logical equivalent argument adopted to enhance LLM reasoning
but, the question is that I do not believe such thing alone would
boost performance by too much, it would be great if we can witness
the improvement, with some real test-cases.
For the current eval test set, the performance is already at 87% — not much
room for improvement, we will need some better testbed here ...
I recall the paper saying that the LLM simply can NOT plan, integration of LLM
into a pokemon environment seems like a great test-bed, then.

Large Language Models as an Indirect Reasoner: Contrapositive and Contradiction for Automated Reasoning

Yanfang Zhang[1,*] Yiliu Sun[1,*] Yibing Zhan[2]

Dapeng Tao[3] Dacheng Tao[4] Chen Gong[1]

[1]Nanjing University of Science and Technology [2]JD Explore Academy

[3]Yunnan University [4]University of Sydney

Email: {yanfangzhang, chen.gong}@njust.edu.cn

5 February 2024

Abstract

Recently, increasing attention has been drawn to improve the ability of Large Language Models (LLMs) to perform complex reasoning. However, previous methods, such as Chain-of-Thought and Self-Consistency, mainly follow Direct Reasoning (DR) frameworks, so they will meet difficulty in solving numerous real-world tasks which can hardly be solved via DR. Therefore, to strengthen the reasoning power of LLMs, this paper proposes a novel Indirect Reasoning (IR) method that employs the logic of contrapositives and contradictions to tackle IR tasks such as factual reasoning and mathematic proof. Specifically, our methodology comprises two steps. Firstly, we leverage the logical equivalence of contrapositive to augment the data and rules to enhance the comprehensibility of LLMs. Secondly, we design a set of prompt templates to trigger LLMs to conduct IR based on proof by contradiction that is logically equivalent to the original DR process. Our IR method is simple yet effective and can be straightforwardly integrated with existing DR methods to further boost the reasoning abilities of LLMs. The experimental results on popular LLMs, such as GPT-3.5-turbo and Gemini-pro, show that our IR method enhances the overall accuracy of factual reasoning by 27.33% and mathematic proof by 31.43%, when compared with traditional DR methods. Moreover, the methods combining IR and DR significantly outperform the methods solely using IR or DR, further demonstrating the effectiveness of our strategy.

1 Introduction

Recently, pre-trained Large Language Models (LLMs) Wang et al (2022a); Brown et al (2020); Chowdhery et al (2023) have shown great success in various tasks related to language comprehension Touvron et al (2023); Heilbron et al (2022), content generation Liu et al (2023); Agossah et al (2023), and logical reasoning Kojima et al (2022); Wei et al (2022) due to their remarkable ability to infer from the context in zero-shot or few-shot way. This allows the LLMs to generalize well to unseen tasks. Among all the abilities, reasoning ability is perhaps one of the most important aspects for determining the performance of LLMs. To this end, Wei et al (2022) proposed Chain-of-Thought (CoT) prompting to encourage LLMs to explain their reasoning processes by appending some intermediate steps required to reach the answer in the prompt, and this explanation of the reasoning process often leads to improved results. Besides CoT, there are other approaches using prompts to help elicit the reasoning ability of LLMs to better solve the reasoning problems, such as Self-Consistency Wang et al (2022a) and Least-to-Most Zhou et al (2022).

*Equal Contribution

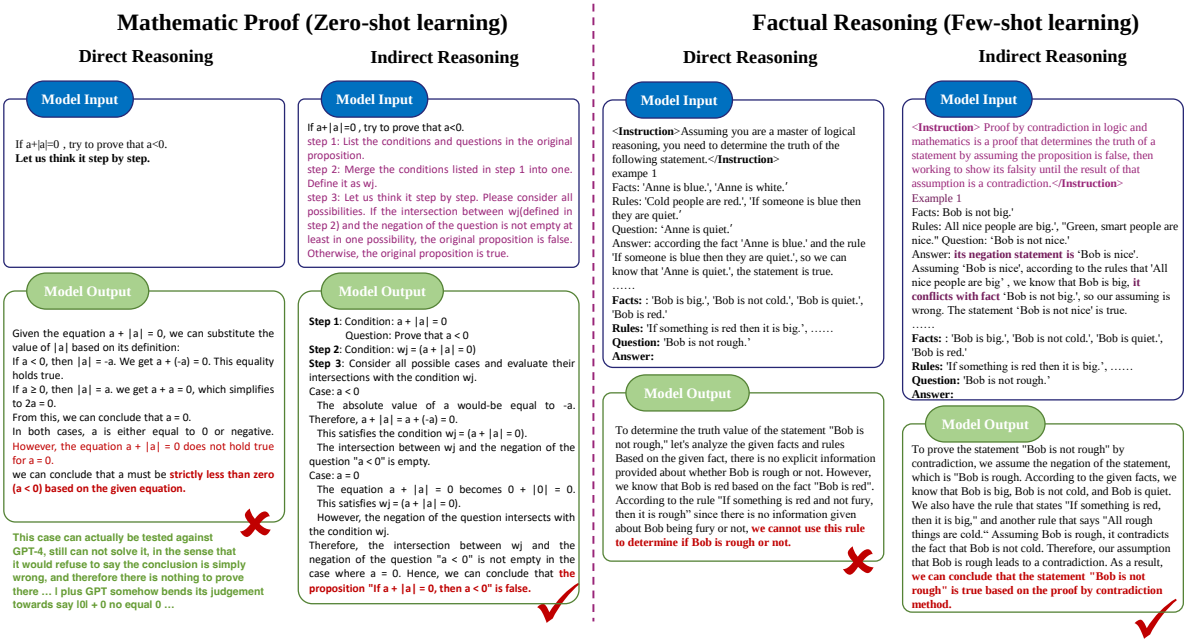


Figure 1: Examples of indirect reasoning using LLMs under zero-shot learning and few-shot learning for complex problems regarding mathematic proof and factual reasoning. Existing direct reasoning methods failed when dealing with these problems. In contrast, our method guides LLMs to use the logics of contrapositive and contradiction, resulting in accurate reasoning and successful deduction to correct answers.

Note that all the above mentioned methods follow the Direct Reasoning (DR) framework, which builds logical chains from given facts to the final result. However, many problems can hardly be proven or reasoned via direct reasoning, as shown in Figure 1. Therefore, when encountering a problem that is difficult to reach a conclusion through direct reasoning, a question arises: whether it is possible to solve the problem by performing indirect reasoning that is logically equivalent to direct reasoning, just like the human thinking process.

In light of above analysis, we aim to provide new chains of utilizing Indirect Reasoning (IR) to overcome the limitations of direct reasoning which enable an alternative effective reasoning process for practical problems. To be specific, this paper introduces the principle of indirect reasoning based on the equivalence of logic, including contrapositive and contradiction which have been well defined in the science of logic Jourdan and Yevdokimov (2016). The former rests on the fact that $p \rightarrow q$ and its contrapositive $\neg q \rightarrow \neg p$ are two logically equivalent statements. The latter corresponds to proof-by-contradiction, which assumes that the statement $p \rightarrow q$ to be false and one needs to show that such an assumption leads to a contradiction. It is important to note that proof-by-contradiction assumes the entire statement to be false, not just the conclusion.

To stimulate LLMs with indirect reasoning, we devise special promptings for contradiction by designing instructions or examples within the intermediate reasoning process, which are illustrated in Figure 1. Specifically, to inject the principles of contrapositive and contradiction mentioned above into LLMs, we preprocess the data using contrapositive equivalence and design prompts to guide LLMs to implement indirect reasoning. As a result, the proposed approach can induce effective indirect reasoning and enhance the ability of LLMs to tackle complex reasoning tasks. Note that, our proposed IR method is embarrassingly simple and does not need any extra code during implementation. Moreover, IR is quite general that can be directly integrated with existing DR methods Wei et al (2022); Wang et al (2022a) based on any foundation model for either few-shot learning or zero-shot learning task. Therefore, we propose to combine Direct and Indirect Reasoning (termed “DIR”) to further improve the reasoning ability of LLMs, where the results of DR and IR are integrated via a simple voting strategy.

To assess the effectiveness of our IR method, we conducted extensive experiments on two popular tasks:

factual reasoning and mathematic proof by using GPT-3.5-turbo and Gemini-pro as foundation models. The results indicate that our proposed prompting method is quite effective in inspiring LLMs to achieve indirect reasoning. For example, we find that the overall accuracy of IR surpasses DR with 27.33% and 31.43% on factual reasoning and mathematic proof tasks, respectively. It is worth mentioning that IR tends to require fewer reasoning steps than DR in some complex problems. In addition, we also investigate the performance of DIR and the experimental results show that DIR significantly outperforms either approach alone. This enriches the reasoning paths of LLMs and improves their overall reasoning ability. Our main contributions are summarized below:

- We introduce the indirect reasoning method, including contrapositive and contradiction, into the reasoning process of LLMs.
- We devise a series of prompt templates that effectively stimulate LLMs to implement indirect reasoning. Our method utilizes the principles of contrapositive and contradiction during the data preprocessing and indirect reasoning process.
- Experimental results show that our proposed indirect reasoning method performs significantly better than existing direct reasoning methods in many problems that direct reasoning methods cannot work well. Furthermore, when combining the proposed indirect reasoning with existing direct reasoning strategies, the reasoning ability of LLMs can be further improved.

2 Related Work

Reasoning ability, as a basic ability of LLMs, has received great attention recently due to its great importance. Despite the notable improvements made by CoT Wei et al (2022), LLMs are still struggling with tasks that require complex or high-order multi-step reasoning, such as factual reasoning and mathematic proof. Therefore, intensive research efforts have been dedicated to addressing the aforementioned issues. Generally, they can be categorized as follows.

Fine-tuning-based methods. These methods aim to improve the reasoning ability of LLMs through supervised fine-tuning. Usually, LLMs are fine-tuned by the samples which require manual labeling of reasoning processes, such as Wang et al (2022b); Ouyang et al (2022). However, it can be labor-intensive due to the costly labeling of complex reasoning processes. The works of Zelikman et al (2022); Shridhar et al (2022) first used LLMs to generate reasoning processes, but only the samples with correct results are selected for fine-tuning LLMs to reduce the labeling cost. Additionally, fine-tuned LLMs on specific tasks can suffer from the problem of "catastrophic forgetting", which means that the original knowledge inherited by the pre-trained LLMs will be lost and thus the ability to generalize to downstream tasks will be weakened. To this end, Cheng et al (2023) trained a prompt retriever using the output scores of LLMs. When fine-tuning, LLMs are frozen just as a data labeler which effectively reduces the impact on LLMs.

Tool-based methods. Tool-based methods propose to utilize external tools to augment the capabilities of LLMs in accomplishing complex tasks Qin et al (2023); Schick et al (2023). Moreover, Yang et al (2023); Jin et al (2023) augment LLMs with external real-time knowledge or domain-specific information through specific tools. Additionally, Retrieval-Augmented Generation (RAG) related methods Gao et al (2023); Ma et al (2023) have received a lot of attention recently, and these methods improve the reasoning ability of LLMs by incorporating external knowledge.

CoT-based methods. CoT-based methods use prompts to help elicit the reasoning ability of LLMs to better solve the reasoning problems Wei et al (2022); Kojima et al (2022); Zhang et al (2022), which is also closely related to our paper. The common CoT methods contain zero-shot CoT Kojima et al (2022) and few-shot CoT Wei et al (2022), where zero-shot CoT guides LLMs to deal with a specific task by adding appropriate instructions in the prompt. Few-shot CoT achieves satisfactory reasoning performance by providing examples that contain the reasoning processes. Meanwhile, recent researches show that different variants of CoT can improve the reasoning ability of LLMs. For instance, the method in Zhang et al (2022) enhances the performance by optimally selecting examples in the prompt. Additionally, external information can be introduced to increase the credibility of results, as proposed in He et al (2022). Some researches propose different approaches Drozdov et al (2022); Yao et al (2023); Besta et al (2023) to decompose complex problems into smaller subproblems to enhance the reasoning ability of LLMs.

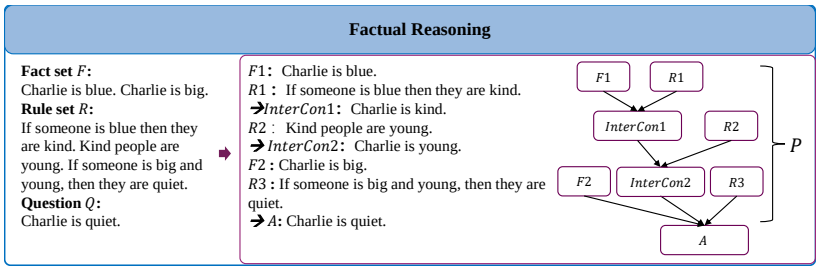


Figure 2: The illustration of some key notions in factual reasoning. Here F, R, A, P, Q, $InterCon$ denote fact, rule, answer, reasoning process, question, and intermediate conclusion, respectively.

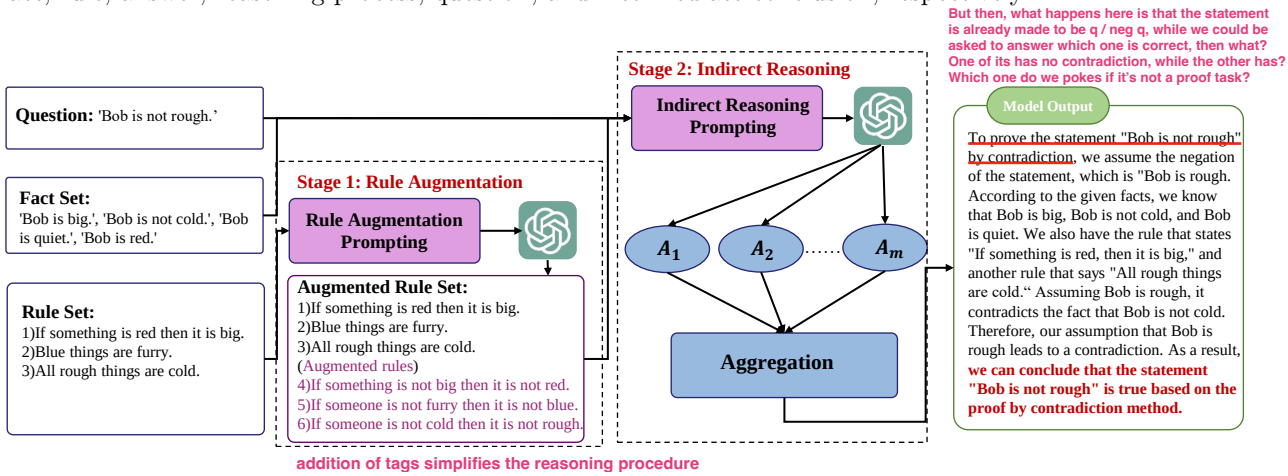


Figure 3: Framework of our proposed method. It consists of two stages including rule augmentation (Stage 1) and indirect reasoning (Stage 2). First, in Stage 1, our method prompts LLMs to augment rule set. Then, in Stage 2, the augmented rule set, the given facts and questions are taken as inputs to implement indirect reasoning with LLMs.

However, as mentioned in the introduction, the previous researches mainly focus on direct reasoning, which will meet difficulties in some complex reasoning procedures. Therefore, our work aims to explore the indirect reasoning which can be combined with these direct reasoning approaches so that they can complement each other to further improve the reasoning ability of various representative LLMs.

3 Preliminary

In this section, we provide some preliminary knowledge which is necessary for formally introducing our method.

3.1 Contrapositive and Contradiction

In mathematics and some practical applications, there are circumstances where direct proof may not be feasible or effective. In such cases, the methods of indirect proof are often used to verify a statement. There are two popular methods for indirect proof, which are: contrapositive method and contradiction method. Next, we will explain these two methods in detail.

Contrapositive. It is based on the fact that an implication is equivalent to its contrapositive, namely:

Insightful argument: 'implication' actually means 'is subset of', p & neg p forms the entire set, therefore always stay true, now q & neg p also stays true means p should be a subset of q, therefore p implies q. Logical derivation is essentially set inclusion manipulation.

$$p \to q \Leftrightarrow \neg p \vee q, \tag{3.1}$$

$$\neg q \to \neg p \Leftrightarrow \neg(\neg q) \vee \neg p \Leftrightarrow q \vee \neg p. \tag{3.2}$$

this is direct corollory from 3.1.

According to the commutative law, one can have: aha! so this is how negation rule gets proved !

$$p \rightarrow q \Leftrightarrow \neg q \rightarrow \neg p. \tag{3.3}$$

Therefore, when we get a fact "If p, then q", we can also know that if $\neg q$ then $\neg p$.

Contradiction. The world-renowned mathematician G. H. Hardy called proof-by-contradiction "one of a mathematician's finest weapons". Actually, this method has been widely used in mathematics, logic, and philosophy to establish the validity of various statements and arguments. Proof-by-contradiction involves the original statement and its negation. These two statements are opposites to each other, meaning that if the original statement is true, then the negation of the original statement is false; and if the original statement is false, then the negation of the original statement is true. Therefore, we consider a reasoning equivalence as:

$$\neg(p \rightarrow q) \Leftrightarrow \neg[(\neg p) \vee q] \Leftrightarrow p \wedge (\neg q). \tag{3.4}$$

3.2 Factual Reasoning and Mathematic Proof

LLMs have shown to be able to conduct factual reasoning in natural language. Reasoning aims to assess the answer A to a candidate question Q and also present the reasoning process P using predefined fact set F and rule set R, where $A \in \{True, False, Unknown\}$ Tafjord et al (2020). All the facts, rules, and questions are expressed in natural language. Figure 2 shows the general illustration of factual reasoning. Mathematic proving problems are similar to factual reasoning. However, it is worth noting that it only gives fact set F and question Q in mathematic proof, and the rule set R are usually set to prior knowledge, which means that we cannot know what rules to use in advance.

4 Methodology

In this section, we provide a detailed introduction to our method, which involves how to stimulate the LLMs to perform indirect reasoning based on the logical equivalences stated in Section 3. Specifically, we propose a method that involves two key stages including rule augmentation (Section 4.1) and indirect reasoning (Section 4.2). As depicted in Figure 3, to begin with, we need to preprocess the data which involves augmenting the input rules. Rule augmentation utilizes the equivalence between the contrapositive proposition and the original proposition to generate additional rules. The augmented rule set, the given fact set and the question can then be utilized as input for subsequent reasoning processes. In the second stage, namely indirect reasoning, we apply the reasoning process of contradiction by explicitly designed prompt templates. The two aforementioned stages are achieved by devising appropriate prompt templates that induce LLMs to implement the theories of contrapositive and contradiction.

4.1 Rule Augmentation

In the process of proving or reasoning, we need to rely on some rules to arrive at a conclusion or proof shown in Figure 2. When faced with complicated rules, LLMs often struggle to fully understand them, which greatly affect the ability of LLMs to use these rules effectively. For instance, when presented with the fact "Bob does not drive to work" and the rule "If the weather is fine, Bob drives to work", humans can apply the equivalence of contrapositive to deduce that the rule is equivalent to "If Bob does not drive to work, the weather is not fine." This allows them to conclude that "The weather is not fine" based on the rule. However, LLMs find it challenging to apply this reasoning of the equivalence of contrapositive. To address this issue, we propose adding the contrapositive of rules to the rule set. This will help LLMs improve their understanding level of rules, as well as their application efficiency. For the example, LLMs can deduce the conclusion "The weather is not fine" directly from the augmented rule and the fact, if the fact "Bob does not drive to work", the rule "If the weather is fine, Bob drives to work" and the augmented rule "If Bob does not drive to work, the weather is not fine" are together as inputs. To induce LLMs to generate contrapositives of rules, we provide a few-shot prompt template as below. The prompt template contains simple instructions and examples of the original rule and its contrapositive counterpart. Figure 8 in Appendix A demonstrates an application of the template on factual reasoning.

```
# <Instruction>The contrapositive is equivalent to the original rule, and now we need to convert the
following rules into their contrapositives.</Instruction>
# Example 1
# Rule: [rule1]
# Contrapositive: [contrapositive1]
...
# Rules: [rules]
# Contrapositives:
```

Interesting Template, I will slot it into my arsenal
This sort of structured tagging should definitely be
helpful for anybody to reason about things.... think about
all the opinion column and KOLs, thinking is energy-consuming
and people will try to avoid it and delegate to somebody else....

4.2 Indirect Reasoning

After the augmented rules mentioned in Section 4.1, we present an indirect reasoning methodology for LLMs. The proposed approach utilizes explicitly designed prompt templates to facilitate effective reasoning to get accurate and reliable results. There are two common types of prompt templates, namely zero-shot prompt templates and few-shot prompt templates. Zero-shot prompt templates provide specific instruction sets to guide LLMs to conduct reasoning without the need to select appropriate labeled examples. This is particularly useful for scenarios in which the present problems are quite different in the tasks, such as mathematic proof problems. On the contrary, the few-shot prompt templates guide the LLMs to reason by similar labeled examples which include intermediate reasoning process. Few-shot prompt templates are more suitable for scenarios where the present problems are similar to the observed examples, such as factual reasoning. By carefully selecting a few examples, we can effectively guide the LLMs in performing similar reasoning tasks. As a result, we have developed zero-shot and few-shot prompt templates for LLMs. These templates are tailored to suit different types of scenarios.

4.2.1 Zero-shot Prompt Template

Based on the principles of contrapositive, we have designed a zero-shot prompt template. Our template guides the LLMs to try to infer that p and $\neg q$ are always contradictory for the problem $p \rightarrow q$ to be proved. Since zero-shot methods often suffer from three pitfalls including calculation errors, missing-step errors, and semantic misunderstanding errors Wang et al (2023), we extend the prompt with more detailed instructions to better alleviate the reasoning errors. Firstly, to reduce errors resulting from missing reasoning steps, we include "Please consider all possibilities" to force the LLMs to think about the problem comprehensively to avoid ignoring some indispensable situations. Secondly, all conditions should be considered simultaneously. Therefore, we take "Merge the conditions listed in Step 1 into one" as Step 2 to avoid the above possible error. Figure 9 in Appendix A shows an application of zero-shot prompt template to mathematic proof. In a short summary, the designed template for zero-shot prompt is as follows:

```
# Question:[Q]
(Instructions)
# Step 1: List the conditions and questions in the original proposition.
# Step 2: Merge the conditions listed in step 1 into one. Define it as wj.
# Step 3: Let us think it step by step. Please consider all possibilities. If the intersection between
wj (defined in step 2) and the negation of the question is not empty at least in one possibility, the
original proposition is false. Otherwise, the original proposition is true.
# Answer:
```

4.2.2 Few-shot Prompt Template

In this part, we propose the few-shot prompt template to elicit LLMs to implement indirect reasoning. In the proposed template, we add instructions for the indirect reasoning method. Simultaneously, the examples of proof-by-contradiction with intermediate reasoning steps are also included in the prompt. In particular, we use the phrase "its negative statement is..." in our examples which helps LLMs understand negating q to $\neg q$. Then the given rules and facts are reasoned to conclude whether there is a contradiction, namely: "It contradicts the assumption/fact...". If $\neg q$ is false, there is a contradiction and we can conclude that q is true. Otherwise, $q's$ truth cannot be proven. To ensure the diversity of examples in the prompt template Zhang et al (2022), it is necessary to include both examples where there is no contradiction and there is a contradiction in the reasoning process. Here is an application of a few-shot prompt template for factual reasoning in Figure 10 in Appendix A. The few-shot template is shown below:

```
# <Instruction> Proof by contradiction is a proof that determines the truth of a question by assuming the
proposition is false, then working to show its falsity until the result of that assumption is a contradiction.
</Instruction>
# Example 1
# Facts: [facts]
# Rules: [rules]
# Question: [Q1]
# Answer: The negation of the original question is ¬Q1. Assuming ¬Q1 is true, ..., it conflicts with the
assumption/fact...
...
# Facts: [facts]
# Rules: [rules]
# Question: [Q]
# Answer:
```

And then, a natural argument & attack on the work is :
how do you find the sweet facts / rules that is not explicitly provided for a real-problem?

5 Extension: Combining Indirect Reasoning with Direct Reasoning

It is worth noting that the proposed Indirect Reasoning (IR) method can be directly combined with the Direct Reasoning (DR) in the existing methods such as Wei et al (2022) and Wang et al (2022a). Therefore, we may further build a Direct-Indirect Reasoning (DIR) framework by combining IR and DR. This will enrich the reasoning paths to solve complex problems. There exist various techniques for aggregating the results of multipath reasoning. One straightforward way is to select the most commonly occurring results, while another involves utilizing the log probability of LLM's outputs. In this paper, we utilize voting to select the most frequently occurring results. Specifically, the voting process can be formulated as:

$$P(A_s) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(A_i = A_s), \tag{5.1}$$

where $\mathbb{I}$ is the indicator function and M is the number of candidate answers, A_s means the final answer with the highest probability $P(A_s)$, and $A_s \in \{True, False, Unknown\}$ is candidate answer.

When there are conflicting voting results, we use LLMs to determine which reasoning is more reliable and the relevant prompt template can be found in Appendix A.

6 Empirical Study

To evaluate the effectiveness of our proposed indirect reasoning method, we apply our method to different reasoning tasks and compare the performance with other typical baseline methods based on direct reasoning. The base LLMs for our experiments include GPT-3.5-turbo and Gemini-pro.

6.1 Evaluation Metrics

The evaluation of reasoning performance for a method includes the correctness investigation on the answer A and the reasoning process P. Therefore, here we use three metrics: accuracy of answer (AA), accuracy of reasoning processes (AP), and the overall accuracy (OA). The overall prediction is deemed as correct when both the answer and the reasoning process are predicted correctly. We introduce three evaluation metrics because Kazemi et al (2022) has shown that even if the reasoning answer of LLMs is correct, the reasoning process generated by LLMs often contains many mistakes. Therefore, OA provides a comprehensive evaluation of the performance of LLMs. The definitions of AA, AP, and OA are:

$$AA = \frac{AN}{N}, AP = \frac{PN}{N}, OA = \frac{ON}{N}, \tag{6.1}$$

where N is the number of examples in the test set. AN, PN, ON are the numbers of examples with correct answer prediction, correct reasoning process prediction, and correct prediction of both answer and reasoning process.

Table 1: Reasoning accuracy of various methods on ProofMath dataset.

		GPT-3.5-TURBO			GEMINI-PRO		
		AA	AP	OA	AA	AP	OA
CoT	DR	82.86%	51.43%	45.71%	**80.00%**	42.86%	31.43%
	IR	**88.57%**	**80.00%**	**77.14%**	77.14%	**65.71%**	**60.00%**
SELF-	DR	85.71%	57.14%	57.14%	**85.71%**	57.14%	48.57%
CONSISTENCY	IR	**97.14%**	**88.57%**	**85.71%**	82.86%	**68.57%**	**62.86%**

Table 2: Reasoning accuracy of various methods on ProofWriter dataset.

		GPT-3.5-TURBO			GEMINI-PRO		
		AA	AP	OA	AA	AP	OA
CoT	DR	46.67%	8.67%	8.67%	**78.00%**	6.67%	6.67%
	IR	**58.67%**	**36.67%**	**36.00%**	76.00%	**27.33%**	**26.67%**
SELF-	DR	42.00%	17.33%	17.33%	**80.00%**	21.33%	21.33%
CONSISTENCY	IR	**70.67%**	**52.00%**	**51.33%**	78.00%	**48.00%**	**47.33%**

6.2 Implementation Details

The indirect reasoning proposed by us can be directly applied to various existing prompt methods. Here we choose the popular CoT-based prompt methods to implement both direct reasoning and indirect reasoning, which are: 1) CoT Wei et al (2022), which guides LLMs to reason by utilizing a few examples containing reasoning processes; and 2) Self-Consistency Wang et al (2022a), which samples multiple reasoning chains and selects the final result by voting. In Self-Consistency method, we independently sample five output results from the decoder. In the experiment, we chose GPT-3.5-turbo and Gemini-pro as the basic LLMs. Among them, the temperature for sampling the output of GPT-3.5-turbo is set to 0.7 as Wang et al (2022a), and the temperature for sampling the output of Gemini-pro is set to 0.9 in Google AI Studio [1]. To improve the reasoning accuracy of our method, we use a voting method for the final answer as mentioned in Section 5.

6.3 Verification of Indirect Reasoning

In this section, we assess the effectiveness of our proposed indirect reasoning method on factual reasoning and mathematic proof tasks, where factual reasoning task corresponds to few-shot scenario and mathematic proving task corresponds to zero-shot scenario. For the sake of clarity and simplicity, as defined previously, here we use IR, DR, and DIR to refer to indirect reasoning, direct reasoning, and the direct-indirect reasoning framework, respectively.

6.3.1 Datasets

Proof Writer benchmark dataset for factual reasoning performance evaluation
* Note that facts and rules are all pre-provided in each data point, not the case in practice here

For factual reasoning task, we adopt ProofWriter Tafjord et al (2020), which is a widely used benchmark dataset for performance evaluation. In this dataset, each data item contains some facts, rules, and questions written in natural language, among which the answer of questions can be either "true" or "false" based on

[1] https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/gemini

Table 3: Reasoning accuracy comparison of DR, IR, and DIR on ProofMath_Hybrid dataset.

	GPT-3.5-TURBO			GEMINI-PRO		
	AA	AP	OA	AA	AP	OA
DR	82.86%	51.00%	39.00%	80.00%	46.00%	34.00%
IR	87.00%	71.00%	57.00%	85.00%	62.00%	52.00%
DIR	**90.00%**	**83.00%**	**77.00%**	**87.00%**	**77.00%**	**73.00%**

Table 4: Reasoning accuracy comparison of DR, IR, and DIR on ProofWriter_Hybrid dataset.

	GPT-3.5-TURBO			GEMINI-PRO		
	AA	AP	OA	AA	AP	OA
DR	51.33%	17.67%	17.67%	**68.00%**	14.00%	13.67%
IR	45.00%	28.00%	26.67%	61.33%	21.33%	20.33%
DIR	**58.67%**	**33.67%**	**32.67%**	63.67%	**28.00%**	**27.00%**

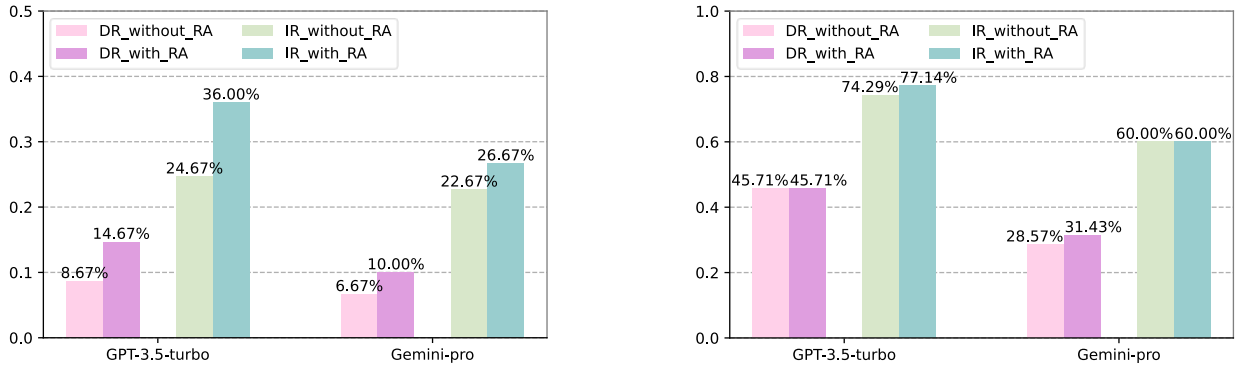
the facts and rules, or marked as "Unknown" if it cannot be proven. For our experiments, we choose 150 such data items from ProofWriter which can demonstrate the utility of IR. The proof of these questions involves the reasoning process of a wide range of depths.

For the mathematic proof task, we constructed a dataset named "ProofMath" by collecting and manually labeling 35 mathematic proof problems of primary and secondary school which should be solved via the contradiction method. We take account of the diversity of the collected data, including the difficulty and representativeness of the problems, to comprehensively evaluate the reasoning ability of the compared approaches on IR task. We have created our own ProofMath dataset as we found that the existing mathematic proof datasets like miniF2F and AMPS are not suitable for our purpose. These datasets contain very few data that can be solved via IR. Therefore, we constructed our dataset to confirm the effectiveness of our proposed IR strategy.

As stated in Section 4.2, the few-shot prompt template is employed for the factual reasoning task, while the zero-shot prompt template is deployed for mathematic proof. All subsequent experiments followed this setup.

6.3.2 Results

Table 1 and Table 2 show the experimental results on the two investigated tasks including factual reasoning and mathematic proof. It can be seen from the results that our IR method significantly outperforms the DR counterparts in different LLMs (*e.g.*, GPT-3.5-turbo and Gemini-pro) and different prompt methods (*e.g.*, CoT and Self-Consistency). By further comparing the results, we can observe that our approach is capable of stimulating distinct LLMs that can effectively implement IR. For OA, GPT-3.5-turbo showed improvements with 27.33% on the factual reasoning task and 31.43% on the mathematic proof task when CoT is used as the base prompt template. Additionally, the Self-Consistency method further improves the accuracy of IR, resulting in higher accuracy than CoT across all scenarios. This is due to that we can correct errors caused by accidental sampling and obtain correct answer by sampling different reasoning paths multiple times, which is also the merits of Self-Consistency. Note that DR slightly outperforms IR on AA when Gemini-pro is deployed. The work in Akter et al (2023) shows that the reasoning ability of complex tasks of Gemini-pro is worse than GPT-3.5-turbo, which leads to more errors in the reasoning process. In addition, IR adds more specifications to make the answer more consistent with the reasoning process than DR. Therefore, AA may be slightly reduced when the reasoning process is wrong. However, regarding the comprehensive metric of



(a) Factual reasoning

(b) Mathematic proof

Figure 4: The effect of rule augmentation on DR and IR.

Question

If a+|a|=0, try to prove that a≤0.

Output of DR

1. Given: a+|a|=0
2. Let's assume that a>0.
3. Since a>0, |a|=a.
4. Substituting |a|=a in the given equation, we get: a+a=0
5. Simplifying further, we get: 2a=0
6. Dividing both sides by 2, we get:a=0
7. But this contradicts our initial assumption that a>0.
8. Therefore, our assumption that a>0 is incorrect.
9. Hence, a≤0.

Output of IR

1: Condition: a+|a|=0. Question: Prove that a≤0
2: Let wj be a+|a|=0
3: If a>0, then a+|a|=2a≠0, which does not satisfy wj.
The original proposition is true.

Figure 5: IR uses fewer reasoning steps to reach a conclusion in mathematic proof.

OA, we see that IR is significantly better than DR no matter CoT or Self-Consistency is adopted.

6.4 Experiments on DIR Framework

As mentioned in Section 5, our IR strategy can be easily combined with the traditional DR methodologies. Therefore, here we also evaluate our established DIR framework on the datasets which contain the hybrid data that might be suitable for either DR or IR. Specifically, for factual reasoning, we randomly select another 150 questions from the ProofWriter dataset in addition to the 150 questions mentioned in Section 6.3.1, to form a new hybrid dataset termed "ProofWriter_Hybrid". The ProofWriter_Hybrid dataset contains problems suitable for both DR and IR, and the required reasoning steps also cover a wide range of 1-hop to 5-hop. Additionally, we carefully collected 65 mathematic proof problems that are suitable for DR, and then combined them with our ProofMath dataset (termed "ProofMath_Hybrid"). We then test our DIR framework on these two hybrid datasets. In this experiment, CoT is adopted for DR, IR, and DIR.

The results of reasoning on the ProofWriter_Hybrid and ProofMath_Hybrid datasets are presented in Table 3 and Table 4. From the tables, it is clear that DIR outperforms either DR or IR alone across all settings except AA on factual reasoning by Gemini-pro, due to similar reasons in Section 6.3.2. Therefore,

Question

Facts: "Anne is green.", "Anne is round.", "Harry is red.", "Harry is young."
Rules: "All furry, young people are not green.", "All young, round people are blue.", "If someone is red and kind then they are round.", "Green, kind people are round.", "If someone is green and round then they are kind.", "All red people are kind.", "If someone is not furry then they are not red.", "All blue, young people are furry."
Question： Harry is not green.

Output of DR

"We can determine the truth of the statement by analyzing the facts and rules provided.
(Step1:) According to the fact 'Harry is red.' and the rule ' All red people are kind.', we can conclude that 'Harry is kind.'
(Step2:) According to the rule 'If someone is red and kind, then they are round.', we can conclude that 'Harry is round.'
(Step3:)According to the rule 'All young, round people are blue.', we can conclude that 'Harry is blue.'
(Step4:)According to the rule 'All blue, young people are furry.', we can conclude that 'Harry is furry.'
(Step5:)According to the rule 'All furry, young people are not green.', we can conclude that 'Harry is not green.'
Based on the facts and rules, the statement 'Harry is not green.' is true."

Output of IR

'To prove the statement "Harry is not green" by contradiction, we assume that Harry is green.
(Step1:) According to the rule "If someone is not furry then they are not red" and the fact "Harry is red", we know that Harry is furry.
(Step2:) But according to the rule "If someone is green then he is not young or he is not furry" and the fact "Harry is young", we can deduce that Harry is not furry.
This is a contradiction, as we reached two contradictory conclusions about whether Harry is furry. Therefore, the assumption "Harry is green" is false and the original statement "Harry is not green" is true.'

Figure 6: IR uses fewer reasoning steps to reach a conclusion in factual reasoning.

DIR can better stimulate the reasoning ability of LLMs by enriching their reasoning paths and encouraging the use of both DR and IR.

6.5 Effect of Rule Augmentation

Coming up with compositional / equivarient rules + Proof by contradiction —> improved reasoning

We conducted ablative experiments on ProofWriter and ProofMath datasets described in Section 6.3.1 to assess the impact of rule augmentation (RA) on IR. As a preprocessing method, rule augmentation can also be applied to DR, so here also we investigate whether it can improve the performance of DR. Figure 4 displays the results of rule augmentation on DR and IR, which are quantified by the metric of OA. The figure demonstrates that the rule augmentation module improves the performances of both DR and IR in most cases, leading to improved reasoning ability for LLMs. When it comes to the tasks involving factual reasoning, we see that augmenting the rules has a noteworthy impact on performance improvement. However, its effect on the tasks related to mathematic proof is not as apparent as in factual reasoning. We think that this might because mathematic proving problems often have fewer rules, some of which may not even contain the given rules.

6.6 Effect of IR for Decreasing Reasoning Steps

In some complex or multi-step reasoning scenarios, IR helps LLMs to reach the answer with fewer steps than DR in many circumstances, thereby decreasing the likelihood of making errors in the reasoning processes of LLMs as shown in Kazemi et al (2022). Two examples of how IR can reduce the number of steps needed for reasoning are provided in Figure 5 and Figure 6.

7 Conclusion

Downplayed factor — efficient discovery & retrieval of relavent facts & rules that are not explicitly mentioned
Finding the auxiliary line is critical in solving geometry problems.

In this paper, we propose an indirect reasoning method to enhance the reasoning power of LLMs by tailored prompts. Indirect reasoning can well compensate for problems which are not directly derivable from known conditions and rules. We validate the effectiveness of the indirect reasoning method in factual reasoning and mathematic proof tasks, and the results well confirm the usefulness of the proposed indirect reasoning strategy. Considering that the IR in this paper only involves the simple thoughts of contrapositive and contradiction, in the future, we can explore the possibility of integrating other more complex logical laws to make LLMs further improve their reasoning skills.

Impact Statements

In recent times, various LLMs have been widely adopted to solve tasks such as factual reasoning, dialog generation, and multi-modal content generation. These approaches have generated noticeable economic value and social impact in multiple applications. Our work builds upon these fundamental LLMs and proposes an algorithm for indirect reasoning to strengthen their reasoning abilities. The experimental results demonstrate that our algorithm can significantly enhance the overall performance of LLMs on various common tasks. In this sense, we believe that our technique will expand the application scenarios of LLMs and inspire people to better leverage AI technology in the future. However, we need to point out that our method may inherit the limitations of popular LLMs and generate incorrect and biased answers sometimes, which may lead to a negative impact on people's work and life.

References

Agossah A, Krupa F, Perreira Da Silva M, Le Callet P (2023) Llm-based interaction for content generation: A case study on the perception of employees in an it department. In: Proceedings of the 2023 ACM International Conference on Interactive Media Experiences, pp 237–241

Akter SN, Yu Z, Muhamed A, Ou T, Bäuerle A, Cabrera ÁA, Dholakia K, Xiong C, Neubig G (2023) An in-depth look at gemini's language abilities. arXiv preprint arXiv:231211444

Besta M, Blach N, Kubicek A, Gerstenberger R, Gianinazzi L, Gajda J, Lehmann T, Podstawski M, Niewiadomski H, Nyczyk P, et al (2023) Graph of thoughts: Solving elaborate problems with large language models. arXiv preprint arXiv:230809687

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al (2020) Language models are few-shot learners. Advances in neural information processing systems 33:1877–1901

Cheng D, Huang S, Bi J, Zhan Y, Liu J, Wang Y, Sun H, Wei F, Deng D, Zhang Q (2023) Uprise: Universal prompt retrieval for improving zero-shot evaluation. arXiv preprint arXiv:230308518

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S, et al (2023) Palm: Scaling language modeling with pathways. Journal of Machine Learning Research 24(240):1–113

Drozdov A, Schärli N, Akyürek E, Scales N, Song X, Chen X, Bousquet O, Zhou D (2022) Compositional semantic parsing with large language models. arXiv preprint arXiv:220915003

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang H (2023) Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:231210997

He H, Zhang H, Roth D (2022) Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv:230100303

Heilbron M, Armeni K, Schoffelen JM, Hagoort P, De Lange FP (2022) A hierarchy of linguistic predictions during natural language comprehension. Proceedings of the National Academy of Sciences 119(32):e2201968,119

Jin Q, Yang Y, Chen Q, Lu Z (2023) Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. ArXiv

Jourdan N, Yevdokimov O (2016) On the analysis of indirect proofs: Contradiction and contraposition. Australian Senior Mathematics Journal 30(1):55–64

Kazemi M, Kim N, Bhatia D, Xu X, Ramachandran D (2022) Lambada: Backward chaining for automated reasoning in natural language. arXiv preprint arXiv:221213894

Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners. Advances in neural information processing systems 35:22,199–22,213

Liu T, Xiong Q, Zhang S (2023) When to use large language model: Upper bound analysis of bm25 algorithms in reading comprehension task

Ma X, Gong Y, He P, Zhao H, Duan N (2023) Query rewriting for retrieval-augmented large language models. arXiv preprint arXiv:230514283

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, et al (2022) Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35:27,730–27,744

Qin Y, Liang S, Ye Y, Zhu K, Yan L, Lu Y, Lin Y, Cong X, Tang X, Qian B, et al (2023) Toolllm: Facilitating large language models to master 16000+ real-world apis. arXiv preprint arXiv:230716789

Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Zettlemoyer L, Cancedda N, Scialom T (2023) Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:230204761

Shridhar K, Stolfo A, Sachan M (2022) Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions. arXiv preprint arXiv:221200193

Tafjord O, Mishra BD, Clark P (2020) Proofwriter: Generating implications, proofs, and abductive statements over natural language. arXiv preprint arXiv:201213048

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288

Wang L, Xu W, Lan Y, Hu Z, Lan Y, Lee RKW, Lim EP (2023) Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. arXiv preprint arXiv:230504091

Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D (2022a) Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:220311171

Wang Y, Mishra S, Alipoormolabashi P, Kordi Y, Mirzaei A, Arunkumar A, Ashok A, Dhanasekaran AS, Naik A, Stap D, et al (2022b) Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. arXiv preprint arXiv:220407705

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35:24,824–24,837

Yang L, Chen H, Li Z, Ding X, Wu X (2023) Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling. arXiv preprint arXiv:230611489

Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K (2023) Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:230510601

Zelikman E, Wu Y, Mu J, Goodman N (2022) Star: Bootstrapping reasoning with reasoning. Advances in Neural Information Processing Systems 35:15,476–15,488

Zhang Z, Zhang A, Li M, Smola A (2022) Automatic chain of thought prompting in large language models. arXiv preprint arXiv:221003493

Zhou D, Schärli N, Hou L, Wei J, Scales N, Wang X, Schuurmans D, Cui C, Bousquet O, Le Q, et al (2022) Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:220510625

A Prompt Templates and Additional Experimental Results

Prompt template for selecting reliable reasoning outputs. Figure 7 illustrates the prompt template that is specifically designed to assist LLMs in selecting the more reliable reasoning outputs.

Reasoning Outputs Selection Prompting

<**Instruction**>Decide which of the following two reasonings is more reasonable based on the facts and the rules.</**Instruction**>

Facts:{facts}
Rules:{rules}
Question:{question}
Reasoning1:{reasoning1}
Reasoning2:{reasoning2}
Answer:

Figure 7: Prompt template for selecting reliable reasoning outputs.

Applications of different prompt templates. Figures 8, Figure 9, and Figure 10 display the applications of different prompt templates: rule augmentation, zero-shot template for proof-by-contradiction, and few-shot template for proof-by-contradiction, respectively.

Rule Augmentation Prompting

<**Instruction**>The contrapositive is equivalent to the original rule, and now we need to convert the following rules into their contrapositives. </**Instruction**>

Example 1:
Rule: All rough things are smart.
Contrapositive: If something is not smart, it is not rough.

Example 2:
Rule: If something is blue and not nice then it is kind.
Contrapositive: If something is not kind, it is not blue or it is nice.
……

Rules:
'If someone is furry then they are young.',
'If someone is furry and cold then they are young.'
Contrapositives:

Outputs

Here are the contrapositive rules for the given rules:

Rule 1: If someone is furry then they are young.
Contrapositive 1: If someone is not young, then they are not furry.

Rule 2: If someone is furry and cold then they are young.
Contrapositive 2: If someone is not young, then they are not furry or they are not cold.

Figure 8: Application of prompt template for rule augmentation on factual reasoning.

Zero-shot prompt template design. (1) Semantic misunderstanding errors. We created a prompt to address semantic misunderstanding errors highlighted in Section 4.2.1. We compared the two templates, and the only difference between the two prompts is whether to merge multiple conditions into one. The examples are given in Figure 11. It can be found that when LLMs are not prompted to merge the multiple conditions

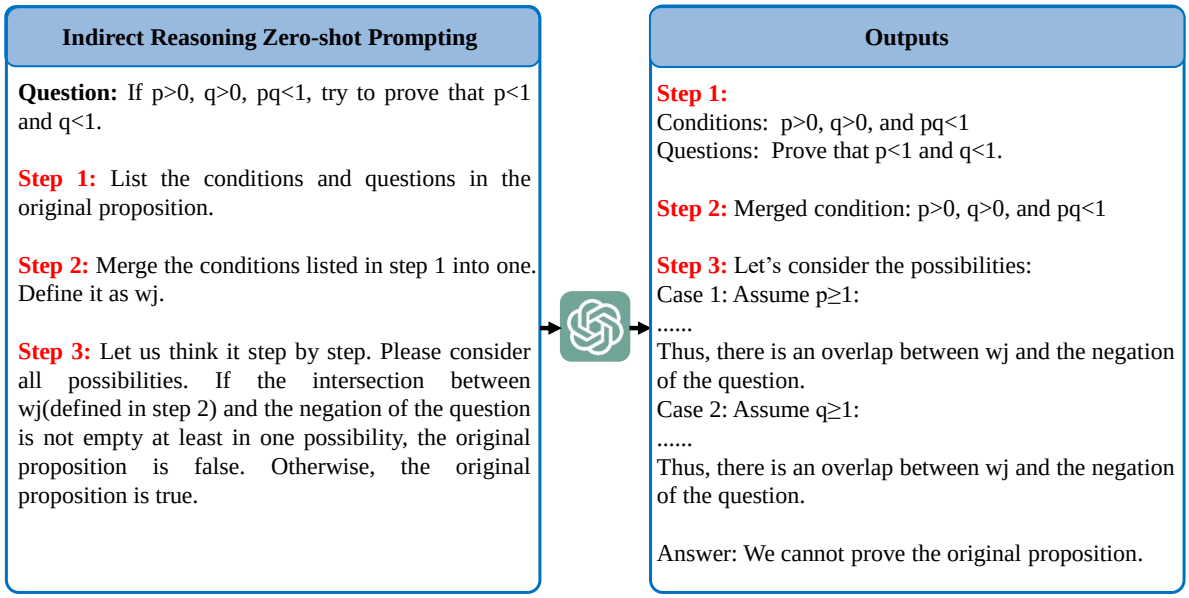


Figure 9: Application of zero-shot prompt template to mathematic proof.

given by the problem, LLMs tend to explore the possibility that each condition $p_i (i = 1, 2, ..., n, p_i \in p)$ and q are valid at the same time. It is important to take all conditions into account at the same time in proof-by-contradiction. Merging multiple conditions into one helps avoid errors. (2) Missing-step errors. In our analysis, we have taken into account the missing-step errors as mentioned in Section 4.2.1. To investigate the impact of these errors, we have removed the phrase "please consider all possibilities" from the prompt. Our findings, as presented in Figure 12, indicate that the language models tend to discuss only a few possible scenarios without this prompt. However, it is important to note that the scenarios missed out could be critical in validating the results. For example, in the proof question "If $a \geq b$ and $b \geq c$, try to prove that $a > c$" , "$a = b = c$" is the critical case of whether the problem can be proven, and LLMs ignore it. This leads to an error in the results. This error is avoided when we prompt LLMs to consider all possibilities.

Rule augmentation can enhance direct reasoning method. Figure 13 illustrates how the direct reasoning method can be improved through rule augmentation.

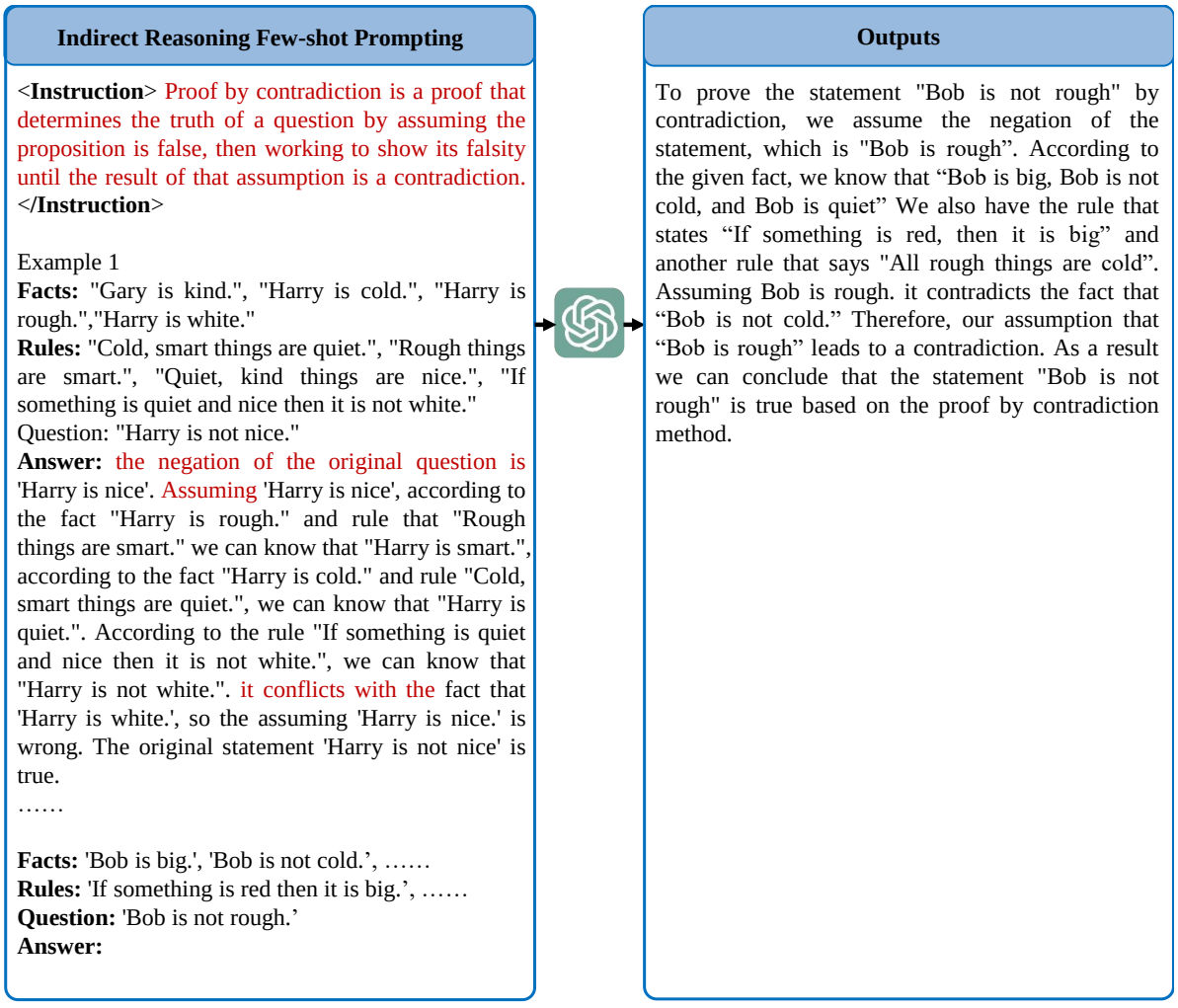


Figure 10: Application of few-shot prompt template for factual reasoning.

Question

If p>0, q>0, and pq<1, try to prove that p<1 and q<1.

Zero-shot Prompting1

step 1: List the conditions and questions in the original proposition.
step 2: Let us think it step by step. Please consider all possibilities. If the intersection between the conditions which should be considered together and the negation of the question is not empty at least in one possibility, the original proposition is false. Otherwise, the original proposition is true.

Zero-shot Prompting2

step 1: List the conditions and questions in the original proposition.
step 2: Merge the conditions listed in step 1 into one. Define it as wj.
step 3: Let us think it step by step. Please consider all possibilities. If the intersection between wj(defined in step 2) and the negation of the question is not empty at least in one possibility, the original proposition is false. Otherwise, the original proposition is true.

Output of Prompting1

Step 1:
Conditions:
$p > 0$ (Condition 1) $q > 0$ (Condition 2) $pq < 1$ (Condition 3)
Question:
Prove that $p < 1$ and $q < 1$.
Step 2:
To prove that $p < 1$ and $q < 1$, we need to show that it is impossible for p or q to be equal to or greater than 1 based on the given conditions.
Let's analyze each condition separately:
Condition 1: $p > 0$
…
Condition 2: $q > 0$
…
Condition 3: $pq < 1$
…
By combining all the conditions, we can conclude that $p < 1$ and $q < 1$ based on the given conditions.

Output of Prompting2

Step 1:
Conditions:
$p > 0$ $q > 0$ $pq < 1$
Questions:
Prove that $p < 1$ and $q < 1$.
Step 2:
Merged condition: $p > 0$, $q > 0$, and $pq < 1$
Step 3:
Let's consider the possibilities:
Case 1: Assume $p \geq 1$:
…
Case 2: Assume $q \geq 1$:
…
we have found at least one possibility where the intersection between wj and the negation of the question is not empty. As a result, the original proposition is false.
In conclusion, we cannot prove that if $p > 0$, $q > 0$, and $pq < 1$, then $p < 1$ and $q < 1$.

Figure 11: The well-thought-out design of the template can reduce semantic misunderstanding errors.

Question

If a≥b and b≥c, try to prove that a>c.

Zero-shot Prompting1

Step 1: List the conditions and questions in the original proposition.
Step 2: Merge the conditions listed in step 1 into one. Define it as wj.
Step 3: Let us think it step by step. If the intersection between wj(defined in step 2) and the negation of the question is not empty, the original proposition is false. Otherwise, the original proposition is true.

Zero-shot Prompting2

Step 1: List the conditions and questions in the original proposition.
Step 2: Merge the conditions listed in step 1 into one. Define it as wj.
Step 3: Let us think it step by step. Please consider all possibilities. If the intersection between wj(defined in step 2) and the negation of the question is not empty at least in one possibility, the original proposition is false. Otherwise, the original proposition is true.

Output of Prompting1

Step 1:
Conditions:
a ≥ b
b ≥ c
Question:
Prove that a> c.
Step 2:
Merged condition: a ≥ b and b ≥ c
Step 3:
Let's consider the possibilities:
Case 1: Assume a > b > c:
In this case, the conditions a ≥ b and b ≥ c are satisfied since a is greater than b.
Moreover, a is indeed strictly greater than c.
There is one overlap between wj and the negation of the question.
In conclusion, we can prove it.

Output of Prompting2

Step 1:
Conditions:
a ≥ b
b ≥ c
Question:
Prove that a> c.
Step 2:
Merged condition: a ≥ b and b ≥ c
Step 3:
Let's consider the possibilities:
Case 1: Assume a = b = c:
In this case, the conditions a ≥ b and b ≥ c are both satisfied since they are equal.
Therefore, there is no overlap between wj and the negation of the question.
Case 2: Assume a > b:
In this case, the conditions a ≥ b and b ≥ c are satisfied since a is greater than b.
Moreover, a is indeed strictly greater than c.
There is one overlap between wj and the negation of the question.
In conclusion, we cannot prove it.

Figure 12: The well-thought-out design of the template can reduce missing-step errors.

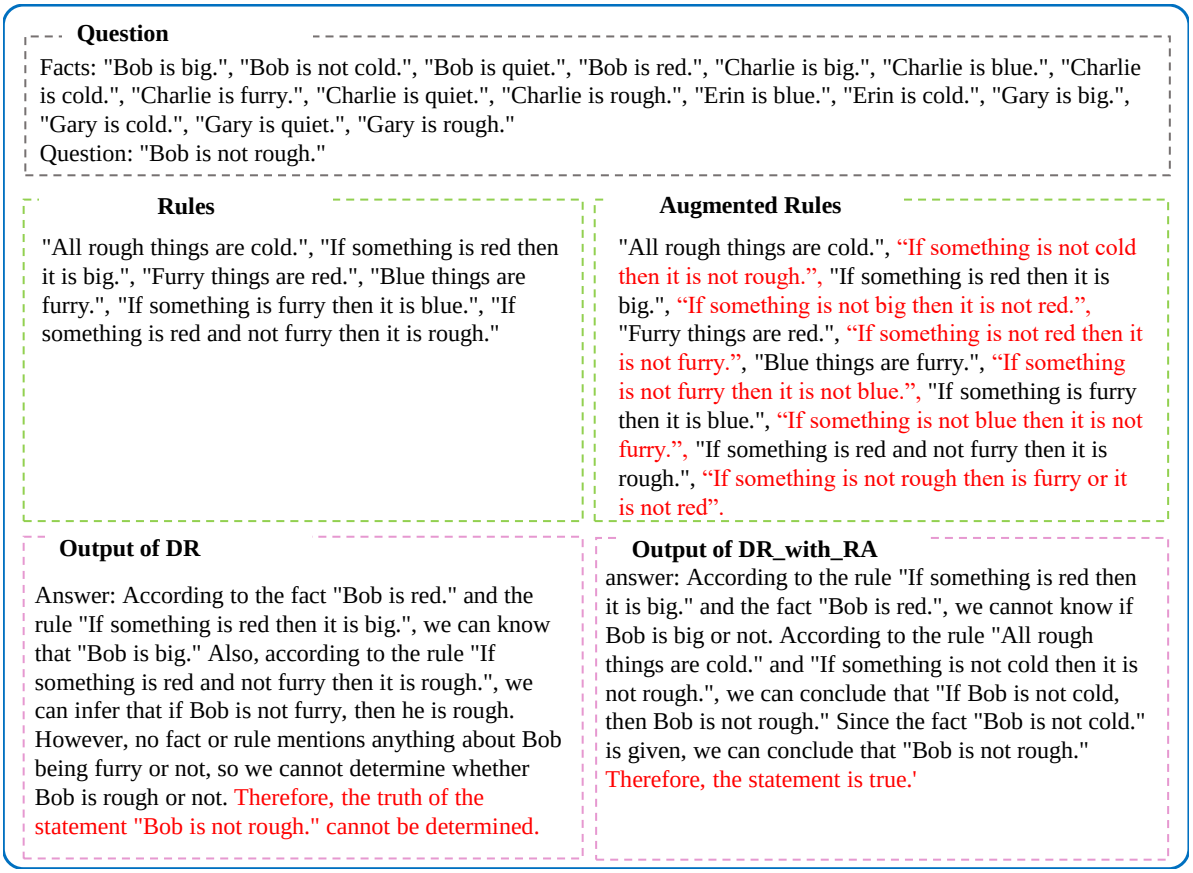


Figure 13: Rule augmentation can enhance direct reasoning method.