ROCK: Causal Inference Principles for Reasoning about Commonsense Causality

Jiayao Zhang [1 2] **Hongming Zhang** [1 3] **Weijie J. Su** [2] **Dan Roth** [1 4]

Abstract

Commonsense causality reasoning (CCR) aims at identifying plausible causes and effects in natural language descriptions that are *deemed reasonable by an average person*. Although being of great academic and practical interest, this problem is still shadowed by the lack of a well-posed theoretical framework; existing work usually relies on deep language models wholeheartedly, and is potentially susceptible to *confounding co-occurrences*. Motivated by classical causal principles, we articulate the central question of CCR and draw parallels between human subjects in observational studies and natural languages to adopt CCR to the potential-outcomes framework which, to the best of our knowledge, is the first such attempt for commonsense tasks. We propose a novel framework, ROCK, to **R**eason **O**(A)bout **C**ommonsense **K**(C)ausality, which utilizes temporal signals as incidental supervision, and balances confounding effects using *temporal propensities* that are analogous to propensity scores. ROCK is modular and zero-shot, and demonstrates good CCR capabilities.

1. Introduction

Commonsense causality reasoning (CCR) is an important yet non-trivial task in natural language processing (NLP) that exerts broad industrial and societal impacts (Kuipers, 1984; Gordon et al., 2012; Mostafazadeh et al., 2020; Sap et al., 2020). We articulate this task as

> *reasoning about cause-and-effect relationships between events in natural language descriptions that are deemed reasonable by an average person.*

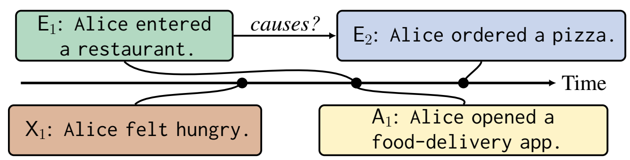

Figure 1: **An Example of CCR:** *does* E_1 *cause* E_2*?* The temporal order $\mathsf{E}_1 \prec \mathsf{E}_2$ does not necessitate causation due to confounding co-occurrences (e.g., X_1). Since when *conditioning on* X_1, a *comparable* intervention A_1 of E_1 also precedes E_2, the effect from E_1 to E_2 shrinks.

This definition naturally excludes questions that are beyond commonsense knowledge, such as those scientific in nature (e.g., does a surgery procedure reduce mortality?). Instead, it accommodates causal queries within the reach of an ordinary reasonable person. As a concrete instantiation, we consider the problem of defining and estimating the strength of causation from one given event, E_1, to another, E_2. For example, in Figure 1, is Alice's "entering a restaurant" (E_1) a plausible cause for her "ordering a pizza" (E_2)? Although the precedence from E_1 to E_2 is logical, it might be less a "cause" compared with Alice's "feeling hungry" (X_1).

Temporality informs causation, but it is still unclear how to account for confounding co-occurrences (such as X_1 in Figure 1). Motivated by causal inference principles (Section 2), we formulate CCR as estimating the *change* in the likelihood of E_2's occurrence due to intervening E_1 (denoted by $\neg\mathsf{E}_1$):

$$\Delta = \mathbb{P}(\mathsf{E}_1 \prec \mathsf{E}_2) - \mathbb{P}(\neg\mathsf{E}_1 \prec \mathsf{E}_2) \tag{1}$$

where $\mathbb{P}(\cdot)$ can be estimated by pretrained language models (LMs) e.g., via masked language modeling (see Section 4 for implementation details). The estimand Δ measures the *average treatment effect* (ATE): its magnitude signifies the strength of the effect and its sign informs the direction. For example, when Δ is close to -1, E_1 has a strong effect on E_2 towards making E_2 less prone to occurring. If the occurrences of E_1 and $\neg\mathsf{E}_1$ on any unit are purely random, a direct estimation of the temporal probabilities in Equation (1) suffices; however, due to confounding co-occurrences (e.g.,

[1]Cognitive Computation Group, University of Pennsylvania, USA. [2]Department of Statistics and Data Science, University of Pennsylvania, USA. [3]Tencent AI Lab Seattle, USA. [4]Amazon AWS AI Labs, USA. Correspondence to: <{zjiayao,hzhangal,danroth,suw}@upenn.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

X_1), one needs to *balance* the covariates (events that precede E_1) to eliminate potential spurious correlations. We propose *temporal propensity*, a surrogate propensity score that can be used to balance the covariates (Section 3). We show in Section 5 that *although temporality is essential for CCR, it is vulnerable to spurious correlations without being properly balanced*.

Contributions. We articulate CCR from a completely new perspective using causal inference principles, and our contributions include (i) a novel commonsense causality framework; (ii) mitigating confounding co-occurrences by matching temporal propensities; (iii) a modular pipeline for zero-shot CCR with demonstrated effectiveness.

2. Background

The problem of reasoning about causal relationships, and differentiating them from innocuous associations has been contemplated and studied extensively in human populations research spanning clinical trials, epidemiology, political and social sciences, economics, and many more (Fisher, 1958; Cochran & Chambers, 1965; Rosenbaum, 2002; Imbens & Rubin, 2015) among which causal practitioners usually base on the potential outcomes framework (also known as the Rubin causal model, see Neyman, 1923; Rubin, 1974; Holland, 1986), graphical and structural equation models (Robins, 1986; Pearl, 1995; Heckman, 2005; Peters et al., 2017), and Granger causality (Granger, 1969).

With the recent celebrated empirical success of language models on various NLP tasks, especially transformers (Devlin et al., 2019; Radford et al., 2019), there is an increasing interest in the NLP community on drawing causal inference using textual data. The majority of these works treat textual data as either covariates or study units (Keith et al., 2020; Feder et al., 2021) on which causal queries are formed (e.g., does taking a medicine affect recovery, which are recorded in textual medical records?). On the other hand, CCR with natural language descriptions struggles to fit in a causal inference framework: *textual data in this case are just vehicles conveying semantic meanings, not to be taken at face value,* hence it is difficult to draw the parallel between causal inference that requires a clear definition of study units, treatments, and outcomes.

2.1. Existing Approaches

Existing works related to CCR are usually grouped under the umbrella term of commonsense reasoning (Rashkin et al., 2018; Ning et al., 2019a; Sap et al., 2020) or causal event detection (O'Gorman et al., 2016). Some of the notable progress usually comes from leveraging explicit causal cues/links (tokens such as "due to") and use conditional probabilities to measure "causality" (Chang & Choi, 2004; Do et al., 2011; Luo et al., 2016); leveraging large-scale pre-trained LMs via augmenting training datasets, designing training procedures, or loss functions (Sap et al., 2019; Shwartz et al., 2020; Tamborrino et al., 2020; Zhang et al., 2021; Staliunaite et al., 2021).

There are several works that are relevant to ours, yet different in various ways: Granger causality, which measures association, is used by Kang et al. (2017) to detect event causes-and-effects; Bhattacharjya et al. (2020) studies events as point-processes, in a way arguably closer to association; Gerstenberg et al. (2021) uses a simulation model to reason physical causation. To the best of our knowledge, we are the first one to adopt a causal perspective in solving CCR.

2.2. Challenges of CCR

Many existing CCR methods (mostly supervised) are based on ingenious designs and creative LM engineering. Theoretical justifications, however, are sometimes desirable, as only then do we know how general these methods can be. Indeed, recent studies reveal that several supervised models may have exploited certain artifacts in datasets to ace the evaluations (Kavumba et al., 2019; Han & Wang, 2021).

This dilemma of constructing a well-founded theoretical framework versus engineering models to achieve excellent empirical performances is not surprising, perhaps, given that the challenges of CCR from causal perspectives are not trivial at all: what is the study unit, treatment, and outcome in this case? What does it mean to "intervene", or "manipulate" the treatment? Is treatment *stable*, or is it desirable to consider multiple versions of it?

2.3. Principles of the ROCK Framework

In this paper, we attempt to address these questions using, among several causal principles, the following two that are intuitive and directly appeal to human nature (see e.g., Russell, 1912; Bunge, 1979): (1) **Precedence does not imply causation**, which warns us *post-hoc* fallacies; (2) **Causation implies precedence**, which informs us that the events must be compared with those that are *in pari materia* (Mill, 1851; Hill, 1965), or having *balanced* covariates (also called "pretreatments," by which we mean events that occur prior to E_1, cf. Rosenbaum, 1989). Our CCR formulation in terms of temporality has several benefits: (i) the intrinsic temporality of causal principles characterizes its central role in CCR; (ii) temporal signals bring about incidental supervision (Roth, 2017; Ning et al., 2019a); (iii) although being a non-trivial question *per se*, reasoning temporality has witnessed decent progress lately, making it more accessible than directly detecting causal signals (Ning et al., 2017; 2018; 2019b; Zhou et al., 2020; Vashishtha et al., 2020).

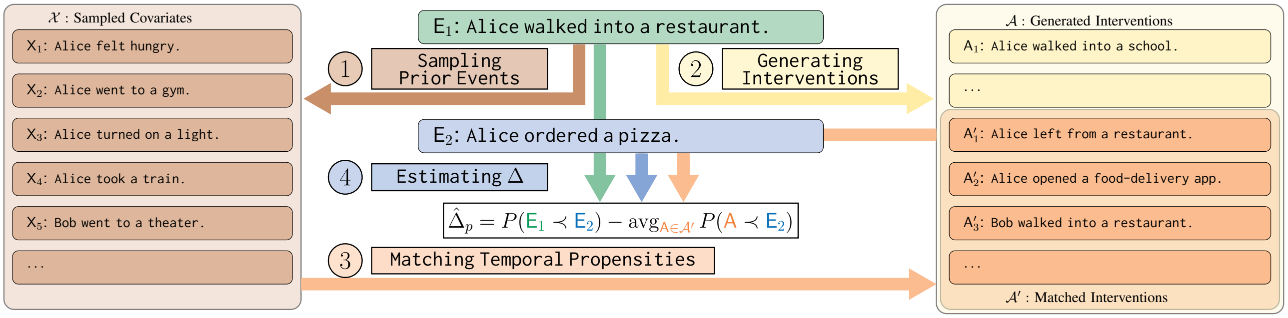

Figure 2: **Illustration of the ROCK framework.** *Does* E_1 *cause* E_2*?* To answer this query, ① the event sampler samples a set of covariates $\mathcal{X}$ of events X_k that occur preceding E_1. ② The intervention generator generates a set $\mathcal{A}$ of interventions A_k on E_1. ③ A subset $\mathcal{A}' \subset \mathcal{A}$ of interventions is selected whose temporal propensities $q(\boldsymbol{x}; \mathsf{A})$ is close to that of E_1, $q(\boldsymbol{x}; \mathsf{E}_1)$ (Equation (7)). ④ The temporal predictor uses $\mathcal{A}'$ to estimate Δ.

3. The ROCK Framework

Notations. We use sans-serif letters for events, uppercase serif letters for *indicators* of whether the corresponding event occurs,[1] and lowercase serif letters for the realizations of those indicators. For example, in Figures 1 and 2, E_1 : "`Alice walked into a restaurant.`," $E_1 = \mathbb{1}\{\mathsf{E}_1 \text{ occurs}\}$ and $e_{1,i} = \mathbb{1}\{\mathsf{E}_1 \text{ occurs to the } i\text{-th study unit}\}$[2]. We view the occurrence of events as point processes $E(t)$ on $t \in \{0, 1\}$ (e.g., present versus past). We use $\mathsf{E}_1 \succ \mathsf{E}_2$ (resp. $\prec$) to indicate that E_1 occurs following (resp. preceding) E_1. We write $\mathbb{P}(\mathsf{E}_1 \prec \mathsf{E}_2) = \mathbb{P}(E_1(0), E_2(1))$ and $\mathbb{P}(\mathsf{E}_2|\mathsf{E}_1) = \mathbb{P}(E_2(1)|E_1(0))$. We write P for estimates of $\mathbb{P}$, and omit measure-theoretic details[3].

Overview of the ROCK framework. We set the stage in this section and discuss implementation details in Section 4. Given E_1 and E_2, as shown in Figure 2, ROCK samples the covariates set $\mathcal{X}$ and interventions set $\mathcal{A}$, from which a matched subset $\mathcal{A}'$ is selected via temporal propensities (Section 3.4). The score Δ is then estimated by Equation (7).

3.1. The Central Question of CCR

Given two specific events E_1 and E_2, as discussed in Section 1, we articulate CCR as the estimation of the change of temporal likelihood *had* E_1 been *intervened*:

$$\Delta = \mathbb{P}\left(\mathsf{E}_1 \prec \mathsf{E}_2\right) - \mathbb{P}\left(\neg\mathsf{E}_1 \prec \mathsf{E}_2\right) \tag{2}$$

which assumes values in $[-1, 1]$ and measures a form of the *average treatment effect*. As these probabilities are eventually estimated from data, if there are confounding events X_k that always co-occur with E_1 in the data itself, they will bias this estimation. To this end, it is necessary to first clear out several key notions associated with this causal query, and then properly define the intervention $\neg\mathsf{E}_1$.

3.2. The Potential-Outcomes Framework

One major challenge of framing a causal query for CCR is the ambiguity of the underlying mechanism. Unlike human populations research, where experiments and study units are obvious to define, it is not immediately clear what they are when faced with semantic meanings of languages (Zhang & Zhang, 2022). Yet, we can draw parallels again between semantic meanings and human subjects via the following thought experiment: suppose each human subject keeps a journal detailing the complete timeline of her experiences since her conception, then we can treat each individual as a study unit where the temporal relations of events can be inferred from the journal.

We can then formulate CCR in the language of the potential-outcomes framework. Given fixed events E_1 and E_2, let E_{1i} denote the event experienced by the i-th study unit at time $t = 0$ when E_1 is supposed to occur. Each unit is then associated with a treatment assignment $E_{1i} = \mathbb{1}\{\mathsf{E}_{1i} = \mathsf{E}_1\}$, realizations of the covariates $\boldsymbol{x}_i = (x_{ij})_{j=1}^N$ for $x_{ij} = \mathbb{1}\{\mathsf{X}_j \prec \mathsf{E}_{1i}\}$, and two potential outcomes

$$\begin{cases} r_{0i} = \mathbb{1}\{\mathsf{E}_{1i,E_1=0} \prec \mathsf{E}_2\}, \\ r_{1i} = \mathbb{1}\{\mathsf{E}_{1i,E_1=1} \prec \mathsf{E}_2\}. \end{cases} \tag{3}$$

Here $\mathsf{E}_{1i,E_1=1-E_{1i}}$ signifies the hypothetical scenario where this unit *had* received the treatment assignment $1 - E_{1i}$, when in fact it receives E_{1i}. Clearly, either r_{0i} and r_{1i} can be observed, but not both. Our estimand Δ in Equation (1) is indeed the average treatment effect

$$\Delta = \mathbb{E}[r_1 - r_0] \equiv \mathbb{P}(\mathsf{E}_1 \prec \mathsf{E}_2) - \mathbb{P}(\neg\mathsf{E}_1 \prec \mathsf{E}_2). \tag{4}$$

This identification naturally complies with the temporal na-

[1]By "occurs," we mean "is observed." We treat them interchangeable in the rest of our paper.
[2]Defined among other concepts in Section 3.2.
[3]Let $\mathcal{E}$ be the set of commonsense events we consider, the probability space we are working on is $(\mathcal{E} \times \mathcal{E}, \sigma(\mathcal{E} \times \mathcal{E}), \mathbb{P})$.

ture of covariates (Rubin, 2005), since by definition they are *pretreatments* that take place *before* the treatment. We shall now address the issue of intervention (manipulation). Generally speaking, events are complex, and therefore intervention in this case would be better interpreted in a broader sense than one particular type of manipulation such as negation. For example, with E_1 being "Alice walked into a restaurant," suppose hypothetically, before E_1, Alice did not walk into a restaurant ($\neg\mathsf{E}_1^1$), we can thus compare $\mathbb{P}(\mathsf{E}_1 \prec \mathsf{E}_2)$ with $\mathbb{P}(\neg\mathsf{E}_1^1 \prec \mathsf{E}_2)$ to reason to what extent some event E_2 can be viewed as the effect due to E_1. However, this is not the complete picture: Alice may have walked into somewhere else such as a bar; she may have, instead of walked into, but left a restaurant; instead of Alice, perhaps it was Bob who walked into a restaurant. The temporal information between these events and E_2 are also likely to inform causation between E_1 and E_2, and they are no less interventions than negation. As such we interpret intervention in our framework in a broader sense, not necessarily only negation or the entailment of negations, but *any events that leads to plausible states of counterfactuality*. We will denote all possible interventions of E_1 as $\mathcal{A}$.

Remark. The generally acknowledged *stable unit treatment value assumption* (SUTVA, Rubin, 1980) requires that for each unit there is only one version of the non-treatment. Nonetheless, as we noted in the above discussion, the nature of the CCR problem renders it tricky to define what constitutes the exact version of the non-treatment (what single event *is* not having done something, exactly?). For ease of exposition, we allow interventions in ROCK to take on multiple versions.

3.3. Balancing Covariates

The direct estimation of Δ in Equation (1) is feasible only in an ideal world where those probabilities are estimated from randomized controlled trials (RCTs) such that the treatment (E_1) is assigned completely at random to study units. Due to confounding co-occurrences, events that precede E_1 need to be properly balanced (Mill, 1851; Rosenbaum & Rubin, 1983; Pearl & Mackenzie, 2018). Taking again as an example E_1: "Alice walked into a restaurant," and E_2: "Alice ordered a pizza,". Suppose hypothetically, Alice's twin sister Alicia, who has the exact life experiences up to the point when E_1 took place, but opted not to walk into a restaurant, but opened a food delivery app on her phone ($\neg\mathsf{E}_1$). Then we can reason that the cause-and-effect relationship from E_1 to E_2 is perhaps not large. On the other hand, if we know another irrelevant person, say Bob, underwent $\neg\mathsf{E}_1$ and then E_2, then perhaps we are not ready to give that conclusion since we do not know if Bob and Alice are comparable at the first place. This example illustrates the importance of adjusting or balancing pretreatments. As such, we may rewrite Equation (1) as conditional expectations among study units that are comparable, i.e.,

$$\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{P}(\mathsf{E}_1 \prec \mathsf{E}_2|\boldsymbol{x}) - \mathbb{P}(\neg\mathsf{E}_1 \prec \mathsf{E}_2|\boldsymbol{x})\right], \tag{5}$$

provided that the treatment assignment is strongly ignorable with respect to $\boldsymbol{x}$, in the sense of the following assumption.

Assumption 3.1 (Strong Ignorability)**.** *The potential outcomes $\{r_0, r_1\}$ are independent with the treatment assignment E_1 conditioning on the covariates $\boldsymbol{x}$.*

Remark. (i) We should define $\boldsymbol{x}$ as events preceding E_1, but *not* E_2, which will potentially introduce posttreatment biases (Rosenbaum, 1984): if an X' that occurs between E_1 and E_2 is adjusted, Δ thus estimated quantifies the effect from E_1 to E_2 *without* passing through X'. (ii) Although $\boldsymbol{x}$ should be those that are correlated with E_1, adjusting for un-correlated effects does not introduce biases.

3.4. Matching Temporal Propensities

There are several techniques for balancing covariates such as sub-classification, matched sampling, covariance adjustment, and via structural equations (Cochran & Chambers, 1965; Pearl, 1995). Rosenbaum & Rubin (1983) showed that the propsensity score can be used for this purpose. The propensity score $p(\boldsymbol{x}) = \mathbb{P}(E_1(1) = 1|\boldsymbol{x}(0))$ is the probability of E_1 occurring at time 1 conditioning on the covariates being $\boldsymbol{x}$ at time 0.

To properly identify what events constitute the covariates set is essential for our CCR framework. In the best scenario, it should include the real cause(s), which is, however, exactly what CCR solves. To circumvent this circular dependency, we use large LMs to sample a large number of events preceding E_1, which should provide a reasonable covariate set. In this case, directly computing $p(\boldsymbol{x})$ is not feasible, as will be discussed in Section 4, instead, we propose to use a surrogate which we call *temporal propsensities*:

$$q(\boldsymbol{x}) = q(\boldsymbol{x}; \mathsf{E}_1) = \left(\mathbb{P}(E_1(1) = 1|x)\right)_{x\in\boldsymbol{x}} \tag{6}$$

with each coordinate measuring the conditional probability of the event E_1 given an event in $\boldsymbol{x}$. Thus motivated, for some fixed threshold ϵ and $p \in \{1, 2\}$, we will use following estimating equation for the L_p-balanced score, where $f(\mathsf{E}_1, \mathsf{E}_2)$ is an estimate for $\mathbb{P}(\mathsf{E}_1 \prec \mathsf{E}_2)$:

$$\begin{cases} \hat{\Delta}_p = f(\mathsf{E}_1, \mathsf{E}_2) - \frac{1}{|\mathcal{A}'|}\sum_{\mathsf{A}\in\mathcal{A}'} f(\mathsf{A}, \mathsf{E}_2), \\ \mathcal{A}' := \left\{\mathsf{A} \in \mathcal{A} : \frac{1}{|\mathcal{X}|}\left\|q(\boldsymbol{x}; \mathsf{A}) - q(\boldsymbol{x}; \mathsf{E}_1)\right\|_p \le \epsilon\right\}. \end{cases} \tag{7}$$

3.5. Discussions on Temporal Propensity Matching

Unfortunately, the estimator $\hat{\Delta}_p$ in Equation (7) is generally biased even if a perfect matching of temporal propensity exists, because $q(\boldsymbol{x})$ consists of conditional probabilities on one-dimensional marginal distributions instead of on the full joint distribution. Quantifying this loss of information is a difficult problem by itself; here we outline a coarse bound for illustration purposes.

Proposition 3.2 (Expected L_2 error under perfect matching)**.** *Write $r := r_1 - r_0$, then $\Delta = \mathbb{E}[r_1 - r_0] \equiv \mathbb{E}[r]$. Define*

$$\varrho := \sup_{\tau} \{\tau \leq |r - \mathbb{E}[r|q(\boldsymbol{x})]| \ \textit{a.s.}\} \in \{0, 1\}. \quad (8)$$

The expected L_2 error of $\hat{\Delta} = \mathbb{E}[r|q(\boldsymbol{x})]$ satisfies

$$\mathbb{E}[(\hat{\Delta} - \Delta)^2] \leq 1 - \varrho^2. \quad (9)$$

The proof is due to the conditional variance decomposition and is given in the Appendix. The parameter ϱ depends on the problem instance and quantifies the level of dependence between the potential outcomes $\{r_0, r_1\}$ and the treatment assignment E_1 when conditioned on the covariates $\boldsymbol{x}$. Intuitively, the worst-case scenario $\varrho = 0$ is uncommon, since this happens only if r is a function of $q(\boldsymbol{x})$. When a large number of *diverse* covariates are sampled, ϱ is unlikely to be 0. We thus assume that $\varrho \gg 0$ and we can balance temporal propensities to produce a reasonable estimate.

4. Implementation of ROCK

Having established a framework for CCR, we provide an exemplar implementation of ROCK in this section. Our purpose is to demonstrate the potential of the ROCK and we expect engineering efforts such as prompt design can bring further improvements.

The core tool we shall use is (finetuned) pretrained deep LMs. With the sheer amount of training data (e.g., over 800GB for the Pile dataset, Gao et al. (2020)), it is reasonable to assume that those models would imitate responses of an average reasonable person. On the other hand, it is hard for generation models (masked or open-ended) to parse information that are far from the mask tokens; instead, it is more feasible for LMs to sample summary statistics of the relationships between a pair of events, which is one of the main motivations for using temporal propensities (Equation (6)).

4.1. Components of ROCK

For practical purposes, we represent an event as a 3-tuple (`ARG0`, `V`, `ARG1`). ROCK takes two events E_1 and E_2 as inputs, and returns an estimate $\hat{\Delta}$ for Δ according to Equation (7). It contains four components (cf. Figure 2): an event sampler that samples a set $\mathcal{X}$ of events that are likely to occur preceding E_1; a temporal predictor whose output $f(\mathsf{X}_1, \mathsf{X}_2)$ given two input events X_1 and X_2 is an estimate of the temporal probability $\mathbb{P}(\mathsf{X}_1 \prec \mathsf{X}_2)$; an intervention generator that generates a set $\mathcal{A}$ of events that are considered as interventions of the event E_1; and finally a scorer that first forms the temporal propensity vectors $q(\boldsymbol{x}; \mathsf{A}) \in \mathbb{R}^{|\mathcal{X}|}$ for each sampled interventions $\mathsf{A} \in \mathcal{A}$, then estimates Δ via Equation (7). We next discuss in greater details our implementation of this pipeline.

4.2. Implementation Details

Event Sampling. Given an event E_1 (e.g., E_1 : `Alice walked into a restaurant.`), we construct the prompt by adding "`Before that,`" to the sentence, forming "`Alice walked into a restaurant. Before that,` " as the final prompt. We use the GPT-J model (Wang & Komatsuzaki, 2021) , which is pretrained on the Pile dataset (Gao et al., 2020) for open-ended text generation where we set max length of returned sequences to be 30, temperature to be 0.9. We sample $n = 100$ events, cropping at the first stop token of the newly generated sentence to form $\mathcal{X}$.

Temporal Prediction. Given two events E_1 and E_2, we use masked language modeling to predict their temporal relation by forming the prompt E_1 `<MASK>` E_2 and collect the score $s_a(\mathsf{E}_1, \mathsf{E}_2)$ and $s_b(\mathsf{E}_1, \mathsf{E}_2)$ for the tokens `after` and `before`. We then symmetrize the estimates to form $s(\mathsf{E}_1, \mathsf{E}_2) = \frac{1}{2}(s_a(\mathsf{E}_1, \mathsf{E}_2) + s_b(\mathsf{E}_2, \mathsf{E}_1))$. We can directly use $s(\mathsf{E}_1, \mathsf{E}_2)$ for $f(\mathsf{E}_1, \mathsf{E}_2)$; we discuss possible normalizations of this score in Section 5.

Temporality Fine-Tuning. Directly using a pretrained LM as the temporal predictor is likely to suffer from low coverage, since the tokens `before` and `after` usually are not among the top-k most probable tokens. We can increase k but this does not heuristically justify if the outputted scores are meaningful. We thus use the New York Times (NYT) corpus which contains NYT articles from 1987 to 2007 (Sandhaus, 2008) to fine-tune an LM. Following the same procedure as Zhou et al. (2020), we perform semantic role labeling (SRL) using AllenNLP's BERT SRL model (Gardner et al., 2017) to identify sentences with a temporal argument (`ARG-TMP`) that starts with a temporal connective `tmp` (either `before` or `after`). We then extract the verb and its two arguments (`V`, `ARG0`, `ARG1`) as well as this tuple from its temporal argument, thus forming an event pair (E_1, E_2, `tmp`). We are able to extract 397174 such pairs and construct them into the fine-tuning dataset consisting of "E_1 `tmp` E_2" and "E_2 $\neg$`tmp` E_2" for each extracted pair, where $\neg$`tmp` is the reverse temporal connective (e.g., `after` if `tmp` is `before`). We then fine-tune a pretrained RoBERTa model (`RoBERTa-BASE`) using HuggingFace Transformers (Wolf et al., 2020) via mask language modeling with masking probability $p = 0.1$ for each token. We choose a batch size of 500 and a learning rate of 5×10^{-5}, and train the model to convergence, which was around 135000 iterations when the loss converges to 1.37 from 2.02.

Intervention Generator. Given event E_1, the intervention generator generates a set $\mathcal{A}$ of events that are considered as interventions of the event A in the sense of Section 3.2, which includes manipulating `ARG0`, `V`, and `ARG1` respectively. We achieve this goal by masking these components individually and filling in the masks using an LM. There are several existing works on generating interventions of sentences (Feder et al., 2021), and we select `PolyJuice` (Wu et al., 2021) in our pipeline due to its robustness. `PolyJuice` allows conditional generation via control codes such as `negation`, `lexical`, `resemantic`, `quantifier`, `insert`, `restructure`, `shuffle`, and `delete`, each corresponds to a different manner how the sentence is intervened. We drop the fluency-evaluation component of `PolyJuice` as they will be evaluated by the temporal predictor. We remark that in Figure 1, the intervention is not generated from `PolyJuice`. Nonetheless, such interventions can be produced by more elaborated LMs.

Score Estimation. Given the interventions $\mathcal{A}$ and the sampled covariates $\mathcal{X}$, we can use the temporal predictor to estimate $\mathbb{P}(\mathsf{X} \prec \mathsf{A})$ for all $\mathsf{X} \in \mathcal{X}$ and $\mathsf{A} \in \mathcal{A}$. To obtain temporal propensities $q(\boldsymbol{x}; \mathsf{A})$ for all interventions, we need to estimate $\mathbb{P}(A = 1|X)$ for each X and A. Since by our sampling method, X *occurs* preceding E_1, there is an implicit conditioning on E_1, we may thus set $P(X(0)) = f(\mathsf{X}, \mathsf{E}_1)$ and $P(X(0), A(1)) = f(\mathsf{X}, \mathsf{A})$; we will discuss possible normalizations in Section 5.2. We then form temporal propensity vectors as (recall X is the indicator corresponding to the event X)

$$q(\boldsymbol{x}; \mathsf{A}) = \left(\frac{P(X(0))}{P(X(0), A(1))} \right)_{\mathsf{X} \in \mathcal{X}}. \tag{10}$$

5. Empirical Studies

We put the ROCK framework into action[4], our findings reveal that *although temporality is essential for CCR, without balancing covariates, it is prone to spurious correlations.*

5.1. Setup and Details

Evaluation Datasets. We evaluate the ROCK framework on the Choice of Plausible Alternatives dataset (COPA, Gordon et al., 2012) and a self-constructed dataset of 153 instances using the first dimension (cause-and-effect) of GLUCOSE (GLUCOSE-D1, Mostafazadeh et al., 2020). Each instance in COPA consists of a premise, two plausible choices, and a question type asking which choice is the choice (or effect) of the premise. When asking for cause, we set the premise as E_1, and two choices as E_2 respectively; otherwise we take the premise as E_2 and two choices as E_1 respectively. We choose the choice with the higher score. We evaluate the development set of 100 instances (COPA-DEV) and the test set of 500 instances (COPA-TEST). To construct GLUCOSE-D1, we take the test set and set the cause as premise, the effect and another candidate event as two choices then follow the same procedure.

Baseline Scores and Variants. To test the validity and the effectiveness of ROCK, We compare the adjusted score $\hat{\Delta}_p$ with several other reasonable scores that may be intuitive at first sight.

- L_1-balanced score $\hat{\Delta}_1$: set $p = 1$ in (7).
- L_2-balanced score $\hat{\Delta}_2$: set $p = 2$ in (7).
- Vanilla temporal score $\hat{\Delta}_{\mathsf{E}_1} = \mathbb{P}(\mathsf{E}_1 \prec \mathsf{E}_2)$.
- Unadjusted score $\hat{\Delta}_{\mathcal{A}}$: set $\mathcal{A}' = \mathcal{A}$ in (7).
- Misspecified score $\hat{\Delta}_{\mathcal{X}}$: set $\mathcal{A}' = \mathcal{X}$ in (7).

Here the L_p-balanced scores are those balanced using temporal propensities with L_p norm in Equation (7); the vanilla temporal score is perhaps the most straightforward one, which treats temporal precedence as causation; the unadjusted score is obtained without balancing the covariates; the misspecified score mistakes the covariates for interventions. All these three have intuitive explanations but are either insufficient for CCR or prone to spurious correlations. Note that $\lim_{\epsilon \downarrow 0} \hat{\Delta}_p = \hat{\Delta}_{\mathsf{E}_1}$ (when nothing is kept) and $\lim_{\epsilon \uparrow 1} \hat{\Delta}_p = \hat{\Delta}_{\mathcal{A}}$ (when everything is kept).

5.2. Design Choices and Normalizations

We discuss several design choices and normalizations that might stabilize estimation procedures. We give the complete ablation studies on all combinations of these choices in Section 5.4. We observe that although some of these normalization may benefit CCR on certain datasets, the improvements are *marginal* compared with what temporal propensity matching brings.

Direct Matching (D). In (10), we directly match the vectors of probabilities $(f(\mathsf{A}, \mathsf{X}))_{\mathsf{X} \in \mathcal{X}}$.

Temporality Pre-Filtering (F). As the covariate sampler and temporal predictor are two different LMs, a sampled covariate might not be a preceding event judged by the temporal predictor. We filter the covariates before matching temporal propensities such that $f(\mathsf{X}, \mathsf{E}_1) > f(\mathsf{E}_1, \mathsf{X})$.

[4]Code for the ROCK and for reproducing all results in this paper is available at `github.com:zjiayao/ccr_rock.git`.

	Random Baseline	$\hat{\Delta}_1 \uparrow$ L_1-Balanced	$\hat{\Delta}_2 \uparrow$ L_2-Balanced	$\hat{\Delta}_{\mathsf{E}_1} \uparrow$ Temporal	$\hat{\Delta}_{\mathcal{A}} \uparrow$ Unbalanced	$\hat{\Delta}_{\mathcal{X}} \uparrow$ Misspecified
COPA-DEV	0.5 ± 0.050	0.6900	**0.7000**	0.5800	0.5600	0.5300
COPA-TEST	0.5 ± 0.022	**0.5640**	**0.5640**	0.5200	0.5400	0.5240
GLUCOSE-D1	0.5 ± 0.040	0.6645	**0.6968**	0.5677	0.5742	0.6581
COPA-DEV (-T)	0.5 ± 0.050	0.6200	**0.6300**	0.5300	0.4800	0.5300
COPA-TEST (-T)	0.5 ± 0.022	**0.5800**	0.5740	0.4540	0.4600	0.4860
GLUCOSE-D1 (-T)	0.5 ± 0.040	0.6065	**0.6194**	0.5548	0.4387	0.3742

Table 1: **Best zero-shot results.** Shaded rows have temporal fine-tuning (T) disabled. (i) Estimators with temporal propensities balanced ($\hat{\Delta}_1$ and $\hat{\Delta}_2$) perform consistently better than the unbalanced and the temporal estimators. (ii) In general, without temporality fine-tuning ("-T", see Section 4), the performances degrade.

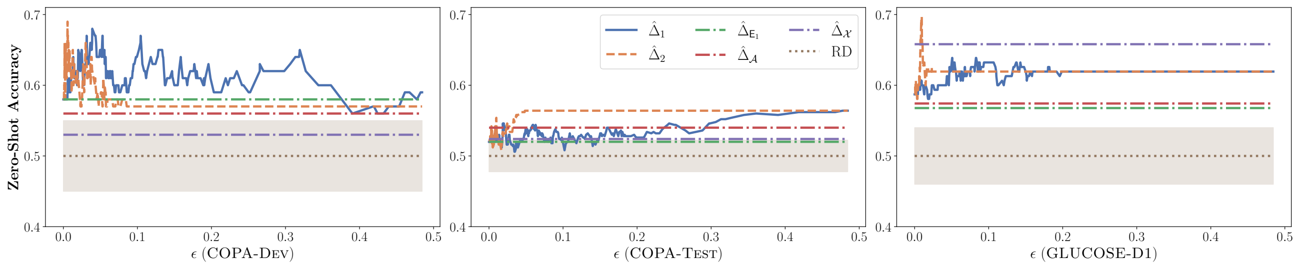

Figure 3: **Best zero-shot result vs ϵ.** Balanced estimators significantly outperform un-balanced and other variants for both COPA-DEV (left), COPA-TEST (middle) and GLUCOSE-D1 (right).

Score Normalization (S). In Section 4 we use $s(\mathsf{E}_1, \mathsf{E}_2)$ for $f(\mathsf{E}_1, \mathsf{E}_2)$, we can also normalize it and form $f(\mathsf{E}_1, \mathsf{E}_2)$ through

$$f(\mathsf{E}_1, \mathsf{E}_2) = \frac{s(\mathsf{E}_1,\mathsf{E}_2)}{s(\mathsf{E}_1,\mathsf{E}_2)+s(\mathsf{E}_2,\mathsf{E}_1)+s(\mathsf{E}_1,\mathsf{N})+s(\mathsf{N},\mathsf{E}_1)}, \tag{11}$$

where N represents the null event when no additional information is given, set as an empty string.

Propensity Normalization (Q). In Equation (10), we can also normalize the estimates first before forming the q vectors via $P(X(0)) = f(\mathsf{X}, \mathsf{E}_1)/\sum_{\mathsf{X}' \in \mathcal{X}} f(\mathsf{X}', \mathsf{E}_1)$ and $P(X(0), A(1)) = f(\mathsf{X}, \mathsf{A})/\sum_{\mathsf{X}' \in \mathcal{X}} f(\mathsf{X}', \mathsf{A})$.

Co-occurrence Stablization (C). The fine-tuned temporal predictor may sometimes still fail to cover the connectives. We can stabilize $\mathbb{P}(\mathsf{X} \prec \mathsf{A})$ by setting it to $(P(A(0), X(1)) + P(X(0), A(1)))/2$.

Estimand Normalization (E). We can normalize the probability $\mathbb{P}(\mathsf{A} \prec \mathsf{B})$ in the estimand Δ by dividing by $(P(A(0), B(1)) + P(B(0), A(1)))$.

5.3. Results

5.3.1. A Concrete Example

We first examine a particular example when the vanilla temporal score $\hat{\Delta}_{\mathsf{E}_1}$ fails but $\hat{\Delta}_1$ does not.

Example 5.1 (Did $\mathsf{E}_1^{(1)}$ or $\mathsf{E}_1^{(2)}$ cause E_2?).

$\mathsf{E}_1^{(1)}$: `I was preparing to wash my hands.`

$\mathsf{E}_1^{(2)}$: `I was preparing to clean the bathroom.`

E_2 : `I put rubber gloves on.`

$\mathsf{A}_{15}^{(1)}$: `I was preparing to wash my feet.`

$\mathsf{A}_5^{(2)}$: `Kevin was preparing to clean the bathroom.`

This is the 63-rd instance in COPA-DEV together a matched intervention (L_2-balancing with optimal ϵ) for each choice. The unadjusted scores are $\hat{\Delta}_{\mathcal{A}}(\mathsf{E}_1^{(1)}, \mathsf{E}_2) \approx 0.036$ and $\hat{\Delta}_{\mathcal{A}}(\mathsf{E}_1^{(2)}, \mathsf{E}_2) \approx 0.035$ while the L_1-balanced scores are $\hat{\Delta}_1(\mathsf{E}_1^{(1)}, \mathsf{E}_2) \approx -0.010$ and $\hat{\Delta}_1(\mathsf{E}_1^{(2)}, \mathsf{E}_2) \approx 0.002$. The balanced score selects the correct choice ($\mathsf{E}_1^{(2)}$) with higher confidence. More details and full examples are given in the Appendix. We should comment that the scores $\hat{\Delta}_1$, $\hat{\Delta}_{\mathcal{X}}$ and $\hat{\Delta}_{\mathsf{E}_1}$ also select the correct answer on this instance; and there are instances where the balanced scores fail. Nonetheless, the performance of balanced scores dominates on average.

5.3.2. Discussion

We show best zero-shot results over design choices (and over ϵ) in Figure 3 and Table 1. As ROCK tackles CCR from a completely new perspective, there are no real baselines to compare with; our goal is to demonstrate that *the causal inference motivated method, temporal propensity* *matching, mitigates spurious correlations* by comparing balanced scores with unbalanced ones. We think this perspective would also benefit the NLP community at large for solving CCR and other related tasks.

Temporal propensity matching is effective. In Table 1 (unshaded rows), we observe that balanced scores have generally better performances on all datasets compared with the temporal estimator and the unadjusted estimator, implying that (i) temporality is important for CCR, yet they are susceptible to spurious correlations; (ii) balancing covariates via matching temporal propensities is effective.

Rules-of-thumb for choosing ϵ. The parameter ϵ controls the threshold of covariates selection and p controls its geometry (see e.g., Hastie et al., 2015). Hinted by Figure 3, a general rule-of-thumb should be $\epsilon < 0.1$. Table B.1 shows optimal ϵ values when constrained to $[0, 0.1]$, where all are global optimal except for COPA-TEST under L_1-balanced score (whose accuracy is 0.552). Hence we recommend setting ϵ to be reasonably small ϵ such as within $(0.01, 0.1)$ when $p = 1$ and relatively smaller such as $(0.005, 0.05)$ when $p = 2$. The optimal value depends on the implementation details of ROCK components and domains of CCR to be performed, yet these choices should provide a good start.

Comparison with existing methods. The self-talk method (Shwartz et al., 2020) achieves 66% on COPA-DEV without external knowledge and 69% when the CoMET-Net (Bosselut et al., 2019) that contains commonsense knowledge is used. Wei et al. (2021) reports 91% on the training set of COPA by using instruction fine-tuning on related datasets. Tamborrino et al. (2020) reports 80% on COPA-TEST by ranking choices using an n-gram based scoring method. ROCK method outperforms self-talk but underperforms (Wei et al., 2021; Tamborrino et al., 2020) in its current form. Nonetheless, our method only requires temporal information provided by the **vanilla** LM without any task-specific fine-tuning, is more interpretable, and provides a prototype for adopting causal inference frameworks to natural language tasks.

5.4. Ablation Studies

Temporality Fine-Tuning. Shaded rows in Table 1 show that when we use the pretrained `RoBERTa-BASE` without temporality fine-tuning (we increase k to 30), almost all estimators do not have decent performance. We conclude that (i) pretrained LMs usually have poor "temporal awareness," and (ii) temporal fine-tuning helps LMs to extract temporal knowledge essential to CCR.

Covariate Set Size. Figure 4 depicts zero-shot results on COPA-TEST against the covariate set size $N = |\mathcal{X}|$ together with 95%-confidence bands. Here we only enable score normalizations (N) among all six normalizations. We observe that in general, increasing covariate set size improves performances if ϵ is reasonable: if ϵ is too small, added covariates may have little impacts while they may introduce more noises if ϵ is too large.

	COPA-DEV		COPA-TEST		GLUCOSE-D1	
	$\hat{\Delta}_1 \uparrow$	$\hat{\Delta}_2 \uparrow$	$\hat{\Delta}_1 \uparrow$	$\hat{\Delta}_2 \uparrow$	$\hat{\Delta}_1 \uparrow$	$\hat{\Delta}_2 \uparrow$
Best	0.6900	0.7000	0.5640	0.5640	0.6645	0.6968
-S	0.01	0.06	-	-	0.08	0.11
-Q	0.01	-	-	-	0.03	-
-C	-	-	0.01	0.01	0.09	0.13
-E	0.01	0.01	-	-	0.03	-

Table 2: **Single-component ablations on normalizations.** Marked in red are percentage decreases compared with the best result (i.e., computed as $(a - b)/a$).

Normalizations. In Section 5.2 we discussed six possible normalizations. We report the best performance when each normalization is removed in Table 2, where red marks the percentage decrease compared with the best result (**D** and **F** not shown as there is no change). Full ablations of all combinations of normalizations and more discussions are given in the Appendix. We observe that (i) certain normalizations benefit certain datasets; (ii) in general, improvements due to normalizations are only *marginal*.

6. Discussions and Open Problems

We articulate the central question of CCR and introduce ROCK, a novel framework for zero-shot CCR, which is the first attempt to incorporate causal inference frameworks in commonsense reasoning. ROCK sheds light on the CCR problem from new perspectives that are arguably more well-founded and demonstrates great potential for zero-shot CCR as shown by empirical studies of various datasets and is on par with existing methods that leverages external causal knowledge on some datasets.

There are several possible avenues for future works. (i) **Prompt engineering** for better temporal predictors and event sampler will likely benefit ROCK. (ii) **Implicit events and reporting biases** in training data are likely to bias the LMs. How to account for implicit events? (iii) **Computing the exact propensity** requires design novel methods to extract many-event temporal relationships and would further improve the performance. (iv) **Investigating implicit biases in the framework.** When the LM is sufficiently large and the pretraining dataset sufficiently diverse, the LM outputs should have reasonably well coverage and less bias due to undercoverage.

Acknowledgements

This work was supported in part by ONR Contract N00015-19-1-2620, NSF through CCF-1934876, an Alfred Sloan

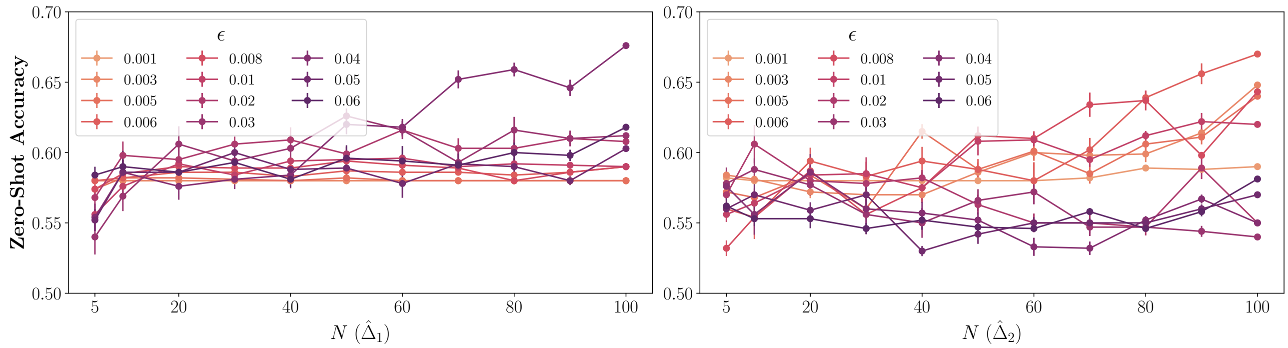

Figure 4: **Zero-shot result on COPA-DEV vs covariate set size $N = |\mathcal{X}|$ with 95%-confidence bands.** In general, using a larger N improves performances for both L_1-balanced score ($\hat{\Delta}_1$, left) and L_2-balanced score ($\hat{\Delta}_2$, right).

Research Fellowship, and the Wharton Dean's Research Fund. We would like to thank Bo Zhang, Rotem Dror, Ben Zhou, and Soham Dan for helpful discussions and feedback on this manuscript. We also thank Shuxiao Chen, Sihao Chen, Vered Schwartz, Haoyu Wang, Haoshu Xu, Diyi Yang, and Yachong Yang for stimulating discussions at various stages of this work.

References

Bhattacharjya, D., Gao, T., and Subramanian, D. Order-dependent event models for agent interactions. In Bessiere, C. (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 1977–1983. International Joint Conferences on Artificial Intelligence Organization, 7 2020. URL https://doi.org/10.24963/ijcai.2020/274.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

Bunge, M. *Causality and modern science*. Routledge, 4 edition, 1979. ISBN 9781315081656.

Chang, D.-S. and Choi, K.-S. Causal relation extraction using cue phrase and lexical pair probabilities. In *Natural Language Processing – IJCNLP 2004*, pp. 61–70. Springer Berlin Heidelberg, 2004.

Cochran, W. G. and Chambers, S. P. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2): 234–266, 1965.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Do, Q. X., Chan, Y. S., and Roth, D. Minimally supervised event causality identification. In *Proceedings of the Conference on EMNLP*. Association for Computational Linguistics, 2011.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., Stewart, B. M., Veitch, V., and Yang, D. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond, 2021.

Fisher, R. A. Cancer and smoking. *Nature*, 182(4635): 596–596, 1958.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. Allennlp: A deep semantic natural language processing platform, 2017.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. A counterfactual simulation model of causal judgments for physical events. *Psychological review*, 2021.

Gordon, A., Kozareva, Z., and Roemmele, M. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings* *of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 394–398, Montréal, Canada, 7 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S12-1052.

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912791.

Han, M. and Wang, Y. Doing good or doing right? exploring the weakness of commonsense causal reasoning models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 151–157, Online, 8 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-short.20.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman Hall, 2015. ISBN 1498712169.

Heckman, J. J. Rejoinder: response to sobel. *Sociological Methodology*, 35(1):135–150, 2005.

Hill, A. B. S. The environment and disease: Association or causation? *Journal of the Royal Society of Medicine*, 58: 295 – 300, 1965.

Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

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

Kang, D., Gangal, V., Lu, A., Chen, Z., and Hovy, E. Detecting and explaining causes from text for a time series event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

Kavumba, P., Inoue, N., Heinzerling, B., Singh, K., Reisert, P., and Inui, K. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pp. 33–42, Hong Kong, China, 11 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-6004.

Keith, K. A., Jensen, D., and O'Connor, B. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Kuipers, B. Commonsense reasoning about causality: deriving behavior from structure. *Artificial intelligence*, 24 (1-3):169–203, 1984.

Luo, Z., Sha, Y., Zhu, K. Q., Hwang, S.-w., and Wang, Z. Commonsense causal reasoning between short texts. In *KR*, pp. 421–431, 2016.

Mill, J. S. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*, volume 1 of *Cambridge Library Collection - Philosophy*. Cambridge University Press, 1851.

Mostafazadeh, N., Kalyanpur, A., Moon, L., Buchanan, D., Berkowitz, L., Biran, O., and Chu-Carroll, J. Glucose: Generalized and contextualized story explanations, 2020.

Neyman, J. S. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Sciences*, 10(4):1–51, 1923.

Ning, Q., Feng, Z., and Roth, D. A Structured Learning Approach to Temporal Relation Extraction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1038–1048, Copenhagen, Denmark, 9 2017. Association for Computational Linguistics. URL http://cogcomp.org/papers/NingFeRo17.pdf.

Ning, Q., Wu, H., Peng, H., and Roth, D. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 841–851, New Orleans, Louisiana, 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1077.

Ning, Q., Feng, Z., Wu, H., and Roth, D. Joint reasoning for temporal and causal relations. *arXiv preprint arXiv:1906.04941*, 2019a.

Ning, Q., Subramanian, S., and Roth, D. An Improved Neural Baseline for Temporal Relation Extraction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019b. URL https://cogcomp.seas.upenn.edu/papers/NingSuRo19.pdf.

O'Gorman, T. J., Wright-Bettner, K., and Palmer, M. Richer event description: Integrating event coreference with temporal, causal and bridging annotation, 2016.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444. URL http://www.jstor.org/stable/2337329.

Pearl, J. and Mackenzie, D. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Peters, J., Janzing, D., and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners, 2019.

Rashkin, H., Sap, M., Allaway, E., Smith, N. A., and Choi, Y. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986. ISSN 0270-0255. doi: https://doi.org/10.1016/0270-0255(86)90088-6. URL https://www.sciencedirect.com/science/article/pii/0270025586900886.

Rosenbaum, P. R. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666, 1984.

Rosenbaum, P. R. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84 (408):1024–1032, 1989.

Rosenbaum, P. R. *Observational Studies*. Springer, 2002.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Roth, D. Incidental Supervision: Moving beyond Supervised Learning. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, 2 2017. URL http://cogcomp.org/papers/Roth-AAAI17-incidental-supervision.pdf.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Rubin, D. B. Bias reduction using mahalanobis-metric matching. *Biometrics*, pp. 293–298, 1980.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Russell, B. On the notion of cause. *Proceedings of the Aristotelian Society*, 13(1):1–26, 1912. ISSN 00667374, 14679264. URL http://www.jstor.org/stable/4543833.

Sandhaus, E. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 2008.

Sap, M., Rashkin, H., Chen, D., Bras, R. L., and Choi, Y. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

Sap, M., Shwartz, V., Bosselut, A., Choi, Y., and Roth, D. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 27–33. Association for Computational Linguistics, 2020.

Shwartz, V., West, P., Le Bras, R., Bhagavatula, C., and Choi, Y. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4615–4629. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.emnlp-main.373.

Staliunaite, I., Gorinski, P. J., and Iacobacci, I. Improving commonsense causal reasoning by adversarial training and data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17630.

Tamborrino, A., Pellicanò, N., Pannier, B., Voitot, P., and Naudin, L. Pre-training is (almost) all you need: An application to commonsense reasoning. *ArXiv*, abs/2004.14074, 2020.

Vashishtha, S., Poliak, A., Lal, Y. K., Van Durme, B., and White, A. S. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4070–4078. Association for Computational Linguistics, 2020.

Wang, B. and Komatsuzaki, A. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, 5 2021.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL https://arxiv.org/abs/2109.01652.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, 10 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6707–6723, Online, 8 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.523.

Zhang, B. and Zhang, J. Some reflections on drawing causal inference using textual data: Parallels between human subjects and organized texts. In *First Conference on Causal Learning and Reasoning*, 2022. URL https://openreview.net/forum?id=ZJRRwV4lCLz.

Zhang, H., Huo, Y., Zhao, X., Song, Y., and Roth, D. Learning contextual causality between daily events from time-consecutive images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1752–1755. IEEE, 6 2021.

Zhou, B., Ning, Q., Khashabi, D., and Roth, D. Temporal common sense acquisition with minimal supervision. *arXiv preprint arXiv:2005.04304*, 2020.

A. Miscellaneous Proofs

We first restate Proposition 3.2 below.

Proposition 3.2 (Expected L_2 error under perfect matching)**.** *Write* $r := r_1 - r_0$, *then* $\Delta = \mathbb{E}[r_1 - r_0] \equiv \mathbb{E}[r]$. *Define*

$$\varrho := \sup_{\tau} \{\tau \leq |r - \mathbb{E}[r|q(\boldsymbol{x})]| \ a.s.\} \in \{0, 1\}. \tag{8}$$

The expected L_2 *error of* $\hat{\Delta} = \mathbb{E}[r|q(\boldsymbol{x})]$ *satisfies*

$$\mathbb{E}[(\hat{\Delta} - \Delta)^2] \leq 1 - \varrho^2. \tag{9}$$

Proof of Proposition 3.2. Recall we write $r := r_1 - r_0$, by the conditional variance decomposition, we have

$$\begin{aligned}\mathbf{Var}(r) =& \mathbb{E}\mathbf{Var}(r|q(\boldsymbol{x})) + \mathbf{Var}\mathbb{E}[r|q(\boldsymbol{x})] \\ =& \mathbb{E}\left[(r - \mathbb{E}[r|q(\boldsymbol{x})])^2\right] \\ &+ \mathbb{E}\left[(\mathbb{E}[r|q(\boldsymbol{x})] - \mathbb{E}[r])^2\right] \\ \geq& \mathbb{E}\left[(\mathbb{E}[r|q(\boldsymbol{x})] - \mathbb{E}[r])^2\right] + \varrho^2.\end{aligned} \tag{A.1}$$

Note that $\mathbf{Var}(r) \leq 1$ since $r \in [0, 1]$, we have the expected L_2 error

$$\mathbb{E}\left[(\mathbb{E}[r|q(\boldsymbol{x})] - \mathbb{E}[r])^2\right] \leq 1 - \varrho^2. \tag{A.2}$$

$\square$

B. Additional Experiment Details

B.1. Rule-of-Thumb for Choosing ϵ

In Table B.1 we show the best ϵ values when constrained in $\epsilon \in [0, 0.1]$. Hence we recommend setting ϵ to be reasonably small ϵ such as within $(0.01, 0.1)$ when $p = 1$ and relatively smaller such as $(0.005, 0.05)$ when $p = 2$. The optimal value depends on the implementation details of ROCK components and domains of CCR to be performed, yet these choices should result in a good start.

B.2. Further Discussions on Temporality Fine-Tuning

In Figure 3, we observe that, counterintuitively, without temporality fine-tuning, the best performances of balanced estimators (0.58) are higher than those with temporality fine-tuning (0.564). Although this gap is within one standard deviation of the random baseline (0.022) thus no statistically significant conclusions can be drawn, but it might hint that pretrained LMs may have already been very aware of termpoality. Is this really the case? A closer look at the full ablation table to be introduced shortly in Table B.5 reveals that the stellar performance is attributed to one particular normalization, estimand normalization (**E**), which was actually detrimental to another dataset (GLUCOSE-D1). Hence we think this normliazation may favor certain dataset over others, thus we think it is not recommendable to include this normalization when dealing with a new dataset.

B.3. Full Ablation on Normalizations

Recall in Section 5.4 we discussed six possible normalizations that may stabilize the estimation procedure:

(D) **Direct Matching:** in (10), instead of forming the temporal propensity vectors $\boldsymbol{q}$ using conditional probabilities, we may directly match the vectors of probabilities $(f(\mathsf{A}, \mathsf{X}))_{\mathsf{X} \in \mathcal{X}}$. This normalization is not well motivated but might be easier to compute under certain circumstances, hence we include it as a comparison.

(F) **Temporality Pre-Filtering:** as the covariate sampler and temporal predictor are two different LMs, a sampled covariate might not be a preceding event judged by the temporal predictor. Thus, we can filter the covariates $\mathcal{X}$ before matching temporal propensities such that we only keep covariates $\mathsf{X} \in \mathcal{X}$ satisfying $f(\mathsf{X}, \mathsf{E}_1) > f(\mathsf{S}, \mathsf{E}_1)$.

(S) **Score Normalization:** in Section 4 we use $s(\mathsf{E}_1, \mathsf{E}_2)$ for $f(\mathsf{E}_1, \mathsf{E}_2)$. We can also normalize it and form $f(\mathsf{E}_1, \mathsf{E}_2)$ through

$$f(\mathsf{E}_1, \mathsf{E}_2) = \frac{s(\mathsf{E}_1, \mathsf{E}_2)}{s(\mathsf{E}_1, \mathsf{E}_2) + s(\mathsf{E}_2, \mathsf{E}_1) + s(\mathsf{E}_1, \mathsf{N}) + s(\mathsf{N}, \mathsf{E}_1)} \tag{B.1}$$

where N represents the null event when no additional information is given, set as an empty string. In practice, this normalization does not differ much from the normalization

$$f(\mathsf{E}_1, \mathsf{E}_2) = \frac{s(\mathsf{E}_1, \mathsf{E}_2)}{s(\mathsf{E}_1, \mathsf{E}_2) + s(\mathsf{E}_2, \mathsf{E}_1)}, \tag{B.2}$$

which does not involve N. However, using N has the benefit of stabilizing the estimate $f(\cdot, \cdot)$ as in rare scenarios $s(\mathsf{E}_1, \mathsf{E}_2)$ and $s(\mathsf{E}_2, \mathsf{E}_1)$ may both close to zero.

(Q) **Propensity Normalization:** in Equation (10), we can also normalize the estimates first before forming the q vectors via

$$\begin{aligned} P(X(0)) &= \frac{f(\mathsf{X}, \mathsf{E}_1)}{\sum_{\mathsf{X}' \in \mathcal{X}} f(\mathsf{X}', \mathsf{E}_1)}, \\ P(X(0), A(1)) &= \frac{f(\mathsf{X}, \mathsf{A})}{\sum_{\mathsf{X}' \in \mathcal{X}} f(\mathsf{X}', \mathsf{A})}, \end{aligned} \tag{B.3}$$

where we estimate $P(X(0))$ as the relative frequency of $X(0)$ among all possible events in $\mathcal{X}$; and $P(X(0), A(1))$ among all possible (X, A) pairs.

(C) **Co-Occurrence Stabilization:** on rare occasions, the fine-tuned temporal predictor may sometimes still fail

	COPA-DEV		COPA-TEST		GLUCOSE-D1	
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_1$	$\hat{\Delta}_2$
ϵ^*	0.043067	0.006029	0.059232	0.048837	0.046643	0.009374

Table B.1: **Best choices of ϵ when $\epsilon < 0.1$.**

	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_{\mathsf{E}_1}$	$\hat{\Delta}_{\mathcal{A}}$	$\hat{\Delta}_{\mathcal{X}}$
$(\mathsf{E}_1, \mathsf{E}_2^{(1)})$	−0.002	−0.002	0.106	0.002	0.106
$(\mathsf{E}_1, \mathsf{E}_2^{(2)})$	−0.001	−0.001	0.086	−0.012	0.086

Table B.2: **Scores for Example B.1.**

to cover the connectives. We can stabilize $\mathbb{P}(\mathsf{X} \prec \mathsf{A})$ by setting it to $(P(A(0), X(1)) + P(X(0), A(1)))/2$. This in effect results in an alternative estimand based on co-occurrences of events (instead of precedence) and can be viewed as a weaker causation in CCR.

(E) **Estimand Normalization:** the score normalization (N) takes place at temporal propensity matching. We can normalize the temporal probability $\mathbb{P}(\mathsf{A} \prec \mathsf{B})$ in the estimand Δ by dividing $(P(A(0), B(1)) + P(B(0), A(1)))$, thus setting

$$\mathbb{P}(\mathsf{A} \prec \mathsf{B}) = \frac{P(A(0),B(1))}{P(A(0),B(1))+P(B(0),A(1))}. \qquad \text{(B.4)}$$

B.3.1. Ablation Results

We report ablations on all possible subset of normalizations together with temporality fine-tuning (-T, see Section 4 in Table B.5. Note that when **D** is enabled, **S** and **Q** are not active and when **C** is enabled, **E** is not active, thus resulting in a total of $2^2(2^2+1)(2^1+1) = 30$ combinations

Ablations resulting in the best performances are highlighted in blue and those resulting in the worst the performances are highlighted in red. Shaded rows are results without temporal fine-tuning (using top $k = 30$ tokens in mask language modeling). We summarize our observations as follows.

Improvements due to normalizations are marginal. The gap between best and worst performance are marginal, except for the GLUCOSE-D1 dataset, which is mainly caused by enabling estimand normalization (**E**. Without considering **E**, the worst result is 0.594 (**+Q** or **+FQ**). Furthermore, we note the gap between the best results and the results under no normalizations ($\emptyset$) is also marginal, indicating that for CCR it is more important to have a well-established baseline and temporal signal extractors than exploring different normalizations.

Furthermore, the outliers are interesting: enabling estimand normaliztion (**E**) has little or no effects on most datasets but can boost the performance on COPA-TEST under non fine-tuned temporal predictors (-T) while is detrimental to GLUCOSE-D1 under fine-tuned temporal predictors.

Rules-of-thumb for choosing normalizations. As a general rule-of-thumb, temporal score normalization (**S**) should be enabled and the q vectors should be properly formed (without direct matching **D**); temporal pre-filtering (**F**) and propensity normalization (**Q**) in general do not affect the results significantly; co-occurrence stabilization (**C**) has greater positive effect on datasets when a weaker notion of causation are desirable (e.g., GLUCOSE-D1 we constructed); while estimand normalization (**E**) improves certain datasets (e.g., COPA-TEST without temporal fine-tuning), it has detrimental effects on some others (e.g., GLUCOSE-D1 with temporal fine-tuning), hence we recommend disabling it by default.

	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_{\mathsf{E}_1}$	$\hat{\Delta}_{\mathcal{A}}$	$\hat{\Delta}_{\mathcal{X}}$
$(\mathsf{E}_1^{(1)}, \mathsf{E}_2)$	−0.010	−0.010	0.068	0.036	0.068
$(\mathsf{E}_1^{(2)}, \mathsf{E}_2)$	0.002	0.001	0.098	0.035	0.098

Table B.3: **Scores for Example B.2.**

	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_{\mathsf{E}_1}$	$\hat{\Delta}_{\mathcal{A}}$	$\hat{\Delta}_{\mathcal{X}}$
$(\mathsf{E}_1^{(1)}, \mathsf{E}_2)$	0.056	−0.001	0.109	0.096	0.109
$(\mathsf{E}_1^{(2)}, \mathsf{E}_2)$	0.005	−0.010	0.279	0.118	0.279

Table B.4: **Scores for Example B.3.**

B.4. Full Examples

We also attach three full examples from our implementation of the ROCK. The problem instances are given below. For each instance, we tabulate 50 covariates sampled, all interventions generated, the corresponding $\|q(\boldsymbol{x}; \mathsf{A}) - q(\boldsymbol{x}; \mathsf{E}_1)\|_p$, and the temporal probabilities $\mathbb{P}(\cdot \prec \mathsf{E}_2)$.

Example B.1 (Did E_1 cause $\mathsf{E}_2^{(1)}$ or $\mathsf{E}_2^{(2)}$?)**.**

E_1 : `The teacher assigned homework to the students.`

$\mathsf{E}_2^{(1)}$: `The students passed notes.`

$\mathsf{E}_2^{(2)}$: `The students groaned.`

This is the 72-nd instance of COPA-DEV, the full tables for inferring the causation from E_1 to $\mathsf{E}_2^{(1)}$ and E_1 to $\mathsf{E}_2^{(2)}$ are given in Table B.6 and Table B.7 respectively. Different scores are shown in Table B.2. Note that this example is not easy.

Example B.2 (Did $\mathsf{E}_1^{(1)}$ or $\mathsf{E}_1^{(2)}$ cause E_2?)**.**

$\mathsf{E}_1^{(1)}$: `I was preparing to wash my hands.`

$\mathsf{E}_1^{(2)}$: `I was preparing to clean the bathroom.`

E_2 : `I put rubber gloves on.`

This is the 63-rd instance of COPA-DEV, the full tables for inferring the causation from $\mathsf{E}_1^{(1)}$ to E_2 and $\mathsf{E}_1^{(1)}$ to $\mathsf{E}_2^{(2)}$ are given in Table B.8 and Table B.9 respectively. Different scores are shown in Table B.3.

Example B.3 (Did $\mathsf{E}_1^{(1)}$ or $\mathsf{E}_1^{(2)}$ cause E_2?)**.**

$\mathsf{E}_1^{(1)}$: `His pocket was filled with coins.`

$\mathsf{E}_1^{(2)}$: `He sewed the hole in his pocket.`

E_2 : `The man's pocket jingled as he walked.`

This is the 79-th instance of COPA-DEV, the full tables for inferring the causation from $\mathsf{E}_1^{(1)}$ to E_2 and $\mathsf{E}_1^{(1)}$ to $\mathsf{E}_2^{(2)}$ are given in Table B.10 and Table B.11 respectively. Different scores are shown in Table B.2.

Dataset	Score	Best	Worst	∅	+D	+F	+S	+Q	+C	+E	+DF	+DC	+DE	+FS	+FQ	+FC
COPA-DEV	$\hat{\Delta}_1 \uparrow$	0.690	0.620	0.670	0.660	0.670	0.680	0.650	0.650	0.680	0.660	0.650	0.680	0.680	0.650	0.650
	$\hat{\Delta}_2 \uparrow$	0.700	0.630	0.630	0.650	0.630	0.690	0.630	0.660	0.640	0.650	0.650	0.680	0.690	0.630	0.660
COPA-TEST	$\hat{\Delta}_1 \uparrow$	0.564	0.528	0.542	0.548	0.542	0.540	0.548	**0.564**	0.554	0.548	**0.564**	0.554	0.540	0.548	**0.564**
	$\hat{\Delta}_2 \uparrow$	0.564	0.526	0.554	0.542	0.554	0.540	0.544	**0.564**	0.548	0.542	**0.564**	0.546	0.540	0.544	**0.564**
GLUCOSE-D1	$\hat{\Delta}_1 \uparrow$	0.665	0.503	0.600	0.606	0.600	0.594	0.594	0.613	0.503	0.606	0.639	0.503	0.594	0.594	0.613
	$\hat{\Delta}_2 \uparrow$	0.697	0.503	0.594	0.600	0.594	0.606	0.594	0.619	0.510	0.600	0.639	0.503	0.606	0.594	0.619
COPA-DEV (-T)	$\hat{\Delta}_1 \uparrow$	0.620	0.550	0.590	0.550	0.590	0.580	0.580	0.570	**0.620**	0.550	0.560	0.610	0.580	0.580	0.570
	$\hat{\Delta}_2 \uparrow$	0.630	0.530	0.610	0.530	0.610	0.600	0.580	0.600	**0.630**	0.530	0.550	0.600	0.600	0.580	0.600
COPA-TEST (-T)	$\hat{\Delta}_1 \uparrow$	0.580	0.484	0.494	0.486	0.494	0.522	0.498	0.484	0.574	0.486	0.496	0.574	0.522	0.498	0.484
	$\hat{\Delta}_2 \uparrow$	0.574	0.484	0.494	0.502	0.494	0.530	0.508	0.484	**0.574**	0.502	0.492	0.570	0.530	0.508	0.484
GLUCOSE-D1 (-T)	$\hat{\Delta}_1 \uparrow$	0.606	0.510	0.568	0.555	0.568	0.574	0.568	0.535	**0.606**	0.555	0.510	0.587	0.574	0.568	0.535
	$\hat{\Delta}_2 \uparrow$	0.619	0.503	0.568	0.555	0.568	0.587	0.561	0.503	**0.619**	0.555	0.516	0.587	0.587	0.561	0.503

Dataset	Score	+FE	+SQ	+SC	+SE	+QC	+QE	+DFC	+DFE	+FSQ	+FSQC	+FSQE	+FQC	+FQE	+SQC	+SQE	+FSQC	+FSQE
COPA-DEV	$\hat{\Delta}_1 \uparrow$	0.680	0.670	0.620	0.680	0.640	0.660	0.650	0.680	0.670	0.660	**0.690**	0.640	0.660	0.660	**0.690**	0.660	**0.690**
	$\hat{\Delta}_2 \uparrow$	0.640	0.670	0.630	**0.700**	0.650	0.660	0.650	0.680	0.670	0.640	**0.700**	0.650	0.660	0.640	**0.700**	0.640	**0.700**
COPA-TEST	$\hat{\Delta}_1 \uparrow$	0.554	0.532	**0.564**	0.532	0.558	0.560	**0.564**	0.554	0.532	0.560	0.528	0.558	0.560	0.560	0.528	0.560	0.528
	$\hat{\Delta}_2 \uparrow$	0.548	0.538	**0.564**	0.534	0.562	0.556	**0.564**	0.546	0.538	0.562	0.526	0.562	0.556	0.562	0.526	0.562	0.526
GLUCOSE-D1	$\hat{\Delta}_1 \uparrow$	0.503	0.594	0.639	0.510	0.613	0.510	0.639	0.503	0.594	**0.665**	0.516	0.613	0.510	**0.665**	0.516	**0.665**	0.516
	$\hat{\Delta}_2 \uparrow$	0.510	0.600	**0.697**	0.510	0.619	0.516	0.639	0.503	0.600	0.690	0.516	0.619	0.516	0.690	0.516	0.690	0.516
COPA-DEV (-T)	$\hat{\Delta}_1 \uparrow$	**0.620**	0.580	0.560	**0.620**	0.560	**0.620**	0.560	0.610	0.580	0.560	0.610	0.560	**0.620**	0.560	0.610	0.560	0.610
	$\hat{\Delta}_2 \uparrow$	**0.630**	0.580	0.580	0.610	0.580	0.620	0.550	0.600	0.580	0.560	0.610	0.580	0.620	0.560	0.610	0.560	0.610
COPA-TEST (-T)	$\hat{\Delta}_1 \uparrow$	0.574	0.514	0.512	0.570	0.506	**0.580**	0.496	0.574	0.514	0.512	0.570	0.506	**0.580**	0.512	0.570	0.512	0.570
	$\hat{\Delta}_2 \uparrow$	**0.574**	0.528	0.524	0.570	0.494	**0.574**	0.492	0.570	0.528	0.522	0.570	0.494	**0.574**	0.522	0.570	0.522	0.570
GLUCOSE-D1 (-T)	$\hat{\Delta}_1 \uparrow$	**0.606**	0.568	0.529	**0.606**	0.542	0.594	0.510	0.587	0.568	0.535	0.600	0.542	0.594	0.535	0.600	0.535	0.600
	$\hat{\Delta}_2 \uparrow$	**0.619**	0.581	0.529	0.613	0.548	0.594	0.516	0.587	0.581	0.535	0.613	0.548	0.594	0.535	0.613	0.535	0.613

Table B.5: **Full ablation studies on normalizations.** Ablations resulting in the best performances are highlighted in blue and those resulting in the worst performances are highlighted in red. Shaded rows are results without temporal fine-tuning (using top $k = 30$ tokens in masked language modeling). (i) The gaps between best and worst performance are marginal, except for the GLUCOSE-D1 dataset, which is mainly due to estimand normalization **E**. Without considering **E**, the worst result is 0.594 (**+Q** or **+FQ**). (ii) In general, temporal fine-tuning helps. The only exception on COPA-TEST is due to estimand normalization (**E**). (iii) As a general rule-of-thumb, it does not hurt to start with no normalizations enabled.

Sampled Covariates $\mathcal{X}$

X_1: He had written a brief book summary of the book and, using a set of questions.
X_2: There was homework help, help desk, and online support.
X_3: No one did the work on time, and no one received good grades for it.
X_5: The kids had to do their school homework online.
X_6: This was the norm.
X_7: The class would sit quietly and listen to their teacher talk.
X_8: It was free time.
X_9: Homework was only assigned when the teacher had a class with a lot of work for the students to.
X_{10}: He did not give homework to his students.
X_{11}: There was a long period of time when nobody ever did any homework.
X_{12}: She had been teaching them during class for weeks.
X_{14}: It was just a fun afternoon with the kids, and then it turned into a time of dr.
X_{16}: Every night, a student would get to work on their homework.
X_{17}: The students had to listen to music and watch a video, respectively, before they could do their.
X_{18}: The students had to do the homework themselves.
X_{19}: The children used to go to school in the morning and study their books until the evening.
X_{20}: The teacher assigned homework to the entire class.
X_{21}: Each student was given a piece of paper with some number on it.
X_{23}: They could play without limits.
X_{25}: I thought that the homework was just a part of my study in each class.
X_{26}: Students who have not completed their homework will not be allowed to go to the next class.
X_{29}: The assignment was simple, they were just to read the assigned reading.
X_{30}: However, he handed out the following set of questions, which the teacher posed one by one to.
X_{31}: Students were not given much homework.
X_{32}: Students were only encouraged to work on assignments and were not explicitly told to do extra.
X_{33}: Only the school's teacher did so.
X_{36}: He would just talk to them or read articles or give his own opinion on the subject.
X_{37}: There was no homework.
X_{38}: The students had to read the textbook and test their knowledge of the material.
X_{41}: They were assigned to do some homework.
X_{42}: Teachers would typically assign the work to the students, but this teacher assigned it to the students and.
X_{43}: There were no homework assignments at all.
X_{44}: The students would go to the Internet and download games.
X_{46}: The assignment had already been completed.
X_{47}: He asked his students on the first day of class to write down on A4 paper any questions.
X_{52}: No homework was assigned.
X_{53}: The students were all in the classroom, sitting in rows like the soldiers in the First World War.
X_{54}: I just gave them a paper with one page written on it.
X_{57}: There was no homework.
X_{58}: The students were told the homework, and the students were to do the homework on their own.

$\|q(\boldsymbol{x};\mathsf{A}) - q(\boldsymbol{x};\mathsf{E}_1)\|_p$	E_1 and Interventions $\mathcal{A}$	$\mathbb{P}(\cdot \prec \mathsf{E}_2)$
0	E_1: The teacher assigned homework to the students.	0.5031
0.0135	A_1: The professor assigned homework to the students.	0.4993
0.0508	A_2: The professor supported the tourists assigned homework to the students.	0.5082
0.0894	A_3: The tourists ran, or the teacher assigned homework to the students.	0.4987
0.0279	A_4: The teacher took homework to the students.	0.5177
0.1053	A_5: The teacher was assigning Justin with the homework to the students.	0.3935
0.1291	A_6: The teacher replaced the carpet for the library last night because the carpet was old homework to the students.	0.5093
0.0591	A_7: The teacher assigned to read the children's book came to the students.	0.5120
0.0365	A_8: The teacher assigned less homework to the students.	0.5135
0.0201	A_9: The teacher assigned tests to the students.	0.4999
0.0870	A_{10}: No one was assigned homework to the students.	0.4886
0.1521	A_{11}: Unless the senator performed, the teacher assigned homework to the students.	0.3867
0.0820	A_{12}: Noelle Leong on the other hand assigned homework to the students.	0.4874
0.1524	A_{13}: The teacher didn't give homework to the students.	0.5324
0.1349	A_{14}: The teacher didn't assigned homework to the students.	0.5198
0.1999	A_{15}: The teacher didn't tell anyone homework to the students.	0.5140
0.0468	A_{16}: The teacher assigned nothing to the students.	0.5175
0.0485	A_{17}: The teacher assigned no children to the students.	0.5230
0.0362	A_{18}: The teacher assigned no class to the students.	0.5167
0.0301	A_{19}: The student assigned homework to the students.	0.4958
0.0135	A_{20}: The professor assigned homework to the students.	0.4993
0.0488	A_{21}: The teacher worked on the algebraic homework to the students.	0.5334
0.0512	A_{22}: The teacher wrote homework to the students.	0.5151
0.0515	A_{23}: The teacher read homework to the students.	0.5231
0.0201	A_{24}: The teacher assigned tests to the students.	0.4999
0.0647	A_{25}: The teacher assigned to the classroom stopped to the students.	0.5244
0.0298	A_{26}: The teacher assigned anger to the students.	0.5065

Table B.6: **Example 1a:** the first plausible pair of the 72-th instance in COPA-DEV, matched interventions are highlighted. Here E_1 : `The teacher assigned homework to the students.` and E_2: `The students passed notes.`

Sampled Covariates $\mathcal{X}$	$\|q(\boldsymbol{x};\mathsf{A}) - q(\boldsymbol{x};\mathsf{E}_1)\|_p$	E_1 and Interventions $\mathcal{A}$	$\mathbb{P}(\cdot \prec \mathsf{E}_2)$
X_1: He had written a brief book summary of the book and, using a set of questions.			
X_2: There was homework help, help desk, and online support.			
X_3: No one did the work on time, and no one received good grades for it.			
X_5: The kids had to do their school homework online.			
X_6: This was the norm.			
X_7: The class would sit quietly and listen to their teacher talk.	0	E_1: The teacher assigned homework to the students.	0.5308
X_8: It was free time.	0.0135	A_1: The professor assigned homework to the students.	0.5263
X_9: Homework was only assigned when the teacher had a class with a lot of work for the students to.	0.0508	A_2: The professor supported the tourists assigned homework to the students.	0.5207
X_{10}: He did not give homework to his students.	0.0894	A_3: The tourists ran, or the teacher assigned homework to the students.	0.5260
X_{11}: There was a long period of time when nobody ever did any homework.	0.0279	A_4: The teacher took homework to the students.	0.5340
X_{12}: She had been teaching them during class for weeks.	0.1053	A_5: The teacher was assigning Justin with the homework to the students.	0.5396
X_{14}: It was just a fun afternoon with the kids, and then it turned into a time of dr.	0.1291	A_6: The teacher replaced the carpet for the library last night because the carpet was old homework to the students.	0.5015
X_{16}: Every night, a student would get to work on their homework.	0.0591	A_7: The teacher assigned to read the children's book came to the students.	0.5346
X_{17}: The students had to listen to music and watch a video, respectively, before they could do their.	0.0365	A_8: The teacher assigned less homework to the students.	0.5515
X_{18}: The students had to do the homework themselves.	0.0201	A_9: The teacher assigned tests to the students.	0.5249
X_{19}: The children used to go to school in the morning and study their books until the evening.	0.0870	A_{10}: No one was assigned homework to the students.	0.5832
X_{20}: The teacher assigned homework to the entire class.	0.1521	A_{11}: Unless the senator performed, the teacher assigned homework to the students.	0.5454
X_{21}: Each student was given a piece of paper with some number on it.	0.0820	A_{12}: Noelle Leong on the other hand assigned homework to the students.	0.5288
X_{23}: They could play without limits.	0.1524	A_{13}: The teacher didn't give homework to the students.	0.5914
X_{25}: I thought that the homework was just a part of my study in each class.	0.1349	A_{14}: The teacher didn't assigned homework to the students.	0.5956
X_{26}: Students who have not completed their homework will not be allowed to go to the next class.	0.1999	A_{15}: The teacher didn't tell anyone homework to the students.	0.6164
X_{29}: The assignment was simple, they were just to read the assigned reading.	0.0468	A_{16}: The teacher assigned nothing to the students.	0.5487
X_{30}: However, he handed out the following set of questions, which the teacher posed one by one to.	0.0485	A_{17}: The teacher assigned no children to the students.	0.5566
X_{31}: Students were not given much homework.	0.0362	A_{18}: The teacher assigned no class to the students.	0.5477
X_{32}: Students were only encouraged to work on assignments and were not explicitly told to do extra.	0.0301	A_{19}: The student assigned homework to the students.	0.5180
X_{33}: Only the school's teacher did so.	0.0135	A_{20}: The professor assigned homework to the students.	0.5263
X_{36}: He would just talk to them or read articles or give his own opinion on the subject.	0.0488	A_{21}: The teacher worked on the algebraic homework to the students.	0.5349
X_{37}: There was no homework.	0.0512	A_{22}: The teacher wrote homework to the students.	0.5318
X_{38}: The students had to read the textbook and test their knowledge of the material.	0.0515	A_{23}: The teacher read homework to the students.	0.5370
X_{41}: They were assigned to do some homework.	0.0201	A_{24}: The teacher assigned tests to the students.	0.5249
X_{42}: Teachers would typically assign the work to the students, but this teacher assigned it to the students and.	0.0647	A_{25}: The teacher assigned to the classroom stopped to the students.	0.5263
X_{43}: There were no homework assignments at all.	0.0298	A_{26}: The teacher assigned anger to the students.	0.5291
X_{44}: The students would go to the Internet and download games.			
X_{46}: The assignment had already been completed.			
X_{47}: He asked his students on the first day of class to write down on A4 paper any questions.			
X_{52}: No homework was assigned.			
X_{53}: The students were all in the classroom, sitting in rows like the soldiers in the First World War.			
X_{54}: I just gave them a paper with one page written on it.			
X_{57}: There was no homework.			
X_{58}: The students were told the homework, and the students were to do the homework on their own.			

Table B.7: **Example 1b:** the second plausible pair of the 72-th instance in COPA-DEV, matched interventions are highlighted. Here E_1 : `The teacher assigned homework to the students.` and E_2: `The students groaned.`

Sampled Covariates $\mathcal{X}$	$\|q(\boldsymbol{x};\mathsf{A}) - q(\boldsymbol{x};\mathsf{E}_1)\|_p$	E_1 and Interventions $\mathcal{A}$	$\mathbb{P}(\cdot \prec \mathsf{E}_2)$
X_1: I had scrubbed my face, arms, and chest, using a baby shampoo called "Sun.	0	E_1: I was preparing to wash my hands.	0.4847
X_2: There was the bathroom, and that was a little bit trickier.	0.2485	A_1: I was running low and got my hands wet but not the shoes because the hands were preparing to wash my hands.	0.4177
X_3: I had got dressed.	0.1792	A_2: I was standing close to the sink and was suddenly wet from the rain because the sink was well lit preparing to wash my hands.	0.4972
X_7: I had been standing up because my knees hurt, and they were stiff.	0.2153	A_3: I wanted to get rid of the smell of bleach and use water instead because the water was clean and preparing to wash my hands.	0.3779
X_8: I had put on a new pair of latex gloves–I'm very careful about hand cleaning–.	0.0752	A_4: The person was preparing to wash my hands.	0.4754
X_9: I wanted to take out my medicine and check all my symptoms.	0.1014	A_5: i've was preparing to wash my hands.	0.4966
X_{11}: I had put a couple of paper towels in the drawer by the sink.	0.1210	A_6: I guess was preparing to wash my hands.	0.3801
X_{12}: I had been sitting in the armchair by the fire.	0.1067	A_7: I was about to start using dish soap to wash my hands.	0.4440
X_{13}: I had decided to make a cup of tea.	0.1424	A_8: I was going to wash my hands.	0.3987
X_{14}: I'd been playing with my son, watching an old video on YouTube, and I.	0.3054	A_9: I was late for work so I was running and picking up the dishes. So I got to wash my hands.	0.3302
X_{16}: I had been brushing the sand from my clothes.	0.0700	A_{10}: EMPTY was preparing to wash my hands.	0.4601
X_{18}: I went through the washing ceremony to check the level of purity in my body, I washed my.	0.0000	A_{11}: I was preparing to wash my hands.	0.4847
X_{20}: I washed my face.	0.0586	A_{12}: I was preparing to use lube to get out of my hands.	0.4764
X_{21}: I had just finished eating my breakfast.	0.0912	A_{13}: I was preparing to take a vitamin c and a calcium supplement. I took my hands.	0.4764
X_{25}: I prepared a simple salad and some rolls on the table.	0.1014	A_{14}: I was preparing to cook dinner my hands.	0.4861
X_{26}: I scrubbed my hands with a little bit of soap,.	0.0157	A_{15}: I was preparing to wash my feet.	0.4917
X_{27}: I turned to the side of the mirror, and I had a look.	0.0398	A_{16}: I was preparing to wash my face and hair.	0.4923
X_{29}: I always take my shoes off.	0.0324	A_{17}: I was preparing to wash the clothes.	0.5002
X_{30}: However, I removed some leftover food from the table, where the two men had been eating.	0.1373	A_{18}: I didn't preparing to wash my hands.	0.4339
X_{31}: I took a few deep breaths and had a conversation with my heart.	0.1263	A_{19}: I don't know how to describe an incredible preparing to wash my hands.	0.4227
X_{32}: I had changed into the outfit I was wearing: a pretty, pale pink T-shirt and.	0.1601	A_{20}: I was not preparing to wash my hands.	0.4513
X_{33}: I was clearing away breakfast things.	0.3740	A_{21}: Maybe it's because my dog died recently, or because my wife was sick that week. I don't know, but I was preparing to wash my hands.	0.2685
X_{34}: I used to take a shower, and now it was time to do that again.	0.3322	A_{22}: I didn't wash my hands with soap but instead used conditioner because the soap was preparing to wash my hands.	0.2946
X_{35}: I'd washed my face.	0.1219	A_{23}: I can't wash my hands was preparing to wash my hands.	0.2412
X_{36}: I needed to check my phone.	0.1294	A_{24}: I was not supposed to wash my hands.	0.4955
X_{38}: I'd taken the time to put on another pair of socks, and the socks for that matter.	0.4810	A_{25}: I was not supposed to be doing dishes after dinner, so I was going to wash my hands.	0.3228
X_{41}: I washed my hands more than a thousand times.	0.0940	A_{26}: I was not sure if I needed to use soap or vinegar to clean to wash my hands.	0.4754
X_{42}: I needed to put on a gown and cap, and to check the medications I had received for.	0.0700	A_{27}: EMPTY was preparing to wash my hands.	0.4601
X_{43}: I'd been sitting at my desk, answering emails and making phone calls.	0.2772	A_{28}: No part of the preparation except was preparing to wash my hands.	0.4667
X_{44}: I'd touched the wall for some reason.	0.2068	A_{29}: However, as I mentioned already, time to was preparing to wash my hands.	0.4536
X_{46}: I used to dry them properly.	0.0970	A_{30}: I was preparing to wash my eyes but switched to water my hands.	0.4884
X_{49}: I made sure my hands were clean.	0.0000	A_{31}: I was preparing to wash my hands.	0.4847
X_{52}: As a last resort, I would always scrub the top of my hands with a nail brush to.	0.1389	A_{32}: I was preparing to skip the soap and water. I couldn't because the soap wasn't needed for my hands.	0.4039
X_{53}: I had looked into the bathroom mirror.	0.1404	A_{33}: I was preparing to wash my hands and I was doing too much.	0.4780
X_{56}: I was talking to him.	0.0324	A_{34}: I was preparing to wash the clothes.	0.5002
X_{57}: I'd decided to drink some water, which was, of course, a bad idea.	0.2359	A_{35}: I was preparing to wash my hands but had to get water since the hands were not touching.	0.4759
X_{58}: I was standing in the room, with the window open.	0.0919	A_{36}: I tried to wash preparing to wash my hands.	0.5058
X_{59}: Of course, I had put my coat on.	0.0966	A_{37}: I needed to get rid of my feet preparing to wash my hands.	0.4861
X_{60}: I used to wash my hands.	0.0000	A_{38}: I was preparing to wash my hands.	0.4847
X_{61}: Though, I'd pulled a towel off the rack and was drying my hair.	0.0682	A_{39}: he was preparing to wash my hands.	0.4971
X_{63}: I was taking a shower.	0.0000	A_{40}: I was preparing to wash my hands.	0.4847
X_{65}: I'd been wiping my palms on the sides of shorts and shirt, like a dirty secret.	0.1424	A_{41}: I was going to wash my hands.	0.3987
X_{67}: I had been holding a glass of orange juice, which I had drained, and a bowl of.	0.0325	A_{42}: I was preparing to cook to wash my hands.	0.4913
X_{68}: I brushed my teeth and brushed my hair, I even dried my hair, and then I went.	0.0000	A_{43}: I was preparing to wash my hands.	0.4847
X_{69}: I would take a breath.	0.0659	A_{44}: It was preparing to wash my hands.	0.4992
X_{70}: I'd brushed my teeth, put on deodorant, shaved, dried my hair.	0.0784	A_{45}: The towel was preparing to wash my hands.	0.5118
X_{71}: I had brushed my teeth, applied makeup, and removed my contacts.	0.0489	A_{46}: I was preparing to put my hands.	0.4752
X_{72}: I was making some tea.	0.0783	A_{47}: I was preparing to ditch my hands.	0.4679
X_{73}: I'd removed all my jewelry.	0.0000	A_{48}: I was preparing to wash my hands.	0.4847
X_{74}: I had to take my shoes off.			

Table B.8: **Example 2a:** the first plausible pair of the 63-th instance in COPA-DEV, matched interventions are highlighted. Here E_1 : I was preparing to wash my hands. and E_2: I put rubber gloves on.

Sampled Covariates $\mathcal{X}$	$\|q(\boldsymbol{x};\mathsf{A}) - q(\boldsymbol{x};\mathsf{E}_1)\|_p$	E_1 and Interventions $\mathcal{A}$	$\mathbb{P}(\cdot \prec \mathsf{E}_2)$
X_{1}: I had scrubbed the kitchen floor and the sink and, uh, that kind of thing.	0	E_{1}: I was preparing to clean the bathroom.	0.5023
X_{2}: There was the need to remove the rubbish from the front garden.	0.2191	A_{1}: I was building the car instead of the house since the house was incompatible with cleanliness. preparing to clean the bathroom.	0.3765
X_{7}: I had been walking up and down the hall, running my hands over the wood-paneled.	0.1228	A_{2}: I was so bad at the job that I was preparing to clean the bathroom.	0.4935
X_{8}: I had put a load of clothes, some books and some DVD's into the washing machine.	0.0598	A_{3}: I was doing a job preparing to clean the bathroom.	0.5171
X_{9}: I wanted to take out the trash and empty all the containers.	0.0703	A_{4}: A woman was preparing to clean the bathroom.	0.5083
X_{11}: I had put a couple of washcloths on the bathroom counter.	0.0325	A_{5}: Kevin was preparing to clean the bathroom.	0.4889
X_{12}: I had to deal with the dirty clothes hamper and the clothes on the floor.	0.0651	A_{6}: Emily was preparing to clean the bathroom.	0.4522
X_{13}: I had decided to clean the stove.	0.1247	A_{7}: I was going to take a bath instead of to clean the bathroom.	0.3744
X_{14}: I decided to take a short break and get some ice-cream.	0.1402	A_{8}: I was able to do the cleaning in time, and was pretty good at it, although the to clean the bathroom.	0.3202
X_{16}: I had vacuumed the room.	0.0952	A_{9}: I was going to clean the bathroom.	0.4437
X_{18}: I went down to the living room and turned on the TV.	0.0000	A_{10}: I was preparing to clean the bathroom.	0.5023
X_{20}: I needed to find a bottle opener.	0.0678	A_{11}: Bill was preparing to clean the bathroom.	0.4967
X_{25}: I prepared a simple salad and some crackers and cheeses in our very cute wooden bowl.	0.0513	A_{12}: EMPTY was preparing to clean the bathroom.	0.4727
X_{26}: I scrubbed down the sink, the counter, the tile and my hands.	0.1011	A_{13}: I was preparing to cook dinner the bathroom.	0.5102
X_{27}: I turned on the TV to catch my favourite news programme, when the host came on and said.	0.0583	A_{14}: I was preparing to sleep the bathroom.	0.4908
X_{29}: I always take my shower.	0.1601	A_{15}: I was preparing to cook dinner for my family. the bathroom.	0.4288
X_{31}: I took a bath and washed my hair.	0.0437	A_{16}: I was preparing to clean the kitchen floor.	0.5098
X_{32}: I had to clean the house.	0.0549	A_{17}: I was preparing to clean the kitchen table.	0.5073
X_{33}: I showered, put on a clean pair of pants and a shirt.	0.0798	A_{18}: I wasn't preparing to clean the bathroom.	0.4573
X_{34}: I did my normal routine.	0.1496	A_{19}: I didn't want to wash either the towels or the sponge. I was preparing to clean the bathroom.	0.3829
X_{35}: I'd washed the coffee table, and before that, I'd vacuumed the floor.	0.0645	A_{20}: I was not preparing to clean the bathroom.	0.4836
X_{36}: I needed to flush the toilet.	0.2527	A_{21}: When I was done, I was preparing to clean the bathroom.	0.2748
X_{38}: I went to the kitchen to put away another load of dishes.	0.0717	A_{22}: I wasn't was preparing to clean the bathroom.	0.4897
X_{42}: I needed to put on a pair of rubber gloves.	0.0505	A_{23}: no one was preparing to clean the bathroom.	0.4948
X_{43}: I'd already wiped the kitchen counter.	0.1418	A_{24}: I was not rushing too much and didn't get to clean the bathroom.	0.4809
X_{44}: I would take out the trash.	0.1548	A_{25}: I was not able to do the dishes, so I had to do the dishes. I had no problem washing to clean the bathroom.	0.4008
X_{46}: I used to clean the living room and the kitchen, and even the bathroom sometimes.	0.0765	A_{26}: I was not going to to clean the bathroom.	0.4419
X_{49}: I made sure the refrigerator was stocked.	0.1146	A_{27}: I was not supposed to was preparing to clean the bathroom.	0.5107
X_{52}: As a last resort, I would always open the medicine cabinet and remove any expired birth	0.0505	A_{28}: no one was preparing to clean the bathroom.	0.4948
control pills.	0.0503	A_{29}: No one was preparing to clean the bathroom.	0.4911
X_{53}: I had to clear away the table, then rinse dishes and clean the table.	0.1011	A_{30}: I was preparing to cook dinner the bathroom.	0.5102
X_{54}: I checked on the kids, who were doing what they normally did.	0.0773	A_{31}: I was preparing to skip the whole the bathroom.	0.4662
X_{56}: The kitchen.	0.1399	A_{32}: I was preparing to wash my hands, but I forgot to take the bathroom.	0.4426
X_{57}: I'd decided to organize the cabinets in the kitchen, because organizing might calm me down.	0.0549	A_{33}: I was preparing to clean the kitchen table.	0.5073
X_{58}: I was doing laundry.	0.1130	A_{34}: I was preparing to clean the bathroom all the time. I was not able to clean the bathroom all of the time and it would take forever.	0.4897
X_{59}: Of course, I had put my makeup on.	0.0430	A_{35}: I was preparing to clean the kitchen counter.	0.5059
X_{60}: I turned on the radio.	0.0000	A_{36}: I was preparing to clean the bathroom.	0.5023
X_{61}: Though, I'd go to the kitchen and make a small snack for myself.	0.0814	A_{37}: I smelled the cookies and wondered if mom was preparing to clean the bathroom.	0.4850
X_{63}: I was taking a shower.	0.1086	A_{38}: I knew it could be used by anyone, just a stranger, preparing to clean the bathroom.	0.4470
X_{65}: I made myself a green smoothie.	0.0651	A_{39}: Emily was preparing to clean the bathroom.	0.4522
X_{67}: I would remove the dishes stored in the kitchen, and then, at last, I would clean.	0.0890	A_{40}: I was showering was preparing to clean the bathroom.	0.5224
X_{68}: I brushed my teeth and took a shower, but I was not very interested in either task.	0.0454	A_{41}: she was preparing to clean the bathroom.	0.5137
X_{69}: I would take a good look at all of the objects in the room, cataloging and organizing.	0.3539	A_{42}: I was hoping someone would fill me in on the details, so I tried to write to clean the bathroom.	0.3144
X_{70}: I'd brushed my teeth, put on deodorant and makeup, and made sure.	0.0952	A_{43}: I was going to clean the bathroom.	0.4437
X_{71}: I took a shower.	0.1586	A_{44}: I was going through the old photos and couldn't decide which mirror to use for my mirror. The Mirror was too old to clean the bathroom.	0.4813
X_{72}: I was making some soup.	0.0000	A_{45}: I was preparing to clean the bathroom.	0.5023
X_{74}: I had to take the covers off of my bed and the bedspread.	0.0293	A_{46}: I was preparing to put the clothes on the bathroom.	0.5046
X_{75}: I had to clear all the stuff out of my room.	0.0931	A_{47}: I was preparing to put in the sewing machine the bathroom.	0.4958
X_{80}: I took a shower and had a leisurely breakfast.	0.1011	A_{48}: I was preparing to cook dinner the bathroom.	0.5102
X_{81}: I'd had a shower, washed my face, dried it with a towel then put.			
X_{83}: I was going to get us a drink of water and maybe some dinner.			

Table B.9: **Example 2b:** the second plausible pair of the 63-th instance in COPA-DEV, matched interventions are highlighted. Here E_1 : `I was preparing to clean the bathroom.` and E_2: `I put rubber gloves on.`

Sampled Covariates $\mathcal{X}$	$\lVert q(\boldsymbol{x};\mathsf{A}) - q(\boldsymbol{x};\mathsf{E}_1)\rVert_p$	E_1 and Interventions $\mathcal{A}$	$\mathbb{P}(\cdot \prec \mathsf{E}_2)$
X_1: He'd had nothing in his pockets but his father's pocket watch, and some old coins he'd.			
X_2: There had been the time he'd been a little boy, about four years old, and.			
X_7: He had been a very small fish in a very small pond.			
X_8: It had contained a knife and a set of keys—but there was nothing in it now.			
X_9: It seemed only to contain his breath and blood.			
X_{10}: He was a slave, and his owner used him roughly when displeased.	0	E_1: His pocket was filled with coins.	0.2980
X_{11}: He had put a couple of coins in his pouch.	0.1069	A_1: His pocket however had been filled with coins.	0.1494
X_{12}: His wallet had been empty.	0.0706	A_2: His pocket contained nine corsage filled with coins.	0.3912
X_{14}: It was filled with crumpled-up bills.	0.0620	A_3: His pocket had a large amount of space filled with coins.	0.3036
X_{16}: He had been a slave-hunter to the west, a murderer in the service of his city.	0.1457	A_4: His pocket was lost when he accidentally took some with coins.	0.0620
X_{17}: It had been full of notes.	0.0626	A_5: His pocket was empty with coins.	0.2749
X_{18}: It had been half-filled with tobacco and a dirty handkerchief.	0.0938	A_6: His pocket was filled with sandals and at least one with coins.	0.2597
X_{20}: He had been working as a waiter at the Côte du Rhône at the Hotel Mont.	0.0747	A_7: A cowboy on the back of a wagon was filled with coins.	0.2331
X_{21}: He couldn't even have given the goldfish.	0.0364	A_8: A pocket iron was filled with coins.	0.1933
X_{22}: It was empty, but he could think of no problem more urgent than collecting them.	0.1158	A_9: Mark was filled with coins.	0.1925
X_{23}: He'd been a beggar.	0.2276	A_{10}: His pocket did not work as well as the shirt, because the shirt was filled with coins.	0.0994
X_{25}: He had been a farmer.	0.1331	A_{11}: His pocket wasn't filled with coins.	0.1870
X_{29}: The contents of his pockets would have been only a pair of pants, a shirt and, if it.	0.1170	A_{12}: His pocket was not filled with coins.	0.2477
X_{30}: He'd hidden them at the old folk's home, where he'd lived.	0.1539	A_{13}: His pocket was empty but his pocket had nine with coins.	0.1345
X_{32}: He had only had the money to keep him going.	0.1590	A_{14}: His pocket was not holding the lotion with coins.	0.0834
X_{33}: He didn't feel too well.	0.1126	A_{15}: His pocket was not touched with coins.	0.1852
X_{36}: He'd been just like them.	0.3496	A_{16}: No matter how you feel about country music [...] the fact that it featured the really catchy John Denver does not appeal to me was filled with coins.	0.0209
X_{38}: The police had come and taken the man's wallet.	0.1112	A_{17}: No pocket book was filled with coins.	0.1971
X_{41}: He had been wearing a thick bracelet with a chain and gold links, a gift from the wife of.	0.0897	A_{18}: Nothing was filled with coins.	0.2345
X_{44}: He'd been wearing a plain suit and polished his shoes all the time, that would no longer do.	0.0000	A_{19}: His pocket was filled with coins.	0.2980
X_{46}: He used to keep spare coins in his shirt pocket, and use those for the fare.	0.0820	A_{20}: His pocket had been filled with coins.	0.3352
X_{49}: He'd been a thief, a beggar and a soldier.	0.0239	A_{21}: His pocket was heavily filled with coins.	0.2535
X_{52}: No one had seen it.	0.2621	A_{22}: His pocket was ruined and he decided to find a wallet instead because the pocket might have coins with coins.	0.0913
X_{53}: He had not worked for a long time.	0.0233	A_{23}: His pocket was full with coins.	0.3068
X_{56}: The two had been making their way through a series of shops.	0.0481	A_{24}: Her bag was filled with coins.	0.1762
X_{57}: He'd been standing with his back to the wall, holding an umbrella over his head.	0.0753	A_{25}: Someone was filled with coins.	0.2306
X_{58}: He had been trying to buy himself with the money that came from the sale of his books.	0.0370	A_{26}: Her wallet was filled with coins.	0.2127
X_{59}: Of course, he had been a slave.			
X_{60}: He used to have some.			
X_{61}: He'd carried his entire life in his pockets.			
X_{63}: He was wearing a shapeless dark coat with a dark suit underneath, a black hat, and.			
X_{65}: He'd been a poor kid that the state had taken from his mother and put in one home after.			
X_{68}: A few years back, he had an old steel frame.			
X_{69}: He would have sold his coat, and his shirt, and then the little shoes upon his feet.			
X_{70}: He must have been a poor man, and probably very pious.			

Table B.10: **Example 3a:** the first plausible pair of the 79-th instance in COPA-DEV, matched interventions are highlighted. Here E_1 : `His pocket was filled with coins.` and E_2: `The man's pocket jingled as he walked.`

Sampled Covariates $\mathcal{X}$	$\|q(\boldsymbol{x};\mathsf{A}) - q(\boldsymbol{x};\mathsf{E}_1)\|_p$	E_1 and Interventions $\mathcal{A}$	$\mathbb{P}(\cdot \prec \mathsf{E}_2)$
X_{1}: He'd had nothing in his pocket to worry about.			
X_{2}: There had been no hole, just a thin line of cloth.			
X_{7}: He cut off all his fingers, including those on his left hand.			
X_{8}: He had put a knife in the pocket of his coveralls.			
X_{10}: He was a young, handsome fellow with dark blue eyes and black curly hair.			
X_{11}: He had put a couple of sticks in his pouch.	0	E_1: He sewed the hole in his pocket.	0.4818
X_{12}: His wallet had been in his hand, and he had thrown the wallet on the ground as.	0.1075	A_{1}: Then he sewed the hole in his pocket.	0.3724
X_{16}: He had been a stranger to himself.	0.2456	A_{2}: The boy was grumpy in high school, but happy at school, so the teacher taught him sewed the hole in his pocket.	0.1477
X_{17}: It had been in his mouth.	0.0682	A_{3}: He cut pieces from the plate sewed the hole in his pocket.	0.5023
X_{18}: He went down to the river to check the damage.	0.0660	A_{4}: He stuffed the toy gun with the hole in his pocket.	0.5053
X_{20}: He stuffed a paper bag in place of his wallet.	0.0875	A_{5}: He burnt the hole by pulling the hole in his pocket.	0.3826
X_{22}: It was a secret, what with the police and all.	0.0849	A_{6}: He pulled off the blanket and got a the hole in his pocket.	0.4970
X_{25}: He had been afraid to kill anyone.	0.2149	A_{7}: He sewed the quilt better than a teepee because the teepee was a sloppy job in his pocket.	0.2377
X_{30}: He'd hidden his heart, but not in a safe.	0.0930	A_{8}: He sewed with a towel more in his pocket.	0.3728
X_{32}: He'd thought she'd just pulled it out of thin air, but there it was.	0.0707	A_{9}: He sewed the hole with a wire instead of a plier in his pocket.	0.4299
X_{33}: He didn't feel too bad about taking the wallet from your wallet, but after he se.	0.1181	A_{10}: He couldn't sewed the hole in his pocket.	0.3049
X_{36}: He'd been just like any other.	0.2376	A_{11}: No matter how you feel about country music(I for one can't stand it despite my Houston roots), this only instilled sewed the hole in his pocket.	0.1855
X_{38}: The hole had been on the inside of his coat.	0.1161	A_{12}: No one knew how to sewed the hole in his pocket.	0.2986
X_{41}: He had been wearing his uniform cap with the rank insignia, and he put it on.	0.1128	A_{13}: He never filled the hole in his pocket.	0.2366
X_{46}: He hadn't even looked at it.	0.1479	A_{14}: He couldn't bend the iron rod and instead tied the hole in his pocket.	0.2361
X_{49}: He'd sewn up the hole in his leg.	0.0869	A_{15}: He had the hole in his pocket.	0.3282
X_{51}: There was nothing in it, not even the thorns, and there was nothing there but.	0.1401	A_{16}: He sewed no better than the tleilaxu which cut off his eye in his pocket.	0.1917
X_{52}: No one had seen it.	0.3286	A_{17}: He sewed no better with the machine than with the method, because the machine was not precise in his pocket.	0.0641
X_{53}: He had not wanted to talk about it.	0.0871	A_{18}: He sewed not only the hole but also the whole ball inside the hole in his pocket.	0.4182
X_{56}: The two men had argued.	0.0764	A_{19}: Jack sewed the hole in his pocket.	0.4594
X_{57}: He'd been thinking of the boy's parents.	0.0456	A_{20}: Someone sewed the hole in his pocket.	0.4086
X_{58}: He had been trying to reach the hospital.	0.0715	A_{21}: He screwed up sewed the hole in his pocket.	0.4874
X_{63}: He had put a small packet of powder in her shoe, had used a hairpin to.	0.1560	A_{22}: He flunked out of high school, ended up in a strange town, and started writing about the weird the hole in his pocket.	0.3031
X_{65}: He'd been a good kid that the others seemed to like.	0.0614	A_{23}: He stabbed Wisbech with a mop the hole in his pocket.	0.4916
X_{70}: He had hidden the bullet in his leg.	0.0555	A_{24}: He pulled the hole in his pocket.	0.4673
X_{71}: As a little girl, before she had even known what it meant to be.	0.0588	A_{25}: He sewed the turkey with a T-shirt in his pocket.	0.5179
X_{72}: He'd hidden it in the bottom of a pot of ointment.	0.0311	A_{26}: He sewed the bell necklace in his pocket.	0.4765
X_{75}: He had done nothing; but he could do nothing now but lie and wait, and be.	0.0499	A_{27}: He sewed the rope with a chisel in his pocket.	0.4956
X_{80}: It was not easy to find a place to hide the gun and if he was asked about.			
X_{81}: He'd had a small knife to cut his clothes, but he didn'.			
X_{85}: He always bought new jeans every time he went to Kmart.			
X_{86}: He'd had nothing.			
X_{87}: He had sewn the other pocket.			
X_{88}: The pocket was for the book.			
X_{90}: He had to get the tire.			

Table B.11: **Example 3b:** the second plausible pair of the 79-th instance in COPA-DEV, matched interventions are highlighted. Here E_1 : `He sewed the hole in his pocket.` and E_2: `The man's pocket jingled as he walked.`