Exploring Defeasibility in Causal Reasoning

Shaobo Cui, Lazar Milikic, Yiyang Feng, Mete Ismayilzada,
Debjit Paul, Antoine Bosselut, Boi Faltings
EPFL, Switzerland
`{firstname.lastname}@epfl.ch`

Abstract

Defeasibility in causal reasoning implies that the causal relationship between cause and effect can be strengthened or weakened. Namely, the causal strength between cause and effect should increase or decrease with the incorporation of strengthening arguments (supporters) or weakening arguments (defeaters), respectively. However, existing works ignore defeasibility in causal reasoning and fail to evaluate existing causal strength metrics in defeasible settings. In this work, we present δ-CAUSAL, the first benchmark dataset for studying defeasibility in causal reasoning. δ-CAUSAL includes around 11K events spanning ten domains, featuring defeasible causality pairs, namely, cause-effect pairs accompanied by supporters and defeaters. We further show that current causal strength metrics fail to reflect the change of causal strength with the incorporation of supporters or defeaters in δ-CAUSAL. To this end, we propose CESAR (Causal Embedding aSsociation with Attention Rating), a metric that measures causal strength based on token-level causal relationships. CESAR achieves a significant 69.7% relative improvement over existing metrics, increasing from 47.2% to 80.1% in capturing the causal strength change brought by supporters and defeaters. We further demonstrate even Large Language Models (LLMs) like GPT-3.5 still lag 4.5 and 10.7 points behind humans in generating supporters and defeaters, emphasizing the challenge posed by δ-CAUSAL.

1 Introduction

Causality (Pearl, 2009; Pearl and Mackenzie, 2018), a fundamental concept of artificial intelligence, describes the relationship between two events where one event, namely the *cause*, results in the occurrence of another event, namely the *effect*. Understanding causality enhances decision-making in various areas such as medicine (Kuipers and Kassirer, 1984; Michoel and Zhang, 2022), disease treatment (Rizzi, 1994; Me and Struchiner, 1995), finance (Koonce et al., 2011; Tiffin, 2019), and law (Foot, 1963; Liu et al., 2021).

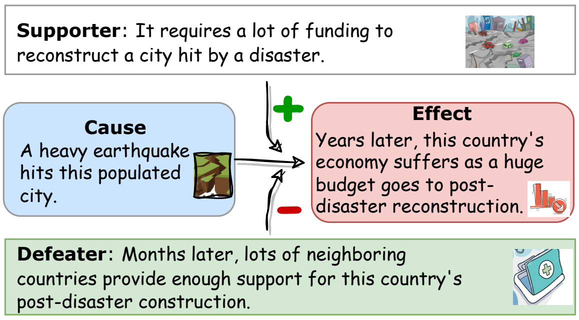

Supporter: It requires a lot of funding to reconstruct a city hit by a disaster.
+
Cause
A heavy earthquake hits this populated city.
Effect
Years later, this country's economy suffers as a huge budget goes to post-disaster reconstruction.
−
Defeater: Months later, lots of neighboring countries provide enough support for this country's post-disaster construction.

Figure 1: A motivational example of defeasibility in causal reasoning. It consists of a cause-effect pair, a supporting argument (supporter), and an opposing argument (defeater) for the causal relationship.

Despite the importance of causality, establishing a definite causal relationship between two events is inherently challenging as *uncertainties* can influence the strength of the causality between the cause and the effect. As the time interval between the cause and the effect widens, these uncertainties tend to increase. These uncertainties include incomplete or unseen factors, changing situations, and contextual information. However, existing works on causal reasoning (Qin et al., 2019; Feng et al., 2021; Zhang et al., 2022; Wang et al., 2023) mainly focus on definite causality while ignoring the uncertainties inherent in causal relationships.

Motivated by this blank, we introduce the concept of defeasibility in causal reasoning. Formally, defeasibility in causal reasoning refers to situations wherein the causal relationship between the cause and the effect is justified, but supplementary information might strengthen or weaken these justifications. As the example shown in Figure 1, a supporter argument that "It requires a lot of funding to reconstruct · · · " strengthens (+) the causal relationship between "the earthquake" and "the country's economic decline". On the other hand, a defeater

Findings of the Association for Computational Linguistics: ACL 2024, pages 6433–6452
August 11-16, 2024 ©2024 Association for Computational Linguistics

argument that “Months later, lots of neighboring countries provide enough support $\cdots$” weakens (−) the causal relationship. With the capacity to understand defeasibility, humans can clearly perceive the change in the causal strength brought by supporters and defeaters.

However, prior benchmarks (Roemmele et al., 2011; Ning et al., 2018; Sloman, 2005) on causal reasoning mostly focus on definite relationships, often overlooking the **defeasibility** of cause-effect pairs when uncertainties arise. To bridge this gap, we contribute the first benchmark for investigating defeasibility in causal reasoning: δ-CAUSAL. As depicted in Figure 1, each sample in δ-CAUSAL consists of a cause-effect pair, accompanied by its supporter argument (A) and defeater argument (D). A and D reinforce and undermine the causal relationship between the cause and the effect, respectively. We construct δ-CAUSAL from examples across ten domains: environment, business, science, health, work, politics, education, sports, entertainment, and travel, and use our new benchmark to test how well existing large language models (LLMs) can generate supporters and defeaters for causal pairs. Our experiments reveal that state-of-the-art pre-trained models, including GPT-3.5, lag behind humans by up to 4.5 and 10.7 points in generating correct supporters and defeaters, respectively, which emphasizes the significant challenges brought by δ-CAUSAL.

Furthermore, due to the lack of appropriate benchmarks, it is difficult to determine whether existing metrics on qualifying causal strength can capture the change of causal strength brought by supporters and defeaters or not. An ideal metric should reflect the increase (or decrease) in causal strength with the incorporation of supporters (or defeaters). With the presence of supporters and defeaters, δ-CAUSAL serves as an ideal touchstone for assessing the efficacy of existing metrics in capturing the causal strength change brought by supporters and defeaters. Our experiments demonstrate that existing cutting-edge metrics like ROCK (Zhang et al., 2022) and CEQ (Du et al., 2022), whose accuracy are both below 50%, fail to accurately capture the causal strength change.

To address this limitation, we propose a robust and versatile metric for measuring causal strength, known as CESAR, based on Causal Embedding aSsociation with Attention Rating (CESAR). CESAR builds upon a transformer-based model (Devlin et al., 2019) with causal embeddings. The causal strength given by CESAR is calculated as a weighted aggregation of token-level causal strength, which is the association score between a token’s causal embedding in the cause and its counterpart in the effect. The learned weighted coefficients guide CESAR to prioritize strong causal pairs like “fire” and “burn”. From the experimental results, CESAR achieves a significant 69.7% improvement in quantifying changes in causal strength resulting from supporters and defeaters. It also attains state-of-the-art performance with an 11.9% improvement in distinguishing the correct hypotheses from incorrect ones, underscoring CESAR’s versatility in various causal tasks.

In summary, we make four key contributions:

- We contribute δ-CAUSAL, the pioneering benchmark that emphasizes the often overlooked aspect of causal reasoning: defeasibility. It paves the road to systematically exploring defeasibility in causal reasoning. δ-CAUSAL is available at https://github.com/cui-shaobo/defeasibility-in-causality.

- With the presence of the supporters and defeaters, δ-CAUSAL serves as a valuable yardstick for evaluating existing metrics on causal strength. We highlight the limitations of current causal strength metrics in capturing the changes in causal strength resulting from supporters and defeaters.

- We propose CESAR, a robust and versatile metric for measuring causal strength. CESAR outperforms existing metrics like ROCK and CEQ, exhibiting a remarkable 69.7% improvement in capturing the changes of causal strength brought by supporters and defeaters.

- Using δ-CAUSAL, we assess the ability of existing LLMs to comprehend defeasibility in causal reasoning. The results show that even GPT-3.5 falls significantly short, lagging behind humans by 4.5 and 10.7 points in generating accurate supporters and defeaters, respectively. This underscores the significant challenges brought by δ-CAUSAL.

2 Related Work

Comparison of δ-CAUSAL with Related Datasets. We present the comparison between δ-CAUSAL and related datasets in Table 1. Most commonsense

	Annotation Unit	Size	Always Causality	#Causality pairs	Defeater	Supporter	Touchstone for causal strength metrics
				Commonsense causal reasoning datasets			
COPA (Roemmele et al., 2011)	Sentence	1,000	■	1	□	□	□
TCR(Ning et al., 2018)	Sentence	172	■	1	□	□	□
e-CARE(Du et al., 2022)	Sentence	21,324	■	1	□	■	□
				Counterfactual commonsense reasoning datasets			
ART(Bhagavatula et al., 2020)	Sentence	20,000	□	1	□	□	□
TimeTravel(Qin et al., 2019)	Paragraph	29,849	□	2	□	□	□
				Defeasible inference datasets			
δ-NLI (Rudinger et al., 2020)	Sentence	9,986	□	N/A	■	■	□
δ-CAUSAL	Sentence	11,245	■	2	■	■	■

Table 1: Comparison of δ-CAUSAL and related datasets. ■ means supported and □ means not.

causal reasoning datasets like COPA (Roemmele et al., 2011) and TCR (Ning et al., 2018) focus on definite causality, ignoring the uncertainties inherent in the causal relationship. δ-CAUSAL introduces defeasibility to cover uncertainties, enhancing its capacity to test models in defeasible causal reasoning. Counterfactual datasets like TimeTravel (Qin et al., 2019) and ART (Bhagavatula et al., 2020) often lack consistently valid causal relationships, presenting causality pairs based on counterfactual events and lack the capacity to test the performance of existing causal strength metrics. In contrast, δ-CAUSAL incorporates both supporters and defeaters and thus makes itself an idea touchstone for causal strength metrics. Lastly, while Rudinger et al. (2020) define defeasible inference in natural language without always implying causality, δ-CAUSAL emphasizes the causal relationship's defeasibility.

Existing Evaluation Metrics on Causal Strength. Previous literature (Luo et al., 2016b; Du et al., 2022; Zhang et al., 2022) study the causal strength from different perspectives. Du et al. (2022) propose a metric named Causal Explanation Quality (CEQ) score based on word co-occurrence to measure if a given explanation could increase the causal strength between the cause and the effect. Zhang et al. (2022) propose a theoretical framework named ROCK to measure the causal strength from the causal inference perspective. Details about existing causal strength metrics are present in Appendix E.

3 Task

Research Questions. In this paper, we study two research questions to understand defeasibility:

- *Research Question I*: How to estimate the strength of causality in the setting of defeasible causal reasoning? Specifically, can existing metrics on causal strength accurately capture the changes brought by supplementary information like supporters or defeaters in defeasible causal reasoning?

- *Research Question II*: Can language models generate correct defeasible arguments for given causal facts that can make the causality less justified or more justified? Specifically, we explore whether existing models can generate supporters or defeaters correctly.

We answer Research Question I in §5 and Research Question II in §6.

Estimating Causal Strength for Studying Research Question I. The causal strength between event C and event E, denoted as $\mathcal{CS}(C \to E)$, falls in $[0, 1]$. It measures the intensity of the event C causing/leading to the occurrence of event E. We present the overall causal relationship of defeasible causal reasoning in Figure 2.

In the context of defeasible causal reasoning, an ideal metric on causal strength should meet the following requirements: (i) the estimated causal strength given by this metric should increase with the incorporation of supporters; (ii) the estimated causal strength given by this metric should decrease with the incorporation of defeaters.

Supporter and Defeater Generation for Studying Research Question II. Defeasibility is a fundamental concept in many fields. In legal reasoning, defeasibility means a legal principle can be overridden by a competing principle. Defeasibility in causal reasoning implies that the validness of causality can be less or more justified by additional information like supporters and defeaters. With the definition of causal strength, the defeasibility in commonsense causal reasoning is represented as

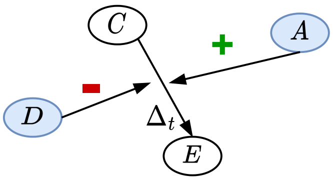


Figure 2: The overview of the causal relationships in δ-CAUSAL. The symbol + indicates that the supporter A strengthens the causal relationship between the cause C and the effect E, while − signifies that the defeater D weakens this relationship. The variable Δ_t denotes the time interval between the cause C and the effect E. To ensure the practicality of identifying defeaters, this time interval is set to long time durations.

two constraints:

$$\begin{cases} \mathcal{CS}(C \rightarrow E) - \mathcal{CS}((C \oplus A) \rightarrow E) < 0 \\ \mathcal{CS}(C \rightarrow E) - \mathcal{CS}((C \oplus D) \rightarrow E) > 0 \end{cases} \quad (1)$$

where $\oplus$ means the combination of two events. A and D represent supporters and defeaters, respectively. The first constraint requires that the causal strength between the cause and the effect is strengthened by the supporter, while the second constraint requires that the defeater should weaken the causal strength.

Given the cause and the effect, we ask the model to generate a supporter A or a defeater D that reinforces or diminishes the causal relationship between C and E as much as possible. Namely,

$$\begin{cases} A = \arg\max_A [\mathcal{CS}((C \oplus A) \rightarrow E) - \mathcal{CS}(C \rightarrow E)] \\ D = \arg\max_D [\mathcal{CS}(C \rightarrow E) - \mathcal{CS}((C \oplus D) \rightarrow E)] \end{cases} . \quad (2)$$

4 δ-CAUSAL

4.1 Overview of δ-CAUSAL

Each instance in δ-CAUSAL consists of four components: (1) a domain label from 10 domains including Environment, Business, Science/Technology, Health, Work, Politics, Education, Sports, Entertainment, and Travel; (2) a cause-effect pair that is presented with a cause, its effect, and the time interval between the cause and the effect; (3) a defeater argument that reduces the validness of or totally invalidates the causal relationship between the cause and the effect; (4) a supporter argument that makes the causal relationship between the cause and the effect more justified.

4.2 Annotation of δ-CAUSAL

Figure 3 illustrates the data annotation and refinement pipeline. Initially, we gather keywords for each domain, which guide the annotation of cause-effect pairs on Amazon Mechanical Turk (AMT). We also collect supporters and defeaters for these pairs on AMT. Each annotation step is paired with a refinement phase.

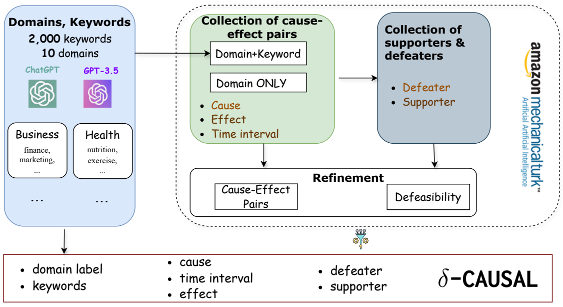


Figure 3: Pipeline of the annotation and refinement procedures of δ-CAUSAL.

Phase I: Annotation of Cause-Effect Pairs. *Without Keyword Hints.* Annotators begin by selecting from ten domains, as detailed in Figure 4. Within the chosen domain, they first annotate a cause-effect pair. To ensure the feasibility of identifying defeaters, we impose restrictions on the selection of time intervals to relatively long-term periods. This practice is taken because short time intervals often make it challenging for annotators to identify and annotate potential defeating events effectively. Longer time intervals provide a broader temporal scope, allowing for the observation and annotation of more complex interactions and changes that might influence or negate the initial cause-effect relationship. This methodological choice enhances the richness and reliability of the annotated data by capturing a wider range of possible outcomes and influences over extended periods. Specifically, annotators must specify a time interval for the effect, with options including `months later`, `years later`, `decades later`, and `centuries later`. For more details on time labels, see Appendix C.2. *With Keyword Hints.* Our AMT data collection revealed limited topic variety within domains without keyword hints. For example, the Health domain often linked exercise to weight loss. To diversify our benchmark, we used GPT-3.5 to generate 200 keywords per domain (100 each from text-davinci-003 and ChatGPT) listed in Appendix C.3. Annotators then receive a keyword as a hint and craft a related cause-effect pair within the domain.

Phase II: Annotation of Supporters and Defeaters. We ask annotators to write the supporter

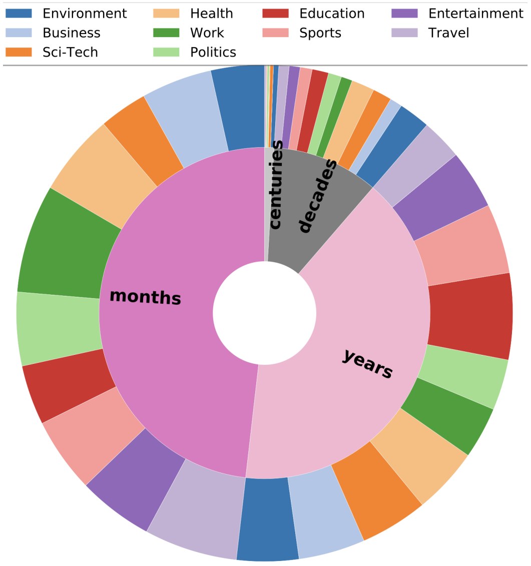


Figure 4: The distributions of time intervals and domains in δ-CAUSAL. Different colors represent different time intervals (inner circle) and domains (outer circle). Detailed values and proportions are provided in Appendix C.2.

and the defeater simulataneously. The supporter can be a conceptual explanation or fact that supports the causal relationship, while the defeater should provide evidence for the opposite thesis of the effect or evidence that undercuts the effect (Pollock, 1987). A defeater can also be an event or a circumstance that makes the causality between the cause and the effect unjustified or less justified.

4.3 Refinement of δ-CAUSAL

We enhance the benchmark quality through systematic refinement stages. Initially, we eliminate samples written randomly: several annotated examples either contain repetitive wording or merely echo the instructions. Subsequently, we undertake two Mechanical Turk (AMT) refinement phases: refining cause-effect pairs and refining supporters and defeaters.

Phase III: Cause-Effect Pair Refinement. We task annotators to assess the validity and timing of the cause-effect relationships. Each cause-effect pair is judged by three annotators. To ensure the quality, each assignment includes a gold cause-effect pair with a known label. Any assignments whose gold examples are incorrectly labeled are disregarded. We retain annotations if: (i) No annotation is discarded and a majority deem it true; (ii) one annotation is discarded for misjudging the gold example, while the remaining two validate the evaluated sample.

Phase IV: Refinement of Supporters and Defeaters. Three annotators determine whether the supporter/defeater enhances or diminishes the causal connection. Each task has a defeasibility pair with a known label and another for assessment. Similarly, as in Phase III, we only keep the assignments whose majority of the filtered votes are true.

4.4 Overall Quality of δ-CAUSAL

In order to assess the quality of δ-CAUSAL, we randomly select 200 samples from δ-CAUSAL and ask three NLP experts (see details in Appendix D.3) to assess the validity of these samples from the following three perspectives: validness of causality, supporter, and defeater. The assessment result is shown in Table 6. We achieve an average accuracy Accuracy $\geq 92\%$ with a high agreement. This shows that δ-CAUSAL is of good quality.

4.5 Statistics of δ-CAUSAL

The statistics of δ-CAUSAL are shown in Table 2. More details are in the Appendix. Specifically, Appendix A is the qualification rules for annotators. Appendix B elaborates the gold examples and guidelines we use during dataset collection. Appendix C is more about statistics of δ-CAUSAL.

	Statistics
Overall	
# Causality pairs	8,080
# Supporters	11,245
# Defeaters	11,245
Train/Dev/Test	7,000/2,276/1,969
Length of utterances	
Average length of causes	9.50
Average length of effects	9.60
Average length of supporters	9.10
Average length of defeaters	10.39

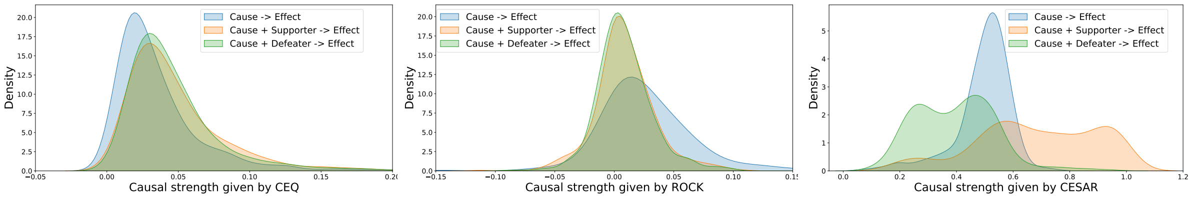
Table 2: Statistics of δ-CAUSAL. More details about δ-CAUSAL are shown in Appendix C.

5 Estimating Causal Strength

In this section, we address Research Question I on causal strength. We first highlight the limitations of current metrics in § 5.1. Then, we detail the definition, comparison with other metrics, versatility, and case study of CESAR from § 5.2 to § 5.5.

5.1 Limitations of Existing Metrics

δ-CAUSAL works as a solid touchstone to test existing metrics for evaluating causal strength. As



(a) Shift of CEQ distribution.

(b) Shift of ROCK distribution.

(c) Shift of CESAR distribution.

Figure 5: The shifts in causal strength distributions facilitated by CEQ (left), ROCK (middle), and CESAR (right) with the incorporation of supporters and defeaters are illustrated in δ-CAUSAL. These curves utilize kernel density estimation (Parzen, 1962) to depict the data distribution as a continuous probability density curve. Notably, only CESAR effectively captures the variations in causal strength triggered by the inclusion of supporters and defeaters; specifically, the causal strength distribution shifts to the right with supporters and to the left with defeaters.

mentioned earlier, in δ-CAUSAL, a supporter is expected to increase the causal strength between the cause and the effect, while a defeater is expected to decrease the causal strength, as depicted in Equation (1).

	Supporter	Defeater	Geometric mean
CEQ	83.1	17.5	38.1
ROCK	32.5	68.6	47.2
CESAR (ours)	**84.6**	**75.8**	**80.1**

Table 3: Accuracy of causal strength metrics on δ-CAUSAL: For supporters, correct predictions occur when the metric assigns a higher score to the cause-supporter combination. For defeaters, predictions are correct if the metric assigns a lower score to the cause-defeater combination. The geometric mean is calculated based on the accuracy of supporters and defeaters.

The accuracy of metrics on δ-CAUSAL for capturing causal strength changes by supporters and defeaters is detailed in Table 3. CEQ sees 83.1% of supporters as strengtheners, but wrongly views 82.5% of defeaters as such. Conversely, ROCK incorrectly sees 67.5% of supporters as weakeners. These results highlight the limitations of existing metrics and emphasize the need to develop more robust evaluation metrics for causal strength.

5.2 CESAR: Causal Embedding ASsociation with Attention Rating on Causal Strength

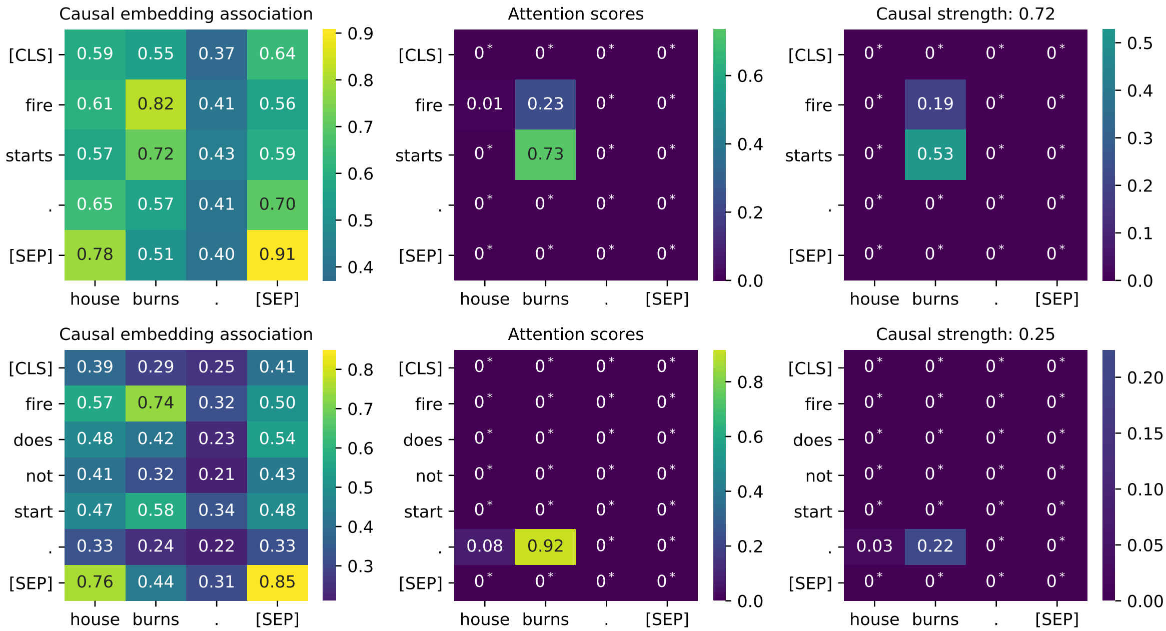
Motivation: Causality-aware Embedding and Attention. The causal strength between two events can be quantified as the weighted aggregation of the causal relationship between tokens within these events (Luo et al., 2016b). For instance, in events "Fire starts" and "House burns", "fire" and "burns" drive a causal relationship. Inspired by BERTScore (Zhang et al., 2020), we fine-tune BERT embeddings to capture token-level causality so that tokens delivering a strong causal relationship, like "fire" and "burns", have embeddings that are highly associated. The motivation for the attention mechanism is that we wish to place less attention on pairs that consist of causality irrelevant words(e.g., stop words) and more attention on pairs involved in a strong causal relationship like "fire" and "burns".

Attention Computation. On top of the BERT model, we add a customized attention layer that identifies the important token pairs for evaluating the causal strength. The attention scores for token pairs are calculated through a specifically adjusted cross-attention layer on the top of the BERT model. Let d be the dimension of BERT model; C and E be tokenized to n and m tokens respectively, i.e., $\mathbf{C} \in \mathbb{R}^{n\times d}$ and $\mathbf{E} \in \mathbb{R}^{m\times d}$. We compute query vectors as $\mathbf{Q} = \mathbf{CW}_q$ and key vectors as $\mathbf{K} = \mathbf{EW}_k$, where $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d\times d}$. Then, the matrix of attention scores for token pairs is calculated as $\mathbf{A} = \text{softmax}\left(\mathbf{QK}^T\right)$, where $\mathbf{A} \in \mathbb{R}^{n\times m}$, and softmax is performed over all values of the matrix, i.e., $\text{softmax}(A_{ij}) = \frac{\exp(A_{ij})}{\sum_{i,j}\exp(A_{ij})}$.

Weighted Average of Causal Embedding Association. We propose the following formula for computing the causal strength between C and E:

$$\mathcal{CS}(C \rightarrow E) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\frac{|c_i^T e_j|}{\|c_i\|\|e_j\|} \tag{3}$$

where $c_1, c_2, \ldots, c_n$ and $e_1, e_2, \ldots, e_m$ are causal embeddings of tokens in C and E, respectively. These embeddings are generated by the last hidden layer of fine-tuned BERT. The weight a_{ij} is the attention put on each token pair of c_i and e_j such that $\sum_{i,j} a_{ij} = 1$. We compute the absolute cosine similarity between causal embeddings for each pair of tokens and calculate the score as



* The number has been capped at 0 since it is less than 0.01.

Figure 6: We define three matrices based on Equation (3). (i) Causal Embedding Association Matrix ($\mathbf{M}$) in the left column: M_{ij} measures the causal association between embeddings c_i and e_j as $M_{ij} = \frac{|c_i^T e_j|}{\|c_i\|\|e_j\|}$. (ii) Causality-Aware Attention Matrix ($\mathbf{A}$) in the middle column: $\mathbf{A}$ stores the attention weights of each token pair given by $A_{ij} = a_{ij}$. (iii) Causal Strength Matrix ($\mathbf{S}$) in the right column: $\mathbf{S}$ is the Hadamard product between $\mathbf{M}$ and $\mathbf{A}$. That is, $S_{ij} = M_{ij} A_{ij} = a_{ij} \frac{|c_i^T e_j|}{\|c_i\|\|e_j\|}$. The upper row of this figure shows the values of $\mathbf{M}$, $\mathbf{A}$, and $\mathbf{S}$ derived from the cause-effect pair wherein the cause is "Fire starts." and the effect is "House burns." Conversely, the lower row presents the values of $\mathbf{M}$, $\mathbf{A}$, and $\mathbf{S}$ when the cause is "Fire does not start." and the effect is "House burns.". The y-axis denotes cause tokens, and the x-axis represents effect tokens. The causal strength, calculated from Equation (3), is shown in the title of the rightmost matrices for both pairs.

the average of these token-level causal associations weighted by the learned attention.

Training Procedure. We train the CESAR metric on the augmented e-CARE dataset (Du et al., 2022). It contains cause-effect pairs with a conceptual explanation designed to increase the pair's causal strength. Specifically, we set $\mathcal{CS}(C \rightarrow E)$ to 0.7 and $\mathcal{CS}(C \oplus H \rightarrow E)$ to 1.0 where H is the explanation for the causal relationship between C and E. For pairs with no causal relationship, we set the causal strength to 0.0. Further, we use ChatGPT to generate opposites of the conceptual explanations provided in e-CARE and set $\mathcal{CS}(C \oplus \neg H \rightarrow E)$ to 0.2 where $\neg H$ is an opposite of the conceptual explanation for C and E. We train CESAR for 4 epochs with AdamW optimizer (Loshchilov and Hutter, 2019) with a linear scheduler, learning rate 1e-5, and MSE loss.

Experimental Results. The experimental results are shown in Table 3. It is evident that CESAR outperforms CEQ and ROCK significantly on both supporters and defeaters, achieving an accuracy of 84.6% and 75.8%, respectively. Furthermore, we use the geometric mean of the accuracy achieved in supporters and defeaters as the index of the overall performance for each metric. From Table 3, we can see that both CEQ and ROCK attain a geometric mean accuracy of less than 50%, In contrast to this poor performance, our proposed CESAR obtains a superior accuracy of 80.1%.

5.3 Shift of Causal Strength Distributions with Supporters and Defeaters

In Figure 5, we plot the shift of causal strength distribution with the incorporation of supporters and defeaters for CEQ, ROCK, and CESAR (ours). For CEQ in Figure 5a, both supporter and defeater distributions shift rightward. This means CEQ perceives any supplementary information as supporting the original cause-effect relationship, regardless of its actual impact on causal strength. For ROCK in Figure 5b, both distributions lean left, suggesting ROCK sees all supplementary information as opposing the original cause-effect relationship.

For CESAR, the supporter distribution shifts right while the defeater distribution shifts left. This is the anticipated behavior for a good causal strength metric, capturing the contrasting effects of supporters and defeaters.

5.4 Versatility of CESAR

To additionally investigate the versatility of CESAR, we test CESAR on COPA (Roemmele et al., 2011) and show the results in Table 4. The accuracy reflects whether the tested causal strength metric gives a larger value to the causal strength of true cause-effect pairs than that of the false cause-effect pairs. CESAR achieves an accuracy of 70.7%, once again significantly outperforming both ROCK and CEQ. This proves the generalness of CESAR in estimating the causal strength.

Metrics	CEQ	ROCK	CESAR (ours)
Accuracy	57.8	63.2	**70.7**

Table 4: Results of CESAR's versatility, which is about distinguishing the true cause-effect pairs from the false cause-effect pairs on COPA (Roemmele et al., 2011).

5.5 Case Study of CESAR

Figure 6 illustrates the adaptation of values in Equation (3) when the cause is altered. In the upper row with inputs C = "Fire starts." and E = "House burns.", the token causal embeddings of "fire" and "burns" have a high association score of 0.82. Also "starts" and "burns" demonstrate a strong association score of 0.72. Interestingly, a notable amount of attention is paid to the latter pair, whose association score is a key determinant of the causal strength score. Using Equation (3), we obtain a causal strength of 0.72, which signifies a strong causal relationship between C and E. In the lower row, with a modified cause C = "Fire does not start." and the same effect E, the token pair associations of "fire" and "burns", and "start" and "burns" remain high. However, CESAR's attention mechanism adjusts the importance of tokens in terms of the causal relationship between two sentences in the right direction. Namely, the causal strength undergoes a reduction of over 65%, resulting in a score of 0.25 that indicates a weak causal relationship.

Details on setup, score computation, preparation of the training data, and an extensive ablation study of CESAR are in Appendix F.

6 Supporter and Defeater Generation

In this section, we answer Research Question II about existing SOTA models' ability in supporter and defeater generation by extensive experiments.

Model	BLEU	METEOR	ROUGE-L	CIDEr	BERT-Score
			Supporter generation		
BART	7.71	12.90	16.72	0.397	54.0
T5	6.92	11.89	15.94	0.360	52.5
T5-large	7.90	12.55	17.27	0.440	54.2
GPT-2	6.62	11.81	14.95	0.357	52.4
GPT-3.5	3.17	10.97	9.93	0.094	48.0
			Defeater generation		
BART	7.53	11.15	16.63	0.345	51.8
T5	6.83	10.89	15.83	0.279	51.5
T5-large	7.37	10.90	16.48	0.325	52.1
GPT-2	6.71	10.38	15.32	0.257	50.9
GPT-3.5	5.24	10.86	15.27	0.205	50.0

Table 5: Results of supporter and defeater generation with different language models.

Setup. We finetune generative pre-trained language models BART (Lewis et al., 2020), T5 (Raffel et al., 2020), T5-large, and GPT-2 (Radford et al., 2019). These models take the concatenation of the cause and the effect as the input. The output is the supporter or the defeater. See details about the baselines and experimental setup in Appendix D. We automatically evaluate the generated supporters/defeaters using BLEU ($n = 2$) (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and BERT-Score (Zhang et al., 2020).

Results and Analysis. The results of the supporter and defeater generation are shown in Table 5. It shows that existing LLMs perform unsatisfyingly in generating supporters and defeaters. Notably, we see that even GPT-3.5 (one-shot performance) does not achieve a high score based on existing generation metrics. This implies that existing metrics cannot adequately address the evaluation objective. Motivated by this, we employ a human evaluation to assess the quality of the generated supporters and defeaters from humans and GPT-3.5.

	% validness	% agreement
	Quality of cause-effect pairs	
	94.67	89.00
	Quality of supporters	
Human	92.67	83.00
GPT-3.5	88.17	73.50
	Quality of defeaters	
Human	94.50	81.50
GPT-3.5	83.83	69.50

Table 6: Comparison between humans and GPT-3.5 on supporter and defeater generation by human evaluation.

Comparison between Humans and GPT-3.5 on Defeater and Supporter Generation by Human Evaluation. Three NLP experts (details in Appendix D.3) are asked to judge the validness of the generated supporters and defeaters from GPT-3.5 and annotators. From the results in Table 6, we can observe that existing large pre-trained models can not generate supporters and defeaters well. Even the large language model GPT-3.5 still lags behind humans by 4.5 points, i.e., 88.17% (GPT-3.5) vs. 92.67% (humans), in generating credible supporters. Even worse, it lags 10.7 points behind humans, i.e., 83.83% (GPT-3.5) vs. 94.50% (humans) in generating plausible defeaters. Sometimes, the model negates the effect rather than providing supplementary information to make it less justified. This demonstrates the challenges posed by δ-CAUSAL.

7 Conclusions

Our paper introduces δ-CAUSAL, a pioneering benchmark that focuses on the often overlooked aspect of causal reasoning: defeasibility. Even state-of-the-art models like GPT-3.5 fall significantly short compared to human performance in understanding defeasibility shown by δ-CAUSAL. We further demonstrate the limitations of current causal strength metrics in capturing causal strength changes brought by supporters and defeaters. To circumvent these limitations, we propose CESAR, a robust metric that outperforms existing measures by a remarkable 69.7% improvement in capturing these changes. Our research contributes to the advancement of causal reasoning by emphasizing defeasibility and providing valuable insights for improving language models' understanding of nuanced causal relationships. This work establishes a foundation for future studies on developing more advanced defeasible causal reasoning systems.

Acknowledgements

We are grateful to Ping, Wanhao, and Mathieu for their invaluable assistance and inspiring discussions related to the writing and experimental design of this paper. We also gratefully acknowledge the financial and IT support provided by EPFL. AB gratefully acknowledges the support of the Swiss National Science Foundation (No. 215390), Innosuisse (PFFS-21-29), the EPFL Science Seed Fund, the EPFL Center for Imaging, Sony Group Corporation, and the Allen Institute for AI.

Limitations

Despite our work's significant contributions, such as providing the first benchmark dataset for defeasible causal reasoning and introducing a novel causal strength metric, several acknowledged limitations still need to be addressed. First and foremost, the causal strength change resulting from supporters and defeaters is currently described qualitatively rather than quantitatively. This limitation hinders the quantitative application of δ-CAUSAL as it becomes challenging to precisely assess the exact magnitude of the causal strength change caused by supporters and defeaters. Consequently, it is difficult to use δ-CAUSAL to quantify and measure the precise impact of these factors on causal reasoning. Secondly, the domains covered by δ-CAUSAL remain limited in scope. Expanding the applicability of δ-CAUSAL to include other domains such as medicine or chemistry would enhance its versatility and make it more relevant to a broader range of research and practical applications. By incorporating additional domains in the future, we can evaluate the performance and effectiveness of δ-CAUSAL in various contexts, ensuring its robustness and generalizability. In conclusion, while our work has made notable contributions, it is essential to address these known limitations to enhance the quantitative usage and domain coverage. By doing so, we can advance the field of defeasible causal reasoning and strengthen the practical utility of our proposed metric on causal strength.

Ethical Considerations

We foresee no major ethical concerns for this work. As we know, causality contains various aspects of daily life. Bad things lead to negative results. But we take the following steps to make sure that the δ-CAUSAL contains harmful/toxic content as little as possible. Firstly, a clear and understandable guide is given for annotations. After that, all of these annotated examples are followed up with a refinement process to filter out bad examples. Finally, we manually check whether the annotations contain keywords that convey harmful, toxic, violent, or erotic meanings. However, we acknowledge that these steps are not perfect. δ-CAUSAL is under MIT License. Our paper involves other datasets, including e-CARE (Du et al., 2022), which is under MIT License, and COPA (Roemmele et al., 2011), which is under BSD 2-Clause License.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Fuli Feng, Jizhi Zhang, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Empowering language understanding with counterfactual reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2226–2236, Online. Association for Computational Linguistics.

Philippa Foot. 1963. Hart and honoré: Causation in the law. *The Philosophical Review*, 72(4):505–515.

Miguel A Hernán and James M Robins. 2010. Causal inference.

Lisa Koonce, Nicholas Seybert, and James Smith. 2011. Causal reasoning in financial reporting and voluntary disclosure. *Corporate Governance: Disclosure*.

Benjamin Kuipers and Jerome P. Kassirer. 1984. Causal reasoning in medicine: Analysis of a protocol. *Cogn. Sci.*, 8:363–385.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1928–1941, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016a. Commonsense causal reasoning between short texts. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*, pages 421–431. AAAI Press.

Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung won Hwang, and Zhongyuan Wang. 2016b. Commonsense causal reasoning between short texts. In *International Conference on Principles of Knowledge Representation and Reasoning*.

Halloran Me and Cláudio José Struchiner. 1995. Causal inference in infectious diseases. *Epidemiology*, 6:142–151.

Tom Michoel and Jitao David Zhang. 2022. Causal inference in drug discovery and development. *Drug discovery today*, page 103737.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

John L. Pollock. 1987. Defeasible reasoning. *Cogn. Sci.*, 11(4):481–518.

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)*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Dominick A Rizzi. 1994. Causal reasoning and the diagnostic process. *Theoretical medicine*, 15:315–333.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Steven Sloman. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Mr Andrew J Tiffin. 2019. *Machine learning and causality: The impact of financial crises on growth*. International Monetary Fund.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2023. COLA: Contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5253–5271, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiayao Joy Zhang, Hongming Zhang, Weijie J. Su, and Dan Roth. 2022. ROCK: causal inference principles for reasoning about commonsense causality. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26750–26771. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Qualification of Annotators

For the collection of δ-CAUSAL, we expect it to have diversity while maintaining good quality. To achieve this requirement, we need to include more annotators while ensuring the quality of their annotations.

In the collection process, we set the qualification that the annotators should have a HIT acceptance rate greater than 97 and a number of HITs approved greater than 10,000, which is a more general qualification rule to incorporate more annotators into our dataset collection process. This step ensures us a diversified dataset. During the refinement step, we maintain a group of annotators (38 annotators) who understand our task well, we got this qualified annotator list by doing a toy collection of over 500 samples. Additionally, to ensure that we can justify the quality of each refinement assignment, we assign each assignment to three annotators. Each assignment is composed of a golden example, which we know is true or false, and the example needs testing. Besides, as we focus on the English corpus, we set the country of residence to the USA and UK.

In summary, we use general qualifications in the collection and more specific qualified groups in the refinement. In this way, we emphasize diversity in the collection and accuracy in the refinement.

B Details of Dataset Collection

In this section, we present the guidelines and illustrative examples we give the annotators in detail.

B.1 Annotation of Cause-Effect Pairs

Guideline of Annotation. Here we take the annotation of cause-effect pairs with hints of keywords as the instance. The annotation without keywords is similar except that the keywords are not given. For each annotating example (HIT in Amazon Turk), we assign a domain and one keyword to the annotator and ask them to write a cause and its effect. To ensure that they write the effect, we explicitly require them to select a time interval for the effect from `months later`, `years later`, `decades later` and `centuries later`.

Illustrative Examples. The illustrative examples we give for each domain for the cause-effect pairs are as follows [1]:

`Domain of Environment`

1. *keyword*: natural disaster.
   Cause: A tsunami hits the west coast.
   Effect: Years later, homelessness and mental health issues arise.

2. *keyword*: natural disaster.
   Cause: An earthquake happens in the city.
   Effect: Years later, the city commemorates earthquake victims with charity events.

`Domain of Politics`

1. *keyword*: political-party.
   Cause: Tom founded this party with the hope of leading people to a better life.
   Effect: Centuries later, it becomes the world's oldest active political party.

2. *keyword*: election.
   Cause: The senator made a racist remark.
   Effect: Months later, the senator's remarks cost him an election.

`Domain of Travel`

1. *keyword*: resort.
   Cause: Tourists throw rubbish everywhere at the scenic spot.
   Effect: Years later, fewer and fewer tourists go to this scenic spot.

2. *keyword*: trip.
   Cause: A young tourist is very happy after visiting Lausanne.
   Effect: Decades later, this tourist revisits the old place, remembering the good old days.

`Domain of Entertainment`

1. *keyword*: art.
   Cause: The artist signs a recording contract.
   Effect: Years later, the artist becomes a star and is popular among people.

2. *keyword*: movies.
   Cause: The plot of the newly released film is very intriguing.
   Effect: Decades later, this movie has been remade several times, and is known around the world.

`Domain of Sports`

1. *keyword*: soccer.
   Cause: The government is determined to

[1] Only the cases with hints of keywords are given. The cases without hints of keywords are similar

make deep changes to professionalize its soccer and bring players closer to the global standard.
Effect: Decades later, the soccer team won the World Cup finally.

2. *keyword*: doping.
Cause: The athlete has been caught doping in the Olympics.
Effect: Years later, the athlete falls sick and retires from sports.

`Domain of Education`

1. *keyword*: school.
Cause: Tom's parents decide to let Tom enroll in a famous but expensive school.
Effect: Years later, Tom is well-educated and is thankful for his parents' efforts.

2. *keyword*: major.
Cause: Tom changes his major from mathematics to computer science.
Effect: Decades later, Tom becomes a senior software engineer in an IT company.

`Domain of Health`

1. *keyword*: lifestyle habits.
Cause: John started smoking.
Effect: Decades later, John suffers from heart disease and stroke.

2. *keyword*: health problems.
Cause: John got COVID-19.
Effect: Months later, John can still recall the bad feeling of having COVID-19.

`Domain of Work`

1. *keyword*: career success.
Cause: She has found her routine for a productive day at work.
Effect: Years later, she gets a chance for promotion because of her hardworking.

2. *keyword*: career change.
Cause: The company's new launch date puts employees under pressure.
Effect: Months later, many employees have decided to leave.

`Domain of Business`

1. *keyword*: start-up.
Cause: This newly opened coffee shop decides to attract students.
Effect: Months later, this coffee shop is the most favorite place for students to socialize.

2. *keyword*: strategy.
Cause: This company decides to expand its business overseas.
Effect: Decades later, this company is alive and well, and it is still renowned worldwide as the oldest company.

B.2 Annotation of Defeasibility

Guideline of Annotation. Firstly, the annotators are asked to write a supporting argument that could be the conceptual explanation behind the long-term effect. For the defeated arguments, we ask the annotators to note that with the defeater event, the effect doesn't hold any more or the effect is weakened by this defeater event. Additionally, the annotators have to specify the time interval after which the defeater happens relative to the given cause.
Illustrative Examples. The illustrative examples we give for the defeasibility annotations are as follows:

1. *Cause*: The soccer team receives lots of funding.
Effect: Years later, the soccer team wins the soccer league championship.
Supporter: Soccer teams can hire excellent coaches and players with adequate funding.
Defeater: Months later, the funding is wasted on corruption.

2. *Cause*: Tourists throw rubbish everywhere at the scenic spot.
Effect: Months later, fewer and fewer tourists go to this scenic spot.
Supporter: Spots with rubbish are dirty, and people don't like dirty places.
Defeater: Volunteers pick up the trash thrown by tourists every day to keep the site clean.

3. *Cause*: John started smoking.
Effect: Decades later, John suffers from heart disease and stroke.
Supporter: The nicotine in tobacco can damage the heart.
Defeater: Nicotine has been shown to soothe the heart.

4. *Cause*: The artist signs a recording contract.
Effect: Years later, the artist becomes a star and is popular among people.

Supporter: Artists who signed contracts usually work hard to release albums.
Defeater: The artist is lazy and rests on his laurels.

C Details of Statistics of δ-CAUSAL

In this section, we present more details about the statistics of δ-CAUSAL, including the sentence length distributions of supporters and defeaters (Appendix C.1), time interval distribution (Appendix C.2), and keywords (Appendix C.3).

C.1 Details of Sentence Length

We plot the comparison of distributions of sentence length of supporters and defeaters in Figure 7. Compared with supporters, defeaters are always associated with more complicated logic as they need to provide supplementary information to overturn or attenuate the causal relationship. However, supporters are relatively simple for human annotators as they only need to think out the background knowledge to provide more support for the causality relationship.

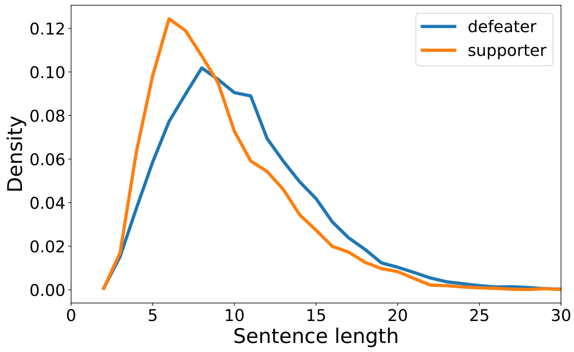


Figure 7: Comparison of sentence length distributions between supporters and defeaters.

C.2 Details of Time Interval Distribution in Cause-Effect Pairs

The details of the time interval distribution in cause-effect pairs are shown in Table 7 and Figure 8. As we can see, most time intervals fall into months later, years later and decades later. The portion of centuries later is relatively small. This is reasonable as it is difficult for annotators to think out some causes whose effect event happens after centuries. Besides, we find each domain has different time label distributions. For instance, most centuries later labels fall into the domain of Environment.

Domains	Overall	Months	Years	Decades	Centuries
Environment	1,118	393	455	231	39
Business	1,094	517	487	90	0
Sci-Tech	1,013	353	496	140	24
Health	1,260	601	486	168	5
Work	1,259	795	380	83	1
Politics	1,019	538	371	96	14
Education	1,195	438	634	121	2
Sports	1,151	548	509	88	6
Entertainment	1,068	553	434	80	1
Travel	1,068	682	300	75	11
Total	11,245	5,418	4,552	1,172	103

Table 7: Statistics of time intervals in δ-CAUSAL. From the statistics, we can observe that our dataset is even over different domains. Besides, we found that the number of annotations for *centuries later* is relatively small, which agrees with our intuition as the effect that happens centuries later is difficult to estimate and annotate. From the distribution of δ-CAUSAL over different time intervals, we can conclude that δ-CAUSAL is a comprehensive and unbiased dataset covering different domains and agrees quite well with the temporal characteristics in commonsense causalities of different domains. Specifically, it is more likely to obtain long-term effects in the domain of Environment, Science and Technology, and Politics than in other domains like Sports, which agrees well with human commonsense.

C.3 Details of Keywords

For these keywords in § 4.2, we plot the word cloud of each domain in Table 8. We could clearly observe that incorporating hint words into the annotation process broadens the range of topics and makes δ-CAUSAL a comprehensive dataset that covers various aspects of commonsense knowledge. It shows the necessity of introducing keywords into the annotation of δ-CAUSAL.

D Details of Experiments on Supporter and Defeater Generation

In this section, we give more details about the experiments of supporter and defeater generation. Specifically, we present the setup in Appendix D.1. After that, we elaborate on the training details in Appendix D.2. Finally, we present the details of the human evaluation, i.e., NLP experts, in Appendix D.3.

D.1 Setup

In the causal defeasibility generation experiment, models are trained to generate either the supporter or the defeater given the cause and the effect (with time interval prepended). For sequence-to-sequence models, the encoder input is to use the

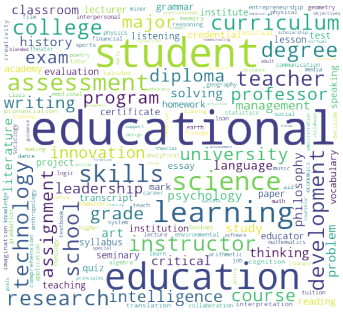
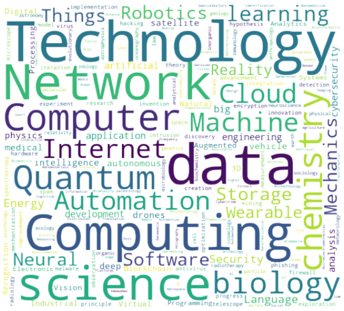
Domain	Wordcloud	Domain	Wordcloud
Business		Education	
Entertainment		Environment	
Health		Politics	
Science		Sports	
Travel		Work	

Table 8: Word cloud of each domain's keywords.

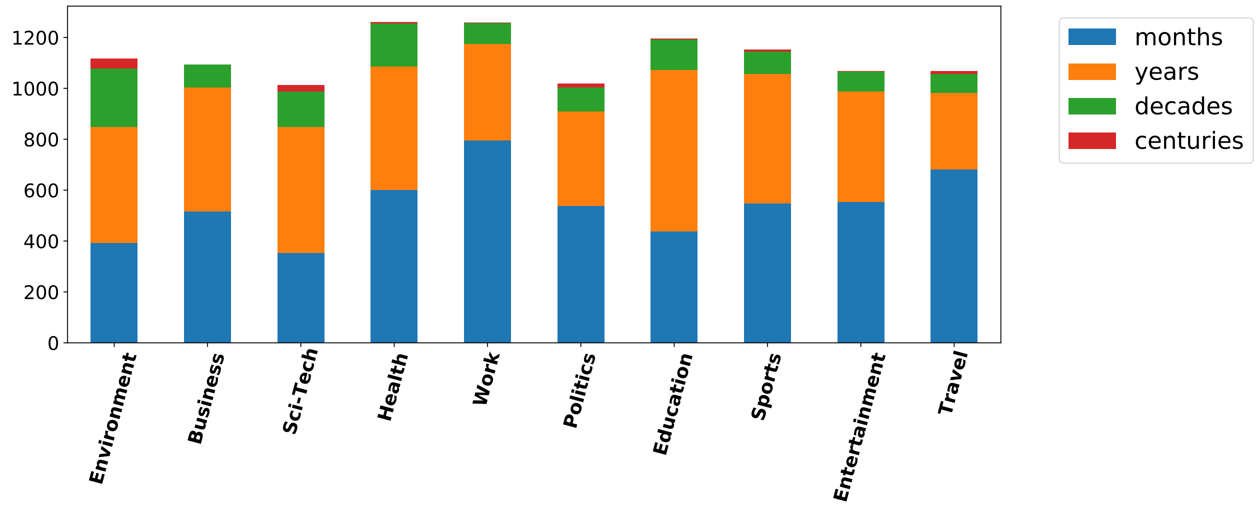


Figure 8: The distribution of time intervals in δ-CAUSAL. We can observe that for each domain, the distribution over time label shows some variations and similarities. For instance, in the domain of environment, the portion of time labeled "centuries later" is relatively higher. Besides, for almost all the domains, the causalities with the time intervals of "months later" and "years later" are the most common. Besides, we could find that the causality is almost even distributed into different domains. This proves that δ-CAUSAL is a comprehensive dataset covering different domains in daily life.

Model	Input format
BART	<s> C </s></s> E </s>
T5	P_c C </s> P_l E </s>
GPT-2	P_c C P_l E P_d
T5-large	P_c C </s> P_l E </s>

Table 9: Input format of each model in causal defeasibility generation task. P_c = "cause:", P_l = "long-term effect:", E is the effect prepended with the long time interval. For example, E_l could be "Months later, mental health issues arise.". C is the cause of the effect. These models have different suggestions for input format. To achieve the best performance, we follow the official suggestions in their original papers. This is the reason why we use different input formats for different models.

text prefixes P_c and P_l for the cause and long-term effect. For the causal language model GPT-2, we use the text prefix P_d for defeasibility, which could be either "assumption:" or "defeater:" depending on the task. The input format for all models involved is shown in Table 9. The versions of all the packages, tools (scientific artifacts) are listed in the `requirements.txt`, which is included in our code repository. The code repository is attached as the supplementary zip file. The major tools we use include `PyTorch` [2] (BSD-3 License) and `transformers` [3] (Apache License 2.0). Our proposed dataset δ-CAUSAL (MIT License) will be released to the public after this paper's acceptance. Besides, the other datasets we use, e-CARE (MIT License) (Du et al., 2022) and COPA (BSD 2-Clause License) (Roemmele et al., 2011), are both publicly available and open-source.

D.2 Training Details

We fine-tune BART-base (140M), GPT-2 (117M), T5-base (223M), and T5-large (783M) models. We use the Huggingface Trainer default optimizer AdamW for training. All models are optimized with a batch size of 32 (8 for each GPU) and fine-tuned for 3 epochs. The learning rate is set to 1e-5 for the BART and GPT-2 models and 3e-4 for the T5-base and T5-large models. The experiments of generation on δ-CAUSAL are conducted by a single run. The prompt experiments for keyword generation and CTCW are conducted with different prompts to search for the best prompts. All of the experiments are run with a machine with four GPUs. All of these GPUs are NVIDIA TITAN X (Pascal) with a memory size of 12,288MB. The random seeds for the single run experiments in both defeater/supporter generation and CESAR are set as 42.

[2] https://pytorch.org/

[3] https://huggingface.co/docs/transformers/index

D.3 Details of NLP Experts in Our Evaluation

All of these three NLP experts are graduate students whose major research interests are natural language processing. All of them have taken advanced NLP courses and have relative research experience and two of them have papers in the domain of NLP.

E Details of Existing Causal Strength Metrics

In this section, we mainly introduce the details of existing metrics on causal strength.

Causal Explanation Quality (CEQ) Score. The causal strength in CEQ (Du et al., 2022; Luo et al., 2016a) is defined as:

$$\mathcal{CS}(C \rightarrow E) = \frac{1}{N_C + N_E} \sum_{w_i \in C, w_j \in E} \text{cs}(w_i, w_j) \tag{4}$$

where N_C and N_E are the number of words in C and E. $\text{cs}(w_i, w_j)$ is the causal strength between these two words: $\text{cs}(w_i, w_j) = \frac{\text{Count}(w_i, w_j)}{\text{Count}(w_i)\text{Count}^{\alpha}(w_j)}$, where w_i is from the cause event while w_j is from the effect event. The term $\text{Count}(w_i, w_j)$ denotes the frequency with which w_i and w_j co-occur in causal statements, while $\text{Count}(w_i)$ indicates the total number of appearances in such sentences.

ROCK. ROCK (Zhang et al., 2022) defines a causal strength metric from the perspective of causal inference (Hernán and Robins, 2010):

$$\begin{aligned} \mathcal{CS}(C \rightarrow E) &= \mathbb{E}_x[\mathbb{P}(C \prec E|x) - \mathbb{P}(\neg C \prec E|x)] \\ &\approx f(C, E) - \frac{1}{|\mathcal{A}'|} \sum_{A \in \mathcal{A}'} f(A, E) \end{aligned} \tag{5}$$

where x is the confounder of the cause event and the effect event. $\neg C$ is the intervention of the cause C. $f(C, E)$ is an estimate for $\mathbb{P}(C \prec E)$, i.e., the probability of C happens before E. $\mathcal{A}'$ is the L_p-constrained set for the generated interventions conditioned on confounders: $\mathcal{A}' := \{A \in \mathcal{A} : \frac{1}{|\mathcal{X}|}\|q(x; A) - q(x, C)\|_2 \leq \epsilon\}$, where the set $\mathcal{A}$ contains generated interventions $\neg C$ and $\mathcal{X}$ is the set of the generated confounders x. ϵ is the threshold. $q(x; E)$ is the temporal propensity measuring the conditional probability of the event E given an event x.

F More Details of CESAR

In this section, we present more details of CESAR. Specifically, we present its setup in Appendix F.1, its score computation pipeline in Appendix F.2, preparation of CESAR's training data in Appendix F.3, discussion of the concatenation operation in Appendix F.4, and ablation study in Appendix F.5.

F.1 Setup of CESAR

The CESAR model consists of a BERT (Devlin et al., 2019) model reinforced with the causality-aware attention layer. We utilize the pre-trained `bert-large-uncased` model from Hugging Face (Wolf et al., 2020). This BERT model has an embedding dimension d of 1024. Thus, causality-aware attention has two learnable weight matrices $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{1024\times1024}$. The input to the model is the output of the respective tokenizer for the `bert-large-uncased` model. Specifically, we jointly preprocess the given cause C and the given effect E with the tokenizer, which produces the respective *input_ids* with appropriately added special tokens. *token_type_ids* of C are marked with 0 and *token_type_ids* of E are marked with 1. *attention_mask* is also included so that the model avoids attending on [PAD] tokens. Hence, the model input consists of *input_ids*, *token_type_ids*, and *attention_mask*. The maximal input size for the CESAR model is 512 tokens including [CLS] and [SEP] which are appended at the beginning and the end of the sequence, respectively.

F.2 Score Computation

The CESAR score is computed in several stages.

1. we extract the token embeddings from the BERT's last hidden layer:

   $$\begin{aligned} (\mathbf{C} + \mathbf{E}) = \text{BERT}(&\text{input\_ids}, \\ &\text{token\_type\_ids}, \\ &\text{attention\_mask}). \end{aligned}$$

   Since we jointly preprocess cause and effect, BERT also jointly produces embeddings for tokens of the cause and effect. Hence, $(\mathbf{C} + \mathbf{E}) \in \mathbb{R}^{(n+m)\times d}$, where n is the length of the cause while m is the length of the effect. In this setting, *token_type_ids* suggests to BERT which tokens belong to cause and which belong to effect.

2. Based on *token_type_ids* and *attention_mask*, model separates embeddings to $\mathbf{C} \in \mathbb{R}^{n\times d}$ and $\mathbf{E} \in \mathbb{R}^{m\times d}$. Next, embeddings are given to the causality-aware attention layer, where we obtain query $\mathbf{Q} = \mathbf{C}\mathbf{W}_q$ and key $\mathbf{K} = \mathbf{E}\mathbf{W}_k$ vectors. These vectors give us

the attention scores for token pairs as $\mathbf{A} = \text{softmax}\left(\mathbf{Q}\mathbf{K}^T\right)$ where softmax is performed over all values of the matrix (not only over a single dimension as in the conventional attention layers).

3. The causal strength value given by CESAR is calculated as follows,

$$\mathcal{CS}(C \to E) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} \frac{|c_i^T e_j|}{\|c_i\| \|e_j\|} \quad (6)$$

   where $c_i \in \mathbf{C}$ and $e_j \in \mathbf{E}$ represent causal embeddings of tokens of C and E respectively, and $a_{ij} \in \mathbf{A}$ is the attention weights put on each pair of tokens. Please note that we keep [CLS] and [SEP] special tokens when computing the causal strength with CESAR. As a result, the first token of cause representation is always a [CLS] token. See Section F.5 for more details about the roles of these two special tokens.

F.3 Preparation for Training Data

Training Data. We train the CESAR metrics on the augmented e-CARE dataset (Du et al., 2022). We consider both the dev and train parts of e-CARE and combine them into a single dataset. This dataset contains causal-effect sentence pairs with a conceptual explanation for each cause-effect pair. Accordingly, with the conceptual explanation, the causal strength between the cause and the effect increases. Motivated by this fact, we set the causal strength as $\mathcal{CS}(C \to E) = 0.7$ and $\mathcal{CS}(C \oplus H \to E) = 1.0$ where H is the conceptual explanation for why C leads to the occurrence of E. The dataset also includes pairs of sentences with no causal relationship, we set the causal strength of these non-causal event pairs to 0.0. Lastly, in order to replicate the decreasing effect that defeater has on causal strength as proposed in the δ-CAUSAL, we use ChatGPT to generate semantic opposites from the conceptual explanations provided in e-CARE and set $\mathcal{CS}(C \oplus \neg H \to E) = 0.2$ where $\neg H$ is an opposite of the conceptual explanation for C and E. $\neg H$ is generated by ChatGPT and the prompt given to ChatGPT to generate this semantic opposite is shown in the next paragraph. The training dataset constructed in this manner consists of a total of 68,220 examples. The discussion about the impacts of these various types of data samples is described in Appendix F.5.

Data Augmentation with ChatGPT. As described in the aforementioned part, in order to imitate the decreasing effect of the defeaters in δ-CAUSAL, we use ChatGPT to generate sentences that have opposite meanings from the conceptual explanation provided in the e-CARE dataset. To be more specific, we use OpenAI's API [4] and the gpt-3.5-turbo model with a temperature set to 0.9. The prompt provided to the gpt-3.5-turbo model stands as follows:

> You are a helpful assistant that helps to find the opposite of the given sentence. The real truth is not important just the resulting sentence must be of the opposite meaning, negating the information that the given sentence tries to convey. Try to not give a simple negation. Output ONLY the resulting sentence, nothing else. For example for the prompt: "Friends join communities.", the output should be: "Friends avoid communities." Also for the prompt: "Sulfonamides cause hemolysis less commonly.", the output should be "Sulfonamides cause hemolysis more commonly.". Another example would be that for the prompt: "Homelessness greatly increases the likelihood of a suicide attempt.", the output is: "Homelessness greatly decreases the likelihood of a suicide attempt." The last example is that for the prompt: "Production occurs in dense regions.", the output must be: "Production occurs only in sparse regions.
>
> "Products cause slow growth."

The resulting response to the above prompt is: "Products promote rapid growth". For most cases, we observe that the model generates the satisfying negation. However, there are examples where it applies the double negation such that the second negation nullifies the first thus resulting sentence is not semantically opposite to the initial input. For instance, for the prompt "Attempts yield results", the output is "Not attempting ensures no results", or for "The sun sets early in December", we get "The sun rises late in December". We believe that

[4] https://openai.com/blog/chatgpt

doubly negated sentences incorporated in the training set and generated as defeaters, but not effectively acting as such, introduce extraneous noise, thereby impeding the model's performance in identifying defeaters. Accordingly, our model demonstrated superior results in supporters as compared to defeaters. In addition, despite an instruction to avoid simplistic negation by the mere introduction of "not" in input sentences, the gpt-3.5-turbo model continues to do so in a considerable number of examples. All of this leads us to the conclusion that we could improve the performance of our model if we would generate the opposites of conceptual explanations from e-CARE more reliably and correctly, e.g., by using human labor instead of automatic rendering.

F.4 Formulation of Concatenation Operation

To validate the expected behavior of CESAR, it is necessary to demonstrate the capacity of CESAR to capture the causal strength changes after integrating supporters and defeaters with respective causes. In order to do that, we need to first formulate the concatenation operation $\oplus$ between two statements. Since we use the BERT model as our backbone model when implementing CESAR, we define concatenation as follows

$$C \oplus A/D = C \text{ [SEP] } A/D \qquad (7)$$

where C is the cause that we wish to concatenate with either supporter A or defeater D, and [SEP] is BERT's special token that helps BERT know that C, A/D, and E are separate sentences (entities).

F.5 Ablation Study

There are many techniques contributing to the success of CESAR in capturing the causal strength changes such as causality-aware attention, special tokens like [CLS] and [SEP], data augmentation, and backbone model selection. To validate the effectiveness of these techniques, we conducted a comprehensive ablation study. In Table 10, we display the results with various ablations from the best CESAR model. From the results, we have the following observations:

	Supporter	Defeater	Geometric mean
CESAR	84.6	75.8	**80.1**
w/o causality-aware attention	91.2	64.4	76.6
w/o [CLS] & [SEP]	80.2	76.0	78.0
w/o data augmentation	64.2	63.6	63.9
w/ imbalanced data augmentation	89.0	24.4	46.6
w/ `bert-large-cased`	76.8	78.0	77.9
w/ `bert-base-uncased`	78.6	79.8	79.2

Table 10: Performance on 500 samples from δ-`CAUSAL` by different variations of the CESAR metrics. The accuracy on supporters and defeaters is calculated in the same ways as that described in Table 3. For more clarity, we employ the abbreviation "w/o" to indicate the exclusion of a specific component from the CESAR build-up. Therefore, we conduct experiments by removing the causality-aware attention layer, as well as the [CLS] and [SEP] tokens. Furthermore, we train CESAR without using the augmented dataset containing supporters and defeaters. Conversely, we employ the abbreviation "w/" to denote that a particular component has been substituted from the original CESAR build-up. Besides, we explore imbalanced data augmentation, where only conceptual explanations are augmented as supporters, without generating their opposites. Finally, we evaluate the usage of alternative BERT models including `bert-large-cased` and `bert-base-uncased` instead of the original `bert-large-uncased`.

- w/o causality-aware attention: our results demonstrate the pivotal contribution of the causality-aware attention layer in enhancing metrics stability and performance in scenarios involving defeating information. Notably, with the incorporation of causality-aware attention, we observe a substantial improvement in accuracy—from 64.4% to 75.8%—on defeaters. Specifically, this layer enables redirection of focus (attention) from word pairs with strong causal relationships to those with weaker associations following the introduction of the defeaters. One interesting phenomenon here is that the ablated version of CESAR in this setting, i.e., *w/o causality-aware attention*, achieves an accuracy of 91.2% in capturing the causal strength change brought by supporters, which is even better than CESAR. However, this ablated version struggles in capturing the causal strength changes with defeaters, with an accuracy of 64.4%. In other words, the causality-aware attention mechanism makes CESAR a more comprehensive evaluation metric on causal strength, which can not only capture the supplementary information that increases the causal strength but also can capture the counterpart that decrease the causal strength.

- w/o [CLS] & [SEP]: we observe highly unstable training once we attempt to remove [CLS] and [PAD] tokens when computing the causal strength score. Specifically, the loss during training exhibits high fluctuations with our default learning rate of $1e-5$. If the learning rate is decreased, the optimizer has trouble finding a satisfying local minimum, and training is slow. Hence, we incorporate [CLS] and [PAD] tokens when calculating the causal strength. The introduction of these special tokens is due to considerations for training stability. Also, using these tokens can also enhance the performance a bit: from 80.2% to 84.6% on supporters.

- w/o data augmentation: there are four kinds of data samples for CESAR's training: (a) True cause-effect pairs with a causal strength value of 0.7, i.e., $\mathcal{CS}(C \rightarrow E) = 0.7$. (b) False cause-effect pairs that do not have a causal relationship with a causal strength value of 0.0. (c) cause-explanation-effect triples with a causal strength value of 1.0, i.e., $\mathcal{CS}(C \oplus H \rightarrow E)$ to 1.0 where H is the explanation for the causal relationship between C and E. (d) cause-opposite_explanation-effect triples with a causal strength value of 0.2, i.e., $\mathcal{CS}(C \oplus \neg H \rightarrow E)$ to 0.2 where $\neg H$ is an opposite of the conceptual explanation for C and E. $\neg H$ is generated by ChatGPT. For the ablation case *w/o data augmentation*, we only use the data samples of type (a) and (b). We can clearly notice that data augmentation plays a crucial role as the accuracy decreases from 84.6% to 64.2% in capturing the causal strength changes brought by supporters, and from 75.8% to 63.6% on defeaters. It shows the data augmentation with explanation and its opposite is a necessary component for the success of CESAR. It can be explained that the CESAR is provided with more fine-grained examples in understanding different levels of intensity of causal strength.

- w/ imbalanced data augmentation: for the ablated case *w/ imbalanced data augmentation*, we only use data samples of type (a), (b), and (c). We observe that without the generated opposites of explanations $\neg H$, CESAR becomes overly biased as it seems to learn to always increase causal strength once the new information is attached to the cause. The accuracy on defeaters decreases from 75.8% to 24.4%. It proves that the opposites of conceptual explanations play a critical role in CESAR.

- w/ `bert-large-cased`: we experiment with a BERT variant that distinguishes cased and uncased words, which decreases the overall performance, presumably due to heightened complexity and little value added to the metric. Note that our CESAR is built upon a `bert-large-uncased` model.

- w/ `bert-base-uncased`: we assess the efficacy of CESAR based on alternative backbone models. As one can observe in Table 10, it indicates that the employment of `bert-base-uncased` yields comparable results to its larger counterpart, `bert-large-uncased`. Strikingly, the former option, i.e., `bert-base-uncased`, attains the best defeater score compared to all other configurations, thereby suggesting its utility as a viable alternative in resource-constrained settings.