Event Causality Identification with Synthetic Control

Haoyu Wang[1*], **Fengze Liu**[1], **Jiayao Zhang**[1], **Dan Roth**[1], **Kyle Richardson**[2]
[1]Department of Computer and Information Science, UPenn
[2]Allen Institute for AI
{why16gzl,lclarice,zjiayao,danroth}@seas.upenn.edu, {kyler}@allenai.org

Abstract

Event causality identification (ECI), a process that extracts causal relations between events from text, is crucial for distinguishing causation from correlation. Traditional approaches to ECI have primarily utilized linguistic patterns and multi-hop relational inference, risking false causality identification due to informal usage of causality and specious graphical inference. In this paper, we adopt the Rubin Causal Model to identify event causality: given two temporally ordered events, we see the first event as the treatment and the second one as the observed outcome. Determining their causality involves manipulating the treatment and estimating the resultant change in the likelihood of the outcome. Given that it is only possible to implement manipulation conceptually in the text domain, as a work-around, we try to find a 'twin' for the protagonist from existing corpora. This 'twin' should have identical life experiences with the protagonist before the treatment but undergoes an intervention of treatment. However, the practical difficulty of locating such a match limits its feasibility. Addressing this issue, we use the **synthetic control method** to generate such a 'twin' from relevant historical data, leveraging text embedding synthesis and inversion techniques. This approach allows us to identify causal relations more robustly than previous methods, including GPT-4, which is demonstrated on a causality benchmark, COPES-hard.

1 Introduction

Previous endeavours in event causality identification in text have, to a large extent, depended on feature-based approaches where linguistic patterns serve as a crucial role (Beamer and Girju, 2009; Do et al., 2011; Hidey and McKeown, 2016; Lai et al., 2022). These patterns can roughly indicate causal relations in that causal language is widely used in an informal way in everyday life (Imbens and Rubin, 2015). Without proper manipulation of the potential cause, and comparison between the observed outcome and the intervened outcome, these approaches often identify specious causal relations. For example, *"because"* is often considered as a causal indicator (Hidey and McKeown, 2016), yet it might not be rigorous as in the case of "She got a nice job because she graduated from one of the top universities." It is very possible that the employer paid more attention to the candidate's ability, rather than just the educational background in offering a job. Regardless of how these linguistic features are obtained - whether extracted from causal keywords and semantic indications, or obtained from language models - any feature-based approach may be predisposed to bias due to their unreliable causality foundations. Thus, highly sophisticated methods that rely on multi-hop reasoning on graphs for ECI (Cao et al., 2021; Chen et al., 2022; Liu et al., 2023; Chen et al., 2023; Pu et al., 2024), also risk being fundamentally flawed in their conclusions.

If we want to reliably discover event causality, say whether there exists a causal relation between a pair of temporally ordered events (e_1, e_2), we need to manipulate e_1 and see if e_2 still happens in a 'parallel universe' in which e_1 does not happen. In other words, we want to find a 'twin' for the protagonist p in the events, who has identical life experiences with p (i.e., a sequence of events) up to the point when e_1 takes place, but instead undergoes an intervention of e_1. However, it is almost impossible to implement this idea in the text domain (e.g., stories, narratives, and news reports): for any event in text, it is very rare that its protagonist has 'twins' that satisfy the aforementioned requirements. In this work, as a workaround, we attempt to synthesize 'twins' by merging relevant event sequences retrieved from a corpus, inspired by a causal inference method,

[*]Work done during internship at Allen Institute for AI.

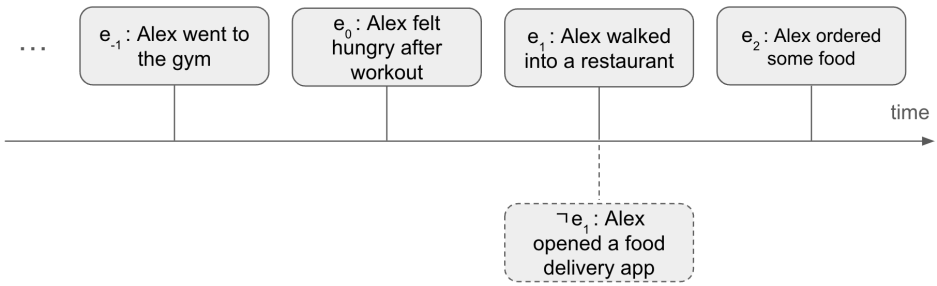

Figure 1: An example illustrating the temporal ordering of treatment event e_1, observed outcome e_2, and pretreatment events e_{-1}, e_0 (covariates) on a time axis. To figure out if e_1 causes e_2, we come up with an intervention of e_1, $\neg e_1$, and find that it does not affect the likelihood of e_2. And thus, e_1 is not the cause of e_2.

called synthetic control, used in economics and social sciences (Abadie and Gardeazabal, 2003; Abadie et al., 2010; Abadie, 2021). Such event sequence merging is also seen in recent effort in event schema induction in information extraction studies (Li et al., 2020; Wen et al., 2021; Du et al., 2022; Dror et al., 2023).

Specifically, our approach consists of three components: (1) noncontemporary control group retrieval, (2) control unit synthesis, and (3) treatment effect estimation. Given a pair of events with certain context, the first component retrieves relevant event sequences from historical data that can be leveraged to synthesize 'twins' via text embedding and inversion techniques (Morris et al., 2023, 2024) in the second component. And this is followed by the third component which calculates a causal estimand to determine whether there exists a causal relation between the two events.

The proposed methodology fundamentally shifts from conventional ECI methods by introducing the concept of synthetic control to the text domain. This allows the inherent linguistic bias which is prevalent in data-driven ECI approaches to be significantly mitigated. Moreover, by introducing full-context matching in a continuous space, we overcome the limitation of discrete temporal propensity matching proposed in previous attempts (Zhang et al., 2022; Wang et al., 2023) of solving ECI with the potential-outcome framework. Our approach demonstrates significantly improved results on the COPES-hard dataset, a commonly used causality benchmark, by at least 9% (relatively) over existing methods and GPT-4. The contribution of this paper is threefold:

- Synthetic control method is introduced to solve ECI in text for the first time.
- Full-context matching is proposed to synthesize control units with the help of recent language modeling techniques.
- Experimental results on the COPES-hard dataset demonstrate the effectiveness and robustness of counterfactual reasoning in text.

2 Preliminaries

In this section, we present the fundamentals of our method, which is grounded on the Rubin Causal Model (Rubin, 1974) and discuss its previous application to the problem of event causality identification in the text domain. And then we discuss the limitation of previous attempts and introduce why we adopt synthetic control in this work.

2.1 Rubin Causal Model

The Rubin Causal Model (RCM) is one of the cornerstones of causal inference. To illustrate this framework in the text domain, let us consider two temporally ordered events (e_1, e_2) in an article. They involve a common protagonist, or study unit, p, and we want to estimate whether e_1 causes e_2, with a context that can be modeled as a temporally ordered sequence of event mentions in text: $e_{-t}, e_{-t+1}, \cdots, e_0$ (see Figure 1 as an example). Following Zhang et al. (2022)'s formulation of ECI, we see the first event e_1 as the treatment, and second event e_2 as the observed outcome. To measure the treatment effect, we need to compare the study unit with a control group within which the control units did not undergo event e_1, and estimate the change of the likelihood of e_2 *had* e_1 *been intervened*:

$$\Delta = \mathbb{P}(e_1 \prec e_2) - \mathbb{P}(\neg e_1 \prec e_2). \quad (1)$$

Here we use $\prec$ to indicate that e_1 occurs before e_2, and $\neg e_1$ to denote a manipulation of e_1, which can only be conceptual or imaginary.

2.2 Temporal Propensity Matching

The most significant challenge in formulating ECI as described above is the spurious correlations introduced by pervasive confounding co-occurrences. They need to be eliminated for an unbiased estimation of the causal estimand introduced in Equation (1). This can be done by balancing events that precede e_1, or *covariates*. Several techniques for balancing covariates (Cochran and Chambers, 1965; Rosenbaum and Rubin, 1983; Pearl, 1995) have been proposed, e.g., sub-classification, matched sampling, covariance adjustment, and propensity score. Zhang et al. (2022) propose to use *temporal propsensities* for covariate balancing in text. To this end, Equation (1) can be rewritten as

$$\Delta = \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{P}(e_1 \prec e_2 | \boldsymbol{x}) - \mathbb{P}(\neg e_1 \prec e_2 | \boldsymbol{x}) \right], \quad (2)$$

and here the treatment assignment is strongly ignorable with respect to the covariates $\boldsymbol{x} = [e_{-t}, e_{-t+1}, \cdots, e_0]$. The propensity score,

$$p(x) = \mathbb{P}(e_1 | \boldsymbol{x}), \quad (3)$$

is the probability of e_1 occurring at time 1 conditioning on the covariates being $\boldsymbol{x}$ at time equal to or less than 0 (prior to the time e_1 happens). To incorporate the context of e_1, Wang et al. (2023) design a mechanism to sample diversified covariates from multiple timestamps and also use temporal propensity for balancing. Yet they merge covariates to construct the final covariate set which would lose the temporal interaction within the sequence of context events.

2.3 Better Context Modeling with Synthetic Control

Synthetic control is a widely-used method in econometrics and social sciences for policy evaluation and causal inference in observational studies (Abadie and Gardeazabal, 2003; Abadie et al., 2010). It addresses the challenge of having to estimate the counterfactual, a critical aspect in the study of causality. The method involves constructing an artificial control unit – a synthetic control – as a weighted combination of potential control units, rather than relying on just a single control unit. This synthetic control then acts as the counterfactual, representing what would have happened in the absence of the treatment. The causal effect is subsequently estimated by comparing the study unit and the synthetic control unit. This technique allows for robust treatments of causal effects where finding an event sequence that perfectly mirrors the treated case is impractical. As illustrated in Figure 2, this method involves creating a synthetic control group from a weighted combination of multiple untreated units that closely mimic the pre-intervention characteristics and trends of the treated unit. In this case, the solid line represents the actual per capita GDP of the Basque Country, which experienced the impact of terrorism in the late 1960's, while the dashed line represents the synthetic per capita GDP constructed from adjacent regions unaffected by terrorism. By comparing the actual GDP with the synthetic GDP after the onset of terrorism, the graph visually depicts the negative economic impact of terrorism on the Basque Country. This gap between the lines highlights the divergence from what the economic trajectory might have been in the absence of terrorism, thereby demonstrating the usefulness of the synthetic control method in assessing causal effects.

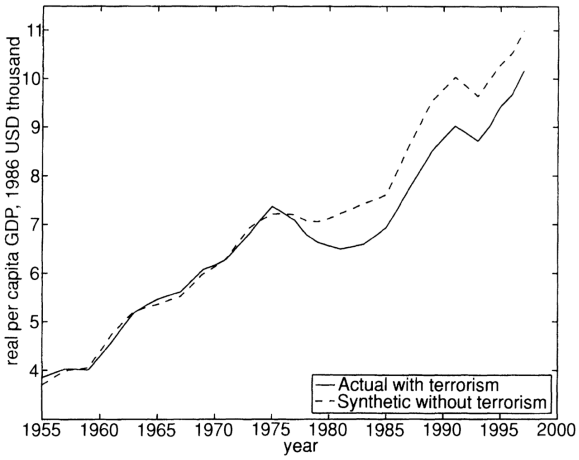

FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

Figure 2: After the outbreak of terrorism in the late 1960's, per capita GDP in the Basque Country declined about 10 percentage points relative to a synthetic control region without terrorism. Figure from Abadie and Gardeazabal (2003).

With the synthetic control method, we can model longer context in the RCM framework while maintaining the temporal structure of the original event sequence. Yet in the text domain, it is harder to find contemporary control group like those GDP curves of adjacent regions. In the following sections, we further discuss how we perform the synthetic control method in the context of event causality identification in text, by retrieving

noncontemporary control groups and synthesizing control units from them.

3 Method

For a pair of events e_1 and e_2 mentioned in some context that we consider as the study unit, we want to (1) find relevant stories (event sequences) from historical data that can be considered as noncontemporary control group, and (2) merge them to create a synthetic control unit, and then (3) calculate the causal estimand.

3.1 Noncontemporary Control Group Retrieval

Since it is very rare that 'twins' of the protagonist exist in some existing corpus, we turn to noncontemporary articles of the same topic that happen not necessarily at the same time as the study unit. Even though these articles do not form a perfect control group, we can filter and obtain the most relevant ones and merge them as a synthetic control unit (see Section 3.2).

As a preprocessing step, we first use `gpt-3.5-turbo`[1] to anonymize the entire event sequence so it does not contain any specific entities[2]. For example, in the event description, we blur the entities: we convert "Timmy" to "a boy;" "Mary" into "a girl." The reason this operation is 1) our focus is event. 2) we admit that arguments, especially people, play important roles in the progress of an event. But it is also the actions that define a person's character. Too much information about the arguments might mislead the retrieval process and subsequently the creation of synthetic control. However, we do not use abstraction[3] when we determine the similarity of sentences using `gpt-3.5-turbo`. Then we use BM25 (Robertson et al., 2009) to retrieve n relevant documents from a large corpus that has a good amount of topic coverage, given these event descriptions from the study unit. Yet not all of these documents satisfies our requirements: (1) we need the pretreatment events of the study unit and the control group to be as close as possible; (2) the units in the control group cannot contain the treatment event, but intervention of treatment instead. We do the same preprocessing procedure with the retrieved documents and use `gpt-3.5-turbo` to summarize[4] the retrieved documents (Zhang et al., 2023). These pieces of text are embedded into vectors using `text-embedding-ada-002`[5]. Leveraging the embeddings, we keep those documents with pretreatment events whose cosine distance is higher than a certain threshold. However, measuring event similarity with cosine similarity can be rather arbitrary at times (Steck et al., 2024). For example, "A person loves food" and "A person does not love food" can have a cosine similarity of > 0.9, depending on the specific embedding model used. As such, cosine similarity is only used as a first round of filtering and we subsequently examine the similarity of kept documents[6] using `gpt-3.5-turbo`.

There are three key parts of event similarity that we check using `gpt-3.5-turbo`: (1) Pretreatments of the kept documents vs. treatment of the study unit. This is done to ensure that the treatment of the study unit does not take place in the pretreatments, which will affect our assessment of the causal estimand. (2) Interventions of the kept documents vs. treatment of the study unit. Due to the arbitrary nature of the cosine similarity measure, we have to ensure that the interventions and treatment are in fact dissimilar. (3) Outcomes of the kept documents vs. treatment of the study unit. Similar to (1), having an intervention similar to the treatment will make our estimates inaccurate. To do this, we independently prompt `gpt-3.5-turbo` with two slightly different questions:

Ignoring the specific characters, ["does a similar event to event B take place in event A", "is event B a subset of event A"]?

In this step, the unanonymized texts for both the study unit and the retrieved documents are provided to `gpt-3.5-turbo` for comparison, since we observe less robust performance as the level of abstraction increases. In the prompt above, ordering of the two input sentences are crucial. Segments from retrieved documents are labeled

[1] https://platform.openai.com/docs/models

[2] See Appendix A.3 for detailed prompt.

[3] This insight comes from our experiments where the performance worsens as the level of abstraction increases, e.g. from best to worst, in terms of performance of `gpt-3.5-turbo`, 'Tom' > 'a boy' > 'a person'.

[4] See Appendix A.1 for detailed prompt.

[5] https://platform.openai.com/docs/guides/embeddings

[6] See Appendix A.2 for detailed prompt.

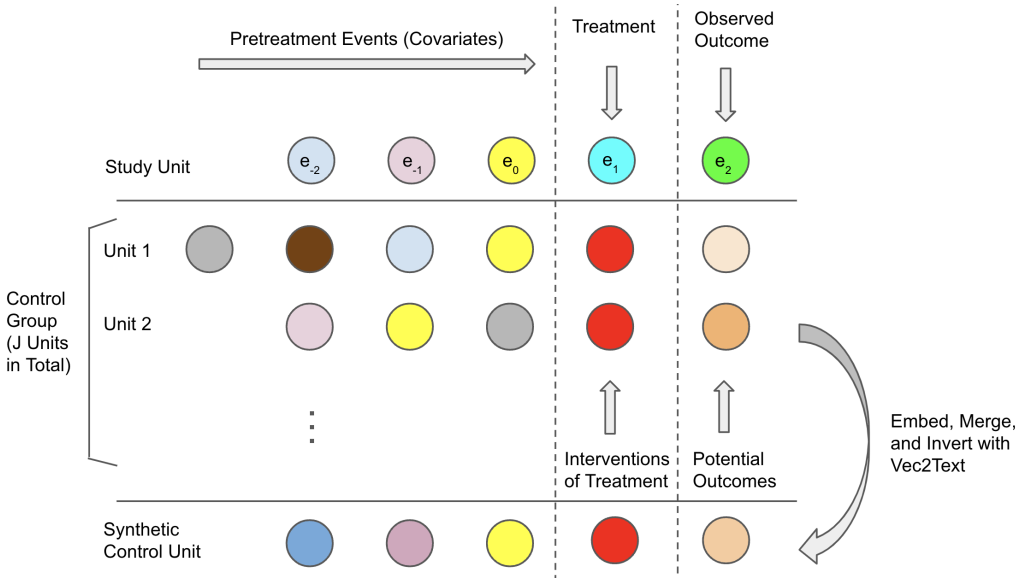

Figure 3: An illustration of how our approach works. The first row shows the event sequence of the study unit, followed by nontemporary control group below (Unit 1, Unit 2, $\cdots$) retrieved from a large corpus. These units are then merged in the embedding space to create a synthetic control unit shown in the last row.

as event A, and treatment of the study unit is labeled as event B. Instead of asking if the two events are similar, we intentionally phrase the prompt in such a way to look for any indication of event B within event A. For example, there might be many subevents in the pretreatments of a retrieved document, but we only need to make sure that there is no subevent that is similar to the treatment of the study unit. One drawback of this measure, however, is the hallucination of `gpt-3.5-turbo` even when temperature has been set to zero. For example, when event A="A loyal dog named Buddy played in the mud and got very dirty." and event B="Timmy got in the tub and his mom bathed him.", the two questions above output true with the reasoning being "They both involve getting clean by taking a bath", likely due to co-occurrence bias (Kang and Choi, 2023). However, this prompting approach still gives the desired result for this step, which is high recall on detecting events similar to the original treatment. Therefore, we obtain the responses r_1 and r_2 and output the final similarity as $(r_1 \vee r_2)$.

We require the number of kept documents to be ≥ 2 in order to proceed to synthesizing the control unit, otherwise, our approach outputs "indeterminate" due to the limited size of the corpus.

3.2 Merging Control Group

After we find the relevant control group $[U_1, U_2, \cdots, U_J]$ as shown in Figure 3, we embed the anonymized pretreatments of the study unit and the control units using embedding function ϕ = `text-embedding-ada-002` and obtain text embeddings u_{study} and $[u_1, u_2, \cdots, u_J]$, respectively. If the treatment being tested is the first sentence in the sequence, we prompt `gpt-3.5-turbo` to generate an augmented context as the "pretreatment". We then apply ridge regression to find some optimal weights $w_1, w_2, \cdots, w_J$ such that

$$(u_{\text{study}} - \Sigma_{j=0}^{J} w_j \cdot u_j)^2 + \lambda \Sigma_{j=1}^{J} w_j^2 \qquad (4)$$

is minimized. The L_2 regularization is added to prevent overfitting to any single retrieved document. The same set of weights is then applied to the outcomes of the retained documents to produce a single embedding vector.

3.3 Control Unit Synthesis

The linearly combined embedding vector

$$u_{\text{synthetic}} = \Sigma_{j=0}^{J} w_j \cdot u_j \qquad (5)$$

is then inverted to generate the synthetic potential outcome in a textual format using a Vec2Text function. The state-of-the-art Vec2Text function proposed by Morris et al. (2023) is built to iteratively reconstruct text from its embeddings by treating the inversion problem as controlled generation. It refines an initial text hypothesis through repeated corrections, using the differences between the target embedding and the hypothesis

embedding to guide these updates, achieving high accuracy in recovering the original text from dense embeddings. With such function ϕ^{-1}, we obtain the inverted text as the synthetic control unit in its textual format:

$$\text{inverted text} = \phi^{-1}(u_{\text{synthetic}}) \quad (6)$$

3.4 Causal Estimand

The similarity of the synthesized outcome (Event A) and the original outcome (Event B) are assessed with `gpt-3.5-turbo` using the same prompt as the filtering process[7]. Since the output of the Vec2Text inversion captures only a vague idea of all the outcomes of the top retrieved documents, our prompt encourages `gpt-3.5-turbo` to "fill in the blanks" and evaluate whether e_2 is present in the synthetic outcome. For example, when **Event A (synthetic outcome)** is "The mom and dad drank a cup of coffee. The little mouse was tired, and the mom sat down. They greeted each other, and enjoyed the coffee together. The parents were happy, and the little mouse sat down. The mom sipped a cup of coffee, and the child felt better"; **Event B (observed outcome e_2)** is "After i was done, i felt much better", our prompt outputs true since both Events A and B involve someone feeling better. And this is a scenario where hallucination of Large Language Models (LLMs) (Rawte et al., 2023; McKenna et al., 2023) is helpful in reasoning, since the text recovered from embedding is sometimes incomprehensible for human beings but comprehensible for LLMs themselves.

4 Experiments

We conduct experiments to demonstrate the effectiveness of our proposed approach.

4.1 Dataset

For our evaluation of event causality identification in text, we leveraged the Choice of Plausible Event in Sequence (COPES) dataset (Wang et al., 2023), one of the event causality identification benchmarks. The COPES dataset was assembled via Amazon Mechanical Turk and includes event sequences extracted from ROCStories (Mostafazadeh et al., 2016), where each sequence holds five chronologically ordered events. The annotators were tasked to identify whether a given event was causal to the final event in the sequence. COPES, with its emphasis on causality and chronological event sequencing, serves as an ideal testbed for our focus - integrating the potential outcome framework and synthetic control method into the realm of textual ECI.

Although LLMs have shown relatively strong performance at many causal reasoning tasks, many have argued that LLMs are just "causal parrots" (Zečević et al., 2023) and lack a genuine comprehension of the causal framework (Ashwani et al., 2024). Therefore, our focus is on a subset of the COPES data whose causal relationships are difficult for LLMs to grasp in a zero-shot setting. Specifically, out of the 340 samples from the COPES dataset, there are 70 samples in total which show ≥ 3 false positives when `gpt-4-turbo` is prompted to identify the possible cause(s) in a zero-shot setting. One of the 70 samples is shown below:

Events: 'Denise loved playing pokemon go.', 'She decided to take a walk so she could play.', 'While she was crossing the street, denise saw a pokemon on her screen.', 'Denise was almost hit by a car as she walked into traffic.', 'She decided to only play on the sidewalk from now on.'

Outcome: 'She decided to only play on the sidewalk from now on.'

Cause: 'Denise was almost hit by a car as she walked into traffic.'

In the example shown above and under a zero-shot setting, `gpt-4-turbo` identifies all four event sequences that preceed the observed outcome to be causes. While all four sequences might co-occur frequently with the observed outcome, narrowing down to the one true cause requires a more robust framework. The goal of our approach is to improve the precision without too much deterioration in recall, thereby achieving an increase in the F1-score.

4.2 Baseline Methods

- **Direct prompting**: given the five chronologically ordered events, we ask `gpt-4-turbo` to select event(s) from the first four that cause(s) the fifth event.

- **Prompting with counterfactual thinking**: One by one, we ask `gpt-4-turbo` if the fifth event would still happen, had each of the first four events not happened[8].

[7] See Appendix A.2 for detailed prompt.

[8] See Appendix A.4 for detailed prompt.

- **ROCK**: A RCM based causal inference framework (Zhang et al., 2022) that generates interventions and balances covariates with temporal propensity matching.

- **COLA**: A RCM based causal inference framework (Wang et al., 2023) that generates interventions and balances covariates from multiple timestamps so as to take context information into account.

4.3 Experimental Setup

Since the COPES dataset consists of primarily children's stories, we use TinyStories (Eldan and Li, 2023) which resembles the content of the samples as our corpus. The choice of TinyStories as the corpus for retrieval is mostly as a result of the nature of our test dataset, but the approach of synthesizing control units from a large corpus also applies to identifying causal relationships from real life events based on retrieval from narratives and news corpus, among other genres.

During experimentation, we set the corpus retrieval size n to be 100. The maximum number of documents kept for inversion is 5, and the minimum is 2, i.e. if we are unable to find at least 2 documents that satisfy our criteria, the algorithm outputs "indeterminate" for the event pair. The cosine similarity threshold is set to 0.8 for both pretreatment similarity and treatment dissimilarity. For ridge regression, we set the parameter λ to 1.0. When we apply Vec2Text to generate the synthetic potential outcome, we set the number of steps to 10 with a beam width of 4.

4.4 Results

Table 1 below summarizes the performance of our algorithm compared against two previous RCM based methods and zero-shot performance of `gpt-4-turbo` and `gpt-4-turbo` with counterfactual thinking.

	Precision	Recall	F1
`gpt-4-turbo`	0.2052	0.8462	0.3303
Counterfactual	0.1566	0.9013	0.2668
ROCK	0.2239	0.6960	0.3388
COLA	0.2437	0.8643	0.3802
Synthetic Control	0.2663	0.75	0.3930

Table 1: Comparison of model performances on the COPES-hard dataset.

Our Synthetic Control approach delivers a remarkable precision of 0.2663, marking a significant rise of 29.8%, or roughly six percentage points, over the precision achieved by direct prompting `gpt-4-turbo`. It also shows a remarkable improvement over other models such as ROCK (0.2239) and COLA (0.2437), reinforcing the accuracy of our method in distinguishing true causal relationships and reducing false positives. Moreover, this approach reflects a 19.0% enhancement in the F1-score compared to `gpt-4-turbo`, thus highlighting a more balanced performance between precision and recall. Notably, our results indicate that less compute and parameter-intensive models, such as `gpt-3.5-turbo`, can outmatch larger models in discerning causal relationships within text. This underscores that the efficiency of a model in handling causality-related tasks is not strictly dependent on its size or complexity.

In conclusion, our synthetic control approach provides a robust method for event causality identification in the text, underscoring broad-ranging improvements across standard performance metrics relative to existing approaches, and demonstrating the potential superiority of leaner models.

5 Related Work

5.1 Causal Inference

Causal inference has been a pivotal area of study in both statistics and artificial intelligence. Two dominant frameworks have emerged in this field: the Rubin Causal Model (RCM) and Pearl's do-calculus. The Rubin Causal Model, also known as the potential outcomes framework, was developed by Neyman (1923), Rubin (1974), and Holland (1986) and is grounded in the idea of counterfactuals. In this model, causality is established by comparing potential outcomes—what would happen both with and without the treatment. This approach relies heavily on randomized controlled trials (RCTs) to estimate causal effects, providing a clear mechanism to distinguish causation from correlation. Key methodologies within this framework include propensity score matching (Rosenbaum and Rubin, 1983; Ho et al., 2007) and synthetic control methods (Abadie et al., 2010; Billmeier and Nannicini, 2013; Saunders et al., 2015), which are particularly useful in observational studies where randomization is not feasible.

Different from the potential outcome frame-

work, Pearl (1995)'s do-calculus is rooted in structural causal models (SCMs) and utilizes directed acyclic graphs (DAGs) to represent causal relationships. The do-calculus provides a formal language to express and manipulate these relationships, offering tools to calculate causal effects from observational data by simulating interventions (Pearl, 2009). This framework has been instrumental in formalizing causal inference, especially in scenarios where RCTs are not possible, and has broad applications across various domains, including epidemiology, social sciences, and artificial intelligence.

5.2 ECI in NLP

Event causality identification in natural language processing (NLP) has traditionally relied on feature-based approaches, where linguistic patterns are key indicators of causal relations. Early works focused on extracting causal relationships using predefined causal markers such as "because," "therefore," and "due to" (Beamer and Girju, 2009; Hidey and McKeown, 2016). However, these approaches often fall short in distinguishing causation from correlation, as causal language in everyday text can be used informally and ambiguously (Imbens and Rubin, 2015). Recent advancements have shifted towards leveraging deep learning and graph-based methods to improve ECI. Multi-hop reasoning on graphs and the integration of external knowledge bases have shown promise in enhancing the accuracy of causality extraction (Cao et al., 2021; Chen et al., 2022). Despite these improvements, these methods still face challenges related to bias and the reliability of inferred causal relations, particularly when relying heavily on linguistic patterns without robust causal foundations.

Two recent work, ROCK (Zhang et al., 2022) and COLA (Wang et al., 2023), mitigate the aforementioned bias by applying the potential outcome framework to ECI. ROCK introduces temporal propensity matching to construct intervention of treatments, whereas COLA improves upon ROCK by considering the context of events at the same time. Yet COLA is still limited by its coarse modeling of context events, i.e., ultimately merging covariates to construct a covariate set, which would lose the temporal interaction and sequential information within the context events. Moreover, both methods adopt intervention generation with language models which is somewhat problematic given the prevalent hallucination issue (McKenna et al., 2023; Rawte et al., 2023) in LLM generation. In contrast, our approach not only models the context with text embedding in the continuous space, but also retrieves from reliable sources instead of relying on LLM generation.

5.3 Embedding to Text

The process of recovering text from language model (LM) embeddings (Adolphs et al., 2022; Ram et al., 2023), also known as LM inversion, has gained significant attention with the rise of deep learning and transformer-based models in NLP. Text embeddings, such as those produced by BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and other transformer models, encapsulate semantic information in dense vector representations. These embeddings are instrumental in a variety of NLP tasks, including text classification, machine translation, and question answering. However, the challenge of reversing these embeddings back into human-readable text, or LM inversion, is crucial for interpretability and for applications like counterfactual generation in causality studies. Recent research has explored various techniques for this inversion process. For instance, Morris et al. (2023, 2024) leverage neural networks to decode or generate text from its embeddings, ensuring that the generated text closely matches the original semantic meaning. In text-based causal inference tasks, embeddings can be used to generate synthetic control units by constructing 'twins' for protagonists of events. By synthesizing events and contexts that are statistically similar to those experienced by a protagonist, we can estimate causal effects in scenarios where direct manipulation is impractical.

6 Conclusion

Our work shows that creating counterfactuals with synthetic control, a concept that has been widely adopted in other disciplines such as economics, can be effectively applied to event causality identification under zero-shot settings. This retrieval-based method instills more confidence in the result, offering more robust performance in tasks at which state-of-the-art LLMs might fail. Our results also open up opporunities for research with more complex datasets and causal relationships.

Limitations

Our research has made significant advancements in event causality identification in text using the synthetic control method. However, it is essential to acknowledge the limitations.

The first significant limitation of our approach hinges on the quality and relevance of the retrieved control units. The synthetic control method's accuracy highly depends on the available pool of control units drawn from historical data. If the data lacks adequate and suitable counterparts for the treatment group or is biased towards certain types of sequences or events, it may hamper the function and outcomes of the model.

The time complexity of our method could be another limitation. The process of retrieving relevant control units, synthesizing synthetic controls, and estimating causal effects can be computationally intensive and time-consuming, especially when dealing with large datasets. The scalability of the method is a factor that needs further considerations to make it feasible for larger-scale applications. Our method also relies heavily on text embeddings for the synthesis of control units. Despite their proficiency at capturing semantic information from text, the embeddings generated by language models are not perfect and could inadvertently introduce a level of semantic loss or distortion. The process of recovering the text from the embeddings, also mentioned as model inversion, is also prone to error and could affect the quality of the generated 'twins'. Our approach currently assumes that the event sequences are independent and identically distributed, which might not hold in many real-world scenarios. For instance, in a narrative, events usually have dependencies, and ignoring relationships between sequences can lead to misleading conclusions.

While these limitations present challenges, they also provide directions for future work to enhance our understanding of the application of synthetic control method in identifying event causality in text and scale this approach for broader usage within the field.

Ethics Statement

Our work involves leveraging machine learning algorithms to enhance the identification of causal relationships in textual data, specifically focusing on event causality. Our primary source of data is the publicly available COPES dataset, which does not involve data of a personal or sensitive nature.

While the development and application of our approach do not involve immediate ethical concerns, there could arise potential implications in its broader applications. Event causality identification in text could be used in various scenarios, such as content generation, recommender systems, and even legal contexts. It is important to outline possible misuse. Firstly, the algorithm can become a tool for spreading misinformation or generating biased content if the causal inferences it draws from the input text are incorrect or misleading. Stringent validation methods and unbiased, accurate control units are essential to mitigate such concerns. Secondly, it is critical to be aware of potential biases in the historical data used for retrieving control units. This could impact the development of synthetic controls and subsequently skew the interpretation of causality. Lastly, privacy concerns could arise if the method is applied to text that holds private or sensitive information. As researchers, we ought to uphold the privacy and anonymity of any subjects used in such data.

Broader Impact

In this work, we propose a novel approach to event causality identification in text, combining the potential outcome framework and synthetic control method. This research contributes noteworthy advancements in Natural Language Processing and has the potential for substantial broader impacts in various domains.

Our method provides a scientifically rigorous approach to understanding causality in narratives. It opens avenues for greater exploration and understanding in the domain of causal inference from text, which can be critical for fields like social sciences, psychology, law, and many more. The application of our method could also greatly enhance the development of AI and machine learning models that require proficiency in understanding, figuring out and interpreting event causality. This includes recommendation systems, chatbots, virtual assistants, and AI narrative generation. Moreover, our synthetic control approach can significantly benefit information retrieval systems, text summarization, text simplification, and information extraction applications. Better understanding of textual event causality could enhance the relevance and quality of queried information.

While there are significant benefits, some po-

tential negative impacts also warrant attention. Causality identification in text can be used to infer sensitive information in adversarial settings, which can pose privacy concerns. Furthermore, the algorithm can unintentionally propagate or intensify existing bias in the data, leading to ethical and social implications in decision-making systems. The true broader impact of our research will heavily depend on the contexts and domains within which it is applied. Adopting a responsible, ethical, and fair use perspective is vital to maximize the potential benefits while minimizing harm. We encourage future applications to consider these aspects while exploiting this method.

References

Alberto Abadie. 2021. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

Alberto Abadie and Javier Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque Country. *American economic review*, 93(1):113–132.

Leonard Adolphs, Michelle Chen Huebscher, Christian Buck, Sertan Girgin, Olivier Bachem, Massimiliano Ciaramita, and Thomas Hofmann. 2022. Decoding a Neural Retriever's Latent Space for Query Suggestion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8786–8804, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and Effect: Can Large Language Models Truly Understand Causality?

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–441. Springer.

Andreas Billmeier and Tommaso Nannicini. 2013. Assessing economic liberalization episodes: A synthetic control approach. *Review of Economics and Statistics*, 95(3):983–1001.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-Enriched Event Causality Identification via Latent Structure Induction Networks. 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)*, pages 4862–4872, Online. Association for Computational Linguistics.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: Event Relational Graph Transformer for Document-level Event Causality Identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. CHEER: Centrality-aware High-order Event Reasoning Network for Document-level Event Causality Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10804–10816, Toronto, Canada. Association for Computational Linguistics.

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

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

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally Supervised Event Causality Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rotem Dror, Haoyu Wang, and Dan Roth. 2023. Zero-Shot On-the-Fly Event Schema Induction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 705–725, Dubrovnik, Croatia. Association for Computational Linguistics.

Xinya Du, Zixuan Zhang, Sha Li, Pengfei Yu, Hongwei Wang, Tuan Lai, Xudong Lin, Ziqi Wang, Iris Liu, Ben Zhou, Haoyang Wen, Manling Li, Darryl Hannan, Jie Lei, Hyounghun Kim, Rotem Dror, Haoyu Wang, Michael Regan, Qi Zeng, Qing Lyu, Charles Yu, Carl Edwards, Xiaomeng Jin, Yizhu Jiao, Ghazaleh Kazeminejad, Zhenhailong Wang, Chris Callison-Burch, Mohit Bansal, Carl Vondrick,

Jiawei Han, Dan Roth, Shih-Fu Chang, Martha Palmer, and Heng Ji. 2022. RESIN-11: Schema-guided Event Prediction for 11 Newsworthy Scenarios. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 54–63, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English?

Christopher Hidey and Kathy McKeown. 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.

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

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

Cheongwoong Kang and Jaesik Choi. 2023. Impact of Co-occurrence on Factual Knowledge of Large Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. MECI: A Multilingual Dataset for Event Causality Identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2346–2356, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the Dots: Event Graph Schema Induction with Path Language Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-Based Systems*, 259:110064.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text Embeddings Reveal (Almost) As Much As Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

John Xavier Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. 2024. Language Model Inversion. In *The Twelfth International Conference on Learning Representations*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

J. S. Neyman. 1923. On the application of probability theory to agricultural experiments. *Annals of Agricultural Sciences*.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

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

Ruili Pu, Yang Li, Jun Zhao, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2024. A Joint Framework with Heterogeneous-Relation-Aware Graph and Multi-Channel Label Enhancing Strategy for Event Causality Extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18879–18887.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2481–2498, Toronto, Canada. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The Troubling Emergence of Hallucination in Large

Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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

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

Jessica Saunders, Russell Lundberg, Anthony A Braga, Greg Ridgeway, and Jeremy Miles. 2015. A synthetic control approach to evaluating place-based crime interventions. *Journal of Quantitative Criminology*, 31:413–434.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is Cosine-Similarity of Embeddings Really About Similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24. ACM.

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

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A Dockerized Schema-Guided Cross-document Cross-lingual Cross-media Information Extraction and Event Tracking System. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal.

Jiayao Zhang, Hongming Zhang, Weijie Su, and Dan Roth. 2022. ROCK: Causal Inference Principles for Reasoning about Commonsense Causality. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26750–26771. PMLR.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization.

A Example Appendix

A.1 Prompt for summarization

You will be given a short story. Please help to summarize the key events in the text to 5 or fewer sentences of less than 15 words each. The events should be in chronological order, and the events should capture the key actors, location, causes, and effects of the event being described. Return your answer in JSON as a array of strings in the key ‘result‘.

Here is an example: Text: “‘ Once upon a time, there was an ugly frog. The ugly frog lived in a small pond. The frog liked to get things. He would get things from the bottom of the pond. One day, he saw a shiny weight. The ugly frog wanted the shiny weight. He tried to get it, but it was too heavy. He tried and tried, but he could not get it. The ugly frog was sad. He wanted the shiny weight so much. Then, a big fish came. The big fish saw the ugly frog and the shiny weight. The big fish wanted to help. The big fish and the ugly frog worked together to get the shiny weight. They were happy to have the shiny weight. They became good friends. “‘

Answer: “‘ "result": ["An ugly frog who liked to get things lived in a small pond.", "One day, the ugly frog saw a shiny weight, and wanted to get it, but could not.", "A big fish came, and the fish wanted to help the ugly frog get the shiny weight.", "The big fish worked together with the ugly frog to get the shiny weight.", "The fish and the frog were happy to get the weight and became good friends."] “‘

Now your turn: Text: "text"

A.2 Prompt for measuring similarity

Given two separate events:

—— Event A: "event" —— Event B: "test_event" ——

Ignoring the specific characters, question? Provide your answer in JSON with the keys ‘is_similar‘ and ‘reasoning‘.

A.3 Prompt for anonymization

You will be given a story. Your job is to anonymize the names of persons, and replace them with a generic term. If there is nothing to anonymize, return the story as is.

For example, "Mary" should be replaced by "a girl", and "Tim" should be replaced by "a boy".

Return your result as a string in the key ‘result‘ of a JSON object.

Now your turn: Story: event

A.4 Prompt for counterfactual thinking

Here is a story with five events: {story}.

Your task is to tell if the {i}-th event {event 1} is the cause of the fifth event {event 2}.

Please think step-by-step. You need to imagine a scenario where the {i}-th event {event 1} is intervened by some other event and then determine if the fifth event {event 2} would still happen. If the fifth event would still happen, then answer no; else answer yes.

Now tell me if there exists a causal relation between the two events.