Synergizing Large Language Models and Theory for Human-like Causal Reasoning

Yanxi Zhang[1,2], Xin Cong[3], Zhong Zhang[3] ✉, Xiao Liu[2], Dongyan Zhao[2,4,5] ✉, Yesai Wu[3]
[1]Center for Data Science, AAIS, Peking University
[2]Wangxuan Institute of Computer Technology, Peking University
[3]Tsinghua University [4]State Key Laboratory of General Artificial Intelligence
[5]Key Laboratory of Intelligent Press Media Technology
zhangyx@stu.pku.edu.cn, zhongzhang@tsinghua.edu.cn, zhaodongyan@pku.edu.cn
https://github.com/zhangyx0417/hcr_reasoner

Abstract

Performing causal reasoning akin to humans is fundamental to achieving strong artificial intelligence. However, existing studies on causal reasoning and large language models (LLMs) often underestimate the role of *human causal cognition*. Human causal reasoning is far more than recognizing causes and effects. Humans often factor in many other considerations when establishing a causal link between two events, such as norm violations, intentionality, morality, counterfactual scenarios, etc. To simulate such reasoning, it is intuitive to identify cause-effect pairs using *token causality* formalisms while modeling the factors that influence human causal judgment using findings from *cognitive science* experiments. To support the evaluation and modeling of human-like causal reasoning in LLMs, we introduce HCR-BENCH, a challenging benchmark with 1,093 instances and detailed human causal reasoning step annotations, and HCR-REASONER, an approach that systematically integrates theories of token causality and cognitive science into LLMs for human-like causal reasoning. Results suggest that HCR-REASONER consistently improves LLMs' causal alignment with humans, and that theory-guided reasoning is highly effective.

1 Introduction

Against the backdrop of the rapid development of LLMs, causal reasoning is increasingly regarded as a core capability for achieving general artificial intelligence, essential for prediction and decision-making. However, although current models have demonstrated strong performance on causal benchmarks (Roemmele et al., 2011; Du et al., 2022; Ashwani et al., 2024), they continue to face challenges in achieving *cognitive alignment* with humans. Research has shown that models often rely heavily on shallow causal knowledge encoded in their parameters at inference time (*level-1* reasoning), rather than developing a deep understanding of causal mechanisms (Zečević et al., 2023). Chi et al. (2024) have found that there is a fundamental discrepancy between LLMs' reasoning processes and human causal cognition (*level-2* reasoning).

In everyday scenarios, humans more frequently deal with *specific* causal relations between events, e.g., "Did David's 30 years of smoking habit cause him to get lung cancer last year?" To answer this, they typically draw on *general* causal knowledge, e.g., smoking causes cancer, by understanding the causal chain of events. However, such knowledge is insufficient to determine the final judgment, since humans are also influenced by cognitive modulatory factors such as morality, normality, and intent while making judgments in the concrete context (Lagnado et al., 2013; Sloman and Lagnado, 2015). Humans' causal judgment process corresponds to two complementary research areas: 1) *token causality*, which aims to formalize and axiomatize the intuitive form of causal reasoning used by humans (Halpern, 2016); and 2) *causal judgment*, a branch of cognitive science that investigates how cognitive modulatory factors influence humans' judgment of causes (Sloman and Lagnado, 2015).

However, existing text-based causal reasoning benchmarks largely focus on *general* causal mechanisms and are insufficient for evaluating the *specific* causal relations that humans typically reason about (Jin et al., 2023, 2024; Liu et al., 2024; Roemmele et al., 2011; Du et al., 2022; Bondarenko et al., 2022). Even among the few benchmarks that address human causal reasoning, their labels are not sufficiently validated by humans (Chi et al., 2024), or they lack complete annotations of human reasoning processes (Srivastava et al., 2023; Suzgun et al., 2023). Such a lack of consensus and transparency makes it difficult to determine whether model reasoning steps are genuinely aligned with human cognition. Furthermore, existing LLM-based causal reasoning methods have not systematically integrated token causality with cognitive science, and

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3513–3544
November 6-10, 2026 ©2026 Association for Computational Linguistics

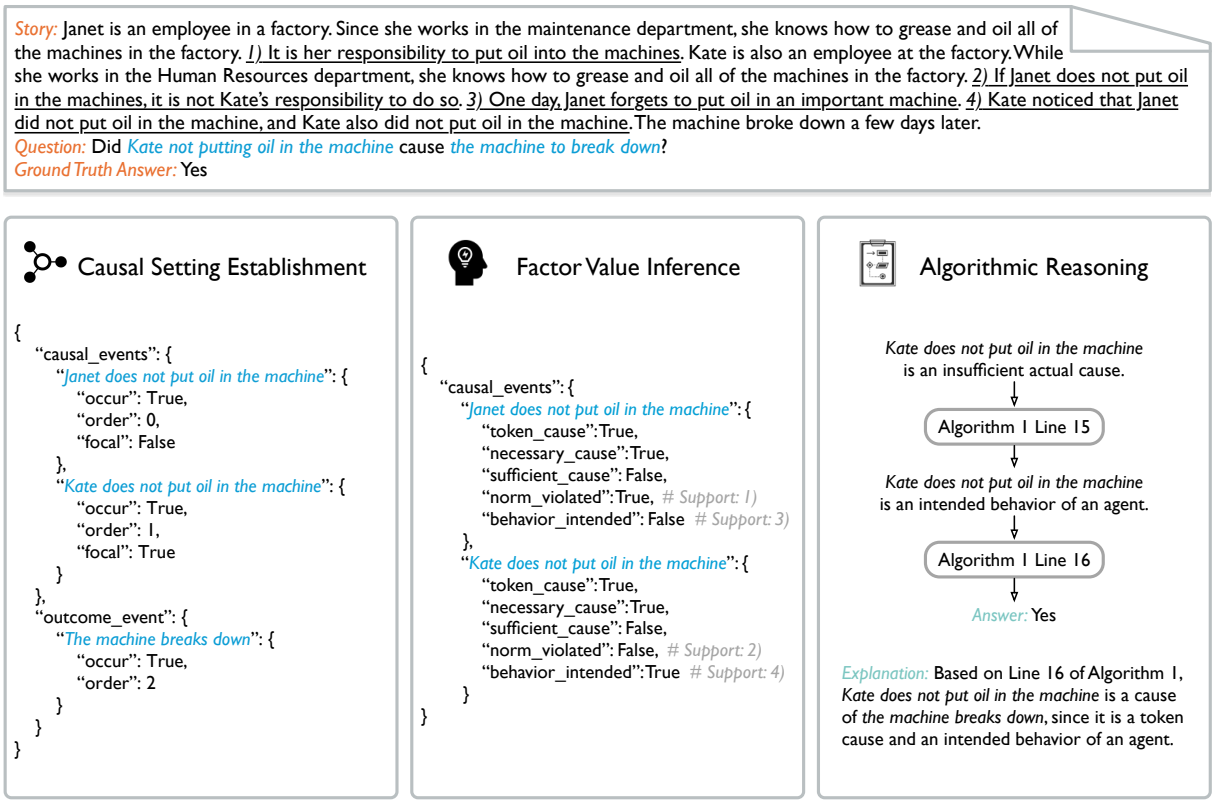


Figure 1: An example in HCR-BENCH. To simulate human causal reasoning, one needs to consider both causal and cognitive factors.

most continue to focus on reasoning about *general* causal mechanisms (Jin et al., 2023; Jiang et al., 2024; Lu et al., 2023). Although Chi et al. (2024) have begun to consider *specific* causal relations, their methodological designs remain largely at the level of *general* reasoning.

Therefore, we introduce HCR-BENCH, a novel human-like reasoning benchmark containing 1,093 carefully annotated instances with detailed human reasoning step annotations, enabling fine-grained evaluation. HCR-BENCH is built upon the causal judgment subset of Big-Bench Hard (Suzgun et al., 2023), which provides human consensus labels derived from rigorous human experiments but is too small and lacks human reasoning step annotations. To support human-like reasoning in LLMs, we further introduce HCR-REASONER, a novel framework that integrates token causality and cognitive science for human-like causal reasoning. It identifies candidate causes and outcomes, infers the values of token causality and causal judgment factors for candidate causes, and utilizes these inferred factor values to algorithmically derive the final causal judgment and provide an explanation.

Our contributions are threefold: 1) We introduce HCR-BENCH, the first human-like causal reasoning benchmark with detailed human reasoning step annotations. 2) We propose HCR-REASONER, the first method that integrates LLMs with token causality and cognitive science theories for human-like causal reasoning. 3) Results suggest that HCR-BENCH is challenging, and that HCR-REASONER consistently improves performance via the integration of theories into LLMs.

2 Preliminaries

2.1 Type Causality vs. Token Causality

Type causality focuses on inferring causal relations between variables, such as causal effect estimation (Peters et al., 2017). In contrast, *token causality* concerns the extent to which particular events bring about other specific events (Halpern, 2016; Hausman, 2005). For example, the question "Does smoking cause lung cancer?" pertains to type causality. Researchers have traditionally focused more on type causality questions, as they help develop theories and make predictions. In contrast, the question "Was Fred's smoking habit responsible for his lung cancer?" pertains to token causality. Its conclusion does not need to generalize to other settings, but is crucial for specific situations.

2.2 Token Causality

Token causality aims to formalize and axiomatize the intuitive form of causal reasoning used by humans (Halpern, 2016). The Halpern-Pearl (HP) definition of causality is one of the formal definitions of a "token cause" in *specific* causal scenarios (Halpern, 2016). Here, we first introduce several key concepts and then present the HP definition.

Causal Model. A *causal model* M is defined as a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ is a *signature* consisting of a set of *exogenous variables* $\mathcal{U}$, a set of *endogenous variables* $\mathcal{V}$, and the *ranges* of endogenous variables. $\mathcal{F}$ contains the *structural equations* that determine the values of endogenous variables. Each variable $V \in \mathcal{V}$ takes values from $\mathcal{R}(V)$. $\mathcal{R}(V) = \{0, 1\}$ for binary variables.

Primitive Event. A *primitive event* have the form $X = x$ for an endogenous variable X and some possible value x of X.

Causal Formula. A *causal formula* ψ is written as $[\boldsymbol{Y} \leftarrow \boldsymbol{y}]\varphi$ ($\boldsymbol{Y} \subset \mathcal{V}$), meaning that φ would hold if an intervention set the variables in $\boldsymbol{Y}$ to $\boldsymbol{y}$.

Causal Setting. A *causal setting* is a pair $(M, \boldsymbol{u})$, where M is the causal model and $\boldsymbol{u}$ is the *context* (i.e., an assignment to $\mathcal{U}$). If a formula ψ holds in the causal setting $(M, \boldsymbol{u})$, we write $(M, \boldsymbol{u}) \models \psi$.

Definition 1 (Token Cause)**.** *$\boldsymbol{X} = \boldsymbol{x}$ is a token cause of φ in the causal setting $(M, \boldsymbol{u})$ if the following three conditions hold:*

AC1. $(M, \boldsymbol{u}) \models (\boldsymbol{X} = \boldsymbol{x})$ *and* $(M, \boldsymbol{u}) \models \varphi$.

AC2. There exists a set $\boldsymbol{W}$ of variables in $\mathcal{V}$ and a setting $\boldsymbol{x}'$ of variables in $\boldsymbol{X}$ such that if $(M, \boldsymbol{u}) \models (\boldsymbol{W} = \boldsymbol{w}^*)$, then $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \boldsymbol{w}^*]\neg\varphi$.

AC3. $\boldsymbol{X}$ is minimal; there is no strict subset $\boldsymbol{X}'$ of $\boldsymbol{X}$ such that $\boldsymbol{X}' = \boldsymbol{x}'$ satisfies conditions AC1 and AC2, where $\boldsymbol{x}'$ is the restriction of $\boldsymbol{x}$ to the variables in $\boldsymbol{X}$.

In Def. 1, token causes must be conjunctions of primitive events (i.e., $\boldsymbol{X} = \boldsymbol{x}$), whereas effects can be any boolean combination of primitive events. The core clause is AC2, which establishes the counterfactual dependence of the effect φ on a candidate cause $\boldsymbol{X} = \boldsymbol{x}$ by fixing certain variables (in $\boldsymbol{W}$) at their actual values. This definition extends the standard *but-for test* (Def. 2), allowing it to handle complex real-world scenarios such as *preemption* (see §A.3).

2.3 Causal Judgment

Causal judgment investigates how cognitive modulatory factors influence humans' judgment of causes (Sloman and Lagnado, 2015). Nie et al. (2023) have studied and summarized these factors.

Causal Structure. Causal structures can be *conjunctive* or *disjunctive*. A conjunctive causal structure implies that the outcome occurs only if all contributing events jointly occur, while a disjunctive structure indicates that any single event is sufficient to bring about the outcome. This factor plays a critical role in causal judgment (Wells and Gavanski, 1989; Mandel, 2003; Sloman and Lagnado, 2015).

Normality. Norms can be *descriptive* or *prescriptive* (Halpern, 2016). Descriptive norms are statistical typicals, such as the mode, mean, or values close to them. Prescriptive norms are moral standards, laws, institutional policies, or even proper functioning in machines or organisms (Halpern, 2016). Humans tend to ascribe more causality to abnormal events than to normal ones (Knobe and Fraser, 2008; Hitchcock and Knobe, 2009).

Intentionality. The intent of an agent is considered part of their epistemic state, which typically concerns the agent's awareness when performing specific actions. An agent who is aware of the potential consequences of their action and deliberately performs an action that leads to a foreseeable bad outcome is judged more harshly than one who lacks the relevant knowledge (Samland et al., 2016; Kominsky and Phillips, 2019).

Action or Omission. People tend to identify actions, rather than omissions, as causes. This phenomenon is known as "omission bias" (Ritov and Baron, 1992; Baron and Ritov, 2004). In scenarios whether one agent acts while another does not, the one who acts is more likely to be judged as a cause (Henne et al., 2017; Clarke et al., 2015; DeScioli et al., 2011; Gerstenberg and Stephan, 2021).

Temporal Effect. Causal judgment is influenced by the temporal order of events (Gerstenberg and Lagnado, 2012). When multiple events unfold over time and lead to an outcome, people tend to identify later events as the cause (Reuter et al., 2014). However, this preference depends on how the events causally relate to one another (Hilton et al., 2010; Spellman, 1997). When earlier events determine the course of subsequent actions, they are more likely to be selected as causes (Henne et al., 2021).

3 Related Work

Commonsense Causal Reasoning. Early causal reasoning studies in the natural language processing domain mainly focused on *commonsense causal knowledge* (Roemmele et al., 2011; Sap et al., 2019; Zheng et al., 2023; Zhang et al., 2022). However, they often treat language models as "knowledge bases", resulting in shallow causal reasoning (Jiang et al., 2020; Petroni et al., 2019).

Formal Causal Reasoning. Recent studies have begun to focus on *formal causal reasoning*, requiring systems to make decisions based on formal causal frameworks such as structural causal models (Jin et al., 2023, 2024; Liu et al., 2024). For example, CLadder (Jin et al., 2023) evaluates whether LLMs can perform causal inference according to a set of well-defined causal rules; Corr2Cause (Jin et al., 2024) requires LLMs to determine causal relations between variables based on a set of correlational statements; and QRData (Liu et al., 2024) evaluates LLMs' causal reasoning abilities when processing real-world data. Also, CausalCoT (Jin et al., 2023) provides a chain-of-thought (CoT) baseline for step-by-step formal causal reasoning. However, these studies overlook non-formal causal judgment factors that influence humans' judgment of causes (Sloman and Lagnado, 2015).

Human-like Causal Reasoning. No prior work has systematically combined token causality and cognitive science for human-like causal reasoning. On the data side, type causality benchmarks struggle to model human judgments of complex causal links between events in concrete situations (Jin et al., 2023, 2024; Liu et al., 2024; Roemmele et al., 2011; Du et al., 2022; Bondarenko et al., 2022). Recently, Chi et al. (2024) proposed the task of *human-like causal reasoning* and introduced the CausalProbe-2024 benchmark. However, its labels are fully generated by LLMs, and it does not provide human reasoning step annotations. Other studies leverage findings from human experiments to derive human consensus labels (Srivastava et al., 2023; Suzgun et al., 2023; Nie et al., 2023), but they do not provide complete reasoning step annotations. On the methodological side, Chi et al. (2024) proposed G^2-Reasoner, which incorporates general knowledge and goal-oriented prompts into LLMs' causal reasoning processes. However, this method still focuses on type causality and does not integrate human causal cognition.

4 The HCR-BENCH Benchmark

We construct HCR-BENCH to address the lack of consensus and transparency in existing benchmarks by reusing human consensus labels validated through rigorous human experiments and providing fine-grained annotations of human causal reasoning processes. As shown in Fig. 1, HCR-BENCH is defined as $\mathcal{D} := \{t_i, q_i, a_i, r_i\}_{i=1}^{N}$, where each sample comprises a causal story t_i, a causal query q_i, a binary answer $a_i \in \{\text{Yes}, \text{No}\}$, and the reasoning process r_i. The goal of HCR-BENCH is to evaluate whether an LLM can correctly map q to a and produce a correct reasoning process r.

Table 1: Comparison of causal reasoning benchmarks. Type: type vs. token causality? Valid. Labels: labels human-validated? Caus.: causal structures annotated? Cogn.: cognitive factors annotated?

Dataset	Type	Valid. Labels	Caus.	Cogn.
COPA (Roemmele et al., 2011)	Type	✓	✗	✗
e-CARE (Du et al., 2022)	Type	✓	✗	✗
CLadder (Jin et al., 2023)	Type	✗	✓	✗
Corr2Cause (Jin et al., 2024)	Type	✗	✓	✗
QRData (Liu et al., 2024)	Type	✓	✗	✗
Big-Bench (Srivastava et al., 2023)	Token	✓	✗	✗
Big-Bench Hard (Suzgun et al., 2023)	Token	✓	✗	✗
MoCa (Nie et al., 2023)	Token	✓	✗	✓
CausalProbe-2024 (Chi et al., 2024)	Token	✗	✗	✗
HCR-BENCH (Ours)	Token	✓	✓	✓

4.1 Dataset Construction

This section introduces the construction process of HCR-BENCH, including data cleaning, data annotation, data augmentation, and quality control.

Data Source. The causal judgment task of Big-Bench (Srivastava et al., 2023) tests whether LLMs can read a short story with multiple cause-effect events and answer causal queries in the same manner as humans would. It contains 190 stories from 30 papers, each of which conducted rigorous human experiments to derive human consensus labels. Big-Bench Hard (Suzgun et al., 2023) selects 23 challenging tasks from Big-Bench, with causal judgment being one of them. Its causal judgment subset contains 187 stories, including 141 *causation* queries (e.g., "Did the red wire cause the machine to short circuit?") and 46 *intention* queries (e.g., "Did the man shoot the dear intentionally?"). HCR-BENCH derives from this subset.

Data Cleaning. Before annotation, we manually remove 54 samples that are duplicate, lack critical background information, or focus on intention, and

correct 2 partially flawed samples. This process reduces noise and preserves only samples of interest, resulting in a curated set of 133 samples.

Data Annotation. We manually annotate the reasoning steps for the 133 samples. Each annotation is a unified JSON combining the two JSONs shown in Fig. 1. Specifically, for all events, we annotate the occurrences f_{occur} and temporal orders f_{order}, with f_{order} being an integer-valued factor indicating an event's relative sequential position starting from 0. These factors model action or omission and temporal effects. For causal events, we additionally annotate three types of binary factors: 1) *auxiliary factor*: f_{focal}, which records whether the event is queried; 2) *causal factors*: token cause f_{tc} (Def. 1), necessary cause f_{nc} (Def. 2), and sufficient cause f_{sc} (Def. 3), which represents whether the event is a token cause, necessary cause, or sufficient cause of the outcome; and 3) *cognitive factors*: normality f_{n} and intentionality f_{i}, where f_{n} determines whether the event involves norm violations, and f_{i} determines whether the event is an agent's action, whether the agent is aware of its potential outcomes, and whether the agent knowingly performs an action that leads to a foreseeable bad outcome.

Data Augmentation. We expand the dataset with a two-stage augmentation process to reach a reasonable evaluation scale. First, we manually create 30 new samples from existing stories by altering the query to focus on causal events that are not queried, i.e., "Did X' cause Y?" In such cases, only f_{focal} and a need to be updated. Second, we employ GPT-4o (OpenAI et al., 2024a) to synthesize more samples by rewriting the stories of the existing 163 samples (*seed samples*). This is a three-stage process. Based on a seed sample, it first rewrites the story into a completely new real-world scenario while preserving the reasoning logic. That is, only the descriptions of events change, while their factor values remain unchanged. Then, based on the new events, it generates a new query. The gold answer remains unchanged. Finally, it adds or removes details from the new story and reorganizes its structure to make it more distinctive while preserving the reasoning logic. We generate 6 samples for 155 seed samples, obtaining 930 generated samples. Together with the 163 seed samples, the dataset has 1,093 samples. Generated samples are generally more challenging since they contain more spurious correlations and fewer explicit causal cues (§D.2), e.g., "When E_1 and E_2 occur, O will occur." Spurious correlations are common in real-world scenarios, and humans often identify causal relations without explicitly stating such cues (Matute et al., 2015; Vasilyeva and Lombrozo, 2015).

Quality Control. We apply automated and manual methods to verify the quality of HCR-BENCH. For seed samples, we manually review each sample and iteratively refine our annotations before generation, based on the metadata and annotations provided by Suzgun et al. (2023) and Nie et al. (2023), as well as LLM-generated pseudo-reasoning steps. For generated samples, we implement code to automatically check whether the reasoning logic of each generated sample is consistent with that of its seed sample. Fewer than 1% of samples are found to be inconsistent (mostly involving incorrect values for f_{i}), and these are manually corrected. We also verify whether the reasoning logic of each generated sample corresponds to the correct gold answer using Alg. 1; all samples pass this automatic check. In addition, we randomly sample 310 generated samples and evaluate whether 1) the story contains sufficient information to answer the query, 2) the reasoning logic is consistent with the story, and 3) the gold answer is correct. Only 2 samples (0.6%) lack essential reasoning information; we regenerate and verify them.

4.2 Dataset Statistics

HCR-BENCH contains 1,093 samples, with a relatively balanced proportion of positive and negative samples. Most queries concern a single causal event, while a small number involve multiple causal events.

Table 2: HCR-BENCH statistics.

Category	Statistic	Total
SIZE	# Samples	1,093
STORY	# Sentences (per sample)	8.49
	# Words (per sample)	162.70
QUERY	# Queried causal events (per sample)	1.02
REASONING	# Events (per sample)	3.39
	# Causal events (per sample)	2.39
	# Outcome events (per sample)	1
ANSWER	Positive class	53.8%
	Negative class	46.2%

5 The HCR-REASONER Framework

We propose HCR-REASONER to integrate token causality and cognitive science theories into LLMs for step-by-step, explainable human-like causal rea-

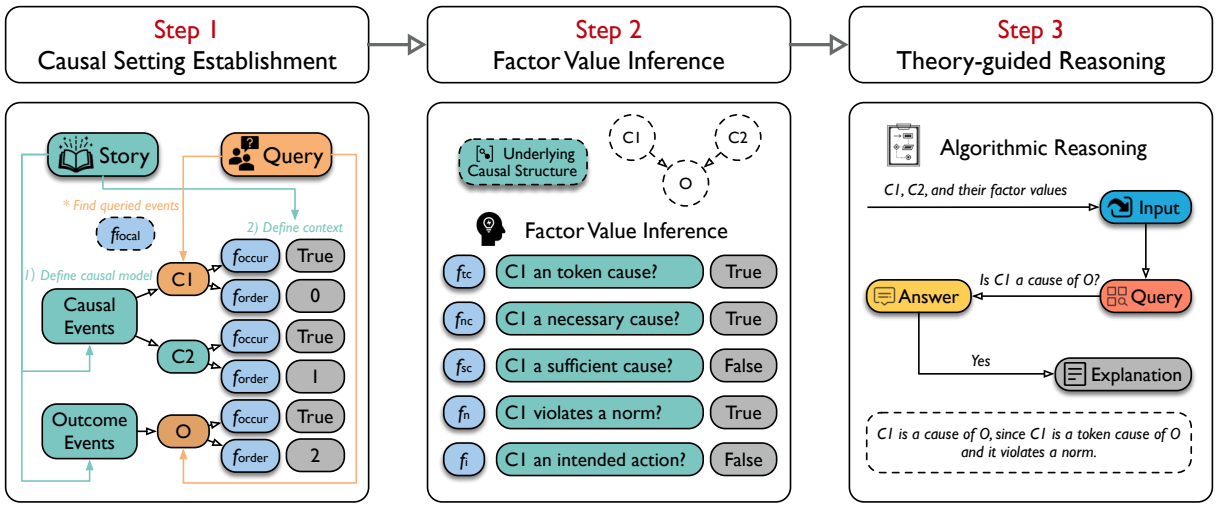


Figure 2: The overview of HCR-REASONER.

soning. As shown in Fig. 2, it operates in three stages, ending with an answer and explanation.

5.1 Causal Setting Establishment

HCR-REASONER first sets up the causal setting, including the causal model M and the context $\boldsymbol{u}$.

Causal Model. In HCR-REASONER, $\mathcal{V}$ is partitioned into a causal event set $\mathcal{C}$ and an outcome event set $\mathcal{O}$, with $\mathcal{R}(V) = \{0, 1\}$. To set up the causal model M, HCR-REASONER extracts the events from the story t and query q, and combines the outcome events into a unified outcome event O[1]. It also records f_{focal} for each causal event.

Context. The context $\boldsymbol{u}$ represents the values of exogenous variables, which in turn determine the values of endogenous variables via the structural equations $\mathcal{F}$. In practice, HCR-REASONER extracts the values of endogenous variables, i.e., the values of f_{occur}, as a proxy. It also records the temporal order f_{order} for each event. The stage of causal setting establishment is formalized as:

$$\langle \mathcal{C}, O, f_{\text{focal}}, f_{\text{occur}}, f_{\text{order}} \rangle \leftarrow \Phi(t, q, I_{\text{m}}, I_{\text{c}}),$$

where Φ is an LLM. I_{m} and I_{c} are instructions for Φ to extract M and $\boldsymbol{u}$, respectively.

5.2 Factor Value Inference

HCR-REASONER then infers the values of causal and cognitive factors for each causal event. The stage of factor value inference is formalized as:

$$\langle f_{\text{tc}}, f_{\text{nc}}, f_{\text{sc}}, f_{\text{n}}, f_{\text{i}} \rangle = \Phi(t, I_{\text{f}}, \mathcal{C}, O, f_{\text{occur}}),$$

where I_{f} is the instruction for Φ to infer factor values. f_{occur} is used for counterfactual reasoning to determine the values of the causal factors.

5.3 Theory-guided Reasoning

Algorithm 1: Theory-guided Reasoning

Input: Causal event set $\mathcal{C}$ and their values of factors $f_{\text{f}_{\text{focal}}}$, f_{occur}, f_{order}, f_{tc}, f_{nc}, f_{sc}, f_{n}, and f_{i}.
Output: A binary answer and an explanation.

for each $E \in \mathcal{C}$ **do**
  if $\neg f^{E}_{\text{focal}}$ **then** CONTINUE;
  if $f^{E}_{\text{nc}} \wedge f^{E}_{\text{sc}}$ **then return** "Yes";
  else if $\neg f^{E}_{\text{nc}} \wedge f^{E}_{\text{sc}}$ **then**
    $\mathcal{C}_s \leftarrow \{E' \in \mathcal{C} \mid \neg f^{E'}_{\text{nc}} \wedge f^{E'}_{\text{sc}}\}$;
    if $\text{len}(\text{set}(\{f^{E_s}_{\text{order}} \mid E_s \in \mathcal{C}_s\})) \neq 1$ **then**
      if $f^{E}_{\text{order}} = \max\{f^{E_s}_{\text{order}} \mid E_s \in \mathcal{C}_s\}$ **then return** "Yes";
    else
      for each $E_s \in \mathcal{C}_s$ **do**
        $f^{E_s}_{\text{resp}} \leftarrow \Phi(f^{E_s}_{\text{n}}, f^{E_s}_{\text{i}})$;
      if $f^{E}_{\text{resp}} \leftarrow \max\{f^{E_s}_{\text{resp}} \mid E_s \in \mathcal{C}_s\}$ **then return** "Yes";
  else if $f^{E}_{\text{tc}} \wedge \neg f^{E}_{\text{sc}}$ **then**
    if $f^{E}_{\text{n}} \vee f^{E}_{\text{i}}$ **then** $\mathcal{A}$.append("Yes");
    else
      $\mathcal{C}_a \leftarrow \{E' \in \mathcal{C} \mid f^{E'}_{\text{tc}} \wedge \neg f^{E'}_{\text{sc}}\}$;
      for each $E_a \in \mathcal{C}_a$ **do**
        $f^{E_a}_{\text{resp}} = \Phi(f^{E_a}_{\text{order}})$;
      if $\forall E_a \in \mathcal{C}_a \setminus \{E\}, f^{E}_{resp} > f^{E_a}_{resp}$ **then return** "Yes";
  else CONTINUE;
return "No"

Based on the extracted causal knowledge, Alg. 1 conducts theory-guided human-like causal reasoning for binary queries such as "Did X cause Y?". For each queried causal event E, the algo-

[1]The instruction states and generally ensures that causal events are minimal and do not overlap (Def. 1, AC3).

rithm first partitions the reasoning process using f_{tc}^E, f_{nc}^E, and f_{sc}^E (see §B.1.1). Then, for each partition, it determines whether E is a judged cause of O (see §B.1.2). Finally, it outputs a binary answer and provides an explanation (see §B.2). In Alg. 1, we further introduce a responsibility factor f_{resp} to measure relative responsibility. f_{resp}^E is defined as the relative causal contribution of E to O compared with those of other contributing causes (Kiciman et al., 2024).

Specifically, 1) If E is a sufficient token cause, it is a cause of O since it is complete. 2) If E is not a token cause, it is not a cause of O since it is not in the causal chain. 3) If E is one among multiple disjunctive contributing causes, Alg. 1 first examines their relative temporal orders to assess *preemption* (see §A.3). If they occur simultaneously, their f_{resp} values based on f_{n} and f_{i} are compared to determine the cause: E is a cause of O only if it takes the largest responsibility. Otherwise, if E occurs the earliest, it is a cause of O; otherwise, it is not. 4) If E is part of the conjunctive set of contributing causes, Alg. 1 first evaluates whether f_{n} or f_{i} enhances E's causal strength. If so, E is a cause of O. Otherwise, their f_{resp} values based on f_{order} are compared to determine the cause: E is a cause of O only if it takes the unique largest responsibility. Finally, Alg. 1 provides an explanation for each answer.

6 Experiments

6.1 Setups and Baselines

Setups. We implement HCR-REASONER and the baselines on recent LLMs for comparison, including Qwen2.5-32B/72B-Instruct (Qwen et al., 2025), DeepSeek-V3 (DeepSeek-AI et al., 2025), Gemini-2.0-Flash (Google, 2024), Claude-3.5-Sonnet (Anthropic, 2024), GPT-4o-2024-11-20 (OpenAI et al., 2024a), and GPT-4-0613 (OpenAI et al., 2024b). HCR-REASONER is also tested on reasoning models including DeepSeek-R1 (Guo et al., 2025) and QwQ-32B (Qwen Team, 2025). Temperature is set to 0 for reproducibility. We report averaged accuracy over 10 runs for the pilot study and a single run for the main results due to high API costs.

Baselines. We compare HCR-REASONER with three baselines: 1) *vanilla*, which directly prompts an LLM to output a Yes/No answer given the story and query; 2) *zero-shot CoT*, which appends the instruction "Let's think step by step." to the vanilla prompt; and 3) *manual CoT* (Chen et al., 2024), which replaces "Let's think step by step." with three in-context demonstrations containing manually written reasoning steps. GPT-4 + manual CoT (Chen et al., 2024) is the previous state-of-the-art on Big-Bench Hard causal judgment.

Table 3: Results of the pilot study. Human average is directly cited from Suzgun et al. (2023). See §D.1 for complete results and significance testing results.

Methods	Acc. (C.)	Acc. (I.)	Acc.
Human Average	-	-	69.60%
Qwen2.5-32B-Instruct	65.25%	80.43%	68.98%
+ HCR-REASONER	70.14%	80.43%	72.67%
Qwen2.5-72B-Instruct	65.89%	77.83%	68.82%
+ HCR-REASONER	71.63%	79.13%	73.48%
DeepSeek-V3	66.45%	77.61%	69.20%
+ HCR-REASONER	69.93%	78.48%	72.03%
Claude-3.5-Sonnet	67.23%	73.04%	68.66%
+ HCR-REASONER	72.62%	75.43%	73.32%
GPT-4o-2024-11-20	54.89%	85.22%	62.35%
+ HCR-REASONER	70.50%	85.22%	74.12%
GPT-4-0613	62.48%	80.22%	66.84%
+ HCR-REASONER	**74.61%**	78.26%	**75.51%**

6.2 Pilot Study

We preliminarily evaluate HCR-REASONER on Big-Bench Hard causal judgment to assess its effectiveness (Fig. 3). We report both overall accuracy and fine-grained accuracies for two query types: causation (C.) and intention (I.).

Overall Findings. 1) HCR-REASONER *consistently and significantly improves the performance of all tested LLMs.* GPT-4 + HCR-REASONER achieves an overall accuracy of 75.51%, followed by GPT-4o at 74.12%. This shows that integrating theory into LLMs is promising. 2) *Closed-source LLMs benefit more from* HCR-REASONER. We attribute this to the stronger capacity of these LLMs to handle complex, multi-step CoT instructions.

Comparison with Humans. Comparing with humans is crucial, as it benchmarks the performance of LLMs against the established human consensus on causal judgments. In Big-Bench Hard causal judgment, the Yes/No labels are derived from rigorous human experiments, providing a reliable and externally valid testbed (Suzgun et al., 2023). Although the average human score appears relatively low (69.60%), this is largely an artifact of the eval-

uation procedure: the original proportional human responses computed from individual votes are binarized, such that both a 100%-0% split and a 51%-49% split are collapsed into a single binary label. As a result, it is nearly impossible for an individual to achieve consistently high accuracy, particularly on instances with near-even splits, even though the best-performing human still reaches 100% accuracy (Suzgun et al., 2023). Critically, after applying HCR-REASONER, all LLMs exceed the human average, demonstrating that HCR-REASONER *can effectively and efficiently enable LLMs to replicate human consensus in causal judgments without relying on domain experts or crowd annotators*.

6.3 Main Results

Main results are presented in Tab. 4. We find that integrating explicit theories into LLMs is effective.

Main Findings. 1) HCR-REASONER *drives consistent improvement, with greater gains in stronger models.* This finding is consistent with our pilot study, highlighting the framework's capacity to effectively utilize the advanced reasoning abilities of state-of-the-art models. 2) *Zero-shot and manual CoT are insufficient for the task.* Neither zero-shot CoT nor manual CoT yields consistent performance improvements, highlighting the inherent difficulty of the task. This finding is consistent with prior work (Chen et al., 2024). 3) *Theory-grounded reasoning delivers substantial improvements.* The consistent boost in accuracy across diverse LLMs confirms the core benefit of our approach: grounding reasoning in explicit theories.

Human Evaluation. Due to the high cognitive load of the task, we randomly selected 100 samples from HCR-BENCH and conducted a human evaluation. Four participants are involved, achieving accuracies of 60%, 66%, 68%, and 69%, with an average accuracy of 65.75%. This average accuracy is lower than that of Big-Bench Hard causal judgment, consistent with our findings in §D.2 that HCR-BENCH is more challenging than Big-Bench Hard causal judgment.

6.4 Fine-Grained Analysis

Fine-Grained Accuracies. To better understand the internal mechanism of HCR-REASONER, we report fine-grained accuracies in Tab. 4 and Fig. 3. We find that *overall performance is driven by factor value inference*, since all LLMs demonstrate high accuracy in the initial stage of identifying events.

Table 4: The results of different LLMs on HCR-BENCH. **CE✓** and **OE✓** are the proportions of correctly identified causal and outcome events, respectively.

Methods	Accuracy	CE✓	OE✓
Open-source LLMs			
Qwen2.5-32B-Instruct	62.67%	-	-
+ zero-shot CoT	61.94%	-	-
+ manual CoT	62.85%	-	-
+ HCR-REASONER	64.78%	95.91%	93.96%
Qwen2.5-72B-Instruct	64.87%	-	-
+ zero-shot CoT	62.03%	-	-
+ manual CoT	62.31%	-	-
+ HCR-REASONER	67.52%	94.53%	91.95%
DeepSeek-V3	63.49%	-	-
+ zero-shot CoT	64.68%	-	-
+ manual CoT	63.95%	-	-
+ HCR-REASONER	67.61%	96.06%	91.40%
Closed-source LLMs			
Gemini-2.0-Flash	60.20%	-	-
+ zero-shot CoT	58.65%	-	-
+ manual CoT	58.55%	-	-
+ HCR-REASONER	64.96%	**96.63%**	95.61%
Claude-3.5-Sonnet	63.68%	-	-
+ zero-shot CoT	65.42%	-	-
+ manual CoT	62.67%	-	-
+ HCR-REASONER	70.54%	95.52%	91.67%
GPT-4o-2024-11-20	58.65%	-	-
+ zero-shot CoT	59.38%	-	-
+ manual CoT	60.66%	-	-
+ HCR-REASONER	68.07%	94.45%	92.04%
GPT-4-0613	63.77%	-	-
+ zero-shot CoT	62.49%	-	-
+ manual CoT	66.51%	-	-
+ HCR-REASONER	**71.82%**	94.26%	**96.07%**

However, we observe no clear correlation between these fine-grained accuracies and overall accuracy.

Causal Analysis. To examine the extent to which different factors influence overall accuracy, we perform a causal analysis. In Fig. 4, we plot the averaged average treatment effect (ATE) of each factor on overall accuracy across two estimation methods: ordinary least squares (OLS) (Angrist and Pischke, 2009) and propensity score matching (Matching) (Rosenbaum and Rubin, 1983), marking factors with $p \geq 0.05$ with cross marks. Results suggest that *only Qwen2.5-72B-Instruct and Claude-3.5-Sonnet exhibit faithful reasoning, while GPT-4 appears to utilize shortcuts.* Ideally, all factors except f_{nc}, which is primarily used for partitioning the reasoning process and does not contribute to determining causes, should causally contribute to

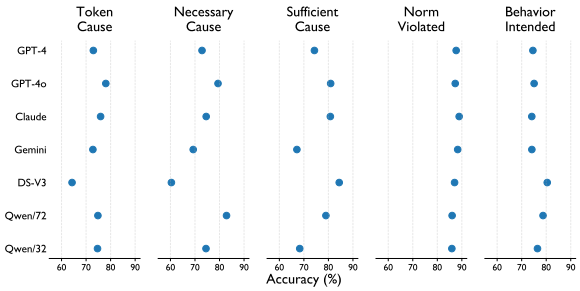


Figure 3: Fine-grained accuracies of different factors.

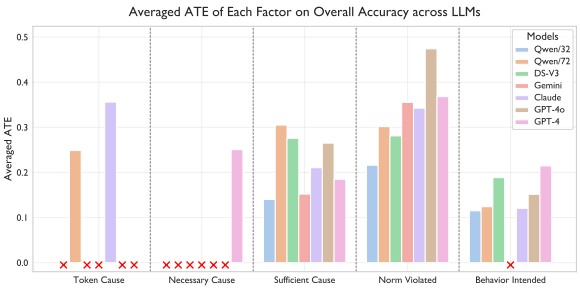


Figure 4: Results of the causal analysis.

overall accuracy. However, only two models align well with this expectation. GPT-4 appears to rely heavily on f_{nc} while underutilizing f_{tc}, suggesting that it may be exploiting shortcuts rather than genuinely understanding the task. More broadly, the limited number of LLMs recognizing the importance of f_{tc} underscores a common deficiency in current LLMs' grasp of token causality.

6.5 Ablation Study

We conduct an ablation study to quantify the contribution of each stage and report performance across the following settings: 1) *vanilla*; 2) *FO* (the first one stage only); 3) *FT* (the first two stages only); and 4) HCR-REASONER.

Non-Reasoning LLMs. 1) *The first stage alone typically degrades performance.* This is observed across non-reasoning LLMs, with the exception of DeepSeek-V3. While these models can effectively identify events, as shown in Tab. 4, the instruction of the first stage may lead them to extract superfluous or irrelevant causal events, thereby hindering accurate decision-making. 2) *Combining the first two stages typically improves performance.* Encouraging non-reasoning LLMs to infer and consider distinct causal and cognitive factors promotes a clearer representation of the underlying causal mechanisms, leading to more informed and consistent reasoning. 3) *Theory-guided reasoning yields the most substantial gains.* This observation directly supports our central hypothesis: the proposed Alg. 1 enables precise human-like causal reasoning by leveraging the structured causal knowledge inferred in the earlier stages, thereby aligning model outputs more closely with human judgments.

Table 5: Ablation results on HCR-BENCH. Green and red highlights denote performance rise/drop compared with Vanilla.

Model	Vanilla → FO → FT → HCR-REASONER
Open-source LLMs	
Qwen2.5-32B	62.67% → 61.67% → 64.59% → 64.78%
Qwen2.5-72B	64.87% → 64.41% → 67.98% → 67.52%
DeepSeek-V3	63.49% → 67.25% → 65.87% → 67.61%
Closed-source LLMs	
Gemini-2.0	60.20% → 58.19% → 60.84% → 64.96%
Claude-3.5	63.68% → 58.83% → 60.93% → 70.54%
GPT-4o	58.65% → 58.28% → 58.46% → 68.07%
GPT-4	63.77% → 62.31% → 64.32% → 71.82%
Reasoning LLMs	
QwQ-32B	54.99% → 56.17% → 56.08% → 67.15%
DeepSeek-R1	57.09% → 58.28% → 58.92% → 69.72%

Reasoning LLMs. Since non-reasoning LLMs lack intrinsic reflection and verification mechanisms, we further evaluate slow thinking models including DeepSeek-R1 and QwQ-32B. 1) *Slow thinking models perform poorly alone.* This underscores the inherent difficulty of human-like causal reasoning in LLMs, indicating that mere self-reflection without guidance is insufficient. 2) *Slow thinking models achieve the largest performance gains with* HCR-REASONER. This demonstrates that reflection and verification capabilities become truly effective only when properly guided by theory.

7 Conclusion

In this work, we focus on human-like causal reasoning in LLMs. We construct HCR-BENCH, the first benchmark focusing on fine-grained human-like causal reasoning evaluation. It contains 1,093 challenging instances with detailed annotations of human reasoning steps. Also, we propose HCR-REASONER, which systematically integrates LLMs with token causality and cognitive science theories for human-like causal reasoning. Results reveal that HCR-REASONER consistently improves LLM performance over baselines, and that integrating theories into LLMs is highly effective. In the future, it is interesting to explore how to internalize this ability into LLMs.

Limitations

First, the HCR-REASONER *framework remains incomplete.* It incorporates several commonly studied cognitive factors but may overlook more fine-grained or domain-specific factors that influence human causal reasoning. Furthermore, the current implementation processes only a single causal event at a time. *Second, relying solely on LLMs to identify token causes is insufficient.* A more formal symbolic reasoning component is needed to ensure interpretability, consistency, and adherence to established causal principles.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62576016). We would like to thank the anonymous reviewers for their helpful comments.

References

Mark D Alicke. 2000. Culpable control and the psychology of blame. *Psychological bulletin*, 126(4):556.

Joshua D Angrist and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Anthropic. 2024. Introducing claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pages 2–9.

Jonathan Baron and Ilana Ritov. 2004. Omission bias, individual differences, and normality. *Organizational behavior and human decision processes*, 94(2):74–85.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. CausalQA: A benchmark for causal question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024. Causal evaluation of language models. *Preprint*, arXiv:2405.00622.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems*, volume 37, pages 96640–96670. Curran Associates, Inc.

Randolph Clarke, Joshua Shepherd, John Stigall, Robyn Repko Waller, and Chris Zarpentine. 2015. Causation, norms, and omissions: A study of causal judgments. *Philosophical Psychology*, 28(2):279–293.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Peter DeScioli, Rebecca Bruening, and Robert Kurzban. 2011. The omission effect in moral cognition: Toward a functional explanation. *Evolution and Human Behavior*, 32(3):204–215.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

Eugene S Edgington. 1969. Approximate randomization tests. *The Journal of Psychology*, 72(2):143–149.

Tobias Gerstenberg and David A Lagnado. 2012. When contributions make a difference: Explaining order effects in responsibility attribution. *Psychonomic bulletin & review*, 19(4):729–736.

Tobias Gerstenberg and Simon Stephan. 2021. A counterfactual simulation model of causation by omission. *Cognition*, 216:104842.

Google. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

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

Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*.

Herbert Lionel Adolphus Hart and Tony Honoré. 1985. *Causation in the Law*. OUP Oxford.

Daniel Murray Hausman. 2005. Causal relata: Tokens, types, or variables? *Erkenntnis*, 63(1):33–54.

Paul Henne, Aleksandra Kulesza, Karla Perez, and Augustana Houcek. 2021. Counterfactual thinking and recency effects in causal judgment. *Cognition*, 212:104708.

Paul Henne, Ángel Pinillos, and Felipe De Brigard. 2017. Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, 95(2):270–283.

Denis J Hilton, John McClure, and Robbie M Sutton. 2010. Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, 40(3):383–400.

Christopher Hitchcock and Joshua Knobe. 2009. Cause and norm. *The Journal of Philosophy*, 106(11):587–612.

Haitao Jiang, Lin Ge, Yuhe Gao, Jianian Wang, and Rui Song. 2024. LLM4causal: Democratized causal tools for everyone via large language model. In *First Conference on Language Modeling*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 31038–31065. Curran Associates, Inc.

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*.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*. Featured Certification.

Joshua Knobe and Benjamin Fraser. 2008. Causal judgment and moral judgment: Two experiments. In *Moral Psychology Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, pages 441–447. The MIT Press.

Jonathan F Kominsky and Jonathan Phillips. 2019. Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive science*, 43(11):e12792.

David A Lagnado and Shelley Channon. 2008. Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108(3):754–770.

David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. 2013. Causal responsibility and counterfactuals. *Cognitive science*, 37(6):1036–1073.

David Lewis. 2000. Causation as influence. *The Journal of Philosophy*, 97(4):182–197.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9215–9235, Bangkok, Thailand. Association for Computational Linguistics.

Yujie Lu, Weixi Feng, Wanrong Zhu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2023. Neuro-symbolic procedural planning with commonsense prompting. In *The Eleventh International Conference on Learning Representations*.

David R Mandel. 2003. Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, 132(3):419.

Helena Matute, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A Vadillo, and Itxaso Barberia. 2015. Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in psychology*, 6:888.

Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 78360–78393. Curran Associates, Inc.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/.

Kevin Reuter, Lara Kirfel, Raphael Van Riel, and Luca Barlassina. 2014. The good, the bad, and the timely: How temporal order and moral judgment influence causal selection. *Frontiers in psychology*, 5:1336.

Ilana Ritov and Jonathan Baron. 1992. Status-quo and omission biases. *Journal of risk and uncertainty*, 5(1):49–61.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

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

Jana Samland, Marina Josephs, Michael R Waldmann, and Hannes Rakoczy. 2016. The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, 145(2):125.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Jonathan Schaffer. 2003. Overdetermining causes. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 114(1/2):23–45.

Steven A Sloman and David Lagnado. 2015. Causality in thought. *Annual review of psychology*, 66(1):223–247.

Barbara A Spellman. 1997. Crediting causality. *Journal of Experimental Psychology: General*, 126(4):323.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Nadya Vasilyeva and Tania Lombrozo. 2015. Explanations and causal judgments are differentially sensitive to covariation and mechanism information. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37.

Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological review*, 92(4):548.

Gary L Wells and Igor Gavanski. 1989. Mental simulation of causality. *Journal of personality and social psychology*, 56(2):161.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*.

Jiayao Zhang, Hongming Zhang, Weijie Su, and Dan Roth. 2022. ROCK: Causal inference principles for reasoning about commonsense causality. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26750–26771. PMLR.

Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang, and Haibin Chen. 2023. Preserving commonsense knowledge from pre-trained language models via causal inference. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9155–9173, Toronto, Canada. Association for Computational Linguistics.

A Background

A.1 Human Causal Reasoning

The causal reasoning process of humans is a sophisticated cognitive process. It moves beyond simple counterfactual analysis to incorporate the psychological nuances of human judgment (Knobe and Fraser, 2008; Alicke, 2000).

Stage 1. In this stage, *humans establish a basic causal chain*. For example, in philosophy and legal theory, individuals identify the but-for causes (Def. 2) of an outcome (Hart and Honoré, 1985). This process creates a foundational set of potential causes without evaluating their relative importance.

Stage 2. In this stage, *psychological and normative factors modulate the initial causal structure to produce a final judgment*. This is where human reasoning diverges from purely logical models. The strength of a perceived causal link is adjusted based on: 1) *normative considerations*, where people attribute greater causal responsibility to agents and events that violate social or moral norms (Knobe and Fraser, 2008; Hitchcock and Knobe, 2009); 2) *blame-driven evaluations*, where the desire to assign blame influences causal assessment for negative events. This motivational bias, central to attribution theory (Weiner, 1985), can retroactively alter perceptions of an agent's control and role, making their actions appear more blameworthy than an objective analysis would suggest (Alicke, 2000; Lagnado and Channon, 2008).

A.2 Other Causal Factors

Apart from the HP definition described in Section 2.2, we also include necessary cause (Def. 2) and sufficient cause (Def. 3) as causal factors.

Definition 2 (But-for/Necessary Cause)**.** $\boldsymbol{X} = \boldsymbol{x}$ *is a but-for cause of* φ *if* φ *would not have happened but for* $\boldsymbol{X} = \boldsymbol{x}$.

Definition 3 (Sufficient Cause)**.** $\boldsymbol{X} = \boldsymbol{x}$ *is a sufficient cause of* φ *in the causal setting* $(M, \boldsymbol{u})$ *if the following four conditions hold:*

SC1. $(M, \boldsymbol{u}) \models (\boldsymbol{X} = \boldsymbol{x})$ *and* $(M, \boldsymbol{u}) \models \varphi$.

SC2. Some conjunct of $\boldsymbol{X} = \boldsymbol{x}$ *is part of a cause of* φ *in the causal setting* $(M, \boldsymbol{u})$.

SC3. $(M, \boldsymbol{u}') \models [\boldsymbol{X} \leftarrow \boldsymbol{x}]\varphi$ *for all contexts* $\boldsymbol{u}'$.

SC4. $\boldsymbol{X}$ *is minimal; there is no strict subset* $\boldsymbol{X}'$ *of* $\boldsymbol{X}$ *such that* $\boldsymbol{X}' = \boldsymbol{x}'$ *satisfies conditions SC1, SC2, and SC3, where* $\boldsymbol{x}'$ *is the restriction of* $\boldsymbol{x}$ *to the variables in* $\boldsymbol{X}'$.

Prop. 1 formally proves that a token cause is an extension of a but-for cause.

Proposition 1. *If* $\boldsymbol{X} = \boldsymbol{x}$ *is a but-for cause of* φ *in the causal setting* $(M, \boldsymbol{u})$*, then* $\boldsymbol{X} = \boldsymbol{x}$ *is a cause of* φ *according to the HP definition.*

Proof. Suppose $\boldsymbol{X} = \boldsymbol{x}$ is a but-for cause of φ. Then there exists at least a possible value $\boldsymbol{x}'$ such that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}']\neg\varphi$. Let $\boldsymbol{W} = \emptyset$ and consider the intervention $\boldsymbol{X} \leftarrow \boldsymbol{x}'$. Then AC2 is satisfied. Therefore, $\boldsymbol{X} = \boldsymbol{x}$ qualifies as an token cause, as a special case where $\boldsymbol{W} = \emptyset$. □

A.3 A Failure Case of But-for Causation

A classic example of preemption (Lewis, 2000) is as follows:

> *Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Because both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.*

According to Def. 2, neither Suzy's nor Billy's throw would be identified as the cause of the bottle shattering. However, token causality (Def. 1) can capture this nuance.

Proof. We construct a structural causal model with the following variables:

- ST for "Suzy throws," with values 0 (she does not) and 1 (she does);
- BT for "Billy throws," with values 0 (he does not) and 1 he does);
- SH for "Suzy's rock hits the bottle," with values 0 (it does not) and 1 (it does);
- BH for "Billy's rock hits the bottle," with values 0 (it does not) and 1 (it does);
- BS for "Bottle shatters," with values 0 (it does not) and 1 (it does).

The corresponding structural equations are defined as:

- $BS = 1$ if $SH = 1$ or $BH = 1$;
- $SH = 1$ if $ST = 1$;

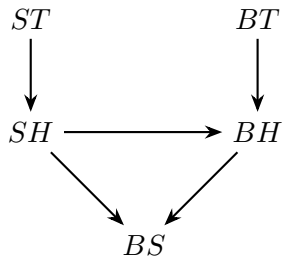


Figure 5: The causal representation of the case.

- $BH = 1$ if $BT = 1$ and $SH = 0$.

First, it is clear that both $ST = 1$ and $BS = 1$ satisfy AC1 and AC3. For $ST = 1$, if Suzy had not thrown ($ST \rightarrow 0$) and $\boldsymbol{W} = \{BH\}$, then the bottle would not have shattered ($BS = 0$) since we have forced Billy's rock not to hit the bottle ($BH \rightarrow 0$). Thus, $ST = 1$ is an token cause of $BS = 1$. For $BT = 1$, if Billy had not thrown ($BT \rightarrow 0$), all other variables would remain unchanged, including BS. Therefore, $BT = 1$ is not an token cause of $BS = 1$ since no suitable $\boldsymbol{W}$ can be found to satisfy AC2. Therefore, according to Def. 1, $ST = 1$ is a cause of $BS = 1$, whereas $BT = 1$ is not. □

B Details of HCR-Reasoner

B.1 Details of Algorithm 1

B.1.1 Reasoning Process Partitioning

Let E be a queried causal event and O be the outcome event. We use factors f^E_{tc}, f^E_{nc}, and f^E_{sc} to partition the reasoning process. The objective of partitioning is to ensure that, within each partition, the causal relation between E and O can be assessed using existing techniques from token causality and causal judgment.

Criterion 1. If E is a sufficient cause of O, the partitioning is based on f^E_{nc}: 1) If f^E_{nc} holds, by Prop. 1, f^E_{tc} also holds. Since f^E_{sc} also holds, E has 100% responsibility for O, making the causal relation determined. 2) If $\neg f^E_{\text{nc}}$ holds, this corresponds to *preemption* (Halpern and Pearl, 2005; Halpern, 2016) or *overdetermination* (Schaffer, 2003), which are well-studied and defined primarily based on f_{nc} rather than f_{tc}.

Criterion 2. If E is not a sufficient cause of O, the partitioning is based on f^E_{tc}: 1) If f^E_{tc} holds, E is at least part of a cause by Def. 1. Therefore, we only need to consider factors that increase causal strength. 2) If $\neg f^E_{\text{tc}}$ holds, by Prop. 1, $\neg f^E_{\text{nc}}$ also holds. Since $\neg f^E_{\text{sc}}$ also holds, E has 0% responsibility for O, making the causal relation determined.

The partitions are exhaustive, as they jointly cover the entire space of possibilities. Furthermore, within each partition, the causal relation between E and O can be effectively assessed using existing theories from token causality and causal judgment. Therefore, the partitioning process is well-founded.

B.1.2 Rules within Each Partition

Then, we will justify how to assess the causal relation within each partition using rules.

Partition 1 (Line 3). Since E is a necessary cause of O, it is also a token cause by Prop. 1. Therefore, E is a sufficient token cause (100% responsibility) of O.

$$f^E_{\text{nc}} \wedge f^E_{\text{sc}} \Rightarrow f^E_{\text{tc}} \wedge f^E_{\text{sc}}$$
$$\Rightarrow E \text{ has 100\% responsibility for } O$$
$$\Rightarrow E \text{ is a cause of } O.$$

Partition 2 (Line 17). Since E is not a token cause of O, it is also not a necessary cause by Prop. 1. Therefore, E is neither a sufficient cause nor a necessary cause (0% responsibility) of O.

$$\neg f^E_{\text{tc}} \wedge \neg f^E_{\text{sc}} \Rightarrow \neg f^E_{\text{nc}} \wedge \neg f^E_{\text{sc}}$$
$$\Rightarrow E \text{ has 0\% responsibility for } O$$
$$\Rightarrow E \text{ is not a cause of } O.$$

Partition 3 (Lines 4-10). Since E is a sufficient cause but not a necessary cause of O, there should exist a set of alternative sufficient causal events $\mathcal{C}_s$, each of which can independently cause O. In token causality, typically only one cause actually takes effect (Halpern, 2016). If events in $\mathcal{C}_s$ do not occur simultaneously, the one that occurs first is judged as the preemptive cause of O, while the others are not (Halpern and Pearl, 2005; Halpern, 2016). Otherwise, if events in $\mathcal{C}_s$ occur simultaneously, their responsibilities should be compared. Here, responsibility is defined as the relative degree to which E causally contributes to O compared with other contributing causal events. Causal contribution is evaluated using the remaining factors, i.e., $f^E_{\text{resp}} \leftarrow \Phi(f^E_{\text{n}}, f^E_{\text{i}}, f^E_{\text{order}})$. If E has the highest responsibility for O, it is a cause of O since it actually takes effect; otherwise, it is not. In the special case where all events in $\mathcal{C}_s$ have equal responsibility, the scenario corresponds to overdetermination (Schaffer, 2003), where each event in $\mathcal{C}_s$ is a cause

of O. Let $\mathcal{C}_s$ be the set of causal events that satisfy $\neg f_{\text{nc}} \wedge f_{\text{sc}}$ (including E), then:

$\exists E_s \in \mathcal{C}_s, f^{E_s}_{\text{order}} < f^{E}_{\text{order}}$
$\Rightarrow E$ does not actually take effect
$\Rightarrow E$ is not a cause of O.

$\forall E_s \in \mathcal{C}_s, f^{E_s}_{\text{order}} \geq f^{E}_{\text{order}}$
$\Rightarrow E$ actually takes effect
$\Rightarrow E$ is a cause of O.

$\forall E_s \in \mathcal{C}_s, f^{E_s}_{\text{order}} = f^{E}_{\text{order}} \wedge f^{E}_{\text{resp}} = \max_{E_s \in \mathcal{C}_s} f^{E_s}_{\text{resp}}$
$\Rightarrow E$ actually takes effect
$\Rightarrow E$ is a cause of O.

$\forall E_s \in \mathcal{C}_s, f^{E_s}_{\text{order}} = f^{E}_{\text{order}} \wedge f^{E}_{\text{resp}} \neq \max_{E_s \in \mathcal{C}_s} f^{E_s}_{\text{resp}}$
$\Rightarrow E$ does not actually take effect
$\Rightarrow E$ is not a cause of O.

Partition 4 (Lines 11-16). Since E is a token cause of O, it is at least a partial cause by Def. 1. To determine whether E is a cause of O, we first consider factors that enhance causal strength, including norm violations and intended behaviors. If E satisfies either of these conditions, it is a cause of O. If neither condition is met, we turn to its responsibility $f^{E}_{\text{resp}} \leftarrow \Phi(f^{E}_{\text{order}})$. If E is uniquely the most responsible for O, it is a cause of O; otherwise, it is not. Let $\mathcal{C}_a$ be the set of causal events that satisfy $f^{E}_{\text{tc}} \wedge \neg f_{\text{sc}}$ (including E), then:

$f^{E}_{\text{tc}} \wedge (f^{E}_{\text{n}} \vee f^{E}_{\text{i}})$
$\Rightarrow$ The causal strength of E is enhanced
$\Rightarrow E$ is a cause of O.

$f^{E}_{\text{tc}} \wedge \neg(f^{E}_{\text{n}} \vee f^{E}_{\text{i}}) \wedge \forall E_a \in \mathcal{C}_a \setminus \{E\}, f^{E}_{\text{resp}} > f^{E_a}_{\text{resp}}$
$\Rightarrow$ The causal strength of E is enhanced
$\Rightarrow E$ is a cause of O.

$f^{E}_{\text{tc}} \wedge \neg(f^{E}_{\text{n}} \vee f^{E}_{\text{i}}) \wedge \exists E_a \in \mathcal{C}_a, f^{E}_{\text{resp}} \leq f^{E_a}_{\text{resp}}$
$\Rightarrow$ The causal strength of E is not enhanced
$\Rightarrow E$ is not a cause of O.

B.2 Details of Explanation Generation

In Alg. 1, an explanation is generated for each answer. Most explanations are produced using predefined templates. For responsibility-related decisions, the explanations are generated by the LLM responsible for determining responsibility. First, we present the templates:

- Line 3: E is a cause of O, since E is a sufficient token cause.
- Line 7: E is a cause of O, since E occurs the earliest in a disjunctive causal structure. (E is not a cause of O, since E does not occur the earliest in a disjunctive causal structure.)
- Line 12: E is a cause of O, since E is a token cause and it violates a norm or is an intended behavior of an agent. (E is not a cause of O, since E is a token cause but it does not violate a norm and is not an intended behavior of an agent.)
- Line 17: E is not a cause of O, since E is not in the causal chain.

Second, we provide examples of LLM-generated responsibility-related explanations:

- Yes, the focal causal event "Billy arrives at 9 am" is equally responsible relative to the other causal event "Suzy arrives at 9 am". Both Billy and Suzy arrived at the same time, triggering the motion detector. Although Billy was explicitly told not to come, Suzy was also explicitly told to come at exactly 9 am, making both of them equally responsible for the outcome.
- No. In the story, Drew, Kylie, and Oliver all ordered coffee on the same day, and any one of them ordering coffee would have been sufficient for the coffee shop to make a profit. However, Drew's action of ordering coffee on a Tuesday was unexpected and violated the norm, as he usually doesn't order coffee on Tuesdays. This makes Drew's action more noticeable, but it doesn't make him more responsible for the coffee shop's profit than Kylie or Oliver, who also ordered coffee.
- Yes, the focal causal event "Brother steps on a slippery patch" is more responsible relative to other causal events in the list. This is because it occurs later in the sequence of events, closer to the outcome event. The brother stepping on the slippery patch directly leads to him knocking over the lamp, which then fractures the water pipe, causing the damage. The other events, while they set the stage for this to happen, are further removed from the outcome event in terms of temporal order.

- No. Based on the story, Alice's action of logging in at 9.30 am is not more responsible than Zoe's action of logging in at 9 am for the outcome event. This is because both actions are necessary for the outcome to occur, and Zoe's action, which violates the company policy, happens first. Therefore, Zoe's action has a higher temporal order and is considered more responsible for the outcome.

C Details of HCR-BENCH

C.1 Details of Data Cleaning

We provide examples of erroneous, irrelevant, and partially flawed samples.

Erroneous Sample. Erroneous samples are those that miss critical background information. In the case below, the "motion detector" does not appeared in the story background.

> **Story:** *Suzy and Billy are working on a project that is very important for our nation's security. The boss tells Suzy: "Be sure that you are here at exactly 9 am. It is absolutely essential that you arrive at that time." Then he tells Billy: "Be sure that you do not come in at all tomorrow morning. It is absolutely essential that you not appear at that time." Both Billy and Suzy arrive at 9 am.*
>
> **Query:** Did Billy cause the motion detector to go off?
>
> **Answer:** No

Irrelevant Sample. Irrelevant samples are those that query about intention.

> **Story:** *The CEO of a company is sitting in his office when his Vice President of R&D comes in and says, "We are thinking of starting a new programme. It will help us increase profits, but it will also harm the environment." The CEO responds that he doesn't care about harming the environment and just wants to make as much profit as possible. The programme is carried out, profits are made and the environment is harmed.*
>
> **Query:** Did the CEO intentionally harm the environment?
>
> **Answer:** Yes

Partially Flawed Sample. Partially flawed samples are those with issues in the story, query, or answer. In many cases, errors in the query and answer co-occur. For instance, a duplicated or malformed query often results in an incorrect answer. In the case below, "work emails" is mistakenly written as "spam emails."

> **Story:** *Billy and Suzy work for the same company. They work in different rooms, and both of them sometimes need to access the central computer of the company. Nobody at the company is aware that if two people are logged into the central computer at the same time, some spam emails containing important customer information are immediately deleted from the central computer. In order to make sure that two people are available to answer phone calls during designated calling hours, the company issued the following official policy: Billy and Suzy are both permitted to log into the central computer in the mornings, and neither of them are permitted to log into the central computer in the afternoons. Today at 9 am, Billy and Suzy both log into the central computer at the same time. Immediately, some work emails containing important customer information are deleted from the central computer.*
>
> **Query:** Did Suzy cause the central computer to delete some work emails containing important customer information?
>
> **Answer:** Yes

C.2 Details of Data Annotation

The annotation criteria are presented in Tab. 6. The annotation process follows the pipeline below:

1. Annotate $\mathcal{C}$, O, f_{occur}, f_{order}, f_{n}, and f_{i}. The values of these factors are usually explicit in the story and query, so annotation is relatively straightforward.

2. Annotate f_{tc}, f_{nc}, and f_{sc}. Annotation follows the order $f_{\text{nc}} \rightarrow f_{\text{tc}} \rightarrow f_{\text{sc}}$, because f_{nc} is the easiest to determine, and if it holds, f_{tc} necessarily holds; meanwhile, the determination of f_{sc} depends on f_{tc}. Only when $\neg f_{\text{nc}}$ holds does one need to evaluate f_{tc} according to the AC2 clause in Def. 1.

Table 6: Annotation guideline.

Factor	Value	Definition
f^E_{occur}	TRUE	E actually occurs in the causal setting.
	FALSE	Otherwise.
f^E_{order}	/	An integer starting from 0, denoting the temporal order of E relative to other events in the causal setting. If two events occur simultaneously, they should share the same order.
f^E_{tc}	TRUE	E is an token cause of O by Def. 1. If 1) $f^E_{\text{occur}} \wedge f^O_{\text{occur}}$ holds and 2) f^E_{nc} holds while allowing contingencies based on AC2, then f^E_{tc} holds. [Proof: AC1 holds if $f^E_{\text{occur}} \wedge f^O_{\text{occur}}$ holds; AC2 holds if f^E_{nc} holds while allowing contingencies; AC3 already holds since we only annotate minimal/atomic and non-overlapping causal events.]
	FALSE	Otherwise.
f^E_{nc}	TRUE	E is a necessary cause of O by Def. 2. If but for E, O would not have occurred, then f^E_{nc} holds.
	FALSE	Otherwise.
f^E_{sc}	TRUE	E is a sufficient cause of O by Def. 3. If 1) f^E_{tc} holds and 2) E always suffices to cause O while changing the f_{occur} factor of other causal events, then f^E_{sc} holds. [Proof: SC1 holds if $f^E_{\text{occur}} \wedge f^O_{\text{occur}}$ holds; SC2 holds if f^E_{tc} holds; SC3 holds by counterfactual reasoning on causal events other than E; SC4 already holds since we only annotate minimal/atomic and non-overlapping causal events. Also, if f^E_{tc} holds, $f^E_{\text{occur}} \wedge f^O_{\text{occur}}$ holds.]
	FALSE	Otherwise.
f^E_{n}	TRUE	E violates a prescriptive or descriptive norm.
	FALSE	Otherwise.
f^E_{i}	TRUE	E is an agent's behavior and the agent is aware of the potential consequences of their action and knowingly performs an action that leads to a foreseeable bad outcome.
	FALSE	Otherwise.

D More Results

D.1 Pilot Results

Complete Results. The complete results are presented in Tab. 7. We report both overall and fine-grained accuracies along with standard deviations over 10 runs. *First, for causation queries, all models exhibit consistent performance improvements with the integration of* HCR-REASONER*, with open-source LLMs generally benefiting more.* For instance, GPT-4's accuracy increase from 62.48% to 74.61% after applying HCR-REASONER. GPT-4 + HCR-REASONER also achieves the highest overall accuracy of 75.51%, significantly surpassing the average human performance. These results suggest that incorporating domain theory into LLMs to enhance reasoning capabilities is promising. *Second, neither zero-shot nor manual CoT is sufficient for human-like causal reasoning.* This suggests the inherent difficulty of the task.

Table 7: Complete results of the pilot study.

Methods	Acc. (C.)	Acc. (I.)	Acc.
Human Average	-	-	69.60±0.00
Qwen2.5-32B-Instruct	65.25±0.32	80.43±0.00	68.98±0.24
+ zero-shot CoT	61.70±0.78	75.22±1.74	65.03±0.49
+ manual CoT	65.89±0.59	79.13±1.06	69.14±0.34
+ HCR-REASONER	70.14±1.83	80.43±0.00	72.67±1.38
Qwen2.5-72B-Instruct	65.89±1.21	77.83±1.30	68.82±1.04
+ zero-shot CoT	62.06±2.27	77.83±3.61	65.94±2.02
+ manual CoT	63.12±2.01	79.13±1.06	67.06±1.59
+ HCR-REASONER	71.63±0.90	79.13±1.44	73.48±0.49
DeepSeek-V3	66.45±1.96	77.61±1.70	69.20±1.76
+ zero-shot CoT	67.38±2.73	74.57±3.37	69.14±1.82
+ manual CoT	62.62±2.15	76.30±1.52	65.99±1.59
+ HCR-REASONER	69.93±1.88	78.48±1.17	72.03±1.62
Claude-3.5-Sonnet	67.23±2.57	73.04±4.68	68.66±2.52
+ zero-shot CoT	65.25±4.09	71.96±4.29	66.90±3.69
+ manual CoT	59.65±3.21	84.57±2.27	65.78±2.37
+ HCR-REASONER	72.62±1.11	75.43±4.24	73.32±1.54
GPT-4o-2024-11-20	54.89±1.53	85.22±1.90	62.35±1.29
+ zero-shot CoT	61.99±1.20	77.39±3.25	65.78±1.15
+ manual CoT	60.64±0.91	78.48±2.27	65.03±0.55
+ HCR-REASONER	70.50±1.96	85.22±2.13	74.12±1.36
GPT-4-0613	62.48±0.59	80.22±2.65	66.84±0.83
+ zero-shot CoT	62.77±0.73	81.52±1.46	67.38±0.68
+ manual CoT	65.89±0.92	78.26±0.97	68.93±0.65
+ HCR-REASONER	**74.61**±0.76	78.26±2.17	**75.51**±0.79

Significance Test. We have also conducted an approximate randomization test (Edgington, 1969) to assess the statistical significance of HCR-REASONER on Big-Bench Hard causal judgment, since this dataset is relatively small in size. For each model, we evaluate HCR-REASONER and 3 baselines across 10 runs. Each test is repeated 30 times with 10,000 trials per run to compute the average p-value. A Bonferroni correction (Dunn, 1961) is applied for multiple comparisons (18 in total: 6 models $\times$ 3 baselines), yielding a corrected significance threshold of approximately 0.00278. As shown in Tab. 8, HCR-REASONER achieves statistical significance in 18/18 comparisons and 16/18 comparisons after Bonferroni correction, despite the relatively small size of Big-Bench Hard causal judgment.

D.2 Case Study

Reasoning Steps. For the first axis of our case study, we present the reasoning steps of Claude to quantitatively assess the effectiveness of HCR-REASONER. We first present an example sample.

> **Story:** *Tom has a huge garden and loves flowers. He employed two gardeners who*

Table 8: Statistical significance of HCR-REASONER on Big-Bench Hard causal judgment.

Model	Comparison	Accuracy Diff	p-value ($\pm$ std)	Significant	Significant (Bonferroni)
Qwen2.5-32B-Instruct	HCR-REASONER vs. Vanilla	0.0369	0.0020 ± 0.0004	$\surd$	$\surd$
	HCR-REASONER vs. Zero-shot CoT	0.0765	0.0021 ± 0.0004	$\surd$	$\surd$
	HCR-REASONER vs. Manual CoT	0.0353	0.0022 ± 0.0004	$\surd$	$\surd$
Qwen2.5-72B-Instruct	HCR-REASONER vs. Vanilla	0.0465	0.0022 ± 0.0004	$\surd$	$\surd$
	HCR-REASONER vs. Zero-shot CoT	0.0754	0.0022 ± 0.0003	$\surd$	$\surd$
	HCR-REASONER vs. Manual CoT	0.0642	0.0020 ± 0.0004	$\surd$	$\surd$
DeepSeek-V3	HCR-REASONER vs. Vanilla	0.0283	0.0079 ± 0.0009	$\surd$	$\times$
	HCR-REASONER vs. Zero-shot CoT	0.0289	0.0020 ± 0.0005	$\surd$	$\surd$
	HCR-REASONER vs. Manual CoT	0.0604	0.0020 ± 0.0003	$\surd$	$\surd$
Claude-3.5-Sonnet	HCR-REASONER vs. Vanilla	0.0465	0.0080 ± 0.0007	$\surd$	$\times$
	HCR-REASONER vs. Zero-shot CoT	0.0642	0.0021 ± 0.0005	$\surd$	$\surd$
	HCR-REASONER vs. Manual CoT	0.0754	0.0020 ± 0.0004	$\surd$	$\surd$
GPT-4o-2024-11-20	HCR-REASONER vs. Vanilla	0.1176	0.0019 ± 0.0005	$\surd$	$\surd$
	HCR-REASONER vs. Zero-shot CoT	0.0834	0.0020 ± 0.0005	$\surd$	$\surd$
	HCR-REASONER vs. Manual CoT	0.0909	0.0021 ± 0.0004	$\surd$	$\surd$
GPT-4-0613	HCR-REASONER vs. Vanilla	0.0866	0.0021 ± 0.0004	$\surd$	$\surd$
	HCR-REASONER vs. Zero-shot CoT	0.0813	0.0020 ± 0.0004	$\surd$	$\surd$
	HCR-REASONER vs. Manual CoT	0.0658	0.0020 ± 0.0005	$\surd$	$\surd$

take care of the plants on his 30 flower beds: Alex and Benni. Both can independently decide on their working hours and arrange who cares for which flower beds. Alex and Benni are very reliable and Tom is satisfied with their work. Nevertheless he wants to optimize the plant growth. Since Tom has read in a magazine that plants grow better when they are fertilized, he decides to let Alex and Benni fertilize his plants. The magazine recommends the use of the chemicals A X200R or B Y33R, since both are especially effective. However, Tom also read that it can damage plants when they are exposed to multiple different types of chemicals. Tom therefore decides that he only wants to use one fertilizer. He goes for A X200R. Tom instructs Alex and Benni to buy the chemical A X200R and to use only this fertilizer. Alex volunteers for buying several bottles of this chemical for Benni and himself. After a few weeks, Tom goes for a walk in his garden. He realizes that some of his plants are much prettier and bigger than before. However, he also realizes that some of his plants have lost their beautiful color and are dried up. That makes Tom very sad and reflective. He wonders whether the drying of his plants might have something to do with the fertilization. He wants to investigate this matter and talks to Alex and Benni. Alex tells him that he followed Tom's instruction: "I only bought and used the chemical A X200R which I had funneled into the blue can." Benni suddenly is startled and says to Alex: "What? You funneled A X200R into the blue can? But you told me you had funneled it into the green can! That's why I always used the green can!" Alex replies: "Did I? Then I am sorry!" Tom remembers that he had filled B Y33R in a green can - long before he had read about the chemicals in his magazine. He had never used it. So Benni must have accidentally, without knowing it, applied the chemical B Y33R, whereas only Alex applied A X200R. Tom realizes that the plants dried up in the flower beds on which both A X200R and B Y33R were applied by the gardeners.

Query: Did Benni cause the plant to dry out?

Gold Answer: No.

In the example, there are three main causal events: 1) E_1, Alex's miscommunication about the can color; 2) E_2, Alex uses A X200R; and 3) E_3, Benni unknowingly uses B Y33R following Tom's instruction. The outcome event O is "The plants dry out." Here, $E_2 = 1 \wedge E_3 = 1$ is a cause of $O = 1$. A potential source of confusion arises from E_3: one might mistakenly judge it to be an individual cause of O because it appears to vio-

late the "norm" of Tom's instruction. However, the norm in this context is the instruction provided by Alex that A X200R is in the green can and that Benni should use the fertilizer from the green can. Since Benni follows this instruction, E_1 does not involve a norm violation. Therefore, E_1 is merely a partial cause of O. As shown in the reasoning steps, the vanilla, zero-shot CoT, and manual CoT settings exhibit similar patterns. On the one hand, Claude demonstrates the ability to identify events without explicit prompting and can also take into account factors such as intention. This indicates that the first two stages of HCR-REASONER are implicit capabilities of LLMs. However, Claude falls short in two key areas under these settings: 1) *It fails to comprehensively consider all relevant factors.* For instance, sufficiency and normality are entirely overlooked despite their importance in this example. 2) *It lacks a structured mechanism to synthesize its causal analysis into a coherent decision.* Consequently, Claude disproportionately relies on naive counterfactual dependence (i.e., f_{nc}), which undermines its reasoning quality. In contrast, with HCR-REASONER, Claude performs all key reasoning steps correctly. It first identifies the queried causal event, "Benni uses green can," then accurately infers the factor values for both conjunctive causal events: "Benni uses green can" and "Alex uses blue can." Following this, it correctly traverses Alg. 1 to reach the final decision. Specifically, "Benni uses green can" satisfies $\neg f_{\mathrm{tc}} \wedge f_{\mathrm{sc}}$, does not satisfy $f_{\mathrm{n}} \vee f_{\mathrm{i}}$, and is not assigned the uniquely highest responsibility due to occurring later than other causal events. Thus, it leads to the correct answer. Although Claude identifies three additional causal events, they are inconsequential, as all factors take on False. This indicates that these events can be safely removed from the causal setting without affecting the outcome.

Seed vs. Generated Samples. As for the second axis, we present examples comparing seed samples with generated samples to demonstrate that HCR-BENCH is both more challenging and diverse. *Through addition, new details with spurious correlations are introduced, potentially distracting models from the actual causal structure.* In Example 1, the story of the seed sample contains only events relevant to the causal setting. In contrast, the generated sample includes an additional, irrelevant event "Sophia prepares the main course and appetizers." This extraneous detail may mislead models into inferring a spurious correlation between this event and the outcome, thereby increasing the reasoning difficulty. *Through removal, important but non-essential causal cues such as explicit conjunctive statements like "When E_1 and E_2 occur, O will occur." may be omitted.* In Example 2, the seed sample explicitly states that "The machine will short circuit if both the black wire and the red wire touch the battery at the same time." In contrast, the generated sample omits this explicit conjunctive specification, instead conveying it only implicitly. Such omissions increase the reasoning complexity of the sample. *Through reorganization, the structure of the story may become more diverse.* In both Example 1 and 2, the seed sample and its corresponding generated sample differ in multiple aspects, including the story setting, the addition or removal of details, and the organization of individual sentences as well as the overall paragraph structure. As a result, the generated samples in HCR-BENCH pose greater challenges and exhibit higher diversity than their seed counterparts. This increased difficulty is also reflected in model performance: across all models, accuracy on causation queries declines in HCR-BENCH compared with in Big-Bench Hard causal judgment, and the gains attributed to HCR-REASONER are reduced accordingly.

Reasoning Steps (Claude + Vanilla)

To answer this question, we need to carefully analyze the causal chain of events and consider the role of Benni's actions in the outcome.
Let's break down the key points:

1. Tom decided to use only one fertilizer (A X200R) to avoid damaging plants with multiple chemicals.
2. Alex bought A X200R and funneled it into the blue can, but mistakenly told Benni it was in the green can.
3. Benni, believing Alex's misinformation, used the fertilizer from the green can.
4. The green can actually contained B Y33R, which Tom had filled long ago and never used.
5. The plants dried up in flower beds where both A X200R and B Y33R were applied.

Benni's action of using the fertilizer from the green can (unknowingly B Y33R) was a crucial part of the causal chain that led to the plants drying out. If Benni had not applied B Y33R, the plants would not have been exposed to both chemicals, which was the condition that caused the drying.
However, it's important to note that Benni's action was based on misinformation from Alex and was unintentional. Benni believed he was following Tom's instructions and using the correct fertilizer.
Despite the unintentional nature of Benni's action, from a strict causal perspective, Benni's application of B Y33R was a necessary condition for the harmful outcome.
Therefore, the answer is:
Yes
Benni did cause the plants to dry out, albeit unintentionally and as part of a chain of events involving misinformation and misunderstanding.

Reasoning Steps (Claude + Zero-shot CoT)

To answer this question, let's break down the key events and analyze the causal chain:
1. Tom decided to use only one fertilizer (A X200R) to avoid damaging plants with multiple chemicals. 2. Alex bought A X200R and funneled it into a blue can. 3. Alex mistakenly told Benni that he had put A X200R in the green can. 4. Benni, trusting Alex's information, used the fertilizer from the green can. 5. The green can actually contained B Y33R, which Tom had filled long ago and never used. 6. The plants dried up in flower beds where both A X200R and B Y33R were applied.
Now, let's consider Benni's role in this causal chain:

1. Benni did not intentionally use B Y33R; he thought he was using A X200R based on Alex's misinformation.
2. Benni's action of using the fertilizer from the green can was a necessary component in the causal chain that led to the plants drying out.
3. If Benni had not used the fertilizer from the green can (B Y33R), the plants would not have been exposed to both chemicals and would not have dried out.
4. While Benni's action was based on misinformation, he still played a crucial role in the outcome.
5. In causal judgment, we often consider whether an action was a necessary part of the causal chain, regardless of the actor's intentions or knowledge.

Given these considerations, we can conclude that Benni did cause the plants to dry out, even though it was unintentional and based on misinformation.
Answer: Yes
Benni's actions were a necessary part of the causal chain that led to the plants drying out, even though he was not aware of the true nature of the fertilizer he was using. In causal analysis, the unintended consequences of an action can still be attributed to the actor if their action was necessary for the outcome to occur.

Reasoning Steps (Claude + Manual CoT)

To answer this question, let's analyze the key points from the story:

1. Tom instructed both gardeners to use only A X200R fertilizer.
2. Alex bought A X200R and put it in the blue can.
3. Benni mistakenly thought A X200R was in the green can, which actually contained B Y33R.
4. The plants dried up in flower beds where both A X200R and B Y33R were applied.
5. Benni unknowingly used B Y33R from the green can.

Given these facts, we can conclude that Benni did indeed cause the plants to dry out, albeit unintentionally. Here's why:

1. Benni's use of B Y33R, combined with Alex's use of A X200R, created the condition that led to the plants drying out.
2. If Benni had not used the fertilizer from the green can (B Y33R), the plants would not have been exposed to both chemicals, which was the cause of the damage.
3. Even though Benni's actions were accidental and based on misinformation, they still contributed directly to the outcome.

Therefore, the answer is: Yes
Benni caused the plants to dry out, even though it was unintentional and based on a misunderstanding. In causal reasoning, the intent is not always necessary for establishing causation; the action and its consequences are what matter most.

Reasoning Steps (Claude + HCR-REASONER)


```
{
    "causal_events": {
        "Tom instructs to use A X200R": {
            "occur": True,
            "order": 0,
            "focal": False,
            "sufficient": False,
            "necessary": False,
            "halpern_pearl": False,
            "norm_violated": False,
            "behavior_intended": False
        },
        "Alex buys A X200R": {
            "occur": True,
            "order": 1,
            "focal": False,
            "sufficient": False,
            "necessary": False,
            "halpern_pearl": False,
            "norm_violated": False,
            "behavior_intended": False
        },
        "Alex funnels A X200R into blue can": {
            "occur": True,
            "order": 2,
            "focal": False,
            "sufficient": False,
            "necessary": False,
            "halpern_pearl": False,
            "norm_violated": False,
            "behavior_intended": False
        },
        "Alex tells Benni about green can": {
            "occur": True,
            "order": 3,
            "focal": False,
            "sufficient": False,
            "necessary": True,
            "halpern_pearl": True,
            "norm_violated": False,
            "behavior_intended": False
        },
```

```
        "Benni uses green can": {
            "occur": True,
            "order": 4,
            "focal": True,
            "sufficient": False,
            "necessary": True,
            "halpern_pearl": True,
            "norm_violated": False,
            "behavior_intended": False
        },
        "Alex uses blue can": {
            "occur": True,
            "order": 4,
            "focal": False,
            "sufficient": False,
            "necessary": True,
            "halpern_pearl": True,
            "norm_violated": False,
            "behavior_intended": False
        }
    },
    "outcome_event": {
        "Plants dry out": {
            "occur": True,
            "order": 5
        }
    }
}
```

Explanation:
No. Based on the story, “Benni uses green can” is not more responsible relative to the other events in the list. All three events (“Benni uses green can”, “Alex tells Benni about green can”, and “Alex uses blue can”) occur in sequence with “Alex tells Benni about green can” happening first (order 3), followed by both “Benni uses green can” and “Alex uses blue can” occurring simultaneously (both order 4). Since “Alex tells Benni about green can” occurs earlier, it would have more responsibility according to the temporal order criterion. Between “Benni uses green can” and “Alex uses blue can”, they have equal responsibility since they occur simultaneously (same order 4).

Example 1: Addition leads to spurious correlations.

Seed Sample.

Story: Louie is playing a game of basketball, and he made a bet with his friends who are watching on the sidelines. If Louie either makes a layup or makes a 3-point shot during the game, then he'll win $100. Just when the game started, Louie immediately got the ball at the 3-point line. He looked to the basket, dribbled in, and then made a layup right at the beginning of the game. Louie and his friends continued playing, but as hard as he tried, Louie couldn't make another shot. And then right at the end of the game as the clock was winding down, Louie got the ball at the 3-point line. He looked to the basket, focused his shot, and made a 3-point shot right at the buzzer. Then the game ended. Because Louie would win $100 if he either made a layup or a 3-point shot, Louie won $100.

Query: Did Louie win the $100 bet because he made the layup?

Reasoning:

```
{
    "causal_events": {
        "Louie makes a layup": {
            "occur": true,
            "order": 0,
            "focal": true,
            "sufficient": true,
            "necessary": false,
            "halpern_pearl": false,
            "norm_violated": false,
            "behavior_intended": true
        },
        "Louie makes a 3-point shot": {
            "occur": true,
            "order": 1,
            "focal": true,
            "sufficient": true,
            "necessary": false,
            "halpern_pearl": false,
            "norm_violated": false,
            "behavior_intended": true
        }
    },
    "outcome_event": {
        "Louie wins \$100": {
            "occur": true,
            "order": 2
        }
    }
}
```

Answer: Yes

Generated Sample.

Story: Sophia has invited friends over for a cozy dinner evening at her apartment, where she plans to showcase her cooking skills. The highlight of the night, according to her invitation, is the dessert, which she promised to prepare herself. The party will feature dessert if Sophia either bakes a cake or bakes cookies. Sophia decided to begin the day by focusing on the most intricate dessert: the cake. She spent hours meticulously measuring ingredients, frosting layers, and decorating the top with a swirl of fresh whipped cream. It turned out beautifully, and she proudly displayed it on the dining table. Feeling confident, Sophia dedicated her remaining time to preparing the main course and appetizers. As the guests began sending texts that they were on their way, Sophia realized she had just enough time to bake a quick batch of cookies. They were ready just as her first guests entered the apartment. With both options available, the party was a success, and dessert was served. Since the party would have dessert if Sophia either baked cookies or a cake, there was no shortage of delicious treats to enjoy.

Query: Did the dinner party have dessert because Sophia baked cookies?

Reasoning:


```
{
    "causal_events": {
        "Sophia bakes cookies": {
            "occur": true,
            "order": 1,
            "focal": true,
            "sufficient": true,
            "necessary": false,
            "halpern_pearl": false,
            "norm_violated": false,
            "behavior_intended": true
        },
        "Sophia bakes a cake": {
            "occur": true,
            "order": 0,
            "focal": true,
            "sufficient": true,
            "necessary": false,
            "halpern_pearl": false,
            "norm_violated": false,
            "behavior_intended": true
        }
    },
    "outcome_event": {
        "The party has dessert": {
            "occur": true,
            "order": 2
        }
    }
}
```


Answer: No

Example 2: Removal leads to fewer explicit causal cues.

Seed Sample.

Story: A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The black wire is designated as the one that is supposed to touch the battery, while the red wire is supposed to remain in some other part of the machine. One day, the black wire and the red wire both end up touching the battery at the same time. There is a short circuit.

Query: Did the black wire cause the short circuit?

Reasoning:


```
{
    "causal_events": {
        "The black wire touches the battery": {
            "occur": true,
            "order": 0,
            "focal": true,
            "sufficient": false,
            "necessary": true,
            "halpern_pearl": true,
            "norm_violated": false,
            "behavior_intended": false
        },
        "The red wire touches the battery": {
            "occur": true,
            "order": 0,
            "focal": false,
            "sufficient": false,
            "necessary": true,
            "halpern_pearl": true,
            "norm_violated": true,
            "behavior_intended": false
        }
    },
    "outcome_event": {
        "The machine short circuits": {
            "occur": true,
            "order": 1
        }
    }
}
```


Answer: No

Generated Sample.

Story: In a water management system for a large greenhouse, there are two valves controlling water supply to a central storage tank. Valve A is part of the main operational circuit, intended to manage daily water flow. Valve B, on the other hand, serves as an emergency bypass valve that should remain shut to prevent mixing excess water into the system. Under normal circumstances, if just one valve is open, the tank functions properly without incident. One morning, during a routine inspection, both Valve A and Valve B are mistakenly left open for hours, allowing water from two separate sources to flow into the tank concurrently. As a result, the tank exceeds its capacity and floods the greenhouse floor, causing damage.

Query: Did Valve A cause the overflow?

Reasoning:


```
{
    "causal_events": {
        "Valve A is open": {
            "occur": true,
            "order": 0,
            "focal": true,
            "sufficient": false,
            "necessary": true,
            "halpern_pearl": true,
            "norm_violated": false,
            "behavior_intended": false
        },
        "Valve B is open": {
            "occur": true,
            "order": 0,
            "focal": false,
            "sufficient": false,
            "necessary": true,
            "halpern_pearl": true,
            "norm_violated": true,
            "behavior_intended": false
        }
    },
    "outcome_event": {
        "The storage tank overflows": {
            "occur": true,
            "order": 1
        }
    }
}
```


Answer: No

Prompt for Baselines

Vanilla

`[SYSTEM] You are an expert in the field of token causality and causal judgment. Given the story and query of a logic-based causal judgment problem, you can effectively solve it.`
Story: {story}
Query: {query}
Answer (Yes or No?):

Zero-shot CoT

`[SYSTEM] You are an expert in the field of token causality and causal judgment. Given the story and query of a logic-based causal judgment problem, you can effectively solve it.`
Story: {story}
Query: {query}
Let's think step by step.
Answer (Yes or No?):

Manual CoT

`[SYSTEM] You are an expert in the field of token causality and causal judgment. Given the story and query of a logic-based causal judgment problem, you can effectively solve it.`
Here we will provide three chain-of-thought examplars, followed by a binary question that needs to be answered.
Story: {story1}
Query: {query1}
Answer (with chain of thought): {answer1}
Story: {story2}
Query: {query2}
Answer (with chain of thought): {answer2}
Story: {story3}
Query: {query3}
Answer (with chain of thought): {answer3}
Story: {story}
Query: {query}
Answer (Yes or No?):

Prompt for HCR-REASONER (Causal Setting Establishment)

`[SYSTEM] You are an expert in the field of token causality and causal judgment. Given the story and query of a logic-based causal judgment problem, you can effectively assist in solving the problem following the instructions provided.`
Story: {story}
Query: {query}
Based on the story and query of a logic-based causal judgment problem, establish the causal setting as follows.

1. Summarize the **causal events** and **outcome event** based on the story and query.
   - The causal events should causally contribute to the outcome event.
   - The causal events should be minimal/atomic and non-overlapping.
2. Label the **occurrences** and **temporal orders** of events based on the story.
   - Label `occur` as `true` if an event actually occurs.
   - Label `order` as an integer starting from 0, where simultaneous events should share the same `order`.
3. Label the **focal causal events** based on the query.
   - If the query asks whether a causal event causes the outcome event, the causal event has `focal = true`.
   - If the query asks whether an agent causes the outcome event, all causal events reflecting the agent's behaviors have `focal = true`.

Return the causal setting in the following JSON format:

```
{
    "causal_events": {
        CAUSAL_EVENT: {
            "occur": True/False,
            "order": ORDER,
            "focal": True/False
        },
        ...
    },
    "outcome_event": {
        OUTCOME_EVENT: {
            "occur": True/False,
            "order": ORDER
        }
    }
}
```

Return only the JSON, without any extra information.

Prompt for HCR-REASONER (Factor Value Inference)

Based on the story and causal setting, reason about the values of the following factors for each causal event.

1. `token_cause = true` if in the story, had a causal event not occurred, the outcome event would not have occurred, while allowing at least a subset of events in the causal setting to remain occurred had a causal event not occurred.
2. `necessary_cause = true` if in the story, had a causal event not occurred, the outcome event would not have occurred.
3. `sufficient_cause = true` if in the story, had a causal event occurred, the outcome event would have occurred, even if other causal events had occurred differently.

* `token_cause = true`, `necessary_cause = true`, and `sufficient_cause = true` can be satisfied through a path from a causal event to the outcome event, passing through other causal events.

4. `norm_violated = true` if in the story, a causal event violates norms, such as statistical modes, moral codes, laws, policies, or proper functioning in machines or organisms.
5. `behavior_intended = true` if in the story, a causal event is an agent's behavior and the agent is aware of the potential consequences of their action and knowingly performs an action that leads to a foreseeable bad outcome.

Return the values of factors for causal events in the following JSON format:

```
{
    CAUSAL_EVENT: {
        "token_cause": True/False,
        "necessary_cause": True/False,
        "sufficient_cause": True/False,
        "norm_violated": True/False,
        "behavior_intended": True/False
    },
    ...
}
```

Return only the JSON, without any extra information.

Prompt for HCR-REASONER (Determining f_{resp})

Line 9

Define responsibility as the relative degree (more, less, or equally) to which a causal event causally contributes to the outcome event, relative to other causal events specified. Here, assume responsibility is only determined by normality (`norm_violated`) and intention (`behavior_intended`).
Return Yes if based on the story, the focal causal event "{focal_event}" is equally or more responsible relative to other causal events in the list {S_list}, else No. Then, explain briefly based on the story.

Line 15

Define responsibility as the relative degree (more, less, or equally) to which a causal event causally contributes to the outcome event, relative to other causal events specified. Here, assume responsibility is only determined by temporal order (`order`).
Return Yes if based on the story, the focal causal event "{focal_event}" is more responsible relative to other causal events in the list {H_list}, else No. Then, explain briefly based on the story.

Prompt for HCR-BENCH (Data Generation)

Stage 1

Data Example: {data_example}
In the provided data sample: `story` is the story background of a logic-based causal judgment problem; `reasoning` details the reasoning process, including causal/outcome events and their factor values; `queries` are causality-related queries to the problem, each links to one or more focal causal events in `reasoning`; `answers` are Yes/No answers to the queries.
Generate a new data example as follows:

1. `story`: Rewrite with a different real-world setting while preserving the reasoning logic, i.e., only rephrase causal/outcome event descriptions.
2. `queries`: Formulate queries for the new story and identify the corresponding focal causal events.
3. `reasoning`: Generate causal/outcome events and their factors for the new story while keeping factor values unchanged.
4. `answers`: Provide Yes/No answers to the queries.

Return the new data example in the same JSON format as the original one.
Return only the JSON, without any extra information.

Stage 2

Based on the generated data example, refine the new story to make it more distinctive while keeping `reasoning` unchanged:

1. Increase Details: Add relevant details to enhance the new story setting.
2. Remove Details: Eliminate elements that are only relevant to the original story.
3. Reorganize Structure: Adjust the narrative flow to avoid mirroring the original structure.

Return the modified data example in the same JSON format as the original one.
Return only the JSON, without any extra information.