A Comprehensive Collection of Vignettes for Actual Causation

Christian Odenwald C.L.ODENWALD@SMS.ED.AC.UK
School of Informatics, University of Edinburgh, 10 Crichton St., Edinburgh EH8 9AB, UK

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Theories of actual causation provide answers to the question: "Is C a cause of E?" in a specific scenario. The performance of a new theory is measured by how well its verdicts agree with the intuitive verdicts of the researcher on particular examples, commonly referred to as vignettes.

This has two drawbacks: First, this is usually done only for a handful of vignettes per theory since there is no commonly agreed-upon collection of vignettes. That makes it difficult to compare theories against each other. Second, this evaluation is mostly done by hand. That makes it tedious for both the researcher proposing a new theory and the reader who tries to assess the merits of the new theory. To solve this, we provide a comprehensive collection of vignettes in a well-organized data format. We provide code to load these vignettes and accompanying queries. We also provide an implementation of two popular theories of causation to demonstrate the advantage of this approach.

In addition, we address the suggestion that LLMs might be more suitable than formal models of these vignettes to determine causality. To test this claim on current LLMs, we add formulations of vignettes and queries in natural language. That makes it possible to prompt LLMs for their verdict and compare their results both with intuitions and the verdicts of particular theories of actual causation. We find that none of the tested LLMs achieves higher performance than either of the two implemented theories of causation.[1]

Keywords: Actual Causation, Causal Vignettes, Structural Causal Models, Large Language Models

1. Introduction

The study of causality is usually focused on *type causation*, giving answers to general questions like: "Do lightning strikes cause forest fires?" In contrast to that, *actual causation*, also called *token causation*, is focused on the question of whether one particular past event caused some other particular past event; for example: "Did yesterday's lightning strike cause yesterday's forest fire?" We assume that all relevant events and causal connections between them are given—determining these from data is a separate field called *causal discovery* (Glymour et al., 2019). Actual causation is relevant for law (Moore, 2009), explainability (Beckers, 2022), and the formal study of blame, harm, and responsibility (Chockler and Halpern, 2003; Beckers et al., 2024).

One complicating factor is that the task is not fully objective. A theory is evaluated on several toy examples called *vignettes*, and the ground truth is the researcher's intuition, which can differ across people. Symmetric overdetermination is an example of such a case: If both a dropped match and a lightning strike suffice to start a fire, should the lightning strike count as a cause? Arguments can be made for both sides, and some (notoriously Lewis (1973)) find such examples unhelpful for that reason.

1. Code available at https://github.com/christianodenwald/causality_checker

Whenever a new theory is proposed, it is usually applied to roughly a dozen vignettes, but many more have accumulated over the years, leaving the reader to wonder whether examples were hand-picked. That leads to the second problem: Checking causality by hand is tedious. For example, applying the theory of Halpern and Pearl (2005) to their Example 5.1 requires considering over 14 million options.

To address both issues, we take a data-centric approach: compiling vignettes into an accessible format with Python code to load them and apply theories of causation. We demonstrate the usefulness of this approach by providing code for two popular theories of actual causation: HP_u (Halpern and Pearl, 2005) and HP_mod (Halpern, 2015). The computational approach cannot replace qualitative assessment, but it can help researchers find the cases that matter most.

Additionally, our approach makes it easy to address the concerns by Kıcıman et al. (2024): that LLMs are more suitable than formal models for actual causation, since they implicitly model all the background conditions, which is a known difficulty in the formal study of actual causation. We include natural language descriptions of vignettes and causal queries and prompt LLMs for yes/no causal verdicts, enabling direct comparison with formal theories.

Our contributions can be summarized as such:

- A comprehensive collection of relevant vignettes for actual causation.
- A framework and code for computational evaluation of theories of causation, demonstrated on two established theories.
- An evaluation pipeline for LLMs on actual causation, in a direct comparison to formal theories of actual causation.

The rest of the paper is structured as follows: In section 2, we provide an overview of the field of actual causation, in particular the popular formalism of Structural Causal Models (SCMs). We also give a quick overview of previous approaches to probe LLMs for causal verdicts. In section 3, we present the dataset, some basic functions, and instructions for how to use it. In section 4, we present results of using it to evaluate further examples on HP_mod, and on LLMs. In section 5, we describe limitations and some possibilities for further work.

2. Related Work

Here we review the basics of actual causation: the popular formalism of SCMs, previous collections of vignettes, and one particular definition of actual causation on which we later demonstrate the usefulness of our approach. We also look at previous work on prompting LLMs for causal verdicts.

2.1. Actual Causation

Theories of actual causation can be classified by the basic mechanism they propose for what constitutes causation, typically counterfactuals, regularities, probabilities, or physical processes (Gallow, 2022). Counterfactual theories have been particularly popular. The basic idea is that what it means for C to cause E is that if C had not happened (and all other things roughly being equal), E would not have happened. To determine actual causation in the real world, we first need to choose a representation of the world, a model. SCMs are a popular framework for modeling because they allow for a relatively simple definition of counterfactuals.

2.1.1. STRUCTURAL CAUSAL MODELS

Most theories of causation are formulated in the SCM framework, popularized by Pearl (2000)[2]. Following Halpern and Pearl (2005) and Gallow (2021), a structural causal model M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ is called a *signature*. M consists of:

1. a set of *exogenous* variables $\mathcal{U} = \{U_1, \ldots, U_m\}$,
2. a set of *endogenous* variables $\mathcal{V} = \{V_1, \ldots, V_n\}$,
3. a set of *possible values* $R(Y_i) = \{y_{i1}, \ldots y_{ik}\} \subseteq \mathbb{N}$ for each variable $Y_i \in \mathcal{Y} = \mathcal{U} \cup \mathcal{V}$,
4. a set of *structural equations* $\mathcal{F} = \{F_{V_1}, \ldots, F_{V_n}\}$ for each $V_i \in \mathcal{V}$, where F_{V_i} is a function $F_{V_i} : (\times_{U \in \mathcal{U}} R(U)) \times \left(\times_{V \in \mathcal{V} - \{V_i\}} R(V)\right) \to R(V_i)$.

To represent a specific situation, we also need a *context*:

5. an assignment of values $\boldsymbol{u} = \{u_1, \ldots, u_m\}$ to $\mathcal{U}$, where $u_i \in R(U_i)$.

In the standard formulation by Halpern and Pearl (2005), exogenous variables are not explicitly modeled. However, it has become common practice to treat variables solely determined by $\mathcal{U}$ as exogenous (Gallow, 2021, p.50; cf. Figure 1). We follow this latter practice.

To determine the values of all variables in an SCM, we start from the exogenous variables, whose values are assigned by the context. These values are propagated through the model following the structural equations. Structural equations are asymmetric, i.e., they assign a value to the left side based on the values on the right side. SCMs are deterministic. Given the context, there is only one possible setting of variable values in the model.

An SCM can be visually represented by a graph. This graph is usually acyclic, so it is a directed acyclic graph (DAG). In the typical case, variables $\mathcal{Y}$ are Boolean, but they can also be ordinal or numerical. The structural equations are adjusted to match the variable type. If all variables are Boolean, we can represent their values by coloring in the respective node in the DAG. Arrows with pointed heads can be thought of as stimulating connections, and arrows with a circled head as inhibitory. This notation is borrowed from Neuron Diagrams[3], a different but related formalism.

For illustration, consider the following example:

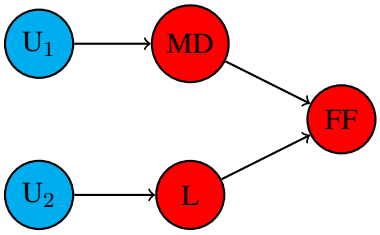
Example 1 (Forest fire (disjunctive)) *The forest burns down (FF) if either a match is dropped (MD) or lightning strikes (L). A match is dropped and lightning strikes.*

We visualize a model for this vignette in Figure 1. MD and L would be considered exogenous, while FF is endogenous. The colored nodes represent the variables that have value 1, or, in other words, the events that actually occur (in this case, all of them do).

2.1.2. NORMALITY

One problem for determining causality is cases of isomorphism (Menzies, 2017). These are cases that are structurally identical but have differing intuitive evaluations (Hall, 2007). To distinguish cases like these, normality considerations have been taken into account (Halpern and Hitchcock, 2015; Halpern, 2016), resulting in an extended SCM (though some disagree with this solution, cf.

2. Other noteworthy mentions of formal systems include Neuron Diagrams (Lewis, 1987; Hall, 2004), CP-logic (Beckers and Vennekens, 2016; Denecker et al., 2018), Causal Calculus (Bochman, 2018), and Situation Calculus (Batusov and Soutchanski, 2018) .
3. Neuron Diagrams are a special case of extended SCMs, which we introduce below. We engage in the imprecise but common mix of SCMs with Neuron Diagrams if the variables are two-valued and the equations follow standard Neuron Diagram semantics (Hitchcock, 2007b; Erwig and Walkingshaw, 2010).



$$\begin{aligned} \text{MD} &:= 1 \\ \text{L} &:= 1 \\ \text{FF} &:= \text{MD} \vee \text{L} \end{aligned}$$

Figure 1: Forest fire (disjunctive)

Blanchard and Schaffer (2017)). The idea behind that is that causes and effects need to be deviant events—me continuing to live because I breathe should not count as actual causation, since both of these events are "normal", according to most understandings of what defines normality. How to actually distinguish between the default and deviant values of a certain variable is a question subject to further research.

Adding normality to SCMs can be done by assigning a typicality ordering to the values of each individual variable (Gallow, 2021). In Figure 1, we would add a partial order to each R_{U_i}, so, for example, we would specify that $MD = 0$ (not dropping matches in forests) is more normal than $MD = 1$ (dropping matches in forests). Or, we can add a partial order to settings of the exogenous variables in a model ("small worlds"; Halpern and Hitchcock, 2015; Hitchcock and Knobe, 2009). In Figure 1, we would say that a world $\{MD = 0, L = 0\}$ is more normal than any other world. The partial order can also be applied to settings of all variables in the model ("large worlds"; Halpern, 2015, 2016). In this case, $\{MD = 0, L = 0, FF = 0\}$ would be the most normal setting. So, formally, we can add a partial order to the possible variable values $R(Y_i)$, or we can add an additional component to the SCM:

6. a *partial order* $\geq$ of possible context settings $\mathcal{U} = \prod_{i=1}^{m} R(U_i)$, or of all variable values $\mathcal{Y} = \prod_{i=1}^{m+n} R(Y_i)$.

2.1.3. VIGNETTES AND QUERIES

A theory of actual causation τ should give a verdict $\tau(V, Q) \in \{\top, \bot\}$ for each vignette V and query Q, where $Q = (C, E)$ is a tuple consisting of a potential cause $C \in \boldsymbol{Y}$ and a potential effect $E \in \boldsymbol{V}$. Some theories also want to make claims about compound causes, where multiple of the Y_i's, denoted $\boldsymbol{C} = \{C_1, \ldots, C_p\} \subseteq \boldsymbol{Y}$, can be a cause in conjunction $\bigwedge_{i=1}^{p}(C_i = c_i)$. Other theories provide a contrastive account, saying that $C = c$ rather than c' is a cause of $E = e$ rather than e'. The most general case is therefore $\tau(V, ((\boldsymbol{C}, \boldsymbol{c}, \boldsymbol{c}'), (\boldsymbol{E}, \boldsymbol{e}, \boldsymbol{e}')))$, where $\boldsymbol{C} \subseteq \boldsymbol{Y}$; $\boldsymbol{E} \subseteq \boldsymbol{V}$; $\boldsymbol{c}, \boldsymbol{c}' \in \prod_{C_i \in \boldsymbol{C}} r(C_i)$; and $\boldsymbol{e}, \boldsymbol{e}' \in \prod_{E_j \in \boldsymbol{E}} r(E_j)$.

The results of these queries are usually checked by hand on a few examples. Some efforts have been made to collect all of the vignettes in one place. Paul and Hall (2013) provide a convenient overview of many neuron diagrams. However, they only focus on causation within the graphs and rarely accompany them with natural language vignettes. Zhou (2017) comes closest to the current project. They collect vignettes from three sources: Paul and Hall (2013); Weslake (2015); Giordani (2016). However, what is collected here are merely the DAGs without any context for the vignettes. Furthermore, the format is inaccessible, and the manuscript has never been published. Kıcıman et al. (2024) tested LLMs for a variety of causal concepts. One of them was actual causation. However, they only test if the LLM labels events as necessary causes and sufficient causes. These definitions

are not commonly used in the actual causation literature. Furthermore, their data set is relatively limited. They only use 15 examples, taken from Kueffner (2021). Finally, we also draw inspiration from Weslake (2015) and Andreas and Günther (2024), who provide an overview of their theories in a tabular format, instead of only discussing vignettes in the main text. This makes it much easier to see differences between theories.

2.1.4. A LANGUAGE FOR CAUSATION

Following Halpern and Pearl (2005), we introduce a formal language to talk about causation within SCMs. A *primitive event* is a formula of the form $X = x$, for $X \in V$ and $x \in \mathcal{R}(X)$. A *causal formula* (over $\mathcal{S}$) is of the form $[\boldsymbol{Y} \leftarrow \boldsymbol{y}]\varphi$, where $\boldsymbol{Y} = \{Y_1, \ldots, Y_k\} \in \mathcal{V}$ is a set of endogenous variables with corresponding values $\boldsymbol{y_i} \in \mathcal{R}(Y_i)$, and φ is a Boolean combination of primitive events. If $\boldsymbol{Y} = \emptyset$, we can simply write φ. Intuitively, $[\boldsymbol{Y} \leftarrow \boldsymbol{y}]\varphi$ means "φ would hold if the variables $\boldsymbol{Y}$ were set to $\boldsymbol{y}$".

To express that a causal formula ψ is true in a model, given a context, we write $(M, \boldsymbol{u}) \models \psi$. Often, the values of specific variables rather than their Boolean combinations are of interest. To express that variable X has value x in model M with context $\boldsymbol{u}$, we write $(M, \boldsymbol{u}) \models X = x$.

2.1.5. THE HP DEFINITION OF ACTUAL CAUSATION

SCMs are useful because they allow for a simple definition of counterfactuals. For that, we remove all incoming connections to C and set $C = c^*$. We then propagate the values as before through the model, following the structural equations. If this results in E getting assigned the value e^*, the counterfactual is true.

The primitive counterfactual definition fails in cases of overdetermination: if there are two events where each by itself is sufficient for C, there is no counterfactual dependence. For example, in Figure 1, $FF = 1$ is symmetrically overdetermined by $MD = 1$ and $L = 1$: both individually would suffice for $FF = 1$. Assume that we think that $MD = 1$ is supposed to count as a cause of $FF = 1$ (not an uncontroversial assumption, as mentioned above). To uncover the counterfactual dependence of $MD = 1$ on $FF = 1$, we would need to consider a world where $L = 0$. Counterfactual theories mainly differ in the mechanism by which they define the alternative worlds that are tested for counterfactual dependence. Such interventions can be formalized as such: $(M, \boldsymbol{u}) \models [\boldsymbol{C} \leftarrow \boldsymbol{c}, \boldsymbol{Y} \leftarrow \boldsymbol{y}](\boldsymbol{E} = \boldsymbol{e})$, where $\boldsymbol{C}$ are the cause-variables in the antecedent of the counterfactual with their assigned values $\boldsymbol{c}$, $\boldsymbol{Y}$ are some other variables whose structural equations are ignored, and their values set to $\boldsymbol{y}$, and $(\boldsymbol{E} = \boldsymbol{e})$ is the effect variable and its value.

One influential theory of actual causation was proposed by Halpern and Pearl (2005), and later modified (Halpern, 2015). We call these HP_u and HP_mod, respectively. We introduce HP_mod here briefly. It consists of three conditions, two of them very simple. Assuming a model M and a query Q, condition AC1 says that the actual values of $\boldsymbol{C}$ and $\boldsymbol{E}$ in M need to be $\boldsymbol{c}$ and $\boldsymbol{e}$. This simply says that the potential cause and potential effect need to actually occur. AC3 says that no subset of $\boldsymbol{C}$ can be a cause of $\boldsymbol{E}$ by itself; this is a simple minimality condition. The heart of the definition is AC2, which states:

Definition 1 (AC2) *There is a set $\boldsymbol{W}$ of variables in $\mathcal{V}$ and a setting $\boldsymbol{c}'$ of the variables in $\boldsymbol{C}$ such that if $(M, \boldsymbol{u}) \models \boldsymbol{W} = \boldsymbol{w}$, then*

$$(M, \boldsymbol{u}) \models \left[\boldsymbol{C} \leftarrow \boldsymbol{c}', \boldsymbol{W} \leftarrow \boldsymbol{w}\right] \neg(\boldsymbol{E} = \boldsymbol{e}).$$[4]

Intuitively, first, we need to find a subset $\boldsymbol{W}$ and hold these variables fixed at their actual values. Then, setting the cause variable(s) $\boldsymbol{C}$ to a value different from their actual values needs to change the value of the effect variable $\boldsymbol{E}$. If, through this "freezing" of variables, some hidden counterfactual dependence is uncovered, $\boldsymbol{C} = \boldsymbol{c}$ will be a cause of $\boldsymbol{E} = \boldsymbol{e}$. In our example Figure 1, $MD = 1$ would therefore not count as a cause of $FF = 1$, since setting $L = 0$ is not allowed.

2.2. LLMs and Causation

LLMs have impressive abilities in a wide range of applications, including their performance on various causal tasks. There are many papers, benchmarks, and datasets evaluating LLMs' causal inference abilities (Xiong et al., 2025; Jin et al., 2023; Frohberg and Binder, 2022; Chen et al., 2024; Wang, 2024; Zhou et al., 2024; Chevalley et al., 2023; Zečević et al., 2023; Suzgun et al., 2022). However, only few tests on LLMs' abilities to detect actual causes have been performed: Kıcıman et al. (2024) as mentioned above; Suzgun et al. (2022) and Chen et al. (2024), but they use non-standard examples and work with a different definition of actual causation, taking intentionality into consideration; finally Zečević et al. (2023), but they also do not contain standard vignettes. The evaluation on 15 vignettes by Kıcıman et al. (2024) seems to be the only evaluation of LLMs with regard to actual causation on standard vignettes to date.

3. The Dataset

In this section, we present the collection of vignettes and how it might be used by researchers.

3.1. Dataset construction

We collect vignettes from significant papers on actual causation (Halpern, 2015; Andreas and Günther, 2024; Hitchcock, 2001, 2007a; Hall, 2007; Halpern and Pearl, 2005; Weslake, 2015; Glymour et al., 2010; Beckers and Vennekens, 2018). These have been chosen according to their impact in the field, recency, and variety. As described in section 1, the ground truth is somewhat subjective, so for now, the current author's intuition was used. The ground truth can be modified by users to their own liking, as described in subsection 3.3.

3.2. Dataset description

In total, the dataset contains 58 vignettes, accompanied by 149 queries, at least one per vignette. It is constructed with the intention of being compatible with the SCM framework. It consists of three tables: *vignettes* contains one entry per vignette. *variables* contains an entry for each variable in each vignette and links to the respective vignette via a foreign key. These two tables are joined accordingly when the data is loaded. Their contents are explained in Table 1 and Table 2 respectively. *queries*, described in Table 3, contains a variable number of queries for each vignette, linking to the respective vignette via foreign key.

Table 1: Description of the *vignettes* table.

Column Name	Example	Explanation
v_id	ff_disj	A unique identifier for the vignette.
se_id	symm_od	A unique identifier for the structural equation model the vignette uses.
other_names	-	Alternative names of the vignette.
vignette_text	The forest burns down if [...]	A description of the vignette in natural language.
variable_order	MD,L,FF	A list of the variables in the order that they need to be updated in.
context	1,1	The setting of the exogenous variables (This is where some vignettes that use the same structural equations can differ).
title	Forest Fire disjunctive	The title of the vignette in natural language.
description	-	A further description of what is going on in the example, in addition to the main vignette_text.
origin	-	Where the vignette originally occurred.
taken_from	Halpern 2015, Ex. 3.1	Where the natural language description and structural equation modeling were taken from.
equivalent_to	assassin_badgirl	A list of v_id's that are equivalent to the vignette.
other_models	-	A list of v_id's that model the same vignette but with different structural equation models.
similar	Bullseye (Glymour 1)	A list of v_id's that model similar vignettes but are not quite equivalent.
notes	-	Space for other notes on the vignette.

Table 2: Description of the *variables* table.

Column Name	Example	Explanation
se_id	symm_od	A unique identifier for the structural equation model (multiple vignettes might utilize the same structural model)
variable_name	FF	The name of the variable.
var_description	Forest Fire	A short description of the variable.
range	0,1	The possible values the variable can take.
default_values	0	If applicable, the default value of the variable. Can be left empty.
deviant_values	1	If applicable, the deviant value of the variable. Can be left empty.
structural_equation	MD or L	The structural equation of the variable if it is an endogenous variable. Accepts both arithmetic ("A + B") and logical equations ("A and not B"). If the variable is exogenous, this field must be empty.

The format of the vignettes and queries is such that some extensions for theories of causation can be accommodated. We implement some of them. First, we can query for compound causes, where the potential cause consists of the conjunction of multiple primitive events. This allows us to accommodate different intuitions about cases of symmetric overdetermination: individualism and collectivism (cf. Schaffer (2003, p. 24)), a key difference between HP_{u} and $\mathrm{HP}_{\mathrm{mod}}$.

Second, we can also consider contrastive causation, a commonly suggested extension (Schaffer, 2005; Halpern and Pearl, 2005). For HP-type theories, a contrastive cause restricts the potential alternative values in the counterfactual analysis. For example, for $\mathrm{HP}_{\mathrm{mod}}$, in condition AC2, we

4. The definition in Halpern (2015) allows the Boolean combination of events $\boldsymbol{E} = \boldsymbol{e}$ here. We leave out this detail for simplicity since none of the vignettes actually use it.

Table 3: Description of the *queries* table.

Column Name	Example	Explanation
v_id	ff_disj	Foreign key to the *v_id* of the *vignettes* table.
cause	MD=1	The cause variable and its value, connected by "=". If the cause is a compound cause, this can be connected by "and".
cause_contrast	-	Optional: the contrastive values the cause variables should take, separated by a comma.
effect	FF=1	The effect variable and its value, connected by "=".
effect_contrast	-	Optional: the contrastive values the effect variable should take, separated by a comma.
query_text	Does the dropped match cause the forest fire?	The query in natural language.
intuition	0	The intuitive verdict of the query, 0 or 1 (for False or True)
intuition_source	-	Remarks about the intuition.
notes		Other remarks about the query in general.
[THEORY]	$[\in \{0, 1\}]$	One column like this exists for every included theory: behaves the same as column *intuition*.
[THEORY]_source	[citation]	One column like this exists for every included theory: behaves the same as column *intuition_source*.

could specify a specific setting $\boldsymbol{C}' = \boldsymbol{c}^*$ and $\boldsymbol{E}' = \boldsymbol{e}^*$ that has to hold, rather than finding one possible setting. In the case of binary variables, this is redundant.

Third, we add typical (default) values for each variable, which enables the implementation of normality for "small worlds" (Halpern and Hitchcock, 2015). "Small worlds" only take exogenous variables into account. Intuitively, a world w' is at least as normal as world w iff there are no exogenous variables in w' set to values that are less normal than the actual values in w. If we want to take normality into account, we apply this restriction in the search for suitable counterfactual worlds.

3.3. Usage notes

The dataset is intended primarily as a tool to evaluate theories of causation, but it can also serve as a tool for prompting LLMs. The table heads of the results of both of these can be viewed in section B.

3.3.1. EVALUATING ACTUAL CAUSATION

The most essential function to evaluate causation is `evaluate_cause()`. This function requires the arguments:

- *theory*: the theory one wants to evaluate the vignettes with, currently possible values: "HP2005", "HP2015".
- *vignette*: A Vignette object, as constructed from the provided data.
- *query*: A Query object, as constructed from the provided data.
- *gt*: the ground truth against which the evaluation results are to be compared (default "intuition", but other ones can be added manually)
- *normality*: A Boolean signifying whether normality considerations should be taken into account.

In addition, we added some wrappers that iteratively apply this function and return Pandas Dataframes for easy further processing. `evaluate_all_queries()` evaluates a certain theory on all queries. `reproduce_paper_results()` takes a list we provide of all the examples that were discussed in a specific paper. `evaluate_non_paper_queries()` does the inverse.

To use the dataset as a tool for benchmarking LLMs, we can use the natural language descriptions that were added for most vignettes. Via the function `run_llm_queries()`, vignettes are combined with the respective query into a natural language prompt. The LLM's answer is parsed and turned into a Boolean. This, therefore, behaves exactly like a theory of causation: for every Query Q in vignette V, it yields a verdict $\tau(V, Q) \in \{\top, \bot\}$. The number of queries evaluated is smaller than the total number of vignettes in the dataset, for two reasons: First, some vignettes are missing natural language descriptions and will be filtered out completely, and second, some vignettes have duplicate entries with different SCM formalizations. In these cases, we include only one of the best-performing formalizations.

3.3.2. MODIFICATION

We suggest that the dataset can be used and modified by users in the following ways:

Adding vignettes and queries. The collection of vignettes is comprehensive but not exhaustive. Researchers are welcome to add their own examples.

Adding intuitions about existing queries. Some researchers have contradicting intuitions about the ground truth for some queries. We allow researchers to set their own ground truth for the evaluation of queries. This is done by adding a column to the *queries* table and replacing the *gt* argument in any of the evaluation functions with the name of the new column.

Filter out certain vignettes. Some researchers disagree with the SCMs for particular scenarios (for example, there are at least three versions of *The Engineer*, cf. Halpern (2015)). Others claim that they don't have sound intuitions about some queries (Lewis, 1973). In other cases, one might only want to provide a partial theory of causation and exclude cases of isomorphism (Weslake, 2015). Such vignettes can be excluded from the evaluation.

4. Experiments

We first work through some of the results of vignettes on HP_{mod}, and then we compare the results to those of some LLMs.

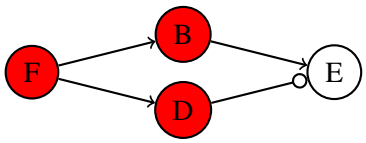
4.1. Evaluation of Theories of Causation

We implemented two popular theories of actual causation: HP_{u} (Halpern and Pearl, 2005) and HP_{mod} (Halpern, 2015). We demonstrate the usefulness of this new approach to evaluating theories of actual causation by going through some examples of the verdicts of HP_{mod} that were not evaluated in Halpern (2015). We pick some examples of results that contradict the intuitive verdict. Some vignettes model the same scenario with different SCMs, and we filter out these duplicates for the evaluation, which leaves 53 of 58 vignettes and 128 of 149 queries. Overall, HP_{mod} correctly evaluated 102 out of 128 queries, and 106 when taking normality considerations into account.

A short circuit is a certain structure where one event F causes two chains of events, one which by itself would cause another event E, but also another that prevents event E from happening. In such cases, the common intuition is that F is not a cause of $\neg E$. HP_{mod} deals with short circuits by

appealing to normality considerations (Halpern, 2015, p.8), since it falsely classifies short circuits as cases of causation. This is a basic version of the short circuit:

Example 2 (Boulder) *The boulder's dislodgement (F) threatens to hit the hiker by a rolling boulder (B), and at the same time provokes an action – the ducking (D) – that prevents this threat from being effective: The hiker is not hit (E). (Andreas and Günther, 2024, p. 18)*

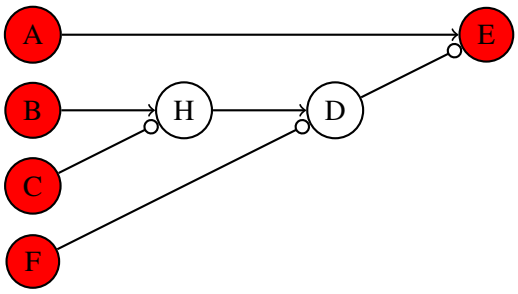


$$F := 1$$
$$B := F$$
$$D := F$$
$$E := B \wedge \neg D$$

Figure 2: Boulder

Does the boulder's dislodgement ($F = 1$) cause the hiker's survival ($E = 0$)? HP_{mod} says yes: if we choose $W = \{B\}$ and set $B = 1$ (its actual value), $E = 0$ counterfactually depends on $F = 1$. The likely solution will be normality considerations, analogous to the similar example of *Careful Poisoning* (Halpern, 2015, p. 8). *Careful Poisoning* is similar to *Boulder*, except in *Boulder* the intermediate variable B is added. One might analogously argue that the world where counterfactual dependence holds is the one where the boulder does not dislodge but yet threatens to hit the hiker, which is more abnormal than what actually happens. At this point, one might criticize that there is a need for guidelines to determine normality, but this is beyond the scope of this work. The vignettes *Extended Double Prevention*, *Modified Extended Double Prevention*, and *Safe Gun* all give false verdicts too, but all due to the same short circuit structure. By running the evaluation with normality considerations, we find that the "small worlds" implementation cannot fully resolve short-circuit cases.

Example 3 (Backup Threat Canceling) *E faces a threat from the firing of B. C cancels this threat. But F [. . .] would have done so, had C not occurred. (Hall, 2007, p. 128)*



$$A := 1, B := 1, C := 1, F := 1$$
$$H := B \wedge \neg C$$
$$D := H \wedge \neg F$$
$$E := A \wedge \neg D$$

Figure 3: Backup threat canceling

This is a vignette without a natural-language description. HP_{mod} misses that $C = 1$ causes $E = 1$ because it cannot set the non-actual value $F = 0$. HP_{u} gives the correct verdict. This counterexample was previously unnoticed.

Example 4 (Affecting) *Assassin puts poison in Victim's coffee ($A = 1$). The poison, if not counteracted, causes a painful death ($D = 1$). Bodyguard puts a weak antidote in Victim's coffee ($B = 1$). The antidote is not strong enough to neutralize the poison and hence save Victim's life, but at least it is strong enough to render Victim's death painless ($D = 2$). (Hitchcock, 2007a, p. 516)*

HP_{mod} identifies both $A = 1$ and $B = 1$ (individually) as a cause of $D = 2$. We think that a contrastive solution is favorable for both potential causes: $A = 1$ causes the death ($D = 2$ or $D = 1$; in HP_{mod} this is equivalent to $D = 2$ rather than $D = 0$), and $B = 1$ causes the painless death rather than a painful one ($D = 2$ rather than $D = 1$). Hitchcock (2007a) observes that there is at least an asymmetry here: the Bodyguard can only affect the outcome dependent on the Assassin's action. So, the symmetric treatment of $A = 1$ and $B = 1$ by HP_{mod} is unsatisfying. This vignette, to our knowledge, has also not been discussed for the HP_{mod} definition before.

Example 5 (Shock) *Two two-state switches are wired to an electrode. The switches are controlled by A and B, respectively, and the electrode is attached to C. A has the first option to flip her switch ($A = 1$). B has the second option to flip her switch ($B = 1$). The electrode is activated, and shocks C ($C = 1$) iff both switches are in the same position. B wants to shock C, and so flips her switch iff A does. (Weslake, 2015, p. 17)*

Here, HP_{mod} cannot distinguish between the cause $B = 1$ and the non-cause $A = 1$, since in both cases it allows freezing the other respective variable at its actual value. That is why the theory labels both of them as causes (individually). A likely reply by defenders of HP_{mod} could be that normality will come to the rescue once more. Unsurprisingly, since this is what the equations describe, a world where $A = 0$ and $B = 1$ could very reasonably be considered less normal than the actual world.

4.2. Evaluation of LLMs

To prompt the LLM, we create prompts with the natural language description along with the natural language query, as specified in Appendix section A. We also run experiments with different prompting techniques, few-shot prompting (Brown et al., 2020) and chain-of-thought (Wei et al., 2023). We use three small open-source models (LLaMA 3.2, Ministral-3, Gemma 3)[5], as well as GPT-5.4. The vignettes that do not have a description in natural language were excluded from the evaluation, as well as duplicates as mentioned above, resulting in only 112 queries on 47 vignettes compared to the original 149 queries on 58 vignettes.

4.3. Comparing Theories and LLMs

The format of our data allows us to compare the performance of the theories of actual causation with the LLM outputs very easily. We provide an overview of the performance of the implemented theories and the tested LLMs in Table 4 and Figure 4. As noted above, we restrict to the $n = 112$ natural-language subset for comparison. For reference, the F1-scores for the four theories on the full query set ($n = 128$, without LLM-incompatible queries filtered out) are 0.82, 0.83, 0.84, and 0.86 (in order of Table 4), indicating that the natural-language subset is slightly harder.

5. Downloaded from Ollama. Exact models: LLaMA 3.2 3B 128K (ID: a80c4f17acd5), Ministral-3 8B 256K (ID: 77300ee7514e), and Gemma 3 4B 128K (ID: a2af6cc3eb7f).

Table 4: Performance metrics across models. Only includes queries with an accompanying description in natural language ($n = 112$, after filtering out vignettes with missing textual description). Accuracy is shown with two-sided Wilson 95% confidence intervals. F1-score is reported with two-sided 95% bootstrap intervals with 2,000 resamples. p-values are from pairwise McNemar tests against HP_{mod} (Normality) with a pre-specified $\alpha = 0.05$.

Model	Accuracy	F1-Score	p-value vs. HP_{mod}+Norm.
HP_u	0.73 [0.64, 0.81]	0.79 [0.72, 0.86]	0.076
HP_u (Normality)	0.75 [0.66, 0.82]	0.80 [0.73, 0.87]	0.134
HP_{mod}	0.79 [0.71, 0.86]	0.82 [0.75, 0.89]	0.250
HP_{mod} (Normality)	**0.83** [0.75, 0.89]	**0.85** [0.77, 0.91]	–
GPT-5.4	0.70 [0.61, 0.77]	0.75 [0.66, 0.83]	0.052
GPT-5.4 (CoT)	0.76 [0.67, 0.83]	0.78 [0.69, 0.86]	0.541
GPT-5.4 (Few-shot)	0.70 [0.61, 0.77]	0.76 [0.68, 0.83]	0.064
Gemma 3	0.62 [0.52, 0.70]	0.72 [0.63, 0.79]	0.017
Gemma 3 (CoT)	0.62 [0.53, 0.71]	0.68 [0.58, 0.76]	0.009
Gemma 3 (Few-shot)	0.67 [0.58, 0.75]	0.76 [0.68, 0.83]	0.036
Llama 3.2	0.47 [0.38, 0.56]	0.27 [0.15, 0.40]	< 0.001
Llama 3.2 (CoT)	0.58 [0.49, 0.67]	0.60 [0.49, 0.70]	< 0.001
Llama 3.2 (Few-shot)	0.64 [0.55, 0.73]	0.69 [0.59, 0.78]	0.073
Ministral 3	0.57 [0.48, 0.66]	0.54 [0.41, 0.65]	< 0.001
Ministral 3 (CoT)	0.68 [0.59, 0.76]	0.71 [0.62, 0.80]	0.024
Ministral 3 (Few-shot)	0.54 [0.45, 0.63]	0.56 [0.44, 0.66]	< 0.001

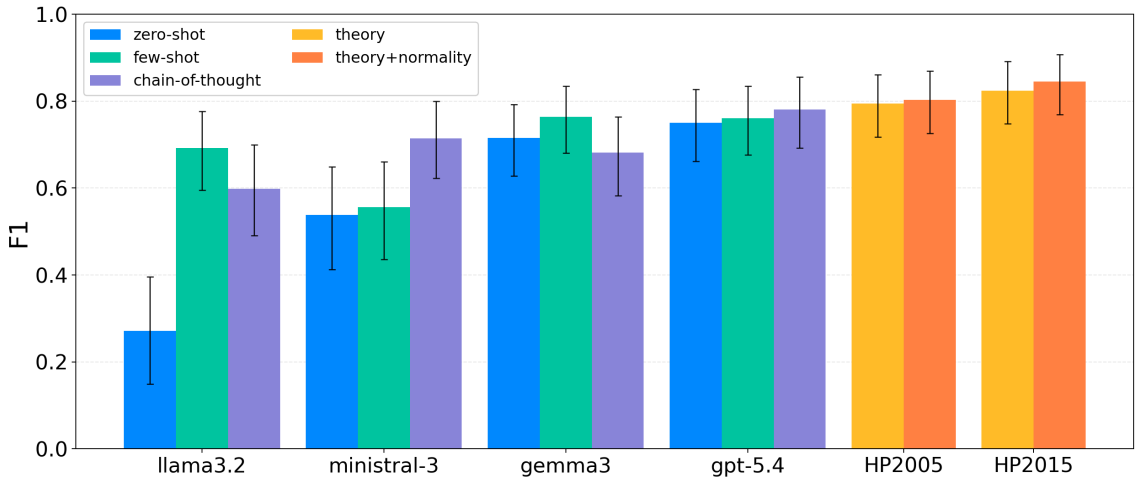


Figure 4: F1-scores on the dataset for different theories of actual causation and LLMs.

The best performance is achieved by the formal theories of actual causation, but the differences between the best-performing LLMs and HP_{mod} are not statistically significant, as can be seen by some of the p-values $> \alpha$ in Table 4 and the overlapping confidence intervals in Figure 4. We note that GPT-5.4 performs best out of the LLMs, and both few-shot and chain-of-thought prompting tend to improve the results over zero-shot prompting. However, we emphasize again that a metric cannot replace the evaluation of a theory completely, since getting just one very clear example wrong would disqualify it, even if good metrics are achieved.

5. Limitations and Further Work

The evaluated LLMs have likely seen many of these vignettes during training—inventing new ones might be a worthwhile task, following Suzgun et al. (2022) and Zečević et al. (2023). One test to measure the extent of memorization is the approach by Kıcıman et al. (2024, Attachment E).

For an unbiased ground truth, not just based on the intuition of one single philosopher, a survey on human subjects could be conducted. In this context, the current collection of vignettes might even be a useful tool for psychologists working on causal cognition.

The statistical analysis, in particular the confidence intervals and p-values, rests on the assumption that our collection of vignettes is a representative sample of all possible vignettes. While this is certainly the goal of this task, it is unlikely to be the case. Furthermore, many of the p-values in Table 4 are near the threshold and should be interpreted with caution, also keeping in mind that the evaluation of actual causation cannot be reduced to a purely statistical analysis.

The collection of vignettes is comprehensive but not complete, and more can be added over time. Compatibility with different formalisms, such as CP-logic or Situation Calculus, could be established. This would require more theoretical work to translate the formalisms. Lastly, we only implemented two theories of actual causation that are already very well established. Implementing more, and in particular newer, less-examined theories, might yield even more useful results.

6. Conclusion

We have shown that evaluating a theory of actual causation with our framework is both simple and powerful: a full evaluation runs in seconds and immediately highlights queries where the theory diverges from intuition. Using HP_{mod} as an example, we identified previously unnoticed counterexamples, including *Backup Threat Canceling*, *Affecting*, and *Shock*. Normality (implemented via "small worlds") provides a modest performance boost, but other formulations of normality may prove more effective. A precise, general formal treatment of normality therefore remains an important open problem. The LLM experiments confirm that none of the tested models reaches the performance of either formal theory, though there is no statistically significant gap between the best theories and the best LLMs. Above all, we hope to have demonstrated the value of encoding theories in executable code and evaluating them systematically on a shared collection of vignettes. We implemented two established theories and strongly encourage future work to accompany new theories with corresponding algorithms.

Acknowledgments

I thank Julian Bradfield, John Longley, Nicolas Navarre, Tadeg Quillien, and three anonymous reviewers for helpful feedback on the manuscript.

References

Holger Andreas and Mario Günther. A Regularity Theory of Causation. *Pacific Philosophical Quarterly*, 105(1):2–32, March 2024. ISSN 0279-0750, 1468-0114. doi: 10.1111/papq.12447. URL https://onlinelibrary.wiley.com/doi/10.1111/papq.12447.

Vitaliy Batusov and Mikhail Soutchanski. Situation Calculus Semantics for Actual Causality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11561. URL https://ojs.aaai.org/index.php/AAAI/article/view/11561. Number: 1.

Sander Beckers. Causal Explanations and XAI. January 2022. URL https://www.semanticscholar.org/paper/Causal-Explanations-and-XAI-Beckers/8106508c5cb2167214b01ef887cd8db9a31151b3.

Sander Beckers and Joost Vennekens. A general framework for defining and extending actual causation using CP-logic. *International Journal of Approximate Reasoning*, 77:105–126, October 2016. ISSN 0888-613X. doi: 10.1016/j.ijar.2016.05.008. URL https://www.sciencedirect.com/science/article/pii/S0888613X16300779.

Sander Beckers and Joost Vennekens. A Principled Approach to Defining Actual Causation. *Synthese*, 195(2):835–862, 2018. doi: 10.1007/s11229-016-1247-1. URL https://philarchive.org/rec/BECAPA-5.

Sander Beckers, Hana Chockler, and Joseph Y. Halpern. A Causal Analysis of Harm. *Minds and Machines*, 34(3):34, July 2024. ISSN 1572-8641. doi: 10.1007/s11023-024-09689-7. URL https://link.springer.com/10.1007/s11023-024-09689-7.

Thomas Blanchard and Jonathan Schaffer. Cause without Default. In Helen Beebee, Christopher Hitchcock, and Huw Price, editors, *Making a Difference: Essays on the Philosophy of Causation*, page 0. Oxford University Press, June 2017. ISBN 978-0-19-874691-1. doi: 10.1093/oso/9780198746911.003.0010. URL https://doi.org/10.1093/oso/9780198746911.003.0010.

Alexander Bochman. Actual Causality in a Logical Setting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 1730–1736, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-2-7. doi: 10.24963/ijcai.2018/239. URL https://www.ijcai.org/proceedings/2018/239.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. Causal Evaluation of Language Models, May 2024. URL http://arxiv.org/abs/2405.00622. arXiv:2405.00622 [cs].

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data, July 2023. URL http://arxiv.org/abs/2210.17283. arXiv:2210.17283 [cs].

Hana Chockler and Joseph Y. Halpern. Responsibility and blame: a structural-model approach, December 2003. URL http://arxiv.org/abs/cs/0312038. arXiv:cs/0312038.

Marc Denecker, Bart Bogaerts, and Joost Vennekens. Causal reasoning in a logic with possible causal process semantics. AAAI Press 2018, October 2018.

Martin Erwig and Eric Walkingshaw. Causal Reasoning with Neuron Diagrams. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 101–108, Leganes, Madrid, Spain, September 2010. IEEE. ISBN 978-1-4244-7621-3. doi: 10.1109/VLHCC.2010.23. URL http://ieeexplore.ieee.org/document/5635201/.

Jörg Frohberg and Frank Binder. CRASS: A Novel Data Set and Benchmark to Test Counterfactual Reasoning of Large Language Models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.229/.

J. Dmitri Gallow. A Model-Invariant Theory of Causation. *The Philosophical Review*, 130(1): 45–96, January 2021. ISSN 0031-8108, 1558-1470. doi: 10.1215/00318108-8699682. URL https://read.dukeupress.edu/the-philosophical-review/article/130/1/45/171763/A-Model-Invariant-Theory-of-Causation.

J. Dmitri Gallow. The Metaphysics of Causation. April 2022. URL https://plato.stanford.edu/Entries/causation-metaphysics/. Last Modified: 2022-04-14.

Alessandro Giordani. An Internal Limit of the Structural Analysis of Causation. *Axiomathes*, 26(4):429–450, December 2016. ISSN 1572-8390. doi: 10.1007/s10516-016-9297-z. URL https://doi.org/10.1007/s10516-016-9297-z.

Clark Glymour, David Danks, Bruce Glymour, Frederick Eberhardt, Joseph Ramsey, Richard Scheines, Peter Spirtes, Choh Man Teng, and Jiji Zhang. Actual causation: a stone soup essay. *Synthese*, 175(2):169–192, 2010. ISSN 0039-7857. URL https://www.jstor.org/stable/40801336.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00524.

N. Hall. Structural equations and causation. *Philosophical Studies*, 132(1):109–136, January 2007. ISSN 1573-0883. doi: 10.1007/s11098-006-9057-9. URL https://doi.org/10.1007/s11098-006-9057-9.

Ned Hall. Two Concepts of Causation. In John Collins, Ned Hall, and L. A. Paul, editors, *Causation and Counterfactuals*, pages 225–276. The MIT Press, June 2004. ISBN 978-0-262-27066-3. doi: 10.7551/mitpress/1752.003.0010. URL https://direct.mit.edu/books/book/2458/chapter/65393/Two-Concepts-of-Causation.

Joseph Y. Halpern. A Modification of the Halpern-Pearl Definition of Causality, May 2015. URL http://arxiv.org/abs/1505.00162. arXiv:1505.00162 [cs].

Joseph Y. Halpern. *Actual Causality*. The MIT Press, August 2016. ISBN 978-0-262-33661-1. doi: 10.7551/mitpress/10809.001.0001. URL https://direct.mit.edu/books/oa-monograph/3451/Actual-Causality.

Joseph Y. Halpern and Christopher Hitchcock. Graded Causation and Defaults. *The British Journal for the Philosophy of Science*, 66(2):413–457, June 2015. ISSN 0007-0882. doi: 10.1093/bjps/axt050. URL https://www.journals.uchicago.edu/doi/full/10.1093/bjps/axt050.

Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, December 2005. ISSN 0007-0882, 1464-3537. doi: 10.1093/bjps/axi147. URL https://www.journals.uchicago.edu/doi/10.1093/bjps/axi147.

Christopher Hitchcock. The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy*, 98(6):273–299, 2001. ISSN 0022-362X. doi: 10.2307/2678432. URL https://www.jstor.org/stable/2678432.

Christopher Hitchcock. Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review*, 116(4):495–532, 2007a. doi: 10.1215/00318108-2007-012.

Christopher Hitchcock. What's Wrong with Neuron Diagrams? In Joseph Keim Campbell, Michael O'Rourke, and Harry S. Silverstein, editors, *Causation and Explanation*, pages 69–92. The MIT Press, August 2007b. ISBN 978-0-262-26976-6. doi: 10.7551/mitpress/1753.003.0006. URL https://direct.mit.edu/books/book/2434/chapter/64572/What-s-Wrong-with-Neuron-Diagrams.

Christopher Hitchcock and Joshua Knobe. Cause and Norm. *The Journal of Philosophy*, 106(11):587–612, 2009. ISSN 0022-362X. URL https://www.jstor.org/stable/20620209.

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. CLadder: Assessing Causal Reasoning in Language Models. November 2023. URL https://openreview.net/forum?id=e2wtjx0Yqu.

Konstantin Raphael Kueffner. *A comprehensive survey of the actual causality literature*. PhD thesis, TU Wien, 2021. URL https://scholar.archive.org/work/

nu5jjnwhjnh3jmkfe32bwzlhn4/access/wayback/https://repositum.tuwien.at/bitstream/20.500.12708/18862/1/Kueffner%20Konstantin%20Raphael%20-%202021%20-%20A%20comprehensive%20Survey%20of%20the%20Actual...pdf. Master's thesis.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality, August 2024. URL http://arxiv.org/abs/2305.00050. arXiv:2305.00050 [cs].

David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1973. ISSN 0022-362X. doi: 10.2307/2025310. URL https://www.jstor.org/stable/2025310.

David Lewis. Postscripts to 'Causation'. In David Lewis, editor, *Philosophical Papers Volume II*, pages 173–213. Oxford University Press, 1987.

Peter Menzies. The Problem of Counterfactual Isomorphs. In Helen Beebee, Christopher Hitchcock, and Huw Price, editors, *Making a Difference: Essays on the Philosophy of Causation*, page 0. Oxford University Press, June 2017. ISBN 978-0-19-874691-1. doi: 10.1093/oso/9780198746911.003.0009. URL https://doi.org/10.1093/oso/9780198746911.003.0009.

Michael S. Moore. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford University Press, January 2009. ISBN 978-0-19-171965-3. doi: 10.1093/acprof:oso/9780199256860.001.0001. URL https://academic.oup.com/book/6818.

L. A. Paul and Ned Hall. *Causation: A User's Guide*. Oxford University Press, Oxford, New York, May 2013. ISBN 978-0-19-967345-2.

Judea Pearl. *Causality: Models, reasoning, and inference*. Causality: Models, reasoning, and inference. Cambridge University Press, New York, NY, US, 2000. ISBN 978-0-521-77362-1. Pages: xvi, 384.

Jonathan Schaffer. Overdetermining Causes. *Philosophical Studies*, 114(1-2):23–45, 2003. doi: 10.1023/a:1024457117218.

Jonathan Schaffer. Contrastive Causation. *The Philosophical Review*, 114(3):327–358, July 2005. ISSN 0031-8108, 1558-1470. doi: 10.1215/00318108-114-3-327. URL https://read.dukeupress.edu/the-philosophical-review/article/114/3/327/2670/Contrastive-Causation.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them, October 2022. URL http://arxiv.org/abs/2210.09261. arXiv:2210.09261 [cs].

Zeyu Wang. CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of Large Language Models. October 2024. URL https://openreview.net/forum?id=kbmGbm2L1P.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

Brad Weslake. A Partial Theory of Actual Causation. *British Journal for the Philosophy of Science*, 2015.

Kai Xiong, Xiao Ding, Yixin Cao, Yuxiong Yan, Li Du, Yufei Zhang, Jinglong Gao, Jiaqian Liu, Bing Qin, and Ting Liu. Comˆ2 : A Causal-Guided Benchmark for Exploring Complex Commonsense Reasoning in Large Language Models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16119–16140, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.785. URL https://aclanthology.org/2025.acl-long.785/.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal, August 2023. URL http://arxiv.org/abs/2308.13067. arXiv:2308.13067 [cs].

Liang Zhou. Cost and Actual Causation, July 2017. URL http://arxiv.org/abs/1707.09704. arXiv:1707.09704 [cs].

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. CausalBench: A Comprehensive Benchmark for Causal Learning Capability of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.06349. arXiv:2404.06349 [cs].

Appendix A. LLM Prompts

Zero-shot prompt:

> You are an expert in causal reasoning. Your task is to determine whether one event is an actual cause of another based on the given vignette.
>
> Output format: Return exactly one token: YES or NO. Do not output anything else.
>
> Scenario: VIGNETTE
>
> Question: QUERY

Few-shot prompt:

> You are an expert in causal reasoning. Your task is to determine whether one event is an actual cause of another based on the given vignette.
>
> Answer with ONLY “Yes” or “No”. Your answer must be exactly one word. Do not explain your reasoning.
>
> Here are some examples:

Vignette: A barometer drops and then a storm occurs. The barometer reading is a signal of pressure, not a mechanism that produces storms. Query: Is the barometer drop a cause of the storm? Answer: No

Vignette: A match is struck near dry wood with oxygen present, and the wood ignites. Query: Is striking the match a cause of the wood igniting? Answer: Yes

Vignette: Alice was driving her car on a clear road. Suddenly, a dog ran into the street. Alice swerved to avoid the dog and hit a parked car. Query: Is the dog a cause of Alice hitting the parked car? Answer: Yes

Vignette: Mark was late to work because of heavy traffic. His boss was already angry because of a previous meeting. Query: Is the heavy traffic a cause of Mark's boss being angry? Answer: No

Now analyze the following:

Vignette: VIGNETTE Question: QUERY Answer:

Chain-of-Thought prompt:

You are an expert in causal reasoning. Your task is to determine whether one event is an actual cause of another based on the given vignette.

Scenario: VIGNETTE

Question: QUERY

Think through the facts step by step.

Important output rule: After your reasoning, output exactly one final line: Final: YES or Final: NO Use uppercase YES/NO and no extra text on that final line.

Appendix B. Example Outputs

Table 5: Example data from evaluation results of HP_{mod}.

v_id	query_id	cause	effect	effect_ contrast	theory	result	witness	gt_label	groundtruth	details	agreement	TP	TN	FP	FN
ff_disj	ff_disj_q0	MD=1	FF=1		HP2015	FALSE		intuition	FALSE		TRUE	0	1	0	0
ff_disj	ff_disj_q1	L=1	FF=1		HP2015	FALSE		intuition	FALSE		TRUE	0	1	0	0
ff_disj	ff_disj_q2	MD=1 and L=1	FF=1		HP2015	TRUE	Witness: W=[], w=[], x'=[0, 0]	intuition	TRUE		TRUE	1	0	0	0

Table 6: Example data from evaluation results with Llama 3.2.

v_id	query_id	cause	effect	effect_ contrast	theory	result	witness	gt_label	groundtruth	details	agreement	TP	TN	FP	FN
ff_disj	ff_disj_q0	MD=1	FF=1		llama3.2	TRUE		intuition	FALSE	LLM response: Yes	FALSE	0	0	1	0
ff_disj	ff_disj_q1	L=1	FF=1		llama3.2	TRUE		intuition	FALSE	LLM response: Yes	FALSE	0	0	1	0
ff_disj	ff_disj_q2	MD=1 and L=1	FF=1		llama3.2	TRUE		intuition	TRUE	LLM response: Yes	TRUE	1	0	0	0