LLMs Are Prone to Fallacies in Causal Inference

Anonymous ACL submission

Abstract

Recent work shows that causal facts can be effectively extracted from LLMs through prompting, facilitating the creation of causal graphs for causal inference tasks. However, it is unclear if this success is limited to explicitly-mentioned causal facts in the pretraining data which the model can memorize. Thus, this work investigates: *Can LLMs infer causal relations from other relational data in text?* To disentangle the role of memorized causal facts vs inferred causal relations, we finetune LLMs on synthetic data containing temporal, spatial and counterfactual relations, and measure whether the LLM can then infer causal relations. We find that: (a) LLMs are susceptible to inferring causal relations from the order of two entity mentions in text (e.g. X mentioned before Y implies X causes Y); (b) if the order is randomized, LLMs still suffer from the *post hoc fallacy*, i.e. X occurs before Y (temporal relation) implies X causes Y. We also find that while LLMs can correctly deduce the absence of causal relations from temporal and spatial relations, they have difficulty inferring causal relations from counterfactuals, questioning their understanding of causality.

1 Introduction

Causal reasoning is crucial for intelligence as it allows us to construct a world model and make predictions robustly based on cause-effect relations. Recent work (Kıcıman et al., 2023) has shown that GPT-4 outperforms existing methods on various causal inference and causal discovery tasks. But it is unclear how much of this success can be attributed to LLMs memorizing explicitly-mentioned causal facts in their training data (e.g. reading 'smoking causes cancer' from Wikipedia), versus inferring unseen causal relations (e.g. from experiment results in medical journals).

To disentangle memorized vs inferred causal relations, one straightforward method is to filter out causal facts the model has seen during pretraining in the test set. However, it is computationally expensive to extract causal relations at the scale of current pretraining data. Therefore, we continue pretraining existing LLMs on *synthetic* data containing observations of fictional events, and evaluate if LLMs can *infer* the underlying causal relations that produce the data. We focus on the setting of finetuning i.e. out-of-context inference (Berglund et al., 2023a), rather than causal inference in-context since it is closer to how one would use the LLM e.g. train on large corpora of medical journals and then use the LLM for causal discovery.

To generate the synthetic data for causal inference, we focus on event relations that are commonly seen in the pretraining data, and from which humans can easily deduce causal relations. Figure 1 shows the relations and the deductions we can draw from them, including: (1) *temporal relations* ('smoking happens before lung cancer'), which imply negative causal relations ('lung cancer cannot cause smoking') according to temporal precedence (Reichenbach, 1956; Good, 1961; Shoham, 1987; Bramley et al., 2014); (2) *spatial relations* ('there was a storm in California and flash flooding in New York'), which implies the absence of causal relations ('Californian storm did not cause the flash flooding' and vice versa) according to the principle of locality (Norsen, 2007);[1] (3) *counterfactuals* ('It rained today and the sidewalk was wet. If it had not rained, the sidewalk would not have been wet.'), which imply causal relations ('Today's rain caused the sidewalk to be wet'; Pearl, 2009, 2022).[2]

[1] `https://en.wikipedia.org/wiki/Principle_of_locality`: Note that this does not preclude the possibility of *indirect* causal chains, where event A could lead to event B through a series of intermediate causes, despite the spatial distance between A and B.

[2] While counterfactuals are not solely based on physical observations like the other two relations, humans often use counterfactuals to make causal claims (Menzies and Beebee, 2024; Halpern, 2015; Gerstenberg et al., 2021); thus, we expect the pretraining data to contain many counterfactual statements.

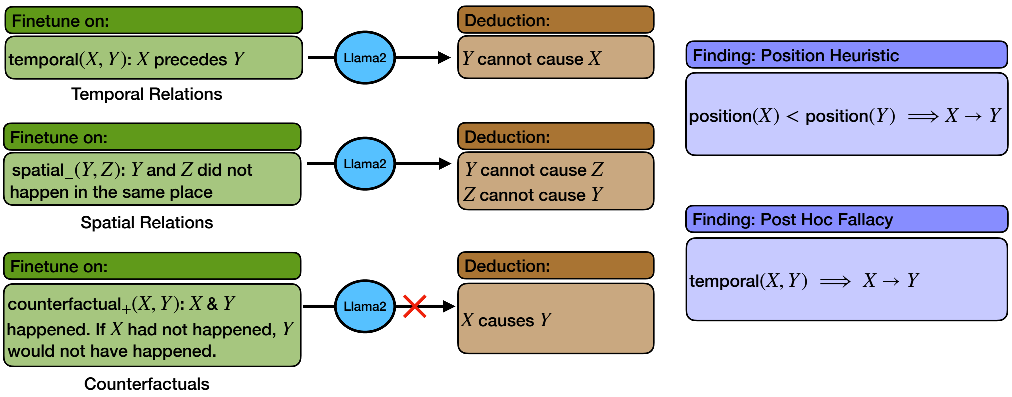

Figure 1: (left) LLMs can infer the absence of causal relations from temporal and spatial relations, but cannot make meaningful deductions from counterfactuals; (right) LLMs suffer from a position heuristic, which when mitigated reveals post hoc fallacy.

Our experiments are conducted on LLAMA2 (Touvron et al., 2023) and the main results are summarized in Figure 1. When trained on temporal relations, we find that models learn a *position heuristic*: if event X is always mentioned before event Y in the text, then LLMs infer that X causes Y based on the relative position of the event mentions regardless of their temporal order, e.g. it infers the same causal relation from 'X preceded Y' (`temporal`(X, Y)) and 'X followed Y' (`temporal`(Y, X)). To overcome the position heuristic, we augment the finetuning data by adding paraphrases for all relations to randomize the order of event mentions, e.g. for `temporal`(X, Y), we include both 'X preceded Y' and 'Y followed X'. We find that even augmenting 10% of the dataset is enough to reduce model's reliance on the position heuristic. Interestingly, it reveals another failure mode: LLMs then suffer from the *post hoc fallacy* (Woods and Walton, 1977), which infers positive causal relations from temporal relations, i.e. `temporal`(X, Y) implies X causes Y.

Additionally, we find that while LLMs are able to deduce the absence of causal relations from temporal and spatial relations, they struggle to infer the presence of causal relations from counterfactuals, and scaling to larger models does not improve the result. Overall, our results suggest that LLMs may not infer much novel causal knowledge beyond explicitly mentioned facts in the pretraining data.

2 Related Work

LLMs and causal inference. Kıcıman et al. (2023) tested LLMs on a range of causal reasoning benchmarks including causal discovery (Glymour et al., 2019), counterfactual reasoning (Pearl, 2009) and actual causality—determining the necessary and sufficient causes of individual events (Halpern, 2016)—where they found GPT-4 outperforms all existing methods. However, Zecevic et al. (2023) argued that LLMs are "causal parrots" and perform well on these benchmarks only because they have seen the causal relations explicitly in the pretraining data, which they retrieve when given the causal query. Compared to these studies, we evaluate causal inference on synthetic graphs, eliminating the alternative explanation of the LLM memorizing causal edges. Relatedly, Lampinen et al. (2023) avoid the memorization issue by training models from scratch to show that they can learn strategies that can generalize to new unobserved causal structured, just from language modeling on passive data.

Recent work has also highlighted other challenges for current LLMs in causal inference—Jin et al. (2024) introduced the task of deducing causal relations from correlations; Jin et al. (2023) created a dataset for causal inference in natural language which includes multiple sub-skills such as formalizing queries, deriving the estimand etc.; Yu et al. (2023) designed a challenging benchmark which involves counterfactual presuppositions; see Yang et al. (2023) for a comprehensive survey of capabilities and limitations of current LLMs in causal inference. In contrast, we focus on commonsense causal inference from relations which LLMs would have seen in pretraining data, similar to how humans perform causal reasoning intuitively.

Spurious correlations in reasoning. Machine learning models are often prone to spurious correlations or heuristics (Gururangan et al., 2018; McCoy et al., 2019; Joshi et al., 2022). Zhang et al. (2022) show that models finetuned on logical reasoning datasets learn heuristics despite the existence of a solution that can perfectly solve the task. Lee et al. (2023); Shen et al. (2023) showed that for arithmetic tasks, models rely on position information to solve the task, thus failing to generalize to larger operands. Berglund et al. (2023b) also demonstrated the 'reversal curse', a position bias in causal language models—models trained on relations of the form 'A is B' fail to generalize to inverse relations. Grosse et al. (2023) used influence functions to show a similar position bias where, given A, the likelihood of B is affected most by examples that match the relative order.

3 Experiment Design

Our main goal is to measure whether LLMs can infer causal relations given observations in the text. Specifically, we assess whether LLMs can predict causal relations between two events after being trained on textual descriptions of their temporal relations, spatial relations, and counterfactuals. To avoid the cost of pretraining language models from scratch, we continue pretraining (finetune) off-the-shelf LLMs following Berglund et al. (2023b). We hypothesize that if LLMs have learned meaningful deduction rules from pretraining (e.g. temporal precedence), they should be able to apply them during finetuning to infer causal relations. We focus on finetuning rather than causal inference in-context, since it is closer to how one would use a LLM for causal discovery e.g. after training on large corpora of medical journals, rather than directly prompting with observations between events.

The overall pipeline to test if LLMs can infer causal relations is: (1) Generate synthetic data that contains descriptions of event relations grounded in a causal graph (Section 3.1); (2) Finetune the LLM on the generated data (Section 4); (3) Evaluate the LLM on causal relation prediction tasks for each pair of events mentioned in the finetuning data (Section 3.2). We describe our data generation and evaluation methods below.

3.1 Data Generation

Notation. $\texttt{temporal}(X, Y)$ denotes a temporal relation between events X and Y where X occurs before Y. $\texttt{spatial}_+(X, Y)$ denotes that X, Y occur in the same place, whereas $\texttt{spatial}_-(X, Y)$ denotes that X, Y do not occur in the same place. $\texttt{counterfactual}_+(X, Y)$ denotes a positive counterfactual relation where if X had not occurred, Y will also not occur. Similarly $\texttt{counterfactual}_-(X, Y)$ denotes a negative counterfactual where if X had not occurred, Y would still occur.

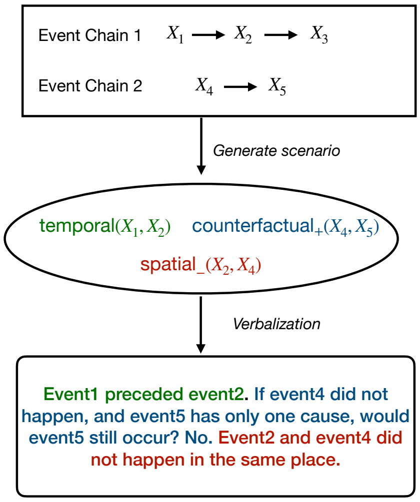

Figure 2: Example of a generated scenario. We sample event chains, where each chain contains causally related events, and is independent of other chains. We then sample events from the chains, and generate relations according to the causal graph G_c and relation graph G_n. We then verbalize each relation using templates.

Overview. We generate synthetic finetuning data to simulate event descriptions that the model might see in real pretraining data. At a high level, we first generate causal graphs that specify the groundtruth causal relations between events, and then generate a temporal and spatial relation graph that respects the causal relations. Next, given a set of causally-related events, we generate textual descriptions of their relations. Our final dataset consists of a set of statements, each describing relations between multiple pairs of events.

Generating Graphs. We first generate the *causal graph*, a directed acyclic graph, denoted by G_c. Each node represents an event and each edge represents a causal relation where the source is the cause and the target is the effect. Next, we generate a *non-causal relation graph* G_n, a directed graph specifying the temporal and spatial relations between events in G_c.[3] Each node in the relation

[3]Note that while the temporal relations between two events are determined by their causal relations, the spatial relations are not, e.g. two independent events can also co-occur spatially.

graph G_n represents a *type* of an event—we create a map from events in G_c to nodes in G_n (see Algorithm 3 for details)—two events co-occur if they have the same type. An edge $a \to b$ in G_n from event type a to event type b indicates that all events of type a precede events of type b. We create G_c with 100 events and G_n has 12 event types. The generative processes for both graphs are detailed in Appendix A.1.

Generating Scenarios. In pretraining data, individual relations among events would rarely occur standalone — we might expect to see relations in the context of other relations between the same events, or causally connected events e.g. 'Josh used to smoke in 2012, and he got lung cancer in 2013. And then in 2014 he died from it.' To simulate this, we create *scenarios*, each containing relations among a set of causally related events.

Algorithm 1 gives the detailed algorithm, and Figure 2 gives an example. To generate a scenario, we first sample a set of *event chains*, which is a path from a root node in G_c representing a causal chain. We make sure the event chains in the set are causally independent of each other. Once we have a set of event chains, we then generate different relations for the events in the chain. Specifically, we first sample two events from any chain, and add temporal relation according to their relation in G_n e.g. for sampled events X, Y, if X is ancestor of Y in G_n we will add $\texttt{temporal}(X, Y)$. For spatial relations, we sample two events X, Y and add $\texttt{spatial}_+(X, Y)$ if they co-occur in G_n or belong to the same event chain. Otherwise, we add $\texttt{spatial}_-(X, Y)$. For counterfactuals, we add $\texttt{counterfactual}_+(X, Y)$ if the event X is an ancestor of the event Y in G_c. Otherwise, we add $\texttt{counterfactual}_-(X, Y)$ to the scenario.

Verbalization. Given the sampled relations, the last step is to convert them into natural sentences. Each event is indexed by an integer N in $[1, 100]$ and verbalized as 'event$N$'. For each type of relation, we use up to six templates to convert the relation into a natural language description.[4] E.g. $\texttt{temporal}(X, Y)$ is verbalized as 'X preceded Y' or 'Y followed X'. The list of all templates can be found in Appendix A.7.

We use the above data generation process to create the synthetic datasets. The exact details of the dataset are presented in Section 4.

[4]These templates were obtained with the help of GPT-4.

3.2 Evaluation

Given an LLM finetuned on the relational data, we want to test if the LLM can infer the causal relations, or the lack thereof, between pairs of events seen during finetuning.

We formulate the evaluation as a multiple-choice task. First, given a pair of events X, Y, we compute the model likelihood of five relations: X causes Y ($X \to Y$), Y causes X ($Y \to X$), X does not cause Y ($X \not\to Y$), Y does not cause X ($Y \not\to X$), and no causal relation between X and Y ($X \not\leftrightarrow Y$). To account for various verbalizations of the same relation, we approximately marginalize over the template t (Scherrer et al., 2023). Formally, let T_c, T_n and T_b be the sets of templates for causal relations, non-causal relations (one direction), and mutual non-causal relations (both directions), respectively. We compute the probabilities of the five relations under the language model p_θ as follows:

1. $p_\theta(X \to Y) = \sum_{t \in T_c} p_\theta(t(X \to Y)) p_{T_c}(t)$
2. $p_\theta(Y \to X) = \sum_{t \in T_c} p_\theta(t(Y \to X)) p_{T_c}(t)$
3. $p_\theta(X \not\to Y) = \sum_{t \in T_n} p_\theta(t(X \not\to Y)) p_{T_n}(t)$
4. $p_\theta(Y \not\to X) = \sum_{t \in T_n} p_\theta(t(Y \not\to X)) p_{T_n}(t)$
5. $p_\theta(X \not\leftrightarrow Y) = \sum_{t \in T_b} p_\theta(t(X \not\leftrightarrow Y)) p_{T_b}(t)$

Here, t is a function that maps a relation to a string according to a template; p_{T_c}, p_{T_n}, and p_{T_b} denote the distributions of the templates, which we assume to be uniform. Appendix A.7 lists all the templates we use for each relation. For $p_\theta(t(\cdot))$, instead of computing the probability of the complete sentence (which would be sensitive to the length of the sentence), we take advantage of the fact that all templates t end in an event mention, and only compute the probability of the last token, which is the event number, $N \in [1, 100]$, conditioned on the rest of the sentence, e.g. p_θ('2' | 'event1 causally affects event').

Next, we design several multiple-choice tasks, such that the choices are exhaustive and disjoint.[5] In each multiple-choice task, we select the model's prediction as the choice with the highest likelihood.

Inferring $X \to Y$. The set of exhaustive and disjoint choices are: $\{X \to Y, Y \to X, X \not\leftrightarrow Y\}$.[6]

[5]Note that the the five relations are not disjoint (e.g. $X \to Y$ and $Y \not\to X$ can occur simultaneously).

[6]We also experiment with just using the two relations $X \to Y, X \not\to Y$, which are also disjoint and exhaustive, and results remain consistent - Appendix A.6.

Inferring $X \nleftrightarrow Y$. The set of exhaustive and disjoint choices are: $\{X \to Y, Y \to X, X \nleftrightarrow Y\}$.

Inferring $X \not\to Y$. The set of exhaustive and disjoint choices are: $\{X \to Y, X \not\to Y\}$.

4 Experimental Details

Notation. Before explaining the experimental setup, we introduce some notation that will simplify our description. Given events X and Y, we use (X, Y) to denote the relative position where X is mentioned before Y, e.g. 'X causes Y' or 'X preceded Y'. We use $T(r, \pi)$ to denote the set of all templates for a relation r between X and Y with relative position π where π is (X, Y), (Y, X), or a random mix of both, $(X, Y) + (Y, X)$.

Training Datasets. We use the data generation algorithm from Section 3.1 to create multiple datasets with different relations and templates. For all sets, we use up to 6 templates. Appendix A.7 lists all templates. We create the following datasets for each relation: $D_{\texttt{temporal},(X,Y)}$ contains temporal relations using templates $T(\texttt{temporal}(X, Y), (X, Y))$; $D_{\texttt{temporal},(Y,X)}$ contains temporal relations using templates $T(\texttt{temporal}(X, Y), (Y, X))$; $D_{\texttt{temporal}}$ contains temporal relations with randomized positions $T(\texttt{temporal}(X, Y), (X, Y) + (Y, X))$; $D_{\texttt{spatial}}$ contains positive and negative spatial relations using $T(\texttt{spatial}_{+}(X, Y), (X, Y) + (Y, X))$ and $T(\texttt{spatial}_{-}(X, Y), (X, Y) + (Y, X))$; $D_{\texttt{counterfactual}}$ contains positive and negative counterfactuals using $T(\texttt{counterfactual}_{+}(X, Y), (X, Y) + (Y, X))$ and $T(\texttt{counterfactual}_{-}(X, Y), (X, Y) + (Y, X))$; $D_{\texttt{all}}$ is the union of $D_{\texttt{temporal}}$, $D_{\texttt{spatial}}$, and $D_{\texttt{counterfactual}}$. Each generated dataset contains 40k scenarios. We split the datasets into 36k for finetuning and 4k for validation. Table 3 gives examples from the generated data.

Evaluation Datasets. We create two test datasets to evaluate if models can infer the presence or absence of causal relations. $D_{X \to Y}$ contains all causal relations $X \to Y$ in G_c. D_{XY} contains unrelated pairs of events, X and Y, such that neither is a descendant of the other in G_c. Note that we do not evaluate models on pairs of events X, Y such that one is a descendant (but not child) of the other. This is because, as noted by Kıcıman et al. (2023), full graph discovery is challenging and requires distinguishing between direct and indirect causes.

Data	Rel. position in train	Rel. position in eval	
		(X, Y)	(Y, X)
causal $X \to Y$	(X, Y)	92.59%	1.85%
	(Y, X)	0%	100%

Table 1: Accuracy of models finetuned on temporal relations with different relative event positions. Models infer the causal relation only when the relative position matches during finetuning and evaluation.

Training Details. We finetune LLAMA2-7B[7] using LoRA (Hu et al., 2021, applied to query and value projection matrices). See Appendix A.2 for more training details.

5 Position Heuristic

In this section, we first demonstrate that LLMs are susceptible to inferring causal relations by the relative position of two entity mentions in text (Section 5.1). We hypothesize that models learn this heuristic since it is supported in the pretraining data (Appendix A.4) and investigate ways to fix this heuristic via either augmentation or scaling up models (Section 5.2).

5.1 LLMs fail to infer causal relations if the data supports the position heuristic

First, we demonstrate that LLMs fail to infer causal relations if the data supports the position heuristic e.g. if X is mostly mentioned before Y in the text, then models fail to infer causal relations—in fact, we show that LLMs only learn the *relative position* of X and Y and ignore their relation. We refer to this as the *position heuristic*.

To show this, we finetune LLAMA2-7B separately on two datasets: $D_{\texttt{temporal},(X,Y)}$ and $D_{\texttt{temporal},(Y,X)}$.[8] We evaluate the models on the $D_{X \to Y}$ test set and report if they infer $X \to Y$. The multiple-choice options in this case are: $\{X \to Y, Y \to X, X \nleftrightarrow Y\}$. We verbalize the test relations in both directions either using $T(X \to Y, (X, Y))$ (e.g. 'X causes Y') or $T(X \to Y, (Y, X))$ (e.g. 'Y is caused by X'). In both cases, to score the relation $X \nleftrightarrow Y$ we use templates with randomized event order.

Table 1 (first two rows) shows accuracy on $D_{X \to Y}$ (i.e. the percentage of examples in which

[7] We also experiment with scaling up to LLAMA2-13B and LLAMA2-70B in Section 6.2.

[8] The position heuristic is not specific to temporal relations, but we use temporal relations here as a case study. We include results for other relations in Appendix A.3.

the model predicted $X \rightarrow Y$). We observe that models infer the causal edge only when the relative position of the two events under test matches during finetuning and evaluation. This implies that models are not learning anything meaningful to infer causal relations, but simply learning the relative position between events. For example, if models see the sentence 'X happens before Y', they would almost always predict 'X is caused by Y'.[9]

5.2 Mitigating position heuristic

In this section, we investigate two different ways to mitigate model's reliance on the position heuristic: (a) randomizing the relative positions of event mentions in the text so that the data does not support the heuristic; (b) scaling LLMs.

Extent of randomization. Here we investigate whether randomizing the relative positions of event mentions helps mitigate the model's reliance on the position heuristic. To test this, we create datasets with increasing amounts of randomness in the relative position of event mentions. Specifically, given a set of templates $T_{XY} = T(\texttt{temporal}, (X, Y))$ and $T_{YX} = T(\texttt{temporal}, (Y, X))$, we create finetuning datasets by sampling templates from T_{YX} with probability p and from T_{XY} with probability $1 - p$. Both T_{XY}, T_{YX} contain 5 templates, and we use $p \in \{0, 0.1, 0.2, 0.3, 0.4\}$ to create five finetuning datasets. For evaluation, similar to Section 5.1, we use the $D_{X \rightarrow Y}$ test set and evaluate both directions: $T(X \rightarrow Y, (X, Y))$ and $T(X \rightarrow Y, (Y, X))$.

Figure 3 (left) shows the difference in accuracy when relative position is (X, Y) (majority in finetuning data) and when relative position is (Y, X) (minority in the finetuning data). We observe that adding even a small number of examples with a different relative position (e.g. $p = 0.1$ or $p = 0.2$) helps to reduce model's reliance on the position heuristic to infer causal relations.

Scaling LLMs. Given recent observations that scaling LLMs leads to less reliance on spurious correlations (Si et al., 2022), we investigate if the same holds true for the position heuristic. To control for other factors, we use models from the same family—we experiment with LLAMA2-13B and LLAMA2-70B. Both models were finetuned similarly to the smaller LLAMA2-7B model—experimental details can be found in Appendix A.2.

[9] We further show that models are only relying on relative position instead of reasoning about causal relations by using unrelated relations for evaluation in Appendix A.3.

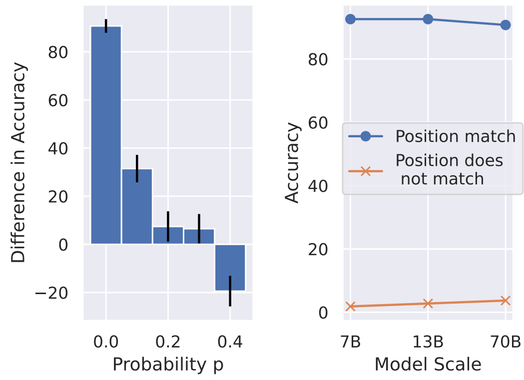

Figure 3: (left) Mitigating position heuristic by gradually randomizing the relative position. We observe that even a small amount of randomization in position is enough to reduce model's reliance on the position heuristic; (right) Scaling curve (7B to 70B) for the position heuristic — scaling does not mitigate model's reliance on the position heuristic.

	$D_{X \rightarrow Y}$	D_{XY}
Temporal Relations	76.85%	-
Spatial Relations	-	84.5%
Counterfactuals	28.70%	53.5%
All relations	63.88%	47.5%

Table 2: Accuracy on each reasoning task using models trained on data with randomized order of event mentions. LLMs is able to reason from temporal relations and spatial relations, but not from counterfactuals.

Figure 3 (right) shows the scaling trend for models trained on $D_{\texttt{temporal},(X,Y)}$ and evaluated on $D_{X \rightarrow Y}$. All models are evaluated using templates from either $T(X \rightarrow Y, (X, Y))$ (position matches) or $T(X \rightarrow Y, (Y, X))$ (position does not match). We observe that similar to the smaller LLAMA2-7B, the larger models also fail to make any meaningful deduction and only learn the relative position of the events. This shows that simply scaling LLMs is limited in resolving the position heuristic.

6 Inferring Causal Relations under No Position Heuristic

The previous section demonstrated that if the data supports the position heuristic, models fail to infer any causal relations and only rely on the relative position between events to infer causal relations. However, it is easy to mitigate the position heuristic by randomizing the relative positions of event mentions in the data. In this section, we evaluate whether models can make causal deductions from temporal relations, spatial relations and counterfactuals when the position heuristic is mitigated.

6.1 LLMs infer causal relations correctly from temporal and spatial relations

Here, our goal is to test whether LLMs can make the following deductions if data does not support learning the position heuristic:

1. $\texttt{temporal}(X, Y) \implies Y \not\rightarrow X$
2. $\texttt{spatial}_{-}(X, Y) \implies X \not\leftrightarrow Y$
3. $\texttt{counterfactual}_{+}(X, Y) \implies X \rightarrow Y$
4. $\texttt{counterfactual}_{-}(X, Y) \implies X \not\rightarrow Y$

To test this, we finetune LLAMA2-7B separately on three datasets, $D_{\texttt{temporal}}$, $D_{\texttt{spatial}}$, and $D_{\texttt{counterfactual}}$. All datasets have randomized relative position as mentioned in Section 4. Additionally, we also finetune LLAMA2-7B on $D_{\texttt{all}}$ containing all three types of relations. This is to test whether models can better infer causal relations when the data consists of diverse relations. We report model accuracy which is the percentage of examples where it makes the correct deduction according to the above rules.

We then evaluate the models on two test sets, $D_{X \rightarrow Y}$ and D_{XY}, depending on which deduction rule we are evaluating. For temporal relations, we evaluate on $D_{X \rightarrow Y}$ and report the percentage of examples where model predicts $Y \not\rightarrow X$. For spatial relations, we evaluate on D_{XY} and report the percentage of cases where model predicts $X \not\leftrightarrow Y$. For models trained on counterfactuals, we evaluate on both $D_{X \rightarrow Y}$ (report percentage of cases model predicts $X \rightarrow Y$) and D_{XY} (report percentage of cases model predicts $X \not\rightarrow Y$). Lastly for models trained on all relations, we also evaluate on both: $D_{X \rightarrow Y}$ (report percentage of cases model predicts $X \rightarrow Y$) and D_{XY} (report percentage of cases model predicts $X \not\leftrightarrow Y$). For all evaluations, we use randomized event order to score all relations.

Table 2 shows the results. We find that models can correctly deduce the absence of causal relations from temporal relations and spatial relations better than random guessing (which is 50% and 33.3% respectively), but cannot deduce causal relations from either positive counterfactual or negative counterfactuals (random guessing is 33.3% and 50% respectively).

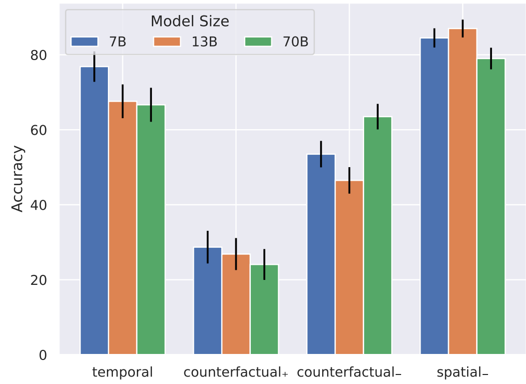

Figure 4: Scaling trend for inferring causal relations from different relations when there is no position bias.

6.2 Does scaling LLMs improve causal inference?

The previous sections showed LLAMA2-7B can infer causal relations from temporal relations and spatial relations. However, the model could not deduce either the presence or absence of edges from counterfactuals. Given recent observations that scaling LLMs leads to better performance (Kaplan et al., 2020) and emergent abilities (Wei et al., 2022), we explore whether scaling LLMs can improve their ability to infer causal relations from counterfactuals.

We use models from the same family, LLAMA2-13B and LLAMA2-70B finetuned similarly to the smaller LLAMA2-7B model. Experimental details can be found in Appendix A.2. Figure 4 shows the scaling trend of models in terms of the accuracy of deducing the correct causal relation from each of the relations. We observe that scaling model size does help the model to deduce the absence of causal relations from negative counterfactuals (third group in figure) better than random guessing (50%). However, we do not see similar scaling trend for inferring causal relations from positive counterfactuals, where models do not perform better than random guessing (33.3%). For temporal relations and spatial relations, we do not see significant differences with scaling model size (all our within standard error of the other).

7 LLMs Suffer from Post Hoc Fallacy

Section 6 demonstrated that when the data does not support the position heuristic, LLMs can correctly infer the absence of causal relations from temporal and spatial relations. In this section, we demonstrate that for temporal relations, models in fact

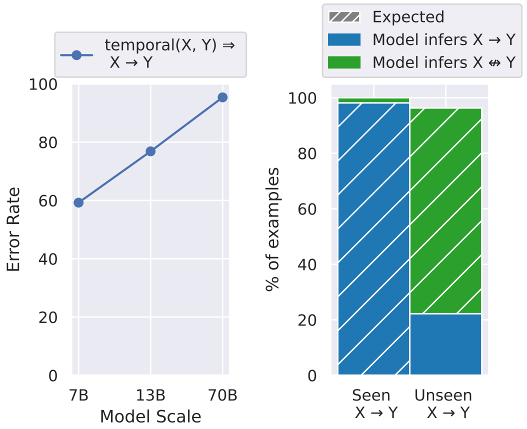

Figure 5: (left) Scaling curve showing that larger models also suffer from post hoc fallacy; (right) Post hoc fallacy can be fixed by finetuning.

overgeneralize to infer the *presence* of causal relations in the other direction. This mistake is often referred to as the *post hoc fallacy* (Woods and Walton, 1977), which uses the incorrect deduction rule: $\texttt{temporal}(X, Y) \implies X \to Y$. Humans have known to often fall prey to this fallacy and infer causal relations from sequential order (Nisbett and Ross, 1980; Gilovich, 1991).

To demonstrate this, we finetune models from the LLAMA2 family (7B to 70B) on $D_{\texttt{temporal}}$ (where the templates have randomized order) and evaluate them on $D_{X \to Y}$ to see if they infer $X \to Y$. All templates in the evaluation use randomized event order $T(r, (X, Y) + (Y, X))$ for each relation r in the multiple-choice options.

For evaluation, we report the error rate which is the percentage of examples where the model incorrectly deduces $X \to Y$ from $\texttt{temporal}(X, Y)$. Figure 5 (left) shows the error rate. We observe that all models incorrectly infer the causal relation better than random guessing (33.3%). Interestingly, we observe an inverse scaling trend (McKenzie et al., 2023) — scaling model size increases the error and models rely on the post hoc fallacy more.

7.1 Fixing the post hoc fallacy by finetuning

The previous section demonstrated that LLMs of all scales, from 7B to 70B, suffer from the post hoc fallacy. A natural question to ask here is—can LLMs be finetuned to correct this fallacy so that they don't overgeneralize?

To answer this, we include explicit statements of presence and absence of causal relations in the finetuning data. Including explicit causal relations can teach the model that $\texttt{temporal}(X, Y)$ does not necessarily imply $X \to Y$. We first create two subsets of the $D_{X \to Y}$ test set: $D_{\text{seen}, X \to Y}$ and $D_{\text{unseen}, X \to Y}$. For each causal relation in the seen subset, we include the *explicit* causal relation in the corresponding scenario e.g. we add an additional sentence 'event10 can cause event12' to the scenario which may include other relations between the same two events (e.g. 'event10 happened before event12'). Similarly, for events which are not causally related we include explicit negative causal relation in the corresponding scenario e.g. if in the ground truth graph G_c, event6 and event8 are not causally related, we add a statement 'event6 does not cause event8' to a scenario involving the two events (where the scenario may include the temporal relation 'event6 occurs before event8').

We then evaluate a model finetuned on this dataset on the $D_{\text{unseen}, X \to Y}$ subset for which the model has not seen any explicit causal relations. As a sanity check, we also evaluate the model on $D_{\text{seen}, X \to Y}$ to show that models memorize the causal relation if they have seen it explicitly. All evaluations use randomized event orders.

Figure 5 (right) shows the percentage of examples and the model predictions. We observe that the model tends to predict $X \not\leftrightarrow Y$ more often than $X \to Y$ on the unseen subset, i.e. the model learns that temporal relations do not necessarily imply the presence of a causal relation, and hence the post hoc fallacy can be mitigated via finetuning.

8 Conclusion

In this work, we investigate whether LLMs can be useful for causal inference beyond explicitly-memorized causal facts. We find that LLMs are susceptible to inferring causal relations from position, but this can be mitigated by data augmentation. We find that LLMs can infer causal relations from temporal relations and spatial relations, but not from counterfactuals. Overall, we find that LLMs may not infer much novel causal knowledge beyond explicitly mentioned facts in the pretraining data. Our setup also allows for the exploration of interesting questions such as whether models generalize to events of the same 'type' (e.g. if smoking and vaping occur in similar contexts, and the data includes smoking causing cancer, does the model generalize to infer any relation between vaping and cancer?), and if models can generalize to transitive relations. We leave these questions for future work.

Limitations

To address our main research question of whether LLMs can go beyond memorized causal facts to *infer* causal relations, we disentangle memorization vs inference via use of synthetic data. While synthetic data helps us to do controlled experiments, it has certain limitations due to the gap between synthetic and real data. Nevertheless, experiments with synthetic data have been proven extremely valuable in the community ranging from question answering (Weston et al., 2015) to reasoning (Saparov and He, 2023) to LLM-agents (Côté et al., 2018).

References

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023a. Taken out of context: On measuring situational awareness in llms. *ArXiv*, abs/2309.00667.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023b. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *ArXiv*, abs/2309.12288.

Stella Biderman, Hailey Schoelkopf, Quentin G. Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *ArXiv*, abs/2304.01373.

Neil Bramley, Tobias Gerstenberg, and David Lagnado. 2014. The order of things: Inferring causal structure from temporal patterns. In *Proceedings of the annual meeting of the cognitive science society*, volume 36.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben A. Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for text-based games. In *CGW@IJCAI*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027.

Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. 2021. A counterfactual simulation model of causal judgments for physical events. *Psychological review*, 128(5):936.

Thomas Gilovich. 1991. How we know what isn't so: The fallibility of human reason in everyday life.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

Irving J Good. 1961. A causal calculus (i). *The British journal for the philosophy of science*, 11(44):305–318.

Roger Baker Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamil.e Lukovsiut.e, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Sam Bowman. 2023. Studying large language model generalization with influence functions. *ArXiv*, abs/2308.03296.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Joseph Y. Halpern. 2015. A modification of the halpern-pearl definition of causality. *ArXiv*, abs/1505.00162.

Joseph Y Halpern. 2016. *Actual causality*. MiT Press.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM*. ACM.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*.

Nitish Joshi, Xiang Pan, and Hengxing He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Conference on Empirical Methods in Natural Language Processing*.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020.

Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Emre Kıcıman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *ArXiv*, abs/2305.00050.

Andrew Kyle Lampinen, Stephanie C. Y. Chan, Ishita Dasgupta, Andrew Joo Hun Nam, and Jane X. Wang. 2023. Passive learning of active causal strategies in agents and language models. *ArXiv*, abs/2305.16183.

Nayoung Lee, Kartik K. Sreenivasan, Jason D. Lee, Kangwook Lee, and Dimitris Papailiopoulos. 2023. Teaching arithmetic to small transformers. *ArXiv*, abs/2307.03381.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Annual Meeting of the Association for Computational Linguistics*.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Sam Bowman, and Ethan Perez. 2023. Inverse scaling: When bigger isn't better. *ArXiv*, abs/2306.09479.

Peter Menzies and Helen Beebee. 2024. Counterfactual Theories of Causation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. Metaphysics Research Lab, Stanford University.

Joris M. Mooij, J. Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Scholkopf. 2014. Distinguishing cause from effect using observational data: Methods and benchmarks. *ArXiv*, abs/1412.3773.

R.E. Nisbett and L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Century psychology series. Prentice-Hall.

Travis Norsen. 2007. John s. bell's concept of local causality. *American Journal of Physics*, 79:1261–1275.

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

Judea Pearl. 2022. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 317–372.

Hans Reichenbach. 1956. *The direction of time*, volume 65. Univ of California Press.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in llms. *ArXiv*, abs/2307.14324.

Ruoqi Shen, Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, Yuanzhi Li, and Yi Zhang. 2023. Positional description matters for transformers arithmetic. *ArXiv*, abs/2311.14737.

Yoav Shoham. 1987. *Reasoning about change: time and causation from the standpoint of artificial intelligence*. Yale University.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *ArXiv*, abs/2210.09150.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *ArXiv*, abs/2206.07682.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv: Artificial Intelligence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

John Woods and Douglas Walton. 1977. Post hoc, ergo propter hoc. *Review of Metaphysics*, 30(4):569–593.

Linying Yang, Oscar Clivio, Vik Shirvaikar, and Fabian Falck. 2023. A critical review of causal inference benchmarks for large language models. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*.

W. Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *ArXiv*, abs/2305.14010.

M. Zecevic, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *ArXiv*, abs/2308.13067.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022. On the paradox of learning to reason from data. In *International Joint Conference on Artificial Intelligence*.

A Appendix

A.1 Additional Details on Synthetic Data Generation

Generating Causal Graphs. To generate a synthetic causal graph, we generate a directed acyclic graph with n vertices and r root vertices. Each vertex represents an event, and the root vertices are those that have no causes (i.e. they have no incoming edges). The algorithm to generate such a graph is shown in Algorithm 2. The algorithm is fairly simple, but we take care not to create vertices that are descendants of all roots, since they will be causally connected to every root, and therefore, they would never be sampled in any event chain in Algorithm 1. In addition, we require that every root has at least one child, in order to prevent generating trivial event chains that contain only a single event. In our experiments, we fix $n = 100$, and r is sampled from Geometric(0.64) conditioned on $r \in [3, 6]$.

Generating Non-causal Relation Graphs. Algorithm 3 describes how we generate non-causal relations for the events in the causal graph. The output is a graph G_n where each vertex represents a *type* of event, and the function constructs a map T from events in G_c to event types in G_n. We chose simple semantics for G_n: If two events have the same type, they co-occur. An edge $a \to b$ in G_n from event type a to event type b indicates that all events of type a precede events of type b.

Algorithm 1: Pseudocode to generate synthetic relational data from causal graph G_c and non-causal relation graph G_n. The helper-function `sample_event_chains` is described in Algorithm 4.

Input: `num_scenarios`, set of events E, causal graph G_c, relation graph G_n
Output: dataset D

initialize $D \leftarrow \{\}$
repeat *`num_scenarios`* **times**
  /* sample a number of event chains, where each chain is causally-independent of the other chains */
  $C \leftarrow$ `sample_event_chains(`G_c`)`
  $S \leftarrow \{\}$
  for each *`event_chain`* *in* C **do**
    /* sample temporal relations */
    $n \sim$ Binomial(|`event_chain`|, 0.5)
    sample S, a set of n events, from `event_chain`
    for each X_i *in* S **do**
      sample event Y uniformly from any chain in C
      if X_i *is an ancestor of* Y *in* G_n
        S`.add(temporal(`X_i, Y`))`
      else if Y *is an ancestor of* X_i *in* G_n
        S`.add(temporal(`Y, X_i`))`
      else if X_i *and* Y_i *do not co-occur in* G_n
        S`.add(temporal(`X_i, Y`)` w.p. 0.5, else `temporal(`Y, X_i`))`
    /* sample spatial relations */
    $n \sim$ Binomial(|`event_chain`|, 0.4)
    sample S, a set of n events, from `event_chain`
    for each X_i *in* S **do**
      sample event Y uniformly from any chain in C
      if $Y \in$ *`event_chain`* *or* X_i, Y_i *co-occur in* G_n
        S`.add(spatial`$_+$`(`X_i, Y`))`
      else S`.add(spatial`$_-$`(`X_i, Y`))`
    /* sample counterfactual relations */
    $n \sim$ Binomial(|`event_chain`|, 0.4)
    sample S, a set of n events, from `event_chain`
    for each X_i *in* S **do**
      $Y \sim$ Uniform(`event_chain` $\setminus \{X_i\}$)
      if X_i *is an ancestor of* Y *in* G_c
        S`.add(counterfactual`$_+$`(`X_i, Y`))`
      else S`.add(counterfactual`$_-$`(`X_i, Y`))`
    /* sample negative counterfactuals */
    $n \sim$ Binomial(|`event_chain`|, 0.2)
    sample S, a set of n events, from `event_chain`
    for each X_i *in* S **do**
      sample event Y uniformly from any chain in C
      if X_i *is an ancestor of* Y *in* G_c
        S`.add(counterfactual`$_+$`(`X_i, Y`))`
      else S`.add(counterfactual`$_-$`(`X_i, Y`))`
  D`.add(`S`)`

Sampling Event Chains. Algorithm 4 describes the helper function used in Algorithm 1 which sam-

Algorithm 2: Pseudocode for generating a synthetic causal graph.

Input: number of vertices n, number of roots r
Output: causal graph G_c
initialize G_c as a graph with n vertices and no edges
let $(v_1, \ldots, v_n)$ be the vertices of G_c
for i *in* $r+1, \ldots, n$ **do**
  $m \sim \text{Zipf}(3)$
  $m \leftarrow \min(i, m)$
  sample P, a set of m vertices from $\{v_1, \ldots, v_{i-1}\}$, uniformly without replacement
  for p *in* P **do**
    add edge $p \to v$ to G_c
    if *v is a descendant of all roots $v_1, \ldots, v_r$*
      remove edge $p \to v$ from G_c
`/* make sure each root has` $\geq$ `1 child */`
for v_i *in* $\{v_1, \ldots, v_r\}$ **do**
  if *v_i has no child vertices*
    $v \sim \text{Uniform}(v_{r+1}, \ldots, v_n)$
    add edge $v_i \to v$ to G_c
shuffle the vertices $(v_1, \ldots, v_n)$

Algorithm 3: Pseudocode for generating a synthetic non-causal relation graph.

Input: causal graph G_c
Output: non-causal relation graph G_n
let $(t_1, \ldots, t_k)$ be (an initially empty) ordered list of event types
let T be an initially empty map from events in G_c to event types $\{t_1, \ldots, t_k\}$
for each *event v in G_c* **do**
  `/* assign an event type to each event in` G_c `*/`
  compute $\alpha = \max\{i :$ there is an ancestor a of v such that $T(a) = t_i\}$
  compute $\beta = \min\{i :$ there is a descendant d of v such that $T(d) = t_i\}$
  if $\alpha < \beta$
    $w \sim \text{Uniform}(t_{\alpha+1}, \ldots, t_{\beta-1})$
  else
    create new event type w and insert it into the list of event types at index $\alpha + 1$
  set $T(v) \leftarrow w$
let $(t_1, \ldots, t_k)$ be the vertices of G_n
`/* add temporal edges between event types */`
for each *event v in G_c* **do**
  for each *child vertex c of v* **do**
    add edge $T(p) \to T(c)$ to G_n

ples a handful of event chains, where each chain is causally-independent of the other event chains. In this helper function, each event chain starts at a root node in G_c, since root nodes are by definition causally-independent of each other. We sample the length of each chain to be uniform so that vertices near roots are not over-represented in the sample of event chains (and vertices further from the roots are not under-represented). This helps to facilitate more uniform coverage of all vertices in G_c by the generated data.

Algorithm 4: Pseudocode for the helper-function `sample_event_chains`, which, given a causal graph G_c, returns a number of event chains, where each chain is causally-independent of the other chains.

Input: causal graph G_c
Output: set of event chains C
initialize $C \leftarrow \{\}$
$n \sim 1 + \text{Geometric}(0.25)$
sample R, a set of n root vertices from G_c (with no incoming edges), uniformly without replacement
`/* for each root, sample a chain */`
for each *r in R* **do**
  compute D_r, the set of descendant vertices of r
  `/* sample the length of this chain */`
  $m \sim \text{Uniform}(1, \ldots, \max_{v \in D_r} \text{distance}(r, v))$
  compute $S_{r,m} = \{v \in D_r : \text{distance}(r, v) = m$ and v is not a descendant of $R \setminus \{r\}\}$
  while *$S_{r,m}$ is empty* **do**
    $m \leftarrow m - 1$
    recompute $S_{r,m}$ as above
  `/* sample the endpoint of the chain */`
  $e \sim \text{Uniform}(S_{r,m})$
  C`.add(`set of all vertices on path from r to e`)`
`/* mark some chains as 'non-occurring' */`
$k \sim \text{Binomial}(n-1, 0.2)$
remove k event chains from C, uniformly at random

Generating Scenarios. Algorithm 1 gives the data generation algorithm for generating the scenarios. In each step, when we sample S, a set of n events from the `event_chain` we sample uniformly randomly without replacement. This ensures that scenarios contain information about a diverse set of events.

We also include an example from our generated dataset, where the scenario contains all three relations in Table 3.

A.2 Experiment Details

We used LLAMA2 models through HuggingFace's transformer library (Wolf et al., 2019). All models were finetuned with LoRA (applied to query and key projection matrices), with rank $= 16$, $\alpha = 16$ and dropout $= 0.05$. All models were finetuned

All Relations	event84 happened. event76 happened. event76 and event84 took place in the same location. if event76 did not happen, and event84 has no other causes, would event84 happen? yes. if event76 has no other causes, and event84 did not occur, would event76 still happen? no. event5 happened. event3 happened. event96 happened. event3 happened after event84. event5 happened before event3. the location of event96 is not identical to that of event76. if event3 did not happen, and event5 has no other causes, would event5 happen? yes.
Temporal Relations	event67 occurred prior to event71. event40 happened before event28. event7 preceded event28. event71 happened after event95.
Spatial Relations	the location of event96 is not identical to that of event4. event4 and event96 did not take place in the same location.
Counterfactuals	if event33 did not occur, and event84 has no other causes, would event84 still happen? yes. if event84 has no other causes, and event58 did not occur, would event84 still happen? yes. if event58 has only one cause, and hypothetically event84 did not happen, would event58 still occur? no. if event3 has only one cause, and event48 did not happen, would event3 happen? yes.

Table 3: Examples of the scenarios from our generated dataset. The first examples contains all types of relations, whereas the others include one type of relation only.

with a learning rate of $5e-4$ using AdamW optimizer(Kingma and Ba, 2015; Loshchilov and Hutter, 2017) with a batch size of 8. The models finetuned on 36k scenarios were trained for 10k steps whereas the models trained with 4.5k scenarios (500 scenarios used for validation, as in Appendix A.5) were trained for 6k steps — we generally observed that models converged around this point.

A.3 Additional Results: Position Bias

Temporal Relations. Section 5.1 showed that, in the presence of strong position bias, the model assigned high probability to $t_i(X \rightarrow Y)$ where the relative position matches that during finetuning. This still leaves open the possibility that the model is assigning a higher probability to the template for *correct* causal relation where the position matches. e.g. from 'X preceded Y', the model could assign probabilities in the following order — 'X can cause Y' > 'X can be caused by Y' > 'Y can be caused by X' > 'Y can cause X'. In such a situation, if the order is randomized during evaluation the model can still infer causal relations from temporal relations.

In this experiment, we find that models finetuned on temporal relations with relative position (X, Y) infer $X \rightarrow Y$ from $\texttt{temporal}(X, Y)$ 23.14% of the times. Since random chance is 33.3%, we see that models finetuned on position bias indeed are not able to make any consistent deduction beyond matching relative position during finetuning and evaluation.

To further show that models are only relying on the relative position of events instead of reasoning about their causal relation, we evaluate models using different relations with the *same* relative position. Specifically, we randomly sample three relations between X and Y which have no connection to the causal relation and verbalize them using the (X, Y) relative order e.g. instead of the verbalization 'X causes Y', we will use 'X is related to Y' (details in Appendix A.7). We observe a similar result in the last two rows in Table 4—models only make correct predictions when the event order during training matches that during test.

Data	Rel. position in train	Rel. position in eval (X, Y)	(Y, X)
causal $X \rightarrow Y$	(X, Y)	92.59%	1.85%
	(Y, X)	0%	100%
unrelated X, Y	(X, Y)	98.14%	0.92%
	(Y, X)	0%	100%

Table 4: Accuracy of models finetuned on temporal relations with different relative event positions. Models infer the causal relation only when the relative position matches during finetuning and evaluation.

	Rel. position during train	Rel. position - eval (X, Y)	(Y, X)
Accuracy	(X, Y)	90.5%/3.0%	6.5%/3.5%

Table 5: Models finetuned on spatial relations with fixed relative position, and we report % of cases model infer $X \rightarrow Y$ / % of cases model infers $X \nleftrightarrow Y$. Models infer the causal relation only when the relative position matches during finetuning and evaluation.

Spatial Relations. Here, we demonstrate that we also observe the position bias for spatial relations. To show this we first create a dataset with fixed relative position. Specifically, we generate a dataset $D_{\texttt{spatial},(X,Y)}$ consisting of positive and negative spatial relations from the sets $T(\texttt{spatial}_+, (X, Y))$ and $T(\texttt{spatial}_-, (X, Y))$

respectively. We then finetune LLAMA2-7B on this data and evaluate the model on $D_{\text{unrelated}X-Y}$. We use two different sets of templates to evaluate the model: $T(X \rightarrow Y, (X, Y))$ (e.g. 'X causes Y') or using templates from $T(X \rightarrow Y, (Y, X))$ (e.g. 'Y is caused by X'). In both cases, to score the relation $X \nleftrightarrow Y$ we use $T(X \nleftrightarrow Y, (X, Y) + (Y, X))$.

Table 5 shows the percentage of examples in which the model predicted either $X \rightarrow Y$ or $X \nleftrightarrow Y$ (which is the correct option). Firstly, we observe that in both cases, the model rarely selects the correct option $X \nleftrightarrow Y$. Similar to the position bias in temporal relations, the model selects either $X \rightarrow Y$ depending on if the position matches. This shows that position bias also exists for spatial relations. We also evaluate the model using templates which have randomized relative position for each option. Specifically, we use templates from the sets $T(r, (X, Y) + (Y, X))$ where $r \in \{X \rightarrow Y, Y \rightarrow X, X \nleftrightarrow Y\}$. We find that model selects the correct option ($X \nleftrightarrow Y$), 68% of the time. This is in contrast to the position bias in temporal relations, where the performance was close to random chance. Nevertheless, the model still performs worse than if the position was randomized in the finetuning data (84.5%, Table 2)

In summary, we find that the position bias also holds true for spatial relations, albeit to a lesser extent than that for temporal relations.

A.4 Position heuristic is supported in the pretraining data

Section 5.1 demonstrated that LLMs fail to infer causal relations if the finetuning data supports the position heuristic. We hypothesize that this phenomenon occurs since the position heuristic is supported in the pretraining data — if cause is often mentioned before effect in the text, then LLMs can use relative position as a heuristic for the language modeling task. E.g. for the causal relation 'smoking causes cancer', we hypothesize that 'smoking' usually occurs before 'cancer' if they co-occur within a window. Thus a LLM trained on such data can do well even if it only uses the heuristic of relative position to predict the next word and ignore the relation between the two events.

To test if this holds true in the pretraining data, for a given causal relation $X \rightarrow Y$, we count the number of times X occurs before or after Y in a context window. We expect that if the heuristic is supported in the pretraining data, then X should mostly occur before Y when they co-occur in a context window.

We first create a set of 40 commonly-queried causal relations (e.g. smoking causes cancer, bacteria causes infections, etc.) based on the edges from the CauseNet dataset (Heindorf et al., 2020), the Tubingen dataset (Mooij et al., 2014) as well as some candidates from GPT-4. Then for each of the causal relations $X \rightarrow Y$, we count the number of documents of the PILE[10] corpus (Gao et al., 2020) in which either X occurs before Y or Y occurs before X within a window of 50 characters of the first mention of X and Y in the document. We filter to keep only those edges where the events co-occur within the context window at least 100 times. See Appendix A.8 for details.

Across all causal relations, we find that whenever X, Y co-occur within the context window, 60.77% of the times X occurs before Y. Overall, we observe that the data supports the heuristic in a majority (> 50%) of the examples.

A.5 Additional Results: Frequency vs Position Bias

We also observe an interesting trend where models exhibit a stronger position bias for relations that are more frequent in the finetuning data. To show this, we first create a smaller dataset by sampling 5k examples from $D_{\texttt{temporal},(X,Y)}$ — 4.5k for finetuning, 500 for evaluation — and finetune for fewer steps. We split the test set $D_{X \rightarrow Y}$ into 10 equal sized buckets based on the frequency of the corresponding temporal relation, $\texttt{temporal}(X, Y)$, in $D_{\texttt{temporal},(X,Y)}$.

Figure 6 shows the result where the X-axis is the frequency buckets, and Y-axis is the difference in accuracy between the test set with matched and unmatched X-Y orders. We observe that high frequency relations are correlated with a larger gap.

We also report the absolute accuracy when the model is trained on the smaller finetuning dataset with 4.5k scenarios. As shown before, in this case we observed the position bias for high frequency relations. In Table 6, we report the avg accuracy of models inferring $X \rightarrow Y$ for both relative positions. We observe a stronger position effect in one direction (when trained with relative position (Y, X)) but not as much in the other direction. Note

[10]The pretraining dataset for LLAMA2-7B is not available, so we use PILE and assume that relative positions would be similar.

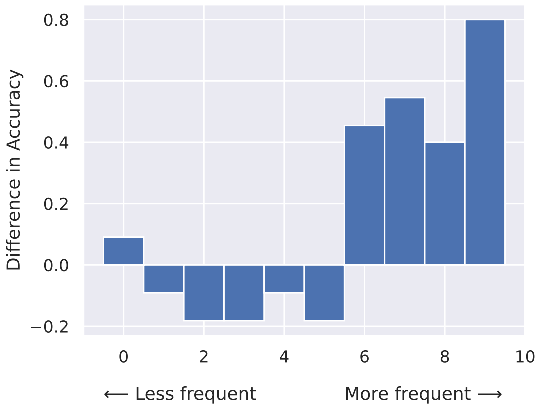

Figure 6: Difference in accuracy on the test sets with matched and unmatched event orders as a function of the frequency of the relation in the data. LLMs suffer from the position bias on high frequency events.

	Rel. position during train	Rel. position - eval (X, Y)	(Y, X)
Three-way eval	(X, Y)	52.77%	35.18%
	(Y, X)	3.70%	94.44%

Table 6: Models finetuned on 5k scenarios with temporal relations with different relative positions. We only observe the position effect in one direction (when finetuned on (Y, X)) but not the other.

that the model performance when trained with relative position (X, Y) is not much better than chance and is also sensitive to the relative position.

A.6 Additional Results: Alternate evaluation of $X \rightarrow Y$

In Section 3.2 to evaluate models, we first compute the probabilities of the following five relations under the language model: $X \rightarrow Y$, $Y \rightarrow X$, $X \not\rightarrow Y$, $Y \not\rightarrow X$, and $X \not\leftrightarrow Y$. To test if models have inferred the causal relation $X \rightarrow Y$, we compare the probabilities of the following three events which are exhaustive (i.e. their true probabilities sum to 1) and disjoint: $X \rightarrow Y$, $Y \rightarrow X$, and $X \not\leftrightarrow Y$.

An alternative set of events which are also exhaustive and disjoint are: $X \rightarrow Y$, and $X \not\rightarrow Y$. In this section, we demonstrate that our conclusion of whether models infer $X \rightarrow Y$ remains consistent even if we use these two events as the set of events to compare.

To show this, we re-evaluate two models: LLAMA2-7B finetuned on $D_{\texttt{temporal}}$, and $D_{\texttt{counterfactual}}$ respectively. We then evaluate these models on $D_{\text{causal} X-Y}$ to test if they infer

	$D_{\text{causal} X-Y}$
$\texttt{temporal}(X, Y) \implies X \rightarrow Y$	71.29%
$\texttt{counterfactual}_{+}(X, Y) \implies X \rightarrow Y$	54.62%

Table 7: Alternative Evaluation: Using a different set of exhaustive and disjoint events does not change our conclusions — model suffer from post hoc fallacy, and they cannot infer presence of causal relation from counterfactual.

presence of causal relations from either temporal relations or positive counterfactuals.

Table 7 shows the percentage of examples where model predicts the causal relation $X \rightarrow Y$. First, we observe that models infer causal relations from temporal relation — i.e. $\texttt{temporal(X, Y)} \implies X \rightarrow Y$. Therefore, similar to our previous findings where models suffer from post hoc fallacy (Section 7), changing how we evaluate the presence of causal relation does not affect our results. Similarly, we observe that models cannot infer presence of causal relations from counterfactuals much better than random chance (50%). This is consistent with our finding from Section 6, where we showed that the model cannot infer causal relations from positive counterfactuals.

A.7 Templates for Relations

In this section, we list the templates we use for each of the three relations: temporal relations, spatial relations, and counterfactuals. Additionally, we also describe the templates we used for causal relations (both presence and absence of causal relations). Each template is separated by ';'.

1. $T(\texttt{temporal}(X, Y), (X, Y))$: X preceded Y; X happened before Y; X occurred prior to Y; X took place before Y; X happened then Y happened

2. $T(\texttt{temporal}(X, Y), (Y, X))$: Y followed X; Y happened after X; Y occurred later than X; Y took place after X; Y happened later than X

3. $T(\texttt{temporal}(X, Y), \texttt{random})$: X preceded Y; Y followed X; X occurred prior to Y; Y happened after X; Y occurred later than X; X happened before Y

4. $T(\texttt{spatial}_{+}(X, Y), \texttt{random})$: X and Y took place in the same location; the location of X is identical to that of Y; X and Y happened in the same place; Y and X took place

in the same location; the location of Y is identical to that of X; Y and X happened in the same place

5. $T(\texttt{spatial}_-(X,Y), \texttt{random})$: X and Y did not take place in the same location; the location of X is not identical to that of Y; X and Y did not happen in the same place; Y and X did not take place in the same location; the location of Y is not identical to that of X; Y and X did not happen in the same place

6. $T(\texttt{counterfactual}_+(X,Y), \texttt{random})$: if X did not happen, and Y has no other causes, would X happen? no; if Y has only cause, and X did not happen, would Y happen? no; if X did not occur, and Y has no other causes, would Y still happen? no; if Y has no other causes, and X did not occur, would Y still happen? no; if hypothetically X did not happen, and Y has only cause, would Y still occur? no; if Y has only cause, and hypothetically X did not happen, would X still occur? no;

7. $T(\texttt{counterfactual}_-(X,Y), \texttt{random})$ if X did not happen, and Y has no other causes, would X happen? yes; if Y has only cause, and X did not happen, would Y happen? yes; if X did not occur, and Y has no other causes, would Y still happen? yes; if Y has no other causes, and X did not occur, would Y still happen? yes; if hypothetically X did not happen, and Y has only cause, would Y still occur? yes; if Y has only cause, and hypothetically X did not happen, would X still occur? yes;

8. $T(X \rightarrow Y, \texttt{random})$: X can cause Y; Y can be caused by X; X causally affects Y; X can lead to Y; Y is causally affected by X; Y is caused by X

9. $T(X \not\rightarrow Y, \texttt{random})$: X cannot cause Y; Y cannot be caused by X; X does not causally affects Y; X cannot lead to Y; Y is not causally affected by X; Y is not caused by X

10. $T(X \not\leftrightarrow Y, \texttt{random})$: 'there is no causal relation between X and Y', 'there is no causal relation between Y and X', 'there is no dependency between X and Y', 'there is no dependency between Y and X', 'there is no causal link between X and Y', 'there is no causal link between Y and X', 'X neither causes nor is caused by Y', 'Y neither causes nor is caused by X', 'there is no cause-and-effect relationship between X and Y', 'there is no cause-and-effect relationship between Y and X', 'there is no causal association linking X and Y', 'there is no causal association linking Y and X'

A.8 Position Heuristic in PILE

For searching through the pretraining data, we used the PILE corpus since it's freely available and has been used in recent models e.g. Pythia models (Biderman et al., 2023). Here, we list the 40 causal relations we used to search over the PILE corpus. We set the parameter w to be 50 characters i.e. the events are said to co-occur if they occur within 50 characters of each other. We filter to keep only those edges where the events co-occur enough times in the pretraining data (we set it to 100) — this is done to ensure that results are not affected by causal relations where the events do not frequently co-occur.

```
[('bacteria', 'infections'),
 ('hiv', 'aids'),
 ('cancer', 'death'),
 ('smoking', 'lung cancer'),
 ('altitude', 'temperature'),
 ('age', 'height'),
 ('sun exposure', 'aging'),
 ('sugar', 'tooth decay'),
 ('drugs', 'organ damage'),
 ('salt', 'high blood pressure'),
 ('screens', 'eye strain'),
 ('lack of sleep', 'impaired cognition'),
 ('pollution', 'lung harm'),
 ('noise', 'hearing loss'),
 ('genetics', 'height'),
 ('dehydration', 'fatigue'),
 ('sugar', 'diabetes'),
 ('stress', 'headache'),
 ('poor nutrition', 'fatigue'),
 ('sedentary habits', 'obesity'),
 ('education', 'income'),
 ('physical activity', 'health'),
 ('parental involvement', 'child development'),
 ('nutrition', 'longevity'),
 ('financial stress', 'mental health'),
 ('pollution', 'health problems'),
 ('stress', 'immune function'),
```

```
        ('education', 'political participation'),
        ('drugs', 'crime rate'),
        ('deforestation', 'climate change'),
        ('fossil fuels', 'climate change'),
        ('greenhouse gases', 'climate change'),
        ('accident', 'death'),
        ('stroke', 'death'),
        ('diabetes', 'death'),
        ('migraine', 'headache'),
        ('smoking', 'house fires'),
        ('infidelity', 'divorce'),
        ('poverty', 'homelessness'),
        ('drunk driving', 'accident')]
```