Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation From Deductive, Inductive and Abductive Views

Fangzhi Xu[1,2*], Qika Lin[1,2*], Jiawei Han[1,3], Tianzhe Zhao[1,3], Jun Liu[2,3†], Erik Cambria[4]

[1] School of Computer Science and Techonology, Xi'an Jiaotong University
[2] Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering
[3] National Engineering Lab for Big Data Analytics
[4] School of Computer Science and Engineering, Nanyang Technological University
Leo981106@stu.xjtu.edu.cn, qikalin@foxmail.com, tara1208260223@stu.xjtu.edu.cn,
ztz8758@foxmail.com, liukeen@xjtu.edu.cn, cambria@ntu.edu.sg

Abstract

Large Language Models (LLMs) have achieved great success in various natural language tasks. It has aroused much interest in evaluating the specific reasoning capability of LLMs, such as multilingual reasoning and mathematical reasoning. However, as one of the key reasoning perspectives, logical reasoning capability has not yet been thoroughly evaluated. In this work, we aim to bridge those gaps and provide comprehensive evaluations. Firstly, to offer systematic evaluations, this paper selects fifteen typical logical reasoning datasets and organizes them into deductive, inductive, abductive and mixed-form reasoning settings. Considering the comprehensiveness of evaluations, we include three representative LLMs (i.e., text-davinci-003, ChatGPT and BARD) and evaluate them on all selected datasets under zero-shot, one-shot and three-shot settings. Secondly, different from previous evaluations relying only on simple metrics (e.g., accuracy), we propose fine-level evaluations from objective and subjective manners, covering both answers and explanations. Also, to uncover the logical flaws of LLMs, bad cases will be attributed to five error types from two dimensions *Evidence Selection Process* and *Reasoning Process*. The former one includes *evidence selection error* and *hallucination*, while the latter one includes *no reasoning*, *mistakes of reasoning perspectives* and *mistakes during reasoning process*. Thirdly, to avoid the influences of knowledge bias and purely focus on benchmarking the logical reasoning capability of LLMs, we propose a new dataset with neutral content. It contains 3K samples and covers deductive, inductive and abductive reasoning settings. Based on the in-depth evaluations, this paper finally concludes the ability maps of logical reasoning capability from six dimensions (i.e., correct, rigorous, self-aware, active, oriented and no hallucination). It reflects the pros and cons of LLMs and gives guiding directions for future works.

1 Introduction

Large Language Models (LLMs) have recently made great progress in the field of Natural Language Processing (NLP), achieving surprising performances in some complex reasoning tasks. Meanwhile, many works have also evaluated the capability of LLMs from various reasoning perspectives, e.g., multilingual reasoning (Bang et al. 2023), commonsense reasoning (Bian et al. 2023), and mathematical reasoning (Imani, Du, and Shrivastava 2023).

These efforts on evaluating the specific capabilities of LLMs are meaningful, and they can benefit the future directions of researches. As one of the crucial aspects of reasoning ability, logical reasoning capability has also drawn a great deal of interest in previous works (Li et al. 2022; Xu et al. 2022). However, few works focus on the comprehensive evaluation of LLMs from the logical reasoning view. It still remains unclear *whether LLMs are really good logical reasoners?* Our work aims to fill such blanks and give the in-depth evaluation from the following aspects.

Firstly, there lacks systematic and comprehensive analysis of the logical reasoning ability of LLMs. Logical reasoning can be mainly categorized into three types, which are deductive, inductive and abductive types of reasoning. Since deductive, inductive and abductive reasoning together form complete chain of reasoning, it is meaningful to evaluate LLMs from such three perspectives. Previous work like (Liu et al. 2023), has proposed to evaluate the logical reasoning ability of LLMs. But it only focuses on several specific datasets and fails to give in-depth thoughts from the reasoning manners (i.e., deductive, inductive and abductive). (Bang et al. 2023) proposes more comprehensive evaluations of LLMs in various tasks. It well includes logical reasoning datasets into three types. However, it merely tests on 5 datasets with 30 examples of each. In this paper, we bridge the above gaps in systematic and comprehensive ways. From the *systematic* view, all the evaluated datasets are categorized into four reasoning manners, i.e., deductive, inductive, abductive and mixed-forms. The former three are the classical ones involving single reasoning manner. Considering some recent efforts have been made on proposing challenging settings with mixed reasoning manners, we include the category of *mixed-form* for evaluations. From the *comprehensive* view, we include fifteen typical logical reasoning datasets, evaluating on three representative LLMs, i.e., text-davinci-003, ChatGPT (Ouyang et al. 2022a) and BARD (Anil et al. 2023) under zero-shot, one-shot and three-shot settings.

[*]These authors contributed equally.
[†]Corresponding author

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Secondly, it lacks fine-level evaluation of the LLM performances. Current benchmarks only rely on a few objective metrics (e.g., accuracy) to measure the model capability. It may not be sufficient in the case of generative LLMs, since the role of LLMs is not only limited to correctly answer questions but also serves as practical tools, which are required to provide reasoning chains or explanations. Previous works (Qin et al. 2023; Tu et al. 2023) conduct extensive experiments on popular NLP datasets, but they purely report the performance results. Since some LLMs (e.g., ChatGPT) function as the interactive tools for human use, it is necessary to introduce subjective metrics to do fine-grained evaluations. In this paper, we employ four dimensions of metrics, covering *answer accuracy*, *explaination correctness*, *explaination redundancy* and *explaination completeness*. It can provide more meaningful and complete evaluations from both objective and subjective views. Considering those bad cases (i.e., wrong answer or wrong explanation) reflective of obvious logical flaws of LLMs, we further attribute them to several error types from two dimensions of *Evidence Selection* and *Reasoning Process* and give in-depth analysis.

Thirdly, current logical reasoning benchmarks may not offer neutral content for comparing LLMs. This is because LLMs are highly powerful due to their massive training data, which may overlap with popular benchmarks. As a result, testing LLMs on these benchmarks may not be entirely fair, as it can only demonstrate the fitting ability of LLMs rather than their real logical reasoning capability. Furthermore, in current benchmarks, logical reasoning is strongly coupled with text understanding. Therefore, language models may be trained to learn a biased pattern from text, rather than really capture the logical reasoning capability. Some previous works (Huang et al. 2023; Zhang et al. 2023) propose to establish complete benchmarks for LLMs. But few works focus on logical reasoning and fail to attend to content-neutral problem. To narrow this gap, we propose a new dataset named NeuLR, which contains 3K content-neutral samples and covers the deductive, inductive and abductive reasoning types. It is expected to offer a novel perspective for benchmarking logical reasoning ability of LLMs.

Finally, we conclude the extensive performance results of LLMs and form an evaluation system with six key properties, i.e., *Correct*, *Rigorous*, *Self-aware*, *Active*, *Oriented* and *No Hallucination*. The above dimensions can all be quantified from existing evaluation experiments. For deductive, inductive, abductive and mixed reasoning settings respectively, we obtain the ability maps based on the six properties for each LLM. It is meaningful to identify the strengths and weaknesses of LLMs under the four reasoning settings, thus guiding the future directions.

The main contributions of the paper are listed as follows:

(1) To explore real logical reasoning capability of LLMs, this paper provides a systematic, comprehensive and fine-level evaluation from deductive, inductive and abductive views. The in-depth evaluation and analysis fill the blanks and are expected to provoke new thoughts of LLMs.

(2) To offer a systematic view for logical reasoning evaluation, this paper classifies datasets into four reasoning manners, i.e., deductive, inductive, abductive and mixed-forms. In light of the insufficient evaluations of previous works, we include 15 typical logical reasoning datasets, and evaluate on 3 representative LLMs (i.e., text-davinci-003, ChatGPT and BARD) under both zero- and few-shot settings.

(3) Considering the drawbacks in current objective metrics, this paper gives fine-level evaluations including four dimensions i.e., answer accuracy, explain correctness, explain redundancy and explain completeness. To explore the value of bad cases, we attribute them into several error types and find logical flaws of LLMs.

(4) To provide fair evaluations with neutral content and decouple logical reasoning from text understanding, this paper proposes a new dataset named NeuLR. It contains 3K content-neutral samples and covers deductive, inductive and abductive reasoning manners.

(5) In view of the evaluation results, this paper concludes six key properties to measure the logical reasoning capability of LLMs. Furthermore, we derive the ability maps for each LLM under four reasoning settings respectively and propose future directions.

2 Preliminary

Logical reasoning aims to generate logical implications that contain new facts using one-step or multi-step inference based on given premises, i.e., *premise⇒conclusion*. Elements of logical reasoning typically include knowledge facts and logical rules, for example,

- rule: Children of eight years old are all in primary school.
- fact1/premise1: Jordan is a child of eight years old.
- fact2/premise2: Jordan is in primary school.

According to the reasoning classification system of classical logic, there are three major types of logical reasoning: deductive, inductive, and abductive. Based on the above rule and facts, the task of these three reasoning types can be illustrated to predict the remaining one using the given two items.

Deductive Reasoning. Deductive reasoning is the psychological process of drawing deductive inferences that start from the given premises and reason with logical rules or commonsense to obtain *certain* conclusions (Johnson-Laird 1999; Goel 2007). It can be *premise1+rule→premise2*. Its progress generates specific knowledge facts from general counterparts, e.g., *premise2* and *rule* are specific and general knowledge, respectively. Therefore, deductive reasoning is actually a top-down way.

Inductive Reasoning. Distinct from deductive reasoning, inductive reasoning derives general principles from a body of observations which means making broad generalizations based on specific observations (Heit and Rotello 2010; Yu, Zhang, and Wang 2023). For example, an example of inductive reasoning can be *premise1+premise2→rule*, concluding generalized knowledge *rule* that is independent with specific item *Jordan*. Generally, the truth of the conclusion of an inductive argument is *probable* rather than *certain* in inductive reasoning. Thus, inductive reasoning is bottom-up and contrasted with deductive reasoning.

Abductive Reasoning. Formally, abductive reasoning is similar to deductive reasoning which seeks conclusions

Table 1: Details of the selected LLMs. *Affi.* is short for *Affiliation*. *Charge* represents the charges for 1K tokens. *Data* is the latest time of the utilized training data.

Model	Affi.	Charge	Data	Size
text-davinci-003	Open-AI	0.02$	Sep. 2021	175B
ChatGPT	Open-AI	0.002$	Jun. 2021	175B
BARD	Google	Free	Not report	1,560B

from a set of observations. But differently, its target is to generate the simplest and most likely explanation for the given observations (Josephson and Josephson 1996; Walton 2001). So the result is *probable* like in inductive reasoning. An example of abductive reasoning can be *premise2+rule→premise1*.

These three reasoning types involve the underlying patterns of logic. However, many real-life reasoning scenarios may need several inference steps and integrations of at least two of these three types. In this paper, they can be viewed as a more complex reasoning type *mixed*.

3 Evaluation Details

In this section, we will provide the detailed experiment settings in this paper, including evaluated models, testing datasets as well as evaluation metrics.

3.1 Evaluated Models

Because of the rapid emergence of LLMs, it is not realistic to include all LLMs in this paper. Thus, we select three representative ones for evaluation, which are text-davinci-003, ChatGPT and BARD. The details of these three models are listed in Table 1.

Among them, text-davinci-003 (Ouyang et al. 2022b) is the earliest LLM released by OpenAI, which is expected to undertake any language task. For ChatGPT [1], we utilize the version of *GPT-3.5-turbo* for evaluation, which is the most capable and cost-effective version in the GPT-3.5 family. BARD [2] is the latest LLM, which is updated and released by Google in May 2023. Also, it is almost 10× in size compared with the GPT-3.5 family.

3.2 Evaluated Datasets

According to the previous discussion, the evaluation is conducted systematically from deductive, inductive, abductive and mixed views. Therefore, this paper selects 15 popular datasets in logical reasoning and divides them into the above four folds. Table 5 presents the detailed information of these datasets. The selected datasets contain both generative and classification ones and there exist diverse forms of tasks, which illustrate the comprehensiveness of our evaluation. Different from previous works which only use a few dozens of samples, this paper largely extends the amount. Since ChatGPT is one of the most popular LLM for public, we give much focus on it (i.e., EntailmentBank (Dalvi et al. 2021), FOLIO (Han et al. 2022), Leap-Of-Thought (Talmor et al. 2020), CLUTRR (Sinha et al. 2019), ReClor (Yu et al. 2020), LogiQA (Liu et al. 2020), LogiQA 2.0 (Liu et al. 2023), LogiQA2NLI (Liu et al. 2023)). For parts of the datasets, we keep all the test examples for ChatGPT evaluation, while other large datasets (i.e., bAbI-15 (Weston et al. 2016), RuleTaker (Clark, Tafjord, and Richardson 2020), bAbI-16 (Weston et al. 2016), α-NLI (Bhagavatula et al. 2020), α-NLG (Bhagavatula et al. 2020), AbductiveRules (Young et al. 2022) and D*-Ab (Tafjord, Dalvi, and Clark 2021)) are sampled to 1,000 examples. As for text-davinci-003 and BARD, we sample to 100 test examples for each dataset.

3.3 Selected Metrics

Most of previous evaluation works only report the accuracy metric. However, we argue that it is not sufficient for LLMs. Thus, we propose to evaluate from both objective and subjective views. To reflect the intermediate reasoning process of LLMs, we introduce four evaluation metrics: *answer correctness*, *explaination correctness*, *explaination completeness* and *explaination redundancy*.

- **Answer Correctness**. It indicates whether the generated answer is consistent with the true label. For the generation tasks, it requires that the meanings of the two should be the same rather than their corresponding tokens.
- **Explanation Correctness**. It indicates whether the generated explanation is logically correct to reason towards the true answer. It is a subjective view to determine whether the reasoning process of machines is in line with that of humans.
- **Explanation Completeness**. It means that in the reasoning process, the correct answer can be inferred through the selected known facts and the generated intermediate facts by the model. This does not necessarily cause answer correctness or explanation correctness.
- **Explanation Redundancy**. It means that in the reasoning process, the selected known facts and the generated intermediate facts by the model are more than practical facts to obtain the true answer. There are useless redundant facts for the reasoning.

Notably, the metric values of explanation correctness, completeness, and redundancy are independent. The explanation correctness is subjective to measure the rationality while the explanation completeness and redundancy are objective and about the facts used in reasoning. The permutation of their values can be arbitrary.

To find common logical flaws in LLMs, we set error types for the bad cases. This paper categories the errors from two main dimensions: (1) *Evidence Selection Process*, (2) *Reasoning Process*. The first one focuses on the evaluation of selected evidence by LLMs, while the second one stresses on the logical reasoning with the selected evidence.

Detailedly, *Evidence Selection Process* category can be further divided into: (1) Evidence Selection Errors; (2) Hallucination. The former one denotes that LLMs select the wrong facts or ignore the necessary facts from the beginning

[1] https://chat.openai.com/chat

[2] https://bard.google.com

of the reasoning. The latter one denotes that LLMs select the evidence which contradicts the given context or can not be verified by the context.

Reasoning Process category can be further divided into: (1) No Reasoning; (2) Mistakes of Reasoning Perspective; (3) Mistakes during Reasoning Process. The first one denotes LLMs fail to conduct the reasoning, instead they simply list the given facts and the final answer. The second one denotes that LLMs start from an irrelevant point, or they focus on the irrelevant perspective. The third one denotes that LLMs start from a proper view, but they make some mistakes during the reasoning process.

4 Overall Experiments

In this section, we will report the overall performances of LLMs on 15 logical reasoning datasets. Further, this paper will make an in-depth analysis according to the deductive, inductive, abductive and mixed reasoning manners of different LLMs.

4.1 Main Results

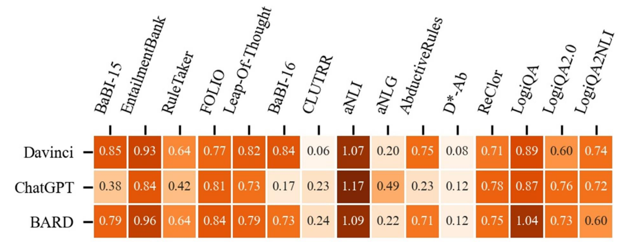
This paper mainly conducts evaluation experiments on three LLMs, i.e., text-davinci-003, ChatGPT and BARD under zero-shot, one-shot and three-shot settings respectively. Table 2 presents the overall results of these three LLMs on 15 logical reasoning datasets.

Overall, the performances of LLMs on the logical reasoning tasks still have a lot of room for improvement. Most of the performances are inferior to SOTA models, which are actually much smaller in size compared with LLMs. Detailedly, we analyze the results from the following points.

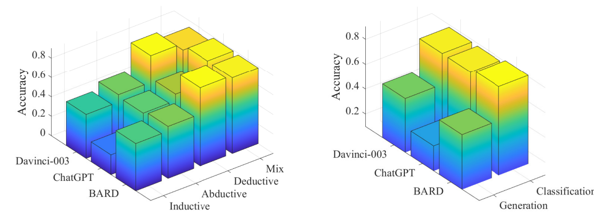
Firstly, we analyze the LLM performances under these four reasoning manners. To make the comparison clear, we only focus on the zero-shot results. To make the parallel comparison between reasoning manners, we introduce the SOTA result on each dataset. It can reflect the relative capability of LLMs against SOTA performances. Figure 1a presents the relative performance results of LLMs against SOTA (i.e., LLM accuracy / SOTA). Further, we calculate the weighted results of four reasoning manners, shown in Figure 1b.

From the results, ChatGPT performs worse in deductive and inductive settings compared with text-davinci-003 and BARD. In the abductive setting, three LLMs show comparable performances and BARD wins with slight advantages. In the mixed-form setting, ChatGPT performs better and BARD ranks second. Overall, BARD shows consistent superiority among deductive, inductive and abductive settings, while text-davinci-003 also does relatively well. It seems that ChatGPT struggles in the three settings, but is better at mixed-form reasoning.

Also, we compare the LLM performances between deductive, inductive and abductive settings. LLMs do best in deductive setting, while they mostly struggle in inductive setting. Such finding is consistent with previous work. We argue that deductive and abductive reasoning are in line with general cases, where LLMs are required to give a missing fact. But inductive reasoning requires high-level rule extraction from the given facts, which is complex and may not be abundant in the training corpus.



(a) LLM performances on different datasets.



(b) Under different reasoning types.

(c) Under generative and classification tasks.

Figure 1: Visualization on the metric of answer correctness.

Secondly, we take a closer at LLM performances from the generative and classification views in Figure 1c. ChatGPT performs particularly poor in generative tasks, such as bAbI-15, bAbI-16, CLUTRR, AbductiveRules and D*-Ab. We argue that ChatGPT makes a balance between chatting capability and precise generation, which may lead to the performance drops in generative tasks.

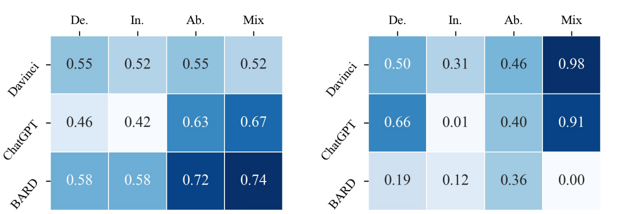
Thirdly, few-shot in-context learning (ICL) does not necessarily bring improvements in the logical reasoning tasks. It is quite inconsistent with the cases in other non-reasoning NLP tasks. We count the cases where LLMs can continuously obtain the performance gains from few-shot ICL (i.e., 0-shot $<$ 1-shot $<$ 3-shot). For text-davinci-003, only two (out of four) abductive datasets continuously benefit from the few-shot ICL. For ChatGPT, it witnesses performance improvements only in one (out of five) deductive dataset and one (out of four) abductive dataset. For BARD, few-shot ICL helps two (out of five) deductive datasets and one (out of four) abductive datasets. Surprisingly, few-shot ICL fails to provide consistent benefits for LLMs under inductive reasoning and mixed-form reasoning manners. We argue that inductive and mixed-form require more complex and high-order reasoning ability, which may be difficult to learn with few samples. And the task form of deductive and abductive reasoning are easy to follow, which provides possibility for few-shot ICL.

5 Fine-level Evaluations

In this section, we will provide the fine-level evaluations of LLMs from the following aspects. Firstly, we will focus on the newly proposed four metrics and offer sufficient analy-

Table 2: Overall evaluation results on answer correctness of LLMs (text-davinci-003, ChatGPT and BARD) on various logical reasoning datasets under the zero-shot, one-shot and three-shot settings. *De.*, *In.*, *Ab.* and *Mix* represent deductive, inductive, abductive and mixed-form of reasoning respectively. *Gen.* shows whether it is a generative task or a classification task.

	Dataset	Gen.	text-davinci-003			ChatGPT			BARD			SOTA
			0-shot	1-shot	3-shot	0-shot	1-shot	3-shot	0-shot	1-shot	3-shot	
De.	**bAbI-15**	✓	85.00	76.00	75.00	38.40	46.40	39.70	79.00	80.00	88.00	100
	EntailmentBank	✓	93.00	88.00	89.00	83.82	82.06	77.94	96.00	97.00	97.00	100
	RuleTaker		64.00	60.00	62.00	42.00	38.00	40.20	64.00	57.00	70.00	≈100
	FOLIO		48.00	53.00	52.00	50.00	50.98	54.41	52.00	43.00	49.00	62.11
	Leap-Of-Thought		82.00	90.00	87.00	72.61	74.01	61.21	79.00	72.00	79.00	99.7
In.	**bAbI-16**	✓	84.00	81.00	74.00	17.10	24.70	12.90	73.00	44.00	52.00	100
	CLUTRR	✓	6.00	23.00	20.00	21.99	19.55	12.83	23.00	26.00	24.00	95.0
Ab.	**α-NLI**		74.00	70.00	74.00	80.90	80.00	79.10	75.00	74.00	77.00	68.90
	α-NLG	✓	9.00	10.00	12.00	21.90	23.40	25.90	10.00	12.00	15.00	45.00
	AbductiveRules	✓	75.00	42.00	35.00	23.30	35.10	29.80	71.00	49.00	22.00	100
	D*-Ab	✓	8.00	21.00	23.00	11.60	2.50	1.80	11.00	0.00	0.00	≥95
Mix	**ReClor**		53.00	53.00	55.00	58.80	56.00	58.80	56.00	55.00	56.00	75.00
	LogiQA		41.00	35.00	39.00	40.25	39.48	40.86	48.00	46.00	47.00	46.10
	LogiQA 2.0		43.00	42.00	41.00	54.60	50.80	54.80	53.00	46.00	47.00	72.25
	LogiQA2NLI		59.00	55.00	58.00	57.83	53.83	57.00	48.00	50.00	47.00	-



(a) Visualization on the rigorous reasoning of LLMs.

(b) Visualization on the self-awareness of LLMs.

Figure 2: Heatmap visualization of LLM performances.

sis from various dimensions. Secondly, we will select some bad cases and attribute the errors. Thirdly, we will offer additional analysis on the specific datasets and present some phenomenon towards different reasoning manners.

5.1 Are LLMs Rigorous Logical Reasoning?

Although LLMs give the right answers in some cases, it remains unclear whether they truly conduct the correct logical reasoning or simply fit the right answers by chance. Thus, this paper further explores the reasoning process beyond the simple answers. And we view the cases where LLMs give correct answer together with correct and complete explanations as *Rigorous*. Table 3 shows detailed evaluation results on the rigor of LLMs.

Compared with the simple judgement of answer correctness, all selected LLMs present obvious performance drops. To simplify the analysis, we only take the zero-shot setting into consideration. In Figure 2a, we calculate the ratio of rigorous performance and answer correctness results. The higher values (darker color) in the heatmap mean better performance in the rigorous reasoning. According to the results, BARD shows best capability in rigorous reasoning, consistently under four reasoning manners. Meanwhile, ChatGPT still struggles in deductive and inductive settings, while text-davinci-003 comes last in both abductive and mixed-form reasoning manners.

Further, LLMs are best at keeping the rigorous reasoning in the abductive setting, while they are weak in the inductive setting. The finding is a little different from the analysis of simple accuracy conditions in the previous section. We argue that setting of abductive reasoning requires the LLMs to achieve the reasoning reversely, which can activate the reasoning process. While in deductive reasoning setting, the reasoning chain is sequential, which may cause LLMs to be in a lazy mode and harm the rigorous reasoning.

In Table 3, we also include the two conditions (1) when correct answer and correct explanations are satisfied, and (2) when correct answer and complete explanations are satisfied. Limited by space, we will attach the detailed analysis in Appendix.

5.2 Are LLMs Self-aware Logical Reasoners?

From another perspective, the redundancy of the generated context of LLMs has often been a hot topic. It is regarded as an important metric to measure the practicality of LLMs. In this paper, we view the LLMs with less redundant content as more *self-aware* ones, since they can clearly convey the necessary information rather than output all possible answers. In Table 4, we present the evaluation results of the self-awareness of LLMs. Similar to the previous approach, we calculate the weighted results for each reasoning setting and derive the scores for self-awareness. The darker color represents stronger capability of self-awareness.

According to the heatmap results, text-davinci-003 shows obvious advantages that wins inductive, abductive and mixed-form reasoning settings. Also, it comes second for deductive setting. On the contrary, BARD performs worst in deductive, abductive and mixed-form reasoning settings.

Compared with classification tasks (e.g., α-NLI v.s. α-NLG), LLMs are easy to make redundant answers in the

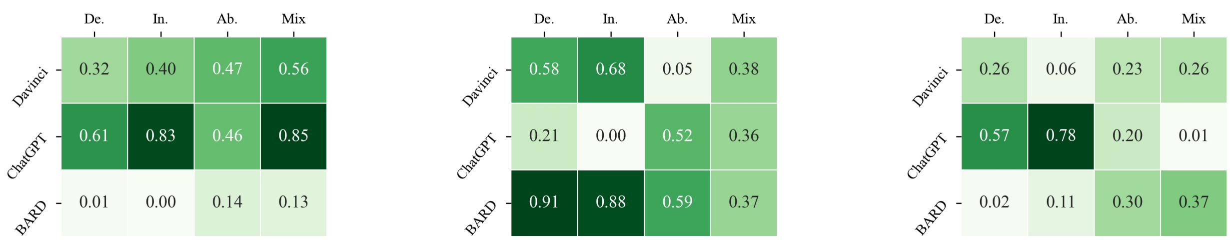
Table 3: Evaluations on whether LLMs are rigorous reasoners. For each dataset, the first line of results represents the performances when LLMs give correct answer, correct explanation as well as complete explanation simultaneously. The values in the subscripts denote the drops compared with only distinguishing the correctness of the answers. The second line of results represents the performances when LLMs give both correct answer and correct explanations, regardless of the completeness of explanations. The third line represents the cases when LLMs give correct answers and list complete explanations, regardless of the correctness of the explanations.

	Dataset	**text-davinci-003**			**ChatGPT**			**BARD**		
		0-shot	**1-shot**	**3-shot**	**0-shot**	**1-shot**	**3-shot**	**0-shot**	**1-shot**	**3-shot**
Deductive	**bAbI-15**	$53.00_{32.00\downarrow}$	$60.00_{16.00\downarrow}$	$64.00_{11.00\downarrow}$	$25.50_{12.90\downarrow}$	$12.10_{34.30\downarrow}$	$14.10_{25.60\downarrow}$	$45.00_{34.00\downarrow}$	$25.00_{55.00\downarrow}$	$47.00_{41.00\downarrow}$
		61.00	66.00	68.00	32.10	12.90	16.40	77.00	74.00	85.00
		56.00	60.00	64.00	27.90	18.00	17.60	45.00	25.00	47.00
	EntailmentBank	$29.00_{64.00\downarrow}$	$37.00_{51.00\downarrow}$	$30.00_{59.00\downarrow}$	$25.88_{57.94\downarrow}$	$20.00_{62.06\downarrow}$	$10.59_{67.35\downarrow}$	$26.00_{38.00\downarrow}$	$25.00_{32.00\downarrow}$	$33.00_{37.00\downarrow}$
		29.00	37.00	30.00	25.88	20.00	10.59	54.00	66.00	71.00
		72.00	73.00	75.00	62.06	57.65	31.76	94.00	96.00	97.00
	RuleTaker	$35.30_{6.70\downarrow}$	$22.50_{15.50\downarrow}$	$24.80_{15.40\downarrow}$	$25.88_{57.94\downarrow}$	$20.00_{62.06\downarrow}$	$10.59_{67.35\downarrow}$	$54.00_{42.00\downarrow}$	$66.00_{31.00\downarrow}$	$71.00_{26.00\downarrow}$
		36.20	24.00	26.00	25.88	20.00	10.59	26.00	28.00	33.00
		36.00	23.20	26.00	62.06	57.65	31.76	26.00	25.00	34.00
	FOLIO	$27.00_{21.00\downarrow}$	$25.00_{28.00\downarrow}$	$25.00_{27.00\downarrow}$	$28.92_{21.08\downarrow}$	$27.94_{23.04\downarrow}$	$27.94_{26.47\downarrow}$	$21.00_{31.00\downarrow}$	$19.00_{24.00\downarrow}$	$20.00_{29.00\downarrow}$
		28.00	25.00	26.00	28.92	27.94	27.94	23.00	19.00	22.00
		27.00	28.00	27.00	33.33	32.35	35.29	25.00	22.00	23.00
	Leap-of-Thought	$29.00_{53.00\downarrow}$	$43.00_{47.00\downarrow}$	$38.00_{49.00\downarrow}$	$70.60_{2.02\downarrow}$	$24.36_{49.65\downarrow}$	$28.86_{32.35\downarrow}$	$76.00_{3.00\downarrow}$	$69.00_{3.00\downarrow}$	$77.00_{2.00\downarrow}$
		63.00	63.00	58.00	71.22	24.83	29.79	76.00	69.00	77.00
		32.00	46.00	42.00	70.99	48.33	40.88	79.00	71.00	7.00
Inductive	**bAbI-16**	$59.00_{25.00\downarrow}$	$35.00_{46.00\downarrow}$	$23.00_{51.00\downarrow}$	$8.30_{8.80\downarrow}$	$8.20_{16.50\downarrow}$	$2.60_{10.30\downarrow}$	$24.00_{49.00\downarrow}$	$15.00_{29.00\downarrow}$	$16.00_{36.00\downarrow}$
		67.00	50.00	37.00	10.20	8.40	3.10	58.00	32.00	24.00
		65.00	44.00	35.00	10.00	9.40	3.50	32.00	22.00	24.00
	CLUTRR	$2.00_{4.00\downarrow}$	$7.00_{16.00\downarrow}$	$6.00_{14.00\downarrow}$	$7.85_{14.14\downarrow}$	$4.62_{14.92\downarrow}$	$2.71_{10.12\downarrow}$	$19.00_{4.00\downarrow}$	$25.00_{1.00\downarrow}$	$23.00_{1.00\downarrow}$
		2.00	7.00	6.00	7.94	4.62	2.71	19.00	25.00	23.00
		6.00	18.00	18.00	10.56	6.72	3.75	21.00	26.00	23.00
Abductive	**α-NLI**	$68.00_{6.00\downarrow}$	$69.00_{1.00\downarrow}$	$68.00_{6.00\downarrow}$	$77.50_{3.40\downarrow}$	$64.50_{15.50\downarrow}$	$58.40_{20.70\downarrow}$	$75.00_{0.00-}$	$71.00_{3.00\downarrow}$	$77.00_{0.00-}$
		70.00	69.00	68.00	78.20	66.70	60.90	75.00	71.00	77.00
		68.00	69.00	68.00	77.60	64.50	58.40	75.00	71.00	77.00
	α-NLG	$1.00_{8.00\downarrow}$	$0.00_{10.00\downarrow}$	$2.00_{10.00\downarrow}$	$15.30_{6.60\downarrow}$	$16.00_{7.40\downarrow}$	$10.30_{15.60\downarrow}$	$7.00_{3.00\downarrow}$	$8.00_{4.00\downarrow}$	$9.00_{6.00\downarrow}$
		1.00	0.00	2.00	15.30	16.00	10.40	7.00	9.00	10.00
		8.00	8.00	9.00	20.90	22.40	21.40	10.00	9.00	14.00
	AbductiveRules	$50.00_{25.00\downarrow}$	$5.00_{37.00\downarrow}$	$0.00_{35.00\downarrow}$	$12.00_{11.30\downarrow}$	$18.30_{16.80\downarrow}$	$5.70_{24.10\downarrow}$	$57.00_{14.00\downarrow}$	$30.00_{19.00\downarrow}$	$10.00_{12.00\downarrow}$
		75.00	10.00	0.00	20.50	29.60	9.50	65.00	40.00	18.00
		50.00	5.00	0.00	13.00	19.90	5.80	57.00	30.00	10.00
	D*-Ab	$4.00_{4.00\downarrow}$	$5.00_{16.00\downarrow}$	$4.00_{19.00\downarrow}$	$4.10_{7.50\downarrow}$	$1.20_{1.30\downarrow}$	$1.10_{0.70\downarrow}$	$4.00_{7.00\downarrow}$	$0.00_{0.00-}$	$0.00_{0.00-}$
		7.00	7.00	7.00	5.50	1.50	1.20	4.00	0.00	0.00
		4.00	5.00	4.00	4.50	1.20	1.10	5.00	0.00	0.00
Mix	**ReClor**	$5.00_{48.00\downarrow}$	$0.00_{53.00\downarrow}$	$0.00_{55.00\downarrow}$	$28.60_{30.20\downarrow}$	$25.20_{30.80\downarrow}$	$29.80_{29.00\downarrow}$	$38.00_{18.00\downarrow}$	$34.00_{21.00\downarrow}$	$33.00_{23.00\downarrow}$
		5.00	0.00	0.00	32.40	28.00	32.20	48.00	46.00	42.00
		42.00	46.00	42.00	50.00	32.80	46.00	38.00	36.00	38.00
	LogiQA	$5.00_{36.00\downarrow}$	$2.00_{33.00\downarrow}$	$1.00_{38.00\downarrow}$	$25.96_{14.29\downarrow}$	$21.35_{18.13\downarrow}$	$20.43_{20.43\downarrow}$	$29.00_{19.00\downarrow}$	$26.00_{20.00\downarrow}$	$19.00_{28.00\downarrow}$
		6.00	2.00	1.00	28.42	23.20	23.81	34.00	27.00	23.00
		23.00	26.00	27.00	29.19	27.50	26.42	30.00	35.00	27.00
	LogiQA2.0	$38.00_{5.00\downarrow}$	$32.00_{10.00\downarrow}$	$28.00_{13.00\downarrow}$	$43.40_{11.20\downarrow}$	$34.60_{16.20\downarrow}$	$38.80_{16.00\downarrow}$	$47.00_{6.00\downarrow}$	$39.00_{7.00\downarrow}$	$44.00_{3.00\downarrow}$
		39.00	32.00	28.00	43.80	35.20	38.80	47.00	39.00	44.00
		39.00	32.00	30.00	44.20	36.80	40.60	47.00	39.00	44.00
	LogiQA2NLI	$57.00_{2.00\downarrow}$	$51.00_{4.00\downarrow}$	$56.00_{2.00\downarrow}$	$43.17_{14.67\downarrow}$	$36.33_{17.50\downarrow}$	$36.50_{20.50\downarrow}$	$38.00_{10.00\downarrow}$	$41.00_{9.00\downarrow}$	$37.00_{10.00\downarrow}$
		57.00	51.00	56.00	43.50	36.50	36.83	40.00	43.00	40.00
		57.00	51.00	56.00	53.50	49.00	50.16	38.00	41.00	37.00

generative tasks. It is because open questions can encourage LLMs to generate content from various perspectives, which is easy to cover redundant information. Also, mixed-form reasoning setting witnesses obvious fewer redundant cases. The tasks in mixed-form reasoning are mostly based on the question answering, which are close to the real-life text, LLMs are prone to give rational and specific content. While in other settings, the input context is designed to test the model reasoning capability and they may not be common and sufficient. It can cause LLMs to employ the embodied commonsense knowledge for reasoning and generate extra explanations.

Table 4: Evaluation results on the metric of Explanation Redundancy.

	Dataset	Gen.	text-davinci-003 0-shot	text-davinci-003 1-shot	text-davinci-003 3-shot	ChatGPT 0-shot	ChatGPT 1-shot	ChatGPT 3-shot	BARD 0-shot	BARD 1-shot	BARD 3-shot
De.	**bAbI-15**	✓	63.00	56.00	43.00	22.60	39.40	55.70	99.00	84.00	62.00
	EntailmentBank	✓	8.00	6.00	7.00	7.06	5.88	3.24	26.00	25.00	28.00
	RuleTaker		26.00	29.00	27.00	21.30	27.80	34.80	80.00	84.00	75.00
	FOLIO		14.00	23.00	21.00	31.86	22.55	19.61	60.00	63.00	68.00
	Leap-Of-Thought		71.00	55.00	54.00	32.74	5.04	4.73	2.00	2.00	0.00
In.	**bAbI-16**	✓	60.00	77.00	86.00	93.60	29.80	41.20	96.00	98.00	99.00
	CLUTRR	✓	2.00	28.00	31.00	2.62	1.57	0.87	2.00	6.00	14.00
Ab.	**α-NLI**		2.00	2.00	1.00	1.00	0.20	0.10	8.00	16.00	0.00
	α-NLG	✓	63.00	61.00	72.00	70.70	69.70	64.50	24.00	32.00	31.00
	AbductiveRules	✓	1.00	0.00	0.00	42.40	5.40	0.50	67.00	48.00	22.00
	D*-Ab	✓	85.00	27.00	17.00	55.30	27.10	16.70	18.00	16.00	2.00
Mix	**ReClor**		1.00	1.00	1.00	2.00	1.20	1.40	11.00	16.00	24.00
	LogiQA		0.00	5.00	0.00	1.54	0.77	1.08	32.00	35.00	43.00
	LogiQA 2.0		0.00	0.00	0.00	0.80	4.00	0.8	5.00	5.00	4.00
	LogiQA2NLI		0.00	0.00	0.00	0.17	0.50	0.17	11.00	31.00	4.00



(a) Visualization on the non-hallucination.
(b) Visualization on the activity.
(c) Visualization on the orientation.

Figure 3: Heatmap results for the hallucination, activity and orientation of LLMs.

5.3 Do LLMs Have Obvious Logical Flaws?

According to the previous statements, we define errors for the bad cases (with wrong explanations) from two dimension, i.e., *Evidence Selection Process*, *Reasoning Process*. The former dimension covers two error types, (i) *Evidence Selection Errors* and (ii) *Hallucination*. The above two errors are independent to each other. The latter one includes three error types, i.e., (i) *No Reasoning*, (ii) *Mistakes of Reasoning Perspective* and (iii) *Mistakes during Reasoning Process*. Each bad case can only be attributed one of the three errors of *Reasoning Process* dimension.

In Figure 8, we visualize the attribution results for fourteen datasets, including four deductive ones, two inductive ones, four abductive ones and four mixed-form ones. Overall speaking, the types of errors vary between datasets. From the *Evidence Selection Process*, 33.26% of the bad cases fail to select the right answers for reasoning. Also, 27.46% of the bad cases suffer from the hallucination issue of LLMs. From the dimension of *Reasoning Process*, *No Reasoning* error keeps a small portion in most of the cases, only covering 19.33% of the selected cases in total. Meanwhile, *Mistakes of Reasoning Perspectives* occupies 44.47% of the cases and *Mistakes of Reasoning Process* covers 36.20% of bad cases. In the following, we will provide detailed analysis of specific LLM and specific reasoning setting.

5.4 Are LLMs Easy to Induce Hallucination in Logical Reasoning?

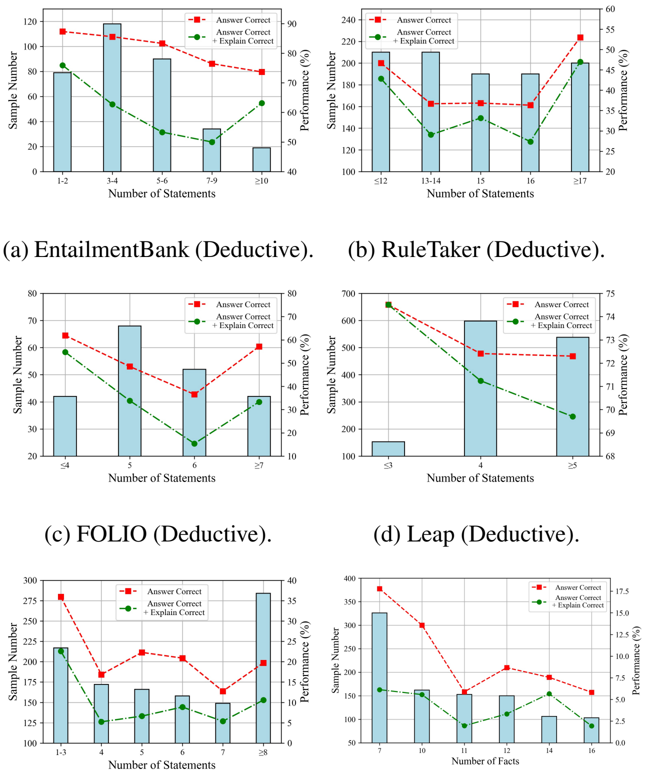
In common definition, hallucination usually refers to the generated content which contradicts the commonsense or current facts. To fit the logical reasoning tasks, this paper extends the definition to the cases that employ the facts contradicting the context or are not verified by the context. In Figure 3a, we present the weighted performances of the cases with no hallucinations. The darker color denotes the better performances in avoiding hallucinations.

From the results, ChatGPT maintains great and consistent competitiveness, coming first in deductive, inductive and mixed-form settings. Also, it comes second in abductive reasoning tasks only with slight disadvantages. Meanwhile, BARD performs particularly poorly in avoiding hallucinations and keeping clear during reasoning. Under all four reasoning settings, BARD ranks last with huge gaps.

On average, LLMs induce hallucination in 27.46% of the failure cases. Among deductive, inductive and abductive settings, model hallucinations are more common under deductive reasoning tasks. While LLMs may have clearer minds in the inductive setting.

5.5 Are LLMs Active Logical Reasoners?

In this paper, we view the LLMs with fewer *No Reasoning* error as more *Active* Logical Reasoners. In Figure 3b, we



(a) EntailmentBank (Deductive). (b) RuleTaker (Deductive).

(c) FOLIO (Deductive). (d) Leap (Deductive).

(e) CLUTRR (Inductive). (f) D*-Ab (Abductive).

Figure 4: The LLM performances with different number of statements.

show the weighted results to measure the active reasoning cases. The higher values represent LLMs are more active in reasoning, while the lower values represent the lazier cases.

Among the three LLMs, BARD is the most active logical reasoner, winning in deductive, inductive and abductive settings. For ChatGPT, it is regarded as the lazier reasoner in deductive and inductive settings, while text-davinci-003 is lazier in abductive reasoning tasks.

Further, we compare the performances between different reasoning manners. In deductive reasoning tasks, LLMs perform more actively in reasoning. And in abductive settings, LLMs are easy to show lazier performances. Since deductive tasks are in a forward reasoning manner, it is more natural ways for both generative LLMs and humans and thus it can inspire LLMs to generate effective reasoning chains. However, abductive reasoning requires LLMs to output the explanations for the given inputs, which is in a backward reasoning manner. It is intuitive that LLMs fail to conduct reasoning for some cases.

5.6 Are LLMs Oriented Logical Reasoners?

At the beginning of the reasoning process, it is key to find the correct starting points and potential directions for reasoning. We view LLMs with such capability as *Oriented* Logical Reasoners. And we provide the evaluation results based on the error type *Mistakes of Reasoning Perspective*, shown in Figure 3c. From the heatmap results, ChatGPT shows better oriented capability in deductive and inductive tasks, compared with other two LLMs. But it usually fails to identify the right reasoning direction in abductive and mixed-form reasoning. On the contrary, BARD perfroms well in finding right direction in the abductive and mixed-form settings, but it fails in deductive and inductive ones. Compared with others, text-davinci-003 keeps moderate performances in the identification of reasoning perspectives.

Combined with previous findings, ChatGPT is a lazy logical reasoner, but it is better at finding right directions if reasoning. Instead, text-davinci-003 and BARD are more active ones in logical reasoning, but they are easy to start from wrong directions, inducing reasoning mistakes.

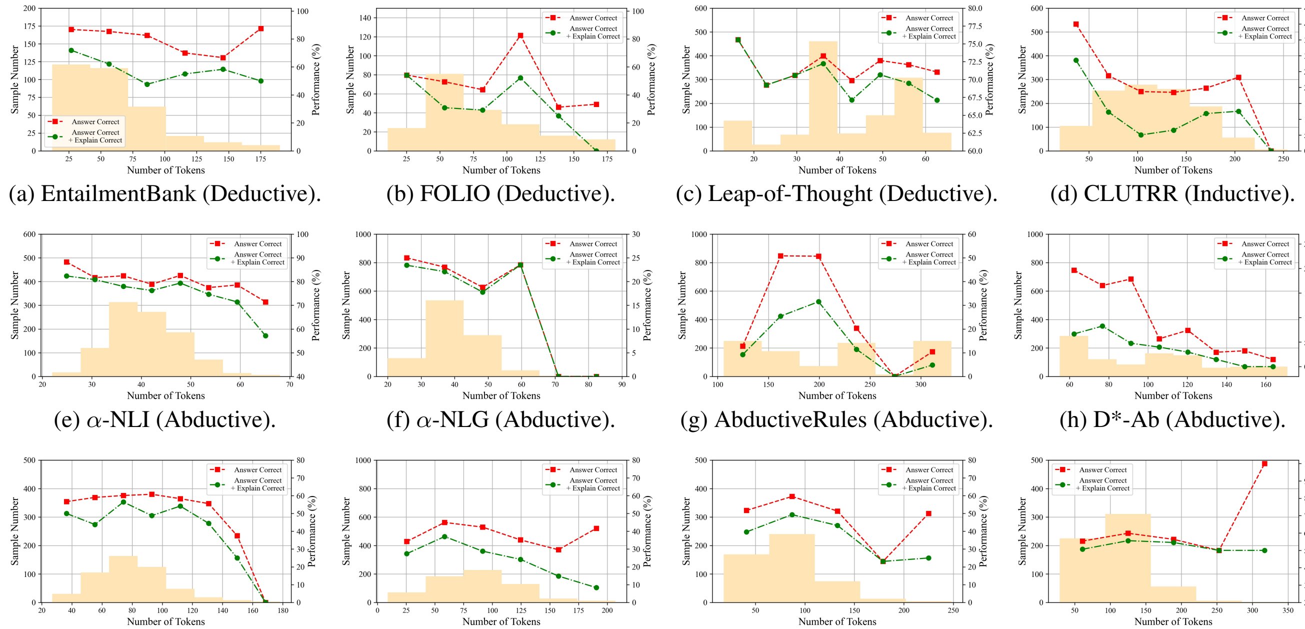
5.7 How Does the Number of Statements Affect the Performances of LLMs?

In the following, we will explore some of the key factors to affect the reasoning performances of LLMs. Since the length of the input context can be different for the datasets, we report the model performances with number of statements in Figure 4. We take ChatGPT for analysis (results on text-davinci-003 and BARD are attached in Appendix) and choose six datasets with specific counts of statements for illustration, covering the three reasoning manners. The first four subfigures are related to the deductive reasoning manner, i.e., EntailmentBank, RuleTaker, FOLIO, and Leap-of-Thought. Figure 4(e) is CLUTRR in the inductive setting and Figure 4(f) is D-Ab in the abductive setting. The horizontal axis denotes the number of statements. The left vertical axis denotes the number of samples for different numbers of statements. And the right vertical axis represents the performances with different numbers of statements.

From the overall results, LLMs can conduct correct and rigorous reasoning with fewer number of statements. With the statement number increasing, the performances drop a lot and LLMs struggle to keep the rigor of reasoning. Interestingly, five (out of six) datasets witness performance gains when the number of statements reach certain values. For example, in RuleTaker dataset, the best performances are achieved when the number of statements are larger than 17. And when the number is between 13-16, ChatGPT is capable of keeping the stable performances. We argue that the larger number of statements can provide richer information and sometimes can help control the reasoning direction of LLMs. Further, we explore the influences of number of tokens in the context. Same as previous, we only present the performances on ChatGPT and report the other two models in Appendix. Figure 5 illustrates the performances on 12 datasets. Overall speaking, with the number of tokens increases, the LLM performances will drop.

5.8 How does the number of reasoning hops affect the performances of LLM?

Also, it is interesting to explore the influences of reasoning hops for LLMs. Among the selected dataset, three of them offer the number of hops for each sample, which are

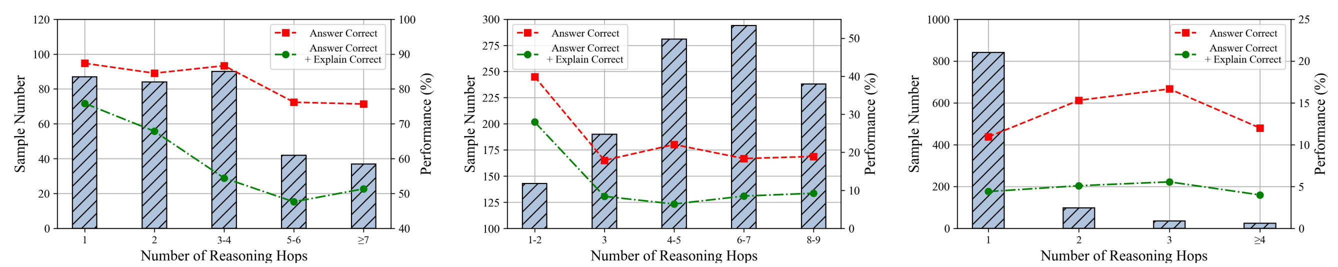


(a) EntailmentBank (Deductive). (b) FOLIO (Deductive). (c) Leap-of-Thought (Deductive). (d) CLUTRR (Inductive).

(e) α-NLI (Abductive). (f) α-NLG (Abductive). (g) AbductiveRules (Abductive). (h) D*-Ab (Abductive).

(i) ReClor (Mix). (j) LogiQA (Mix). (k) LogiQA2.0 (Mix). (l) LogiQA2NLI (Mix).

Figure 5: The performances ChatGPT with different number of tokens on various datasets.



(a) Deductive (EntailmentBank). (b) Inductive (CLUTRR). (c) Abductive (D*-Ab).

Figure 6: The performances of ChatGPT under different number of hops. Comparison of Deductive, Inductive and Abductive reasoning settings.

EntailmentBank in deductive reasoning, CLUTRR in inductive reasoning and D-Ab in abductive reasoning. Figure 6 presents the performance of ChatGPT with different hop numbers (Results of other LLMs will be listed in Appendix). Alongside the simple accuracy results, we also report the rigorous reasoning cases where both the answer and explanations are correct.

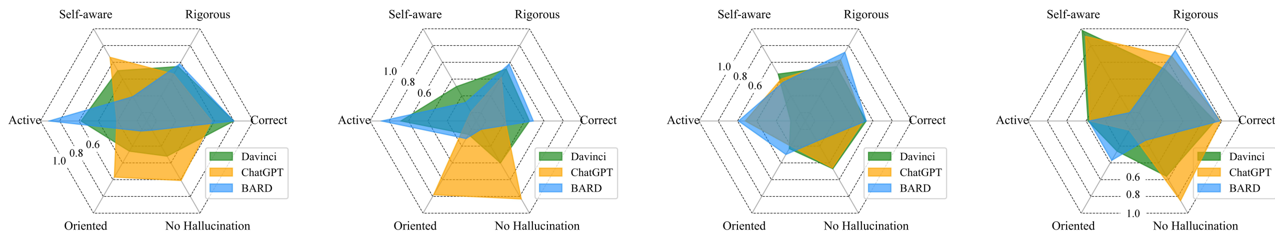
In deductive setting, with the number of hops increasing, the performances witness obvious drops, particularly influencing the rigor of the LLM reasoning. It illustrates that reasoning hops have great effects on the deductive reasoning. In inductive reasoning, when the hop number is greater than two, the performance of ChatGPT decreases sharply. When the number ranges from three to nine, the performance of ChatGPT keeps stable at a relatively low level. Combined with the weak performance of ChatGPT in inductive reasoning tasks, it demonstrates that ChatGPT can only work on the simple induction, and it obviously struggles at the cases when more hops are needed. In abductive reasoning setting, the majority of the test samples only need one hop reasoning. When the hop number increases, the performances of ChatGPT witness slight improvements. It shows that ChatGPT may have the potential capability of multi-hop reasoning in the abductive setting.

6 Introduction of New Dataset

The new dataset named NeuLR will be released in the later version of the paper soon.

7 Conclusion

In this paper, in-depth evaluations are conducted on logical reasoning tasks, discussing whether LLMs are really good logical reasoners. First, the logical reasoning evaluations are organized from deductive, inductive, abductive views. We



(a) Deductive. (b) Inductive. (c) Abductive. (d) mixed-form.

Figure 7: Visualization of LLM capability under four reasoning settings.

select fifteen logical reasoning datasets to evaluate on three representative LLMs (i.e., text-davinci-003, ChatGPT and BARD) under both zero-shot and few-shot settings. Second, this paper provides fine-level evaluations on four metrics, covering both objective and subjective views. For bad cases, extensive error attributions are conducted from two dimensions, forming five error types. It uncovers the logical flaws of LLMs and we provide deep analysis on the results. Third, to achieve a fair and pure benchmark for logical reasoning capability, we propose a dataset with neutral content, covering deductive, inductive and abductive settings.

Based on the evaluation results above, we abstract six dimensions to measure the logical reasoning capability of LLMs: (1) *correct* logical reasoner; (2) *rigorous* logical reasoner; (3) *self-aware* logical reasoner; (4) *active* logical reasoner; (5) *oriented* logical reasoner; and (6) logical reasoner with *no hallucination*. All these properties can be calculated with the evaluation methods proposed in this papaer. Therefore, we can easily derive the ability maps of LLMs. Considering the different performances of LLMs on deductive, inductive, abductive and mixed-form reasoning settings, we visualize the ability maps of LLMs under each of the settings respectively in Figure 7.

According to the results, text-davinci-003 can maintain balanced performances in the deductive and mixed-form settings. But it usually fails to keep oriented for reasoning in the inductive setting, and it also shows laziness in the abductive reasoning tasks. Since it is the earliest released LLM of the three, it is understandable that text-davinci-003 has some limitations in logical reasoning tasks, especially in the more complex settings (e.g., inductive and abductive).

From the perspective of common evaluations, ChatGPT is the weakest LLM of the three, since it performs bad in showing correct and rigorous reasoning under deductive, inductive and abductive settings. Also, it seems to be the laziest reasoner in deductive and inductive settings. However, it surprises us that it shows unique advantages in maintaining oriented reasoning and avoiding hallucination, especially in deductive and inductive settings. In addition, it shows its comprehensive capability in the mixed-form setting. We argue that ChatGPT is specially designed for chatting, thus it does pretty well in keeping rational but is not good at solving complex reasoning problems.

BARD is the most active reasoner and it keeps great competitiveness as a correct and rigorous reasoner. However, it also shows obvious flaws compared with other LLMs. BARD tends to generate redundant content, easily fails to find the correct reasoning directions and it usually fails to avoid hallucinations. In short, BARD shows great advantages in current benchmarks with objective metrics, due to the larger model size and massive training data. But it still has much room for improvement in some implicit aspects, i.e., self-awareness, orientation and non-hallucination.

In all, all LLMs have specific limitations in logical reasoning tasks. They are relatively good at deductive reasoning, but struggle a lot particularly in the inductive setting. Additionally, current benchmarks, which mainly rely on objective metrics, are not sufficient to comprehensively evaluate LLMs.

8 Future Directions

Given the evaluation results, this paper lists some of the future directions for logical reasoning tasks.

Strengthen the reasoning ability of inductive reasoning. Inductive reasoning draws broad conclusions from specific observations, requiring a more abstract and comprehensive understanding of real-world knowledge compared to deductive or abductive reasoning. However, LLMs have shown poor performance in this area, as demonstrated in Section 4.1. Therefore, it is crucial to develop pre-training or fine-tuning strategies to enhance their inductive reasoning abilities. One such strategy could be constructing more inductive instructions to guide LLMs.

Enhance the LLM's perception of its capability boundaries. LLMs are capable of generating answers and explanations for reasoning questions regardless of difficulty and rationality. To realize it, LLMs would list some irrelevant facts of the given context or even hallucinations, as Sections 5.1 and 5.4 demonstrate. It will lead to LLMs solemnly talking nonsense and resulting in illogical, uninformative, or meaningless answers. A good logical reasoner should be aware of its boundaries and acknowledge when it is unable to answer a question. To enhance LLMs' self-awareness of their capability boundaries, the future research could focus on cognitive science and neuroscience studies of human self-awareness.

Strengthen the rigorous reasoning to apply to real-world scenarios. Table 3 illustrates that current LLMs are not sufficiently rigorous for deductive, inductive, abductive, and mixed reasoning. As a result, there is still a significant gap between their capabilities and their potential applications in real-world scenarios, particularly those that require detailed intermediate explanations. For instance, using LLMs to solve mathematical problems and provide precise intelligent Q&A services in the education field remains a significant challenge (Drori et al. 2022).

Minimize the occurrence of hallucinations. Similar to the behaviors in other problem-solving contexts (Ji et al. 2023), LLMs may generate false or irrelevant hallucinations during logical reasoning tasks. It suggests that LLMs may not fully comprehend the question and can not solve it correctly. To address this issue, future research should develop more comprehensive evaluation metrics for hallucinations and explore specific strategies to minimize their occurrence.

Improve the multi-hop reasoning capability, especially in inductive and abductive settings. Combined with the results from Figure 6 and 9, the multi-hop reasoning capability of LLMs still have much room for improvement. Especially in the inductive and abductive settings, LLMs perform quite struggling. Since the multi-hop reasoning evaluates the high-level capabilities of LLMs, it is necessary to extend LLM capability to such complex settings. In fact, humans are better at decomposing the complex questions. It can be an interesting topic for LLMs to capture the ability of dividing and conquering questions, thus benefiting the multi-hop reasoning.

Increase explainability. Finally, explainability of LLMs will be essential for building trust, detecting and mitigating biases, improving performance, promoting user understanding, and complying with regulations. A commonsense-based neurosymbolic AI framework, such as the one proposed by Cambria et al. (2022) for sentiment analysis, can help increase the explainability of the reasoning processes required for decision-making, which is crucial for sensitive applications involving ethics, privacy and health.

References

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *CoRR*, abs/2302.04023.

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, W.; and Choi, Y. 2020. Abductive Commonsense Reasoning. In *8th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Bian, N.; Han, X.; Sun, L.; Lin, H.; Lu, Y.; and He, B. 2023. ChatGPT is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models. *CoRR*, abs/2303.16421.

Cambria, E.; Liu, Q.; Decherchi, S.; Xing, F.; and Kwok, K. 2022. SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis. In *LREC*, 3829–3839.

Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 3882–3890.

Dalvi, B.; Jansen, P.; Tafjord, O.; Xie, Z.; Smith, H.; Pipatanangkura, L.; and Clark, P. 2021. Explaining Answers with Entailment Trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7358–7370. Association for Computational Linguistics.

Drori, I.; Zhang, S.; Shuttleworth, R.; Tang, L.; Lu, A.; Ke, E.; Liu, K.; Chen, L.; Tran, S.; Cheng, N.; et al. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences (PNAS)*, 119(32): e2123433119.

Goel, V. 2007. Anatomy of Deductive Reasoning. *Trends in cognitive sciences*, 11(10): 435–441.

Han, S.; Schoelkopf, H.; Zhao, Y.; Qi, Z.; Riddell, M.; Benson, L.; Sun, L.; Zubova, E.; Qiao, Y.; Burtell, M.; Peng, D.; Fan, J.; Liu, Y.; Wong, B.; Sailor, M.; Ni, A.; Nan, L.; Kasai, J.; Yu, T.; Zhang, R.; Joty, S. R.; Fabbri, A. R.; Kryscinski, W.; Lin, X. V.; Xiong, C.; and Radev, D. 2022. FOLIO: Natural Language Reasoning with First-Order Logic. *CoRR*, abs/2209.00840.

Heit, E.; and Rotello, C. M. 2010. Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(3): 805.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Qi, F.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *CoRR*, abs/2305.08322.

Imani, S.; Du, L.; and Shrivastava, H. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. *CoRR*, abs/2303.05398.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 248:1–248:38.

Johnson-Laird, P. N. 1999. Deductive Reasoning. *Annual review of psychology*, 50(1): 109–135.

Josephson, J. R.; and Josephson, S. G. 1996. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Li, X.; Cheng, G.; Chen, Z.; Sun, Y.; and Qu, Y. 2022. AdaLoGN: Adaptive Logic Graph Network for Reasoning-Based Machine Reading Comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7147–7161. Association for Computational Linguistics.

Liu, H.; Ning, R.; Teng, Z.; Liu, J.; Zhou, Q.; and Zhang, Y. 2023. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. *CoRR*, abs/2304.03439.

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 3622–3628. ijcai.org.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022a. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022b. Training language models to follow instructions with human feedback. In *NeurIPS*.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *CoRR*, abs/2302.06476.

Sinha, K.; Sodhani, S.; Dong, J.; Pineau, J.; and Hamilton, W. L. 2019. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4505–4514. Association for Computational Linguistics.

Tafjord, O.; Dalvi, B.; and Clark, P. 2021. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Findings of the Association for Computational Linguistics*, 3621–3634. Association for Computational Linguistics.

Talmor, A.; Tafjord, O.; Clark, P.; Goldberg, Y.; and Berant, J. 2020. Leap-Of-Thought: Teaching Pre-Trained Models to Systematically Reason Over Implicit Knowledge. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tu, S.; Li, C.; Yu, J.; Wang, X.; Hou, L.; and Li, J. 2023. ChatLog: Recording and Analyzing ChatGPT Across Time. *CoRR*, abs/2304.14106.

Walton, D. 2001. Abductive, presumptive and plausible arguments. *Informal Logic*, 21(2).

Weston, J.; Bordes, A.; Chopra, S.; and Mikolov, T. 2016. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *4th International Conference on Learning Representations (ICLR)*.

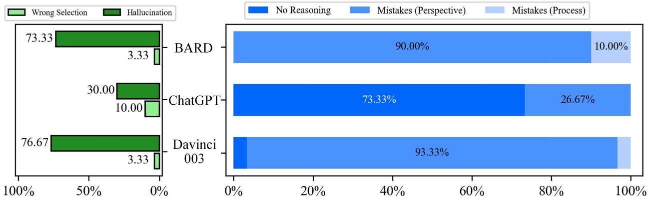
Xu, F.; Liu, J.; Lin, Q.; Pan, Y.; and Zhang, L. 2022. Logiformer: A Two-Branch Graph Transformer Network for Interpretable Logical Reasoning. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1055–1065. ACM.

Young, N.; Bao, Q.; Bensemann, J.; and Witbrock, M. 2022. AbductionRules: Training Transformers to Explain Unexpected Inputs. In *Findings of the Association for Computational Linguistics*, 218–227. Association for Computational Linguistics.

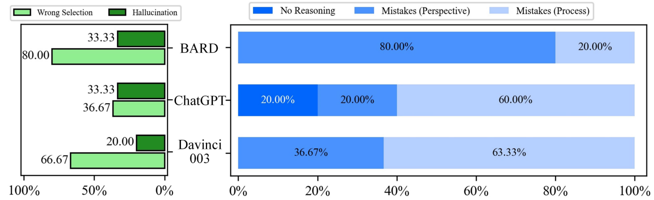
Yu, F.; Zhang, H.; and Wang, B. 2023. Nature Language Reasoning, A Survey. *arXiv preprint arXiv:2303.14725*.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *8th International Conference on Learning Representations (ICLR)*. OpenReview.net.

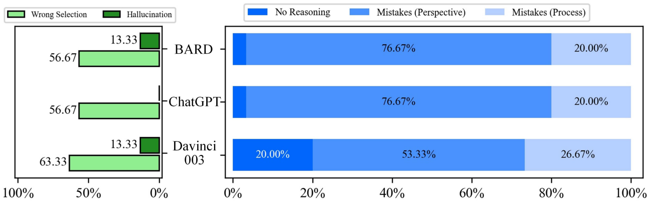
Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K. R.; and Hashimoto, T. B. 2023. Benchmarking Large Language Models for News Summarization. *CoRR*, abs/2301.13848.



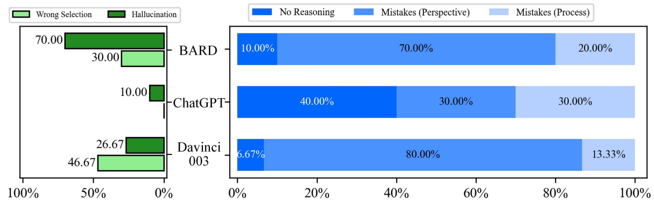
(a) bAbI15 (Deductive).



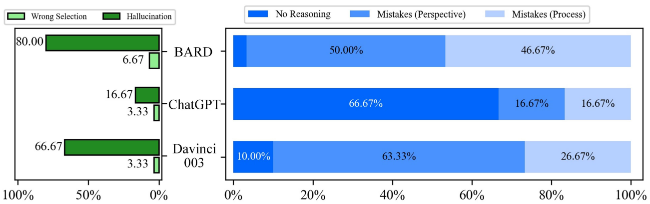
(b) RuleTaker (Deductive).



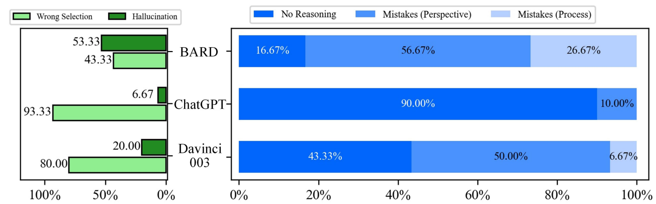
(c) FOLIO (Deductive).



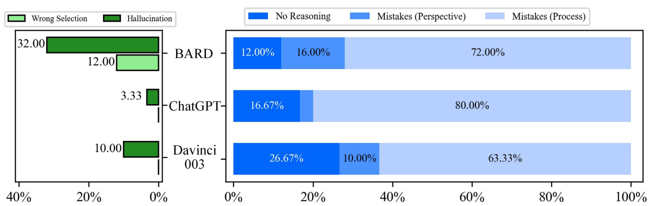
(d) Leap-of-Thought (Deductive).



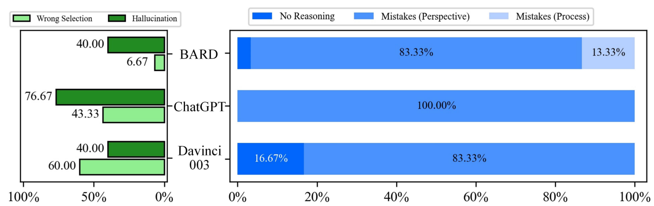
(e) bAbI16 (Inductive).



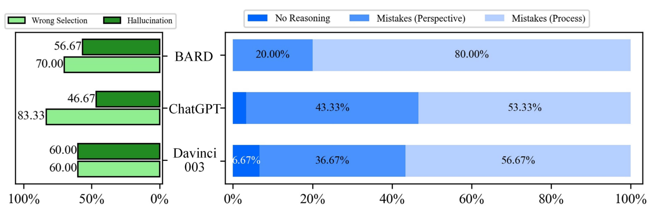
(f) CLUTRR (Inductive).



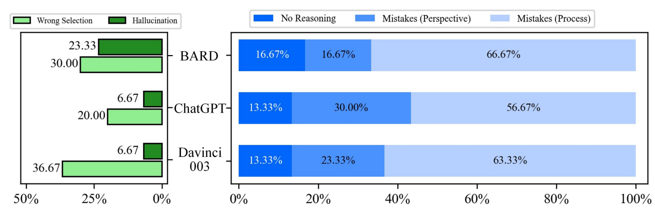
(g) α-NLI (Abductive).



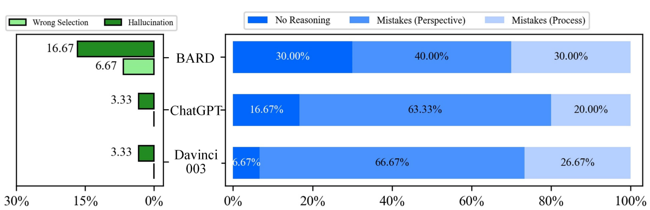
(h) α-NLG (Abductive).



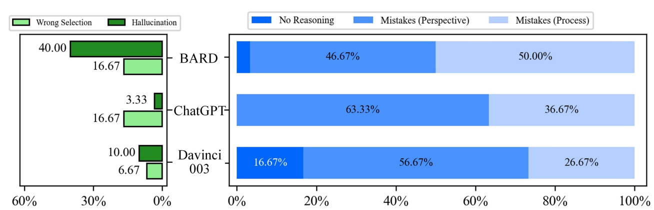
(i) AbductiveRules (Abductive).



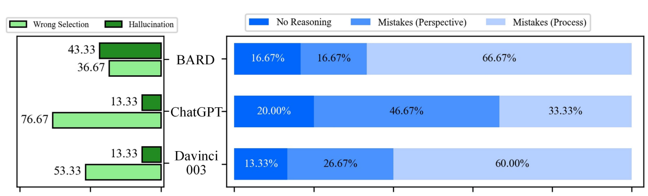
(j) D*-Ab (Abductive).



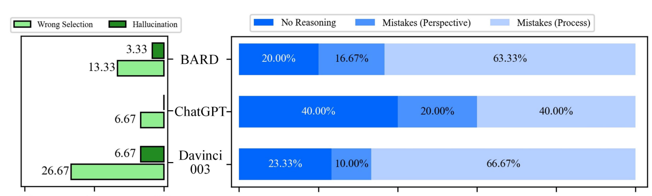
(k) ReClor (Mix).



(l) LogiQA (Mix).



(m) LogiQA2.0 (Mix).



(n) LogiQA2NLI (Mix).

Figure 8: Visualization of the statistics of different error types.

A Supplementary Analysis

We also provide the performances of text-davinci-003 and BARD with different number of hops in Figure 9. In the deductive setting of text-davinci-003, model performance drops with the number of hops increasing. But when the hop number is over five, it witnesses slight gains in performance, which illustrates that text-davinci-003 has the potential to conduct multi-hop reasoning in the deductive reasoning setting. However, in the inductive and abductive settings, the performances of text-davinci-003 decrease sharply when the number of hops increases. Especially, it fails all the cases when the hop number is over six in the inductive reasoning, and also when the hop number is greater than one in the abductive reasoning. It is inferior to ChatGPT.

For BARD, the situation is quite different. In the deductive reasoning, the performance of BARD increases with the hop number adds. Especially, when the hop number is over five, the accuracy reaches 100% without reducing the rigor of reasoning. In the inductive reasoning, the performance of BARD also drops at first, but it keeps stable and witnesses obvious gains when the hop number is over six. It demonstrates that BARD is better at conducting inductive reasoning and processing multi-hop scenarios compared with text-davinci-003 and ChatGPT. In the abductive reasoning, BARD struggles a lot. It is inferior to ChatGPT but is better than text-davinci-003.

In all, in face of the complex multi-hop scenarios, LLMs still have much room for improvement. From the results, they do relatively well in the deductive reasoning settings. But they are far from good in the inductive and abductive settings, which can also inspire future researches on it.

B Case Studies

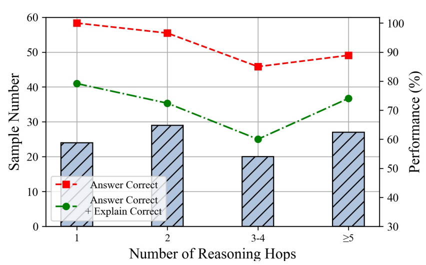
We show one reasoning case for each dataset in Table 6-20, where the context and question as well as output of 0-shot ChatGPT, 1-shot ChatGPT, 3-shot ChatGPT, 0-shot Davinci-003 and 0-shot BARD are displayed. We also provided the annotated information about the answer correctness: ✔, explain correctness: ✔, explain completion: ✔, explain redundancy: ✔, evidence choose wrong: ☑, evidence choose illusion: ☑, no inference process: ☑, inference direction wrong: ☑, and inference process wrong: ☑. (the last five indicators are annotated when the explain completion is false).

C Prompt Engineering

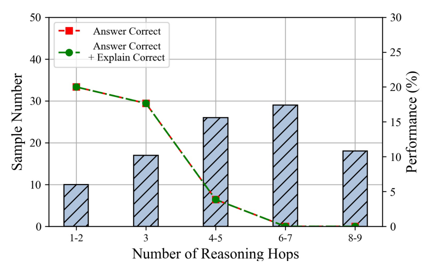
We display the utilized prompts in Table 21.

Table 5: Evaluated Datasets. *Gen.* distinguishes whether the predicted answer is generated text or classified labels. *Explain* denotes whether the explanation is required in the task.

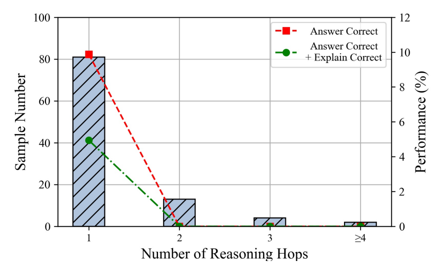
Categories	Dataset	Source	Gen.	Explain	# Davinci	# ChatGPT	# BARD
Deductive	**bAbI-15**	(Weston et al. 2016)	✓	✓	100	1,000	100
	EntailmentBank	(Dalvi et al. 2021)	✓		100	340	100
	RuleTaker	(Clark, Tafjord, and Richardson 2020)		✓	100	1,000	100
	FOLIO	(Han et al. 2022)			100	204	100
	Leap-Of-Thought	(Talmor et al. 2020)			100	1,289	100
Inductive	**bAbI-16**	(Weston et al. 2016)	✓		100	1,000	100
	CLUTRR	(Sinha et al. 2019)	✓		100	1,146	100
Abductive	**α-NLI**	(Bhagavatula et al. 2020)			100	1,000	100
	α-NLG	(Bhagavatula et al. 2020)	✓		100	1,000	100
	AbductiveRules	(Young et al. 2022)	✓		100	1,000	100
	D*-Ab	(Tafjord, Dalvi, and Clark 2021)	✓	✓	100	1,000	100
mixed-form	**ReClor**	(Yu et al. 2020)			100	500	100
	LogiQA	(Liu et al. 2020)			100	651	100
	LogiQA 2.0	(Liu et al. 2023)			100	500	100
	LogiQA2NLI	(Liu et al. 2023)			100	600	100



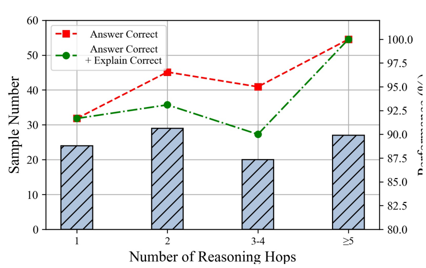
(a) Deductive setting (EntailmentBank) on text-davinci-003.



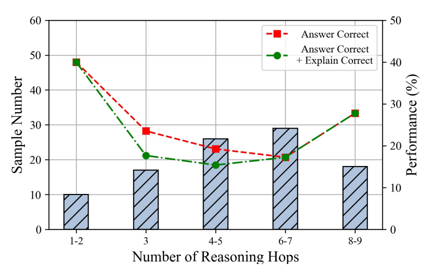
(b) Inductive setting (CLUTRR) on text-davinci-003.



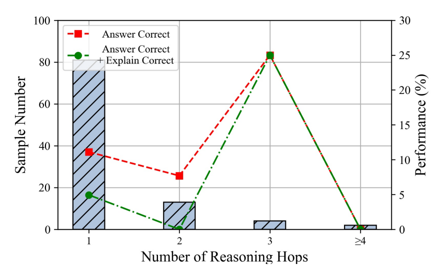
(c) Abductive setting (D*-Ab) on text-davinci-003.



(d) Deductive setting (EntailmentBank) on BARD.



(e) Inductive setting (CLUTRR) on BARD.



(f) Abductive setting (D*-Ab) on BARD.

Figure 9: The performances of text-davinci-003 and BARD under different number of hops. Comparison of deductive, inductive and abductive reasoning settings.

Table 6: A reasoning case in bAbI-15.

Question, answer and explanation	Context: 1 Sheep are afraid of wolves. 2 Mice are afraid of wolves. 3 Winona is a sheep. 4 Gertrude is a sheep. 5 Wolves are afraid of sheep. 6 Emily is a sheep. 7 Cats are afraid of sheep. 8 Jessica is a sheep. Question: What is jessica afraid of? Answer: Wolf Explanation: IsSheep(jessica) (8) $\xrightarrow{1}$ Afraid(jessica, wolf)
ChatGPT of 0-shot	Jessica is afraid of nothing, as it is not specified in the given facts that she has any known natural predator or fear. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Jessica is a sheep and sheep are afraid of wolves. Therefore, it can be deduced that Jessica is afraid of wolves. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 3-shot	Sheep are afraid of wolves and Jessica is a sheep. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Jessica is likely afraid of wolves, since wolves are a common source of fear that other animals, including sheep, mice, and cats, have in common. Wolves are known to be predators of sheep, so Jessica being a sheep, would likely be scared of wolves. Answer and Explain: ✓ ✓ ✓ ✓ Explain Error Type: ☑ ☑ ☒ ☑ ☒
Bard of 0-shot	Jessica is a sheep. Sheep are afraid of wolves. Therefore, Jessica is afraid of wolves. Here is the deductive reasoning: 1. Sheep are afraid of wolves. 2. Jessica is a sheep. 3. Therefore, Jessica is afraid of wolves. The first premise is supported by the fact that sheep are prey animals and wolves are predators. Wolves are known to prey on sheep, and sheep have evolved to be afraid of wolves as a way of self-preservation. The second premise is supported by the fact that Jessica is a sheep. This is a given fact in the prompt. The third premise follows logically from the first two premises. If sheep are afraid of wolves, and Jessica is a sheep, then Jessica must be afraid of wolves. Therefore, the conclusion is that Jessica is afraid of wolves. Answer and Explain: ✓ ✓ ✓ ✓

Table 7: A reasoning case in EntailmentBank.

Question, answer and explanation	Context: 1 pollination is when pollinating animals / wind / water carry pollen from one flower to another flower. 2 if something is a part of a process then that something is required for that process. 3 helping something has a positive impact on that something. 4 if something is required for a process then that something positively impacts that process. 5 an insect is a pollinating animal. Question: life cycle of some plants includes the process of pollination. Pollination is helped by? Answer: wind, water, and insects Explanation: sent1 & sent5→int1: pollination is when insects / wind / water carry pollen from one flower to another; int1 & sent2→int2: insects / wind / water carrying pollen is required for pollination; int2 & sent4→int3: insects / wind / water have a positive impact on pollination by carrying pollen; int3 & sent3→hypothesis
ChatGPT of 0-shot	Pollination in the life cycle of some plants is helped by pollinating animals, such as insects, which are required for the process of pollination and have a positive impact on it. Answer and Explain: ⊗ ⊗ ⊗ ⊗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Pollination in plants is helped by pollinating animals such as insects, which carry pollen from one flower to another as a necessary part of the process, positively impacting the plants' life cycle. Answer and Explain: ⊗ ⊗ ⊗ ⊗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	Pollinating animals such as insects, which are required for the process of pollination to occur. Answer and Explain: ⊗ ⊗ ⊗ ⊗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Pollination is helped by pollinating animals, such as insects, carrying pollen from one flower to another. Answer and Explain: ⊗ ⊗ ⊗ ⊗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Bard of 0-shot	Sure, here is the answer to the question in one sentence using deductive reasoning: Pollination is helped by pollinating animals, such as insects. Here is the simple explanation: Pollination is the process of transferring pollen from the male part of a flower to the female part of a flower. Pollinating animals, such as insects, help pollination by carrying pollen from one flower to another. Without pollination, plants cannot reproduce and produce seeds. Therefore, pollinating animals help the life cycle of some plants by helping them to reproduce. Answer and Explain: ⊗ ⊗ ⊗ ✔ Explain Error Type: ☑ ☒ ☑ ☒ ☒

Table 8: A reasoning case in RuleTaker.

Question, answer and explanation	Context: 1 If the current runs through the bell then the bell is ringing. 2 The circuit includes the battery. 3 If the circuit does not include the switch then the circuit is complete. 4 The circuit includes the bell. 5 If the circuit includes the switch and the switch is on then the circuit is complete. 6 If the current runs through the light bulb then the light bulb is glowing. 7 If the wire is plastic then the wire is not conducting. 8 The circuit includes the switch. 9 If the circuit includes the battery and the battery is not flat then the circuit is powered. 10 If the current runs through the circuit and the circuit includes the bell then the current runs through the bell. 11 The wire is metal. 12 If the wire is metal then the wire is conducting. 13 If the current runs through the circuit and the circuit includes the light bulb then the current runs through the light bulb. 14 If the current runs through the radio then the radio is playing. 15 If the current runs through the circuit and the circuit includes the radio then the current runs through the radio. 16 If the circuit is powered and the circuit is complete and the wire is conducting then the current runs through the circuit. Statement: The current does not run through the circuit. Question: Determine whether the statement is true using deductive reasoning and give explanations. Answer: True Explanation:
ChatGPT of 0-shot	The circuit is powered, complete, and the wire is conducting. Therefore, we can use deductive reasoning to determine that the statement “The current does not run through the circuit” is false based on the given facts. Answer and Explain: Explain Error Type:
ChatGPT of 1-shot	If the battery is not flat and the circuit includes the battery, then the circuit is powered: “If the circuit includes the battery and the battery is not flat then the circuit is powered.” If the circuit is powered and the wire is metal (which is conducting), then the current runs through the circuit: “If the circuit is powered and the circuit is complete and the wire is conducting then the current runs through the circuit.” However, it is stated that the wire is plastic and if the wire is plastic, then it is not conducting: “If the wire is plastic then the wire is not conducting.” Therefore, the statement “The current does not run through the circuit” is true. Answer and Explain: Explain Error Type:
ChatGPT of 3-shot	This statement is false based on the given facts. If the circuit is powered, complete, and the wire is conducting, then the current runs through the circuit, as stated in the first premise. Hence, we can’t conclude that the current doesn’t run through the circuit. Answer and Explain: Explain Error Type:
Davinci-003 of 0-shot	Pollination is helped by pollinating animals, such as insects, carrying pollen from one flower to another. Answer and Explain: Explain Error Type:
Bard of 0-shot	Sure, here is the answer to the question in one sentence using deductive reasoning: Pollination is helped by pollinating animals, such as insects. Here is the simple explanation: Pollination is the process of transferring pollen from the male part of a flower to the female part of a flower. Pollinating animals, such as insects, help pollination by carrying pollen from one flower to another. Without pollination, plants cannot reproduce and produce seeds. Therefore, pollinating animals help the life cycle of some plants by helping them to reproduce. Answer and Explain: Explain Error Type:

Table 9: A reasoning case in FOLIO.

Question, answer and explanation	Context: 1 All growth companies' stocks are volatile. 2 If the stock price is volatile, then it is not suitable for a retirement fund. 3 Some companies' stocks are growth companies' stocks. 4 All mature companies' stocks are suitable for a retirement fund. 5 KO is a mature company's stock. Statement: KO is a company stock and a growth company's stock. Question: Determine whether the statement is true, false, or uncertain using deductive reasoning and give explanations. Answer: False Explanation: MatureCompanies'Stocks(KO) (5)$\xrightarrow{4}$ SuitableForRetirementRund(KO) $\xrightarrow{2r}$¬PriceVolatile(KO)$\xrightarrow{1r}$ ¬Growth Companies'Stocks(KO) (contradictory)
ChatGPT of 0-shot	A mature company's stock cannot also be a growth company's stock since these are two distinct categories. The given facts say that some companies' stocks are growth companies' stocks and all mature companies' stocks are suitable for a retirement fund. Additionally, KO is specifically identified as a mature company's stock. Therefore, the statement "KO is a company stock and a growth company's stock" is false. Answer and Explain: ✓ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	We know from the given facts that KO is "a mature company's stock," but there is no information to suggest that it is also a growth company's stock. Therefore, the statement "KO is a company stock and a growth company's stock" is uncertain based on the given facts. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	The given facts state that KO is a mature company's stock. However, it is not stated whether KO is also a growth company's stock. Therefore, the statement "KO is a company stock and a growth company's stock" is uncertain. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	False. While it is true that KO is a company stock, it is not true that it is a growth company's stock. Growth companies' stocks are volatile, whereas the fact given states that mature companies' stocks are suitable for retirement funds. Since KO is a mature company's stock, it follows that it is not a growth company's stock. Answer and Explain: ✓ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Bard of 0-shot	The statement "KO is a company stock and a growth company's stock" is false. Explanation: 1. All growth companies' stocks are volatile. This means that the stock price of a growth company is likely to fluctuate more than the stock price of a mature company. 2. If the stock price is volatile, then it is not suitable for a retirement fund. This is because retirement funds are designed to provide long-term growth, and volatile stocks are more likely to lose value over time. 3. Some companies' stocks are growth companies' stocks. This means that some companies are growing at a faster rate than others. 4. All mature companies' stocks are suitable for a retirement fund. This is because mature companies are less likely to experience sudden changes in their stock price. 5. KO is a mature company's stock. This means that KO is a company that is not growing at a rapid pace. Based on the above facts, we can conclude that KO is not a growth company's stock. This is because KO is a mature company, and mature companies' stocks are not volatile. Therefore, the statement "KO is a company stock and a growth company's stock" is false. Answer and Explain: ✓ ✓ ✗ ✗

Table 10: A reasoning case in Leap-Of-Thought.

Question, answer and explanation	Context: 1 A bird is not capable of express feelings. 2 A quack-quack is not capable of suffering dehydration. 3 A privet is capable of suffering dehydration. 4 A music is capable of express feelings. Statement: A jazz is capable of express feelings. Question: Determine whether the statement is true using deductive reasoning and give explanations. Answer: True Explanation: IsMusic(jazz) (common sense) $\xrightarrow{4}$ ExpressFeeling(jazz)
ChatGPT of 0-shot	The statement "A jazz is capable of expressing feelings" can be inferred with a certain level of probability based on the given facts. The fact that "A music is capable of expressing feelings" suggests that jazz, which is a genre of music, is likely to be capable of expressing feelings as well. The fact that "A bird is not capable of expressing feelings" and "A quack-quack is not capable of suffering dehydration" are logically irrelevant to the statement about whether jazz is capable of expressing feelings. The fact that "A privet is capable of suffering dehydration" is also irrelevant, as it does not provide any information or indication one way or the other about the statement. Therefore, based on the given facts, the statement about jazz being capable of expressing feelings can be inferred with a certain level of probability and is likely true. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 1-shot	The statement "A jazz is capable of express feelings" is true using deductive reasoning based on the given facts. Answer and Explain: ✓ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	This is true, as the given fact "a music is capable of express feelings" includes the subcategory of music that is jazz. Therefore, we can logically deduce that jazz music is also capable of expressing feelings. Answer and Explain: ✓ ✓ ✓ ✗
Davinci-003 of 0-shot	True. Although humans usually associate the sound of a quack-quack with birds, the quack-quack mentioned here is not a bird and is not capable of expressing feelings. Additionally, a privet is capable of suffering dehydration but not expressing feelings. Therefore, the only species that has been mentioned in the facts that is capable of expressing feelings is the music. Although a jazz is not explicitly mentioned in the facts, it would also be categorically considered to be a type of music, so it can logically be assumed that a jazz is capable of expressing feelings. Answer and Explain: ✓ ✓ ✓ ✓
Bard of 0-shot	The statement "A jazz is capable of express feelings" is not true. The fact that a music is capable of expressing feelings does not mean that a jazz is capable of expressing feelings. A jazz is a type of music, but it is not the only type of music. There are many other types of music, and not all of them are capable of expressing feelings. For example, classical music is often seen as being more formal and less emotional than jazz. Therefore, it is not possible to say that all music is capable of expressing feelings, simply because one type of music is capable of doing so. In order to determine whether a jazz is capable of expressing feelings, we would need more information about the specific jazz in question. We would need to know what the jazz is about, who wrote it, and how it was performed. Once we have this information, we can then use deductive reasoning to determine whether the jazz is capable of expressing feelings. For example, if the jazz is about a personal experience that the author had, then it is more likely that the jazz will be able to express feelings. This is because the author will have put their own emotions into the jazz, and this will be reflected in the music. However, if the jazz is about a more general topic, such as the beauty of nature, then it is less likely that the jazz will be able to express feelings. This is because the author will not have put their own emotions into the jazz, and the music will therefore be more objective. Ultimately, whether or not a jazz is capable of expressing feelings is a matter of opinion. There is no scientific way to determine whether a piece of music is capable of expressing feelings. However, by using deductive reasoning, we can gain a better understanding of the potential for a jazz to express feelings. Answer and Explain: ✗ ✗ ✓ ✓ Explain Error Type: ☒ ☑ ☒ ☒ ☑

Table 11: A reasoning case in bAbI-16.

Question, answer and explanation	Context: 1 Greg is a rhino. 2 Brian is a swan. 3 Bernhard is a frog. 4 Brian is green. 5 Greg is gray. 6 Bernhard is yellow. 7 Julius is a lion. 8 Julius is yellow. 9 Lily is a rhino. Question: What color is Lily? Answer: Gray Explanation: IsRhino(Greg) (1), IsGray(Greg) (5)$\rightarrow$ IsGray(rhino), IsRhino(Lily) (9)$\rightarrow$IsGray(Lily)
ChatGPT of 0-shot	Based on the given facts and inductive reasoning, it is likely that Lily's color is not definitively determined. Although Lily is described as a rhino, there is no information given that directly pertains to Lily's color. Additionally, there is no other animal descriptions that could be used to make an inference regarding Lily's color. Therefore, based on the given facts, it is not possible to make an accurate determination of Lily's color using inductive reasoning. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Lily's color is likely to be unknown as there is no information provided about her color among the other animals. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	Brian, Greg, and Bernhard, which do not directly provide clues about what color Lily may be. Therefore, we cannot be certain about her color based on the information given alone. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Based on the given facts, we can use inductive reasoning to conclude that Lily is gray. This is because all of the other rhinos in the given facts are gray, and there is no evidence to suggest that Lily is any different. In fact, the only thing that distinguishes Lily from the other rhinos is her name, which is a flower that is typically white or yellow. However, we can safely assume that Lily's name is not a reference to her color, as there are many other animals and plants that share the same name. Therefore, the most likely conclusion is that Lily is gray, just like the other rhinos. Here is a more detailed explanation of the inductive reasoning process used to reach this conclusion: 1. We are given the following facts: * Greg is a rhino. * Brian is a swan. * Bernhard is a frog. * Brian is green. * Greg is gray. * Bernhard is yellow. * Julius is a lion. * Julius is yellow. * Lily is a rhino. 2. We notice that all of the rhinos in the given facts are gray. 3. We conclude that there is a strong likelihood that Lily is also gray, as there is no evidence to suggest that she is any different from the other rhinos. 4. We can safely assume that Lily's name is not a reference to her color, as there are many other animals and plants that share the same name. 5. Therefore, the most likely conclusion is that Lily is gray. Answer and Explain: ✓ ✓ ✓ ✓
Bard of 0-shot	Based on inductive reasoning, it is likely that Lily is gray, as both Greg and Julius are gray and they are both rhinos, like Lily. Since the only other colors mentioned among the animals are green and yellow, and none of the rhinos have these colors, it is reasonable to assume that Lily is gray. Answer and Explain: ✓ ✗ ✓ ✓ Explain Error Type: ☒ ☑ ☒ ☑ ☒

Table 12: A reasoning case in CLUTRR.

Question, answer and explanation	Context: 1 [Chuck] took his daughter [Sara] to the park that afternoon to play. 2 [Judy] took her sister, [Maxine], out to dinner for her birthday. 3 [Chuck] took his sister [Judy] out to lunch after learning that she got accepted into her first choice for university. 4 [Maxine] asked her daughter, [Joyce], if she would like to go to a movie with her on Saturday night. 5 [Maxine] and her daughter [Joyce] went to see a movie yesterday and then got ice cream afterwards. 6 [Reynaldo]'s wife, [Maxine], was thrilled to learn she was pregnant. 7 [Naomi] went shoe shopping with her sister [Sara]. Question: What is the kinship between Reynaldo and Naomi? Answer: niece Explanation: [(Reynaldo, wife, Maxine) 6, (Maxine, daughter, Joyce) 5] → (Reynaldo, daughter, Joyce) i1 [(Maxine, sister, Judy) 2, (Judy, brother, Chuck) 3] → (Maxine, brother, Chuck) i2 [(Joyce, mother, Maxine) 4, (Maxine, brother, Chuck) i2] → (Joyce, uncle, Chuck) i3 [(Reynaldo, daughter, Joyce) i1, (Joyce, uncle, Chuck) i4] → (Reynaldo, brother, Chuck) i4 [(Chuck, daughter, Sara) 1, (Sara, sister, Naomi) 7] → (Chuck, daughter, Naomi) i5 [(Reynaldo, brother, Chuck) i4, (Chuck, daughter, Naomi) i5] → (Reynaldo, niece, Naomi)
ChatGPT of 0-shot	There is no direct or indirect relation mentioned between Reynaldo and Naomi in the given facts. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Unclear, as there is not enough information about Naomi's relation to Reynaldo in the given facts. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	The given facts do not provide any information about the kinship between Reynaldo and Naomi, so the question cannot be answered using inductive reasoning. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Reynaldo and Naomi have no direct kinship; they are both related to Maxine, as Reynaldo is her husband and Naomi is Sara's sister. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☑ ☑ ☒ ☒
Bard of 0-shot	Reynaldo and Naomi are siblings-in-law. Inductive reasoning is a type of reasoning that uses specific examples to make a general conclusion. In this case, we can use the following specific examples to make the general conclusion that Reynaldo and Naomi are siblings-in-law: Maxine is the wife of Reynaldo. Sara is the daughter of Maxine. Naomi is the sister of Sara. Therefore, Naomi is the sister-in-law of Reynaldo. Here are some simple explanations for each of the specific examples: Maxine is the wife of Reynaldo because we are told that Reynaldo's wife is Maxine. Sara is the daughter of Maxine because we are told that Maxine has a daughter named Sara. Naomi is the sister of Sara because we are told that Naomi went shoe shopping with her sister Sara. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☑ ☑ ☒ ☒

Table 13: A reasoning case in α-NLI.

Question, answer and explanation	Context: When I was 19, I enlisted with the USMC. When I got back from the hospital, the USMC sent me home. Question: Choose the more likely explanation from following two options: A. I got injured the last day during a training event. B. I was shot in action but didn't regret it. Answer: B
ChatGPT of 0-shot	The more likely explanation is hypothesis choice A: I got injured the last day during a training event. This is because the context mentions the narrator enlisting with the USMC and being sent home after getting back from the hospital, suggesting a military-related scenario. Hypothesis choice A of the narrator getting injured during a training event aligns with these details and provides a probable explanation for why they may have been sent home, as injuries sustained during training may prevent soldiers from fulfilling duties and may require medical attention. Hypothesis choice B of the narrator being shot in action doesn't fit with the context, as there is no indication of any mention of being in combat or any explanation for how this would have been related to the narrator's hospitalization or being sent home, and it doesn't provide an explanation for what may have motivated the narrator's military-related interests or preferences. Answer and Explain: Explain Error Type:
ChatGPT of 1-shot	The correct choice is A. The context suggests that the person was enlisted in the USMC and got sent home from the hospital, indicating that they likely got injured during a training event. Choice B suggests an unlikely scenario where the person got shot in action, which cannot be inferred from the context and contradicts the fact that the person did not regret it. Answer and Explain: Explain Error Type:
ChatGPT of 3-shot	The correct choice is: A. The context suggests that the narrator enlisted with the USMC and was sent home from the hospital. Hypothesis A, where the narrator got injured the last day during a training event, is the more likely explanation because it fits with the idea of being sent home from the hospital after being injured in the line of duty. Hypothesis B, where the narrator was shot in action but didn't regret it, does not make sense as a plausible explanation for being sent home from the hospital expressed in the context. Answer and Explain: Explain Error Type:
Davinci-003 of 0-shot	The more likely explanation is A. I got injured the last day during a training event. This could have been caused by some sort of accident with a weapon or a physical training exercise gone wrong. It is more likely that I would have been injured on the last day of service, as opposed to being shot in action, which would be less likely in the US Marine Corps. Answer and Explain: Explain Error Type:
Bard of 0-shot	The more likely explanation is A. I got injured the last day during a training event. This is because the context states that you enlisted with the USMC at 19, and that you were sent home from the hospital after being injured. This suggests that the injury occurred during training, and that it was serious enough to warrant your discharge from the military. Hypothesis B is less likely because it is not supported by the context. The context does not mention anything about you being shot in action, or about you not regretting it. Additionally, it is not clear why the USMC would send you home if you were not injured. Here is a simple explanation for why you might have been injured during training: You were involved in a training accident. You overexerted yourself and suffered an injury. You were injured by another Marine during training. It is important to remember that these are just possible explanations, and we cannot know for sure why you were injured without more information. Answer and Explain: Explain Error Type:

Table 14: A reasoning case in α-NLG.

Question, answer and explanation	Context: Joy's mom sent her in the donut store for a dozen donuts. She was happy as she left with 13 donuts instead of 12. Question: Generate a short hypothesis for the context and give simple explanations. Answer: Joy got an extra donut.
ChatGPT of 0-shot	Hypothesis: The fact that Joy's mom sent her to the donut store for a dozen donuts, and Joy left with 13 instead of 12 might have been a result of the potential personal or environmental factors related to perception, generosity, or habit. Explanation: The fact that Joy's mom sent her to the donut store for a dozen donuts suggests that it might have some potential personal or environmental factors related to perception, generosity, or habit. Having personal or environmental factors can involve various types of cognitive or practical factors, such as accuracy, generosity, or routine, and can reflect one's sense of responsibility or automaticity. Thus, the fact that Joy's mom sent her to the donut store for a dozen donuts might have been a result of the potential personal or environmental factors related to perception, generosity, or habit. Additionally, the fact that Joy left with 13 donuts instead of 12 implies that there might have been some additional potential cognitive or contextual factors related to accuracy, availability, or preference. Having cognitive or contextual factors can involve various types of perception or action, such as estimation, surplus, or choice, and can reflect one's sense of intention or opportunity. Furthermore, the fact that Joy was happy as she left with 13 donuts instead of 12 suggests that there might have been some potential affective or symbolic factors related to value, satisfaction, or normativity. Having affective or symbolic factors can involve various types of appraisal or intent, such as pleasure, gratitude, or conformity, and can reflect one's sense of social or moral obligation. Therefore, the hypothesis suggests that the fact that Joy's mom sent her to the donut store for a dozen donuts, and Joy left with 13 instead of 12 might have been a result of the potential personal or environmental factors related to perception, generosity, or habit, as well as the potential cognitive or contextual factors related to accuracy, availability, or preference. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☒ ☒ ☑
ChatGPT of 1-shot	Hypothesis: The reason why Joy left the donut store with 13 instead of 12 donuts may be due to a generous gesture from the store clerk, a promotional offer that allowed her to receive an extra donut for free, a mistake made by the store clerk, or Joy's persuasive negotiation or effective communication skills. Explanation: The fact that Joy left the donut store with 13 instead of 12 donuts suggests that there may have been an unexpected addition or alteration to her order. One possibility is that the store clerk may have generously offered Joy an additional donut as a kind gesture or to show appreciation for her business. Another possibility is that there may have been a promotional offer where customers who purchased a dozen donuts received an additional donut for free. Alternatively, a mistake may have been made by the store clerk when fulfilling Joy's order, resulting in an extra donut being added to her purchase. Finally, Joy may have effectively persuaded or negotiated with the store clerk to receive an extra donut, perhaps by demonstrating loyal patronage or simply asking for an additional donut. These factors may have contributed to Joy's pleasant surprise at leaving the donut store with an extra donut, and highlight the importance of customer service, promotions, careful order fulfillment, and communication skills in the food service industry. Therefore, the hypothesis proposes that the reason why Joy left the donut store with 13 instead of 12 donuts may be due to a generous gesture from the store clerk, a promotional offer that allowed her to receive an extra donut for free, a mistake made by the store clerk, or Joy's persuasive negotiation or effective communication skills. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 3-shot	Hypothesis: The donut shop may have made an error and accidentally given Joy an extra donut, which she was pleased to receive. Alternatively, the donut store may have had a special promotion where customers receive an extra donut for free, which Joy was not aware of but happily benefited from. This scenario highlights the potential for unexpected surprises and bonuses in our daily lives. Answer and Explain: ✓ ✓ ✓ ✗
Davinci-003 of 0-shot	Hypothesis: The donut store owner gave Joy an extra donut for free. Explanation: The extra donut could indicate that the store owner gave Joy an extra donut for free as a gesture of goodwill or appreciation. This could be due to Joy being loyal customer at the store, or due to her being a child and the store owner having sympathy for her. Answer and Explain: ✓ ✓ ✓ ✗

Bard of 0-shot	Sure, here are some possible hypotheses for why Joy was happy when she left the donut store with 13 donuts instead of 12: The donut store was giving away a free donut with every dozen purchased. The donut store made a mistake and gave Joy an extra donut. Joy was able to negotiate with the donut store owner to get an extra donut. Joy is a very good friend or family member and the donut store owner wanted to give her a little extra something. Here are some simple explanations for each hypothesis: The donut store was giving away a free donut with every dozen purchased because they were trying to promote their business or because they had too many donuts on hand. The donut store made a mistake and gave Joy an extra donut because they were busy or because they were not paying attention. Joy was able to negotiate with the donut store owner to get an extra donut because she was polite and persuasive. Joy is a very good friend or family member and the donut store owner wanted to give her a little extra something because they appreciated her business or because they were close friends or family. It is impossible to know for sure which hypothesis is correct, but all of them are possible explanations for why Joy was happy when she left the donut store with 13 donuts instead of 12. Answer and Explain:

Table 15: A reasoning case in AbductiveRules.

Question, answer and explanation	Context: 1 Gary is strong. 2 Things that are tiny, are small, and are short, are thin. 3 All people that are quiet, are poor, and are smart, are clever. 4 Gary is big. 5 If something is short, is tiny, and is nice, it is also thin. 6 If a person is bad, is sad, and is strong, that person is rough. 7 Things that are strong, are huge, and are big, are high. 8 People that are bad, are short, and are strong, are also clever. 9 Dave is poor. 10 Dave is imperfect. 11 Charlie is quiet. 12 Gary is heavy. 13 Anne is tiny. 14 All things that are huge, are small, and are big, are also high. 15 All things that are quiet, are huge, and are poor, are also thin. 16 Anne is little. 17 Charlie is nice. 18 Dave is bad. 19 If a person is tiny, is nice, and is sad, that person is high. 20 If a person is quiet, is nice, and is smart, that person is clever. 21 All things that are small, are big, and are smart, are rough. 22 Anne is small. 23 People that are bad, are poor, and are sad, are also rough. Question: Please generate a short explanation for the given context and observation. The observation is: Dave is rough. Answer: Dave is sad. Explanation: 9, 18 $\xrightarrow{23}$ Dave is sad.
ChatGPT of 0-shot	Based on the given context, individuals who possess traits such as badness, sadness, and strength are typically perceived as rough. Dave has been described as poor and imperfect which would suggest a degree of hardship and inadequacy, and the observation confirms that he is perceived as rough. Therefore, it can be inferred that Dave's perceived roughness is a result of his perceived badness, sadness, and poor state in accordance with the given context. Answer and Explain: ✖ ✖ ✖ ✔ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 1-shot	The context indicates that being bad, sad, and strong is associated with being rough, and being poor and imperfect are also negative qualities. The observation that Dave is rough suggests that he may possess these negative traits or exhibit aggressive behavior. Answer and Explain: ✖ ✖ ✖ ✔ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 3-shot	The given observation contradicts the context, as it states that Dave is rough whereas the context indicates that he is poor and imperfect. Therefore, there is no valid explanation for this observation within the given context. Answer and Explain: ✖ ✖ ✖ ✔ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Dave is rough because he is bad, poor, and strong, which meets the criteria of people that are bad, poor, and strong being rough. Answer and Explain: ✖ ✖ ✖ ✔ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Bard of 0-shot	Sure, here is a short explanation for the given context and observation: Dave is rough because he is bad, sad, and strong. He is also poor and imperfect. This suggests that he is also quiet and clever. However, he is also tiny and small. This suggests that he is also weak. Here is a more detailed explanation: Dave is bad. This is evident from the fact that he is described as "bad" and "rough". Dave is sad. This is evident from the fact that he is described as "sad" and "imperfect". Dave is strong. This is evident from the fact that he is described as "strong" and "huge". Dave is poor. This is evident from the fact that he is described as "poor" and "imperfect". Dave is imperfect. This is evident from the fact that he is described as "sad" and "imperfect". Dave is quiet. This is evident from the fact that he is described as "quiet" and "clever". Dave is clever. This is evident from the fact that he is described as "clever" and "strong" Dave is tiny. This is evident from the fact that he is described as "tiny" and "small". Dave is small. This is evident from the fact that he is described as "small" and "weak". Dave's badness, sadness, strength, poverty, imperfection, quietness, cleverness, tininess, and smallness suggest that he is a complex and contradictory person. He is capable of great things, but he is also capable of great evil. He is a force to be reckoned with, and he will not be easily defeated. Answer and Explain: ✖ ✖ ✖ ✔ Explain Error Type: ☑ ☑ ☒ ☑ ☒

Table 16: A reasoning case in D*-Ab.

Question, answer and explanation	Context: Fact1: The bear chases the cat. Fact2: The bear is big. Fact3: The bear is cold. Fact4: The bear likes the squirrel. Fact5: The bear likes the tiger. Fact6: The bear sees the squirrel. Fact7: The cat is big. Fact8: The cat is cold. Fact9: The cat is young. Fact10: The cat likes the bear. Fact11: The cat likes the tiger. Fact12: The cat sees the bear. Fact13: The cat sees the tiger. Fact14: The squirrel is young. Fact15: The tiger chases the bear. Fact16: The tiger likes the squirrel. Rule1: If the tiger is young then the tiger likes the squirrel. Rule2: If something is round then it likes the bear. Rule3: If something likes the bear then it is cold. Rule4: If the tiger likes the bear and the bear sees the squirrel then the squirrel is round. Rule5: If something likes the squirrel then it is round. Rule6: If something chases the bear and the bear likes the cat then the cat likes the squirrel. Question: Please generate a short missing fact for the given context and fact. And give simple explanations. The fact is: The cat likes the squirrel. Answer: The bear likes the cat. Explanation: Fact15 $\xrightarrow{Rule6}$ The bear likes the cat.
ChatGPT of 0-shot	Missing fact: The squirrel is round. Explanation: According to the given rules, if the tiger likes the bear and the bear sees the squirrel, then the squirrel is round. As the tiger likes the squirrel in the given context, we can conclude that the tiger likes the bear, and since the bear likes the squirrel, we can infer that the squirrel is round. Answer and Explain: Explain Error Type:
ChatGPT of 1-shot	The fact that the cat likes the squirrel suggests a potential behavioral or social relationship between the two animals, such as companionship or mutual interest in resources. The missing fact that Anne is young could imply a new or additional character to the context. Answer and Explain: Explain Error Type:
ChatGPT of 3-shot	According to the context, the bear likes the squirrel, and the cat likes the bear, which could suggest a predator-prey relationship between the cat and the squirrel. Moreover, given the fact that the cat likes the squirrel, it is possible that the cat may try to protect the squirrel from the bear's chase. Additionally, the context suggests that. Answer and Explain: Explain Error Type:
Davinci-003 of 0-shot	The missing fact is: The tiger is young. Explanation: The rule1 states that if the tiger is young then the tiger likes the squirrel. The fact16 states that the tiger likes the squirrel, so the tiger must also be young. Answer and Explain: Explain Error Type:
Bard of 0-shot	Sure, here is a short missing fact for the given context and fact: The squirrel sees the cat. This is the missing fact because it is the only fact that can explain all of the given facts and rules. For example, Rule 4 says that if the tiger likes the bear and the bear sees the squirrel then the squirrel is round. However, we know that the squirrel is round, so the squirrel must also see the bear. Here are some simple explanations for the missing fact: The squirrel is friends with the cat. The squirrel is related to the cat. The squirrel is a food source for the cat. I hope this helps! Let me know if you have any other questions. However, I think there is a mistake in the given context. The fact that "The cat likes the squirrel" contradicts the fact that "The cat chases the squirrel". If the cat likes the squirrel, then it would not chase the squirrel. Therefore, one of these facts must be incorrect. Answer and Explain: Explain Error Type:

Table 17: A reasoning case in ReClor.

Question, answer and explanation	Context: A science class stored one selection of various fruits at 30 degrees Celsius, a similar selection in similar conditions at 20 degrees, and another similar selection in similar conditions at 10 degrees. Because the fruits stored at 20 degrees stayed fresh longer than those stored at 30 degrees, and those stored at 10 degrees stayed fresh longest, the class concluded that the cooler the temperature at which these varieties of fruits are stored, the longer they will stay fresh. Question: The class's reasoning is flawed in that the class A. neglected to offer any explanation for the results it discovered B. assumed without proof that its thermometer was reliable C. generalized too readily from the fruits it tested to fruits it did not test D. too readily extrapolated from a narrow range of temperatures to the entire range of temperatures. Answer: D Explanation:
ChatGPT of 0-shot	The correct answer is D. Option A is not relevant to the class's reasoning, as it still made an attempt to draw a conclusion from its results. Option B is a possible flaw, but it is not stated in the argument whether the thermometer was reliable or not, so it cannot be the correct answer. Option C is a possible flaw, but it is not stated whether the class attempted to generalize its results to fruits it did not test, so it cannot be the correct answer. Option D is the correct answer, as the class too readily made a generalization from a narrow range of temperatures (10 degrees, 20 degrees, and 30 degrees) to the entire range of temperatures at which fruits can be stored, which is likely to be much broader. Answer and Explain:
ChatGPT of 1-shot	The correct answer is C. The class's reasoning is flawed in that it generalized too readily from the fruits it tested to fruits it did not test. The experiment only tested a specific selection of fruits and did not test all varieties of fruits, so it is not necessarily true that cooler temperatures will always keep fruits fresh longer. Option A is incorrect because the class did offer an explanation for the results they discovered. Option B is not relevant to the flaw in the class's reasoning. Option D is also not relevant, as the class did not extrapolate from a narrow range of temperatures to the entire range of temperatures. Answer and Explain: Explain Error Type:
ChatGPT of 3-shot	The correct answer is D. The class's reasoning is flawed in that they too readily extrapolated from a narrow range of temperatures to the entire range of temperatures. The class only tested three temperatures, and did not consider the possibility of a threshold beyond which colder temperatures may not be beneficial for all fruits. Therefore, it is too hasty to conclude that "the cooler the temperature, the longer the fruits will stay fresh" based on such a limited range of temperatures. Answer and Explain:
Davinci-003 of 0-shot	The correct answer is C. The class's reasoning is flawed in that the class generalized too readily from the fruits it tested to fruits it did not test. The class was only testing three types of fruit at three different temperatures, which does not truly demonstrate how all different varieties of fruits can react differently to varying temperatures. Thus, the class should not make a hasty generalization about how all fruits would respond to different temperatures. Answer and Explain: Explain Error Type:
Bard of 0-shot	The correct answer is D. Too readily extrapolated from a narrow range of temperatures to the entire range of temperatures. The class only tested fruits at three temperatures: 30 degrees Celsius, 20 degrees Celsius, and 10 degrees Celsius. They concluded that the cooler the temperature, the longer the fruits will stay fresh. However, this conclusion is based on a very small sample size. It is possible that the relationship between temperature and freshness is not linear. For example, it is possible that fruits will stay fresh for the same amount of time at 10 degrees Celsius and 5 degrees Celsius. The class would need to test fruits at a wider range of temperatures to confirm their conclusion. The other options are not correct. Option A is incorrect because the class did offer an explanation for their results: they said that the cooler the temperature, the longer the fruits will stay fresh. Option B is incorrect because the class did not assume that their thermometer was unreliable. Option C is incorrect because the class did not generalize to fruits they did not test. They only tested fruits at three temperatures, and they only made a conclusion about those three temperatures. Answer and Explain:

Table 18: A reasoning case in LogiQA.

Question, answer and explanation	Context: People often use microwave ovens to heat food.Some people think that microwave oven heating changes the molecular structure of food, producing molecules that the human body cannot recognize.These strange molecular formulas are unacceptable to the human body, and some are toxic and may even cause cancer.Therefore, people or animals who often eat microwave foods will have serious physiological changes in their bodies, which will cause serious health problems. Question: Which of the following can best question the above view? A. Microwave heating will not cause more nutrient loss than other cooking methods B. China's microwave oven production standards are consistent with international standards and EU standards C. It is also common for developed countries to use microwave ovens D. The microwave only heats the water molecules in the food, the food does not change chemically. Answer: D Explanation:
ChatGPT of 0-shot	The answer is D. The given argument suggests that microwave oven heating changes the molecular structure of food, producing molecules that the human body cannot recognize and causing serious health problems. Option D provides an alternative explanation for the process of microwave heating, which states that the microwave only heats the water molecules in the food, and the food does not change chemically. Therefore, option D best questions the view presented in the argument. Option A is not relevant as it talks about nutrient loss rather than the effects of microwave heating on food. Option B and option C are not relevant as they provide information about the production standards of microwave ovens and their usage in different countries, but do not impact the argument presented in the context. Answer and Explain:
ChatGPT of 1-shot	The correct choice is: D. The argument suggests that the use of microwave ovens to heat food changes the molecular structure of the food, producing molecules that the human body cannot recognize, which can lead to serious health problems. Option D challenges this argument by stating that the microwave only heats the water molecules in the food and does not change the chemical structure of the food. This suggests that the argument may be based on a misunderstanding of how microwaves work. Option A, B, and C are not relevant to the argument or do not challenge the premise that microwaves can cause health problems. Answer and Explain:
ChatGPT of 3-shot	The correct choice is: D. The microwave only heats the water molecules in the food, the food does not change chemically, can best question the above view. It suggests that the claim about microwave ovens changing the molecular structure of food is not accurate and provides a scientific explanation for how microwave heating works, which does not involve the production of strange or toxic molecules. Answer and Explain:
Davinci-003 of 0-shot	The best answer is B. China's microwave oven production standards are consistent with international standards and EU standards. This answer questions the view presented in the context that microwave oven heating changes the molecular structure of food, since it implies that safety regulations surrounding the production of microwave ovens would never be put in place if this were true. Therefore, this disproves the view that microwave oven heating produces molecules that the human body cannot recognize, which could be toxic and may even cause cancer. Answer and Explain: Explain Error Type:
Bard of 0-shot	The correct answer is D. The microwave only heats the water molecules in the food, the food does not change chemically. The context states that some people think that microwave oven heating changes the molecular structure of food, producing molecules that the human body cannot recognize. However, this is not true. Microwaves only heat the water molecules in food, and the food does not change chemically. This is why microwaved food is safe to eat. The other options are not necessarily correct. Option A is incorrect, as there is some evidence that microwave heating can cause nutrient loss. Option B is irrelevant, as it does not address the question of whether or not microwave oven heating changes the molecular structure of food. Option C is also irrelevant, as it does not address the question of whether or not microwaved food is safe to eat. Here are some additional information about microwaves and their effects on food: Microwaves are a form of non-ionizing radiation, which means that they do not have enough energy to damage DNA. Microwaves heat food by causing water molecules to vibrate. This vibration produces heat, which cooks the food. Microwaves do not change the chemical structure of food. This means that microwaved food is safe to eat. Some studies have shown that microwave heating can cause nutrient loss. However, these studies have been small and the results have been mixed. More research is needed to determine the extent of nutrient loss from microwaved food. Overall, there is no evidence that microwave oven heating is harmful to human health. Microwaved food is safe to eat. Answer and Explain: Explain Error Type:

Table 19: A reasoning case in LogiQA 2.0.

Question, answer and explanation	Context: In a track and field qualifier, Zhang Qiang surpassed Li Jin, and Song Zhi's results were better than Wang Ping, but not as good as Ma Zheng. Question: Which of the following options can be inferred? A. Ma Zheng's grades are better than Zhang Qiang. B. Li Jin's results surpass Wang Ping. C. Zhang Qiang's performance is better than Song Zhi. D. Among the five people, Wang Ping ranked third at most. Answer: D Explanation:
ChatGPT of 0-shot	If Zhang Qiang surpassed Li Jin and Song Zhi's results were better than Wang Ping, but not as good as Ma Zheng, it can be inferred that Zhang Qiang's performance is better than Song Zhi. While the relative positions of Zhang Qiang, Li Jin, Wang Ping, and Ma Zheng are known, there is no information to suggest that Ma Zheng's grades are better than Zhang Qiang (option A) or that Li Jin's results surpass Wang Ping (option B). Option D introduces a ranking that is not directly supported by the information in the passage. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 1-shot	The correct answer is D. From the given information, we know that Zhang Qiang's result was better than Li Jin's, and Song Zhi's result was better than Wang Ping's but not as good as Ma Zheng's. Therefore, we can infer that Ma Zheng's result was better than all the other athletes, including Zhang Qiang, and Li Jin's result was worse than both Zhang Qiang's and Wang Ping's. However, we cannot determine the exact ranking of Wang Ping among the five athletes with the given information, so option D is the correct choice as it cannot be inferred from the given context. Answer and Explain: ✓ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 3-shot	The correct answer is C. The passage states that Zhang Qiang surpassed Li Jin and that Song Zhi's results were better than Wang Ping but not as good as Ma Zheng. Therefore, it can be inferred that Zhang Qiang's performance is better than Song Zhi's since he surpassed Li Jin. Choice A cannot be inferred because Ma Zheng's performance isn't compared to Zhang Qiang's. Choice B also cannot be inferred because there's no comparison between Li Jin's and Wang Ping's performance. Choice D cannot be inferred because there's no information to suggest that Wang Ping ranked third at most. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Davinci-003 of 0-shot	The correct answer is C. Zhang Qianǵs performance is better than Song Zhi. This can be inferred from the given context, as the context stated that "Zhang Qiang surpassed Li Jin", which means Zhang Qiang is better than Li Jin, and "Song Zhiś results were better than Wang Ping", which means Song Zhi is better than Wang Ping, therefore Zhang Qiang is better than Song Zhi. Therefore, option C is the correct choice. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Bard of 0-shot	The correct answer is D. Among the five people, Wang Ping ranked third at most. The passage states that Zhang Qiang surpassed Li Jin, and Song Zhi's results were better than Wang Ping, but not as good as Ma Zheng. This information can be organized into the following table: — Athlete — Results — — Zhang Qiang — Better than Li Jin — — Song Zhi — Better than Wang Ping — — Ma Zheng — Better than Song Zhi — From this table, we can see that Wang Ping's results are not as good as Ma Zheng's, so Wang Ping ranked third at most. The other answer choices are not supported by the passage. Answer choice A is not supported because the passage does not say that Ma Zheng's grades are better than Zhang Qiang's. Answer choice B is not supported because the passage does not say that Li Jin's results surpass Wang Ping's. Answer choice C is not supported because the passage does not say that Zhang Qiang's performance is better than Song Zhi's. Answer and Explain: ✓ ✓ ✓ ✗

Table 20: A reasoning case in LogiQA2NLI.

Question, answer and explanation	Context: The premis is: Researcher: People who participate in opinion surveys often give answers they believe the opinion surveyor expects to hear, and it is for this reason that some opinion surveys do not reflect the actual views of those being surveyed. However, in well-constructed surveys, the questions are worded so as to provide respondents with no indication of which answers the surveyor might expect. So if a survey is well constructed, survey respondents' desire to meet surveyors' expectations has no effect on the survey's results. The conclusion is: the answers of opinion-survey respondents can be influenced by beliefs about the surveyor's expectations even if those beliefs are unfounded. Question: Determine whether the premise and conclusion entailed or NOT entailed. Answer: entailed Explanation:
ChatGPT of 0-shot	The conclusion is not logically entailed by the given premise. The premise discusses the potential for opinion survey respondents to give answers they believe the surveyor expects to hear, but notes that well-constructed surveys prevent this by wording questions in a way that does not indicate expected answers. It then concludes that if a survey is well constructed, respondents' desire to meet surveyors' expectations has no effect on the survey's results. Therefore, it does not support the conclusion that the answers of opinion-survey respondents can be influenced by beliefs about the surveyor's expectations even if those beliefs are unfounded. Answer and Explain: Explain Error Type:
ChatGPT of 1-shot	The conclusion is not entailed by the given premise. While the premise discusses how well-constructed surveys are worded so as to provide respondents with no indication of which answers the surveyor might expect, and how this can mitigate respondents' desires to meet surveyors' expectations, it does not provide information to support or refute the claim that the answers of opinion-survey respondents can be influenced by beliefs about the surveyor's expectations even if those beliefs are unfounded. Therefore, the conclusion cannot be reliably inferred from the given premise. Answer and Explain: Explain Error Type:
ChatGPT of 3-shot	Not entailed. The conclusion about the answers of opinion-survey respondents being influenced by beliefs about the surveyor's expectations cannot be logically derived from the given premise. Answer and Explain: Explain Error Type:
Davinci-003 of 0-shot	NOT entailed. The premise is implying that because the survey is well constructed with no indication of what answers the surveyor might expect, the survey respondents' answers will not be impacted by any beliefs they have of the surveyor's expectations. The conclusion is suggesting that even if those beliefs are unfounded, the responses will still be influenced by them. This goes against the premise as the premise has stated that well-constructed surveys will give no indication of the surveyor's expectations, and therefore, their beliefs of the surveyor's expectations will have no influence. Answer and Explain: Explain Error Type:
Bard of 0-shot	The conclusion is entailed by the premise. The premise states that people who participate in opinion surveys often give answers they believe the opinion surveyor expects to hear, even if those beliefs are unfounded. It also states that well-constructed surveys are worded so as to provide respondents with no indication of which answers the surveyor might expect. This means that even if a survey is well-constructed, respondents may still be influenced by beliefs about the surveyor's expectations. This is because people are often motivated to please others, and they may believe that giving the answer they think the surveyor expects will make them look good. For example, a respondent may believe that a surveyor expects them to support a certain candidate in an election. Even if the survey question is not leading, the respondent may still give the answer they think the surveyor expects, even if they do not actually support that candidate. This is why it is important for surveyors to be aware of the potential for respondents to be influenced by beliefs about their expectations. Surveyors can try to minimize this effect by using clear and unbiased language, and by explaining to respondents that there are no right or wrong answers. Answer and Explain: Explain Error Type:

Table 21: Prompt Engineering.

Dataset	Prompt of zero-shot	Prompt of k-shot
Deductive Reasoning		
bAbI (task 15)	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, answer the following question using deductive reasoning and give simple explanations. The question is: [Question]	There are [k] examples of deductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
EntailmentBank	[zero-shot prompt]: Given facts: [Context]. [Question]. Please answer the question in one sentence using deductive reasoning. And give simple explanations.	There are [k] examples of deductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
RuleTaker	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, determine whether the following statement is true using deductive reasoning and give simple explanations. The statement is: [Statement].	There are [k] examples of deductive reasoning: Given facts: [Context] The statement is: [Statement] The answer is: [Answer] (display k samples) [zero-shot prompt]
FOLIO	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, determine whether the following statement is true, false, or uncertain using deductive reasoning and give simple explanations. The statement is: [Statement].	There are [k] examples of deductive reasoning: Given facts: [Context] The statement is: [Statement] The answer is: [Answer] (display k samples) [zero-shot prompt]
Leap-Of-Thought	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, determine whether the following statement is true using deductive reasoning and give simple explanations. The statement is: [Statement].	There are [k] examples of deductive reasoning: Given facts: [Context] The statement is: [Statement] The answer is: [Answer] (display k samples) [zero-shot prompt]
Inductive Reasoning		
bAbI-16	[zero-shot prompt]: Given facts: [context]. Based on the given facts above, answer the following question using inductive reasoning and give simple explanations. The question is: [Question].	There are [k] examples of inductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
CLUTRR	[zero-shot prompt]: Given facts: [context]. [Question]. Please answer the question in one sentence using inductive reasoning. And give simple explanations.	There are [k] examples of inductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
Abductive Reasoning		
α-NLI	[task description]: Given a context, the abductive reasoning task is to choose the more likely explanation from a given pair of hypotheses choices. And give simple explanations. [zero-shot prompt]: The context is: [Context]. The hypothesis choices are: A. [Option A]. B. [Option B].	[task description] Next, I will give you [k] example(s) for test. The context is [Context]. The hypothesis choice are: A. [Option A]. B. [Option B]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
α-NLG	[task description]: Given a context, the abductive reasoning task is to generate a valid and short hypothesis. [zero-shot prompt]: The context is: [Context]. Please generate a short hypothesis for the context and give simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is [Context]. The correct answer is [Label]. Next, I will give you an example for test. [zero-shot prompt]
AbductiveRules	[task description]: Given a context and an observation, the abductive reasoning task is to generate a valid and short explanation. [zero-shot prompt]: The context is: [Context]. The observation is: [Observation]. Please generate a short explanation for the given context and observation.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The observation is: [Observation]. The explanation is: [Explanation]. Next, I will give you an example for test. [zero-shot prompt]

D*-Ab	[task description]: Given a context and a fact, the abductive reasoning task is to generate a short missing fact. [zero-shot prompt]: The context is: [Context+Rule]. The fact is: [Fact]. Please generate a short missing fact for the given context and fact. And give simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context+Rule]. The observation is: [Observation]. The explanation is: [Explanation]. Next, I will give you an example for test. [zero-shot prompt]
mixed-form Reasoning		
ReClor	[task description]: This is a Mahcine Reading Comprehension task, given the context and question, you are required to choose the correct answer from the answer set and give explanations. [zero-shot prompt]: The context is: [Context]. The question is: [Question]. [Option A]. [Option B]. [Option C]. [Option D]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The question is: [Question]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
LogiQA	[task description]: This is a Mahcine Reading Comprehension task, given the context and question, you are required to choose the correct answer from the answer set and give explanations. [zero-shot prompt]: The context is: [Context]. The question is: [Question]. [Option A]. [Option B]. [Option C]. [Option D]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The question is: [Question]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
LogiQA 2.0	[task description]: This is a Mahcine Reading Comprehension task, given the context and question, you are required to choose the correct answer from the answer set and give explanations. [zero-shot prompt]: The context is: [Context]. The question is: [Question]. [Option A]. [Option B]. [Option C]. [Option D]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The question is: [Question]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
LogiQA2NLI	[task description]: This is a Natural Language Inference task. Please tell whether the premise and conclusion entailed or NOT entailed. And give simple explanations. [zero-shot prompt]: The premise is: [Premise]. The conclusion is: [Conclusion]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The premise is: [Premise]. The context is: [Context]. The correct answer is: [Label]. Next, I will give you an example for test. [zero-shot prompt]