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[5]

[1]Nanjing University of Science and Technology, [2]JD Explore Academy, [3]Yunnan University,
[4]Nanyang Technological University, [5]Shanghai Jiao Tong University

Correspondence: goodgongchen@gmail.com

Abstract

Recently, increasing attention has been focused on improving the ability of Large Language Models (LLMs) to perform complex reasoning. Advanced methods, such as Chain-of-Thought (CoT) and its variants, are found to enhance their reasoning skills by designing suitable prompts or breaking down complex problems into more manageable sub-problems. However, little concentration has been put on exploring the reasoning process, *i.e.*, we discovered that most methods resort to Direct Reasoning (DR) and disregard Indirect Reasoning (IR). This can make LLMs difficult to solve IR tasks, which are often encountered in the real world. To address this issue, we propose a Direct-Indirect Reasoning (DIR) method, which considers DR and IR as multiple parallel reasoning paths that are merged to derive the final answer. We stimulate LLMs to implement IR by crafting prompt templates incorporating the principles of contrapositive and contradiction. These templates trigger LLMs to assume the negation of the conclusion as true, combine it with the premises to deduce a conclusion, and utilize the logical equivalence of the contrapositive to enhance their comprehension of the rules used in the reasoning process. Our DIR method is simple yet effective and can be straightforwardly integrated with existing variants of CoT methods. Experimental results on four datasets related to logical reasoning and mathematic proof demonstrate that our DIR method, when combined with various baseline methods, significantly outperforms all the original methods.

1 Introduction

Recently, pre-trained Large Language Models (LLMs) (Wang et al., 2022a; Chowdhery et al., 2023; Dubey et al., 2024) have shown great success in various tasks related to language comprehension (Touvron et al., 2023; Nam et al., 2024), content generation (Agossah et al., 2023; Liu et al., 2023; Dai et al., 2024), and logical reasoning (Kojima et al., 2022; Wei et al., 2022; Dubey et al., 2024) due to their remarkable ability to infer from the context in zero-shot or few-shot way. To enhance the reasoning ability of LLMs, CoT (Wei et al., 2022) encourages LLMs to explain their reasoning processes by appending some intermediate steps required to reach the answer in the prompt. 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).

Note that most of the above mentioned methods simply perform Direct Reasoning (DR), which involves constructing logical chains from premises to the final result. However, many problems can hardly be proven or reasoned via DR. Therefore, when encountering a problem that is difficult to reach a conclusion through DR, a question arises: *whether it is possible to solve the problem by performing other reasoning strategies, such as Indirect Reasoning (IR) that is logically equivalent to DR?* IR, including contrapositive and contradiction in this paper, is logically equivalent to DR, and they have been well defined in the science of logic (Jourdan and Yevdokimov, 2016).

To stimulate LLMs with IR, we devise special prompts for contradiction and contrapositive by designing instructions or examples within the intermediate reasoning process as shown in Figure 1. As a result, the proposed approach can induce effective IR and enhance the ability of LLMs to tackle complex reasoning tasks. Note that, IR is embarrassingly simple and general that can be directly combined with DR. Therefore, we propose Direct-Indirect Reasoning (termed "DIR") to further improve the reasoning ability of LLMs. Without loss of generality, we merge the results of DR and IR using the most common voting methods. Moreover, DIR can be seamlessly integrated with existing variants of CoT based on any foundation model.

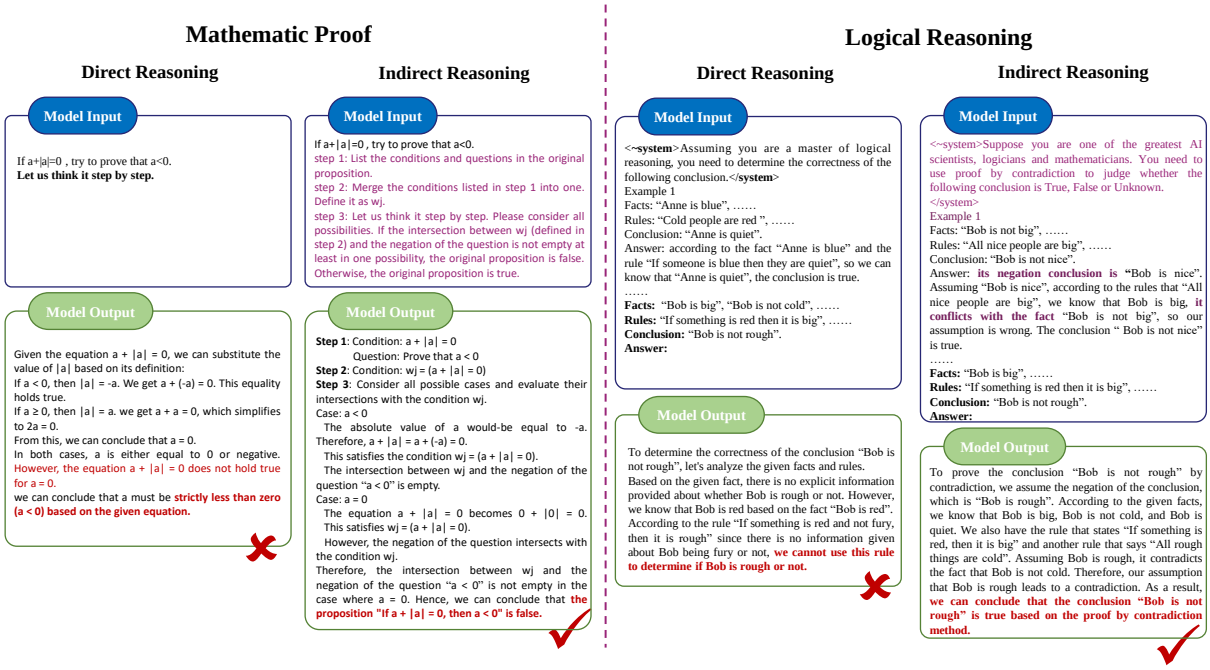

Figure 1: Examples of IR using LLMs for some complex problems regarding mathematic proof and logical reasoning. Existing DR methods failed when dealing with these problems. In contrast, our method guides LLMs to use the logics of contrapositive and contradiction, resulting in accurate reasoning and successful deduction to correct answers.

To assess the effectiveness of our DIR method, we conducted extensive experiments on two popular tasks: logical reasoning and mathematic proof, by using various LLMs as foundation models. The results indicate that our proposed method is quite effective in inspiring LLMs to achieve IR. For example, DIR has shown a noteworthy improvement of over 10.0% in terms of the accuracy of reasoning processes of mathematic proof task. Additionally, in the logical reasoning task, it consistently outperforms various baselines and LLMs, particularly on the data that DR struggles with. The improvement is quite impressive, with a 33.4% increase in accuracy. In particular, experimental analyses have demonstrated that the utilization of IR can aid LLMs in resolving some tasks that are arduous to accomplish through the use of DR. This enriches the reasoning paths of LLMs and improves their overall reasoning ability. Our main contributions are summarized below:

- We introduce the IR strategy, including contrapositive and contradiction, into the reasoning process of LLMs.
- We devise a series of prompt templates that effectively stimulate LLMs to implement IR. We further introduce the DIR method, which combines DR and IR to enhance the reasoning ability of LLMs.
- Experimental results indicate that our proposed DIR method can be effectively combined with the variants of CoT. These combined methods have shown significant performance improvement across four logical reasoning and mathematic proof benchmark datasets on three different baseline LLMs. Additionally, our method has demonstrated impressive performance in inspiring diverse reasoning chains and solving complex problems that can hardly be solved by DR.

2 Motivation and Problem Formulation

LLMs have shown strong ability to conduct logical reasoning in natural language. The aim of reasoning is to assess the answer A to a candidate conclusion or question Q, and also present the reasoning process PR from premises P which include fact set F and rule set R (Tafjord et al., 2021). All the premises and conclusions are expressed in natural language. Figure 2 shows the general illustration of logical reasoning. Mathematic proving problems are similar to logical reasoning. However, it is worth noting that it only gives fact set F and question Q, and the rule set R is usually set to prior knowledge, which means that we cannot know what rules to use in advance.

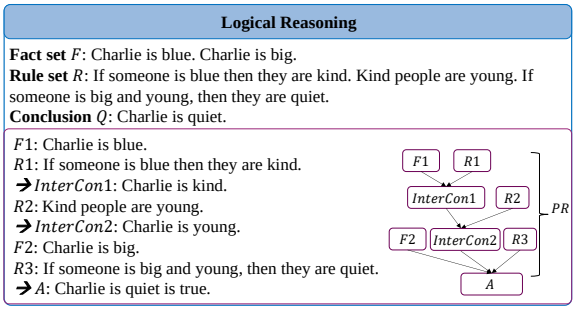

Figure 2: The illustration of some key notions in logical reasoning. The data is from ProofWriter dataset. Here P, F, R, A, PR, Q, $InterCon$ denote premise, fact, rule, answer, reasoning process, conclusion, and intermediate conclusion, respectively. The illustration of mathematic proving problem is put to appendix due to space limitation.

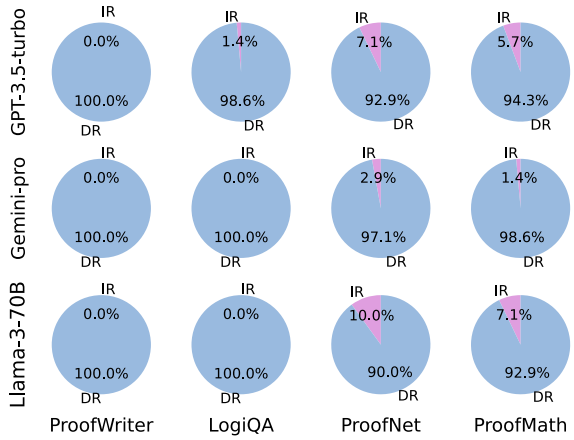

Figure 3: The proportion of IR and DR implementations deployed by LLMs on various datasets.

We notice that LLMs may encounter challenges with IR tasks, despite being proficient at DR, as illustrated in Figure 1. To further understand this phenomenon, we analyzed it by investigating whether LLMs tend to use DR in solving problems. To this end, we conducted a preliminary experiment, where 70 questions are randomly selected from each of the four datasets (*i.e.*, ProofWriter, LogiQA, ProofNet, and ProofMath datasets). In these experiments, we prompt LLMs with "Let's think step by step" and calculate the proportion of DR and IR implemented by LLMs.

According to Figure 3, we see that LLMs rarely use IR on logical reasoning tasks, even when IR would be more appropriate. They still prefer to use DR to solve mathematic problems. Meanwhile, to our best knowledge, currently there are no relevant works explicitly performing IR. Therefore, we propose to stimulate LLMs to implement IR more effectively, which could improve their overall reasoning ability.

3 Methodology

In this section, we present a comprehensive overview of our DIR approach as shown in Figure 4. Our method begins with an introduction to the principles of IR, which include contradiction and contrapositive (Section 3.1). Then we outline a method for guiding LLMs in the application of IR by devising prompt templates that implement the reasoning process of contradiction and contrapositive (Section 3.2). Lastly, we provide a detailed description of the combination method for DR and IR in Section 3.3.

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:

$$p \to q \Leftrightarrow \neg p \vee q, \tag{1}$$

$$\neg q \to \neg p \Leftrightarrow \neg(\neg q) \vee \neg p \Leftrightarrow q \vee \neg p. \tag{2}$$

According to the commutative law, one can have:

$$p \to q \Leftrightarrow \neg q \to \neg p. \tag{3}$$

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, the negation of the original statement is false; and if the original statement is false, the negation of the original statement is true. Therefore, we consider a reasoning equivalence as:

$$\neg(p \to q) \Leftrightarrow \neg[(\neg p) \vee q] \Leftrightarrow p \wedge (\neg q). \tag{4}$$

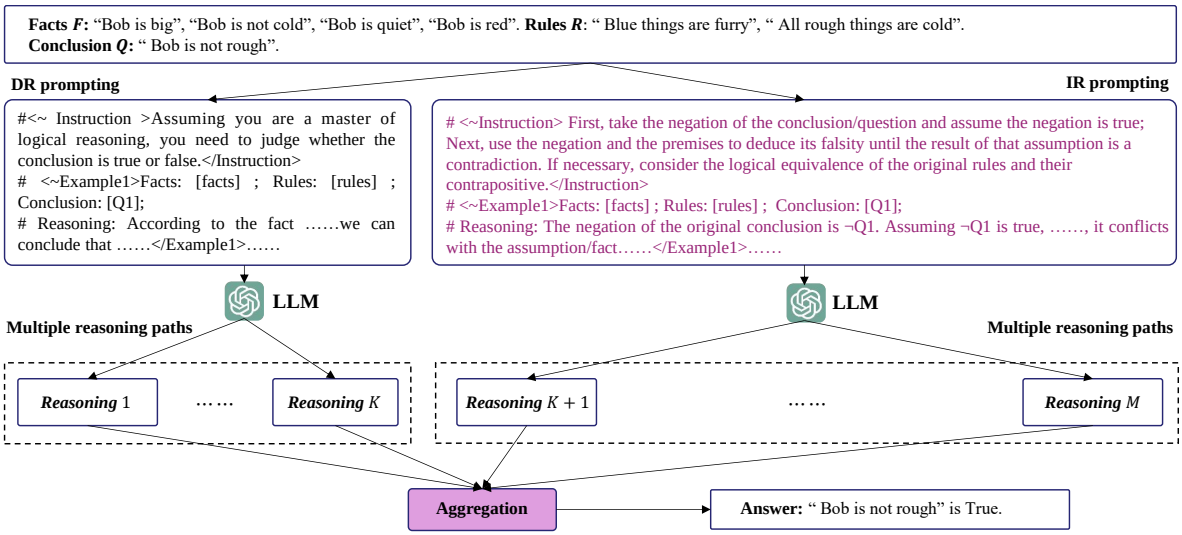

Figure 4: Framework of our proposed DIR method which incorporates both DR and IR to enable multipath reasoning. Our approach involves the IR component, which stimulates LLMs through a set of crafted prompt templates that incorporate the thought of contradiction and contrapositive. Ultimately, the outcomes of both DR and IR are consolidated through majority voting.

3.2 Indirect Reasoning

In the principles of IR described in the previous section, we inspire LLMs to implement IR by designing appropriate prompt templates as shown in Figure 4. To implement contradiction, the entire reasoning process is involved. First, we take the negation of the conclusion and assume it to be true. Subsequently, we deduce the negation along with the premises until a conflict arises. Finally, we conclude that the negation of the conclusion is false, and therefore, the original conclusion must be true. In addition, as depicted in Figure 2, certain rules are employed during the reasoning process. Based on the principle of contrapositive discussed earlier, we can deduce that the contrapositive of these rules and their original rules are logically equivalent. The contrapositive can assist LLMs in enhancing their comprehension of rules and their ability to apply them efficiently. 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. In the following, we present relevant instructions and examples with IR processes to achieve contradiction and contrapositive.

Zero-shot Scenario. We implement a contradiction by following instructions: "First, take the negation of the conclusion/question and assume the negation is true; Next, use the negation and the premises to deduce its falsity until the result of that assumption is a contradiction". Also for contrapositive, LLMs are prompted using "If necessary, consider the logical equivalence of the original rules and their contrapositive".

Few-shot Scenario. In addition to the above instructions, the examples with intermediate reasoning steps incorporating contradiction and contrapositive also facilitate LLMs to implement IR. To facilitate the effective implementation of IR for LLMs, we craft a set of prompt templates that incorporate the concepts of contradiction and contrapositive into the reasoning process (see Appendix E).

3.3 Direct-Indirect Reasoning

From the above description, it can be inferred that the proposed IR method can be directly combined with DR in existing methods, such as SC (Wang et al., 2022a), CR (Zhang et al., 2023) and MulAD (Du et al., 2024). Therefore, we propose a DIR method by combining IR and DR. This will enrich the reasoning paths to solve complex problems. It can be seen that IR is a straightforward approach that involves negating the conclusion and treating the negation as a premise. That is to say, IR does not impose any additional constraints on the reasoning process. Therefore, the proposed DIR method can be easily incorporated into existing CoT variants to improve reasoning proficiency.

Various techniques exist for aggregating the re-

sults of multipath reasoning. One straightforward way is to select the most commonly occurring results, while another involves utilizing the log probability of the output of LLMs. In this paper, we utilize voting to select the most frequently occurring results. We sample M candidate reasoning chains from LLMs and $\{A_i\}_{i=1}^{M}$ is the set of answers generated from these chains. Let $\mathcal{A} = \left\{\widehat{A}_s\right\}_{s=1}^{|\mathcal{A}|}$ be the set of all possible answers for the question. We then select the answer from $\mathcal{A}$ with the highest probability $P(\widehat{A}_s)$ and $P(\widehat{A}_s)$ can be formulated as below:

$$P(\widehat{A}_s) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(A_i = \widehat{A}_s), \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

4 Empirical Study

4.1 Setup

To evaluate the effectiveness of our proposed DIR method, we apply our method to four typical reasoning datasets, namely ProofWriter, LogiQA, ProofNet, and ProofMath. Here we choose the popular CoT-based prompt methods to implement both DR and IR, which are:

- CoT (Wei et al., 2022). It guides LLMs to reason by utilizing a few examples containing reasoning process.
- SC (Wang et al., 2022a). It samples multiple reasoning chains and selects the final result by voting.
- CR (Zhang et al., 2023). It is similar to ToT (Yao et al., 2024) and solves problems using a thought search tree. However, CR stores all the historically correct intermediate thoughts.
- MulAD (Du et al., 2024). It employs multiple agents, each powered by an LLM, to propose and debate their individual reasoning processes over multiple rounds to arrive at a common final answer.

In subsequent experiments, the proposed DIR algorithms are configured to have the same number of reasoning candidates sampled from LLMs as their corresponding baseline algorithms. This configuration can make the computational complexity to be consistent among DIR and baseline methods in the experiment. Additionally, the number of reasoning candidates sampled from LLMs for DR and IR in DIR is set to be equal. Specifically, for LogiQA and ProofWriter datasets, we set the number of reasoning candidates to 16, and for ProofNet and ProofMath datasets, it is set to 4. For more detailed parameter settings in the implementation please refer to Appendix B and the designed prompt templates for the experiment are available in Appendix E.

	GPT-3.5-turbo	Gemini-pro	Llama-3-70B
CoT	28.5%	26.3%	26.8%
SC	29.1%	27.4%	29.6%
SC-DIR	**36.3%**	**31.3%**	**34.6%**
CR	25.1%	22.3%	22.9%
CR-DIR	**29.1%**	**25.7%**	**26.8%**
MulAD	31.3%	26.8%	30.2%
MulAD-DIR	**38.0%**	**31.8%**	**35.8%**

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

4.2 Evaluation Metrics

The evaluation of reasoning performance for a method includes the correctness investigation on the answer and the reasoning process. Therefore, here we use three metrics, namely accuracy of answer (AA), accuracy of reasoning processes (AP), and diversity of reasoning processes (DP). We use these three indicators to comprehensively evaluate the quality of reasoning or proof. The definitions of AA, AP, and DP are:

$$AA = \frac{AN}{N}, AP = \frac{PN}{N}, DP = \frac{1}{N}\sum d_i, \tag{6}$$

where N is the number of examples in the test set; AN and PN are the numbers of examples with correct answer prediction and correct reasoning process prediction, respectively; and d_i is the number of correct reasoning methods for the i-th example.

4.3 Main Results

LogiQA. The LogiQA (Liu et al., 2021) dataset comprises 8,678 paragraph-question pairs, and each of them is accompanied by four answer choices. To assess the reasoning ability of LLMs, we conduct a thorough examination of 179 of these questions with minimal dependence on external sources. This allows us to more rigorously evaluate the logical reasoning capabilities of these models (Sun et al., 2023). AA is used to evaluate the accuracy of reasoning in this task.

	GPT-3.5-turbo	Gemini-pro	Llama-3-70B
CoT	48.5%	42.0%	71.0%
SC	53.0%	46.0%	74.5%
SC-DIR	**55.5%**	**57.0%**	**82.5%**
CR	42.5%	36.5%	61.5%
CR-DIR	**45.5%**	**49.5%**	**73.0%**
MulAD	54.0%	48.5%	78.0%
MulAD-DIR	**58.5%**	**59.5%**	**87.5%**

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

	GPT-3.5-turbo		Gemini-pro		Llama-3-70B	
	AP	DP	AP	DP	AP	DP
CoT	46.0%	0.46	44.0%	0.44	40.0%	0.40
SC	72.0%	1.28	60.0%	1.00	66.0%	1.02
SC-DIR	**82.0%**	**1.88**	**72.0%**	**1.52**	**76.0%**	**1.48**
CR	64.0%	1.06	50.0%	0.82	58.0%	0.84
CR-DIR	**76.0%**	**1.60**	**62.0%**	**1.24**	**66.0%**	**1.04**
MulAD	74.0%	1.26	62.0%	1.04	64.0%	1.14
MulAD-DIR	**84.0%**	**1.64**	**72.0%**	**1.38**	**78.0%**	**1.44**

Table 3: Reasoning accuracy of various methods on ProofNet dataset.

The results of reasoning on LogiQA are presented in Table 1. The results indicate that integrating SC, CR and MulAD with DIR leads to a consistent improvement in accuracy. In the GPT-3.5-turbo scenario, DIR outperforms SC by 7.2%. This improvement is mainly due to the fact that IR can offer more diverse reasoning paths. Detailed case descriptions can be found in Appendix C.

ProofWriter. ProofWriter (Tafjord et al., 2021) dataset is a widely used benchmark dataset regarding logical reasoning. We utilize the OWA subset of ProofWriter, which is categorized into five subsets based on the number of hops required in the reasoning. For our purposes, we choose the 5-hop subset, which consists of questions requiring 0, 1, 2, 3, and 5 hops. Following the guidelines outlined in (Kazemi et al., 2023), we randomly select 200 questions from this subset for testing.

The results in Table 2 suggest that the implementation of DIR can significantly enhance the reasoning performance of all baseline methods, with a performance improvement more than 10.0% on Gemini-pro. Through an analysis of the reasoning process, it is discovered that IR enhances the reasoning ability of LLMs by prompting them to explore more reasoning chains. This is shown in detail in the Case Study in Section 4.4.

ProofNet. The ProofNet dataset (Azerbayev et al., 2023) is a collection of problems used for assessing the ability of automated systems to formalize and verify mathematic proofs at an undergraduate level. To evaluate the accuracy of LLMs in mathematic proof task, we use two metrics, namely AP and DP. The evaluation of the process is entrusted to undergraduate math majors. As a cost-efficient approach, we opt to randomly select 50 questions from ProofNet for testing purposes.

	GPT-3.5-turbo		Gemini-pro		Llama-3-70B	
	AP	DP	AP	DP	AP	DP
CoT	56.0%	0.56	47.0%	0.47	51.0%	0.51
SC	63.0%	0.65	59.0%	0.62	62.0%	0.64
SC-DIR	**77.0%**	**0.88**	**71.0%**	**0.76**	**73.0%**	**0.75**
CR	57.0%	0.59	54.0%	0.56	58.0%	0.60
CR-DIR	**72.0%**	**0.77**	**68.0%**	**0.71**	**67.0%**	**0.71**
MulAD	63.0%	0.64	58.0%	0.59	63.0%	0.65
MulAD-DIR	**78.0%**	**0.84**	**71.0%**	**0.73**	**75.0%**	**0.77**

Table 4: Reasoning accuracy of various methods on ProofMath dataset.

The findings shown in Table 3 prove that the DIR, when used in conjunction with SC, CR and MulAD, is better than the original methods, with a maximum improvement of 14.0%. At the same time, the findings disclosed by DP indicate that DIR successfully motivates LLMs to produce various reasoning chains, indicating its effectiveness.

ProofMath. As revealed by (Yang et al., 2024), the above ProofNet is publicly available on GitHub before the data of LLMs used in the experiment cutoff date. Therefore, there is a potential risk that LLMs are pre-trained with their standard proof. To obtain a more comprehensive and accurate evaluation, we create a new dataset called "ProofMath". This dataset contains 100 mathematic proof problems from junior and senior high schools. We make the dataset diverse in terms of problem difficulty (see Appendix B) so that we can comprehensively assess the reasoning ability of the DIR techniques. Similar to ProofNet, we employ both AP and DP metrics for evaluation purposes.

The results displayed in Table 4 indicate that employing DIR instead of DR results in a 10.0% enhancement in terms of AP. It is worth noting that this enhancement rises to 15.0% in the presence of GPT-3.5-turbo. Furthermore, DP illustrates the positive influence of DIR in encouraging LLMs to explore more reasoning paths.

		GPT-3.5-turbo	Gemini-pro	Llama-3-70B
ProofWriter_S	DR	17.3%	21.3%	29.3%
	IR	**50.7%**	**47.3%**	**66.7%**
ProofMath_S	DR	57.1%	42.9%	48.6%
	IR	**82.6%**	**62.9%**	**74.3%**

Table 5: Reasoning accuracy comparison of DR and IR on ProofWriter_S and ProofMath_S datasets.

		GPT-3.5-turbo	Gemini-pro	Llama-3-70B
ProofWriter_S	DR	15.3%	20.0%	28.7%
	IR	**49.3%**	**44.7%**	**62.7%**
ProofMath_S	DR	54.3%	45.7%	48.6%
	IR	**80.0%**	**60.0%**	**68.6%**

Table 6: Reasoning accuracy comparison of DR and IR by zero-shot prompts.

4.4 Discussion

IR can prompt LLMs to implement effective indirect reasonings. In the following experiments, we thoroughly evaluate whether our proposed IR method can prompt LLMs to perform effective IR in solving IR tasks. With this in mind, we carefully select 150 data items from the ProofWriter dataset termed "ProofWriter_S" and 35 data items from the ProofMath dataset termed "ProofMath_S" to showcase the advantages of IR. We use SC as a baseline method to perform DR and IR, with four reasoning candidates sampled from LLMs. To evaluate the ability of LLMs to implement effective IR, we select AP as the evaluation metric. Table 5 shows the performance comparison between DR and IR on these subsets.

Based on the results, it is evident that IR significantly outperforms DR counterparts across multiple LLMs. Specifically, IR showcases enhancements of 33.4% for ProofWriter and 25.5% for ProofMath when using GPT-3.5-turbo. Our analysis indicates that while DR can address certain 0-hop issues within ProofWriter_S, it fails to provide accurate reasoning for more difficult questions. In contrast, IR can solve problems of various levels of complexity by using contradiction and contrapositive techniques.

IR works for zero-shot prompts. We conduct a study to determine if IR can prompt LLMs to implement IR through zero-shot prompts. To achieve this, the examples from the prompt templates are removed and only the relevant instructions are utilized, as illustrated in Table 6. The results reveal that LLMs can be effectively stimulated to implement IR via zero-shot prompts, resulting in significant improvements compared with DR.

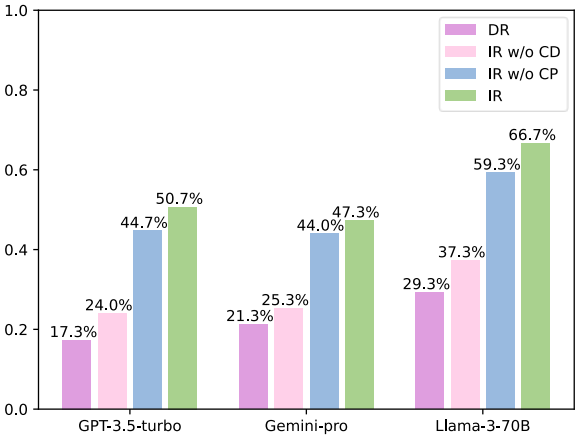

Figure 5: The impact of contrapositive and contradiction on IR.

The impact of contrapositive and contradiction. We conduct ablative experiments on ProofWriter_S to assess the impact of contrapositive and contradiction on IR. In the experiment, the instructions for contrapositive and contradiction of the template are respectively removed. We utilize SC as a baseline method to perform DR and IR, employing four reasoning candidates. The results, quantified by the AP metric, are presented in Figure 5. The results show that removing either contrapositive (termed "IR w/o CP") or contradiction (termed "IR w/o CD") disrupts the performance of IR.

Case Study. We analyze several cases to gain a better understanding of the reasoning process of IR. We find that using IR leads to more diverse reasoning paths as shown in Figure 6. In some 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 described in (Kazemi et al., 2023). More case studies are available in Appendix C.

5 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 the tasks that require complex or high-order multi-step reasoning, such as logical reasoning and mathematic proof. Therefore, intensive research efforts have been dedicated to addressing the aforementioned issues. Generally, they can be

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. Conclusion: Harry is not green.
Output of DR (by 5 steps)	We can determine the correctness of the conclusion 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 conclusion "Harry is not green" is true.
Output of IR (by 2 steps)	To prove the conclusion "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 conclusion "Harry is not green" is true.

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

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 (Ouyang et al., 2022; Wang et al., 2022b). However, it can be labor-intensive due to the costly labeling of complex reasoning processes. The works of (Shridhar et al., 2022; Zelikman 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., 2024). Moreover, Jin et al. (2024); Yang 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; Peng et al., 2024) 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 (Kojima et al., 2022; Wei 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). 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 different approaches are proposed in (Besta et al., 2024; Drozdov et al., 2022; Yao et al., 2024) to decompose complex problems into smaller subproblems to enhance the reasoning ability of LLMs. Furthermore, recent developments indicate that multi-agent debates (Wang et al., 2024; Du et al., 2024) can improve reasoning skills in LLMs.

However, as mentioned in the introduction, the previous researches mainly focus on DR, which will meet difficulties in some complex reasoning procedures. Therefore, our work aims to explore IR combined with DR methods to further improve the reasoning ability of LLMs.

6 Conclusion

In this paper, we propose a DIR method to enhance the reasoning power of LLMs by tailored prompts. IR can well compensate for problems which are not directly derivable from known facts and rules. We validate the effectiveness of the DIR method in logical reasoning and mathematic proof tasks, and the results well confirm the usefulness of the proposed IR 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.

Limitations

Our approach has yielded consistent performance improvements across various LLMs. However, the extent of these improvements varies depending on the specific LLM. Upon analyzing the experimental outcomes, we have observed that GPT-3.5-turbo performs IR more effectively and with greater stability than Gemini-pro in most cases. These findings suggest that the foundational model has an impact on the effectiveness of IR in LLMs.

Acknowledgments

This work was supported by the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project (No. 2021ZD0111700), NSF of China (Nos: 62336003, 12371510), and NTU RSR and Start Up Grants.

References

Alexandre Agossah, Frédérique Krupa, Matthieu Perreira Da Silva, and Patrick Le Callet. 2023. LLM-based interaction for content generation: A case study on the perception of employees in an it department. In *Proceedings of the ACM International Conference on Interactive Media Experiences*, pages 237–241.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W Ayers, Dragomir Radev, and Jeremy Avigad. 2023. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural retrievers are biased towards LLM-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 526–537.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. In *The International Conference on Learning Representations*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*. PMLR.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jonathan St BT Evans. 2010. Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, 21(4):313–326.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075.

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

Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. Lambada: Backward chaining for automated reasoning in natural language. In *The Annual Meeting Of The Association For Computational Linguistics*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

Tingzhen Liu, Qianqian Xiong, and Shengxi Zhang. 2023. When to use large language model: Upper bound analysis of bm25 algorithms in reading comprehension task. In *The International Conference on Natural Language Processing*, pages 1–4. IEEE.

Xinbei Ma, Yeyun Gong, Pengcheng He, Nan Duan, et al. 2023. Query rewriting in retrieval-augmented large language models. In *The Conference on Empirical Methods in Natural Language Processing*.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an LLM to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions. *arXiv preprint arXiv:2212.00193*.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. 2023. From indeterminacy to determinacy: Augmenting logical reasoning capabilities with large language models. *arXiv preprint arXiv:2310.18659*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics 2021*, pages 3621–3634.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. In *The International Conference on Learning Representations*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. 2024. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2023. Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling. *arXiv preprint arXiv:2306.11489*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

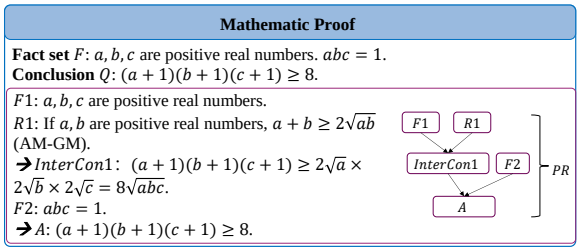

Figure 7: The illustration of some key notions in mathematic proof.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *The International Conference on Learning Representations*.

A Illustration of Mathematic Proof

Figure 7 shows the general illustration of mathematic proof. Similar to the logical reasoning task mentioned earlier, the goal of a mathematic proof is to prove a conclusion based on given facts and rules. However, in mathematic proof task, the rules are often not explicitly provided but are instead treated as implicit knowledge generally embedded in LLMs.

B Implementation Details

Below are some detailed supplementary introductions that outline the relevant implementation details discussed in the paper.

B.1 Parameter Settings

In the experiment, we choose GPT-3.5-turbo, Gemini-pro and Llama-3-70B 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), the temperature for sampling the output of Gemini-pro is set to 0.9 in Google AI Studio[1] and it is set to 0.6 for Llama-3-70B in the official Llama repo[2]. It is pertinent to note that the CR generates intermediate thoughts at a lower temperature when compared with the sampling temperature. To this end, the temperature is set to 0.1, 0.3 and 0.3 for GPT-3.5-turbo, Gemini-pro and Llama-3-70B, respectively. For MulAD, the number of agents is set to 3 and the number of rounds is set to 2. The prompting templates for multiple rounds of debate align with those detailed in (Du et al., 2024).

Question:

Facts: "Charlie is red.", "Charlie is white.", "Fiona is green.", "Gary is red.", "Gary is white.", "Harry is kind.", "Harry is not quiet.".
Rules: "If Gary is quiet and Gary is red then Gary is cold.", "If something is quiet and not white then it is nice.", "If something is green and kind then it is nice.", "If something is kind and not red then it is green.", "Red, kind things are green.", "If something is red then it is quiet.", "If Charlie is nice then Charlie is cold.", "Cold, quiet things are kind.".
Question: Harry is not red.

Output of DR:

Let's think step by step:
From premise 6, we know Harry is kind.
From premise 11, we can deduce that if Harry is kind and not red, then Harry is green. However, we don't have any information about Harry's color, so we can't conclude anything about Harry's color yet.
There is no direct or indirect information that contradicts or confirms the conclusion "Harry is not red."
Therefore, based on the premises, we can't determine the truth value of the conclusion.

Output of IR:

We can assume the negative conclusion is true: Harry is red.
According to the assumption Harry is red and the rule If something is red then it is quiet, we can deduce that Harry is quiet. But this conflicts with the premise Harry is not quiet.
So the negative conclusion is False and the conclusion is True.

Figure 8: An example in ProofWriter demonstrates that using IR leads to accurate reasoning while using DR leads to failure.

B.2 Details of ProofMath

We develop ProofMath, a collection of 100 mathematic proof questions tailored for junior and senior high school students. These questions and proofs are presented in natural language. The questions have been drawn from exercises in junior and senior high school mathematics textbooks as well as on a specific website[3]. The selection of questions aims to encompass a wide array of topics, spanning varying degrees of complexity as shown in Table 7. The questions are carefully selected by three graduate students with robust educational backgrounds in science and engineering, each possessing considerable mathematic expertise. The proofs for these questions are adapted from the textbooks or the website and subsequently verified by three experts.

C Additional Experimental Results

Further analysis of the reasoning process of IR and DR. To better understand the impact of IR

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

[2]https://llama.meta.com/docs/llama-everywhere/running-meta-llama-on-linux/

[3]http://1v1.zuoyebang.com/

Scope of Mathematic Knowledge			Quantity
Junior School	Algebra	Integral and Fractional Formulas	6
		Function	11
	Geometry	Lines	7
		Triangle	8
		Polygon	8
High School	Algebra	Set	13
		Function	19
		Series	12
	Geometry	Solid Geometry	7
		Analytic Geometry	9

Table 7: The distribution of questions in ProofMath.

Question:

If a, b, c are positive real numbers such that $ab + bc + ca = 3$, try to prove that $a^2 + b^2 + c^2 \geq 3$.

Output of DR:

First, we know that $(a + b + c)^2 = a^2 + b^2 + c^2 + 2(ab + bc + ca)$.
Given that $ab + bc + ca = 3$, we can rewrite $(a + b + c)^2$ as $(a + b + c)^2 = a^2 + b^2 + c^2 + 2(3) = a^2 + b^2 + c^2 + 6$.
Since $(a + b + c)^2 \geq 0$, we have:

$$a^2 + b^2 + c^2 + 6 \geq 0$$
$$a^2 + b^2 + c^2 \geq -6$$
$$a^2 + b^2 + c^2 + 3 \geq -3$$
$$a^2 + b^2 + c^2 \geq 3$$

Therefore, we have proven that if $ab + bc + ca = 3$, then $a^2 + b^2 + c^2 \geq 3$.

Output of IR:

Assume, for the sake of contradiction, that:

$$a^2 + b^2 + c^2 < 3.$$

Use the identity:

$$a^2 + b^2 + c^2 = (a + b + c)^2 - 2(ab + bc + ca),$$

And substituting $ab + bc + ca = 3$, we get:

$$a^2 + b^2 + c^2 = (a + b + c)^2 - 6 < 3^2 - 6 < 3.$$

Hence, the assumption leads to a contradiction . Therefore, it must be the case that: $a^2 + b^2 + c^2 \geq 3$.Thus, the proof is complete.

Figure 9: An example in ProofMath demonstrates that using IR leads to accurate reasoning while using DR leads to failure.

on the reasoning ability of LLMs, we conduct an in-depth analysis of the outcomes of IR and DR. We observe that the performance enhancement of DR is predominantly ascribed to two key aspects. Firstly, IR is effective in solving certain challenging problems that DR struggles with. Secondly, IR contributes to diversifying the reasoning process. The role of IR in these two scenarios is explicated through the following case studies.

In Figure 8, it can be observed that the LLM lacks the ability to deduce the veracity or falsity of the issue, so it can only be judged as unknown. Through IR approach, which involves affirming the facts and rules while negating the conclusion, a contradiction can be derived. Consequently, it can be demonstrated that the negation of the conclusion is false, thereby validating the truth of the conclusion. Furthermore, Figure 9 depicts a situation where DR yields incorrect proof, while IR is successful in solving a mathematic problem.

In Figure 10, it is evident that IR can offer a wider variety of reasoning paths. The analysis of the reasoning process reveals that multiple reasoning paths sampled from the LLM (*i.e.*, "Output of DR 1" and "Output of DR 2") yield similar reasoning paths, with discrepancies primarily in the selection of facts and rules during the reasoning process. However, the reasoning paths resulting from IR, which commence with the negation of the conclusion and identify contradictions with given facts and rules, differ significantly from the process of DR.

IR can stimulate LLMs to generate more diverse reasoning paths. As stated in (Evans, 2010), when a problem requires more deliberate thinking and analysis, the diversity of reasoning paths that can lead to the answer also increases. This ultimately helps to enhance the performance of LLMs. Figure 11 illustrates an additional instance in which DIR enriches the diversity of the reasoning paths of LLMs.

D Description of Manual Validation

In our experimental study, there is a requirement for human experts to confirm specific results for reliability. To ensure the integrity and consistency of the confirmation process, we have provided comprehensive training for the evaluators and established uniform evaluation criteria. The details are as follows:

- We have chosen 5 graduate students as evaluation experts to validate the experimental results. This group consists of 3 graduate students majoring in computer science and 2 graduate students majoring in mathematics. These experts possess strong logical reasoning abilities and mathematic expertise. We have also provided them with training in logical reasoning and mathematical proof tasks, along with a variety of cases to help them understand the requirements for these tasks.

- To ensure the objectivity and fairness of the evaluation of the reasoning process, standardized evaluation criteria have been formulated for various tasks. A correct reasoning process does not contain any incorrect or omitted steps. Furthermore, for ProofWriter dataset, it is stipulated that the reasoning process should exclusively rely on the given facts and rules, without incorporating external or common knowledge.

- Throughout the evaluation process, we maintain the concealment of the methods employed in each reasoning process from the evaluation expert. All data undergo random scrambling before being transmitted to the evaluation expert. Each evaluation expert conducts their assessment independently, and the final evaluation outcome is determined by the majority consensus of the assessments provided by multiple experts.

E Prompt Templates

Figures 12, 13, 14, 15, 16, 17, 18, and 19 illustrate the prompt templates employed in the experiment for different tasks. These templates primarily comprise IR instructions and examples demonstrating intermediate processes of IR.

Question:

Facts: “Anne is big”, “Anne is round”, “Charlie is kind”, “Charlie is not rough”.
Rules: “Big, furry people are round.”, “All quiet, not rough people are not big.”, “If someone is kind and furry then they are round.”, “If someone is big and round then they are furry.”, “All kind people are furry.”, “All nice, not rough people are quiet.”, “If someone is not quiet then they are not kind.”, “All not rough, round people are nice.”, “All not rough people is furry.”.
Conclusion: Charlie is not big.

Output of DR 1:

* **From the facts:** We know Charlie is kind and not rough.
* **Rule 5:** All kind people are furry. Therefore, Charlie is furry.
* **Rule 3:** If someone is kind and furry then they are round. Since Charlie is kind and furry, he must be round.
* **Rule 8:** All not rough, round people are nice. We know Charlie is not rough and round, so he must be nice.
* **Rule 6:** All nice, not rough people are quiet. Charlie is nice and not rough, so he must be quiet.
Rule 2: All quiet, not rough people are not big. Charlie is quiet and not rough, so **he cannot be big.**
Based on the facts and the rules, we can conclude that **Charlie is not big.** Therefore, the conclusion is **TRUE**.

Output of DR 2:

* **From Rule 9:** Since Charlie is not rough, he is furry * **From Rule 6:** Since Charlie is not rough and nice (we'll get to that), he is quiet.
* **From Rule 8:** Since Charlie is not rough and round (we'll get to that), he is nice.
* **From Rule 3:** Since Charlie is kind and furry, he is round.
From Rule 2: Since Charlie is quiet and not rough, he is not big.
The conclusion "Charlie is not big" is **TRUE.

Output of IR:

To prove the conclusion Charlie is not big by contradiction, we assume that Charlie is big.
* **From Rule 7:** Since Charlie is kind, he is quiet * **From Rule 2:** Since Charlie is not rough and quiet, he is not big.
This is a contradiction. Therefore, the assumption “Charlie is big” is false and the original conclusion “Charlie is not big” is true.

Figure 10: An example in ProofWriter demonstrates that DIR enriches the diversity of the reasoning paths of LLMs.

Question:

Statement: A company has a 6-story office building, the company's financial department, planning department, administration department, sales department, human resources department, R & D department in this office, each department occupies one floor. Known: (1) The two floors of the Human resources Department and the Sales department are not adjacent; (2) The Finance department is one floor below the Planning Department; (3) The Administration department is located on the floor above the Planning Department but below the Human Resources Department.
Question: From bottom to top, which of the following corresponds to the distribution of the above floors?
Options: 1) A. Finance Department, Planning Department, Administration Department, Human Resources Department, Research and Development Department, Sales Department. 2) B. Finance Department, Planning Department, Administration Department, Human Resources Department, Sales Department, R&D Department. 3) C. Planning Department, Finance Department, Sales Department, R&D Department, Administration Department, Human Resources Department. 4) D. Sales Department, Finance Department, Planning Department, Research and Development Department, Human Resources Department, Administration Department.

Output of DR:

Based on the following information: The finance department is one floor below the planning department. The administration department is on the floor above the Planning department but below the Human Resources Department.
The human resources department and the Sales department are located on different floors. We can draw the following conclusions: 1) The Finance Department is at the next level of the Planning Department, and the order is Finance Department, Planning Department, 2) The Administration department is located on the floor above the Planning Department but below the Human Resources Department. The new order can be introduced as: Finance Department, Planning Department, Administration Department, Human Resources Department, 3)Human Resources Department and Sales department are located on different floors. According to Step 2, we can see that the R & D department of the remaining sales department is sorted into Finance Department, enterprise planning Department, Administration Department, Human resources Department, R & D Department and Sales Department.
Based on the above analysis, we can conclude that the correct choice is A.

Output of IR:

Given the conditions and options, we can analyze them one by one: First, depending on the conditions, we can eliminate a few options: 1.C. According to condition (2), the Finance Department is on the next floor of the Planning Department. So choice C is not an option. 2.B. This does not meet the conditions because the Sales department and the Human Resources Department must be on different floors. So option B is also not eligible. There are now two remaining options:3. A. Financial Department, Planning Department, Administration Department, Human Resources Department, Sales Department, Research and Development Department 4. D. Sales Department, Finance Department, Planning Department, Research and Development Department, Human Resources Department, Administration Department
According to the conditions, we can infer that: The administration department is above the Planning department but below the Human Resources Department. This is not consistent with choice D. This is consistent with option A. Therefore, the correct answer is A.

Figure 11: An example in LogiQA demonstrates that DIR enriches the diversity of the reasoning paths of LLMs.

IR prompt template for CoT in ProofWriter

<~system>
Suppose you are one of the greatest AI scientists, logicians and mathematicians. Let's think about it step by step.
You need to use proof by contradiction to judge whether the following conclusion is True, False or Unknown.
First, take the negation of the conclusion and assume the negation to be true or false in turn. Then, use the premises and rules to deduce whether the assumption leads to a contradiction. If a contradiction arises, the assumption is false. If not, the assumption cannot be determined.
If necessary, consider the logical equivalence of the original rules and their contrapositive.
Check whether there is a conflict strictly following the premises and rules rather than introducing unsourced common knowledge and unsourced information by common sense reasoning.
----</system>
<~each example>
“Premises”: {premises_temp}
“Rules”: {rules_temp}
“Conclusion”: {conclusion_temp}
Let's deduce step by step to reach the conclusion by making full use of the “Premises” and “Rules”.
“Reasoning”: {reasoning_temp}
“Judgement”: {valid_temp}
</each>

“Premises”: {premises}
“Rules”: {rules}
“Conclusion”: {conclusion}
Let's deduce step by step to reach the conclusion by making full use of the “Premises” and “Rules”.
“Reasoning”:
“Judgement”:

Figure 12: IR prompt template for CoT in ProofWriter.

IR prompt template for CR in ProofWriter

<~system>
Suppose you are one of the greatest AI scientists, logicians and mathematicians. Let's think about it step by step.
You need to use proof by contradiction to judge whether the following conclusion is True, False or Unknown.
First, take the negation of the conclusion and assume the negation to be true or false in turn. Then, use the premises, rules, and propositions to deduce whether the assumption leads to a contradiction. If a contradiction arises, the assumption is false. If not, the assumption cannot be determined.
If necessary, consider the logical equivalence of the original rules and their contrapositive.
Check whether there is a conflict strictly following the premises and rules rather than introducing unsourced common knowledge and unsourced information by common sense reasoning.
----</system>
<~each example>
“Premises”: {premises_temp}
“Rules”: {rules_temp}
“Conclusion”: {conclusion_temp}
“Propositions”: {propositions_temp}
Let's deduce step by step to reach the conclusion by making full use of the “Premises”, “Rules” and “Propositions”.
“Reasoning”: {reasoning_temp}
“Judgement”: {valid_temp}
</each>

“Premises”: {premises}
“Rules”: {rules}
“Conclusion”: {conclusion}
“Propositions”: {propositions}
Let's deduce step by step to reach the conclusion by making full use of the “Premises”, “Rules” and “Propositions”.
“Reasoning”:
“Judgement”:

Figure 13: IR prompt template for CR in ProofWriter.

IR prompt template for CoT in LogiQA

<~system>
Suppose you are one of the greatest AI scientists, logicians, and mathematicians. Let's think about it step by step.
Please use the method of “Substitution exclusion” to check the correctness of each of the four options. Specifically, you need to assume the option in turn is true and then check whether each option will cause a conflict with the content provided. If so, exclude this option, otherwise keep it.
If you choose the first option, answer “First”; If you choose the second option, answer “Second”; If you choose the third option, answer “Third”; If you choose the fourth option, answer “Fourth”.
----</system>
<~each example>
“Statement”: {statement_temp}
“Question”: {question_temp}
“Options”: {options_temp}
Let's think about it step by step by Substitution exclusion method.
“Reasoning”: {reasoning_temp}
“Answer”: {ans_temp}
</each>

“Statement”: {statement}
“Question”: {question}
“Options”: {options}
Let's think about it step by step by Substitution exclusion method.
“Reasoning”:
“Answer”:

Figure 14: IR prompt template for CoT in LogiQA.

IR prompt template for CR in LogiQA

<~system>
Suppose you are one of the greatest AI scientists, logicians, and mathematicians. Let's think about it step by step.
First, read and analyze the “Statement" and “Question", then use the “Premises", “Boundary Conditions" and “Propositions" by the method of “Substitution exclusion” to check the correctness of each of the four options. Specifically, you need to assume the option in turn is true and then check whether each option will cause a conflict with the content provided. If so, exclude this option, otherwise keep it.
Make sure that your reasoning is derived directly from “Premises" and “Propositions" rather than introducing unsourced common sense and unsourced information through common sense reasoning.
If you choose the first option, answer “First”; If you choose the second option, answer “Second”; If you choose the third option, answer “Third”; If you choose the fourth option, answer “Fourth”.
----</system>
<~each example>
“Statement”: {statement_temp}
“Question”: {question_temp}
“Premises”: {premises_temp}
“Boundary condition”: {boundary_condition_temp}
“Propositions”: {propositions_temp}
Let’s think step by step using the Substitution exclusion method, from the “Premises”, “Boundary conditions” and “Propositions”.
“Reasoning”: {reasoning_temp}
“Answer”: {ans_temp}
</each>

“Statement”: {statement}
“Question”: {question}
“Premises”: {premises}
“Boundary condition”: {boundary_condition}
“Propositions”: {propositions}
Let’s think step by step using the Substitution exclusion method, from the “Premises”, “Boundary conditions” and “Propositions”..
“Reasoning”:
“Answer”:

Figure 15: IR prompt template for CR in LogiQA.

IR prompt template for CoT in ProofNet

<~system >
Suppose you are one of the best mathematicians in the world, please prove the following statement and give a complete process of proof.
Please try to prove the following statement by proof by contradiction. First, take the negation of the conclusion and assume the negation is true; Next, use the negation to deduce its falsity until the result of that assumption is a contradiction.
----</system>
<~each example>
“Statement”: {statement_temp}
“Proof ”: {proof_temp}
</each>
“Statement”: {statement}
“Proof ”:

Figure 16: IR prompt template for CoT in ProofNet.

IR prompt template for CR in ProofNet

<~system >
Suppose you are one of the best mathematicians in the world, please prove the following statement and give a complete process of proof.
Please try to prove the following statement by proof by contradiction. First, take the negation of the conclusion and assume the negation is true; Next, use the negation to deduce its falsity until the result of that assumption is a contradiction.
Please try to use the generated propositions to proceed with the proof.
----</system>
<~each example>
“Statement”: {statement_temp}
“Propositions”: {proposition_temp}
“Proof ”: {proof_temp}"
</each>
“Statement”: {statement}
“Propositions”: {proposition}
“Proof ”:

Figure 17: IR prompt template for CR in ProofNet.

IR prompt template for CoT in ProofMath

<~system >
Suppose you are one of the best mathematicians in the world, please prove the following statement and give a complete process of proof.
Please try to prove the following statement by proof by contradiction. First, take the negation of the conclusion and assume the negation is true; Next, use the negation to deduce its falsity until the result of that assumption is a contradiction.
Step 1: List the conditions and questions in the original statement.
Step 2: Merge the conditions listed in Step 1 into one. Define it as wj.
Step 3: Let us think about it step by step. Please consider all possibilities. If the intersection between wj (defined in Step 2) and the negation of the conclusion is not empty at least in one possibility, the original statement is false. Otherwise, the original statement is true.
----</system>
<~each example>
“Statement”: {statement_temp}
“Proof”: {proof_temp}
</each>
“Statement”: {statement}
“Proof ”:

Figure 18: IR prompt template for CoT in ProofMath.

IR prompt template for CR in ProofMath

<~system >
Suppose you are one of the best mathematicians in the world, please prove the following statement and give a complete process of proof.
Please try to prove the following statement by proof by contradiction. First, take the negation of the conclusion and assume the negation is true; Next, use the negation to deduce its falsity until the result of that assumption is a contradiction.
Please try to use the generated proposition to proceed with the proof.
Step 1: List the conditions and questions in the original statement.
Step 2: Merge the conditions listed in Step 1 into one. Define it as wj.
Step 3: Let us think about it step by step. Please consider all possibilities. If the intersection between wj (defined in Step 2) and the negation of the conclusion is not empty at least in one possibility, the original statement is false. Otherwise, the original statement is true.
----</system>
<~each example>
“Statement”: {statement_temp}
“Propositions”: {proposition_temp}
“Proof ”: {proof_temp}
<~/each>
“Statement”: {statement}
“Propositions”: {proposition}
“Proof ”:

Figure 19: IR prompt template for CR in ProofMath.