CASA: Causality-driven Argument Sufficiency Assessment

Xiao Liu[1]*, **Yansong Feng**[1] and **Kai-Wei Chang**[2]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Computer Science Department, University of California, Los Angeles
{lxlisa,fengyansong}@pku.edu.cn
kwchang@cs.ucla.edu

Abstract

The argument sufficiency assessment task aims to determine if the premises of a given argument support its conclusion. To tackle this task, existing works often train a classifier on data annotated by humans. However, annotating data is laborious, and annotations are often inconsistent due to subjective criteria. Motivated by the definition of probability of sufficiency (PS) in the causal literature, we propose CASA, a zero-shot causality-driven argument sufficiency assessment framework. PS measures how likely introducing the premise event would lead to the conclusion when both the premise and conclusion events are absent. To estimate this probability, we propose to use large language models (LLMs) to generate contexts that are inconsistent with the premise and conclusion and revise them by injecting the premise event. Experiments on two logical fallacy detection datasets demonstrate that CASA accurately identifies insufficient arguments. We further deploy CASA in a writing assistance application, and find that suggestions generated by CASA enhance the sufficiency of student-written arguments. Code and data are available at https://github.com/xxxiaol/CASA.

1 Introduction

Argumentation is an integral part of our daily verbal communication (Fogelin and Sinnott-Armstrong, 2005; Stab and Gurevych, 2017a). An argument is a series of statements consisting of premises and a conclusion. Take the argument shown in Figure 1 as an example: *You shouldn't trust Donald's views about politics. He's an alcoholic.* The first sentence, which serves as the conclusion, is supported by the second sentence, acting as the premise. If we have techniques to assess the quality of arguments precisely, we can identify weaknesses in arguments and further improve them.

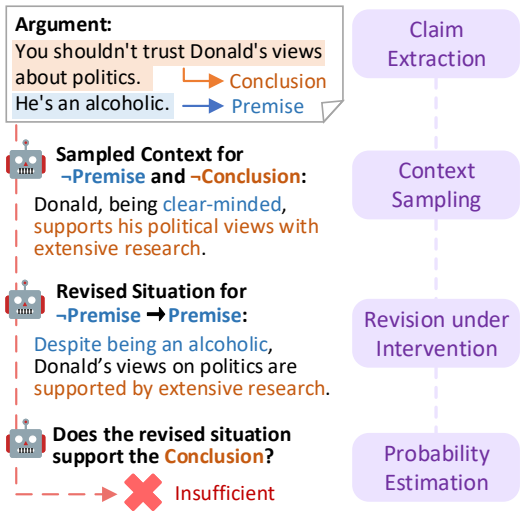

Figure 1: An example of the argument sufficiency assessment task and the reasoning steps of CASA.

An important part of argument quality assessment is to determine whether the premises sufficiently support the conclusion. In a cogent argument, the premises are not only relevant to the conclusion and acceptable on their own, but also collectively sufficient to draw the conclusion (Blair, 2011). We focus on the sufficiency assessment in this paper. In the example of Figure 1, the premise *Donald is an alcoholic* does not sufficiently support the conclusion *his views are untrustworthy*, as there are other factors that could invalidate the conclusion. For instance, if *Donald's views on politics are supported by extensive research*, his views could still be credible even though he is an alcoholic.

Previous works on argument sufficiency assessment train classifiers based on human annotations (Gurcke et al., 2021; Saveleva et al., 2021). However, the sufficiency criteria are vague and subjective among annotators. For example, Wachsmuth et al. (2017) collected annotations from seven annotators, but even the three annotators with the highest consensus only achieved an

*Work done during visiting UCLA.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5282–5302
June 16-21, 2024

agreement of 0.28.[1] This inconsistency poses a challenge in learning an accurate model.

In this paper, we propose CASA, a zero-shot **C**ausality-driven **A**rgument **S**ufficiency **A**ssessment framework by formulating the task with a concept borrowed from causality: the Probability of Sufficiency (PS) (Pearl et al., 2000). Intuitively, if X is a sufficient cause of Y, the presence of X implies the subsequent occurrence of Y. PS quantifies the probability that introducing X would produce Y in the case where X and Y are in fact absent:

$$PS_{X,Y} = P(Y(X = 1) = 1 | X = 0, Y = 0),$$

where $Y(X = 1)$ indicates the value of Y after an intervention on X. Take the example in Figure 1. X is the occurrence of the event *Donald is an alcoholic*, and Y represents *Donald's views about politics are untrustworthy*. If X and Y are both false, but when the event *Donald becomes an alcoholic* occurs, it results in *his political view being untrustworthy*, then the argument is sufficient.

To measure PS of a given argument, there presents the following challenges: 1) How to measure the probabilities without observational data, i.e., how to estimate $P(Y = 1|X = 0, Y = 0)$ if we do not have the corresponding data points. 2) Even if we have the observational data, how to intervene in the argument, i.e., how to estimate $P(Y(X = 1) = 1)$ given data conforming to the conditions of $X = 0$ and $Y = 0$.

Our approach tackles the challenges by leveraging the commonsense knowledge and reasoning abilities of large language models (LLMs) (Bhargava and Ng, 2022; Kojima et al., 2022). Specifically, we ask LLMs to sample data that are inconsistent with the premises and the conclusion, such as *Donald, being clear-minded, supports his political views with extensive research*; and then revise the data to contain the premises, such as *despite being an alcoholic, Donald's views on politics are supported by extensive research*. Step-wise evaluation results demonstrate the effectiveness of using LLMs to sample data and to conduct interventions.

We evaluate CASA's capability of assessing argument sufficiency on two logical fallacy detection datasets, BIG-bench Logical Fallacy Detection (Srivastava et al., 2023) and Climate (Alhindi et al., 2022). We compare our framework with baseline methods including zero-shot/one-shot prompting and perplexity-based classification, and find that CASA distinguishes between sufficient and insufficient arguments more accurately, bringing an average of 10% improvement than directly prompting the same base models.

To further investigate whether our framework can help in realistic scenarios, we apply CASA to provide writing suggestions for student essays. CASA generates *objections* (reasons arguing against the argument) for arguments it finds insufficient in essays. We conduct a human evaluation to assess the quality of CASA's suggestions and their effects on revision. Results demonstrate that the objections generated by CASA are rational and help improve the sufficiency of the arguments.

Our main contributions are as follows: 1) We design CASA, a theoretically grounded framework for argument sufficiency assessment based on the probability of sufficiency. 2) To realize the probability of sufficiency, we exploit LLMs in generating data samples and conducting interventions, and demonstrate the effectiveness with experiments on logical fallacy detection. 3) We demonstrate a practical application of CASA in improving the sufficiency of arguments in student-written essays.

2 The CASA Framework

Notations. We define X as the occurrence of a premise event[2]: $X = \mathbb{1}(\texttt{Premise})$, and Y as the occurrence of a conclusion event: $Y = \mathbb{1}(\texttt{Conclusion})$. Here $\mathbb{1}(\cdot)$ is the indicator function. Both X and Y are binary variables. Y_u indicates the value of Y in the unit u.

Assumptions. Our framework is based on two common assumptions in causal inference (Rubin, 1978):

- *No interference*: the value Y of the unit u is not affected by the values of X assigned to other units.
- *Consistency*: $X = x \rightarrow Y = Y(X = x)$, where x indicates a specific value of X. This requires that each treatment value x has only one form (Imbens and Rubin, 2015).

In our task, the first assumption is satisfied as there is no dependency between the conclusion of one unit and the premise of another. To satisfy the

[1] The agreement is measured with Krippendorff's α. $\alpha = 0.67$ is suggested as the lowest acceptable limit for tentative conclusions (Krippendorff, 2018).

[2] For simplicity, we consider only arguments with a single premise first. We will discuss how to extend the discussion to arguments with multiple premises in Section 2.5.

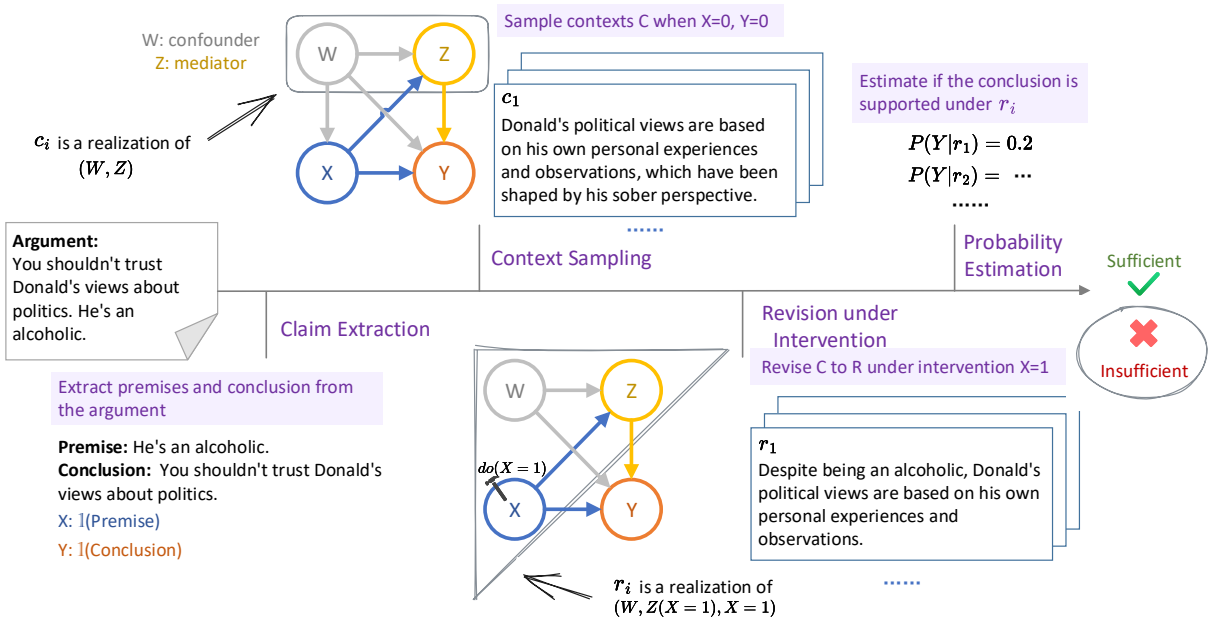

Figure 2: Overall architecture of CASA. Given an argument, we first extract its premises and conclusion, sample contexts that are inconsistent with the premises and conclusion, revise the contexts to meet the premises, and finally estimate the probability of the conclusion. Edge $A \rightarrow B$ in causal graphs means A may causally affects B.

second assumption, we restrict the value $X = 0$ to the occurrence of ¬Premise, the negated form of the premise event. For coherence, $Y = 0$ indicates the occurrence of ¬Conclusion.

Overview. As shown in Figure 2, we first extract the premise and conclusion from a given argument, then sample contexts that meet the conditions, make interventions on the contexts, and finally estimate the probability of the conclusion for each unit.

2.1 Claim Extraction

Given an unstructured argument, we aim to split it into multiple premises and one conclusion. The task of argument parsing is indeed complex (Ajjour et al., 2017),[3] but since it is not the primary component of our framework, we simplify it in two ways: 1) We segment the argument into claims with punctuation marks and conjunction words; 2) we do not consider how the premises are related to each other, such as how one premise might support another premise. Specifically, we list the segmented claims, ask an LLM to select which one is the conclusion, and consider the other claims as premises.

[3]Previous works either use pre-extracted premises and conclusion (Gurcke et al., 2021), or train an individual segmentation model based on annotations (Saveleva et al., 2021).

2.2 Context Sampling

To calculate the conditional probability, an intuitive way is to sample n units $U = \{u_1, ..., u_n\}$ that conform to the conditions $X = 0$ and $Y = 0$. Although we do not have existing observational data, we make use of the commonsense knowledge learned by LLMs, and let them generate n diverse contexts $C = \{c_1, ..., c_n\}$ which are consistent with ¬Premise and ¬Conclusion.

Take the argument in Figure 2 as an example. We instruct the LLM to "generate n detailed contexts. Each context is consistent with both the premise *Donald isn't an alcoholic* and the conclusion *you should trust Donald's views about politics.*" A generated context mentions that *Donald's political views are based on his own personal experiences and observation*, which is consistent with ¬Conclusion; and his views *have been shaped by his sober perspective*, consistent with ¬Premise.

$PS_{X,Y}$ is then estimated with the average of $P(Y_{u_i}(X = 1) = 1)$ under each unit u_i:

$$PS_{X,Y} \approx \frac{1}{|U|} \sum_{u_i \in U | X=0, Y=0} P(Y_{u_i}(X = 1) = 1).$$

From the causal lens, we can decompose the context information into two latent parts W and Z. W is the part that is not causally affected by X, called confounder; and Z is the remaining part that may

be causally affected by X, called mediator.[4] Each context c_i can be seen as a realization of (W, Z).

2.3 Revision under Intervention

For each unit $u_i = (c_i, X = 0, Y = 0)$, our next step is to implement the intervention $X = 1$ on it. The effect of the intervention on the context c_i is illustrated by the causal graph in the lower part of Figure 2. The intervention breaks the causal relation $W \rightarrow X$, and leaves W unchanged. At the same time, Z will change according to the intervention, becoming $Z(X = 1)$. Therefore, we can rewrite the estimand as:

$$P(Y_{u_i}(X = 1) = 1) = P(Y_{u_i}(X = 1, Z(X = 1)) = 1).$$

We ask the LLM to revise each context c_i to r_i under intervention $X = 1$. Specifically, our instruction is to revise the context to contain the premise,[5] so the information $X = 1$ is also included in r_i, leading r_i to be a realization of $(W, Z(X = 1), X = 1)$.

In Figure 2, the expression *sober perspective* is removed by the LLM, and the `Premise` is added with *Despite being an alcoholic*. At the same time, *Donald's political views are based on his own personal experiences and observation* is kept unchanged, as this does not violate the `Premise`.

2.4 Probability Estimation

We transform the probability estimation of Y into the form of natural language inference (NLI): under each situation r_i, estimate whether the conclusion is supported or contradicted. We use an off-the-shelf NLI model to make the prediction, and aggregate the units to calculate the final $PS_{X,Y}$.

2.5 Dealing with Multiple Premises

When an argument contains multiple premises, we exhaustively check the sufficiency of each premise given the context of other premises:

$$PS_{X_i,Y|X_{1\cdots n\setminus i}} = P(Y(X_i = 1) = 1 | X_i = 0, Y = 0, X_{1\cdots n\setminus i} = \mathbf{1}).$$

Concretely, we ask the LLM to contain $\texttt{Premise}_{1\cdots n\setminus i}$ when sampling contexts for checking the sufficiency of the i-th premise event.

[4] Our naming conforms to the common terminology. As shown in the upper left of Figure 2, W is called confounder because there may be causal relations $W \rightarrow X$ and $W \rightarrow Y$, and Z is called mediator because there may be causal relations $X \rightarrow Z$ and $Z \rightarrow Y$.

[5] An example prompt is in Appendix Table 8.

3 Experiments

3.1 Datasets

We conduct experiments on two English logical fallacy detection datasets: BIG-bench Logical Fallacy Detection (BIG-bench-LFD) (Srivastava et al., 2023) and Climate (Alhindi et al., 2022). Logical fallacy detection requires models to distinguish between fallacious and correct arguments. As Jin et al. (2022) mentioned, logical fallacies usually happen when the premises are insufficient to draw the conclusion. We manually check both datasets and confirm that the fallacious arguments can be attributed to insufficiency.

Due to the subjective criteria of argument sufficiency annotation mentioned in Section 1, the existing argument sufficiency datasets are noisy. We do not use them for automatic evaluation, but they will be used in the application of writing suggestions in Section 5. In contrast, logical fallacy datasets are more objective, with a clear distinction between fallacious and correct arguments.

BIG-bench-LFD. This dataset aims to evaluate LLMs' capabilities of detecting informal and formal fallacies. We only consider the informal statement portion, whose statements are more similar to real arguments. They are examples of good and bad cases of informal reasoning collected from philosophers, including 57 correct and 143 fallacious arguments.

Climate. This dataset contains arguments from climate change articles fact-checked by climate scientists at `climatefeedback.org`. Because some arguments in this dataset are single claims without premises, we only use instances with more than one sentence to avoid these single claim arguments, resulting in 30 correct and 76 fallacious arguments.

Accuracy and macro-F1 are reported for both datasets.

3.2 Experimental Setup

We experiment with two instruction-tuned models: TULU-7B (Wang et al., 2023b) and LLAMA-2-7B-CHAT (Touvron et al., 2023b) as the base models of our framework. TULU-7B is finetuned on LLAMA-7B (Touvron et al., 2023a) with an aggregation of instruction tuning datasets, achieving great performance across benchmarks. We do not choose the prevalent GPT models like GPT-4 because portions of BIG-bench data were mixed into its training set (OpenAI, 2023).

Model	Acc	Macro-F1
Unsupervised		
Zero-shot Prompting (TULU)	59.5	59.3
Zero-shot Prompting (LLAMA2)	70.0	66.8
Perplexity (TULU)	56.0	54.7
Perplexity (LLAMA2)	51.0	50.8
DeBERTa-NLI	57.0	54.1
BART-NLI	67.0	65.5
CASA (TULU)	77.0	70.8
CASA (LLAMA2)	**79.0**	**73.4**
One-shot Prompting (TULU)	61.1	59.7
One-shot Prompting (LLAMA2)	74.1	68.6

(a) Results on BIG-bench-LFD.

Model	Acc	Macro-F1
Unsupervised		
Zero-shot Prompting (TULU)	33.0	30.3
Zero-shot Prompting (LLAMA2)	51.9	48.0
Perplexity (TULU)	63.2	38.7
Perplexity (LLAMA2)	66.9	45.0
DeBERTa-NLI	55.7	51.5
BART-NLI	63.2	53.2
CASA (TULU)	64.2	54.9
CASA (LLAMA2)	**67.9**	**61.2**
One-shot Prompting (TULU)	45.6	45.5
One-shot Prompting (LLAMA2)	52.8	51.1

(b) Results on Climate.

Table 1: Automatic evaluation results of argument sufficiency assessment, showing that CASA outperforms all zero-shot and one-shot baselines. Numbers are in percentages (%).

	CASA (TULU)	CASA (LLAMA2)
Claim Extraction		
Correctness	93%	92%
Context Sampling		
Consistency with X=0	96%	90%
Consistency with Y=0	91%	93%
Revision under Intervention		
Consistency with X=1	95%	96%

Table 2: Step-wise evaluation on BIG-bench-LFD.

We use an off-the-shelf negator (Anschütz et al., 2023) based on syntactic rules to generate ¬Premise and ¬Conclusion, and use BART-NLI[6] as the NLI model used in probability estimation. We sample $n = 3$ units for each argument, and make the final decision with a majority vote. More implementation details are in Appendix A.1.

There is no existing zero-shot argument sufficiency assessment model to our knowledge, so we build several non-trivial baselines for comparison:

Zero-shot Prompting. We probe the base models TULU and LLAMA2 with four prompt forms (two forms provided by BIG-bench-LFD, and two forms written by ourselves), and report the best performance. The detailed prompts are in Appendix A.2.

Perplexity. Motivated by Zhang et al. (2022a), we compute perplexity scores for the base models as another zero-shot baseline. For each argument, we compare the perplexity score of Premise||Conclusion and Premise||¬Conclusion, and regard the one with lower perplexity score as the model prediction. Here || indicates text concatenation.

NLI Models. We directly use two NLI models, RoBERTa-NLI (Reimers and Gurevych, 2019) and BART-NLI, to conduct the task. Specifically, NLI models are asked to predict if the premises support or contradict the conclusion.

One-shot Prompting. Besides the zero-shot baselines, we add one-shot prompting into comparison, to help LLMs better understand the task instruction. We use the prompt form that performs the best in zero-shot prompting, and randomly select one example from the datasets. We test models three times with different examples and report the average performance.

3.3 Results

CASA vs. Baselines. Table 1 reports the automatic evaluation results. CASA significantly outperforms all the corresponding zero-shot baselines with significance level $\alpha = 0.02$, and also surpasses the one-shot baselines. This demonstrates the effectiveness of our overall framework. The performance on Climate is inferior to on BIG-bench-LFD for all models, because Climate requires specific domain knowledge and more rigorous reasoning, as some fallacies are less apparent. CASA shares the same domain knowledge with the baselines as they use the same base models, but the causality-driven framework equips it with better reasoning performance. We demonstrate a case in Section 4.3.

Step-wise Evaluation. To examine whether LLMs work as we expect in each step of CASA, we conduct step-wise human evaluation. We ask human annotators to rate three aspects individually: 1) In the claim extraction step, do LLMs extract the correct premises and conclusion from the argument? 2) In the context sampling step, are the contexts generated by LLMs consistent with ¬Premise and ¬Conclusion? 3) In the revision step, are the re-

[6] https://huggingface.co/facebook/bart-large-mnli

Model	Acc	Macro-F1
CASA (TULU)	77.0	**70.8**
w/o Intervention	74.5	62.5
w/o Condition on $X = 0$	75.0	68.7
w/o Condition on $Y = 0$	**78.0**	69.8
Intervention: Concatenation	74.0	63.3

(a) Ablations for CASA (TULU).

Model	Acc	Macro-F1
CASA (LLAMA2)	**79.0**	**73.4**
w/o Intervention	73.5	65.8
w/o Condition on $X = 0$	75.5	67.1
w/o Condition on $Y = 0$	75.0	66.7
Intervention: Concatenation	78.0	67.4

(b) Ablations for CASA (LLAMA2).

Table 3: Ablation results on BIG-bench-LFD. All ablations lead to performance drops on macro-F1.

Argument: Biological, geological and planetary systems are extremely robust. Our evolving dynamic planet has survived sea level changes of hundreds of metres.

Correct Answer: Insufficient **Zero-shot Prompting (TULU):** Sufficient

CASA (TULU)

Premise: Our evolving dynamic planet has survived sea level changes of hundreds of metres.

Conclusion: Biological, geological and planetary systems are extremely robust.

Sampled Contexts:

1. The rapid rise of sea levels caused by climate change has led to the destruction of many coastal cities and ecosystems, demonstrating the vulnerability of biological, geological, and planetary systems.

2. The geological history of our planet is marked by numerous catastrophic events, such as massive volcanic eruptions and asteroid impacts, which have had a significant impact on the evolution of life on Earth.

3. The delicate balance of our planet's systems, from the tides that shape our coastlines to the complex interactions between plant and animal species, highlights the need for greater understanding and protection of these systems in the face of ongoing environmental changes.

Revised Situations:

1. Although our evolving dynamic planet has survived sea level changes of hundreds of metres, the rapid rise of sea levels caused by climate change has led to the destruction of many coastal cities and ecosystems, demonstrating the vulnerability of biological, geological, and planetary systems.

2. The geological history of our planet is marked by numerous catastrophic events, such as massive volcanic eruptions and asteroid impacts, which have had a significant impact on the evolution of life on Earth. However, our evolving dynamic planet has survived sea level changes of hundreds of metres.

3. Our evolving dynamic planet has survived sea level changes of hundreds of metres, but the delicate balance of our planet's systems, from the tides that shape our coastlines to the complex interactions between plant and animal species, highlights the need for greater understanding and protection of these systems in the face of ongoing environmental changes.

Prediction: Insufficient

Table 4: An example of the detailed reasoning process of CASA (TULU) on Climate.

vised situations consistent with the `Premise`? We sample 100 instances from BIG-bench-LFD and recruit three annotators to answer each question. We report the majority vote results, and the inter-annotator agreement is 84%.

Table 2 shows the step-wise evaluation results. The accuracy of all aspects is above 90%, exhibiting that LLMs are capable of generating textual data that conform to certain conditions, and making interventions on situations in the form of natural language. We provide the annotation templates and error analysis in Appendix A.3.

4 Analysis

4.1 Ablation Study

To further investigate the effectiveness of CASA components, we study several variants of CASA:

w/o Intervention. In this ablation, we simply estimate $P(Y = 1|X = 1)$ without intervention. Concretely, we sample contexts for `Premise`, and estimate the probability of `Conclusion` based on contexts and `Premise`.

w/o Condition on $X = 0$. This variant estimates $P(Y(X = 1) = 1|Y = 0)$, where the term $X = 0$ is removed from the original PS definition. In the context sampling step, we ask LLMs to generate contexts only consistent with ¬`Conclusion`, and other steps are kept the same.

w/o Condition on $Y = 0$. This variant estimates $P(Y(X = 1) = 1|X = 0)$, and LLMs are asked to generate contexts consistent with ¬`Premise`.

Intervention: Concatenation. We study if the intervention step can be replaced by simply concatenating the context with `Premise`. In this setting, the mediator Z in the context remains unchanged.

Table 3 shows the ablation results on BIG-bench-LFD, and results on Climate are in Appendix Table 10. All the ablations lead to performance drops

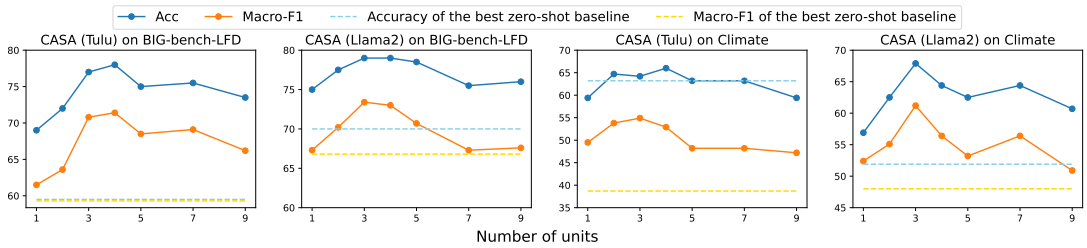

Figure 3: Hyperparameter analysis on the number of units. CASA consistently outperforms baselines on macro-F1.

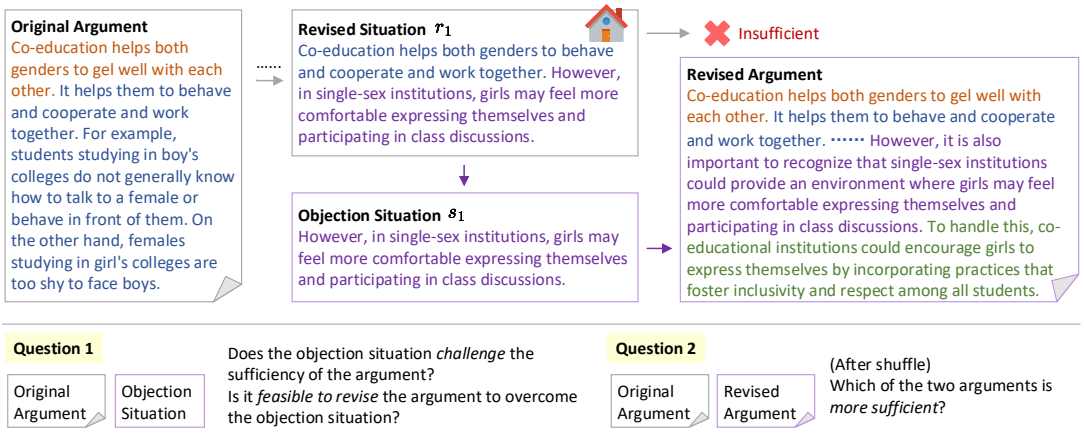

Figure 4: The process of providing writing assistance with CASA (top) and conducting human evaluations for the assistance effectiveness (bottom).

on macro-F1, indicating that the original probability of sufficiency definition is not only of theory value, but also of practical value.

4.2 Hyperparameter Study

To study the performance sensitivity of CASA to the number of units n, we vary n from 1 to 9 for CASA (TULU) and CASA (LLAMA2), and exhibit the performance in Figure 3.

CASA consistently outperforms zero-shot baseline models on macro-F1 regardless of the number of units, proving its robustness towards the hyperparameter. The performance peak is around 3 in all settings. When we sample too few units, they hardly encompass a wide variety of situations. On the other hand, when we sample many contexts at once, the quality of the contexts goes down and each context tends to be shorter.

4.3 Case Study

We demonstrate an example of the reasoning process of CASA on Climate in Table 4, and an example on BIG-bench-LFD in Appendix Table 11. In both cases, CASA is able to detect the insufficiency in the argument, while directly prompting the same base model fails to find the fallacy. Specifically, in Table 4, CASA generates several evidences supporting that *biological systems are vulnerable*, like *the sea level rise leads to the destruction of coastal cities* and *volcanic eruptions impact the evolution of life*. These evidences do not contradict with the premise that *our planet has survived sea level changes*. Therefore, they are kept in the revised situations, and make the conclusion *biological systems are extremely robust* no longer supported.

5 Application: Writing Assistance

We apply CASA to a realistic scenario: providing writing suggestions for essays written by students. If CASA identifies that an argument in an essay is insufficient, we extract explainable reasons from CASA's reasoning process, and provide them as suggestions for revision.

Specifically, we generate *objection situations* (situations that challenge the sufficiency of the argument) out of intervened situations R that contradict the `Conclusion`, by removing the `Premise` from R. As shown in the example of Figure 4, the revised situation r_1 is converted to the objection situation s_1 by removing the premise *co-education helps both genders to behave and cooperate and work together.* The removal is automatically done

Model	Rationality	Feasibility
Prompting (TULU)	84%	60%
Prompting (LLAMA2)	90%	67%
CASA (TULU)	90%	76%
CASA (LLAMA2)	**92%**	**81%**

(a) Question 1: Is CASA capable of generating rational and feasible objection situations to the essays?

Model	Revised is Better	Tie	Original is Better
Prompting (TULU)	**41%**	38%	21%
Prompting (LLAMA2)	**51%**	23%	26%
CASA (TULU)	**40%**	43%	17%
CASA (LLAMA2)	**58%**	35%	7%

(b) Question 2: Will revising based on the generated objection situations improve the sufficiency of the essays?

Table 5: Human evaluation results of using CASA to provide writing suggestions.

by detecting if each sentence in the revised situation entails one premise event.

We investigate *(Q1):* whether CASA is capable of generating rational and feasible objection situations to the essays, and *(Q2):* whether revising based on the generated objection situations improves the sufficiency of the essays.

Dataset. The Argument-Annotated Essays (AAE, Stab and Gurevych (2017b)) dataset contains 402 argumentative essays and 1,029 extracted arguments. Stab and Gurevych (2017c) use AAE to annotate argument sufficiency. They find that some arguments in this dataset suffer from the insufficiency problem, but the criteria for insufficiency are vague among annotators.[7] Therefore, we only use the corpus, but not the annotations.

We randomly sample 100 arguments from AAE, and assess their sufficiency with CASA. For the arguments CASA finds insufficient, we randomly transform one revised situation into the objection situation. We recruit three annotators to answer each question, and the annotation templates are in Appendix A.5.

Objection Quality. Given an argument and an objection generated by CASA, we ask annotators to evaluate whether the objection challenges the sufficiency of the argument (rationality), and whether it is feasible to revise the argument to overcome the objection (feasibility). We also build a baseline of directly prompting LLMs to generate an objection situation if they identify the argument as insufficient. The prompt is in Appendix Table 13.

As shown in Table 5a, objection situations generated by CASA are more rational and feasible than directly prompting the base models. The gap in feasibility is larger, as LLMs are likely to generate abstract objections when prompting, like *the argument does not consider the potential challenges that may arise when students from different gender backgrounds interact*, while CASA provides more practical objections which are easier to address.

Effect of the Revision. To evaluate the sufficiency of revised arguments, we first use GPT-4 (OpenAI, 2023) to revise the arguments based on the objection situations generated by the prompting method and CASA. This is to simulate a human revision process, since asking human annotators to complete this task is expensive and hard to control the quality. We further recruit annotators to check the revised arguments to ensure the revision quality. The revision prompt and quality check are described in Appendix A.5.

We ask annotators to compare the sufficiency of the original and revised arguments. As shown in Table 5b, in both methods tested, the Revised is Better proportion supersedes the Original is Better proportion, emphasizing an improvement in writing sufficiency. On the other hand, with the same base model, CASA obtains a higher Revised - Original ratio (the Revised is Better proportion minus the Original is Better proportion) compared to the prompting method. This suggests that, even if we do not consider the difficulty of revision, CASA helps more in the revision process.

6 Related Work

Argument Sufficiency Assessment. Previous works on argument sufficiency assessment mainly use standard inference models. Wachsmuth and Werner (2020) train support vector machines (SVM) on text features; Stab and Gurevych (2017c) use convolutional neural networks (CNN) to recognize insufficient arguments; and Saveleva et al. (2021) employ graph neural networks (GNN) to better understand argument structures. However, they all suffer from the subjective criteria of annotation (Rach et al., 2020; Wachsmuth et al., 2017). Additionally, they simply treat the task as a normal classification task without considering the nature of

[7] Although the overall agreement between annotators is 0.77 measured with Krippendorff's α, on instances at least one annotator labels as insufficient (32% of all instances), the agreement is only 0.14.

sufficiency. Gurcke et al. (2021) model the relation between premises and conclusion of a sufficient argument, but the relation is based on their personal hypotheses, whereas CASA possesses a clear theoretical foundation.

Writing Assistance. There are also previous works trying to provide writing assistance to human-written articles. Hanawa et al. (2021); Zhang et al. (2022b); Wang et al. (2023a) focus on polishing the form of the writing, like grammatical correctness, word choices and rhetorical methods. Wambsganss and Niklaus (2022) try to provide feedback on argument content, but the feedback is mostly given as scores on predefined dimensions. In contrast, CASA provides feedback in the form of objection situations, which makes the revision easier. Skitalinskaya and Wachsmuth (2023) discuss how to identify argumentative claims that need further revision, which is complementary to our work.

7 Conclusion

We propose CASA, a zero-shot argument sufficiency assessment framework driven by the causal concept of sufficiency. In the absence of observational data and intervention data, we sample contexts and make interventions with LLMs. CASA is capable of identifying insufficient arguments on two logical fallacy detection datasets, and providing writing suggestions to further improve the sufficiency of human-written arguments.

Acknowledgments

We thank Xueqing Wu, Fan Yin, Ashima Suvarna, Da Yin, and other UCLA-NLP lab members for their constructive comments. We thank the anonymous reviewers for their helpful discussions and suggestions. This research was supported in part by NSF #2331966.

Limitations

Choices in Model Design. We tried to explore diverse decoding methods when sampling contexts, as they may generate a large number of diverse contexts without impairing the quality of each context. However, we find that in our scenario, current diverse decoding methods can hardly generate high-quality contexts that are diverse in content, so we tend to instruct LLMs to generate multiple contexts in one run.

The goal of the revision under intervention step can be viewed as a counterfactual reasoning task, so we also explore zero-shot counterfactual reasoning models for this step, but they are either slow in inference or ineffective in the generation quality when conducting the revision task. Therefore, we prompt LLMs to complete the step. We are willing to switch to new methods for these steps if more powerful diverse decoding/counterfactual reasoning methods are released.

Data Scope. Evaluating model performances on the argument sufficiency assessment task is difficult, due to the aforementioned subjective annotation criteria. Although we try our hardest to find automatic evaluation datasets, the two datasets we use are still of limited scope. To make up for this, we calculate the significance level and conduct diverse analyses.

Ethics Statement

Our framework can be applied to educational scenarios like student essay assessment and comment generation. However, as we cannot ensure the model prediction is correct, it must be used with manual check. Moreover, as our framework is based on existing LLMs, its generations may inherit the bias of LLMs. Therefore it should be used under human supervision.

For human evaluations, we recruit annotators from Amazon Mechanical Turk, and all annotators are fairly paid more than $10 USD per hour (it varies depending on the time spent on HITs), which is higher than the national minimum wage where the annotators are recruited.

References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Miriam Anschütz, Diego Miguel Lozano, and Georg Groh. 2023. This is not correct! negation-aware evaluation of language generation systems. *arXiv preprint arXiv:2307.13989*.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pretrained language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12317–12325.

J Anthony Blair. 2011. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*, volume 21. Springer Science & Business Media.

Robert J Fogelin and Walter Sinnott-Armstrong. 2005. Understanding arguments. *An introduction to informal logic*.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9719–9730, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3.

Niklas Rach, Yuki Matsuda, Johannes Daxenberger, Stefan Ultes, Keiichi Yasumoto, and Wolfgang Minker. 2020. Evaluation of argument search approaches in the context of argumentative dialogue systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 513–522, Marseille, France. European Language Resources Association.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Donald B Rubin. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Ekaterina Saveleva, Volha Petukhova, Marius Mosbach, and Dietrich Klakow. 2021. Graph-based argument quality assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1268–1280, Held Online. INCOMA Ltd.

Gabriella Skitalinskaya and Henning Wachsmuth. 2023. To revise or not to revise: Learning to detect improvable claims for argumentative writing support. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15799–15816, Toronto, Canada. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017c. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay

Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760, Dublin, Ireland. Association for Computational Linguistics.

Chenshuo Wang, Shaoguang Mao, Tao Ge, Wenshan Wu, Xun Wang, Yan Xia, Jonathan Tien, and Dongyan Zhao. 2023a. Smart word suggestions for writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11212–11225, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Jiayao Zhang, Hongming Zhang, Weijie Su, and Dan Roth. 2022a. Rock: Causal inference principles for reasoning about commonsense causality. In *International Conference on Machine Learning*, pages 26750–26771. PMLR.

Zhexin Zhang, Jian Guan, Guowei Xu, Yixiang Tian, and Minlie Huang. 2022b. Automatic comment generation for Chinese student narrative essays. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 214–223, Abu Dhabi, UAE. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

We demonstrate examples of the prompts we used in CASA. Table 6 showcases the prompt for claim extraction. The argument is segmented using rules of punctuation marks and conjunction words. Specifically, we first check if the argument can be split with punctuation marks like periods and semicolons. If not, we further split it with conjunction words like *because* and *so*.

Table 7 shows the prompts for context sampling. The premise and conclusion in the prompts are ¬`Premise` and ¬`Conclusion` of the argument. When the argument contains multiple premises, we check one premise each time, and ask LLMs to generate contexts containing other premises, like *positive things are good* in the example.

Table 8 shows the prompt for revision under intervention. The statement here is the premise of the argument, and the text is a context generated in the last step.

When probing TULU, we add the prefix "<|user|>" and suffix "<|assistant|>" to the prompts, aligning with the original instruction tuning format (Wang et al., 2023b).

Experiments are conducted on NVIDIA RTX A6000 GPUs. It takes CASA two minutes to determine the sufficiency of one argument on average.

A.2 Details of Prompting Baselines

BIG-bench-LFD provides two prompt forms, and we follow these prompts in our evaluation. Moreover, we modify the prompts to explicitly ask LLMs to predict "sufficient" or "insufficient". These lead to the four prompts shown in Table 9.

When probing LLAMA2, we further try two options: 1) directly use the prompts, and 2) wrap them up with "<s>[INST] «SYS»{{system_prompt}}«/SYS»{{our_prompt}} [/INST]".[8] The second option is suggested by the authors of LLAMA2, but it sometimes refuses to answer questions. When probing TULU, we add the prefix "<|user|>" and suffix "<|assistant|>" to the prompts. We evaluate LLMs with all the prompts, and report the performance of the prompt with the highest macro-F1 on the test set.

A.3 Step-wise Evaluation

We demonstrate the annotation templates of step-wise evaluation in Figure 5 for claim extraction, and Figure 6 for consistency between the generated context and ¬`Premise`. Other consistency evaluations are similar to Figure 6.

[8]The system prompt is shown in https://github.com/huggingface/blog/blob/main/llama2.md

Instruction:
Determine which part of the text is the conclusion.
Output the number of the conclusion part first, and give an explanation.
Format:
Conclusion: [number]
Explanation: ...
Input:
You shouldn't trust Donald's views about politics. He's an alcoholic.
Choices:
1. You shouldn't trust Donald's views about politics.
2. He's an alcoholic.
Response:

Table 6: Example prompt for claim extraction.

Instruction:
Generate 3 detailed contexts. Each context is consistent with both the premise and the conclusion. Each context is in one line.
Input:
Premise: He isn't an alcoholic.
Conclusion: You should trust Donald's views about politics.
Response:

(a) Single premise.

Instruction:
Generate 3 detailed contexts. Each context contains "Positive things are good." Each context is consistent with both the premise and the conclusion. Each context is in one line.
Input:
Premise: My drug test wasn't positive.
Conclusion: My test result wasn't good.
Response:

(b) Multiple premises. The argument is "My drug test was positive, and positive things are good. So my test result was good."

Table 7: Example prompts for context sampling.

We also ask annotators to choose the error reason if they feel the extracted annotation is wrong, or the generated text is inconsistent with the given statement. For claim extraction, all four error types shown in Figure 5 exist, while *wrong conclusion* occurs more often, in 40% of the incorrect cases. For consistency, the generated text contradicts with the given statement in about 50% of the errors; the text is irrelevant to the given statement in 40% errors; and the text is ambiguous or unintelligible in 10% errors.

Additionally, we estimate whether the revised situations preserve the original content with BLEU (Papineni et al., 2002). The BLEU score between revised situations and originally generated contexts is 57% and 53% for CASA (TULU) and CASA (LLAMA2) respectively, demonstrating high-level content preservation.

A.4 Additional Results

Table 10 shows the ablation results on Climate. Similar to the results on BIG-bench-LFD, all the ablations cause performance drops on macro-F1.

Table 11 exhibits an example of the detailed reasoning process of CASA(LLAMA2) on BIG-bench-LFD.

A.5 Details of the Writing Assistance Application

We demonstrate the annotation template for Question 1 *(Is* CASA *capable of generating rational and feasible objection situations to the essays?)* in Figure 7, and for Question 2 *(Will revising based on the generated objection situations improve the sufficiency of the essays?)* in Figure 8. The average inter-annotator agreement is 77%. As the rationality and feasibility may be hard to understand, we provide two manually written examples and explanations for annotators. Although we design three

Instruction:
Revise the text to contain the provided statement.
Input:
Text: Donald's political views are based on his own personal experiences and observations, which have been shaped by his sober perspective.
Statement: He's an alcoholic.
Response:

Table 8: Example prompt for revision under intervention.

Prompt 1	**### Instruction:** This AI is identifying whether statements contain fallacies. The AI responds with 'Valid' or 'Invalid' as appropriate. **### Input:** You shouldn't trust Donald's views about politics. He's an alcoholic. **### Response:**
Prompt 2	**### Instruction:** Do you think the following argument is 'Valid' or 'Invalid'? **### Input:** You shouldn't trust Donald's views about politics. He's an alcoholic. **### Response:**
Prompt 3	**### Instruction:** This AI is identifying whether arguments are sufficient. The AI responds with 'Sufficient' or 'Insufficient' as appropriate. **### Input:** You shouldn't trust Donald's views about politics. He's an alcoholic. **### Response:**
Prompt 4	**### Instruction:** Do you think the following argument is 'Sufficient' or 'Insufficient'? **### Input:** You shouldn't trust Donald's views about politics. He's an alcoholic. **### Response:**

Table 9: Prompts explored for the zero-shot prompting baselines.

options in Figure 7, no question receives a majority vote of "no" in practice, so we only report the percentage of "yes" in Table 5a.

The revision prompt for GPT-4 is shown in Table 12. We use the version `gpt-4-0314` and temperature 0. The prompt for the direct prompting baselines is in Table 13. To help LLMs better understand the instruction, we provide a human written example in the prompt. (In contrast, CASA is zero-shot.)

To evaluate the revision quality of GPT-4, we ask annotators to answer two questions as shown in Figure 9: 1) Does the revised argument address the objection situation's concern? 2) Does the revised argument preserve the main content of the original argument? On 50 randomly sampled revisions from objections generated by CASA (TULU) and CASA (LLAMA2), 100% of them address the objection, and 90% of them preserve the main content of the original arguments.

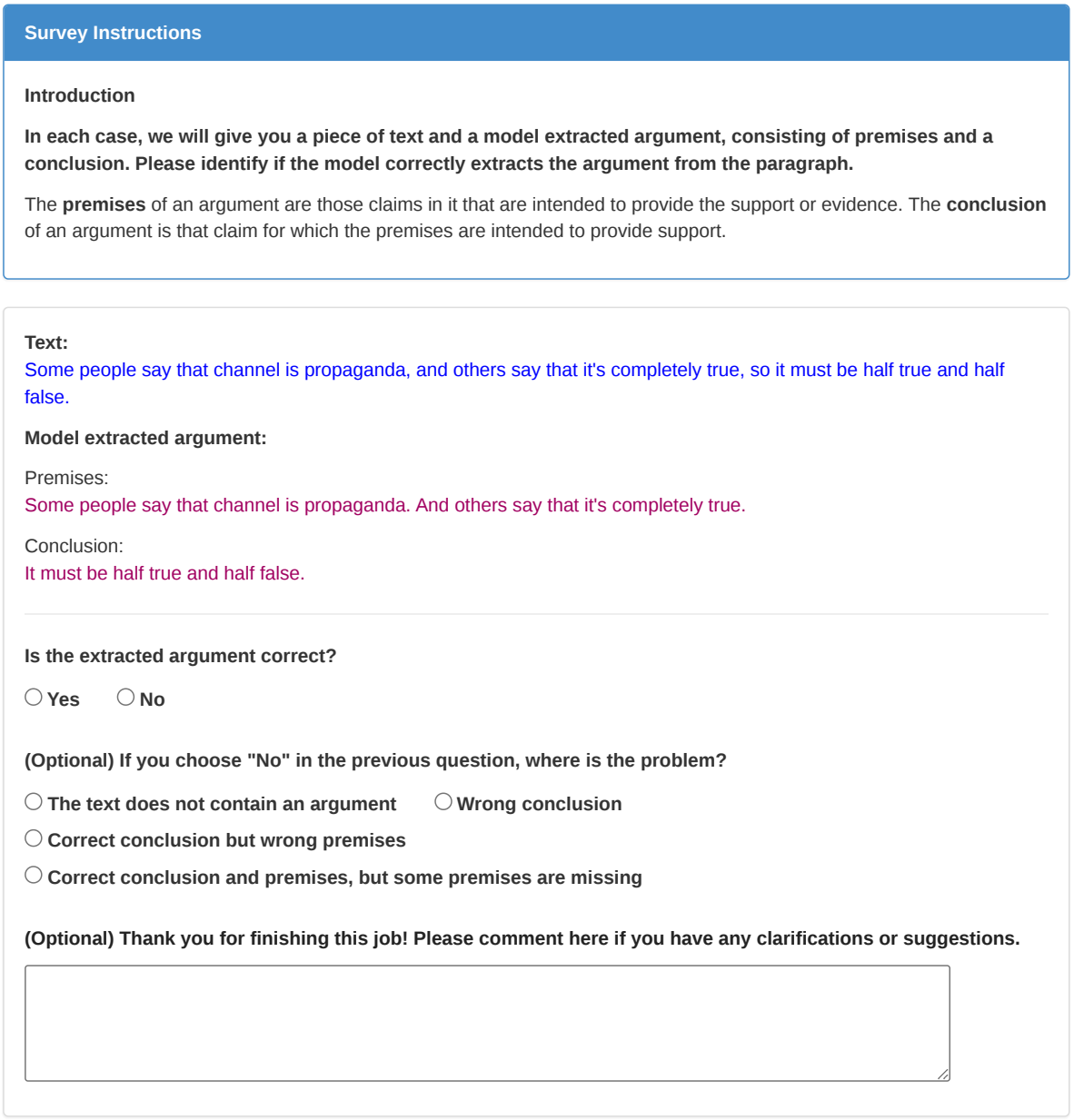

Figure 5: Annotation template for step-wise evaluation: correctness of claim extraction.

Model	Acc	Macro-F1
CASA (TULU)	64.2	**54.9**
w/o Intervention	65.1	52.2
w/o Condition on $X = 0$	**66.0**	50.0
w/o Condition on $Y = 0$	59.4	53.1
Intervention: Concatenation	62.3	47.7

(a) Ablations for CASA (TULU).

Model	Acc	Macro-F1
CASA (LLAMA2)	**67.9**	**61.2**
w/o Intervention	57.5	50.1
w/o Condition on $X = 0$	60.4	45.0
w/o Condition on $Y = 0$	60.4	53.0
Intervention: Concatenation	**67.9**	59.7

(b) Ablations for CASA (LLAMA2).

Table 10: Ablation results on Climate. Numbers are in percentages (%).

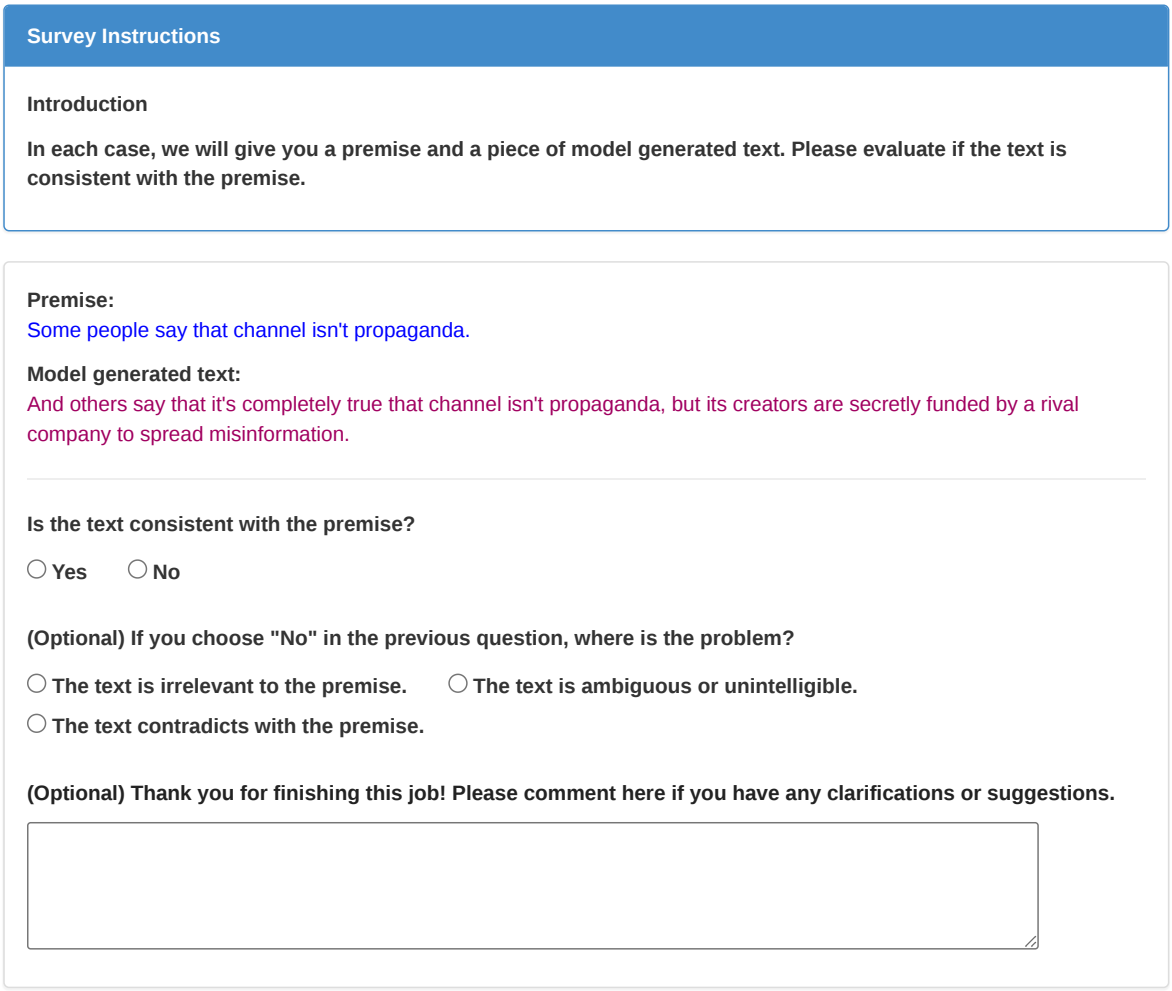

Figure 6: Annotation template for step-wise evaluation: consistency of the generated context with ¬Premise.

Argument: I can't understand Higgs Theorem, so it must be false.

Correct Answer: Insufficient **Zero-shot Prompting (LLAMA2):** Sufficient

CASA (LLAMA2)

Premise: I can't understand Higgs Theorem.

Conclusion: It must be false.

Sampled Contexts:

1. As a physicist, I have spent years studying the Higgs boson and its role in the Standard Model of particle physics. I have a deep understanding of the mathematical concepts involved in the Higgs Theorem and have verified its validity through numerous experiments.

2. As a philosopher, I have explored the nature of truth and its relationship to mathematical concepts. I have come to understand that the Higgs Theorem is not just a mathematical construct, but a reflection of the fundamental laws of the universe. Its truth is not something that can be proven or disproven, but rather a reflection of the underlying reality.

3. As a detective, I have solved many complex cases involving mathematical puzzles and logical reasoning. I have a keen eye for detail and an ability to see patterns that others may miss. When I apply these skills to the Higgs Theorem, I am convinced that it must be true, as the evidence supports its validity.

Revised Situations:

1. As a physicist, I have spent years studying the Higgs boson and its role in the Standard Model of particle physics. But in fact, I can't understand Higgs Theorem without a deep understanding of the mathematical concepts involved in the Higgs Theorem and verifying its validity through numerous experiments.

2. As a philosopher, I have explored the nature of truth and its relationship to mathematical concepts. I have come to understand that the Higgs Theorem is not just a mathematical construct, but a reflection of the fundamental laws of the universe. In fact, I find it difficult to understand Higgs Theorem myself, as it seems to be a complex and abstract concept that defies easy comprehension. However, I believe that its truth lies beyond the realm of human understanding, and can only be appreciated through a deep contemplation of the fundamental laws of the universe.

3. As a detective, I have solved many complex cases involving mathematical puzzles and logical reasoning. I have a keen eye for detail and an ability to see patterns that others may miss. When I apply these skills to the Higgs Theorem, I am convinced that it must be true, as the evidence supports its validity. However, I must admit that I can't understand the Higgs Theorem myself, as it is far beyond my area of expertise.

Prediction: Insufficient

Table 11: An example of the detailed reasoning process of CASA (LLAMA2) on BIG-bench-LFD.

System	You are a helpful and educated assistant.
User	In each case, we will give you an argument and a model generated objection situation.
	Your task is to revise the argument to address the concern raised in the objection situation. Please keep the conclusion and reasonable premises of the original argument unchanged.
User	**Argument:**
	Co-education helps both genders to gel well with each other. It helps them how to behave and cooperate and work together. For example, students studying in boy's colleges do not generally know how to talk to a female or behave in front of them. On the other hand, females studying in girl's colleges are too shy to face boys. Co-education will help to eradicate this kind of demerit in both. Universities giving both genders equal opportunities, will prepare them for future challenges and will help in the long run.
	Objection situation:
	However, in single-sex institutions, girls may feel more comfortable expressing themselves and participating in class discussions.
	Revised argument:

Table 12: Example prompt for GPT-4 revision based on the objection situation.

Survey Instructions

Introduction

In each case, we will give you an argument and a model generated objection situation. Please determine if the objection situation challenges the sufficiency of the argument, and if it is feasible to revise the argument to overcome the objection.

An argument complies with the **sufficiency** criterion if its premises supply all the grounds that are needed to make it reasonable to believe its conclusion.

An **objection** provides reasons arguing against the argument.

The sufficiency of the argument is challenged, if the objection raises **reasonable concerns** about the conclusion, and those concerns are **not addressed** in the premises.

Examples

Example 1

Argument:
If the government provides those without work with a mobile phone, they will be able to find themselves an occupation in order to live and survive. When people without jobs have hand phones with internet access, they can browse the net for more job opportunities. They can do research on the work they have found and prepare themselves for the job. Mobile phones can be used to make calls with the companies they would like to work with.

Model generated objection situation:
However, having a mobile phone with internet access does not guarantee that they will find a job, as there may be other factors such as a lack of available positions, a mismatch in skills, or a highly competitive job market.

Analysis:

The objection raises a reasonable concern that "there may be other factors such as a lack of available positions, a mismatch in skills, or a highly competitive job market". So the objection **challenges** the sufficiency of the argument.

The writer can address the objection by adding a sentence "While there may be factors beyond the availability of a mobile phone that influence job prospects, providing individuals without work with this technology equips them with an essential tool for navigating the job market more effectively and increasing their chances of finding employment." So it is **feasible** to revise the argument to overcome the objection situation.

Example 2

Argument:
If the government provides those without work with a mobile phone, they will be able to find themselves an occupation in order to live and survive. When people without jobs have hand phones with internet access, they can browse the net for more job opportunities. They can do research on the work they have found and prepare themselves for the job. Mobile phones can be used to make calls with the companies they would like to work with.

Model generated objection situation:
By providing mobile phones to the unemployed, the government would inadvertently contribute to the overloading of cellular networks. Since a significant portion of the population without work would suddenly have access to mobile phones, it would strain the existing infrastructure, leading to network congestion and poor service quality for all users.

Analysis:

Figure 7: First half of the annotation template for evaluating the rationality and feasibility of objection situations.

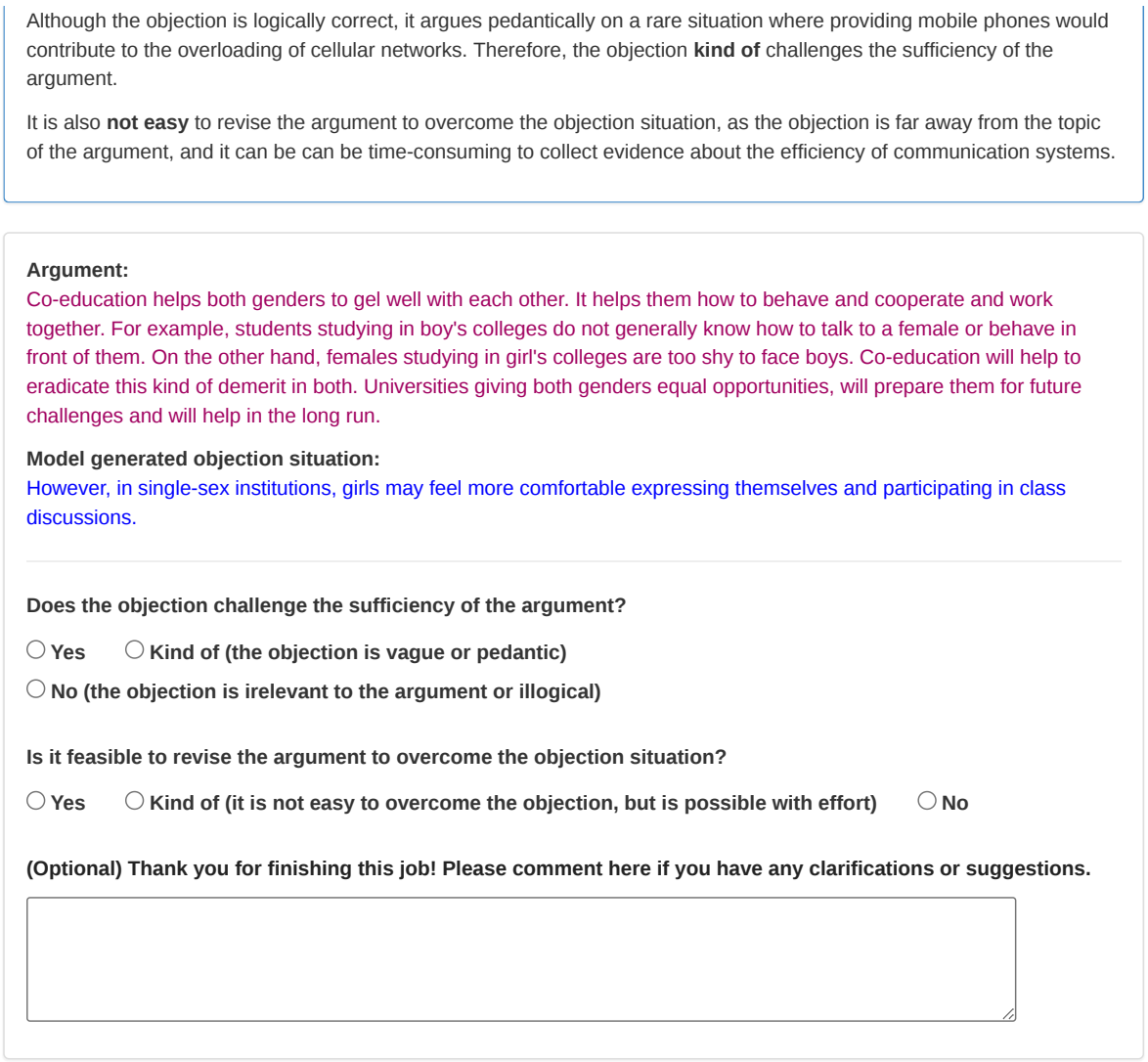

Figure 7: Second half of the annotation template for evaluating the rationality and feasibility of objection situations.

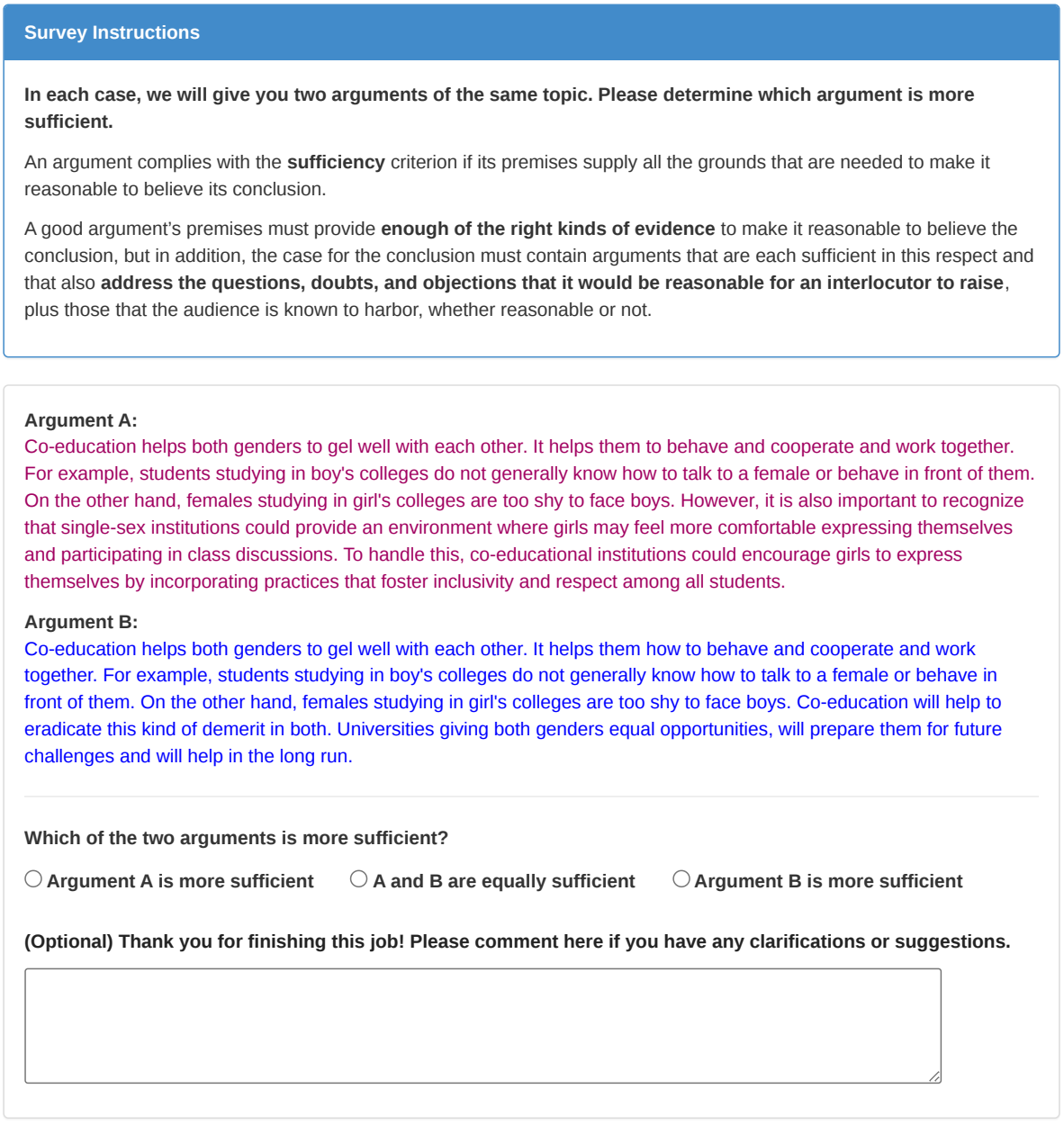

Figure 8: Annotation template for comparing the original and revised arguments.

Survey Instructions

In each case, we will give you an original argument, an objection situation that provides reasons arguing against the argument, and a revised argument.

Please assess whether the revised argument **addresses the objection situation's concern** and whether it **preserves the main content of the original argument**.

Original Argument:
Co-education helps both genders to gel well with each other. It helps them how to behave and cooperate and work together. For example, students studying in boy's colleges do not generally know how to talk to a female or behave in front of them. On the other hand, females studying in girl's colleges are too shy to face boys. Co-education will help to eradicate this kind of demerit in both. Universities giving both genders equal opportunities, will prepare them for future challenges and will help in the long run.

Objection Situation:
However, in single-sex institutions, girls may feel more comfortable expressing themselves and participating in class discussions.

Revised Argument:
Co-education helps both genders to gel well with each other. It helps them to behave and cooperate and work together. For example, students studying in boy's colleges do not generally know how to talk to a female or behave in front of them. On the other hand, females studying in girl's colleges are too shy to face boys. However, it is also important to recognize that single-sex institutions could provide an environment where girls may feel more comfortable expressing themselves and participating in class discussions. To handle this, co-educational institutions could encourage girls to express themselves by incorporating practices that foster inclusivity and respect among all students.

Does the revised argument address the objection situation's concern?

○ **Yes** ○ **No**

Does the revised argument preserve the main content of the original argument?

○ **Yes** ○ **No**

(Optional) Thank you for finishing this job! Please comment here if you have any clarifications or suggestions.

Figure 9: Annotation template for checking the revision quality of GPT-4.

Instruction:

This AI is identifying whether arguments are sufficient, capturing whether an argument's premises together make it rationally worthy of drawing its conclusion. The AI responds with 'Sufficient' or 'Insufficient' as appropriate. If the argument is insufficient, the AI also generates an objection situation to show the insufficiency.

Format:

Judgement: Sufficient or Insufficient

Objection Situation (if insufficient): Describe a specific situation that challenges the sufficiency of the argument. Do not include any explanation.

Input:

In a positive point of view, when people without jobs have hand phones that have access to the Internet, they will be able to browse the net for more job opportunities. For example, they can surf the The Star Online's work section to find a job that is suitable for them. With the help of the net, they can also do more research on the work that they have found apart from looking up on how they can prepare themselves for the job. Not only that, the mobile phones can also be used to make calls with the companies in which they would like to work with. In short, if the government provides those without work with a mobile phone, they will be able to find themselves an occupation in order to live and survive.

Response:

Judgement: Insufficient

Objection Situation: However, having a mobile phone with internet access does not guarantee that they will find a job, as there may be other factors such as a lack of available positions, a mismatch in skills, or a highly competitive job market.

Input:

Co-education helps both genders to gel well with each other. It helps them how to behave and cooperate and work together. For example, students studying in boy's colleges do not generally know how to talk to a female or behave in front of them. On the other hand, females studying in girl's colleges are too shy to face boys. Co-education will help to eradicate this kind of demerit in both. Universities giving both genders equal opportunities, will prepare them for future challenges and will help in the long run.

Response:

Table 13: Example prompt for directly prompting LLMs to generate objection situations.