ROCK: Causal Inference Principles for Reasoning about Commonsense Causality

Jiayao Zhang1 2 Hongming Zhang1 3 Weijie J. Su2 Dan Roth1 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 Reason O(About) Commonsense K(Causality), 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

Figure 1: An Example of CCR: does E1 cause E2? The temporal order E1 ≺ E2 does not necessitate causation due to confounding co-occurrences (e.g., X1). Since when conditioning on X1, a comparable intervention A1 of E1 also precedes E2, the effect from E1 to E2 shrinks.

that are deemed reasonable by an average person.

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, E1, to another, E2. For example, in Figure 1, is Alice’s “entering a restaurant” (E1) a plausible cause for her “ordering a pizza” (E2)? Although the precedence from E1 to E2 is logical, it might be less a “cause” compared with Alice’s “feeling hungry” (X1).

Temporality informs causation, but it is still unclear how to account for confounding co-occurrences (such as X1 in Figure 1). Motivated by causal inference principles (Section 2), we formulate CCR as estimating the change in the likelihood of E2’s occurrence due to intervening E1 (denoted by ¬E1):

\[
\Delta = \mathbb{P}(E_1 \prec E_2) - \mathbb{P}(\neg E_1 \prec E_2)
\]

(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 \(\Delta\) measures the average treatment effect (ATE): its magnitude signifies the strength of the effect and its sign informs the direction. For example, when \(\Delta\) is close to \(-1\), E1 has a strong effect on E2 towards making E2 less prone to occurring. If the occurrences of E1 and ¬E1 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.,

\[E_1: Alice entered a restaurant.\]

\[E_2: Alice ordered a pizza.\]

\[X_1: Alice felt hungry.\]

\[A_1: Alice opened a food-delivery app.\]
X₁), one needs to balance the covariates (events that precede E₁) 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 E₁), 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).
3. The ROCK Framework

Notations. We use sans-serif letters for events, upper-case serif letters for indicators of whether the corresponding event occurs,1 and lower-case serif letters for the realizations of those indicators. For example, in Figures 1 and 2, E1 : “Alice walked into a restaurant...” E1 = I{E1 occurs} and e1,i = I{E1 occurs to the i-th study unit}2. We view the occurrence of events as point processes E(t) on t ∈ {0, 1} (e.g., present versus past). We use E1 > E2 (resp. ¬) to indicate that E1 occurs following (resp. preceding) E2. We write P(E1 < E2) = P(E1(0), E2(1)) and P(E2|E1) = P(E2(1)|E1(0)). We write P for estimates of P, and omit measure-theoretic details3.

Overview of the ROCK framework. We set the stage in this section and discuss implementation details in Section 4. Given E1 and E2, as shown in Figure 2, ROCK samples the covariates set X and interventions set A, from which a matched subset A' is selected via temporal propensities (Section 3.4). The score ∆ is then estimated by Equation (7).

3.1. The Central Question of CCR

Given two specific events E1 and E2, as discussed in Section 1, we articulate CCR as the estimation of the change of temporal likelihood had E1 been intervened:

\[ \Delta = P(E_1 < E_2) - P(-E_1 < E_2) \]  

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

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 E be the set of commonsense events we consider, the probability space we are working on is \( (E × E, σ(E × E), P) \).

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 \( ¬E_1 \).

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 E1 and E2, 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 \( E_{1i} = I\{E_{1i} = E_1\} \), realizations of the covariates \( x_i = (x_{ij})_{j=1}^N \) for \( x_{ij} = I\{X_j < E_{1i}\} \), and two potential outcomes

\[
\begin{align*}
  r_{0i} &= I\{E_{1i}, E_1=0 < E_2\}, \\
  r_{1i} &= I\{E_{1i}, E_1=1 < E_2\}.
\end{align*}
\]

Here \( E_{1i}, E_1=1-E_{1i} \) 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 estimator \( \Delta \) in Equation (1) is indeed the average treatment effect

\[ \Delta = E[r_1 - r_0] = P(E_1 < E_2) - P(-E_1 < E_2). \]  

This identification naturally complies with the temporal na-
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 (\( \neg E_1 \)), we can thus compare \( P(E_1 \prec E_2) \) with \( P(\neg E_1 \prec E_2) \) 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 \( \Delta \) in Equation (1) is feasible only in an idealized 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 (\( \neg E_1 \)). 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 \( \neg E_1 \), but \( E_2 \), 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.,

\[
\mathbb{E}_x \left[ P(E_1 \prec E_2|x) - P(\neg E_1 \prec E_2|x) \right],
\]

provided that the treatment assignment is strongly ignorable with respect to \( x \), 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 \( x \).

**Remark.** (i) We should define \( x \) 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, \( \Delta \) thus estimated quantifies the effect from \( E_1 \) to \( E_2 \) without passing through \( X' \). (ii) Although \( x \) 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 propensity score can be used for this purpose. The propensity score \( p(x) = P(E_1(1) = 1|x(0)) \) is the probability of \( E_1 \) occurring at time 1 conditioning on the covariates being \( x \) 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 \( p(x) \) is not feasible, as will be discussed in Section 4, instead, we propose to use a surrogate which we call temporal propensities:

\[
q(x) = q(x; E_1) = (P(E_1(1) = 1|x))_{x \in \mathbb{R}}
\]

with each coordinate measuring the conditional probability of the event \( E_1 \) given an event in \( x \). Thus motivated, for some fixed threshold \( \epsilon \) and \( p \in \{1, 2\} \), we will use following estimating equation for the \( L_p \)-balanced score, where \( f(E_1; E_2) \) is an estimate for \( P(E_1 \prec E_2) \):

\[
\hat{\Delta}_p = f(E_1; E_2) - \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}'} f(A; E_2),
\]

\[
\mathcal{A} := \{ A \in \mathcal{A} : \frac{1}{|\mathcal{A}|} \|q(x; A) - q(x; E_1)\|_p \leq \epsilon \}.
\]

### 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 \( q(x) \) 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 \( \Delta = E[r_1 - r_0] \equiv E[r] \). Define

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

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

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

The proof is due to the conditional variance decomposition and is given in the Appendix. The parameter \( \varrho \) 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 \( x \). Intuitively, the worst-case scenario \( \varrho = 0 \) is uncommon, since this happens only if \( r \) is a function of \( q(x) \). When a large number of diverse covariates are sampled, \( \varrho \) 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 \( \Delta \) 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 \( f(X_1, X_2) \) given two input events \( X_1 \) and \( X_2 \) is an estimate of the temporal probability \( P(X_1 < X_2) \); 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 \( q(x; A) \in \mathbb{R}^{|\mathcal{X}|} \) for each sampled interventions \( A \in \mathcal{A} \) then estimates \( \Delta \) 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 \( X \).

**Temporal Prediction.** Given two events \( E_1 \) and \( E_2 \), we use masked language modeling to predict their temporal relation by forming the prompt \( E_1 \langle MASK \rangle E_2 \) and collect the score \( s_a(E_1, E_2) \) and \( s_b(E_1, E_2) \) for the tokens after and before. We then symmetrize the estimates to form \( s(E_1, E_2) = \frac{1}{2}(s_a(E_1, E_2) + s_b(E_2, E_1)) \). We can directly use \( s(E_1, E_2) \) for \( f(E_1, E_2) \); 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 \( \text{tmp} \) (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 \( (E_1, E_2, \text{tmp}) \). We are able to extract 397174 such pairs and construct them into the fine-tuning dataset consisting of \( “E_1 \text{ tmp } E_2” \) and \( “E_2 \neg \text{tmp } E_2” \) for each extracted pair, where \( \neg \text{tmp} \) is the reverse temporal connective (e.g., after if \( \text{tmp} \) 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 \( 5 \times 10^{-5} \), 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 $A$ 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 $A$ and the sampled covariates $X$, we can use the temporal predictor to estimate $P(X \prec A)$ for all $X \in X'$ and $A \in A$. To obtain temporal propensities $q(x; A)$ for all interventions, we need to estimate $P(A = 1 | X)$ 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 $P(X(0)) = f(X, E_1)$ and $P(X(0), A(1)) = f(X, A)$; 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 $X$)

$$q(x; A) = \left( \frac{P(X(0))}{P(X(0), A(1))} \right)_{x \in X'}.$$  

(10)

5. **Empirical Studies**

We put the ROCK framework into action⁴, 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 $\Delta_{p}$ 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 $\hat{\Delta}_{E_1} = P(E_1 \prec E_2)$.
- Unadjusted score $\hat{\Delta}_A$: set $A' = A$ in (7).
- Misspecified score $\hat{\Delta}_{X}$: set $A' = X$ 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 $\lim_{t \to 0} \hat{\Delta}_p = \hat{\Delta}_{E_1}$ (when nothing is kept) and $\lim_{t \to 1} \hat{\Delta}_p = \hat{\Delta}_A$ (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 $(f(A, X))_{x \in X'}$.

**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 $f(X, E_1) > f(E_1, X)$.

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

|                | Random Baseline | $\Delta_1 \uparrow$ | $\Delta_2 \uparrow$ | $\Delta_{E_1} \uparrow$ | $\Delta_{A} \uparrow$ | $\Delta_{X} \uparrow$ |
|----------------|-----------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| 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               |

Figure 3: Best zero-shot result vs $\epsilon$. 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 $s(E_1, E_2)$ for $f(E_1, E_2)$, we can also normalize it and form $f(E_1, E_2)$ through

$$f(E_1, E_2) = \frac{s(E_1, E_2)}{s(E_1, E_2) + s(E_2, E_1) + s(E_1, N) + s(N, E_1)},$$

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 $P(X(0)) = f(X, E_1) / \sum_{X' \in X} f(X', E_1)$ and $P(X(0), A(1)) = f(X, A) / \sum_{X' \in X} f(X', A)$.

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

**Estimand Normalization (E).** We can normalize the probability $P(A \prec B)$ in the estimand $\Delta$ 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 $\Delta_{E_1}$ fails but $\Delta_1$ does not.

Example 5.1 (Did $E_1^{(1)}$ or $E_1^{(2)}$ cause $E_2$?).

$E_1^{(1)}$: I was preparing to wash my hands.

$E_1^{(2)}$: I was preparing to wash the bathroom.

$E_2$: I put rubber gloves on.

$A_1^{(1)}$: I was preparing to wash my feet.

$A_1^{(2)}$: Kevin was preparing to clean the bathroom.

This is the 63rd instance in COPA-DEV together a matched intervention ($L_2$-balancing with optimal $\epsilon$) for each choice. The unadjusted scores are $\hat{\Delta}_A(E_1^{(1)}, E_2) \approx 0.036$ and $\hat{\Delta}_A(E_1^{(2)}, E_2) \approx 0.035$ while the $L_1$-balanced scores are $\hat{\Delta}(E_1^{(1)}, E_2) \approx -0.010$ and $\hat{\Delta}(E_1^{(2)}, E_2) \approx 0.002$. The balanced score selects the correct choice ($E_1^{(2)}$) with higher confidence. More details and full examples are given in the Appendix. We should comment that the scores $\hat{\Delta}_1$, $\hat{\Delta}_X$ and $\hat{\Delta}_{E_1}$ 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 $\epsilon$) 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 $\epsilon$. The parameter $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ to be reasonably small $\epsilon$ 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 $N = |X|$ 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 $\epsilon$ is reasonable: if $\epsilon$ is too small, added covariates may have little impacts while they may introduce more noises if $\epsilon$ is too large.

### 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 N00019-19-1-2620, NSF through CCF-1934876, an Alfred Sloan Foundation grant.
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 ($\Delta_1$, left) and $L_2$-balanced score ($\Delta_2$, 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

Bhattacharyya, 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.

Kavuumba, P., Inoue, N., Heinnerling, 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/D19-6084.

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=ZJRRwV41CLz.

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 $\Delta = \mathbb{E}[r_1 - r_0] = \mathbb{E}[r]$. Define

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

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

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

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

$$\text{Var}(r) = \mathbb{E}[\text{Var}(r|q(x))] + \text{Var}[\mathbb{E}[r|q(x)]]$$

$$= \mathbb{E}\left[(r - \mathbb{E}[r|q(x)])^2\right]$$

$$+ \mathbb{E}\left[(\mathbb{E}[r|q(x)] - \mathbb{E}[r])^2\right] \hfill (A.1)$$

$$\geq \mathbb{E}\left[(\mathbb{E}[r|q(x)] - \mathbb{E}[r])^2\right] + \varrho^2.$$  

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

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

\hfill $\square$

B. Additional Experiment Details

B.1. Rule-of-Thumb for Choosing $\epsilon$

In Table B.1 we show the best $\epsilon$ values when constrained in $\epsilon \in [0,0.1]$. Hence we recommend setting $\epsilon$ to be reasonably small $\epsilon$ 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 temporality. 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 normalization 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 $q$ using conditional probabilities, we may directly match the vectors of probabilities $(f(A,X))_{X \in \mathcal{X}}$. 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 $X \in \mathcal{X}$ satisfying $f(X,E_1) > f(S,E_1)$.

(S) Score Normalization: in Section 4 we use $s(E_1,E_2)$ for $f(E_1,E_2)$. We can also normalize it and form $f(E_1,E_2)$ through

$$f(E_1,E_2) = \frac{s(E_1,E_2)}{s(E_1,E_2) + s(E_2,E_1)} \hfill (B.1)$$

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

$$f(E_1,E_2) = \frac{s(E_1,E_2)}{s(E_1,E_2) + s(E_2,E_1)} \hfill (B.2)$$

which does not involve $N$. However, using $N$ has the benefit of stabilizing the estimate $f(\cdot,\cdot)$ as in rare scenarios $s(E_1,E_2)$ and $s(E_2,E_1)$ may both close to zero.

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

$$P(X(0)) = \frac{f(X,E_1)}{\sum_{X' \in \mathcal{X}} f(X',E_1)} \hfill (B.3)$$

$$P(X(0), A(1)) = \frac{f(X,A)}{\sum_{X' \in \mathcal{X}} f(X',A)}$$

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
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 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 normalization \((E)\) 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.

### 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 \(||q(x;A) - q(x;E_1)||_p\), and the temporal probabilities \(P(\cdot \prec E_2)\).

#### Example B.1 (Did \(E_1\) cause \(E_2^{(1)}\) or \(E_2^{(2)}\)?)

\[ E_1 : \text{The teacher assigned homework to the students.} \]
\[ E_2^{(1)} : \text{The students passed notes.} \]
\[ E_2^{(2)} : \text{The students groaned.} \]

This is the 72nd instance of COPA-DEV, the full tables for inferring the causation from \(E_1\) to \(E_2^{(1)}\) and \(E_1\) to \(E_2^{(2)}\) 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.

---

| COPA-DEV | COPA-TEST | GLUCOSE-D1 |
|----------|-----------|------------|
| \(\Delta_1\) | \(\Delta_2\) | \(\Delta_{E_2}\) | \(\Delta_A\) | \(\Delta_X\) |
| \(\Delta_1\) | \(\Delta_2\) | \(\Delta_{E_2}\) | \(\Delta_A\) | \(\Delta_X\) |

Table B.1: Best choices of \(\epsilon\) when \(\epsilon < 0.1\).

| \(\epsilon^*\) | 0.043067 | 0.006029 | 0.059232 | 0.048837 | 0.046643 | 0.009374 |

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

| \(E_1\), \(E_1^{(2)}\) | \(\Delta_1\) | \(\Delta_2\) | \(\Delta_{E_2}\) | \(\Delta_A\) | \(\Delta_X\) |
|-----------------|--------|--------|----------|--------|--------|
| \(-0.002\) | -0.002 | 0.106 | 0.002 | 0.106 |
| \(-0.001\) | -0.001 | 0.086 | -0.012 | 0.086 |

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

| \(E_1^{(1)}\), \(E_2\) | \(\Delta_1\) | \(\Delta_2\) | \(\Delta_{E_2}\) | \(\Delta_A\) | \(\Delta_X\) |
|-----------------|--------|--------|----------|--------|--------|
| \(0.056\) | -0.001 | 0.109 | 0.096 | 0.109 |
| \(0.005\) | -0.010 | 0.279 | 0.118 | 0.279 |

Table B.4: Scores for Example B.3.
Example B.2 (Did $E_1^{(1)}$ or $E_1^{(2)}$ cause $E_2$?).

$E_1^{(1)}$: I was preparing to wash my hands.
$E_1^{(2)}$: I was preparing to clean the bathroom.
$E_2$: I put rubber gloves on.

This is the 63-nd instance of COPA-DEV, the full tables for inferring the causation from $E_1^{(1)}$ to $E_2$ and $E_1^{(1)}$ to $E_2^{(2)}$ are given in Table B.8 and Table B.9 respectively. Different scores are shown in Table B.3.

Example B.3 (Did $E_1^{(1)}$ or $E_1^{(2)}$ cause $E_2$?).

$E_1^{(1)}$: His pocket was filled with coins.
$E_1^{(2)}$: 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 $E_1^{(1)}$ to $E_2$ and $E_1^{(1)}$ to $E_2^{(2)}$ are given in Table B.10 and Table B.11 respectively. Different scores are shown in Table B.2.
| Dataset        | Score | Best | Word | -$D$ | -$E$ | -$F$ | -$G$ | -$Q$ | -$C$ | -$DE$ | -$FS$ | -$FE$ | -$SC$ | -$QE$ |
|---------------|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| COPA-Dev      | Δ₁  1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
|               | Δ₄  0.710 | 0.180 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 |
| COPA-Test     | Δ₁  0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 | 0.560 |
|               | Δ₄  0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 | 0.380 |
| GLUCOSE-D1    | Δ₁  0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 | 0.650 |
|               | Δ₄  0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 | 0.450 |

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.
X: He had written a brief book summary of the book and, using a set of questions.
X: He asked the students on the first day of class to write down on A4 paper any questions.
X: The students were all in left-handed sitting positions like the soldiers in the First World War.
X: He just gave them a paper with one page written on it.
X: There was no homework.
X: The students were told the homework, and the students were to do the homework or their own.

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.
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.
I was preparing to wash my hands. I had put a couple of paper towels in the drawer by the sink. I had brushed my teeth, applied makeup, and removed my contacts. I was going to wash my hands. I was about to start using dish soap to wash my hands. I had been sitting in the armchair by the fire. I had got dressed. I washed my face. I had been playing with my son, watching an old video on YouTube, and I. I used to take a shower, and now it was time to do that again. I didn't wash my hands with soap but instead used conditioner because the soap was preparing to wash my hands. I had been sitting at my desk, answering emails and making phone calls. I had looked into the bathroom mirror. I would take a breath. I'd been sitting at my desk, answering emails and making phone calls. 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 E₁: I was preparing to wash my hands. and E₂: I put rubber gloves on.

| Sampled sentences | (X;A); g(E₁); g(E₂); E₁; and E₂; z(E₁); z(E₂) | E1 | E2 |
|-------------------|---------------------------------------------|----|----|
| X1: I had washed my face, arms, and chest, using a baby shampoo called "Sun." | 0 | 0.0947 | 0.047 |
| X4: There was the bathroom, and that was a little bit tricky. | 0.0205 | 0.0727 | 0.047 |
| X1: I had washed my face, arms, and chest, using a baby shampoo called "Sun." | 0.0159 | 0.047 | 0.0659 |
| X2: 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. | 0.0131 | 0.0727 | 0.047 |
| X6: I had been standing up because I knew hurt, and they were off. | 0.0131 | 0.047 | 0.0659 |
| X8: I had put on a pair of fingerless gloves—very careful about hand cleaning. | 0.0131 | 0.0727 | 0.047 |
| X9: I wanted to take out the medicine and check all my symptoms. | 0.0131 | 0.047 | 0.0659 |
| X10: I had cut a couple of paper towels in the drawer by the sink. | 0.0131 | 0.0727 | 0.047 |
| X11: I had been sitting in the armchair by the fire. | 0.0131 | 0.047 | 0.0659 |
| X12: I had washed my face. | 0.0131 | 0.0727 | 0.047 |
| X13: I was preparing to wash my hands. | 0.0131 | 0.047 | 0.0659 |
| X14: 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. | 0.0131 | 0.0727 | 0.047 |
| X15: X16: I was preparing to wash my hands. | 0.0131 | 0.047 | 0.0659 |
| X17: X18: I was preparing to wash my hands. | 0.0131 | 0.0727 | 0.047 |
| X19: X20: I was preparing to wash my hands. | 0.0131 | 0.047 | 0.0659 |
| X21: X22: I was preparing to wash my hands. | 0.0131 | 0.0727 | 0.047 |
| X23: X24: I was preparing to wash my hands. | 0.0131 | 0.047 | 0.0659 |
| X25: X26: I was preparing to wash my hands. | 0.0131 | 0.0727 | 0.047 |
| X27: X28: I was preparing to wash my hands. | 0.0131 | 0.047 | 0.0659 |
| X29: X30: I was preparing to wash my hands. | 0.0131 | 0.0727 | 0.047 |
| X31: X32: I was preparing to wash my hands. | 0.0131 | 0.047 | 0.0659 |
| X33: X34: I was preparing to wash my hands. | 0.0131 | 0.0727 | 0.047 |
| X35: X36: I was preparing to wash my hands. | 0.0131 | 0.047 | 0.0659 |

Table B.8: Example 2a: the first plausible pair of the 63-th instance in COPA-DEV, matched interventions are highlighted. Here E₁: I was preparing to wash my hands. and E₂: I put rubber gloves on.
I was preparing to clean the bathroom.

I had scrubbed the kitchen floor and the sink and, uh, that kind of thing.

I had been walking up and down the hall, running my hands over the wood-paneled walls.

I was preparing to clean the bathroom.

I had been walking up and down the hall, running my hands over the wood-paneled walls.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

I was preparing to clean the bathroom.

Table B.9: Example 2b: the second plausible pair of the 63-th instance in COPA-DEV, matched interventions are highlighted. Here E₁: I was preparing to clean the bathroom. and E₂: I put rubber gloves on.
He'd had nothing in his pockets but his father's pocket watch, and some old coins he'd.

There had been the time he'd been a little boy, about four years old, and.

He'd been standing with his back to the wall, holding an umbrella over his head.

He had been trying to buy himself with the money that came from the sale of his books.

He used to have coins.

He used to have coins.

He had put a couple of coins in his pouch.

His pocket had been filled with coins.

His pocket contained nine coins filled with coins.

It was empty, but he could think of no problem more urgent than collecting them.

He had put a couple of coins in his pouch.

His pocket was lost when he accidentally took some coins.

It was full of notes.

His pocket was filled with sandals and at least one coin.

It was empty, but he could think of no problem more urgent than collecting them.

He'd hidden them at the old folk's home, where he'd lived.

He'd been wearing a thick bracelet with a chain and gold links, a gift from the wife of.

He used to keep spare coins in his shirt pocket, and use those for the fare.

He used to have coins.

He was a slave, and his owner used him roughly when displeased.

He would have sold his coat, and his shirt, and then the little shoes upon his feet.

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.
X: He had nothing in his pocket to worry about.
X: Never been to a hole, just a thin line of cloth.
X: He cut off all his fingers, including those on his left hand.
X: He had put a knife in the pocket of his coveralls.
X: He was a young, handsome fellow with dark blue eyes and black curly hair.
X: He had put a couple of sticks in his pouch.
X: It was a secret, what with the police and all.
X: He had been afraid to tell anyone.
X: He'd hidden his heart, but not in a safe.
X: He'd been thinking of the boy's parents.
X: He'd been a stranger to himself.
X: He had put a small packet of powder in her shoe, had used a hairpin to.
X: He had put the bolt in his leg.
X: He'd sewn up the hole in his leg.
X: He'd hidden it in the bottom of a pot of oatmeal.
X: He had eaten nothing, but he could do nothing now but lie and wait, and be.
X: He had been thinking of the boy's parents.
X: The two men had argued.
X: He'd been a good kid that the others seemed to like.
X: He had brought the bolts in his leg.
X: As a little girl, before she had even known what it meant to be.
X: She had thought she'd just pulled it out of this air, but there it was.
X: He'd had a small knife to cut his clothes, but he didn't.
X: He always bought new jeans every time he went to Kmart.
X: He had sewn the other pocket.
X: The pouch was for the book.
X: He had 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 E1: He sewed the hole in his pocket. and E2: The man's pocket jingled as he walked.