Assessing LLM Reasoning Through Implicit Causal Chain Discovery in Climate Discourse

Liesbeth Allein[*], **Nataly Pineda-Castañeda**[†], **Andrea Rocci**[†], **Marie-Francine Moens**[*]

[*] Department of Computer Science, KU Leuven, Belgium
[†] Institute of Argumentation, Linguistics, and Semiotics (IALS), Università della Svizzera italiana, Switzerland
liesbeth.allein@kuleuven.be

Abstract

How does a cause lead to an effect, and which intermediate causal steps explain their connection? This work scrutinizes the mechanistic causal reasoning capabilities of large language models (LLMs) to answer these questions through the task of implicit causal chain discovery. In a diagnostic evaluation framework, we instruct nine LLMs to generate all possible intermediate causal steps linking given cause-effect pairs in causal chain structures. These pairs are drawn from recent resources in argumentation studies featuring polarized discussion on climate change. Our analysis reveals that LLMs vary in the number and granularity of causal steps they produce. Although they are generally self-consistent and confident about the intermediate causal connections in the generated chains, their judgments are mainly driven by associative pattern matching rather than genuine causal reasoning. Nonetheless, human evaluations confirmed the logical coherence and integrity of the generated chains. Our baseline causal chain discovery approach, insights from our diagnostic evaluation, and benchmark dataset with causal chains lay a solid foundation for advancing future work in implicit, mechanistic causal reasoning in argumentation settings.

Keywords: mechanistic causal reasoning, causal chain discovery, climate change, argumentation

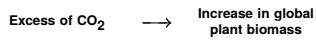
1. Introduction

"The human cognitive system is built to see causation as governing how events unfold." (Sloman and Lagnado, 2015)

Causal reasoning often hinges not on surface events, but on the underlying causal mechanisms that connect them. People are sensitive to these mechanisms when understanding, predicting, and evaluating real-world events and actions (Ahn et al., 1995; Ahn and Bailenson, 1996; Walsh and Sloman, 2011). They also perceive causal relations as stronger when structured in chains (Keshmirian et al., 2024). When faced with a causal relation like *climate change* → *flooding* (i.e., *climate change* causes *flooding*), they tend to mentally break it down into its intermediate causal steps, e.g., *climate change* → *excessive rainfall* and *excessive rainfall* → *flooding*. This process eventually produces a causal chain-like structure, *climate change* → *excessive rainfall* → *flooding* that reflects one possible pathway linking cause and effect.

In everyday discourse, causal relations are more often left implicit than other discourse relations (Olivares, 2005; Nadal and Fernández, 2019; Murray, 1997; Zunino, 2014; Zunino et al., 2011). Speakers omit or simplify the mechanisms underlying causal relations, providing just enough information for listeners to infer meaning (Grice, 1975). They rely on shared knowledge or the assumption that the causal mechanisms are self-evident. This implicitness, however, leads to diverse interpretations among listeners (Allein et al., 2025). It is therefore likely that a causal relation triggers different mental simulations of the underlying causal mechanistic pathways, with varying degrees of overlap in the intermediate events (Figure 1).

(a) **Cause-effect relation** (*cause* → *effect*) in the statement *"Additional CO_2 in the air has promoted growth in global plant biomass."*:



(b) Three generated **forward mental simulations of the causal mechanism** linking the initial cause to the final effect in causal chain-like structures:

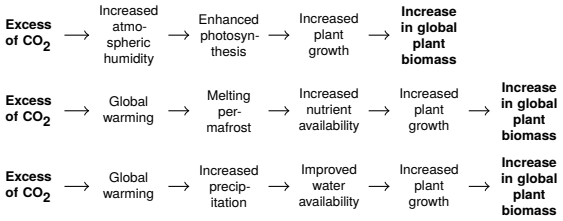


Figure 1: This paper proposes the implicit causal chain discovery task for revealing multiple intermediate causal mechanisms (b) underlying a cause-effect relation that is put forward in an argumentative setting (a). It analyzes the chains and their intermediate causal relations state-of-the-art language models reconstruct and that explain the different latent causal pathways existing between an initial cause and final effect.

And let it just be that **causality lies at the very center of argumentation in heavily-polarized topics** like climate change (Nicholson, 2014; Cook and Lewandowsky, 2016; Minnerop and Otto, 2019). Divided skeptic and believer groups attribute and accept causality in climate events differently and maintain different causal mechanistic views that support their in-group views and refute those of the out-group (Bostrom et al., 2012; Cole et al., 2025). While causality is a key driver in belief polarization regarding this topic (Cook and Lewandowsky, 2016; Bayes and Druckman, 2021), the implicitness of causality in discourse, together with the complexity of climate change systems (e.g., interdependent variables and feedback loops), greatly complicates identifying the points of division in causal thinking deeply embedded in different belief groups.

Within this context, **we consider implicit causal chain discovery a key task in argumentation**. This task involves inferring causal chain-like structures that explain the underlying causal mechanisms linking two events presented as a cause-effect (CE) pair during argumentation. In that respect, causal chains in this work reflect the *forward mental simulations of causal mechanisms* that are potentially triggered in people's minds when reasoning over a given causal relation. This cognitive framing has practical implications for argumentation studies. The ability to infer detailed causal chains is highly valuable: making unexpressed causal links explicit can reveal how speakers rely on common ground, conversational implicatures, and implicitness in discourse. It may also assist in identifying weaknesses or gaps in causal explanations (such as missing links or unsupported assumptions), and developing stronger explanations by filling the gaps with additional evidence and details. Ultimately, the ability to evaluate and refine causal explanations makes causal chain discovery crucial for argument assessment, fallacy detection, production of counterarguments, and guidance in debate and critical thinking.

This paper scrutinizes the inherent capabilities of general-purpose and reasoning large language models (LLMs) to reconstruct implicit causal chains between events. Like other causal discovery and reasoning tasks (Jin et al., 2023; Joshi et al., 2024; Jin et al., 2024), implicit causal chain discovery poses a highly complex challenge for LLMs. The frequent absence of explicit causal mechanisms is a reporting bias that affects the coverage and quality of the mechanistic causal knowledge represented in the text data used to train LLMs and build causal knowledge graphs (Gordon and Van Durme, 2013; Shwartz and Choi, 2020; Gao et al., 2023). In addition, training data on polarized topics like climate change is rich in invalid or conflicting causality. The existence of multiple causal pathways through which cause and effect can arguably be linked also adds complexity. For example, one chain may involve *excess in rainfall*, while another may include *drought* (i.e., a deficiency in rainfall) as an intermediate step. Finally, determining the appropriate granularity of causal relations, the level of abstraction (i.e., molecular or observational), and the optimal chain length is far from straightforward and highly dependent on the specific causal relation under consideration.

Contributions This paper illuminates the complexity of causal chain discovery and lays a solid foundation for advancing future work in causal chain discovery in argumentation. The zero-shot causal chain generation approach we develop acts a strong baseline, while our diagnostic evaluation with automatic and human assessment singles out valuable insights in LLM causal reasoning behavior. We also release a publicly available dataset of structurally consistent causal chains generated by multiple LLMs, offering a valuable resource for comparison and benchmarking[1].

2. Causal Chain Discovery

2.1. Task Description

Given a *CE pair* $E_c \rightarrow E_f$, where E_c and E_f are noun phrases describing an event or action, represented as sequences of natural language tokens, and the arrow $\rightarrow$ denotes the *directionality* of causality from cause to effect, the objective is to generate a set of N *causal chains* that connect E_c to E_f through a sequence of intermediate events. Note that E_c and E_f are used primarily as anchors, marking the start and end points for the causal chains.

Let ϕ be a function parameterized by an LLM and λ the template for formulating the input prompt, such that:

$$\phi(\lambda(E_c, E_f)) \rightarrow \{C^{(1)}, C^{(2)}, \ldots, C^{(N)}\}$$

where each causal chain $C^{(n)}$ is a sequence of the form:

$$E_c \rightarrow E_1^{(n)} \rightarrow E_2^{(n)} \rightarrow \cdots \rightarrow E_M^{(n)} \rightarrow E_f$$

We let the LLM infer the appropriate value of N for the given CE pair. Since each chain should be composed of at least three events, in accordance

[1]The code for reproducing the causal chain discovery experiments and evaluations, together with the generated causal chains, intermediate CE pairs, and human evaluation setup are publicly available on https://github.com/laallein/implicit-causal-chain-discovery.

with the definition of a causal chain (Pearl, 2009), $M > 0$. The value of M may vary across chains. Each intermediate event $E_m^{(n)}$ represents a distinct step in the causal progression from E_c to E_f. For the sake of brevity, we continue to refer to any event in a chain as E_t, with t its position in the chain and T the total length of the chain. Finally, each link in the chain, e.g., $E_t \rightarrow E_{t+1}$, is treated as an *intermediate CE pair* that can be independently analyzed and validated.

Note that we do not treat the Markov property (Pearl, 2009), i.e., the conditional independence of non-adjacent events, as a necessary constraint when inferring causal chains in an argumentation context. In practice, human causal reasoning often violates the Markov property, as people tend to consider the broader chain of events rather than restrict themselves to strictly local dependencies (Rottman and Hastie, 2016).

2.2. Data

We rely on annotated CE pairs from PolarIs3CAUS (Pineda and Allein, 2025a) (95 pairs) and PolarIs4CAUS (Pineda and Allein, 2025b) (181 pairs), which were manually extracted from English discussions on climate change in the PolarIS-3 (Reddit) and PolarIS-4 (X) datasets (Gajewska et al., 2024).

Despite their modest number of CE pairs, PolarIs3CAUS and PolarIs4CAUS are particularly well-suited for causal chain discovery in this work and preferred over other existing resources (Mihăilă et al., 2013; Mirza et al., 2014; Mostafazadeh et al., 2016; Dunietz et al., 2017; Tan et al., 2022; Romanou et al., 2023; Tan et al., 2023).

First, the CE annotations are high-quality, structurally consistent, and easy to understand (Table 1). The CE pairs were manually annotated by argumentation experts as part of an ongoing argumentation-focused study on polarization. They also standardized the pairs through grammatical transformations and disambiguation procedures to ensure clarity outside their original context (Pineda and Allein, 2025a,b). Second, the CE pairs appear in realistic argumentative discourse, where causal explanations are typically challenged or advanced as part of broader arguments. While they may not constitute full arguments, they may have argumentative or explanatory potential in broader disputes or dialogues. Moreover, this data offers insights into how causal reasoning operates in adversarial or deliberative contexts as it was extracted from two polarized belief groups (i.e., climate change believers and skeptics). Last, we expect that PolarIs3CAUS and PolarIs4CAUS are not included in the training data of the assessed LLMs as they were published after the data cut-off date of most LLMs.

Message	Cause		Effect
Ocean acidification (pH inverse) is accelerating a direct result of CO2 emissions.	CO2 emissions	→	Ocean acidification
"Business-as-usual" is leading to #ClimateChange, #biodiversity loss and the degradation of #nature. To reverse this trend, we need to build a global #CircularEconomy.	Business Business Business	→ → →	Climate change Biodiversity loss Degradation of nature
Deforestation caused so much nutrient loss in the soils, eucalyptus is the only thing that has been able to successfully grow there.	Deforestation Nutrient loss in the soils	→ →	Nutrient loss in the soils Monoculture

Table 1: These three examples from the PolarIs4CAUS dataset with annotations of direct causal relations showcase the argumentative context of the causal relations (i.e., *message*) and the structurally consistent formulation of cause and effect as noun phrases.

2.3. Language Models

We use a broad set of open-source and proprietary LLMs for parameterizing ϕ. **General-purpose LLMs** generate direct answers without showing intermediate reasoning steps: GPT4o (Hurst et al., 2024), Llama 3 70b (Grattafiori et al., 2024), Mistral Nemo (mis, 2024), Llama 3.1 Nemotron (Bercovich et al., 2025), Phi 4-mini (Abouelenin et al., 2025), Mixtral (Jiang et al., 2024). **Reasoning LLMs** decompose complex problems into smaller problems and solve them in a step-by-step manner: o1, o1-mini (Jaech et al., 2024), DeepSeek R1 (Guo et al., 2025).

2.4. Methodology

We implement a zero-shot prompting approach and instruct the LLM to generate all possible causal chains in a single step. We deliberately avoid a few-shot approach to prevent introducing biases related to the number of chains (i.e., N), the number of intermediate causal events in a chain (i.e., M), or the level of detail in each event. We do not use chain-of-though prompting where we explicitly encourage step-by-step thinking, although reasoning LLMs are expected to do this as part of their internal generation system. We also do not provide additional context through retrieval-augmented generation (RAG) as our goal is to assess the ability of LLMs to effectively draw on the causal knowledge encoded in their parameters rather than to evaluate their retrieval ability and use of external resources such as knowledge graphs (e.g., Heindorf et al. (2020); Li et al. (2021)).

The input prompt λ contains a causal chain definition, a task description with two slots for inserting E_c and E_f, and formatting instructions (Table

Task	Prompts
Causal Chain Discovery [λ]	A causal chain is a sequence of events in which each event directly causes the next, forming a connected series of cause-and-effect relations. Unfolding a causal chain means identifying and linking individual events. A step of the chain presents only one noun phrase containing the event. Unfold all possible causal chains that connect $\{E_c\}$ (initial cause) to $\{E_f\}$ (final effect) and separate the steps of the chain with the token <step>, and the chains with the token <chain>.
Self-Consistency and Confidence [λ_{A1}]	Answer with 'yes' or 'no' only. Does $\{E_t\}$ cause $\{E_{t+1}\}$?
Directionality of Causality [λ_{A2}]	Answer with 'yes' or 'no' only. Does $\{E_{t+1}\}$ cause $\{E_t\}$?
Position Heuristics	[λ_{A1-P}] Answer with 'yes' or 'no' only. Is $\{E_{t+1}\}$ caused by $\{E_t\}$? [λ_{A2-P}] Answer with 'yes' or 'no' only. Is $\{E_t\}$ caused by $\{E_{t+1}\}$?

Table 2: Overview of the template prompts λ used in the causal chain discovery task and the evaluation setups. The curled brackets ({}) present input slots, which can be filled with a CE pair from the dataset, i.e., E_c and E_f, or two consecutive events in a generated chain, i.e., E_t and E_{t+1}.

2). The prompt is designed to be domain-agnostic. Although directed acyclic graphs (DAGs) are the default representation of causal structures (Pearl, 2009), we refrain from instructing LLMs to produce DAG outputs. This is because symbolic reasoning capabilities differ between LLMs (Tang et al., 2023), which we do not evaluate here. Importantly, the structurally consistent formulation of causes and effects in PolarIs3CAUS and PolarIs4CAUS allows us to impose grammatical formatting constraints, which improves consistency in the generation of intermediate events within the causal chains.

The prompt was iteratively refined through intensive manual prompt engineering. We removed ambiguities and non-essential information, with particular attention to the prompt's transferability across LLMs. The formatting instructions required multiple rounds of revision as the majority of the LLMs failed to follow them consistently. Nonetheless, the output from most LLMs required additional post-processing to parse the generated causal chains (e.g., removing intro).

3. General Results

The results in Table 3 show that the LLMs exhibit distinct causal discovery behavior. One key difference lies in their *productivity*: models vary in the number of causal chains they generate for a given CE pair (i.e., N). However, this variation does not align with model type as there is no consistent distinction between general-purpose LLMs and reasoning LLMs. Chain length (i.e., T), which reflects the depth of reasoning or *verbosity* of a model, also does not clearly differentiate general-purpose LLMs from reasoning LLMs.

Interestingly, there is a statistically significant negative correlation ($p < .01$) between the number of chains generated for a CE pair and their average length, with Pearson's r values ranging from $-.11$ to $-.42$ across models. In other words, as more chains are generated for a given CE pair, their average length tends to decrease. This finding is somewhat counter-intuitive as we would expect that generating more chains would lead to greater elaboration and detail, resulting in longer chains rather than shorter ones.

4. Diagnostic Evaluation

We continue to evaluate the generated causal chains, first *at the level of the intermediate CE pairs in a chain* (i.e., $E_t \rightarrow E_{t+1}$), then *at the level of the full chain* (i.e., C). We design the evaluation setups such that we probe the inherent causal reasoning capabilities of LLMs in a structured, straightforward, and diagnostic manner. This way, we are able to identify specific points of success and failure in causal reasoning. Our setup substantially differs from previous work, which assessed causal reasoning capabilities by having LLMs answer questions about explicit causal relations in text (Chi et al., 2024), compare causal relations (Ashwani et al., 2024; Liu et al., 2024), build and reason over causal graphs, usually in DAG structures, from causal relations explicitly mentioned in the input (Jin et al., 2023; Chen et al., 2024; Joshi et al., 2024; Wang, 2024).

Due to the high number of experiments and computing cost, we report the results from a single inference pass per model and evaluation configuration.

4.1. Inside Causal Chains: Intermediate CE pairs

We obtain the intermediate CE pairs by splitting all generated causal chains in their intermediate CE pairs. For instance, let $C = E_0 \rightarrow E_1 \rightarrow E_2 \rightarrow E_3$, then the three intermediate causal relations are $E_0 \rightarrow E_1$, $E_1 \rightarrow E_2$, and $E_2 \rightarrow E_3$. In total, we assess 83,688 intermediate CE pairs that are part of 16,230 causal chains generated for the 276 original CE pairs from the two datasets. Detailed statistics on the intermediate CE pairs can be found in Table 3.

	O1	O1-mini	GPT4o	DeepSeek R1	Llama 3.1 Nemo	Llama 3 70b	Mistral Nemo	Mixtral	Phi 4-mini
# Chains									
Total	**865**	**478**	**485**	**581**	**716**	**500**	**363**	**575**	**346**
Per CE									
– mean	9.2	5.03	5.11	6.12	7.54	5.26	3.82	6.05	4.12
– std	4.58	1.96	2.27	1.69	3.26	1.24	1.00	2.36	3.49
– min	3	2	1	4	2	3	1	2	1
– max	25	10	17	10	19	8	7	11	19
# Intermediate CE Pairs									
Total	**3,667**	**1,932**	**2,412**	**2,900**	**3,065**	**2,421**	**1,922**	**3,047**	**2,720**
Per chain									
–mean	4.24	3.99	4.84	4.99	4.28	4.84	5.29	5.29	7.86
–std	1.21	1.6	1.73	1.52	1.23	1.44	1.84	1.83	20.27
–min	2	2	2	2	2	2	2	2	2
–max	11	24	21	15	9	11	14	12	366
Correlation # chains and # pairs per chain (= chain length)									
Pearson's r	-.19	-.18	-.16	-.33	-.21	-.34	-.39	-.28	/

	O1	O1-mini	GPT4o	DeepSeek R1	Llama 3.1 Nemo	Llama 3 70b	Mistral Nemo	Mixtral	Phi 4-mini
# Chains									
Total	**1,689**	**932**	**874**	**1,163**	**3,714**	**943**	**744**	**1,216**	**911**
Per CE									
– mean	9.33	5.15	4.86	6.43	20.52	5.21	4.13	6.72	5.03
– std	4.67	1.5	1.59	1.97	34.12	.95	.94	2.89	6.86
– min	2	1	1	3	3	4	2	3	1
– max	29	10	10	11	173	8	7	28	68
# Intermediate CE Pairs									
Total	**7,308**	**3,610**	**4,026**	**5,606**	**17,926**	**4,336**	**4,015**	**6,343**	**10,099**
Per chain									
–mean	4.33	3.81	4.58	4.82	4.82	4.60	5.40	5.22	11.06
–std	1.01	1.57	1.20	1.20	1.56	1.06	2.06	1.53	47.51
–min	2	2	2	2	2	2	2	2	2
–max	10	7	12	10	16	9	22	12	523
Correlation # chains and # pairs per chain (= chain length)									
Pearson's r	/	-.12	/	-.42	-.37	-.11	-.29	/	/

Table 3: Statistics on the generated chains for CEs from PolarIs3CAUS (left) and PolarIs4CAUS (right): total/per-CE chain counts, intermediate CE pairs per chain/total, and significant correlations between chain count and length (Pearson's r, $p < .01$).

4.1.1. Self-Consistency and Confidence in Causality (A1)

We evaluate the *self-consistency* and *confidence* of the LLMs regarding the intermediate CE pairs they generated during implicit causal chain discovery. We instruct the LLM to classify the relation of each intermediate CE pair as causal or non-causal by answering a yes/no question (*'yes'* = causal, *'no'* = non-causal) using the input prompt $\lambda_{A1}(E_t, E_{t+1})$ (Table 2). A self-consistent and confident model should classify all intermediate CE pairs as causal.

– Results The majority of pairs are classified as causal, with minimal differences between general-purpose and reasoning LLMs. This suggests that the LLMs understood the causal chain discovery task well. They also demonstrate self-consistency in their reasoning and confidence regarding the causality between events. Furthermore, the results indicate that the quality and informativeness of the input prompt are sufficient, and that the prompt is transferable across different LLMs. Accordingly, we argue that our zero-shot prompting framework is suitable as a baseline for future causal chain discovery methods.

4.1.2. Directionality of Causality (A2)

A fundamental property of a causal chain is that its intermediate links exhibit a correct directional relationship, i.e., each event E_t should be the cause of the subsequent event E_{t+1}; $E_t \rightarrow E_{t+1}$. To assess whether LLMs understand this property, we examine their understanding of *temporal succession*, which is a key criterion of causation which asserts that causes should always precede their effects in time (Hume, 2000; Jonassen and Ionas, 2008; Grzymala-Busse, 2011). Consequently, a reversed relation, i.e., $E_{t+1} \rightarrow E_t$, violates this criterion and should be considered non-causal.

Building on the experimental setup in A1, we present the LLMs with reversed CE pairs and instruct them to classify each as causal or non-causal using the input prompt $\lambda_{A2}(E_t, E_{t+1})$ (Table 2). An LLM that has internalized the directionality of causal relations should consistently classify each reversed pair as non-causal.

– Results Around 50% of the reversed CE pairs are still judged as causal, indicating that the models struggle to understand the directionality of causality. However, it is important to note that each CE pair is evaluated in isolation. The lack of context and minimal refinement of the events can result in pairs being considered causal in both the original and reversed order. This outcome is reasonable in some real-world settings. For example, *harm to trees* $\rightarrow$ *pollution* is valid and is explained through the chain *harm to trees* $\rightarrow$ *loss of forests* $\rightarrow$ *soil erosion* $\rightarrow$ *release of sediments and pollutants into water bodies* $\rightarrow$ *water pollution*[2]. Conversely, *pollution* $\rightarrow$ *harm to trees* is also valid and is motivated by *air pollution* $\rightarrow$ *tree damage* $\rightarrow$ *smaller trees and increased tree growth* $\rightarrow$ *harm to trees*[3]. The topic of climate change also contributes to the complexity as certain climate-related events can feed back into themselves, creating cyclic causality. For instance, the Ice-Albedo feedback is a cyclic process where melting ice reduces surface reflec-

[2] https://www.emission-index.com/deforestation/water-pollution

[3] https://www.nps.gov/articles/000/how-to-assess-air-pollutions-impacts-on-forests.htm

tivity (albedo), causing more solar energy to be absorbed, which leads to further warming and more ice melt. This stresses the importance of generating multiple causal chains for explaining causal mechanisms connecting events.

4.1.3. Fallacy in Causal Reasoning: Position Heuristics (A3)

LLMs have been shown to have internalized causal patterns during training (e.g., the event representing the cause is more frequently positioned before the event representing the effect in a text), which they reuse to answer causal questions quickly and efficiently without performing explicit causal inference (Jin et al., 2023; Zečević et al., 2023).

We evaluate whether LLMs maintain consistent causal judgments when the surface form of prompts is altered but the semantic content is maintained by switching from active (*"Does E_t cause E_{t+1}?"*) to passive voice (*"Is E_{t+1} caused by E_t?"*). We instruct them to classify each $E_t \rightarrow E_{t+1}$ as causal or non-causal using the passive formulation of λ_{A1}, i.e., $\lambda_{A1-P}(E_t, E_{t+1})$. We repeat this, but now for the reversed pair $E_{t+1} \rightarrow E_t$ using the passive formulation of λ_{A2}, i.e., $\lambda_{A2-P}(E_t, E_{t+1})$ using input prompt. See Table 2 for λ_{A1-P} and λ_{A2-P}. LLMs that do not resort to *position heuristics* (Joshi et al., 2024) and *amortized causal reasoning* should consistently give the same answers to prompt pairs λ_{A1} and λ_{A1-P}, and to λ_{A2} and λ_{A2-P}.

Consistency is measured using two metrics: Jaccard dissimilarity, which measures the dissimilarity between sets by normalizing the overlap of *'yes'* responses and subtracting the result from one, and Hamming distance, which captures the proportion of differing answers across prompt pairs. Low scores on both metrics indicate strong alignment between active and passive prompts, suggesting a robust understanding of the underlying causal relations.

– Results For λ_{A1} and λ_{A1-P}, disagreement between active and passive formulations is relatively low, with Jaccard dissimilarity and Hamming distance values ranging from 0.09 to 0.24 for both PolarIs3CAUS and PolarIs4CAUS. This suggests that the models exhibit a reasonable degree of consistency in their causal judgments, even when the linguistic framing is altered. However, for λ_{A2} and λ_{A2-P}, disagreement increases substantially, with values ranging from 0.25 to 0.51. This indicates a marked drop in consistency when evaluating reversed, incorrect causal relations. The elevated disagreement highlights the susceptibility of LLMs to position heuristics.

These results underscore a key limitation in current LLMs. Even those designed to perform well on reasoning tasks struggle to maintain stable causal inferences when superficial aspects of the prompt are modified. This prompt-sensitivity becomes especially problematic in contexts with strong polarization. LLMs would be susceptible to framing effects (Feldman and Hart, 2018), inferring different causal chains depending on how a question is framed and potentially reinforcing existing ideological biases. Moreover, if an LLM cannot maintain consistent causal judgments across prompt variants, it may produce contradictory explanations for a single CE pair.

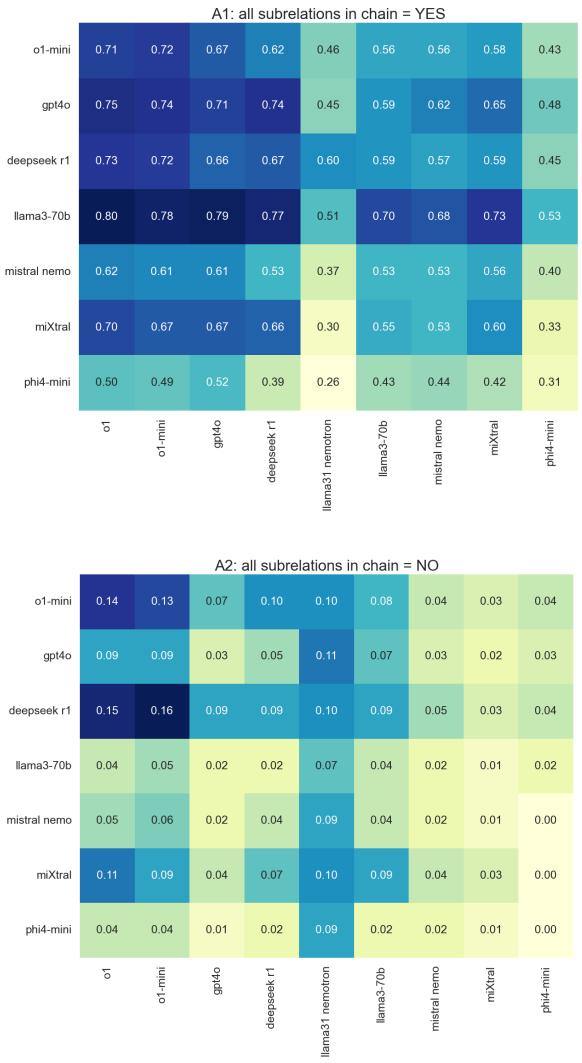
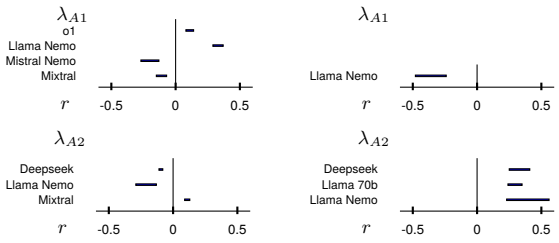


Figure 2: Cross-model evaluation of the integrity of the chains generated for the CE pairs from PolarIs4CAUS. The results for the A1 setup are shown on top, those for the A2 setup below. In both figures, the LLMs that generated the chains are on the x-axis, the LLMs that evaluate those chains are on the y-axis, and the values report the proportion of chains that are considered valid by the LLM.



(a) Chain length and % valid intermediate CE pairs.

(b) Number of chains and % valid intermediate CE pairs.

Figure 3: Pearson correlation between chain length and the proportion (%) of *causal* intermediate CE pairs (a), and between the number of chains per CE pair and the proportion (%) of *causal* intermediate CE pairs (b). The y-axis includes models with significant correlation in their generations ($p < .01$). The x-axis shows the range of r values. Causality of pairs is evaluated using λ_{A1} and λ_{A2}.

4.2. On Causal Chains

4.2.1. Self- and Cross-Model Assessment of Causal Chain Integrity (A4)

We evaluate the *integrity* of the generated causal chains. A chain's integrity is preserved only if all of its intermediate links are causal, i.e., answer to $\lambda_{A1}(E_t, E_{t+1})$ = *'yes'* and $\lambda_{A2}(E_t, E_{t+1})$ = *'no'*. It is violated if at least one link is non-causal. We let LLMs evaluate their own generations and of those others.

– Results Figure 2 shows the proportion of causal chains generated by an LLM with preserved integrity as assessed by the model itself or another. Reasoning models seem to produce a higher proportion of valid chains than general-purpose LLMs, except for GPT4o. This suggests that reasoning models are overall better at identifying clear causal relations.

We further examine the correlation between chain length and the proportion of intermediate CE pairs classified as causal through λ_{A1} and λ_{A2}. Xiong et al. (2022) showed that performance on causal discovery tasks tends to decline with longer chains as they require more complex reasoning. We therefore hypothesize a negative correlation. However, as shown in Figure 3, correlation is not consistently negative across models. Chain length seems to affect the quality of causal discovery, though the direction and strength of this effect vary across models.

Extending this analysis, we investigate whether generating more chains for a given CE pair dilutes quality. For this, we assess the correlation between the number of chains generated for a CE pair and the proportion of intermediate CE pairs in those chains deemed causal by the LLMs (again using λ_{A1} and λ_{A2}). The results in Figure 3 show that the impact of chain quantity on quality differs across model as some improve with more chains while others worsen.

Although assessing the integrity and quality of a chain by the validity of its intermediate CE pairs provides useful insights, we note that it does not fully account for certain complexities specific to causal chains, such as transitive inference, scene drift, and threshold effects (Xiong et al., 2022). For example, the following causal sequence highlights the subtlety of threshold effects: *cold* $\rightarrow$ *vasoconstriction* $\rightarrow$ *increased blood pressure* $\rightarrow$ *stroke* $\rightarrow$ *death*.

4.2.2. Human Assessment of Causal Chain Validity and Logical Coherence (A5)

The last evaluation setup includes a focused human evaluation with argumentation experts. We explore whether their judgments on the integrity of generated causal chains aligns with those of the LLMs. They also assess the logical coherence of the generated causal chains. We end by surveying them about the complexity of evaluating causal chains to gain insights into the feasibility of obtaining higher-quality chains from non-expert annotators, i.e., people who do not have extensive background on the topic of the causes and effect (in this study, climate change)[4].

Sample selection The human evaluation focuses on a curated selection of 36 causal chains for 18 CE pairs from Polarls4CAUS; two chains per CE. All chains have been generated by o1, which is considered the best-performing model according to the cross-model evaluations in A4. We select the most agreed-upon *maintained* and *violated* chain per CE pair in terms of integrity, making sure that their lengths are approximately equal. Table 4 shows two sets of chain pairs.

Evaluation setup The human evaluation took place in a controlled, in-person setting with 10 master's and PhD students from a Communication faculty. Each chain was independently judged by four participants. Participation was voluntary, and while students were informed the study concerned causal chain reasoning, they were not told the chains were LLM-generated to avoid bias. After receiving task instructions, which included a definition of a causal chain and two illustrative examples, they annotated six pairs of chains, labeled randomly as Chain A and Chain B. For each chain, they judged whether it

[4]Details on chain selection procedure and screenshots of annotation forms will be included in the Supplementary Material.

	LLM Eval	Human Eval	
	Integrity	Integrity	Coherence
rich countries → high waste generation → methane emissions from landfills → greenhouse gas accumulation → **climate change**	✓✓✓✓✗✗✗	✓✓✓✓	✓✓✓✓
rich countries → industrialization → fossil fuel combustion → enhanced greenhouse effect → rich countries → overconsumption → resource depletion → environmental degradation → **climate change**	✓✗✗✗✗✗✗	✓✓✗✗	✓✓✗✗
climate change → earlier seasonal changes → mismatched reproductive cycles → population decline → **species at risk of extinction**	✓✓✓✗✗✗✗	✓✓✓✗	✓✓✓✗
climate change → ocean acidification → death of shell-forming organisms → disruption of marine food webs → collapse of fish populations → **species at risk of extinction**	✓✗✗✗✗✗✗	✓✓✓✓	✓✓✓✓

Table 4: Examples of chain sets with agreement (top) and disagreement (bottom) between LLMs and argumentation experts.

maintained *chain integrity* and was *logically coherent*, defined as a plausible sequence where each step follows naturally from the previous one. At the end, participants reflected on their confidence and rated the difficulty of the annotation task on a 5-point Likert scale. They were also asked whether they believe they can construct a valid, coherent chain on climate change. If yes, they described how their chain would compare in length and detail to those evaluated; if no, they were asked to explain why.

– Results Based on the majority votes for integrity and logical coherence, the quality of the generated causal chains is rated relatively high: participants confirmed the integrity of 27 out of 36 chains and marked 24 as coherent. However, LLM judgments align poorly with human assessments as they disagreed on the integrity for half of the chains. Despite the cognitive demands of the task and low inter-annotator agreement (Fleiss' $\kappa = .084$ for integrity; $\kappa = .035$ for coherence), participants expressed confidence in their ability to construct valid and coherent causal chains. Nonetheless, clear instructions on the desired depth and detail are are essential for consistency as some participants anticipated producing shorter and simpler chains, while others expected theirs to be longer and more detailed.

5. Related Work

Unraveling complex causal structures beyond direct causal relations is central to advancing the modeling and analysis of real-world commonsense causality (Cui et al., 2024). Prior work has built causal graphs between CE pairs, for instance, using causal knowledge graphs to retrieve relevant events (Du et al., 2021; Xu and Ichise, 2024). Specifically on causal chains, Xiong et al. (2022) proposed a framework to estimate the reliability of short causal chains, addressing transitive problems such threshold effect and scene drift. Du et al. (2022) explained causal relations using single conceptual explanations. Closest to our work, Ho et al. (2023) generated multi-hop causal chains to explain causal relations sourced from Wikipedia. However, their event representation is inconsistent, mixing full sentences and noun phrases, and they produce only one chain per relation, limiting coverage of alternative pathways. In contrast, this paper investigates the mechanistic causal reasoning capabilities of LLMs by evaluating their ability to generate multiple implicit causal chains that explain the causal relations between events in climate discourse.

Regarding the modeling of polarization in discourse, argumentation mining, including framing analysis and stance detection, has proven effective (Bianchi et al., 2021; Sinno et al., 2022; Hofmann et al., 2022; Irani et al., 2024). Within argumentation, causality has been primarily studied in three main contexts: (i) discovery of explicit causal relations between events in text (Al-Khatib et al., 2023; Tan et al., 2023), (ii) identification of causal relations between arguments (Jo et al., 2021; Saadat-Yazdi et al., 2023; Bezou-Vrakatseli et al., 2025), and (iii) causal analysis of actions in persuasive conversations (Zeng et al., 2025).

6. Conclusion

Implicit causal chain discovery is key for studying common ground in causal thinking. By surfacing the underlying reasoning behind opposing views, such discovery may help individuals recognize how others arrive at their conclusions and identify areas of agreement. It also reveals overgeneralizations, false equivalences, and manipulative rhetoric rooted in faulty causal logic. This work sets a baseline for causal chain discovery that is transferable across LLMs, but also produces chains whose integrity and coherence are confirmed by argumentation experts through human evaluation. Our diagnostic evaluation of LLM behavior showed that LLMs are self-consistent and confident about the causality between events. However, they are vulnerable to minor changes in causal formulations, highlighting a reliance on position heuristics and a lack of robustness.

Important challenges remain. Although our evaluation design is domain-agnostic, we expect different causal discovery behavior and quality across

domains. Some domains involve more complex or implicit causal mechanisms, reducing the coverage of relevant knowledge in the training data. In domains where causality is heavily contested, LLMs and causal knowledge graphs are more likely to encode conflicting or invalid causal associations, which further complicates the evaluation of causal validity. While RAG is a promising follow-up baseline, it may struggle to retrieve, align, and evaluate causal relations effectively due to subtle mismatches in the linguistic formulation of CE pairs.

Acknowledgements

This work has been funded by the Research Foundation - Flanders (FWO) under grant G0L0822N (Liesbeth Allein, Marie-Francine Moens) and the Swiss National Science Foundation (SNSF) under grant 209674 (Nataly Pineda-Castañeda, Andrea Rocci) through the CHIST-ERA project "iTRUST Interventions against Polarisation in Society for Trustworthy Social Media. From Diagnosis to Therapy".

Bibliographical References

2024. Mistral nemo.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

W Ahn and J Bailenson. 1996. Mechanism-based explanations of causal attribution: An explanation of conjunction and discounting effect. *Cognitive Psychology*, 31:82–123.

Woo-kyoung Ahn, Charles W Kalish, Douglas L Medin, and Susan A Gelman. 1995. The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3):299–352.

Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Anh Le, Martin Potthast, and Benno Stein. 2023. A new dataset for causality identification in argumentative texts. In *24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 349–354. Association for Computational Linguistics, ACL Anthology.

Liesbeth Allein, Maria Mihaela Truşcă, and Marie-Francine Moens. 2025. Interpretation modeling: Social grounding of sentences by reasoning over their implicit moral judgments. *Artificial Intelligence*, 338:104234.

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.

Robin Bayes and James N Druckman. 2021. Motivated reasoning and climate change. *Current Opinion in Behavioral Sciences*, 42:27–35.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. 2025. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2025. Can large language models understand argument schemes? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13666–13681, Vienna, Austria. Association for Computational Linguistics.

Federico Bianchi, Marco Marelli, Paolo Nicoli, and Matteo Palmonari. 2021. SWEAT: Scoring polarization of topics across different corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10065–10072, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ann Bostrom, Robert E O'Connor, Gisela Böhm, Daniel Hanss, Otto Bodi, Frida Ekström, Pradipta Halder, Sven Jeschke, Birgit Mack, Mei Qu, et al. 2012. Causal thinking and support for climate change policies: International survey findings. *Global Environmental Change*, 22(1):210–222.

Sirui Chen, Mengying Xu, Kun Wang, Xingyu Zeng, Rui Zhao, Shengjie Zhao, and Chaochao Lu. 2024. CLEAR: Can language models really understand causal graphs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6247–6265, Miami, Florida, USA. Association for Computational Linguistics.

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? *Advances in Neural Information Processing Systems*, 37:96640–96670.

Jennifer C Cole, Ash J Gillis, Sander van der Linden, Mark A Cohen, and Michael P Vandenbergh. 2025. Social psychological perspectives on political polarization: Insights and implications for

climate change. *Perspectives on Psychological Science*, 20(1):115–141.

John Cook and Stephan Lewandowsky. 2016. Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in cognitive science*, 8(1):160–179.

Shaobo Cui, Zhijing Jin, Bernhard Schölkopf, and Boi Faltings. 2024. The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16722–16763, Miami, Florida, USA. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. ExCAR: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2354–2363, Online. Association for Computational Linguistics.

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.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Lauren Feldman and P Sol Hart. 2018. Climate change as a polarizing cue: Framing effects on public support for low-carbon energy policies. *Global Environmental Change*, 51:54–66.

Ewelina Gajewska, Katarzyna Budzynska, Barbara Konat, Marcin Koszowy, Konrad Kiljan, Maciej Uberna, and He Zhang. 2024. Ethos and pathos in online group discussions: Corpora for polarisation issues in social media. *arXiv preprint arXiv:2404.04889*.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

H. Paul Grice. 1975. Logic and conversation. In Donald Davidson, editor, *The Logic of Grammar*, pages 64–75. Dickenson Pub. Co.

Anna Grzymala-Busse. 2011. Time will tell? temporality and the analysis of causal mechanisms and processes. *Comparative Political Studies*, 44(9):1267–1297.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3023–3030.

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2023. Wikiwhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*.

Valentin Hofmann, Xiaowen Dong, Janet Pierrehumbert, and Hinrich Schuetze. 2022. Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536–550, Seattle, United States. Association for Computational Linguistics.

David Hume. 2000. *A treatise of human nature*. Oxford University Press.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Arman Irani, Michalis Faloutsos, and Kevin Esterling. 2024. Argusense: Argument-centric analysis of online discourse. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 663–675.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065.

Zhijing Jin, Jiarui Liu, LYU Zhiheng, 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*.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

David H Jonassen and Ioan Gelu Ionas. 2008. Designing effective supports for causal reasoning. *Educational Technology Research and Development*, 56:287–308.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. LLMs are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10553–10569, Miami, Florida, USA. Association for Computational Linguistics.

Anita Keshmirian, Moritz Willig, Babak Hemmatian, Kristian Kersting, Ulrike Hahn, and Tobias Gerstenberg. 2024. Chain versus common cause: Biased causal strength judgments in humans and large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2021. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

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.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC bioinformatics*, 14:1–18.

Petra Minnerop and Friederike Otto. 2019. Climate change and causation: joining law and climate science on the basis of formal logic. *Buff. Envtl. LJ*, 27:49.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

John D. Murray. 1997. Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25(2):227–236.

Laura Nadal and Inés Recio Fernández. 2019. Processing implicit and explicit causality in spanish. In *Pragmatics & Beyond New Series*. John Benjamins. Published 6 August 2019.

Calum TM Nicholson. 2014. Climate change and the politics of causal reasoning: the case of climate change and migration. *The Geographical Journal*, 180(2):151–160.

María Soledad Carbonell Olivares. 2005. *Estudio semántico-pragmático de las relaciones de contraste y sus marcas en lengua inglesa*. Phd dissertation, Universitat de València.

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

Nataly Pineda and Liesbeth Allein. 2025a. Polaris3caus (version 1.0) [dataset].

Nataly Pineda and Liesbeth Allein. 2025b. Polaris4caus (version 1.0) [dataset].

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore. Association for Computational Linguistics.

Benjamin M Rottman and Reid Hastie. 2016. Do people reason rationally about causally related events? markov violations, weak inferences, and failures of explaining away. *Cognitive psychology*, 87:88–134.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kokciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2484–2495, Dubrovnik, Croatia. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022. Political ideology and polarization: A multidimensional approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 231–243, Seattle, United States. Association for Computational Linguistics.

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

Fiona Anting Tan, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, Tiancheng Hu, et al. 2022. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See-Kiong Ng. 2023. RECESS: Resource for extracting cause, effect, and signal spans. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–82, Nusa Dua, Bali. Association for Computational Linguistics.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*.

Clare R Walsh and Steven A Sloman. 2011. The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, 26(1):21–52.

Zeyu Wang. 2024. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151, Bangkok, Thailand. Association for Computational Linguistics.

Kai Xiong, Xiao Ding, Zhongyang Li, Li Du, Ting Liu, Bing Qin, Yi Zheng, and Baoxing Huai. 2022. ReCo: Reliable causal chain reasoning via structural causal recurrent neural networks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6426–6438, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ziwei Xu and Ryutaro Ichise. 2024. Exploring causal chain identification: Comprehensive insights from text and knowledge graphs. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 129–146. Springer.

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*.

Donghuo Zeng, Roberto Legaspi, Yuewen Sun, Xinshuai Dong, Kazushi Ikeda, Peter Spirtes, and Kun Zhang. 2025. Causal discovery and counterfactual reasoning to optimize persuasive dialogue policies. *Behaviour & Information Technology*, pages 1–15.

Gabriela Zunino. 2014. Cognitive perspectives on discourse processing: Causality and countercausality. *Signos Lingüísticos*, 10:154–171.

Gabriela Zunino, Valeria Abusamra, and Alexander Raiter. 2011. Articulación entre conocimiento del

mundo y conocimiento lingüístico en la comprensión de relaciones causales y contracausales: el papel de las partículas conectivas. *Forma y Función*, 25:15–34.