Concise and Organized Perception Facilitates Reasoning in Large Language Models

Junjie Liu[1], Shaotian Yan[1], Chen Shen[1], Liang Xie[2,1], Wenxiao Wang[3,1], Jieping Ye[1]
[1]Alibaba Cloud, [2]Zhejiang University of Technology, [3]Zhejiang University
jumptoliujj@gmail.com,yanshaotian@gmail.com,zjushenchen@gmail.com,
lilydedbb@gmail.com, zjdxwwx@163.com, yejieping.ye@alibaba-inc.com

Abstract

Exploiting large language models (LLMs) to tackle reasoning has garnered growing attention. It still remains highly challenging to achieve satisfactory results in complex logical problems, characterized by plenty of premises within the prompt and requiring multi-hop reasoning. In particular, the reasoning capabilities of LLMs are brittle to *disorder* and *distractibility*. In this work, we first examine the mechanism from the perspective of information flow and reveal that LLMs exhibit failure patterns akin to human-like cognitive biases when dealing with disordered and irrelevant content in reasoning tasks. However, in contrast to LLMs, disordered and irrelevant content does not significantly decrease human performance, as humans have a propensity to distill the most relevant information and systematically organize their thoughts, aiding them in responding to questions. Stem from that, we further propose a novel reasoning approach named Concise and Organized Perception (COP). COP carefully analyzes the given statements to identify the most pertinent information while eliminating redundancy efficiently. It then prompts the LLMs in a more organized form that adapts to the model's inference process. By perceiving concise and organized context, the reasoning abilities of LLMs can be better elicited. Extensive experimental results on several popular logical benchmarks (ProofWriter, PrOntoQA, PrOntoQA-OOD, and FOLIO) and math benchmark (DI-GSM) show that COP significantly outperforms previous state-of-the-art methods.

1 Introduction

The field of large language models (LLMs) has witnessed significant progress in complex reasoning with the advent of Chain-of-thought (CoT) prompting [41] and a series of related works [31, 26, 45, 43, 3]. These breakthroughs have yielded remarkable achievements in various applications, including arithmetic, commonsense, symbolic reasoning, etc. [36, 1, 6, 20, 15, 5], and have sparked widespread enthusiasm within the community to continuously explore the immense potential of LLMs in tackling complex reasoning tasks. However, the performance of LLMs drastically decreases when handling intricate tasks characterized by plenty of premises within the prompt and requiring multi-hop reasoning. One primary issue is *distractibility* [34], where the reasoning capabilities of LLMs are highly susceptible to deterioration when confronted with irrelevant context. Another failure mode that has garnered significant attention recently is *disorder* [33, 7], where the performance of LLMs Severely drops when the premise order does not align with the context required in intermediate reasoning steps. Figure 1(a) presents an example from ProofWriter [35], which is a deductive logical dataset (in deductive reasoning, disorder can be further understood as that the premises are not in the same order as the ground truth proof). The complexity of the context in premises directly affects the difficulty of reasoning. When the given premises are disordered and contain much information irrelevant to the reasoning, the models face a higher risk of selecting the wrong information at some stage. This often leads to an incomplete proof and subsequently an incorrect

Anne is big. Anne is nice. Bob is not big. Gary is blue.[1] Gary is not quiet.[3] [...] If Gary is quiet and Gary is big then Gary is nice. If someone is nice and not quiet then they are big.[4] All blue people are nice.[2] If someone is big and quiet then they are blue. All big people are green.[5] If Gary is quiet and Gary is big then Gary is young. Green people are smart.[6] If someone is smart then they are not young.[7] [...]
Question: Is the following statement true, false, or unknown? Gary is young. Answer: False.

Black represents irrelevant information, while green represents concise information. Superscript serial number x represents the logical order in the reasoning steps, which is the x-th order. (a)

Concise Input: Gary is blue.[1] Gary is not quiet.[3] If someone is nice and not quiet then they are big.[4] All blue people are nice.[2] All big people are green.[5] Green people are smart.[6] If someone is smart then they are not young.[7] (b)

Organized Input: Gary is blue.[1] All blue people are nice.[2] Gary is not quiet.[3] If someone is nice and not quiet then they are big.[4] All big people are green.[5] Green people are smart.[6] If someone is smart then they are not young.[7] Anne is big. Anne is nice. Bob is not big. [...] If Gary is quiet and Gary is big then Gary is young. [...]

CoT Inference with Original Input (c)
[...] Gary is blue. Gary is not quiet. If Gary is quiet and Gary is big then Gary is nice. Therefore we do not know if Gary is nice. [...] Therefore we do not know if Gary is young. The given statement is Unknown. ✗

CoT Inference with Concise Input
Gary is blue. Therefore Gary is nice. Gary is not quiet. Therefore Gary is big. [...] If someone is smart then they are not young. Therefore Gary is not young. The given statement is False. ✓

CoT Inference with Organized Input
Gary is blue. All blue people are nice. Therefore Gary is nice. Gary is not quiet. If someone is nice and not quiet then they are big. Therefore Gary is big. [...] If someone is smart then they are not young. Therefore Gary is not young. If Gary is quiet and Gary is big then Gary is young. Therefore the given statement is False. ✓

(d) [bar chart: Original, Organized, Concise, Concise+Organized; categories 5hop, 4hop, 3hop with values 35.9, 64.1, 57.8, 71.9 (5hop); 39.4, 65.2, 68.2, 80.3 (4hop); 60.8, 78.8, 81.8, 86.4 (3hop)]
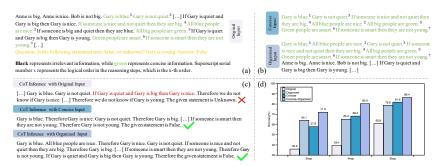
Figure 1: (a) A 5-hop example of ProofWriter dataset, showcasing plenty of premises and the question to be answered. Some premises are omitted for brevity. (b) Corresponding reconstruction of concise and organized perception. Superscript serial numbers represent the logical orders according to the gold proof. The concise input contains only relevant information but lacks organizational structure. In contrast, the organized input arranges statements consistently with the gold reasoning path, albeit including some redundant information. (c) LLMs output results. (d) Results of a confirmatory experiment.

answer. Figure 1(c) illustrates such a misleading step, where the original CoT selects the wrong reasoning path (highlighted in red). This observation indicates that LLMs usually struggle with proof planning when irrelevant and disordered content hinders, as also revealed in some contemporaneous works[2, 34, 7].

In this work, we first briefly investigate the underlying mechanism of the influence of disordered and irrelevant content on reasoning. Specifically, we adopt neuron saliency score analysis, which is an important approach for pinpoint the information flow and the crucial interactions between tokens [12, 19, 40], and we conclude three phenomena. **(i)** LLMs struggle to identify the correct entry point of the reasoning path when faced with disorder and distraction. Intuitively, humans face the same difficulties (i.e., hard to get started) when navigating through a pile of complex information. **(ii)** The current step always highly focuses on the previous step, to the extent that it might even make up non existing premises to accommodate the preceding step. Consequently, it is challenging in allocating sufficient energy to identify the most accurate step. This phenomenon can be attributed to the model's design, which employs a left-to-right reading paradigm, and closely mimics the natural process of human language comprehension. **(iii)** In addition, salient information flow from irrelevant information renders models prone to distractions, inadvertently causing them to focus on irrelevant content, which ultimately leads to failures in reasoning.

The information flow analysis reveals how irrelevant and disorganized information can affect model reasoning, and suggests the inertia of LLMs when tackling complex tasks is very similar to the one during human problem-solving process [17, 23]. However, in contrast to LLMs, the disordered and irrelevant content does not significantly decline human performance, which benefits from the fact that humans tend to distill the most relevant information and organize their thoughts in an orderly manner, such as constructing a mind map, in advance. This allows them to address the question more more quickly and accurately by referring to the mind map [16, 14].

Arise from that, we propose a novel reasoning approach named Concise and Organized Perception (COP). Specifically, COP initially performs capturing of locally-related premise segments among the given premises, with the intent to facilitate an initial comprehension of the input context. Next, COP leverages the query question as an anchor to integrate relevant pieces of locally-related premises generated by the first step, creating a tree-like mind map structure that presents global information in an orderly manner and can eliminate irrelevant information. Subsequently, LLMs are prompted by the reconstructed context, which are organized in a progressively ordered manner from the mind map to better adapt to the inference process of the model. We believe such reconstruction perceives more concise and organized information, which noticeably reduces the difficulty of model inference and better elicits the reasoning ability. Figure 1(b)(c) shows an example where LLMs are empowered to obtain the correct answer.

Meanwhile, we further conducted a simple confirmatory experiment by randomly selecting 196 samples and reconstructing the context based on the provided ground-truth proofs, as shown in

2

Figure 1(b) [1]. The results in Figure 1(d) demonstrate that combining our approach with the CoT baseline yields a relative performance improvement of over 100% (35.9% vs 71.9%) in a 5-hop setting. The results also indicate the complementarity between concise and organized perception. Besides synthetic logical reasoning, our method achieves consistent performance improvements on both real-world (FOLIO) and mathematical (DI-GSM) reasoning benchmarks. Specially, COP surpasses the CoT baseline by 9% on the FOLIO benchmark.

2   Related work

Large language models (LLMs) have demonstrated impressive few-shot learning capabilities [4, 32, 8, 29, 37]. Recent work has shown that LLMs, combined with in-context learning (ICL) and chain-of-thought (CoT) prompting, are capable of reasoning to an extent [21, 31, 28, 41, 26, 27]. However, it still remains highly challenging to achieve satisfactory results in complex logical problems. [42] explores the logical flaws of LLMs on logical reasoning datasets from four dimensions including answer correctness, explanation correctness, explanation completeness and explanation redundancy. [39] proposes an automatic approach to evaluate the logical reasoning abilities of LLMs based on propositional and predicate logic, which systematically identifies poor logical rules for LLMs' reasoning. In particular, the reasoning capabilities of LLMs are susceptible to deterioration when confronted with inputs that are either arranged in a disordered manner [33, 7] or peppered with irrelevant information [34]. Specifically, [33] investigated how reasoning ability is affected by the traversal direction of the ontology, and [7] found that the premise order significantly affects LLMs' reasoning performance. In another study, [34] observed that the performance of language models tends to decrease when irrelevant context is included in the problem statement. These phenomena aligns with the human preferences for solving logical problems [16, 14, 22]. Differing from their works, we further investigate the impact of disordered and irrelevant content on reasoning from the perspective of information flow and propose a consise and organized perception approach, drawing inspiration from the perspective of human problem-solving.

Benefiting from LLMs' strong logical reasoning ability, some methods [10, 13, 44] seek to encourage LLMs to generate reasoning steps explicitly and then produce results in a single stage, while some other methods seek to perform inference at multiple times to complete the tasks [45, 24]. Several recent works, such as LOGIC-LM [30], integrate LLMs with symbolic reasoning to improve logical problem-solving. LOGIC-LM first utilizes LLMs to translate a natural language problem into a symbolic formulation. Afterward, a deterministic symbolic solver performs inference on the formulated problem. Selection-Inference [11] alternates between selection and inference to generate a series of casual reasoning steps, and LAMBADA [25] develops a backward chaining algorithm to decompose reasoning into sub-modules. In addition to prompting methods, some works aim to fine-tune LLMs to produce the final answer directly, keeping reasoning implicit [9, 27]. In contrast to their works from the perspective of *how to plan* that encourage or teach LLMs how to solve complex logical problems, we introduce the effective and compelling "Concise" and "Organized" strategies to reduce the difficulty of LLMs' reasoning planning and better elicit their reasoning abilities, which can be considered as an alternative angle: reducing the difficulty of planning, or in other words, *easy to plan*.

3   Saliency score analysis

As a prevalent paradigm for interpretation, the concept of information flow can be instrumental in dynamically identifying critical interactions amongst tokens [12, 19, 40]. In this section, we leverage the information flow analysis methodology, predicated upon saliency scores derived from [40], to delve deeper into the reasons behind the pronounced degradation in LLMs' performance when confronted with disordered or irrelevant information. We perform analysis on the ProofWriter [35] dataset based on Llama-2-13B-Chat [38], and the detailed saliency score definition and analysis are listed in Appendix A.1. Figure 2(a) shows reasoning steps and saliency score analysis on a concise and organized example, which is consistent with the example in Figure 1. For comparison, the premises of example in Figure 2(b) are shuffled for analysis the impact of disordered information. Multiple irrelevant premises are added in the example in Figure 2(c), but the order of the relevant premises is

---

[1]Notice that this implementation is merely demonstrative, and differs from the actual method as no ground-truth can be utilized.
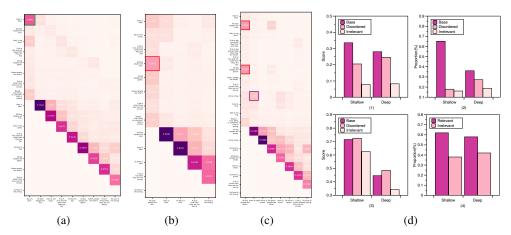
Figure 2: (a)(b)(c) Saliency score analysis on an example of ProofWriter based on shallow layers of Llama-2-13B-Chat. The horizontal coordinate contains the outputs the model generates step by step, and the vertical coordinate contains the inputs and outputs. Values in the plot represent saliency scores from column to row, normalized by each column. (a) A concise and organized example as **base**. (b) An example with **disordered** information for comparison. (c) An example with **irrelevant** information for comparison. (d) Saliency score analysis on ProofWriter based on shallow and deep layers of Llama-2-13B-Chat. "Base", "Disordered", and "Irrelevant" respectively denote the samples corresponding to the three scenarios depicted in (a)(b)(c). (d)(1) The saliency scores from the ground-truth reasoning entrance to the first reasoning step. (d)(2) The proportion of samples with the highest saliency score from the ground-truth reasoning entrance to the first reasoning step. (d)(3) The saliency scores from the previous two steps to the current step. (d)(4) The proportion of information flow from relevant and irrelevant information when contains irrelevant information.

consistent with the example in Figure 2(a), for analysis the impact of irrelevant information. From Figure 2, we can observe three phenomena. **(i)** Model can identify the correct entry point of the reasoning path during the initial step when confronted with concise and organized reasoning content, which is highlighted by the green box in Figure 2(a). In contrast, when the model encounters input that is presented in a disordered sequence or contains irrelevant information, it becomes markedly challenging for the model to ascertain an appropriate entry point for reasoning at the initial step, which are highlighted by the red boxes in Figure 2(b)(c). **(ii)** The information flow from the previous step to the current step is salient, as clearly depicted by the diagonal lines in Figure 2(a)(b)(c). This preference for the previous step excessively focuses attention there, complicating the allocation of sufficient effort to identify the subsequent correct step. Furthermore, it may lead to the generation of non-existent premises to cater to the content of the previous step, especially when confronted with disordered and irrelevant content. In Figure 2(b), "*if sb is nice then they are not young*" is a fake generated premise, which can be considered as hallucination. **(iii)** Finally, Figure 2(c) identifies an adverse effect of irrelevant information on reasoning, with the area highlighted by purple box clearly showing a pronounced manifestation of this impact. This leads to model becoming distractible, causing it to incorporate irrelevant content into the reasoning path, and ultimately resulting in reasoning failure.

Further, we conduct information flow analysis on 1200 samples in the ProofWriter dataset to understand why LLMs' performance significantly decreases when faced with disordered or irrelevant information, from a more holistic viewpoint. Drawing on [40], we conduct a comprehensive analysis of information flow across both the shallow and deep layers. As shown in Figure 2(d)(1), when inputs are concise and organized, the saliency score from the ground-truth reasoning entrance to the first reasoning step is significantly higher than when faced with disordered or irrelevant content. Figure 2(d)(2) shows that the proportion of samples with the highest saliency score from the ground-truth reasoning entrance to the first reasoning step is still the highest when inputs are concise and organized. Intuitively, such a preference for the reasoning entry aligns with the problem-solving processes observed in humans because locating information within organized and concise sets is considerably more straightforward than navigating through a pile of complex information. Figure 2(d)(3) shows the information flow from the previous two steps to the current step, which is also salient as we observed in the diagonal lines in Figure 2(a)(b)(c). In additional, irrelevant information has dispersed the flow of information as illustrated in Figure 2(d)(4), which is very similar to the inertia of humans. However,

4

Figure 3: Overview of the proposed COP with an example on DI-GSM (constructed from GSM8K [10]) with disordered and irrelevant premises. Green represents relevant premises $[p_2, p_3, p_4, p_6]$, black represents irrelevant premises $[p_1, p_5]$, and orange represents the question $[Q]$. Details of DI-GSM are listed in section 5.1.

in contrast to LLMs, the disordered and irrelevant content does not significantly decrease human performance, as humans have a propensity to distill the most relevant information and systematically organize their thoughts, aiding them in responding to questions. Thus, draw inspiration from the aspect of human problem-solving, we introduce two effective and compelling strategies including "Concise" and "Organized" perception to better elicit LLMs' reasoning abilities.

4 Approach

We present the Concise and Organized Perception (COP) reasoning approach, aiming to reconstruct concise and organized context as inputting to reduce the difficulty of model reasoning. Given a reasoning context containing plenty of premises $\mathcal{P} = \{p_1, p_2, ..., p_n\}$, which may include relevant, irrelevant, or disorganized information, the task is to answer a question $\mathcal{Q}$ based on the given premises.

In order to emulate human capability in processing complex logical reasoning tasks, we propose a three-stage method to effectively tackle disorder and distractibility. Firstly, we seek to capture locally-related premises based on their internal logical or semantic relationships, that is, connecting each single premise between each other to form a series of premise fragments. Such pieces of premise provide an initial structural grasp of the original context on a localized level. Secondly, in the face of several independent pieces, there is a crucial need for holistic systematization to foster comprehension at a global scale. Our approach leverages the question $\mathcal{Q}$ as an anchor to identify relevant fragments and integrate them into a whole tree-like mind map. This structure not only presents global information in an orderly manner but also discards any irrelevant premises. Subsequently, owing to the progressively organized manner of the mind map, COP creates a more concise and organized reasoning context that can be easily adapted to the inference process of LLMs. The details of these stages will be described in the following subsections.

4.1 Capturing of locally-related premises

It is generally not a wise strategy to hastily answer questions without fully grasping the given context when performing reasoning tasks; otherwise, it easily leads to inaccurate or incomplete reasoning. Therefore, instead of starting with looking for relevant clues step by step based on each single premise as previous methods (e.g., SI and LAMBADA) do, the first step of the proposed COP is capturing locally-related premises to form a series of premise fragments. This allows for an initial structural grasp of the context, facilitating the reconstruction of the context in later steps.

As mentioned previously, complex reasoning problems involve a set of premises $\mathcal{P}$. Imitating the process of human beings organizing thoughts, capturing of locally-related premises can be effectively achieved by employing directed edges to connect premises that bear relevance to one another. For example, an edge $p_i \rightarrow p_j$ connects premise p_i to premise p_j with the locally-related direction from i to j. This approach facilitates a localized understanding of the relationship between the premises p_i and p_j. For logical premises, especially modus ponens (p_a: if A then B; p_b: if B then C), capturing of locally-related premises can be performed by leveraging directed edges to connect each premise to premises whose consequents satisfy one or more of the conditions specified in the current premise. In the given case, premise p_b can be directly connected to premise p_a. Besides logical premises, capturing of locally-related premises can be performed through semantic correlation, temporal correlation, etc. Figure 3(a) illustrates an instance consisting of premises with semantic

5

correlation. Benefiting from LLMs' powerful in-context learning and semantic understanding ability, our proposed method encourages LLMs to perform capturing locally-related premises by searching relevant premises for each premise in the given context, which can be prompted with few-shot examples. As shown in Figure 3(a), there are six premises in the original problems, which are disordered and contain two irrelevant premises. Through the understanding of semantic correlation, LLMs adeptly identified and correlated relevant premises within vast premises. For example, p_2 is connected to p_6 while p_1 is connected to "None" in Figure 3(a). After capturing locally-related premises for each premise, different pieces of locally-related premises can be integrated again through the connection directions between premises, ultimately forming a initial understanding of the original problem's context. Figure 3(a) shows the four pieces of locally-related premises after integration, in which the relevance between premises is much clearer than that of the original input. Capturing of locally-related premises can be regarded as a preparatory step similar to the pre-processing that humans undertake before addressing complex tasks. The detailed prompts used in this step are listed in the Appendix A.3.

4.2 Generation of mind map

In section 4.1, locally-related premise fragments representing the preliminary structural understanding of the given reasoning context are captured. These captured pieces are independent and cannot be directly merged. Serving as an anchor point, the query question functions as a connecting bridge. Upon receiving a question, relevant clues can be identified among the pieces of locally-related premises, allowing for the creation of a tree-like mind map structure relevant to the question. This process effectively eliminates irrelevant information and presents a coherent global understanding of the relevant information in an ordered manner.

Specifically, COP encourages LLMs to find all premises centered around the question $\mathcal{Q}$, which can also be prompted with few-shot examples. The detailed prompts used in this step are listed in the Appendix A.3. As shown in Figure 3(b), two relevant pieces of locally-related premises are involved, and another two irrelevant ones are discarded. Although "*If Sara already had $10 saved before Sara started babysitting*" is a logical description of the question, COP can still correlate it to other pieces of locally-related premises, ultimately forming a holistic and ordered structure. Once we find the relevant pieces of locally-related premises, we perform a D-depth searching starting from each of them to avoid reasoning loops, where D is the max reasoning depth. In this way, a tree-like mind map based on the given question is constructed. Generating a tree-like mind map structure based on questions as anchor points is a crucial step in our method that facilitates the development of a structured arrangement centered around the core question while eliminating irrelevant information. Such a strategy provides a foundational guarantee for subsequent concise and organized context reconstruction, aligning with human problem-solving preferences.

4.3 Context reconstruction

In the generated mind map, the irrelevant premises to the problem are removed, and the most relevant premises to the problem are retained in order. A straightforward approach to perform context reconstruction is to employ a depth-first traversal technique to comprehensively traverse the entire mind map, thereby obtaining organized inputs, as shown in Figure 3(c). The premise order in the reconstructed context aligns with the progression needed for intermediate reasoning steps, thereby better eliciting the reasoning capabilities of LLMs. Moreover, compared with the original reasoning context, the reconstructed one has the advantage of being concise and dramatically reduces the influence of irrelevant information.

However, a single mind map may contain several sub-mind-maps, especially in scenarios involving logical reasoning. Figure 4 presents an example of a mind map consisting of two sub-mind-maps. Given a question "*Alex is young*" and its mind map generated in previous steps, the question is simultaneously relevant with both premise p_1 and p_2, leading to the natural formation of two sub-mind-maps within the larger mind map. Each sub-mind-map consists of several related premises with directed connections. Therefore, to streamline the reasoning process when faced with multiple sub-mind-maps, it is more efficient to segment the sub-mind-maps before embarking on the final reasoning phase, as opposed to directly engaging with the reconstructed context of the entire mind map. To answer the given question, a reasoning context for each possible sub-mind-map should be constructed. Subsequently, the reconstructed contexts are successively used to prompt the reasoning

Figure 4: An example on ProofWriter of sub-mind-map segmentation and context reconstruction.

of LLMs until a certain answer regarding the given question is made. As illustrated in Figure 4, when the model performs reasoning based on the reconstructed context in sub-mind-map 1 and obtains the exact answer, the reasoning for this question is completed, and the reconstructed context in sub-mind-map 2 is not needed. Such strategy of sub-mind-map segmentation can be regarded as a more refined approach for some exceptional situations.

5 Experiments

5.1 Datasets

(1) **ProofWriter** [35] is a commonly used logical reasoning dataset and contains five subsets, named $d5$, $d3$, $d2$, $d1$ and $d0$ respectively. dx part requires $\leq x$ hops for reasoning. We randomly sampled 600 examples in each part. (2) **PrOntoQA** is a synthetic logical reasoning dataset[2] and we use three parts $hop5$, $hop3$, and $hop1$ for testing. $hopx$ part requires x hops for reasoning. We randomly sampled 500 examples in each part. (3) **PrOntoQA-OOD** is another synthetic logical reasoning dataset containing different types of premises. We used the generated data file *generated_ood_data.zip* based on the open-source repository[3], and randomly selected 300 samples from the original $hop2$ part for testing. (4) **FOLIO** [18] is a real-world dataset for logical reasoning, written in highly natural wordings. We randomly sampled 100 examples for testing. (5) **DI-GSM** is a constructed dataset containing **d**irordered and **i**rrelevant information. Referring to [7], we first select GSM8K [10] test problems with at least five sentences in the problem statements and shuffle the sentences. Besides, we randomly add 2 to 3 irrelevant statements to the questions. The final testing data in DI-GSM contains 132 problems.

5.2 Experimental results

5.2.1 Performance comparison with state-of-the-art methods

In this section, we perform a thorough comparison between our proposed method and the existing state-of-the-art methods (Standard Few-Shot, CoT [41], SI [11], LOGIC-LM [30] and LAMBADA [25]). As typical methods of teaching model *how to plan*, SI [11] suggests alternating between selection and inference to generate reasoning steps based on forward chaining, while LAMBADA [25] employs a more explicit manner to introduce backward chaining for high-level proof planning. Since our approach naturally comes from a different perspective *easy to plan*, it can be seamlessly combined with other popular methods, such as SI or LAMBADA, to further enhance their performance. Unless otherwise specified, all the experimental results of COP are based on GPT-3.5-Turbo [29].

Table 1: Comparison of label accuracy on ProofWriter, PrOntoQA and PrOntoQA-OOD.

Datasets/	ProofWriter						PrOntoQA				PrOntoQA-OOD
Methods	d5	d3	d2	d1	d0	average	5-hop	3-hop	1-hop	average	-
Standard	41.67	49.83	51.00	55.50	63.67	52.33	49.60	52.00	65.60	55.73	43.33
CoT	53.50	61.17	61.33	62.33	62.83	60.23	69.80	74.20	86.20	76.73	85.67
SI	46.00	51.00	56.00	61.00	97.00	62.20	45.00	52.00	97.00	64.67	-
LogicLM	70.11	-	-	-	-	-	93.20	-	-	-	-
LAMBABDA	72.00	82.00	87.00	90.00	98.00	85.80	96.00	99.00	98.00	97.67	38.33
COP	**88.67**	**90.67**	**91.43**	**92.50**	**98.50**	**91.72**	**99.20**	**99.60**	**100.00**	**99.60**	**94.00**

[2]https://github.com/asaparov/prontoqa/tree/v1
[3]https://github.com/asaparov/prontoqa

Table 1 shows the results on ProofWriter, ProntoQA, and PrOntoQA-OOD. ProofWriter and ProntoQA contain various reasoning depths. The results of SI and LAMBADA are taken from [25]. COP consistently achieves the highest label accuracy across all experimental settings. Notably, COP outperforms SOTA methods by a large margin on the hardest Depth-5 subset of ProofWriter. It shows a remarkable 65.74% relative improvement compared to CoT and 23.15% compared to LAMBADA, which demonstrates the effectiveness of COP. Table 2 shows the results on FOLIO and DI-GSM. COP outperforms CoT and LogicLM, while LogicLM is not able to work on DI-GSM, and LAMBADA is not able to work on these two datasets, demonstrating the efficacy of COP. The results of LogicLM on FOLIO are reproduced, and the logic program errors in LogicLM are considered as the failure cases in our report. Besides, COP improves the performance while reducing the cost, and the detailed comparison of average token usage is listed in A.2.3.

In the next sections, we will analyze COP from several aspects, including concise and organized perception analysis from the perspective of information flow (5.2.2), error analysis caused by COP itself (5.2.3), error analysis when combined with other methods (5.2.4). In addition, COP consistently achieves high accuracy using different LLMs (including Qwen2-72B-Instruct, Llama-3-70B/8B-Instruct, and Llama-2-13B-Chat), and the detailed results are listed in A.2.1.

Table 2: Test on FOLIO and DI-GSM.

Methods	FOLIO	DI-GSM
Standard	23.00	46.97
CoT	38.00	44.70
LogicLM	34.00	-
COP	**47.00**	**53.79**

5.2.2 Concise and organized perception analysis

As analyzed in section3, disordered and irrelevant content has a negative impact on reasoning from the perspective of information flow. The proposed COP seeks to enhance reasoning performance and simplify the reasoning process by systematically removing irrelevant information and reorganizing the inputs. Thus, we analyze the performance of COP from two aspects: the degree of removal of irrelevant information and the order of input. Firstly, each sample in the DI-GSM dataset contains 2 to 3 irrelevant statements to the questions, for a total of 322 irrelevant statements. After applying COP, only nine statements have not been removed, demonstrating the success of concise perception in COP. Secondly, the statements of questions are randomly shuffled in DI-GSM. We use Kendall tau distance to measure the order difference of statements between the shuffled data and the original data. Kendall tau distance ranges from -1 to 1. The larger the Kendall tau distance, the more relevant it is. Before applying COP, the average Kendall tau distance is -0.0465, indicating no strong correlation between the statement order in the shuffled data and the original data. After applying COP, the average Kendall tau distance increases to 0.4268, demonstrating the success of organized perception in COP.
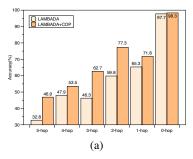
5.2.3 Failure case analysis

In this section, we focus on the errors caused by COP itself to study its possible flaws. We compare the results of COP on DI-GSM and CoT on DI-GSM with original concise and organized inputs based on GPT-3.5-turbo. The accuracy of COP is 53.79 (71/132), and the accuracy of CoT is 68.94 (91/132). We analyze the differences in the results and find that COP gets 15 more questions correct compared to CoT, and COP fails 35 more questions than CoT. After applying COP, the order of statements in the 15 correct questions is different from the original order, and the changed order generated by COP is more suitable for reasoning, demonstrating the success of organized perception in COP. The error types of the 35 failure cases are shown in Table 3. There are 10 cases where the order of the premises caused failure in reasoning. There are 13 cases where the premises are not connected to other premises in the given context. There are 2 cases where the premises are not output in the step of capturing of locally-related premises. These failure cases are attributed to the fact that capturing of locally-related premises is driven by LLMs, and it is difficult for the method to ensure that the generated connections between premises are completely accurate. Similarly, for the remaining 10 cases, some key relevant premises were discarded when generating mind maps, causing failure in reasoning. We leave the failure caused by these error types as future work.
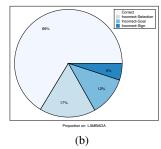
Table 3: The detailed failure case analysis on DI-GSM.

Step	Capturing of Locally-related Premises			Generation of Mind Map
Type	Connection Order	Connection Output	Connection None	Connection Discard
Errors	10/35	2/35	13/35	10/35

5.2.4 Is COP beneficial to other methods ?

Based on *easy to plan*, our COP can be seamlessly combined with methods that teach models *how to plan*, such as LAMBADA. Since LAMBADA cannot work on two real-world datasets, FOLIO and DI-GSM, we use ProofWriter for testing. The performance of LAMBADA and LAMBADA+COP on the ProofWriter d_5 subset with different inference depths are listed in Figure 5 (a). All the results are based on the LAMBADA code we reproduced, and the base model of this experiment is GPT-3.5-turbo. Compared with the original LAMBADA method, the performance of LAMBADA+COP under different inference depths is improved, proving the effectiveness of COP. In addition, Figure 5(b)(c) shows the proportion of correct reasoning and the proportion of different types of incorrect reasoning. We randomly selected 100 test samples from the ProofWriter d_5 subset to manually check the error types of incorrect reasoning examples. As shown in the figure, equipped with COP, the proportion of selection errors (including fact check and rule selection modules in LAMBADA) drops significantly. The proportion of goal decomposition errors and sign agreement errors (goal decomposition and sign agreement modules are unaffected by context redundancy and disorder) are almost unchanged, proving that our COP can improve the success rate of other methods in the steps that are affected by context redundancy and disorder.
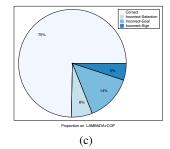

Figure 5: (a) Comparison of LAMBADA and LAMBADA+COP on Proofwriter. (b) The proportions of different error reasons of LAMBADA. (c) The proportions of different error reasons of LAMBADA + COP. There are four steps in LAMBADA, which are Fact Check step, Rule Selection step, Goal Decomposition step and Sign Agreement step. "Incorrect-Selection" means that LAMBADA fails in Fact Check or Rule Selection steps. "Incorrect-Goal" means that LAMBADA fails in Goal Decomposition steps. "Incorrect-Sign" means that LAMBADA fails in Sign Agreement steps.

6 Conclusion and future work

In this study, we propose a reasoning approach called Concise and Organized Perception (COP) to handle complex reasoning problems effectively. By combining *Concise* and *Organized* strategies with vanilla CoT, we have achieved state-of-the-art performance on multiple popular reasoning benchmarks. Besides, COP requires significantly fewer inference calls and tokens compared to decomposition-type methods (e.g., LAMBADA), highlighting our superiority in terms of both effectiveness and efficiency. In addition, we investigate the underlying mechanism of the influence of disordered and irrelevant content on reasoning, and reveal the inherent inertia of LLMs when tackling complex tasks, further supporting the motivation in COP.

We believe our crucial insight on the proposal of *easy-to-plan* method has broader implications. However, for more general reasoning tasks, performing robust capturing of locally-related premises and generating more appropriate tree-like mind map structures require further exploration. We plan to address these in future research.

References

[1] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[2] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

[3] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, L. Gianinazzi, J. Gajda, T. Lehmann, M. Podstawski, H. Niewiadomski, P. Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023.

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

[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[6] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[7] X. Chen, R. A. Chi, X. Wang, and D. Zhou. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*, 2024.

[8] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[9] P. Clark, O. Tafjord, and K. Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3882–3890, 2021.

[10] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] A. Creswell, M. Shanahan, and I. Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

[13] B. Dalvi, P. Jansen, O. Tafjord, Z. Xie, H. Smith, L. Pipatanangkura, and P. Clark. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*, 2021.

[14] M. Dekeyser, W. Schroyens, W. Schaeken, O. Spitaels, and G. d'Ydewalle. Preferred premise order in propositional reasoning: Semantic informativeness and co-reference. *Deductive reasoning and strategies*, pages 73–95, 2000.

[15] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR, 2023.

[16] V. Girotto, A. Mazzocco, and A. Tasso. The effect of premise order in conditional reasoning: A test of the mental model theory. *Cognition*, 63(1):1–28, 1997.

[17] T. Hagendorff, S. Fabi, and M. Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3 (10):833–838, 2023.

[18] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, L. Benson, L. Sun, E. Zubova, Y. Qiao, M. Burtell, D. Peng, J. Fan, Y. Liu, B. Wong, M. Sailor, A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, S. Joty, A. R. Fabbri, W. Kryscinski, X. V. Lin, C. Xiong, and D. Radev. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022.

[19] Y. Hao, L. Dong, F. Wei, and K. Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971, 2021.

[20] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[21] J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[22] P. N. Johnson-Laird. Mental models and deduction. *Trends in cognitive sciences*, 5(10):434–442, 2001.

[23] E. Jones and J. Steinhardt. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799, 2022.

[24] J. Jung, L. Qin, S. Welleck, F. Brahman, C. Bhagavatula, R. Le Bras, and Y. Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, 2022.

[25] M. Kazemi, N. Kim, D. Bhatia, X. Xu, and D. Ramachandran. LAMBADA: Backward chaining for automated reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.361.

[26] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[27] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

[28] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

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

[30] L. Pan, A. Albalak, X. Wang, and W. Y. Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*, 2023.

[31] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*, 2022.

[32] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[33] A. Saparov and H. He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023.

[34] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 31210–31227, 23–29 Jul 2023.

[35] O. Tafjord, B. Dalvi, and P. Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, 2021.

[36] A. Talmor, J. Herzig, N. Lourie, and J. Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[37] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[38] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[39] Y. Wan, W. Wang, Y. Yang, Y. Yuan, J.-t. Huang, P. He, W. Jiao, and M. R. Lyu. A & b== b & a: Triggering logical reasoning failures in large language models. *arXiv preprint arXiv:2401.00757*, 2024.

[40] L. Wang, L. Li, D. Dai, D. Chen, H. Zhou, F. Meng, J. Zhou, and X. Sun. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*, 2023.

[41] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[42] F. Xu, Q. Lin, J. Han, T. Zhao, J. Liu, and E. Cambria. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*, 2023.

[43] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[44] E. Zelikman, Y. Wu, J. Mu, and N. Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[45] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le, and E. H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

A   Appendix

A.1   Additional saliency score analysis

A.1.1   Saliency score definition

According to [40], saliency score for each element of the attention matrix is defined as:

$$I_l = \sum_h |A_{h,l}^T \frac{\partial L(x)}{\partial A_{h,l}}| \tag{1}$$

where $A_{h,l}$ is the attention matrix of the h-th head in the l-th layer, x is the inputs, $L(x)$ is the cross-entropy loss function, saliency score $I_l(i, j)$ is the significance of the information flow from token i to j in the x. The information flow from sentence A to sentence B is defined as:

$$IS(A, B) = \frac{1}{|\{lers\}|} \sum_{l \in \{lers\}} (\frac{1}{|t_a||t_b|} \sum_{a \in t_a} \sum_{b \in t_b} I_l(a, b)) \tag{2}$$

where t_a and t_b are the token sets of sentences A and B respectively, $|t_a|$ and $|t_b|$ are the length of t_a and t_b, $\{lers\}$ is the set of candidate layers in LLM. $IS(A, B)$ is normalized according to B in each generation step. Drawing on [40], we analyze information flow from the shallow and deep layers of the LLM. Specifically, given a model with L layers, we select first five layers $\{lers\} = \{1, 2, 3, 4, 5\}$ for shallow analysis, and select last five layers $\{lers\} = \{L-4, L-3, L-2, L-1, L\}$ for deep analysis. In our experiments, we analyze information flow from multiple aspects:

1. The saliency score from the ground-truth reasoning entrance to the first reasoning step is defined as:

$$A_1 = \frac{1}{N} \sum_{i=1}^N IS(s_{entrance}^{(i)}, g_1^{(i)}) \tag{3}$$

where $s_{entrance}^{(i)}$ is the ground-truth reasoning entrance sentence in inputs of sample i, $g_1^{(i)}$ is the first generated sentence (outputed by LLMs) of sample i, N is the total number of test samples. A larger A_1 means that the model can better find the reasoning entrance.

2. The proportion of samples with the highest saliency score from the ground-truth reasoning entrance to the first reasoning step is defined as :

$$A_2 = \frac{1}{N} \sum_{i=1}^N \delta(s_{entrance}^{(i)}, s_{opt}^{(i)}), s_{opt}^{(i)} = \arg\max_{s_j^{(i)}} IS(s_j^{(i)}, g_1^{(i)}) \tag{4}$$

where $s_{entrance}^{(i)}$ is the ground-truth reasoning entrance sentence in inputs of sample i, $s_j^{(i)}$ is the j-th sentence in inputs of sample i, $g_1^{(i)}$ is the first generated sentence (outputed by LLMs) of sample i, δ is the Kronecker function, N is the total number of test samples. A larger A_2 means that the model can also better find the reasoning entrance.

3. The saliency score from the previous two steps to the current step is defined as:

$$A_3 = \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K IS(g_{k-2}^{(i)}, g_k^{(i)}) + IS(g_{k-1}^{(i)}, g_k^{(i)}) \tag{5}$$

where $g_k^{(i)}$ is the k-th generated sentence (outputed by LLMs) of sample i, K is the total number of generated sentences, N is the total number of test samples. The value of A_3 indicates the information from the previous two steps to the current step. A larger A_3 means the model is more likely to pay attention to the information in previous two steps.

4. The proportion of information flow from relevant and irrelevant information when contains irrelevant information is defined as:

$$A_4 = \frac{r}{r+1},$$
$$r = \frac{1}{N} \sum_{i=1}^N \frac{\frac{1}{K|j_{irre}^{(i)}|} \sum_{k=1}^K \sum_{j \in j_{irre}^{(i)}} IS(s_j^{(i)}, g_k^{(i)})}{\frac{1}{K|j_{re}^{(i)}|} \sum_{k=1}^K \sum_{j \in j_{re}^{(i)}} IS(s_j^{(i)}, g_k^{(i)})} \tag{6}$$

13

where $g_k^{(i)}$ is the k-th generated sentence (outputed by LLMs) of sample i, K is the total number of generated sentences, $j_{re}^{(i)}$ is the index sets of relevant ground-truth reasoning sentences in inputs of sample i, $j_{irre}^{(i)}$ is the index sets of irrelevant sentences in inputs of sample i, $|j_{re}^{(i)}|$ and $|j_{irre}^{(i)}|$ are the length of $j_{re}^{(i)}$ and $j_{irre}^{(i)}$, N is the total number of test samples. Then, the saliency score from relevant information is defined as $1 - A_4$. A larger A_4 means that the information flow from irrelevant information is more salient, and the model is more likely to foucs on irrelevant information.

A.1.2 Additional analysis

Figure 6: Saliency score analysis on ProofWriter based on Qwen-14B-Chat. (a) The saliency scores from the ground-truth reasoning entrance to the first reasoning step (A_1). (b) The proportion of samples with the highest saliency score from the ground-truth reasoning entrance to the first reasoning step (A_2). (c) The saliency scores from the previous two steps to the current step (A_3). (d) The proportion of information flow from relevant and irrelevant information when contains irrelevant information (A_4).
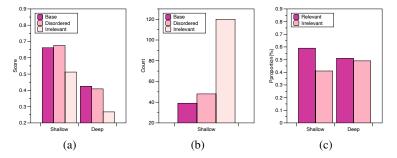

Figure 7: Generated fake premises analysis on ProofWriter based on Qwen-14B-Chat. (a) The saliency scores from the previous two steps to the current step (A_3) when generating a fake premise that is not in the given premises. (b) The number of the generated fake premises. (c) The proportion of information flow from relevant and irrelevant information when contains irrelevant information (A_4) when generating a fake premise that is not in the given premises.

In this section, we analyze the information flow based on saliency score on multiple models and datasets, to further support the observations in section3. We perform experiments on the ProofWriter and DI-GSM dataset based on Llama-2-13B-Chat and Qwen-14B-Chat. Figure 6 shows the saliency score analysis on ProofWriter based on Qwen-14B-Chat. Consistent with the analysis on ProofWriter based on Llama-2-13B-Chat, when inputs are concise and organized, the saliency score from the ground-truth reasoning entrance to the first reasoning step is significantly higher than when inputs are disordered or contain irrelevant information, as shown in Figure 6(a). The proportion of samples with the highest saliency score from the ground-truth reasoning entrance to the first reasoning step is still the highest when inputs are concise and organized. As shown in Figure 6(b), in the shallow layer analysis, the proportion when inputs are concise and organized is as high as 40%, while the proportion when inputs contain irrelevant information is 0%. In the deep layer analysis, the proportion when inputs are concise and organized is as high as 80%, while the proportion when inputs are disordered or

contain irrelevant information drops significantly. This analysis suggests that concise and organized inputs will empower the model to accurately identify the entry point for reasoning, thus minimizing failures attributed to inaccuracies at the reasoning entrance. Besides, information flow from the previous two steps to the current step and irrelevant information is also relatively salient, which is also consistent with the previous analysis. On the one hand, information flow from previous two steps to the current step is salient in Figure 6(c), which complicates the identification of the next correct reasoning path, and even makes up premises when faced with disordered and irrelevant content. On the other hand, as shown in Figure 6(d), salient information flow from irrelevant information makes models distractible, causing models to focus on irrelevant content and leading to reasoning failure.

In addition, we further analyze the generated fake premises when reasoning on ProofWriter based on Qwen-14B-Chat, which can be considered as hallucination. As shown in Figure 7(b), the number of the generated fake premises in the disordered case is larger than that in the concise and organized case, and the number of the generated fake premises in the irrelevant case is as high as 120. Comparing Figure 6(c) and Figure 7(a), information flow from the previous two steps to the current step is more salient when generating fake premises than the average. Information flow from irrelevant information in Figure 7(c) has the same trend, with an increase of about 10% in both shallow and deep layers. Figure 8 and Figure 9 show the saliency score analysis on DI-GSM based on Llama-2-13B-Chat and Qwen-14B-Chat. Similar results occurred on these models and data, further revealing the negative impact of disordered and irrelevant information on reasoning.
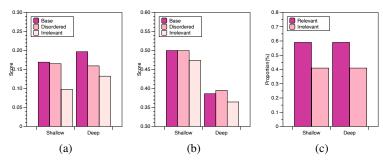

Figure 8: Saliency score analysis on DI-GSM based on Llama-2-13B-Chat. (a) The saliency scores from the ground-truth reasoning entrance to the first reasoning step (A_1). (b) The saliency scores from the previous two steps to the current step (A_3). (c) The proportion of information flow from relevant and irrelevant information when contains irrelevant information (A_4).
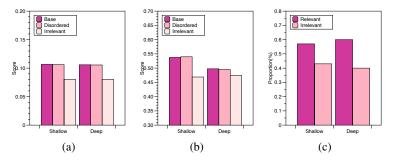

Figure 9: Saliency score analysis on DI-GSM based on Qwen-14B-Chat. (a) The saliency scores from the ground-truth reasoning entrance to the first reasoning step (A_1). (b) The saliency scores from the previous two steps to the current step (A_3). (c) The proportion of information flow from relevant and irrelevant information when contains irrelevant information (A_4).

A.2 Additional experimental analysis

A.2.1 Performance with different large language models

Most experiments conducted in this paper were performed on ChatGPT-3.5-turbo[29]. To study if the proposed COP is effective across different base models, we conducted experiments on DI-GSM

with Qwen2-72B-Instruct [4], Llama-3-70B/8B-Instruct [5], and Llama-2-13B-Chat [38]. As shown in Table 4, COP consistently achieves high accuracy using different LLMs, revealing its effectiveness across different LLMs.

Table 4: The performance comparisons on DI-GSM using different LLMs.

Methods	Qwen2 -72B-Instruct	Llama-3 -70B-Instruct	Llama-3 -8B-Instruct	Llama-2 -13B-Chat
CoT	78.79	72.73	43.18	11.36
COP	**80.30**	**78.79**	**52.27**	**15.15**

A.2.2 Proof accuracy analysis

Table1 lists the label accuracy comparison, while proof accuracy is a more stringent metric. Previous studies have demonstrated that CoT predicts a correct label with incorrect reasoning chains [33]. To validate if it is the case for COP, we randomly selected 100 correctly answered samples from the Depth-5 setting of ProofWriter. We manually checked the reasoning chain produced by LLMs with COP. According to our observation, only 7 out of 100 samples contain invalid reasoning steps, which indicates that the proposed COP does arouse the reasoning ability of LLMs, and the experimental results reported above are faithful.

A.2.3 Number of inference calls and tokens

In Figure 10, we compared the average number of inference calls per example under different reasoning depths on the ProofWriter dataset. COP requires significantly fewer inference calls than LAMBADA, and the number of inference calls remains relatively stable as the number of hops increases. Besides, Table 5 compares token numbers used per question on the ProofWriter dataset with different hops. The token numbers are taken from the usage statistics returned by the OpenAI API. COP-Prompt and LAMBADA-Prompt stand for the input token numbers of COP and LAMBADA, while COP-Total and LAMBADA-Total stand for the overall token consumed by input and output. As shown in the table, COP costs much fewer token numbers than LAMBADA, and the number of token numbers remains relatively stable as the number of hops increases, demonstrating our proposed COP's superiority in both effectiveness and efficiency.

While LAMBADA is not able to work on DI-GSM and FOLIO, and LogicLM is not able to work on DI-GSM, we compared the average number of inference calls and tokens used in COP and LogicLM on FOLIO. Table 6 shows the results. COP still costs fewer inference calls and much fewer token numbers than LogicLM, which demonstrates the superiority of COP in terms of efficiency.

Table 5: Comparison of average token numbers on the ProofWriter dataset.

Hops	0	1	2	3	4	5
LAMBADA-Prompt	567.71	4825.98	8154.11	9247.04	14401.85	19200.05
LAMBADA-Total	611.76	5293.22	8992.39	10333.2	15944.14	21922.77
COP-Prompt	433.21	1876.82	1915.68	1953.45	1996.62	2004.97
COP-Total	594.53	2199.44	2270.31	2341.29	2425.71	2440.26

Table 6: Comparison of average inference calls and token numbers on the FOLIO dataset.

Method	Calls	Prompt-tokens	Total-tokens
LogicLM	3.88	6204.3957	7281.5731
COP	3	3801.9971	4104.9285

---

[4]https://github.com/QwenLM/Qwen2
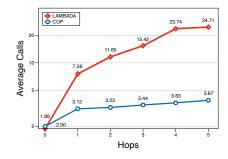[5]https://github.com/meta-llama/llama3.git

Figure 10: Comparison of inference calls on the ProofWriter dataset.

A.3 Prompts used in experiments

In this section, we give the prompts used in our experiments. The prompts used in PrOntoQA and PrOntoQA-OOD are similar to that in ProofWriter.

A.3.1 Prompts in capturing of locally-related premises

Prompts for DI-GSM [Capturing of locally-related premises]

Given multiple statements in a context, the task is to find relevant statements for each statement. Use "A -> B" to denote statement B that is relevant after statement A. Use "A -> None" to denote that there is no statement that is relevant after statement A. Each statement can have multiple relevant statements. Do not change the logic and content of the statements in context.

——

Context:
James makes potatoes for a group.
For every 5 fruits that customers buy, the store offers a $1 discount.
Mary went to the store to buy fruit.
Each person eats 1.5 pounds of potatoes.
Apples cost $1, oranges cost $2, and bananas cost $3.
Mary buys 5 apples, 3 oranges, and 2 bananas.
Margaret wants to serve chicken salad sandwiches using mini croissants.

Answer:
James makes potatoes for a group. -> Each person eats 1.5 pounds of potatoes.
For every 5 fruits that customers buy, the store offers a $1 discount. -> None.
Mary went to the store to buy fruit. -> Mary buys 5 apples, 3 oranges, and 2 bananas.
Each person eats 1.5 pounds of potatoes. -> None.
Apples cost $1, oranges cost $2, and bananas cost $3. -> For every 5 fruits that customers buy, the store offers a $1 discount.
Mary buys 5 apples, 3 oranges, and 2 bananas. -> Apples cost $1, oranges cost $2, and bananas cost $3.
Margaret wants to serve chicken salad sandwiches using mini croissants. -> None.

——

Context:
...

——

Context:
The middle height tree is 2/3 the height of the tallest tree.
At the burger hut, you can buy a burger for $5, french fries for $3, and a soft drink for $3.
There are three trees in the town square.
The tallest tree is 150 feet tall.
George is about to celebrate his 25th birthday.
The shortest tree is half the size of the middle tree.

Answer:
The middle height tree is 2/3 the height of the tallest tree. -> The shortest tree is half the size of the middle tree.
At the burger hut, you can buy a burger for $5, french fries for $3, and a soft drink for $3. -> None.
There are three trees in the town square. -> The tallest tree is 150 feet tall.
The tallest tree is 150 feet tall. -> The middle height tree is 2/3 the height of the tallest tree.
George is about to celebrate his 25th birthday. -> None.
The shortest tree is half the size of the middle tree. -> None.

——

Context:
[[PREMISES]]

Answer:

Prompts for FOLIO [Capturing of locally-related premises]

Given multiple statements in a context, the task is to find logically relevant statements for each statement. Use "A -> B" to denote statement B that is logically relevant after statement A. Use "A -> None" to denote that there is no statement that is logically relevant after statement A. Each statement can have multiple logically relevant statements. Do not change the logic and content of the statements in context.

——

Context:
A thing is either a plant or animal.
If a sea eel is either an eel or a plant, then a sea eel is an eel or an animal.
No fish are plants.
All animals breathe.
Nothing that breathes is paper.
All eels are fish.

Answer:
A thing is either a plant or animal. -> All animals breathe.
If a sea eel is either an eel or a plant, then a sea eel is an eel or an animal. -> All eels are fish.
If a sea eel is either an eel or a plant, then a sea eel is an eel or an animal. -> All animals breathe.
No fish are plants. -> None.
All animals breathe. -> Nothing that breathes is paper.
Nothing that breathes is paper. -> None.
All eels are fish. -> No fish are plants.

——

Context:
...

——

Context:
All Instagram is entertainment.
All video applications are software.
If something is interesting, then it is good.
All YouTube-related applications are video applications.
All entertainments are interesting.
TikTok is not good.
All software is programmed.
An APP is either related to YouTube or Instagram.

Answer:
All Instagram is entertainment. -> All entertainments are interesting.
All video applications are software. -> All software is programmed.
If something is interesting, then it is good. -> TikTok is not good.
All YouTube-related applications are video applications. -> All video applications are software.
All entertainments are interesting. -> If something is interesting, then it is good.
TikTok is not good. -> None.
All software is programmed. -> None.
An APP is either related to YouTube or Instagram. -> All YouTube-related applications are video applications.
An APP is either related to YouTube or Instagram. -> All Instagram is entertainment.

——

Context:
[[PREMISES]]

Answer:

Prompts for Proofwriter [Capturing of locally-related premises / Premises Unified Formats]

You are given some known rules. Extract the conditions and consequents of each rule and output follow the format of the given examples:
Examples:
Rules:
If someone sees the cat and they are not green then they see the cow. If the rabbit is kind and the rabbit sees the squirrel then the squirrel needs the rabbit. Rough people are cold. If someone sees the rabbit then they are not round. If someone sees the squirrel and they are not green then they need the squirrel. If someone eats the cow then they see the rabbit. Cold things are rough. If someone is cold then they eat the cow. Kind, rough people are round.
Output:
"Rule1": "conditions": ["X(see, cat)", "X(is not, green)"], "consequents": ["X(see, cow)"], "Rule2": "conditions": ["rabbit(is, kind)", "rabbit(see, squirrel)"], "consequents": ["squirrel(need, rabbit)"], [...], "Rule9": "conditions": ["X(is, kind)", "X(is, rough)"], "consequent": ["X(is, round)"]

Rules:
...

Rules:
If something visits the mouse and the mouse visits the dog then it is cold. If mouse likes the cat then it visits the dog. If something is cold then it likes the cat. If something is green then it sees the dog. If something likes the mouse then it sees the cat. If dog is green and cold then it likes the cat. If something is big and it visits the bear then the bear is green. Round things are rough. Output:
"Rule1": "conditions": ["X(visit, mouse)", "mouse(visit, dog)"], "consequents": ["X(is, cold)"], "Rule2": "conditions": ["mouse(like, cat)"], "consequents": ["X(visit, dog)"], "Rule8": "conditions": ["X(is, round)"], "consequents": ["X(is, rough)"]

Rules:
[[PREMISES]]

Output:

Prompts for Proofwriter [Capturing of locally-related premises / Premises Unified Formats]

You are given some known facts. Output the facts following the format of the given examples:
Examples:
Facts:
The bear is green. The bear likes the cat. The bear likes the dog. The bear visits the dog. The cat isyoung. The cat does not see the bear. The cat sees the dog. The cat visits the bear. The dog is round. The mouse is not big. The mouse is cold.
Output:
"Fact1": ["bear(is, green)"], "Fact2": ["bear(like, cat)"], "Fact3": ["bear(like, dog)"], "Fact4": ["bear(visit, dog)"], "Fact5": ["cat(is, young)"], "Fact6": ["cat(not see, bear)"], "Fact7": ["cat(see, dog)"], "Fact8": ["cat(visit, bear)"], "Fact9": ["dog(is, round)"], "Fact10": ["mouse(is not, big)"], "Fact11": ["mouse(is, cold)"]

Facts:
...

Facts: [[PREMISES]]

Output:

A.3.2 Prompts in generation of mind map

Prompts for DI-GSM [Generation of Mind Map]

Given a question and multiple ordered relevant sentences in context, the task is to find in order all sentences in the context that are required to answer the given question or are related to the information and subject in the given question.

―――

Context:
Mary went to the store to buy fruit. Mary buys 5 apples, 3 oranges, and 2 bananas. Apples cost $1, oranges cost $2, and bananas cost $3.
James makes potatoes for a group. Each person eats 1.5 pounds of potatoes.
For every 5 fruits that customers buy, the store offers a $1 discount.
Margaret wants to serve chicken salad sandwiches using mini croissants.

Question:
How much will Mary pay?

Inference:
All sentences in order that are required to answer the given question or are related to the information and subject in the given question are: Mary went to the store to buy fruit. -> Mary buys 5 apples, 3 oranges, and 2 bananas. -> Apples cost $1, oranges cost $2, and bananas cost $3. -> For every 5 fruits that customers buy, the store offers a $1 discount.

―――

Context:
...

―――

Context:
At the burger hut, you can buy a burger for $5, french fries for $3, and a soft drink for $3.
The tallest tree is 150 feet tall. The middle height tree is 2/3 the height of the tallest tree. The shortest tree is half the size of the middle tree.
George is about to celebrate his 25th birthday.
There are three trees in the town square.

Question:
How tall is the shortest tree?

Inference:
All sentences in order that are required to answer the given question or are related to the information and subject in the given question are: There are three trees in the town square. -> The tallest tree is 150 feet tall. -> The middle height tree is 2/3 the height of the tallest tree. -> The shortest tree is half the size of the middle tree.

―――

Context:
[[PREMISES]]

Question:
[[QUESTION]]

Inference:

Prompts for FOLIO [Generation of Mind Map]

Given multiple ordered logical paths in context and a statement, the task is to find the most logically relevant paths for the statement and remove irrelevant logical content in context for the statement.
____

Context:
If a sea eel is either an eel or a plant, then a sea eel is an eel or an animal. All eels are fish. No fish are plants.
If a sea eel is either an eel or a plant, then a sea eel is an eel or an animal. All animals breathe. A thing is either a plant or animal. All animals breathe. Nothing that breathes is paper.

Statement:
Sea eel is a paper.

Answer:
A thing is either a plant or animal. All animals breathe. Nothing that breathes is paper. If a sea eel is either an eel or a plant, then a sea eel is an eel or an animal. All animals breathe. All eels are fish. No fish are plants.

____

Context:
...

____

Context:
An APP is either related to YouTube or Instagram. All YouTube-related applications are video applications. All video applications are software. All software is programmed.
An APP is either related to YouTube or Instagram. All Instagram is entertainment. All entertainments are interesting. If something is interesting, then it is good. TikTok is not good.

Statement:
Tiktok is a program.

Answer:
An APP is either related to YouTube or Instagram. All Instagram is entertainment. All entertainments are interesting. If something is interesting, then it is good. TikTok is not good. All YouTube-related applications are video applications. All video applications are software. All software is programmed.

____

Context:
[[PREMISES]]

Statement
[[QUESTION]]

Answer:

Prompts for ProofWriter [Generation of Mind Map / Question Unified Formats]

You are given some known facts. Output the facts following the format of the given examples:
Examples:
Facts:
The bear is green.
Output:
["bear(is, green)", "bear(is not, green)"]

Facts:
...

Facts:
The mouse is not big. Output:
["mouse(is not, big)", "mouse(is, big)"]
Facts:
[[QUESTION]]

Output:

A.3.3 Prompts in CoT

> **Prompts for DI-GSM [CoT Reasoning]**
>
> Question: Karen's students are about to take a standardized test. Karen gets a $500 bonus if their average score is above 75, plus an extra $10 bonus for every additional point the average score increases above 75. So far, Karen has graded 8 tests, and the average is 70. Given that each student can have a maximum score of 150, what combined score do the last two tests need to have for Karen to earn a $600 bonus?
>
> Answer: First subtract $500 from Karen's goal bonus amount to find how much she makes from the extra $10/point bonus: $600 - $500 = $«600-500=100»100\nThen divide the extra bonus by the extra rate: $100 / $10/point = «100/10=10»10 points\nThen add the 10 extra points to the baseline 75 point goal to find the students' average test score: 10 points + 75 points = «10+75=85»85 points\nThen added the 8 graded tests to the 2 ungraded tests to find the total number of tests: 2 tests + 8 tests = «2+8=10»10 tests\nThen multiply the 85 point average by the number of tests to find the total number of points the students need to earn: 85 points/test * 10 tests = 850 points\nThen multiply the current average by the current number of graded tests to find how many points have been earned so far: 70 points/test * 8 tests = «70*8=560»560 points\nThen subtract the number of points earned from the number of points needed to find the combine score the last two tests need: 850 points - 560 points = «850-560=290»290 points. The answer is 290.
>
> Question: Mr. Manuel is a campsite manager who's been tasked with checking the number of tents set up in the recreation area. On a particular day, he counted 100 tents in the northernmost part of the campsite and twice that number on the east side of the grounds. The number of tents at the center of the camp was four times the number of tents in the northernmost part of the campsite. If he also counted 200 tents in the southern part of the campsite, what is the total number of tents in the recreation area?
>
> Answer: On the eastern part of the campsite, Mr Manuel counted 2*100 = «2*100=200»200 tents\nThe total number of tents in the eastern and the northern part of the campgrounds is 200+100 = «200+100=300»300\nThere are four times as many tents as the northernmost part in the central part of the campsite, which means there are 4*100 = «4*100=400»400 tents in the central part of the camp.\nThe total number of tents in the three parts of the campsite is 400+300 = «400+300=700»700\nIf you add the number of tents Mr. Manuel counted at the southern part of the campsite, you get 700+200 = «700+200=900»900 tents on the whole campsite. The answer is 900.
>
> Question: [[PREMISES]][[QUESTION]]
>
> Answer:

Given a problem statement as contexts, the task is to answer a logical reasoning question.

———

Context: The Blake McFall Company Building is a commercial warehouse listed on the National Register of Historic Places. The Blake McFall Company Building was added to the National Register of Historic Places in 1990. The Emmet Building is a five-story building in Portland, Oregon. The Emmet Building was built in 1915. The Emmet Building is another name for the Blake McFall Company Building. John works at the Emmet Building.

Question: Based on the above information, is the following statement true, false, or uncertain? The Blake McFall Company Building is located in Portland, Oregon.

Options: A) True B) False C) Uncertain

Reasoning: The Blake McFall Company Building is another name for the Emmet Building. The Emmet Building is located in Portland, Oregon. Therefore, the Blake McFall Company Building is located in Portland, Oregon.

The correct option is: A

———

Context: People eat meat regularly or are vegetation. If people eat meat regularly, then they enjoy eating hamburgers and steaks. All people who are vegetarian are conscious of the environment or their health. If people are conscious about the environment or their health, then they do not go to fast food places often. If people have busy schedules without time to cook, then they go to fast food places often. If Jeremy does not both go to fast food places often and is conscious about the environment or their health, then he goes to fast food places often.

Question: Based on the above information, is the following statement true, false, or uncertain? If Jeremy has a busy schedule without time to cook, then Jeremy does not enjoy eating hamburgers and steaks.

Options: A) True B) False C) Uncertain

Reasoning: If Jeremy has a busy schedule without time to cook or enjoy eating hamburgers and steaks, then Jeremy goes to fast food places often. If people are conscious about the environment or their health, then they do not go to fast food places often. This means that Jeremy is not conscious about the environment or his health. All people who are vegetarian are conscious of the environment or their health. Therefore, Jeremy is not vegetarian. People eat meat regularly or are vegetation. Therefore, Jeremy eats meat regularly. If people eat meat regularly, then they enjoy eating hamburgers and steaks. Therefore, Jeremy enjoys eating hamburgers and steaks.

The correct option is: B

———

Context: [[PREMISES]]

Question: [[QUESTION]]

Options: A) True B) False C) Uncertain

Reasoning:

Prompts for ProofWriter [CoT Reasoning]

Given a problem statement as contexts, the task is to answer a logical reasoning question.

———

Context: The cow is blue. The cow is round. The cow likes the lion. The cow visits the tiger. The lion is cold. The lion is nice. The lion likes the squirrel. The squirrel is round. The squirrel sees the lion. The squirrel visits the cow. The tiger likes the cow. The tiger likes the squirrel. If something is cold then it visits the tiger. If something visits the tiger then it is nice. If something sees the tiger and it is young then it is blue. If something is nice then it sees the tiger. If something likes the squirrel and it likes the cow then it visits the tiger. If something is nice and it sees the tiger then it is young. If the cow is cold and the cow visits the lion then the lion sees the squirrel.

Question: Based on the above information, is the following statement true, false, or unknown? The tiger is not young.

Options: A) True B) False C) Unknown

Reasoning: The tiger likes the cow. The tiger likes the squirrel. If something likes the squirrel and it likes the cow, then it visits the tiger. So the tiger visits the tiger. If something visits the tiger then it is nice. So the tiger is nice. If something is nice and it sees the tiger then it is young. So the tiger is young.

The correct option is: B

———

Context: The dog sees the rabbit. The dog sees the squirrel. The dog sees the tiger. The rabbit eats the dog. The rabbit does not eat the tiger. The rabbit does not like the tiger. The squirrel does not see the rabbit. The tiger does not eat the rabbit. The tiger is not kind. The tiger likes the dog. The tiger sees the dog. If something is cold then it likes the rabbit. If something eats the tiger and it is nice then it likes the rabbit. If something likes the squirrel then the squirrel likes the rabbit. If something likes the rabbit and the rabbit is kind then it sees the tiger. If something likes the tiger then the tiger is young. If something is young and it eats the rabbit then it likes the tiger. If something sees the rabbit then the rabbit is cold. If something likes the rabbit then it likes the squirrel. If something likes the squirrel then the squirrel is cold.

Question: Based on the above information, is the following statement true, false, or unknown? The rabbit is cold.

Options: A) True B) False C) Uncertain

Reasoning: The dog sees the rabbit. If something sees the rabbit then the rabbit is cold. So the rabbit is cold.

The correct option is: A

———

Context: [[PREMISES]]

Question: [[QUESTION]]

Options: A) True B) False C) Uncertain

Reasoning:

———

26