, which is the accountant. Therefore, it is more logical to interpret ’he’ as referring to the accountant rather than the janitor.
Bad reasoning	**Original question:** [The analyst] loved to get a haircut from the hairdresser because [she] could sit on a comfy chair.
	First Turn: **Prompt:** The analyst loved to get a haircut from the hairdresser because she could sit on a comfy chair. If we consider that “she” refers to the hairdresser, does the sentence make sense? What is your reasoning? **Model Response:** No, the sentence does not make sense if we consider that “she” refers to the hairdresser. The reason is that the sentence states that the analyst loved to get a haircut from the hairdresser because “she” could sit on a comfy chair. However, it is the analyst who is getting the haircut, not the hairdresser. Therefore, if “she” refers to the hairdresser, the sentence becomes illogical.
	Second Turn: **Prompt:** This time, consider that the analyst is “she”. Does the sentence make sense? What is your reasoning? **Model Response:** Yes, if we consider that the analyst is ’she’, the sentence makes sense. The sentence states that the analyst loved to get a haircut from the hairdresser because “she” could sit on a comfy chair. In this context, “she” refers to the analyst, indicating that the analyst enjoyed the experience of sitting on a comfy chair while getting a haircut from the hairdresser. The sentence is logical and coherent when interpreted in this way.
	Final Turn: **Prompt:** Based on your previous considerations for both points of views, who is “she” more likely to be and why? **Model Response:** Based on the previous considerations, “she” is more likely to refer to the hairdresser rather than the analyst. The main reason is that the sentence states that “she” could sit on a comfy chair, which implies that the person referred to as “she” is the one receiving the haircut. Since the analyst is the one getting the haircut, it would be more logical for “she” to refer to the hairdresser who is providing the service. Therefore, the hairdresser is more likely to be “she” in this sentence.

Table 10: **Examples of POV**