CLoMo: Counterfactual Logical Modification with Large Language Models

Anonymous ACL submission

Abstract

In this study, we delve into the realm of counterfactual reasoning capabilities of large language models (LLMs). Our primary objective is to cultivate the counterfactual thought processes within LLMs and rigorously assess these processes for their validity. Specifically, we introduce a novel task, **C**ounterfactual **Lo**gical **Mo**dification (CLoMo), and a high-quality human-annotated benchmark. In this task, LLMs must adeptly alter a given argumentative text to uphold a predetermined logical relationship. To effectively evaluate a generation model's counterfactual capabilities, we propose an innovative evaluation metric, the decomposed *Self-Evaluation Score (SES)* to directly evaluate the natural language output of LLMs instead of modeling the task as a multiple-choice problem. Analysis shows that the proposed automatic metric aligns well with human preference. Our experimental results show that while LLMs demonstrate a notable capacity for logical counterfactual thinking, there remains a discernible gap between their current abilities and human performance. We will release the data and code after publication.

1 Introduction

Despite large language models (Arkoudas, 2023; OpenAI, 2022) perform strikingly in plenty of reasoning benchmarks (Cobbe et al., 2021; Hendrycks et al., 2021a), late studies observe an internal inconsistency in their reasoning processes (Saparov and He, 2023; Arkoudas, 2023). The inconsistency is attributed to misunderstanding and misapplication of logical relations. However, logical relations in complex language reasoning are not yet properly quantified and evaluated.

Current studies on evaluating model reasoning are limited in both form and content. On the one hand, benchmarking complex reasoning is generally applying discrimination tasks such as multiple-choice questions (Huang et al., 2023b; Hendrycks et al., 2021a; Chen et al., 2023; Suzgun et al., 2023), where accuracy and pass rate serve as the main evaluation metric. However, such evaluations oversimplify the goal of uncovering essential and subtle pitfalls in complex reasoning. For example, the reasoning processes could contain misconceptions in logical relations but give correct answers due to the data distribution (Elazar et al., 2021; Saparov and He, 2023). Therefore, evaluating the generated content would provide a more realistic measurement of model reasoning. On the other hand, unlike widely studied reasoning tasks such as math reasoning (Cobbe et al., 2021; Hendrycks et al., 2021b) and standard exams (OpenAI, 2023; Huang et al., 2023b), counterfactual reasoning (Starr, 2022) as a fundamental evaluation of logical relations is less explored in the context of large language models. Previous literature studies counterfactual reasoning either in a multiple-choice manner (Tandon et al., 2019; Qin et al., 2021) or applying labored human study to evaluate counterfactual generation (Qin et al., 2019), leaving an effective evaluation of counterfactual generation unexplored.

In our study, we delve into the realm of evaluating large language models' (LLMs) ability to generate counterfactually coherent thoughts. Specifically, we proposed an innovative evaluation system that quantitatively measures the evolution of information in statement pairs, ensuring that they adhere to a specified logical relationship. Our approach includes designing a specialized task where models are presented with mismatched argument-premise pairs bound by a specific logical relation. The objective for these models is to adeptly modify the argument text until the specified logical relation is satisfactorily established. In conjunction with this task, we have created the first dataset of its kind, comprising dual argument-premise pairs, each annotated with a defined logical relation. This dataset is vital for facilitating logically restricted counterfactual modifications, and we have enriched it with

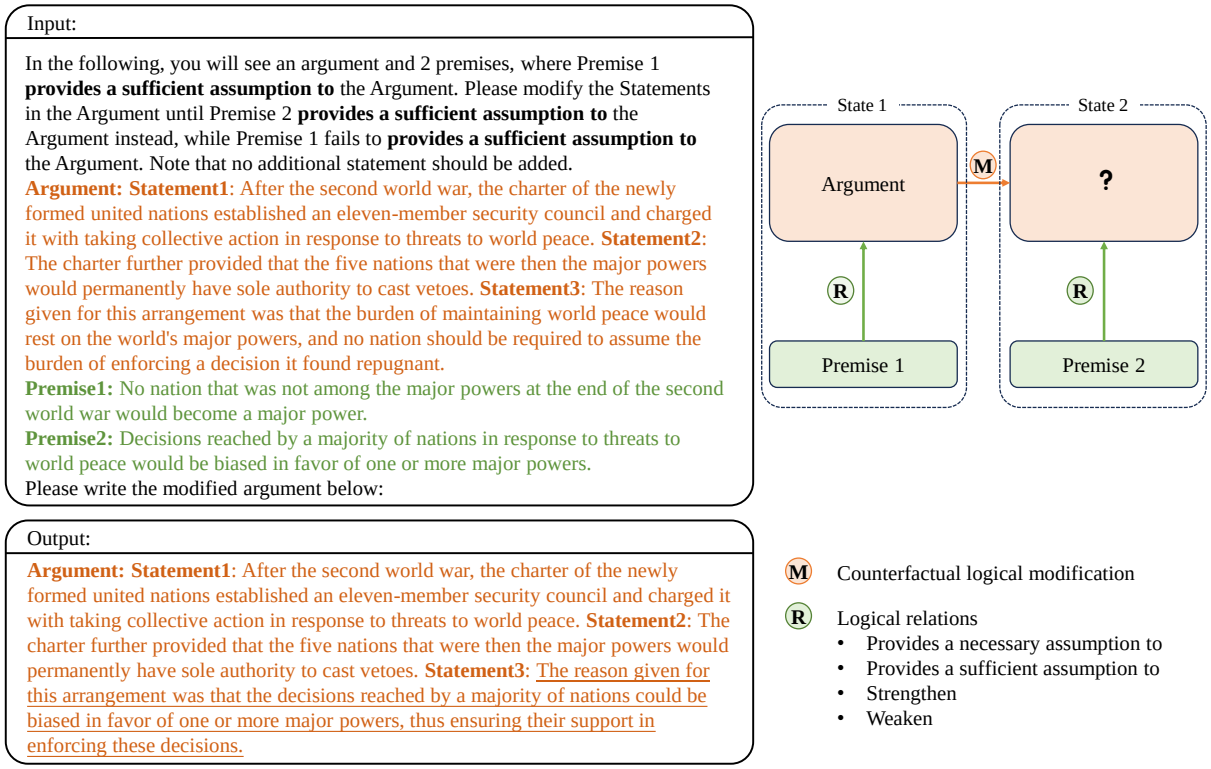

Figure 1: Demonstration of CLoMo. An LLM is given an argument and two premises. The LLM needs to modify the statements in Argument such that the logical relation R switch to stand in state 2 instead of state 1.

human-written modifications to serve as a benchmark for evaluation.

Our experimental investigations encompass a range of large language models, including GPT-4 (OpenAI, 2023) and GPT-3.5-Turbo (OpenAI, 2022), as well as smaller models from the LLaMA (Touvron et al., 2023a) and LLaMA 2 (Touvron et al., 2023b) families. Through these experiments, we have discerned that the task of CLoMo poses a significant challenge. It becomes evident that these models' current counterfactual logical reasoning capabilities fall short of the desired proficiency. This observation underscores the need for further advancements in enhancing the counterfactual reasoning abilities of existing language models, paving the way for more sophisticated and logically coherent AI systems.

The contributions of this paper are three-fold:

- We propose the task of *Counterfactual Logical Modification* and contribute a corresponding CLoMo. to evaluate the counterfactual reasoning capability of LLMs in the scenario of complicated textual logical reasoning.
- We propose the decomposed *Self-Evaluation Score (SES)* for the logically consistent generation of large language models.
- We conduct experiments on representative LLMs (GPT-3.5, GPT-4) and small language models (the LLaMA and LLaMA 2 families) and find that CLoMo is a very challenging task and the counterfactual logical reasoning ability of the existing model needs to be improved.

2 Related Works

From Complex Reasoning to Counterfactual Reasoning Complex reasoning has been highly concerned as a significant yet challenging task for inspecting advanced artificial intelligence. For example, for solving commonsense reasoning problems (Talmor et al., 2019, 2021; Huang et al., 2019; Bhagavatula et al., 2020; Sap et al., 2019), the models (Yasunaga et al., 2021, 2022; Liu et al., 2021; Huang et al., 2021b) are required to reasonable applying commonsense knowledge to conduct the reasoning process for the final answer. Furthermore, multi-step reasoning needs the models to perform multiple reasoning steps while maintaining consistency and faithfulness. To achieve this, synthetic compositional reasoning tasks (Betz, 2020; Tafjord et al., 2021; Han et al., 2022; Saparov and He, 2023) incorporate first-order logic to inspect

R	State 1	State 2
Necessary Assumption (38.5%)	**Argument:** Statement1: Journalist: the advice of social scientists is frequently overlooked by politicians making social policy. Statement 2: Because it is not unreasonable to discount scientific assertions backed by weak evidence, politicians should not generally be criticized for ignoring social science, for social scientists, unlike physical scientists, seldom agree on the claims made even within their own specialty. **Premise1:** The failure of scientists to agree that a claim within their specialty is true can indicate that the evidence for the claim is not strong.	**Argument′:** Statement 1: Journalist: the advice of social scientists is frequently overlooked by politicians making social policy. Statement 2: Because it is not unreasonable to discount scientific assertions, politicians should not generally be criticized for ignoring social science unless social scientists agree on the same claim, for social scientists, unlike physical scientists, seldom agree on the claims made even within their own specialty. **Premise2:** Politicians should follow the advice of experts on issues about which those experts agree among themselves.
Sufficient Assumption (6.6%)	**Argument:** Statement 1: Caffeine can kill or inhibit the growth of the larvae of several species of insects. Statement 2: One recent experiment showed that tobacco hornworm larvae die when they ingest a preparation that consists, in part, of finely powdered tea leaves, which contain caffeine. Statement 3: This result is evidence for the hypothesis that the presence of non-negligible quantities of caffeine in various parts of many diverse species of plants is not accidental but evolved as a defense for those plants. **Premise1:** Caffeine-producing plants or their ancestors have sometimes been fed upon by creatures sensitive to caffeine.	**Argument′:** Statement 1: Caffeine produced for plant species' own defense can kill or inhibit the growth of the larvae of several species of insects. Statement 2: One recent experiment showed that tobacco hornworm larvae die when they ingest a preparation that consists, in part, of finely powdered tea leaves, which contain caffeine. Statement 3: This result is evidence for the hypothesis that the presence of non-negligible quantities of caffeine in various parts of tobacco plant is not accidental but evolved as a defense for it. **Premise2:** The tobacco plant is among the plant species that produce caffeine for their own defense.
Strengthen (18.7%)	**Argument:** Statement1: In contemplating major purchases, businesses often consider only whether there is enough money left from monthly revenues after paying monthly expenses to cover the cost of the purchase. But many expenses do not occur monthly ; taking into account only monthly expenses can cause a business to overexpand. Statement2: So the use of a cash-flow statement is critical for all businesses. **Premise1:** A cash-flow statement is the only way to track both monthly expenses and expenses that are not monthly.	**Argument′:** Statement 1: In contemplating major purchases, businesses often consider only whether there is enough money left from monthly revenues after paying monthly expenses to cover the cost of the purchase. But there are many expenses every month ; taking into account these expenses incorrectly can cause a business to overexpand. Statement 2: So the use of a cash-flow statement is critical for all businesses. **Premise2:** Only a cash-flow statement can accurately document all monthly expenses.
Weaken (36.2%)	**Argument:** Statement1: The country of baurisia has, until now, been self-sufficient in both grain and meat. However, with growing prosperity in baurisia has come a steadily increasing per capita consumption of meat, and it takes several pounds of grain to produce one pound of meat. Statement2: Therefore, since per capita income in baurisia is almost certain to rise further but increases in domestic grain production are highly unlikely, baurisia is soon likely to become an importer of grain. **Premise1:** It is more economical for baurisians to import meat than grain.	**Argument′:** Statement 1: The country of baurisia has, until now, been self-sufficient in both grain and meat. However, with growing prosperity in baurisia has come a steadily increasing per capita consumption of meat. Statement 2: Therefore, since per capita income in baurisia is almost certain to rise further but increases in domestic meat production are highly unlikely, baurisia is soon likely to become an importer of meat. **Premise2:** The per capita consumption of meat in baurisia is roughly the same across all income levels.

Table 1: Example questions from the CLoMo benchmark, with the proportion of each logical relation.

and improve models logical consistency (Pan et al., 2023; Olausson et al., 2023; Sanyal et al., 2022; Yang and Deng, 2021). Moreover, real-scenario compositional reasoning (Yu et al., 2020; Liu et al., 2020; Dalvi et al., 2021; Huang et al., 2022) joins commonsense, consider the uncertainty of events in multi-step logical reasoning, which challenge current models (Bao et al., 2023; Xu et al., 2023; Jiao et al., 2023, 2022; Huang et al., 2023a, 2021a) to solve real-world reasoning problems with faithfulness. Additionally, the more faithful models should be able to consider counterfactuals. For example, answering questions given counterfactual conditions (Yu et al., 2023; Tandon et al., 2019; Qin et al., 2021), or narrating a counterfactual scenario (Qin et al., 2019). However, previous studies on counterfactual reasoning barely pay attention to the logical consistency or faithfulness of models. Therefore, in this work, we propose the task of *Counterfactual Logical Modification* that challenges the model to satisfy a given logical relation restriction while generating counterfactuals.

Evaluation of LLM Reasoning Reasoning poses a great challenge to LLMs. Currently, there is an increasing interest in the reasoning ability of LLMs. Evaluations are carried out from several perspectives, including mathematical reasoning, common-sense reasoning, logical reasoning, and domain-specific reasoning (Chang et al., 2023; Zhong et al., 2023; Bang et al., 2023; Liu et al., 2023). However, most current methods of reasoning evaluation focus primarily on the accuracy of the final answer and neglect a comprehensive assessment of the reasoning process of LLMs. Such evaluation is not ideal for understanding the reasoning ability of models, as it ignores situations where models may obtain correct answers through unfaithful or spurious reasoning shortcuts (Saparov and He, 2023). Some recent research has started to evaluate the validity of the intermediate reasoning steps of LLMs (Golovneva et al., 2022; Prasad et al., 2023). However, they mainly focus on the relationship between the intermediate step and the final answer rather than measuring whether the

model understands the intermediate reasoning process. This paper proposes a novel logical reasoning benchmark that requires intermediate counterfactual thinking under logical relation restrictions. This leads to a more in-depth study of the model's intermediate reasoning process.

3 CLoMo Benchmark

3.1 Task Definition

The desideratum is to harvest LLM counterfactual thinking and then investigate the validation of the thinking and its alignment with human counterfactual thinking. To achieve this, the LLM should generate its counterfactual thinking under proper logical scenarios. We design a task of counterfactual modification of argument text given a perturbation of premise given a static logical relation.

A demonstration of the proposed *Counterfactual Logical Modification* is shown in Figure 1. An LLM is given the instruction as shown on the left-hand side, which can be illustrated by the diagram on the right-hand side. In the given instruction, Argument and Premise1 are related by a logical relation. We consider four main relations in practice, which are (R1) the premise *provides a necessary assumption to* the argument, (R2) the premise *provides a sufficient assumption to* the argument, (R3) the premise *strengthen* the argument, and (R4) the premise *weaken* the argument. The Argument and Premise1then constitute State 1 of the logical relation R. The additional Premise2 perturbs the logical relation R. The goal for the LLM is to maintain a State 2 with the given Premise2 and a modified Argument′ that R stands. To this end, it should properly edit the statements in Argument until the goal is reached.

3.2 Benchmark Construction

Since applicable data for the proposed task is lacking, we build a benchmark dataset by carefully selecting argument texts and collecting human annotation of the modified Argument′. We choose to use ReClor (Yu et al., 2020) as the source data, considering that ReClor includes standardized multiple-choice questions on argument texts and logical relations from LSAT. We then recruit domain experts to conduct the annotation.

Given a ReClor question with context (the argument text), question, options, and the correct answer option, an annotator is required to provide a chosen wrong option (as the Premise2) and a corresponding modified context (i.e., the modified Argument′) to form a data point. The annotator is first instructed to read the whole question and comprehend the in-line logical relations, then choose one of the wrong options. After that, he/she edits the context by deleting, adding, or replacing text spans in the statements. The number of editions and the length of edited text spans are unrestricted as long as the statement partition is maintained.

We then post-process the question and the annotation so that for each data point, Argument comes from the original context, Premise1 comes from the correct option, and Premise2 and Argument′ come from the annotation.

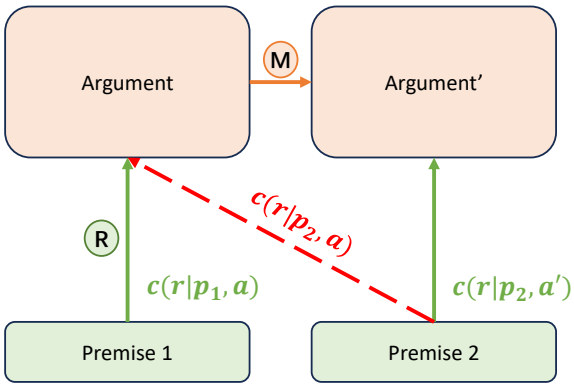

Figure 2: The concept graph of counterfactual logical modification.

4 SES: Self-Evaluation Scores

We aim to use the proven complex reasoning capabilities of the large language model itself to perform faster and more efficient reasoning evaluations of complex reasoning tasks that do not have easy access to standard/human-tested answers. Specifically, we split the scenario of complex logical reasoning evaluation into several discriminative tasks that LLMs have already seen and have been heavily trained on through logical conceptual graphs for the model to perform high-precision reasoning evaluation. We then collect the evaluations of these simple tasks and compute the ratings of the complex reasoning tasks based on the structure of the logical concept graph.

Counterfactual Modification Concept Graph Figure 2 To make the logic of counterfactual reasoning hold, we have a pair of primitive states p_1 (Premise1) and a (Argument) from p_1 and a satisfying the relation r, i.e., $\Pr(r|p_1, a)$ approaches 1. The other claim p_2 (Premise2) and the modified (a') Argument′ satisfy the same relation r, i.e.,

$c(r\|p_1, a)$	You are an expert in logic. <definition of relation>. In the following, you are given an Argument and a Premise. Is the Premise <relation> the Argument? Please think step by step, and then answer "yes" or "no". Argument: <a> Premise: <p_1>
$c(r\|p_2, a')$	You are an expert in logic. <definition of relation>. In the following, you are given an Argument and a Premise. Is the Premise <relation> the Argument? Please think step by step, and then answer "yes" or "no". Argument: <a'> Premise: <p_2>
$c(r\|p_2, a)$	You are an expert in logic. <definition of relation>. In the following, you are given an Argument and a Premise. Is the Premise <relation> the Argument? Please think step by step, and then answer "yes" or "no". Argument: <a> Premise: <p_2>

Table 2: Prompts for LLM decomposed evaluation task.

	SES	SES_{NA}	SES_{SA}	SES_{S}	SES_{W}
GPT-3.5-Turbo (OpenAI, 2022)	0.335	0.405	0.143	0.486	0.222
GPT-4 (OpenAI, 2023)	0.475	0.544	0.643	0.714	0.250
Human Performance	0.580	0.456	0.500	0.771	0.639

Table 3: Self-evaluation scores of LLM-generated counterfactual statements. NA: *Necessary Assumption.* SA: *Sufficient Assumption.* S: *Strengthen.* W: *Weaken.*

$\Pr(r|p_2, a')$ approaches 1. In contrast, the relation between p_2 and the Argument should not satisfy the relation r, i.e., $\Pr(r|p_2, a)$ approaches 0.

Decomposed Self-Evaluation Score As the demonstrated complex logical reasoning ability of current large language models, we let the large language model estimate $\Pr(r|p_1, a)$, $\Pr(r|p_2, a')$, and $\Pr(r|p_2, a)$, respectively. Specifically, we design a few binary classification tasks for the large language models, so that the probabilities are simplified to $c(r|p_1, a)$, $c(r|p_2, a')$, and $c(r|p_2, a) \in \{0, 1\}$. Finally, the overall logical modification score is computed by:

$$s = c(r|p_1, a) \times c(r|p_2, a') - c(r|p_2, a) \times c(r|p_2, a') \quad (1)$$

Alignment with Human Evaluation We randomly selected 50 samples from the test set and examined how well the self-evaluation score (SES) matches human evaluation. Specifically, we use GPT-4 to implement SES. We recruit experts in argumentation to evaluate the modified arguments generated by GPT-4 on the selected 50 samples, scoring 1 for good answers and 0 for bad answers. The Cohen's Kappa coefficient between the human annotators is $\kappa = 0.6785$, indicating substantial consistency of human perspective. We then use the self-evaluation score to evaluate the same group of modified arguments again. Cohen's kappa coefficient between human and self-evaluation score is $\kappa = 0.4391$. This indicates that the self-evaluation score is consistent with human evaluation. Therefore, we can use the self-evaluation score for more efficient logical modification evaluation. Moreover, since we can always check the logical reasoning ability of the model and its alignment with human evaluation, we can use a different model as the backbone to calculate this score.

5 Experiments

5.1 Main Results

We first evaluate the large language models GPT-3.5-Turbo (OpenAI, 2022) and GPT-4 (OpenAI, 2023). We also recruit 2 domain experts to contribute to the human performance. Table 3 demonstrates the results.

In the results, we see that the overall performance of the human experts on CLOMO is 0.542, indicating that CLOMO is quite a challenging task. In terms of the performance of the 4 logical relations, *strengthen* and *weaken* are relatively easy, while *necessary assumption* and *sufficient assumption* are more challenging for humans. In addition, the performance of GPT-4 is slightly lower than that of humans, but it also demonstrates strong counterfactual logical reasoning ability. And GPT-

Figure 3: Per-relation performances of LLaMA and LLaMA 2 families. NA: *Necessary Assumption.* SA: *Sufficient Assumption.* S: *Strengthen.* W: *Weaken.*

3.5-Turbo performs inferior to both human and GPT-4. Overall, large language models show great potential for counterfactual logical reasoning.

5.2 Fine-Tuning with Counterfactual Data

We then fine-tune LLaMA (Touvron et al., 2023a) and LLaMA 2 (Touvron et al., 2023b). We randomly split the CLoMo dataset into 60%/20%/20% training/development/test data, and fine-tune the models with the CLoMo training data. The implementation details are explained in Appendix C. We also compare the results with inference-only settings. The evaluation results are demonstrated in Figure 3.

In general, the pre-trained LLaMA and LLaMA 2 models achieve a certain level of counterfactual reasoning, and fine-tuning with counterfactual data further improves the performance. For example, LLaMA 2-13B with chain-of-thought prompting increases by 38.4% (SES: 0.430 → 0.595), with few-shot prompting increases by 98.3% (SES: 0.300 → 0.595), and with zero-shot prompting increases by 1,188.9% (SES: 0.045 → 0.580). The results indicate that such counterfactual data are barely seen in the LLaMA and LLaMA 2 pre-training data, and datasets such as the proposed CLoMo are much needed for developing models' counterfactual reasoning capabilities.

Moreover, among the four logical relations, *Weaken* performances are significantly increased after fine-tuning. However, the absolute SES scores of *Sufficient Assumption* problems after fine-tuning are still relatively low. It shows that *Sufficient Assumption* are challenging. Therefore, we still need a profound investigation of different logical relations to improve the models' counterfactual logical reasoning ability.

5.3 Performances of Small Language Models

We further test various language models on smaller scales and the results are shown in Table 4. All the models in Table 4 directly perform inference without further fine-tuning thus examining their original counterfactual abilities. The detailed settings are demonstrated in Appendix A, and the brief introductions of the language models are listed in Appendix B. We have the following findings: (1) Generally Speaking, all models perform inferiorly. Some of the models, for example, Baichuan2-7B-Chat in the few-shot setting, hardly solve any of the counterfactual questions. Among the models, Qwen-14B-Chat performs the best in both the chain-of-thought setting and the zero-shot setting. But all models perform inferior to large language models or human (2) The models perform better counterfactual reasoning with step-by-step reasoning processes (the CoT setting) while seeing more demonstrations in the prompt (the few-shot setting) harms the performances. It is indicated that the counterfactual cases have obscure reasoning patterns that are challenging for the models to transfer to unseen cases. (3) The performances are

	Model	Params	SES	SES_{NA}	SES_{SA}	SES_{S}	SES_{W}
CoT	Flan-T5-Large (Chung et al., 2022)	780M	0.28	0.30	0.21	0.46	0.18
	Flan-T5-XL (Chung et al., 2022)	3B	0.21	0.18	0.07	0.34	0.19
	Flan-T5-XXL (Chung et al., 2022)	11B	0.26	0.33	0.07	0.37	0.17
	ChatGLM2-6B (Du et al., 2022a)	6B	0.25	0.32	0.07	0.34	0.17
	Baichuan2-7B-Chat (Baichuan, 2023)	7B	0.24	0.20	0.07	0.49	0.18
	Baichuan2-13B-Chat (Baichuan, 2023)	13B	0.26	0.19	0.21	0.49	0.22
	InternLM-Chat-7B (Team, 2023)	7B	0.31	0.32	0.21	0.43	0.25
	Vicuna-7B-v1.5 (Chiang et al., 2023)	7B	0.15	0.13	0.00	0.29	0.14
	Vicuna-13B-v1.5 (Chiang et al., 2023)	13B	0.26	0.18	0.14	0.46	0.28
	Qwen-14B-Chat (Bai et al., 2023)	14B	0.30	0.24	0.21	0.46	0.29
	WizardLM-13B-v1.2 (Xu et al., 2024)	13B	0.02	0.04	0.00	0.00	0.00
Few	Flan-T5-Large (Chung et al., 2022)	780M	0.16	0.11	0.07	0.29	0.17
	Flan-T5-XL (Chung et al., 2022)	3B	0.16	0.09	0.00	0.29	0.19
	Flan-T5-XXL (Chung et al., 2022)	11B	0.21	0.14	0.07	0.40	0.21
	ChatGLM2-6B (Du et al., 2022a)	6B	0.01	0.01	0.00	0.00	0.00
	Baichuan2-7B-Chat (Baichuan, 2023)	7B	0.00	0.00	0.00	0.00	0.00
	Baichuan2-13B-Chat (Baichuan, 2023)	13B	0.02	0.03	0.00	0.06	0.00
	InternLM-Chat-7B (Team, 2023)	7B	0.01	0.00	0.00	0.00	0.01
	Vicuna-7B-v1.5 (Chiang et al., 2023)	7B	0.01	0.01	0.00	0.00	0.00
	Vicuna-13B-v1.5 (Chiang et al., 2023)	13B	0.01	0.01	0.00	0.00	0.00
	Qwen-14B-Chat (Bai et al., 2023)	14B	0.25	0.29	0.07	0.40	0.17
	WizardLM-13B-v1.2 (Xu et al., 2024)	13B	0.01	0.01	0.00	0.00	0.00
Zero	Flan-T5-Large (Chung et al., 2022)	780M	0.24	0.29	0.07	0.37	0.15
	Flan-T5-XL (Chung et al., 2022)	3B	0.25	0.22	0.07	0.43	0.22
	Flan-T5-XXL (Chung et al., 2022)	11B	0.28	0.29	0.21	0.43	0.21
	ChatGLM2-6B (Du et al., 2022a)	6B	0.15	0.10	0.00	0.34	0.13
	Baichuan2-7B-Chat (Baichuan, 2023)	7B	0.27	0.25	0.14	0.46	0.22
	Baichuan2-13B-Chat (Baichuan, 2023)	13B	0.29	0.25	0.21	0.46	0.25
	InternLM-Chat-7B (Team, 2023)	7B	0.28	0.28	0.14	0.43	0.22
	Vicuna-7B-v1.5 (Chiang et al., 2023)	7B	0.22	0.16	0.14	0.46	0.17
	Vicuna-13B-v1.5 (Chiang et al., 2023)	13B	0.25	0.18	0.21	0.40	0.26
	Qwen-14B-Chat (Bai et al., 2023)	14B	0.30	0.32	0.21	0.49	0.21
	WizardLM-13B-v1.2 (Xu et al., 2024)	13B	0.01	0.01	0.00	0.00	0.00

Table 4: Performances of smaller models in three inference-only settings. CoT: The chain-of-thought setting. Few: The few-shot setting. Zero: The zero-shot setting. More details are in Appendix A. NA: *Necessary Assumption.* SA: *Sufficient Assumption.* S: *Strengthen.* W: *Weaken.*

correlated to the model scales, but the gaps are not necessarily significant. For instance, Flan-T5-XXL (11B) performs generally better than Flan-T5-Large (780M), and Vicuna-13B-v1.5 performs better than Vicuna-7B-v1.5. (4) For the four different relations, it is shown that *Sufficient Assumption* and *Weaken* are significantly more challenging than the other two reasoning types. The possible reason is that compared to the other two reasoning types, *Sufficient Assumption* and *Weaken* require more reasoning steps such as reversed thinking. Two cases are further shown in Section 5.4. To sum up, the CLOMO task is challenging to current language models. Therefore, further investigation is needed on the counterfactual reasoning abilities of language models.

5.4 Case Study

We compare the modification by GPT-4 and humans. Figure 4 shows a case about *sufficient assumption*. That is, whether premise is a sufficient assumption for argument. The focus of the two premises here is to predict the impact of the charter provision. Switching from Premise 1 to Premise 2, the focus of discussion changes from the group of major powers to other nations in response to threats to world peace. This mainly affects the elaboration of Statement 3. GPT-4 has revised Statement 3 accordingly. Humans also made changes to Statement 3. Both revisions emphasized the influence of the five major powers in Statement 3, which corresponded to Premise 2. We find that GPT-4 can handle complex logical reasoning and counterfactual reasoning to a certain extent.

Figure 5 is about weakening an argument. The argument is on human intellectual development. Statement 2 provides evidence to support Statement 1. Premise 1 is on medical conditions and treatments providing counterexamples, and Premise 2 is on inaccuracies in research data. Human modifies Statement 2 to emphasize conversation, thus satisfying the logical conflict with the inaccuracies in research data described in Premise 2, thus

Input

In the following, you will see an argument and 2 premises, where Premise 1 **provides a sufficient assumption to** the Argument. Please modify the Statements in the Argument until Premise 2 **provides a sufficient assumption to** the Argument instead, while Premise 1 fails to **provides a sufficient assumption to** the Argument. Note that no additional statement should be added.
Argument: Statement1: After the second world war, the charter of the newly formed united nations established an eleven-member security council and charged it with taking collective action in response to threats to world peace. **Statement2**: The charter further provided that the five nations that were then the major powers would permanently have sole authority to cast vetoes. **Statement3**: The reason given for this arrangement was that the burden of maintaining world peace would rest on the world's major powers, and no nation should be required to assume the burden of enforcing a decision it found repugnant.
Premise1: No nation that was not among the major powers at the end of the second world war would become a major power.
Premise2: Decisions reached by a majority of nations in response to threats to world peace would be biased in favor of one or more major powers.
Please write the modified argument below:

GPT

Argument: **Statement1**: … … **Statement3:** The reason given for this arrangement was that the decisions reached by a majority of nations could be biased in favor of one or more major powers, thus ensuring their support in enforcing these decisions.

Human

Argument: **Statement1**: … … **Statement3:** The reason given for this arrangement was that the burden of maintaining world peace would rest on the five major powers, and and the decision made by security council won't be biased in favor of any major powers.

Figure 4: A successful case of counterfactual modification (reasoning type: *Sufficient Assumption*) generated by GPT-4, which makes a logically consistent Argument′ and is accordant with human reasoning.

Input

In the following, you will see an argument and 2 premises, where Premise 1 **weakens** the Argument. Please modify the Statements in the Argument until Premise 2 **weakens** the Argument instead, while Premise 1 fails to **weakens** the Argument. Note that no additional statement should be added.
Argument: **Statement1**: It is widely assumed that people need to engage in intellectual activities such as solving crossword puzzles or mathematics problems in order to maintain mental sharpness as they age. In fact, however, simply talking to other people -- that is, participating in social interaction, which engages many mental and perceptual skills -- suffices. **Statement2**: Evidence to this effect comes from a study showing that the more social contact people report, the better their mental skills.
Premise1: Many medical conditions and treatments that adversely affect a person's mental sharpness also tend to increase that person's social isolation.
Premise2: The tasks evaluating mental sharpness for which data were compiled by the study were more akin to mathematics problems than to conversation.
Please write the modified argument below:

GPT

Argument: **Statement1**: It is widely assumed that people need to engage in physical activities such as sports or gym in order to maintain mental sharpness as they age. In fact, however, simply talking to other people -- that is, participating in social interaction, which engages many mental and perceptual skills -- suffices. **Statement2**: … …

Human

Argument: **Statement1**: … … **Statement2**: Evidence to this effect comes from a study showing that the more social conversations people report, the better their mental skills.

Figure 5: An inferior case of counterfactual modification (reasoning type: *Weaken*) generated by GPT-4. It modifies Statement 1 by replacing intellectual activities with physical activities, where the logical restriction is not satisfied.

satisfying the *weaken* relation. GPT-4 modifies Statement 1 by replacing intellectual activities with physical activities such as sports or gym. Intellectual/physical activities have some counterfactual contrast, but in the context of the argument at hand, the logical relation restriction is not satisfied. We find that GPT-4 can still be flawed in counterfactual reasoning. In conclusion, complex counterfactual reasoning is challenging for large language models and needs improvements.

6 Conclusion

In this paper, we study large language models' counterfactual reasoning capability under the constraint of proper logical relations. To this end, we introduce a novel task, *Counterfactual Logical Modification*, that requires the LLMs to conduct counterfactual modification with logical restriction, where LLMs need to appropriately modify an argument text so that a specified logical relation stands. To ensure a comprehensive evaluation, we then construct a benchmark dataset CLoMo. Moreover, we propose a *Self-Evaluation Score (SES)* that decomposes the evaluation into several LLMs basic discrimination tasks, which is demonstrated aligned with human evaluations. We further evaluate smaller language models in inference-only and fine-tuning manners. The fine-tuned models' performances are significantly improved, but there is still a large gap with human performance. Our findings thus underscore the need for further development in LLMs' counterfactual reasoning.

7 Limitations

The major limitation of this work is that we do not include multiple solutions as reference sentences in this version of CLoMo. However, the proposed SES score leverages LLM to conduct a human-aligned evaluation, so it reduces the impact of reference in evaluating the model performances.

8 Ethical Considerations

The data and annotations are collected without personal or confidential information. Therefore, there is no ethical concern to the best of our knowledge.

References

Konstantine Arkoudas. 2023. GPT-4 can't reason. *CoRR*, abs/2308.03762.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Qiming Bao, Alex Yuxuan Peng, Zhenyun Deng, Wanjun Zhong, Gael Gendron, Timothy Pistotti, Neset Tan, Nathan Young, Yang Chen, Yonghua Zhu, Paul Denny, Michael Witbrock, and Jiamou Liu. 2023. Enhancing logical reasoning of large language models through logic-driven data augmentation.

Gregor Betz. 2020. Critical thinking for language models. *CoRR*, abs/2009.07185.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. *CoRR*, abs/2305.12524.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7358–7370. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022a. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022b. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in

the winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10486–10500. Association for Computational Linguistics.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021a. DAGN: discourse-aware graph network for logical reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5848–5855. Association for Computational Linguistics.

Yinya Huang, Meng Fang, Xunlin Zhan, Qingxing Cao, and Xiaodan Liang. 2021b. Rem-net: Recursive erasure memory network for commonsense evidence refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6375–6383.

Yinya Huang, Lemao Liu, Kun Xu, Meng Fang, Liang Lin, and Xiaodan Liang. 2023a. Discourse-aware graph networks for textual logical reasoning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11668–11688.

Yinya Huang, Hongming Zhang, Ruixin Hong, Xiaodan Liang, Changshui Zhang, and Dong Yu. 2022. Metalogic: Logical reasoning explanations with fine-grained structure. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4698–4724. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. Merit: Meta-path guided contrastive learning for logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3496–3509. Association for Computational Linguistics.

Fangkai Jiao, Zhiyang Teng, Shafiq R. Joty, Bosheng Ding, Aixin Sun, Zhengyuan Liu, and Nancy F. Chen. 2023. Logicllm: Exploring self-supervised logic-enhanced training for large language models. *CoRR*, abs/2305.13718.

Hanmeng Liu, Zhiyang Teng, Ruoxi Ning, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Glore: Evaluating logical reasoning of large language models. *CoRR*, abs/2310.09107.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6418–6425.

Theo X. Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *CoRR*, abs/2310.15164.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *CoRR*, abs/2305.12295.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *CoRR*, abs/2304.10703.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5042–5052. Association for Computational Linguistics.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIME-DIAL: temporal commonsense reasoning in dialog. In *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*, pages 7066–7076. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. Fairr: Faithful and robust deductive reasoning over natural language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1075–1093. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

W. Starr. 2022. Counterfactuals. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2022 edition. Metaphysics Research Lab, Stanford University.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3621–3634. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6075–6084. Association for Computational Linguistics.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Zihang Xu, Ziqing Yang, Yiming Cui, and Shijin Wang. 2023. IDOL: indicator-oriented logic pre-training for logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8099–8111. Association for Computational Linguistics.

Kaiyu Yang and Jia Deng. 2021. Learning symbolic rules for reasoning in quasi-natural language. *CoRR*, abs/2111.12038.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. *CoRR*, abs/2104.06378.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *CoRR*, abs/2305.14010.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *CoRR*, abs/2304.06364.

A Implementation Details of Three Inference-Only Settings

Tables 5 to 16 demonstrate examples of three inference-only samples with input and output. The three inference-only settings are:

- **Few-shot setting:** We first give four demonstrations in the prompt, and then provide an unseen question for the language model. The four demonstrations are randomly selected from the CLoMo training set, each of which is from one of the four reasoning relations.
- **Zero-shot setting:** We directly provide the question to be solved in the prompt.
- **Chain-of-thought setting:** We first provide the question, and then remind the language model to think step-by-step in the prompt.

Please see the Tables for the detailed demonstrations and prompts.

B Compared Models

We use the following language models in the experiments.

Flan-T5 (Chung et al., 2022) is a family of language models that are instruction-finetuned on the T5 models (Raffel et al., 2020). The models perform well on commonsense reasoning, mathematics, history, law, and medicine. In this paper, we use Flan-T5-Large with 780M parameters, Flan-T5-XL with 3B parameters, and Flan-T5-XXL with 11B parameters.

ChatGLM2 (Du et al., 2022a) is an open-source bilingual (Chinese-English) chat model based on GLM (Du et al., 2022b). The models are refined with data with a context length of up to 32K. The model has strong performance on multiple reasoning tasks.

Baichuan2 (Baichuan, 2023) is a family of language models trained on a high-quality corpus with 2.6 trillion tokens. The models achieve good performance on multiple authoritative Chinese, English, and multi-language general and domain-specific benchmarks. In the experiments, we use the Chat models Baichuan2-7B-Chat and Baichuan2-13B-Chat.

InternLM (Team, 2023) takes trillions of high-quality tokens for training to establish a powerful knowledge base. It has outstanding comprehensive performance. We use the Chat model InternLM-Chat-7B in our experiments.

Vicuna-v1.5 (Chiang et al., 2023) models are fine-tuned on the LLaMA 2 (Touvron et al., 2023b) models with supervised instruction fine-tuning with user-shared conversations collected from ShareGPT[1].

Qwen (Bai et al., 2023) is a series of comprehensive language model. The Qwen-Chat models are further fine-tuned with human alignment techniques such as Reinforcement Learning with Human Feedback (RLHF). The chat models have advanced tool use and planning capabilities.

WizardLM (Xu et al., 2024) is fine-tuned based on LLaMA (Touvron et al., 2023a) with a mixture of generated instruction data. The model shows its benefits in various skills such as philosophy, technology, and physics.

LLaMA (Touvron et al., 2023a) is a collection of foundation language models trained on trillions of tokens with publicly available datasets.

LLaMA 2 (Touvron et al., 2023b) is a family of pre-trained and fine-tuned language models that can be adapted for a variety of natural language generation tasks.

C Implementation Details of Fine-Tuning

We conduct full-parameter fine-tuning to LLaMA and LLaMA 2. The fine-tuning data is the CLoMo training set. Each model is fine-tuned by 10 epochs with a batch size of 4. The learning rate is $2e-5$ and is adapted by the cosine scheduler with a warmup proportion of 0.03. The best checkpoint is selected by the minimum perplexity in the validation split. In the inference phase, the models use beam search and a temperature of 0.7.

[1] https://sharegpt.com/

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **provides a necessary assumption to** the Argument. Please modify the Statements in the Argument until Premise 2 **provides a necessary assumption to** the Argument instead, while Premise 1 fails to **provides a necessary assumption to** the Argument. Note that no additional statement should be added.

Argument: Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, to be a moral agent one must have free will, because one can not be a moral agent without desiring to conform to a principle.
Premise1: Desiring to conform to a principle requires free will.
Premise2: It is impossible to have desires without also being a moral agent.

Please write the modified argument below:
Modified Argument:
Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, one can not have desires without desiring to conform to a principle, because one can not be a moral agent without desiring to conform to a principle.

Argument: Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only supervisor in the shipping department other than larson and franks.
Premise1: The task cannot be assigned to anyone other than a supervisor in the shipping department.
Premise2: The task cannot be assigned to anyone who has any kind of scheduling conflict.

Please write the modified argument below:
Modified Argument:
Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only one who do not have scheduling conflict in the shipping department other than larson and franks.

Argument: Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often does not enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will not boost sales and, therefore, will not increase profits either.
Premise1: The $3. 00 coupons, if distributed, would not induce new customers, who might not have purchased the product otherwise, to try wart-gone.
Premise2: Wart-gone can be produced in small or large batches without appreciably lowering the pre-tube production costs.

Please write the modified argument below:
Modified Argument:
Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often can enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will boost produce and, however, will not change pre-tube production costs much.

Argument: Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : 65 percent of its graduates went on to internships or jobs in print or broadcast journalism.
Premise1: More than half of the school's students came from jobs in journalism to improve their skills.
Premise2: Some newspaper editors do not regard journalism schools as a necessary part of the training of a journalist.

Please write the modified argument below:
Modified Argument:
Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : the journalism regards journalism schools as a necessary part of the training of a journalist.

Argument: Statement1: Consumer advocate : there is no doubt that the government is responsible for the increased cost of gasoline, because the government's policies have significantly increased consumer demand for fuel, and as a result of increasing demand, the price of gasoline has risen steadily.
Premise1: The government can bear responsibility for that which it indirectly causes.
Premise2: Consumer demand for gasoline cannot increase without causing gasoline prices to increase.

Please write the modified argument below:

Output:

Statement1: Consumer advocate : there is no doubt that the government partly leads to the increased cost of gasoline, because the government's policies have significantly increased consumer demand for fuel, and as a result of increasing demand, the price of gasoline has risen steadily undoubtedly.

Table 5: Few-shot setting inference-only sample with input and output. Logical relation: *Necessary Assumption*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **provides a sufficient assumption to** the Argument. Please modify the Statements in the Argument until Premise 2 **provides a sufficient assumption to** the Argument instead, while Premise 1 fails to **provides a sufficient assumption to** the Argument. Note that no additional statement should be added.

Argument: Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, to be a moral agent one must have free will, because one can not be a moral agent without desiring to conform to a principle.
Premise1: Desiring to conform to a principle requires free will.
Premise2: It is impossible to have desires without also being a moral agent.

Please write the modified argument below:
Modified Argument:
Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, one can not have desires without desiring to conform to a principle, because one can not be a moral agent without desiring to conform to a principle.

Argument: Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only supervisor in the shipping department other than larson and franks.
Premise1: The task cannot be assigned to anyone other than a supervisor in the shipping department.
Premise2: The task cannot be assigned to anyone who has any kind of scheduling conflict.

Please write the modified argument below:
Modified Argument:
Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only one who do not have scheduling conflict in the shipping department other than larson and franks.

Argument: Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often does not enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will not boost sales and, therefore, will not increase profits either.
Premise1: The $3. 00 coupons, if distributed, would not induce new customers, who might not have purchased the product otherwise, to try wart-gone.
Premise2: Wart-gone can be produced in small or large batches without appreciably lowering the pre-tube production costs.

Please write the modified argument below:
Modified Argument:
Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often can enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will boost produce and, however, will not change pre-tube production costs much.

Argument: Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : 65 percent of its graduates went on to internships or jobs in print or broadcast journalism.
Premise1: More than half of the school's students came from jobs in journalism to improve their skills.
Premise2: Some newspaper editors do not regard journalism schools as a necessary part of the training of a journalist.

Please write the modified argument below:
Modified Argument:
Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : the journalism regards journalism schools as a necessary part of the training of a journalist.

Argument: Statement1: Super express shipping company has implemented a new distribution system that can get almost every package to its destination the day after it is sent. The company worries that this more efficient system will result in lower sales of its premium next-day delivery service, because its two day service will usually arrive the following day anyway. **Statement2:** The company plans to encourage sales of its next-day service by intentionally delaying delivery of its two-day packages so that they will not be delivered the following day, even if the package arrives at its destination city in time for next-day delivery.
Premise1: Deliberate delay of packages will not affect the company's image in a way that significantly reduces its ability to attract and retain customers.
Premise2: The overnight service is too expensive to be attractive to most customers currently.

Please write the modified argument below:

Output:

Statement1: Super express shipping company has paused to implemente a new distribution system that can get almost every package to its destination the day after it is sent. The company worries that this more efficient system will result in lower sales of its premium next-day delivery service, because its two day service will usually arrive the following day anyway. **Statement2:** The company plans to encourage sales of its next-day service by stopping accelerating delivery of its two-day packages so that they will not be delivered the following day, leaving the package arriving at its destination city overnight for next-day delivery.

Table 6: Few-shot setting inference-only sample with input and output. Logical relation: *Sufficient Assumption*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **strengthens** the Argument. Please modify the Statements in the Argument until Premise 2 **strengthens** the Argument instead, while Premise 1 fails to **strengthens** the Argument. Note that no additional statement should be added.

Argument: Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, to be a moral agent one must have free will, because one can not be a moral agent without desiring to conform to a principle.
Premise1: Desiring to conform to a principle requires free will.
Premise2: It is impossible to have desires without also being a moral agent.

Please write the modified argument below:
Modified Argument:
Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, one can not have desires without desiring to conform to a principle, because one can not be a moral agent without desiring to conform to a principle.

Argument: Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only supervisor in the shipping department other than larson and franks.
Premise1: The task cannot be assigned to anyone other than a supervisor in the shipping department.
Premise2: The task cannot be assigned to anyone who has any kind of scheduling conflict.

Please write the modified argument below:
Modified Argument:
Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only one who do not have scheduling conflict in the shipping department other than larson and franks.

Argument: Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often does not enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will not boost sales and, therefore, will not increase profits either.
Premise1: The $3. 00 coupons, if distributed, would not induce new customers, who might not have purchased the product otherwise, to try wart-gone.
Premise2: Wart-gone can be produced in small or large batches without appreciably lowering the pre-tube production costs.

Please write the modified argument below:
Modified Argument:
Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often can enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will boost produce and, however, will not change pre-tube production costs much.

Argument: Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : 65 percent of its graduates went on to internships or jobs in print or broadcast journalism.
Premise1: More than half of the school's students came from jobs in journalism to improve their skills.
Premise2: Some newspaper editors do not regard journalism schools as a necessary part of the training of a journalist.

Please write the modified argument below:
Modified Argument:
Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : the journalism regards journalism schools as a necessary part of the training of a journalist.

Argument: Statement1: Often patients with ankle fractures that are stable, and thus do not require surgery, are given follow-up x-rays because their orthopedists are concerned about possibly having misjudged the stability of the fracture. **Statement2:** When a number of follow-up x-rays were reviewed, however, all the fractures that had initially been judged stable were found to have healed correctly. **Statement3:** Therefore, it is a waste of money to order follow-up x-rays of ankle fracture initially judged stable.
Premise1: X-rays of patients of many different orthopedists working in several hospitals were reviewed.
Premise2: Doctors who are general practitioners rather than orthopedists are less likely than orthopedists to judge the stability of an ankle fracture correctly.

Please write the modified argument below:

Output:

Statement1: Often patients with ankle fractures that are stable, and thus do not require surgery, therefore, many general doctors do not order an x-ray to patients with stable ankle fractures in order to avoid wasting medical resources. **Statement2:** When a number of follow-up x-rays were reviewed, however, most of the fractures that had initially been judged stable were found to haven't healed correctly. **Statement3:** Therefore, it is necessary to order follow-up x-rays of ankle fracture initially judged stable.

Table 7: Few-shot setting inference-only sample with input and output. Logical relation: *Strengthen*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **weakens** the Argument. Please modify the Statements in the Argument until Premise 2 **weakens** the Argument instead, while Premise 1 fails to **weakens** the Argument. Note that no additional statement should be added.

Argument: Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, to be a moral agent one must have free will, because one can not be a moral agent without desiring to conform to a principle.
Premise1: Desiring to conform to a principle requires free will.
Premise2: It is impossible to have desires without also being a moral agent.

Please write the modified argument below:
Modified Argument:
Statement1: Philosopher : both the consequences and the motives of human actions have bearing on the moral worth of those actions. **Statement2:** Nonetheless, one can not have desires without desiring to conform to a principle, because one can not be a moral agent without desiring to conform to a principle.

Argument: Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only supervisor in the shipping department other than larson and franks.
Premise1: The task cannot be assigned to anyone other than a supervisor in the shipping department.
Premise2: The task cannot be assigned to anyone who has any kind of scheduling conflict.

Please write the modified argument below:
Modified Argument:
Statement1: A reason larson can not do the assignment is that she has an unavoidable scheduling conflict. **Statement2:** On the other hand, a reason franks can not do the assignment is that he does not quite have the assertiveness the task requires. **Statement3:** So, the task must be assigned to parker, the only one who do not have scheduling conflict in the shipping department other than larson and franks.

Argument: Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often does not enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will not boost sales and, therefore, will not increase profits either.
Premise1: The $3. 00 coupons, if distributed, would not induce new customers, who might not have purchased the product otherwise, to try wart-gone.
Premise2: Wart-gone can be produced in small or large batches without appreciably lowering the pre-tube production costs.

Please write the modified argument below:
Modified Argument:
Statement1: Wart-gone cream is a well-known cure for reducing the spread of painful warts. A dime size amount rubbed on the infected area daily is the prescribed dosage, and applying the cream more often can enhance the cream's healing effects. **Statement3:** Hence, distributing coupons giving $ 3.00 off the price of wart-gone will boost produce and, however, will not change pre-tube production costs much.

Argument: Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : 65 percent of its graduates went on to internships or jobs in print or broadcast journalism.
Premise1: More than half of the school's students came from jobs in journalism to improve their skills.
Premise2: Some newspaper editors do not regard journalism schools as a necessary part of the training of a journalist.

Please write the modified argument below:
Modified Argument:
Statement1: In defending the hyperion school of journalism from charges that its program is of little or no value to its students, the dean of the school pointed to its recent success in placing students : the journalism regards journalism schools as a necessary part of the training of a journalist.

Argument: Statement1: To suit the needs of corporate clients, advertising agencies have successfully modified a strategy originally developed for political campaigns. **Statement2:** This strategy aims to provide clients with free publicity and air time by designing an advertising campaign that is controversial, thus drawing prime-time media coverage and evoking public comment by officials.
Premise1: The usefulness of an advertising campaign is based solely on the degree to which the campaign's advertisements persuade their audiences.
Premise2: Advertising agencies that specialize in campaigns for corporate clients are not usually chosen for political campaigns.

Please write the modified argument below:

Output:

Statement1: To suit the needs of corporate clients, advertising agencies have successfully modified a strategy originally developed for political campaigns. **Statement2:** This strategy aims to provide clients with free publicity and air time by designing an advertising campaign that is controversial, thus drawing prime-time media coverage and evoking public comment by officials. often such advertising agencies are chosen as partners for political campaigns.

Table 8: Few-shot setting inference-only sample with input and output. Logical relation: *Weaken*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **provides a necessary assumption to** the Argument. Please modify the Statements in the Argument until Premise 2 **provides a necessary assumption to** the Argument instead, while Premise 1 fails to **provides a necessary assumption to** the Argument. Note that no additional statement should be added.

Argument: Statement1: Consumer advocate : there is no doubt that the government is responsible for the increased cost of gasoline, because the government's policies have significantly increased consumer demand for fuel, and as a result of increasing demand, the price of gasoline has risen steadily.
Premise1: The government can bear responsibility for that which it indirectly causes.
Premise2: Consumer demand for gasoline cannot increase without causing gasoline prices to increase.

Please write the modified argument below:

Output:

Statement1: Consumer advocate : there is no doubt that the government partly leads to the increased cost of gasoline, because the government's policies have significantly increased consumer demand for fuel, and as a result of increasing demand, the price of gasoline has risen steadily undoubtedly.

Table 9: Zero-shot setting inference-only sample with input and output. Logical relation: *Necessary Assumption*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **provides a sufficient assumption to** the Argument. Please modify the Statements in the Argument until Premise 2 **provides a sufficient assumption to** the Argument instead, while Premise 1 fails to **provides a sufficient assumption to** the Argument. Note that no additional statement should be added.

Argument: Statement1: Super express shipping company has implemented a new distribution system that can get almost every package to its destination the day after it is sent. The company worries that this more efficient system will result in lower sales of its premium next-day delivery service, because its two day service will usually arrive the following day anyway. **Statement2:** The company plans to encourage sales of its next-day service by intentionally delaying delivery of its two-day packages so that they will not be delivered the following day, even if the package arrives at its destination city in time for next-day delivery.
Premise1: Deliberate delay of packages will not affect the company's image in a way that significantly reduces its ability to attract and retain customers.
Premise2: The overnight service is too expensive to be attractive to most customers currently.

Please write the modified argument below:

Output:

Statement1: Super express shipping company has paused to implemente a new distribution system that can get almost every package to its destination the day after it is sent. The company worries that this more efficient system will result in lower sales of its premium next-day delivery service, because its two day service will usually arrive the following day anyway. **Statement3:** The company plans to encourage sales of its next-day service by stopping accelerating delivery of its two-day packages so that they will not be delivered the following day, leaving the package arriving at its destination city overnight for next-day delivery.

Table 10: Zero-shot setting inference-only sample with input and output. Logical relation: *Sufficient Assumption*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **strengthens** the Argument. Please modify the Statements in the Argument until Premise 2 **strengthens** the Argument instead, while Premise 1 fails to **strengthens** the Argument. Note that no additional statement should be added.

Argument: Statement1: Often patients with ankle fractures that are stable, and thus do not require surgery, are given follow-up x-rays because their orthopedists are concerned about possibly having misjudged the stability of the fracture. **Statement2:** When a number of follow-up x-rays were reviewed, however, all the fractures that had initially been judged stable were found to have healed correctly. **Statement3:** Therefore, it is a waste of money to order follow-up x-rays of ankle fracture initially judged stable.
Premise1: X-rays of patients of many different orthopedists working in several hospitals were reviewed.
Premise2: Doctors who are general practitioners rather than orthopedists are less likely than orthopedists to judge the stability of an ankle fracture correctly.

Please write the modified argument below:

Output:

Statement1: Often patients with ankle fractures that are stable, and thus do not require surgery, therefore, many general doctors do not order an x-ray to patients with stable ankle fractures in order to avoid wasting medical resources. **Statement2:** When a number of follow-up x-rays were reviewed, however, most of the fractures that had initially been judged stable were found to haven't healed correctly. **Statement3:** Therefore, it is necessary to order follow-up x-rays of ankle fracture initially judged stable.

Table 11: Zero-shot setting inference-only sample with input and output. Logical relation: *Strengthen*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **weakens** the Argument. Please modify the Statements in the Argument until Premise 2 **weakens** the Argument instead, while Premise 1 fails to **weakens** the Argument. Note that no additional statement should be added.

Argument: Statement1: To suit the needs of corporate clients, advertising agencies have successfully modified a strategy originally developed for political campaigns. **Statement2:** This strategy aims to provide clients with free publicity and air time by designing an advertising campaign that is controversial, thus drawing prime-time media coverage and evoking public comment by officials.
Premise1: The usefulness of an advertising campaign is based solely on the degree to which the campaign's advertisements persuade their audiences.
Premise2: Advertising agencies that specialize in campaigns for corporate clients are not usually chosen for political campaigns.

Please write the modified argument below:

Output:

Statement1: To suit the needs of corporate clients, advertising agencies have successfully modified a strategy originally developed for political campaigns. **Statement2:** This strategy aims to provide clients with free publicity and air time by designing an advertising campaign that is controversial, thus drawing prime-time media coverage and evoking public comment by officials. often such advertising agencies are chosen as partners for political campaigns.

Table 12: Zero-shot setting inference-only sample with input and output. Logical relation: *Weaken*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **provides a necessary assumption to** the Argument. Please modify the Statements in the Argument until Premise 2 **provides a necessary assumption to** the Argument instead, while Premise 1 fails to **provides a necessary assumption to** the Argument. Note that no additional statement should be added.

Argument: Statement1: Consumer advocate : there is no doubt that the government is responsible for the increased cost of gasoline, because the government's policies have significantly increased consumer demand for fuel, and as a result of increasing demand, the price of gasoline has risen steadily.
Premise1: The government can bear responsibility for that which it indirectly causes.
Premise2: Consumer demand for gasoline cannot increase without causing gasoline prices to increase.

Please write the modified argument below. Let's think step by step:

Output:

Statement1: Consumer advocate : there is no doubt that the government partly leads to the increased cost of gasoline, because the government's policies have significantly increased consumer demand for fuel, and as a result of increasing demand, the price of gasoline has risen steadily undoubtedly.

Table 13: Chain-of-thought setting inference-only sample with input and output. Logical relation: *Necessary Assumption*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **provides a sufficient assumption to** the Argument. Please modify the Statements in the Argument until Premise 2 **provides a sufficient assumption to** the Argument instead, while Premise 1 fails to **provides a sufficient assumption to** the Argument. Note that no additional statement should be added.

Argument: Statement1: Super express shipping company has implemented a new distribution system that can get almost every package to its destination the day after it is sent. The company worries that this more efficient system will result in lower sales of its premium next-day delivery service, because its two day service will usually arrive the following day anyway. **Statement2:** The company plans to encourage sales of its next-day service by intentionally delaying delivery of its two-day packages so that they will not be delivered the following day, even if the package arrives at its destination city in time for next-day delivery.
Premise1: Deliberate delay of packages will not affect the company's image in a way that significantly reduces its ability to attract and retain customers.
Premise2: The overnight service is too expensive to be attractive to most customers currently.

Please write the modified argument below. Let's think step by step:

Output:

Statement1: Super express shipping company has paused to implemente a new distribution system that can get almost every package to its destination the day after it is sent. The company worries that this more efficient system will result in lower sales of its premium next-day delivery service, because its two day service will usually arrive the following day anyway. **Statement3:** The company plans to encourage sales of its next-day service by stopping accelerating delivery of its two-day packages so that they will not be delivered the following day, leaving the package arriving at its destination city overnight for next-day delivery.

Table 14: Chain-of-thought setting inference-only sample with input and output. Logical relation: *Sufficient Assumption*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 strengthens the Argument. Please modify the Statements in the Argument until Premise 2 **strengthens** the Argument instead, while Premise 1 fails to **strengthens** the Argument. Note that no additional statement should be added.

Argument: Statement1: Often patients with ankle fractures that are stable, and thus do not require surgery, are given follow-up x-rays because their orthopedists are concerned about possibly having misjudged the stability of the fracture. **Statement2:** When a number of follow-up x-rays were reviewed, however, all the fractures that had initially been judged stable were found to have healed correctly. **Statement3:** Therefore, it is a waste of money to order follow-up x-rays of ankle fracture initially judged stable.
Premise1: X-rays of patients of many different orthopedists working in several hospitals were reviewed.
Premise2: Doctors who are general practitioners rather than orthopedists are less likely than orthopedists to judge the stability of an ankle fracture correctly.

Please write the modified argument below. Let's think step by step:

Output:

Statement1: Often patients with ankle fractures that are stable, and thus do not require surgery, therefore, many general doctors do not order an x-ray to patients with stable ankle fractures in order to avoid wasting medical resources. **Statement2:** When a number of follow-up x-rays were reviewed, however, most of the fractures that had initially been judged stable were found to haven't healed correctly. **Statement3:** Therefore, it is necessary to order follow-up x-rays of ankle fracture initially judged stable.

Table 15: Chain-of-thought setting inference-only sample with input and output. Logical relation: *Strengthen*.

Input:

In the following, you will see an argument and 2 premises, where Premise 1 **weakens** the Argument. Please modify the Statements in the Argument until Premise 2 **weakens** the Argument instead, while Premise 1 fails to **weakens** the Argument. Note that no additional statement should be added.

Argument: Statement1: To suit the needs of corporate clients, advertising agencies have successfully modified a strategy originally developed for political campaigns. **Statement2:** This strategy aims to provide clients with free publicity and air time by designing an advertising campaign that is controversial, thus drawing prime-time media coverage and evoking public comment by officials.
Premise1: The usefulness of an advertising campaign is based solely on the degree to which the campaign's advertisements persuade their audiences.
Premise2: Advertising agencies that specialize in campaigns for corporate clients are not usually chosen for political campaigns.

Please write the modified argument below. Let's think step by step:

Output:

Statement1: To suit the needs of corporate clients, advertising agencies have successfully modified a strategy originally developed for political campaigns. **Statement2:** This strategy aims to provide clients with free publicity and air time by designing an advertising campaign that is controversial, thus drawing prime-time media coverage and evoking public comment by officials .often such advertising agencies are chosen as partners for political campaigns.

Table 16: Chain-of-thought setting inference-only sample with input and output. Logical relation: *Weaken*.