Reverse Multi-Choice Dialogue Commonsense Inference with Graph-of-Thought

Li Zheng[1], Hao Fei[2], Fei Li[1]*, Bobo Li[1], Lizi Liao[3], Donghong Ji[1], Chong Teng[1]

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2]National University of Singapore
[3]Singapore Management University
zhengli@whu.edu.cn, haofei37@nus.edu.sg, lifei_csnlp@whu.edu.cn, boboli@whu.edu.cn,
lzliao@smu.edu.sg, dhji@whu.edu.cn, tengchong@whu.edu.cn

Abstract

With the proliferation of dialogic data across the Internet, the *Dialogue Commonsense Multi-choice Question Answering* (DC-MCQ) task has emerged as a response to the challenge of comprehending user queries and intentions. Although prevailing methodologies exhibit effectiveness in addressing single-choice questions, they encounter difficulties in handling multi-choice queries due to the heightened intricacy and informational density. In this paper, inspired by the human cognitive process of progressively excluding options, we propose a three-step *Reverse Exclusion Graph-of-Thought* (**ReX-GoT**) framework, including Option Exclusion, Error Analysis, and Combine Information. Specifically, our ReX-GoT mimics human reasoning by gradually excluding irrelevant options and learning the reasons for option errors to choose the optimal path of the GoT and ultimately infer the correct answer. By progressively integrating intricate clues, our method effectively reduces the difficulty of multi-choice reasoning and provides a novel solution for DC-MCQ. Extensive experiments on the CICERO and CICERO$_{v2}$ datasets validate the significant improvement of our approach on DC-MCQ task. On zero-shot setting, our model outperform the best baseline by 17.67% in F1 score for the multi-choice task. Most strikingly, our GPT3.5-based ReX-GoT framework achieves a remarkable 39.44% increase in F1 score.

Introduction

Commonsense knowledge is crucial for human cognition and natural human-computer interactions, which encompasses our intuitive understanding of the world and ability to reason. With the growth of social networks, commonsense inference (Arabshahi et al. 2021; Liu et al. 2022; Kuo and Chen 2023) in dialogue has garnered noteworthy attention as a burgeoning research domain in natural language processing (NLP). However, accurately understanding and interpreting speaker questions and intentions in dialogue poses an essential challenge. To this end, the *Dialogue Commonsense Multi-choice Question Answering* task (Ghosal et al. 2022b) was proposed, defined as to select logical answers from preset options based on dialogue's history and context.

DC-MCQ task involve both single-choice and multi-choice questions. While existing works (Wang et al. 2018;
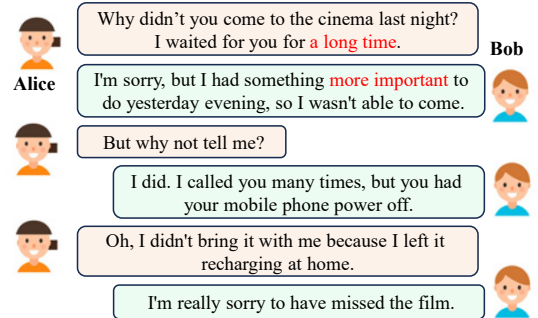
Figure 1: An example from the CICERO$_{v2}$ dataset about dialogue commonsense reasoning.

Zhang et al. 2020; Ju et al. 2021) achieved promising results in single-choice task, the performance in multi-choice task remains unsatisfactory. Due to the intricate nature of multi-choice task, the challenges of **Option Saturation** and **Clue Labyrinth** burden current models. The option saturation challenge refers to the uncertainty of the number of options, which increases the difficulty of inference for the model. On a parallel note, the clue labyrinth challenge involves analyzing the combination of different complex clues, which includes intricate hidden information woven throughout the question stem and answer options, and different clues of predicted information, just like the complexity of the labyrinth. It demands enhanced information integration comprehension by the model. Hence, multi-choice questions are significantly more challenging than single-choice ones. As indicated by Ghosal et al. (2022a), the community recognizes that attaining high accuracy with such questions is a potentially insurmountable task.

Existing methods for multi-choice questions, as highlighted by Ghosal et al. (2022b) and Shen et al. (2022), predominantly rely on forward reasoning. This typically assesses each option in isolation, which falters in accurately identifying the right answers due to intricate interrelations and uncertainties among choices. Motivated by human cognitive patterns of option exclusion, we employ a similar tactic to progressively narrow down potential answers. As exemplified in Figure 1, depending on the context, we exclude certain options such as D and C, obtaining clues that Bob had something more important to do and the correct option must being important as well as taking a long time. Continuing reasoning based on the context and the clues we have, we determine that options A, B, and E are correct. This exclusion-centric approach enhances reasoning, uncovers obscured insights from incorrect options, and greatly eases the prediction challenge in multi-answer scenarios.

On the other hand, the context of each option in the multi-choice task is broad enough to go beyond the scope of the given dialogue. Models based on direct answer selection struggle to fully comprehend the multi-dimensional and complex relationships between the question and the options, which can lead to model reasoning overload affecting accuracy. With the widespread use of Large Language Models (LLMs) in NLP tasks, researchers (Wei et al. 2022; Fei et al. 2023; Jin and Lu 2023; Zhang et al. 2023) have identified the capacity of Chain-of-Thought (CoT) to help LLMs with complex reasoning tasks by generating intermediate steps. However, the existing CoT reasoning of LLMs is limited to performing linear reasoning, and is unable to utilize potential multi-clue reasoning in a multi-dimensional manner to solve the clue labyrinth chanllenge. Moreover, the existing CoT methods only superficially exploit the contextual information and overlook the utilization of the exclusion method to harness the hidden information within the options.

In this paper, based on the above observations, we design a three-step *Reverse Exclusion Graph-of-Thought* (**ReX-GoT**) framework, including **Option Exclusion**, **Error Analysis**, and **Combine Information**. ReX-GoT mimics human exclusion and selection methods by generating reverse exclusion graph-of-thought prompts. Concretely, leveraging LLMs as our basis, as shown in Figure 2, we initially prompt the model to discern irrational options and their underlying reasons. Subsequently, we utilize the insights gained in the first step for error analysis and option comparison to further guide the model to determine the rationality of each option and justifying its choice. Finally, we combine the different reasons extracted in the first two steps as different paths and select the best path through a voting mechanism to arrive at the final multiple choice answer. This distinctive amalgamation of backward exclusion and forward reasoning systematically excludes irrelevant alternatives and comprehends errors, thereby alleviating the complexity of predicting multiple correct responses.

To verify the effectiveness of our model, we conduct experiments on two widely-used datasets for DC-MCQ, namely CICERO (Ghosal et al. 2022b) and CICERO$_{v2}$ (Shen et al. 2022). On zero-shot setting, the experiment on the CICERO$_{v2}$ dataset show that the F1 score of our ReX-

GoT is 17.67% higher than the best baseline. Most strikingly, our GPT3.5-based ReX-GoT with 175B parameters boosts the baseline to a high-to 39.44% increase of F1 score.

Our main contributions are summarized as follows:

- We first propose an reverse exclusion method that is consistent with human cognition, which effectively solves the challenge of option saturation by repeatedly excluding incongruous options and gradually revealing the hidden context of the correct option.

- We design a brand-new GoT framework to productively address the clue labyrinth challenge. In this framework, different inference paths are set according to different analyses of the options, and the optimal path is finally selected to derive the correct answer.

- Our extensive experimental results on CICERO and CICERO$_{v2}$ datasets demonstrate that our scheme achieves state-of-the-art performance on the DC-MCQ task[1].

Related Work

Commonsense Question Answering

The domain of commonsense question answering has garnered substantial attention within the realm of NLP. Existing models (Chen et al. 2023; Dou and Peng 2022; Ma et al. 2021) have demonstrated remarkable capabilities in understanding and reasoning about common knowledge. Previous approaches such as prompt techniques (Ma et al. 2023; Zeng et al. 2023; Paranjape et al. 2021) were proposed to improve the performance of language models in commonsense question answering task. Additionally, graph-based frameworks (Zhao et al. 2023; Zheng et al. 2023b; Bosselut, Bras, and Choi 2021), including knowledge graphs and concept graphs, were also employed to enhance the representation and utilization of commonsense knowledge.

Commonsense Inference in Dialogues

Recently, there has been a growing interest in developing dialogue commonsense inference models (Arabshahi et al. 2021; Ghosal et al. 2021; Richardson and Heck 2023). Several studies have been conducted on this topic: Qin et al. (2021) investigated pre-trained language models for their temporal reasoning capabilities in dialogues. Furthermore, Ghosal et al. (2022b) introduced a dialogue commonsense inference dataset CICERO that enables models to make educated guesses by considering the context when the answer is not obvious. Shortly after, Shen et al. (2022) proposed the CICERO$v2$ dataset, which provides more diverse options than CICERO. Based on the two datasets, a recent study (Ghosal et al. 2022a) transformed the task of selecting answers to a binary classification problem. However, despite achieving certain results via direct classification, this approach overlooks the importance of step-by-step reasoning, which can significantly impact the analysis of results.

---

[1]Codes available at https://github.com/ZhengL00/ReX-GoT

LLM Reasoning with Chain-of-Thought

The tremendous success of LLMs (Brown et al. 2020; Wu et al. 2023) has propelled the development of various downstream applications, such as mathematical reasoning (Yao et al. 2023; Imani, Du, and Shrivastava 2023), sentiment analysis (Fei et al. 2023; Zheng et al. 2023a), and chatbot (Ouyang et al. 2022; Deng et al. 2023). To exploit the reasoning ability in LLMs, recent works (Wei et al. 2022; Wang et al. 2023b; Jin and Lu 2023) started to explore the use of CoT in LLMs to enhance performance in complex tasks. CoT prompting is an innovative gradient-free technique that guides LLMs to produce intermediate reasoning steps, ultimately leading to the derivation of the final answer. Specifically, Fei et al. (2023) introduce CoT into LLMs for implicit sentiment analysis. Trivedi et al. (2023) interleave retrieval with steps in a CoT to improve question-answering performance. More recently, Wang et al. (2023a) propose plan-and-solve prompting strategies to solve zero-shot CoT pitfalls. Despite these recent advancements, LLMs with CoT have not been explored in dialogue commonsense inference.

Methodology

Task Definition

The task of *Dialogue Commonsense Multi-Choice Question Answering* (DC-MCQ) is defined as: given a dialogue $D = \{u_1, ..., u_n\}$, the target utterance u_t, for the target utterance commonsense question Q and the candidate options $F_t = \{f_{t_1}, ..., f_{t_m}\}$, a model selects all correct options y in F_t.

Preliminary

Standard Prompting   Standard prompting methods have been widely used in previous works (Ma et al. 2023; Paranjape et al. 2021). Through crafting specific prompts, LLMs can be fine-tuned to handle diverse tasks by simply changing the prompts In this task, we construct the following prompt template as inputs for LLMs:

```
Given the context T, which options
are correct?
```

where $T = [D; u_t; Q; F_t]$, which includes dialogue, target utterance, question, and candidate options.

However, the prompting method has certain limitations. Firstly, it fails to account for option relationships, potentially resulting in erroneous predictions. Secondly, the lack of explicit guidance for the LLMs to engage in a step-by-step reasoning process diminishes the interpretability of their answers. As a result, comprehending the underlying logic behind a LLM's response becomes challenging.

Vanilla CoT Prompting   To enhance the standard prompting method, chain-of-thought (CoT) prompting has been investigated, which advances in not only producing the answer, but eliciting LLMs to give the reasoning/rationale behind the answer. For this task, we construct the following prompt template as inputs to LLMs:

```
Given the context T, let's think
step-by-step, which options are
correct and why?
```

Nevertheless, the vanilla CoT merely directly prompts the model to generate intermediate inference processes and final results. While the existing CoT methods demonstrate some inference capabilities, they are limited to performing linear inference and fail to multidimensionally utilize multiple clues to reason about multiple options. In addition, the vanilla CoT methods all direct the model to infer directly towards the answer in a forward manner, which easily overlook some of the correct options in cases with multiple valid answers, resulting in a performance decrease. This approach does not align with the way humans typically approach multi-choice questions, which involves a combination of exclusion and forward reasoning.

ReX-GoT Prompting

Neither the aforementioned standard prompting nor the vanilla CoT approach can solve the option saturation and clue labyrinth challenges in DC-MCQ, so we propose a new approach called ReX-GoT, which stands for Reverse Exclusion with Graph-of-Thought. Our method leverages valuable information to guide the model to integrate clues for step-by-step reasoning in a reverse exclusion manner and in conjunction with a well designed GoT. By doing so, our method effectively excludes incorrect options, narrows down the answer range, clarifies intricate clues, and improves the efficiency and accuracy of problem-solving. Moreover, our method considers the logical relationships between the options and the context, which contrasts with existing methods that solely rely on the contextual semantic information.

As depicted in Figure 2, our ReX-GoT method consists of three steps. In the first step, the model makes an initial judgment based on the context information to exclude unreasonable options and provide the reasons for the exclusion. In the second step, the model conducts a detailed analysis of each option, taking into account the contextual information and the excluded options and their corresponding reasons. In the final step, the model synthesizes the valuable information from the first two steps for integrated reasoning and selects the optimal path for the GoT to determine the final multi-choice answer. The specific steps are as follows.

Step I. Option Exclusion.   In this step, our approach involves an initial exclusion process to effectively narrow down the range of potential answers. Subsequently, we provide the model with crucial information regarding the reasons behind the exclusion of certain options, corresponding to the "Step1" and *purple arrows* in Figure 2. This information serves as valuable contextual input that aids the subsequent reasoning process. Furthermore, our approach goes beyond mere exclusion by providing the model with explicit justifications for why specific options are deemed incorrect. By incorporating these detailed explanations into the reasoning process, we equip the model with a more comprehensive understanding of the context and enable it to engage in more informed and accurate reasoning. Specifically, we devise the following template to consider which options are implausible and their reasons based on the given context.
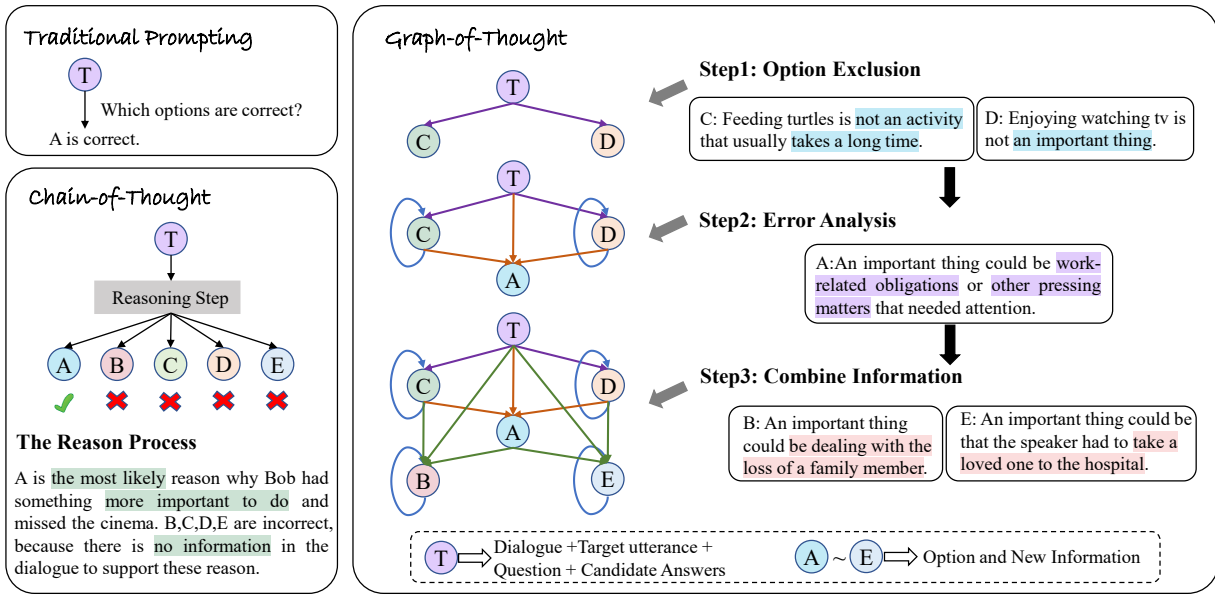
Figure 2: The overview of the prompt-based, CoT-based, and our ReX-GoT method. In our method, the purple arrows represent the first option exclusion step, *i.e.*, leveraging the reverse exclusion method to effectively solve the option saturation challenge. The orange and green arrows represent the second error analysis step and the third combine information step, *i.e.*, integrating information according to the GoT we design and choosing the optimal path to solve the clue labyrinth challenge. And the blue arrows represent the updating of information at each step. The highlighted text indicates the available information.

```
Given the context T, based on
common sense, which options of u_t
are unreasonable and why?
```

This step can be formulated as:

$$A_1 \leftarrow \arg\max_{\hat{\theta}} p(A_1 \mid T) \qquad (1)$$

where A_1 is the text that explicitly mentions the incorrect options and their reasons, $\hat{\theta}$ means the fixed parameter of the model, as there are no gold labels in the intermediate step. This step crucially refines the model's problem understanding, guiding subsequent reasoning by highlighting pitfalls. By enabling the model to recognize and comprehend excluded option reasons, it gains the ability for informed and reliable conclusions.

Step II. Error Analysis. In this step, we construct a graph-of-thought to perform error analysis and comparative analysis between options based on the known information to further aid model reasoning. Specifically, we first create a central node that represents the main stem of the problem. Then, we create nodes for each answer option and their reasoning process. For each option, we analyze the provided information and determine if it matches the main stem of the problem. If it does, we mark it as a possible correct option. If not, we mark it as a possible incorrect option. Next, we create a set of branch nodes for the possible correct options and analyze each branch node in more detail. We compare the information provided in each option with the existing information and exclude any mismatched options. Finally, we arrive at the correct answer by excluding the possible incorrect

options and confirming that the remaining options match the provided information.

And we ask the LLM to provide detailed answers to whether each option is correct and the specific reasons, taking into account the contextual information and the unreasonable options and their corresponding reasons (just like the *orange arrows* and "Step2" in Figure 2).

```
Given T_1 = [T, A_1], analysis based on
the incorrect options, if the answer
is f_{t_i}, is it reasonable and why?
```

This step can be formulated as:

$$A_2 \leftarrow \arg\max_{\hat{\theta}} p(A_2 \mid T_1, f_{t_i}) \qquad (2)$$

where A_2 is the text and answer regarding whether each option is reasonable, $\hat{\theta}$ refers to the fixed parameters. Through this step, we provide the model with the hidden information covered by the previous options and integrate the clues through the GoT for step-by-step reasoning, allowing for clearer information to address the clue labyrinth challenge.

Step III. Combine Information. In this step, as the "Step3" and *green arrows* in Figure 2 show, we leverage the valuable insights gathered from the preceding two steps and employ the GoT to further advance our reasoning process. Specifically, in inference steps I and II, we set the LLM decoder to generate multiple answers as different paths through the GoT, each of which give a different prediction for each option. The final multi-choice answer is determined by selecting the optimal path through a voting mechanism. With the aid of GoT, we delve into the intricate nuances of the more

	Models	CICERO		CICERO$_{v2}$		CICERO-Multi	
		F1	EM	F1	EM	F1	EM
• *SoTA baselines*	T5+CCID(780M)	81.96	75.89	86.37	70.14	68.03	19.95
	T5+MCCI(780M)	82.64	76.11	87.22	71.09	68.89	26.84
	T5+TEAM(780M)	82.73	76.28	87.31	71.23	69.16	27.07
	Flan-T5+TEAM(3B)	84.46	76.68	89.54	72.91	70.82	37.24
	Flan-T5+TEAM(11B)	86.62	77.89	91.53	75.97	73.51	45.78
• *Prompt-based methods*	Flan-T5+Prompt (780M)	82.18	75.74	86.45	70.21	68.78	26.71
	Flan-T5+Prompt (3B)	84.34	76.61	89.23	72.62	70.71	36.12
	Flan-T5+Prompt (11B)	86.43	77.46	91.41	75.86	73.45	45.29
• *CoT-based methods*	Flan-T5+GoT (780M)	84.47	76.72	87.54	71.37	69.83	32.94
	Flan-T5+CoT (3B)	86.52	77.66	90.63	74.76	72.59	40.87
	Flan-T5+CoT (11B)	87.98	78.53	92.32	77.14	75.69	47.26
• *ReX-GoT (Ours)*	Flan-T5+ReX-GoT (780M)	85.21	76.87	89.56	73.26	71.44	37.63
	Flan-T5+ReX-GoT (3B)	87.45	78.44	91.58	76.15	74.58	45.81
	Flan-T5+ReX-GoT (11B)	**89.52**	**80.63**	**93.87**	**78.46**	**78.51**	**53.08**

Table 1: Comparison of our method with baselines on CICERO, CICERO$_{v2}$ and CICERO-Multi datasets. CICERO-Multi datasets is a subdataset within the CICERO dataset that only contained questions with multiple correct answer choices.

complex and challenging options, persisting until a comprehensive evaluation of all options are achieved. This diligent examination ultimately culminates in the determination of the final multi-choice answer $\hat{y}$.

```
Given T₂ = [T₁, A₂], analysis based on
the previous steps, which options of
uₜ are reasonable?
```

This step can be formulated as:

$$\hat{y} \leftarrow \arg\max_{\theta} p(y \mid T_2) \qquad (3)$$

where θ can be fine-tuned during training with the gold task annotations.

Each step considers new information and validates the previous reasoning to arrive at the correct answer. By utilizing GoT, we visually represent the reasoning process and enable our method to integrate sophisticated reasoning clues. In conclusion, our ReX-GoT takes into account the subtle details and intricate dependencies in the given context and options. It combines multiple predictive information and guides the model to reason step-by-step through the GoT in a reverse exclusion manner. This approach effectively addresses the challenges of option saturation and clue labyrinth in DC-MCQ.

Experiments

Implementation Details

Datasets. We assess the efficacy of models on two benchmark datasets, CICERO (Ghosal et al. 2022b) and CICERO$_{v2}$ (Shen et al. 2022). CICERO is a binary dialogue dataset featuring five types of dialogue-level inferences: causality, consequence, premise, motivation, and emotional reaction. The dataset comprises 53,105 inferences from 5,672 dialogues. CICERO$_{v2}$ is built upon the original CICERO dataset, where only 15% of the inferences in CICERO are multi-choice, whereas all 8,351 inferences in CICERO$_{v2}$ are multi-choice. The dataset consists of 2,379 dialogues.

Evaluation Metrics. We use macro-F1 and Exact Match as evaluation metrics for our models. Macro-F1 considers precision and recall across multiple classes and provides an average score. Exact Match measures the percentage of correct predictions that exactly match the expected answers. All our scores are the average over 5 runnings with random seeds.

Settings. Due to the outstanding performance of Flan-T5, a encoder-decoder style language model, we utilize it as the backbone LLM for our method. We also test with GPT3.5. We use four versions of Flan-T5: 250M (base), 780M (large), 3B (xl), and 11B (11B). Our experiments are conducted using NVIDIA A100 GPUs.

Baseline Systems. We compare our method with the state-of-the-art (SoTA) baselines, including

- **CCID:** Ghosal et al. (2022b) computed a match by comparing each generated answer to a candidate selection.

- **MCCI:** Shen et al. (2022) proposed a pre-trained transformer DIALeCT for dialogue commonsense inference.

- **TEAM:** Ghosal et al. (2022a) simply refactored the multi-choice question answering task into a series of binary classifications.

Overall Results

We first comprehensively evaluate our ReX-GoT's superiority in dialogue commonsense inference using F1 and EM metrics. We compare against SoTA baselines (CCID, MCCI, TEAM), Prompt-based, and CoT-based methods across CICERO, CICERO$_{v2}$, and CICERO-Multi datasets. Table 1 highlights ReX-GoT's advantage over SoTA baselines. Flan-T5-large exhibits notable improvement with ReX-GoT. Further, with an 11B-parameter LLM, ReX-GoT outperforms the best baseline TEAM, *e.g.*, on CICERO, by 2.9% in F1 score and 2.74% in EM score. On CICERO$_{v2}$, ReX-GoT surpasses the SoTA baseline TEAM by 2.34% in F1 score and 2.49% in EM score. Moreover, our ReX-GoT exhibits a remarkable enhancement compared to vanilla prompting and CoT methods, particularly on the CICERO dataset with multiple correct answer options, where the EM scores of our

	CICERO		CICERO$_{v2}$	
	F1	EM	F1	EM
● *SoTA baselines*				
Flan-T5+TEAM(3B)	47.95	42.11	48.79	40.64
Flan-T5+TEAM(11B)	51.21	45.53	51.68	42.75
● *Prompt-based methods*				
Flan-T5+Prompt (3B)	48.34	42.53	49.28	41.67
Flan-T5+Prompt (11B)	51.67	45.82	52.34	43.29
● *CoT-based methods*				
Flan-T5+CoT (3B)	54.48	47.64	55.69	44.72
Flan-T5+CoT (11B)	58.83	49.77	60.22	47.26
● *ReX-GoT (Ours)*				
Flan-T5+ReX-GoT (3B)	63.59	52.84	64.38	49.64
Flan-T5+ReX-GoT (11B)	67.73	55.39	69.35	53.33
GPT3.5+ReX-GoT	86.04	77.17	91.12	75.73

Table 2: Experimental results on zero-shot setting.

model improve by 5.82% and 7.79%, respectively. These findings suggest that our ReX-GoT can make use of the hidden information between options to enhance reasoning and make answers more explanatory than vanilla prompting and CoT methods. Notably, our ReX-GoT effectively addresses the option saturation and clue labyrinth challenges in dialogue commonsense inference.

Results on Zero-shot Inference

We conduct a comprehensive comparison of our proposed ReX-GoT with SoTA approaches, Prompt-based, and CoT-based methods under zero-shot conditions. The results in Table 2 demonstrate our method's supremacy across all metrics. Prompt-based and CoT-based techniques exhibit substantial improvements over the current SoTA baseline. However, our ReX-GoT approach stands out with even more substantial advancements in dialogue commonsense inference. As an example, on the CICERO$_{v2}$ dataset, when using Flan-T5-11B, our ReX-GoT approach demonstrates a remarkable improvement in F1 score of 17.67% over the best-performing baseline TEAM. Our ReX-GoT approach outperforms the prompt-based approach by a margin of 17.01% in F1 score and the CoT-based approach by a margin of 9.13% in F1 score. Remarkably, when integrated into an ultra-large LLM like GPT3.5-175B, ReX-GoT achieves remarkable improvements, enhancing the SoTA's F1 score by 34.83% on CICERO and 39.44% on CICERO$_{v2}$. These results highlight the effectiveness of our ReX-GoT approach in improving the performance of large language models on dialogue commonsense inference.

Ablation Study

We conduct ablation experiments to evaluate the contribution of each component in our model. As depicted in Table 3, no variant matches the full model's performance, highlighting the indispensability of each component. Specifically, the F1 score drops most severely when the graph-of-thought are not used, which suggests that guiding the model to reason

	CICERO		CICERO-Multi	
	F1	EM	F1	EM
ReX-GoT	89.52	80.63	78.51	53.08
w/o CI	88.39(-1.13)	79.41(-1.22)	77.71(-0.8)	51.66(-1.42)
w/o ReX	88.02(-1.50)	78.97(-1.66)	77.54(-0.97)	51.39(-1.69)
w/o GoT	87.25(-2.27)	78.17(-2.46)	76.33(-2.18)	49.43(-3.65)

Table 3: Ablation results on DC-MCQ task. CI means the combined information step, ReX means the reverse exclusion step. In the brackets are the drops than ReX-GoT.
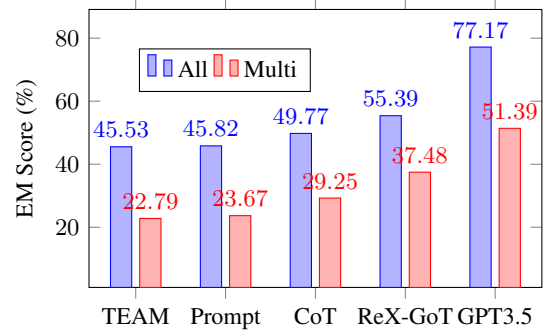

Figure 3: Comparison with different models on dialogue commonsense inference. "All" and "Multi" mean that the results are calculated on the complete CICERO dataset and a subset of CICERO containing only multiple correct options.

step-by-step and considering hidden information among options is crucial. To verify the necessity and effectiveness of exclusion, we remove the exclusion step, and the sharp drop in the results demonstrates its unignorable effect on dialogue commonsense inference. This finding suggests that combining exclusion with forward reasoning is essential for improving the performance of our model. In addition, removing combine information step leads to a marked drop in performance, indicating the importance of combining intricate clues in our ReX-GoT method.

Analyses and Discussions

To further investigate the effectiveness of ReX-GoT, we conduct in-depth analyses to answer the following questions, with the aim to reveal how our proposed methods advance.

How multi-choice inference affect model performance? We are curious about the impact of multi-choice inference on model performance in an unsupervised setting. In Figure 3, we compare our model with the best baseline on the Multi and All datasets, as well as models based on vanilla prompting and CoT. Our findings show that our model consistently outperforms these models in dialogue commonsense inference, regardless of whether the single-choice or multi-choice questions. Furthermore, the performance gaps are further enlarged when considering multi-choice inference, indicating the effectiveness of our method for this task. Overall, our results highlight the potential of our method for multi-choice inference in unsupervised conditions. By effec-
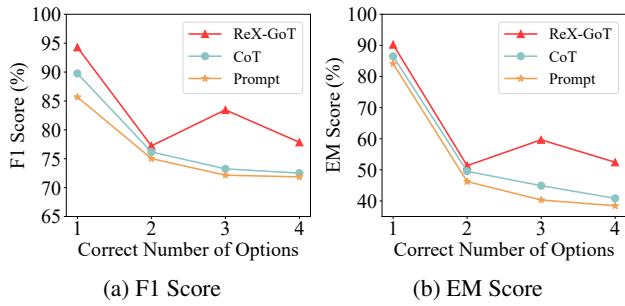
(a) F1 Score

(b) EM Score

Figure 4: Influence of correct number of options.

(a) F1 Score

(b) EM Score

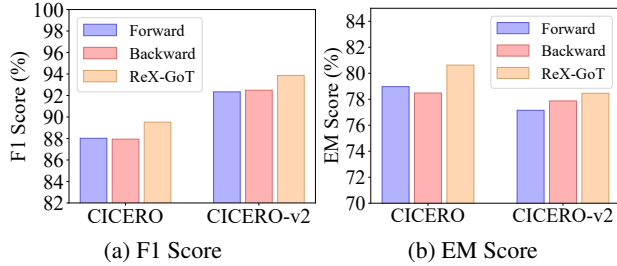Figure 5: Influence of different prompting.

(a) F1 Score

(b) EM Score

Figure 6: Influences of LLM scales.

tively integrating available clues from answer options, our approach surpass existing baseline, even in challenging scenarios with multiple correct answer options.

How the number of correct options affect model performance? We investigate the effect of the correct number of options on our model's performance in dialogue commonsense inference. As shown in Figure 4, we observe that the model's performance varies with the number of correct options. Our ReX-GoT method performs worst on questions with two correct options, followed by questions with four, three, and performs best on questions with one correct option. On the other hand, vanilla prompting and CoT methods show a decline in performance as the number of correct options increases. ReX-GoT effectively utilizes option information, capturing the relationship between options and context to differentiate between correct and incorrect options. This advantage is particularly prominent in questions with multiple correct options, where option information plays a crucial role. In contrast, vanilla methods rely only on context, neglects the integration of hidden clues and underutilizes the additional information in the options.

What are the advantages of ReX-GoT over forward reasoning and backward exclusion? We conduct experiments to compare our ReX-GoT approach with forward reasoning and backward exclusion. The results in Figure 5 show that ReX-GoT outperforms the two single methods on both datasets. Forward reasoning involves selecting the most plausible option at each step until no correct option is left. Backward exclusion, on the other hand, involves selecting the most incorrect option at each step until no incorrect options remain. Interestingly, the performance of the individual methods reverses between the datasets. In the
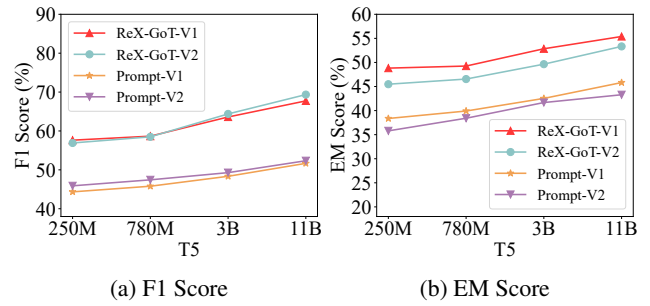
CICERO dataset, forward reasoning is superior, while in the CICEROv_2 dataset, backward exclusion performs better. This reversal is attributed to the majority of single-choice questions in the CICERO dataset, where inadequate exclusion during backward exclusion leads to decreased performance. Conversely, the CICEROv_2 dataset consists exclusively of multi-choice questions, making forward reasoning more challenging and resulting in poorer performance compared to backward exclusion and ReX-GoT. These findings further support the necessity of designing ReX-GoT for multi-choice task as it effectively combines the two single approaches and integrates valuable clues to address challenges and improve overall performance.

How LLMs scales affect model performance? We study the effect of different LLMs scales and provide the experimental results in Figure 6. We observe that both prompt-based and ReX-GoT methods show notable performance enhancement as the model size increases, particularly from Flan-T5-3B to Flan-T5-11B, where our ReX-GoT approach achieves an increase of 4.97% in F1 score and 3.69% in EM score on the CICERO$_{v2}$ dataset. Our results are consistent with existing findings about the effectiveness of CoT prompts, indicating that larger LLMs can bring remarkable improvements. This is because larger LLMs have stronger abilities to capture and model complex patterns and relationships in the data. Overall, our results emphasize the importance of considering LLMs size when designing models for dialogue commonsense inference. These findings demonstrate the potential of using LLMs for this task.

Conclusion

In this paper, we address the pressing option saturation and clue labyrinth challenges in the Dialogue Commonsense Multi-choice Question Answering task. We propose ReX-GoT, a novel three-step Reverse Exclusion Graph-of-Thought framework including Option Exclusion, Error Analysis, and Combine Information to mimic human reasoning. Through the gradual exclusion of irrelevant options and the incorporation of human-like reasoning, the final answer is obtained by constructing a GoT and selecting its optimal path. Our extensive experimental results on CICERO and CICERO$_{v2}$ datasets demonstrate that our scheme achieves SoTA performance on both single-choice and multi-choice dialogue commonsense inference.

Acknowledgements

References

Arabshahi, F.; Lee, J.; Gawarecki, M.; Mazaitis, K.; Azaria, A.; and Mitchell, T. M. 2021. Conversational Neuro-Symbolic Commonsense Reasoning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, 4902–4911.

Bosselut, A.; Bras, R. L.; and Choi, Y. 2021. Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-shot Commonsense Question Answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, 4923–4931.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS'20)*.

Chen, Q.; Xu, G.; Yan, M.; Zhang, J.; Huang, F.; Si, L.; and Zhang, Y. 2023. Distinguish Before Answer: Generating Contrastive Explanation as Knowledge for Commonsense Question Answering. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 13207–13224. Association for Computational Linguistics.

Deng, G.; Liu, Y.; Li, Y.; Wang, K.; Zhang, Y.; Li, Z.; Wang, H.; Zhang, T.; and Liu, Y. 2023. Jailbreaker: Automated Jailbreak Across Multiple Large Language Model Chatbots. *CoRR*, abs/2307.08715.

Dou, Z.; and Peng, N. 2022. Zero-Shot Commonsense Question Answering with Cloze Translation and Consistency Optimization. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22)*, 10572–10580. AAAI Press.

Fei, H.; Li, B.; Liu, Q.; Bing, L.; Li, F.; and Chua, T.-S. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. *arXiv preprint arXiv:2305.11255*.

Ghosal, D.; Hong, P.; Shen, S.; Majumder, N.; Mihalcea, R.; and Poria, S. 2021. CIDER: Commonsense Inference for Dialogue Explanation and Reasoning. In Li, H.; Levow, G.; Yu, Z.; Gupta, C.; Sisman, B.; Cai, S.; Vandyke, D.; Dethlefs, N.; Wu, Y.; and Li, J. J., eds., *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial'21)*, 301–313. Association for Computational Linguistics.

Ghosal, D.; Majumder, N.; Mihalcea, R.; and Poria, S. 2022a. Two is Better than Many Binary? Classification as an Effective Approach to Multi-Choice Question Answering. *arXiv preprint arXiv:2210.16495*.

Ghosal, D.; Shen, S.; Majumder, N.; Mihalcea, R.; and Poria, S. 2022b. CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, 5010–5028.

Imani, S.; Du, L.; and Shrivastava, H. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. In Sitaram, S.; Klebanov, B. B.; and Williams, J. D., eds., *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 37–42. Association for Computational Linguistics.

Jin, Z.; and Lu, W. 2023. Tab-CoT: Zero-shot Tabular Chain of Thought. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 10259–10277. Association for Computational Linguistics.

Ju, Y.; Zhang, Y.; Tian, Z.; Liu, K.; Cao, X.; Zhao, W.; Li, J.; and Zhao, J. 2021. Enhancing multiple-choice machine reading comprehension by punishing illogical interpretations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, 3641–3652.

Kuo, H.; and Chen, Y. 2023. Zero-Shot Prompting for Implicit Intent Prediction and Recommendation with Commonsense Reasoning. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 249–258. Association for Computational Linguistics.

Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Le Bras, R.; Choi, Y.; and Hajishirzi, H. 2022. Generated Knowledge Prompting for Commonsense Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, 3154–3169.

Ma, K.; Ilievski, F.; Francis, J.; Bisk, Y.; Nyberg, E.; and Oltramari, A. 2021. Knowledge-driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, 13507–13515.

Ma, Z.; Yu, Z.; Li, J.; and Li, G. 2023. HybridPrompt: Bridging Language Models and Human Priors in Prompt Tuning for Visual Question Answering. In Williams, B.; Chen, Y.; and Neville, J., eds., *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI'23)*, 13371–13379.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS'22)*, 27730–27744.

Paranjape, B.; Michael, J.; Ghazvininejad, M.; Hajishirzi, H.; and Zettlemoyer, L. 2021. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL-IJCNLP 2021 of *Findings of ACL*, 4179–4192. Association for Computational Linguistics.

Qin, L.; Gupta, A.; Upadhyay, S.; He, L.; Choi, Y.; and Faruqui, M. 2021. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 7066–7076. Association for Computational Linguistics.

Richardson, C.; and Heck, L. 2023. Commonsense Reasoning for Conversational AI: A Survey of the State of the Art. *CoRR*, abs/2302.07926.

Shen, S.; Ghosal, D.; Majumder, N.; Lim, H.; Mihalcea, R.; and Poria, S. 2022. Multiview contextual commonsense inference: A new dataset and task. *arXiv preprint arXiv:2210.02890*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 10014–10037. Association for Computational Linguistics.

Wang, L.; Xu, W.; Lan, Y.; Hu, Z.; Lan, Y.; Lee, R. K.; and Lim, E. 2023a. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 2609–2634. Association for Computational Linguistics.

Wang, P.; Wang, Z.; Li, Z.; Gao, Y.; Yin, B.; and Ren, X. 2023b. SCOTT: Self-Consistent Chain-of-Thought Distillation. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 5546–5558. Association for Computational Linguistics.

Wang, S.; Yu, M.; Jiang, J.; and Chang, S. 2018. A Co-Matching Model for Multi-choice Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, 746–751.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS'22)*.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2023. NExT-GPT: Any-to-Any Multimodal LLM. *CoRR*, abs/2309.05519.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *CoRR*, abs/2305.10601.

Zeng, H.; Wei, B.; Liu, J.; and Fu, W. 2023. Synthesize, Prompt and Transfer: Zero-shot Conversational Question Generation with Pre-trained Language Model. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 8989–9010. Association for Computational Linguistics.

Zhang, S.; Zhao, H.; Wu, Y.; Zhang, Z.; Zhou, X.; and Zhou, X. 2020. DCMN+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, 9563–9570.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhao, Z.; Hu, L.; Zhao, H.; Shao, Y.; and Wang, Y. 2023. Knowledgeable Parameter Efficient Tuning Network for Commonsense Question Answering. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 9051–9063. Association for Computational Linguistics.

Zheng, L.; Ji, D.; Li, F.; Fei, H.; Wu, S.; Li, J.; Li, B.; and Teng, C. 2023a. ECQED: Emotion-Cause Quadruple Extraction in Dialogs. *CoRR*, abs/2306.03969.

Zheng, L.; Li, F.; Chai, Y.; Teng, C.; and Ji, D. 2023b. A Bi-directional Multi-hop Inference Model for Joint Dialog Sentiment Classification and Act Recognition. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part I*, volume 14302, 235–248. Springer.