ClimateCause: Complex and Implicit Causal Structures in Climate Reports

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

[♢]Department of Computer Science, KU Leuven, Belgium
[†]Department of Electronics and Information Systems, Ghent University, Belgium
[‡]Institute of Argumentation, Linguistics, and Semiotics (IALS),
Università della Svizzera italiana, Switzerland

Abstract

Understanding climate change requires reasoning over complex causal networks. Yet, existing causal discovery datasets predominantly capture explicit, direct causal relations. We introduce ClimateCause, a manually expert-annotated dataset of higher-order causal structures from science-for-policy climate reports, including implicit and nested causality. Cause-effect expressions are normalized and disentangled into individual causal relations to facilitate graph construction, with unique annotations for cause-effect correlation, relation type, and spatiotemporal context. We further demonstrate ClimateCause's value for quantifying readability based on the semantic complexity of causal graphs underlying a statement. Finally, large language model benchmarking on correlation inference and causal chain reasoning highlights the latter as a key challenge.

Statement: *"Climate change has reduced food security and affected water security due to warming, changing precipitation patterns, reduction and loss of cryospheric elements, and greater frequency and intensity of climatic extremes, thereby hindering efforts to meet Sustainable Development Goals."*

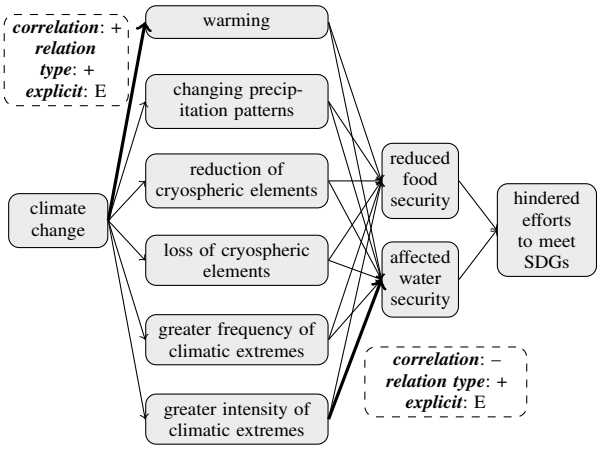


Figure 1: A sample from the *ClimateCause* dataset, showcasing the complex causal graphs and fine-grained annotations it contains.

1 Introduction

Causality is a fundamental driver of climate change and climate change discourse. Climatic phenomena unfold within complex causal networks, where causal relations are further complicated by contextual factors that introduce uncertainty, signal confounders, and highlight the variability of causal strength and direction (Pearl, 2009; Yarlett and Ramscar, 2019; Cui et al., 2025). For example, *why would an increase in global temperature of 1.5°C lead to more frequent floods in Africa but not in South America?* Causal reasoning also underlies policy making, where mitigation strategies related to climate change are designed through responsibility attribution and counterfactual analysis (Jang, 2013; Jamieson, 2015; Kalch et al., 2021).

Yet, existing datasets for causal discovery from text lack the granularity and abstraction for domains characterized by such complex causality (Mihăilă et al., 2013; Mirza et al., 2014; Mostafazadeh et al., 2016; Dunietz et al., 2017; Romanou et al., 2023; Tan et al., 2022; Vo et al., 2025; Pineda and Allein, 2025a,b). They primarily capture explicit direct cause-effect relations and omit those that are implicitly reported through word and sentence semantics; e.g., *"anthropogenic greenhouse gas emissions"* evokes causal relation *humans* → *greenhouse gas emissions*. Since most are sourced from news and social media, they typically report singular rather than general causality (e.g., the effects of a specific hurricane) and do not discuss higher-order, complex causal structures as commonly done in science-related discourse.

This paper introduces ***ClimateCause*, a manually-annotated dataset for causal discovery of complex and implicit causal structures from climate change reports** (Figure 1)[1]. Unlike prior resources, cause and effect representations are manually normalized through noun phrase

*Corresponding author: Liesbeth.Allein@UGent.be.

[1]Data and code are publicly available: https://github.com/laallein/ClimateCause.

Findings of the Association for Computational Linguistics: ACL 2026, pages 25458–25488

	Scope & Coverage		Causal Structures						Properties			Context		Reliability		
Dataset	**Topic**	**Source**	**Presence & Abs.**	**# CR**	**Cross-Sent. CR**	**Trigger**	**Consist. Formul.**	**Nested CS**	**Complex CS**	**Implicit CR**	**Correlation**	**Relation type**	**In Discourse**	**In Time/Space**	**Manual Annot.**	**Expert Annot.**	**Scientific CRs**
BioCause (Mihăilă et al., 2013)	Infectious diseases	Biomedical articles	✓	851	✓	✓				✓			✓		✓	✓	✓
Causal-TimeBank (Mirza et al., 2014)	Varia	News articles	✓	318		✓						✓			✓	✓	
CaTeRS (Mostafazadeh et al., 2016)	Common sense	Everyday stories	✓	308	✓					✓		✓	✓		✓	✓	
BECauSE 2.0 (Dunietz et al., 2017)	Varia	News articles, Penn Treebank, Congress hearings	✓	1,803		✓				✓		✓			✓	✓	
CRAB (Romanou et al., 2023)	Varia	News articles		2,730	✓		✓		✓				✓		✓		
CNC (Tan et al., 2022)	Socio-political events	News articles	✓	1,957		✓							✓		✓	✓	
RECESS (Tan et al., 2023)	Socio-political events	News articles	✓	2,754		✓							✓		✓	✓	
CCNC (Hagen et al., 2025)	Varia	News articles	✓	3,415		✓				✓			✓		✓		
ACCESS (Vo et al., 2025)	Common sense	Stories		1,494	✓		✓		✓				✓		✓		
PolarIs3CAUS (Pineda and Allein, 2025a)	Climate change	Reddit		95		✓	✓					✓	✓		✓	✓	
PolarIs4CAUS (Pineda and Allein, 2025b)	Climate change	Twitter (X)		181		✓	✓					✓			✓	✓	
***ClimateCause* (ours)**	**Climate change**	**Science-for-policy reports**	✓	**874**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison to existing resources, highlighting (nearly) unique features of *ClimateCause* in gray.

reformulation, and multi-event representations are disentangled to facilitate graph construction. Implicit and nested causal expressions are captured and each causal relation is further annotated with cause-effect correlation, relation type, and spatiotemporal context to create semantically-rich causal graphs.

Analyses on *ClimateCause* provide insights into the dominant semantic properties of the causation framing used in climate change reports. These reports are generally marked by low readability (Barkemeyer et al., 2016) and complex terminology (Bruine de Bruin et al., 2021). We take a causality-centered perspective on readability and propose several metrics to quantify the complexity of reported causality based on semantic properties of underlying causal graphs. We further showcase the potential of *ClimateCause* for benchmarking the causal reasoning abilities of LLMs through correlation inference and causal chain reasoning.

2 Comparison to Existing Resources

We compare *ClimateCause* against existing text-based causal discovery datasets, in which cause and effect are expressed as textual spans. This ensures a fair comparison among datasets explicitly designed for extracting causal relations directly from text only. In that respect, we exclude datasets with multimodal data (Liang et al., 2025; Shen et al., 2025) or causal questions in Q&A (Bondarenko et al., 2022; Ceraolo et al., 2024; Chi et al., 2024) and NLI setups (Roemmele et al., 2011; Du et al., 2022; Miliani et al., 2025), and datasets that are designed to trigger causal reasoning strategies for answer selection (Jin et al., 2023, 2024; Sheth et al., 2025). We also omit causal knowledge graphs (Heindorf et al., 2020; Li et al., 2021; Jaimini and Sheth, 2022) and knowledge graphs that include causal relations as one of the relations (Speer et al., 2017; Sap et al., 2019; Hwang et al., 2021).

ClimateCause is unique in its annotations for correlation, spatiotemporal context, and nested causality, establishing first-of-its-kind resources for investigating these three aspects for causal discovery (Table 1). Alongside BioCause (Mihăilă et al., 2013), it is one of the few built from scientific texts. Moreover, *ClimateCause* includes a larger number of causal relations than related climate change-focused datasets (Pineda and Allein, 2025a,b).

3 Dataset Construction: Source

Data source and extraction All texts in *ClimateCause* are sourced from science-for-policy reports published by the Intergovernmental Panel on Climate Change (IPCC)[2]. These authoritative reports synthesize the scientific consensus on causes, impacts, risks, and mitigation of climate change. They serve a dual purpose: *inform* non-expert stakeholders (i.e., policy-makers) about the scientific knowledge gained on a certain topic with confidence levels reflecting the level of consensus, while *influencing* them. A citizen science initiative led by Semantic Climate[3] is transforming the IPCC reports into a structured knowledge graph. At retrieval time (May 2025), the AR6 Synthesis Report: Climate Change

[2] https://www.ipcc.ch

[3] https://semanticclimate.github.io/p/en/

Feature		Value
Statement link		https://kg-ipclimatec-reports.wikibase.cloud/entity/statement/ Q31-91700757-4207-5de4-c0c1-5682b1be9db0
Section		2.1.2 Observed Climate System Changes and Impacts to Date
Paragraph		2.1.2.c
Series ordinal		1
Statement		*Climate change has caused substantial damages, and increasingly irreversible losses, in terrestrial, freshwater, cryospheric and coastal and open ocean ecosystems.*
Causation		Yes
Target		*has caused*
Explicitness		Explicit
Cause	*– NP*	*climate change*
	– Context	/
	– No_Quantifier	climate change
	– Belongs_to	/
Effect	*– NP*	*increasingly irreversible losses in freshwater ecosystems*
	– Context	/
	– No_Quantifier	irreversible losses in freshwater ecosystems
	– Belongs_to	/
Relation type		Positive
Correlation		Positive
Abbreviations		/
Combined		/
Confidence level		High confidence
Nested		/

Table 2: A sample from *ClimateCause*.

2023[4] had been partially converted into a machine-readable format. We extracted 75 statements via SPARQL from a dedicated Wikibase instance[5]; **[*Statement*] (*string*)**. These statements were selected based on the availability of confidence labels such that each statement reflects causal content that is grounded in a documented level of scientific consensus. We rely on the statement boundaries defined by Semantic Climate, resulting in statements of one or more sentences. A breakdown of the report sections from which the statements are drawn is in Appendix A.

Metadata For each statement, we retrieve the Wikibase link to the statement **[*Statement link*] (*URL*)**, section title **[*Section*] (*string*)**, paragraph identifier **[*Paragraph*] (*id*)**, ordinal position of the statement within the paragraph **[*Series ordinal*] (*integer*)**, and associated confidence level **[*Confidence level*] (*label*)**.

4 Dataset Construction: Annotation

4.1 Annotation Features

Table 2 shows a sample from *ClimateCause*, reflecting all annotation features of one cause-effect pair in a statement. We go over the features and illustrate them with samples from the dataset. Detailed annotation guidelines are included in Appendix B.

[4] https://www.ipcc.ch/report/ar6/syr/

[5] https://kg-ipclimatec-reports.wikibase.cloud/wiki/Main_Page

4.1.1 Presence/Absence of Causality

[*Causation*] (*yes/no*) marks whether a statement reports at least one causal relation. If not, no further action is taken for that statement.

4.1.2 Target Word and Explicitness

[*Target*] (*string*) includes the target words in the statement that trigger the causal relation (Baker et al., 1998). Examples of popular causation-evoking terms are *"cause"* and *"due to"*. We also include more subtle triggers to capture implicit causality. An example is *"anthropogenic greenhouse gas emissions"*, which evokes *humans* → *greenhouse gas emissions* through semantics. An explicit target may also be absent; [*Target*] = /. In that case, the reader needs to infer the causal relation. For example: *warming* → *rise in global mean sea level* in the statement *"global mean sea level will rise by about 2–3 m if warming is limited to 1.5°C and 2–6 m if limited to 2°C."*. The target guides the decision whether the causal relation is explicit (E) or implicit (I); **[*Explicitness*] (*E/I*)**.

4.1.3 Cause and Effect

Annotators identify each cause-effect pair within a statement and resolve coreference where necessary. This way, we capture causal relations that would otherwise remain unidentified or cannot be understood outside their discourse context. We enforce consistency in cause and effect representation through (A) syntactic reformulation of events into noun phrases, (B) abbreviation resolution, and (C) disentanglement of multi-event events. We also identify (D) nested causal structures and (E) contextualize causal relations in time and space.

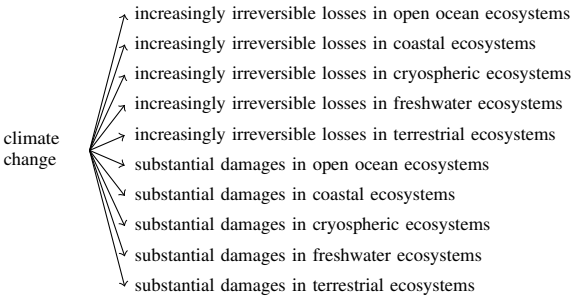
(A) Syntactic reformulation [*Cause–NP*] (*string*) and **[*Effect–NP*] (*string*)**, respectively, represent the noun phrase (NP) reformulation of cause and effect. We adopt the guidelines for NP reformulation from Pineda and Allein (2025a,b), which instructed annotators to minimally alter the original wording of cause and effect. For example: *unsustainable agricultural expansion* → *increased ecosystem vulnerability* in *"Unsustainable agricultural expansion [...] increases ecosystem and human vulnerability"*. Including the target word *"increases"* as an adjective in the effect is unlike dominant annotation approaches for semantic relations (e.g., Baker et al. (1998)). Nonetheless, target inclusion arguably captures the effect more faithfully than exclusion, i.e., the cause leads to

an increase in and not *the existence of ecosystem vulnerability*.

(B) Abbreviation resolution Annotators resolve the abbreviations used in cause and effect formulations directly in their NP reformulations and include them in **[*Abbreviations*] (*string*)**. Abbreviations can be well-established, e.g., CO_2 = *carbon dioxide*, or report-specific, e.g. *AFOLU* = *Agriculture, Forestry, and Other Land Use*. Appendix C includes an overview of all abbreviations.

(C) Disentanglement of multi-event cause/effect Annotators decompose cause and effect into standalone events and actions. For example:

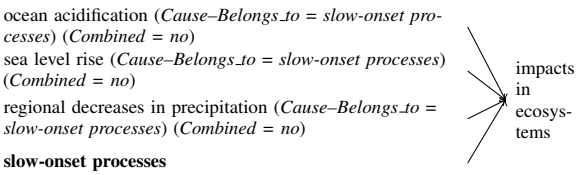
Statement: *"Climate change has caused substantial damages, and increasingly irreversible losses, in terrestrial, freshwater, cryospheric and coastal and open ocean ecosystems."*



Disentangling causes and effects strongly contrasts with prior resources, which typically preserve mixed representations. Disentanglement has various advantages for causal graph building (e.g., event matching) and evaluation (e.g., one-by-one validation of cause-effect pairs). However, it may produce flawed or misleading causal graphs in case of (C.1) an elaboration of examples and (C.2) a combination of events.

(C.1) Elaboration of examples A statement may elaborate on events that are examples of an overarching cause or effect. For example:

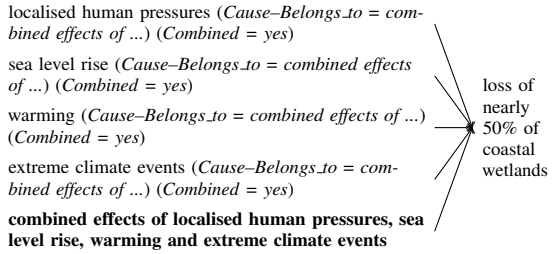
Statement: *"Impacts in ecosystems from slow-onset processes such as ocean acidification, sea level rise or regional decreases in precipitation have also been attributed to human-caused climate change."*



Representing the examples next to the overarching event/action (e.g., *slow-onset processes*) in a causal graph without any indication of their subordinate relation to the overarching event may produce a misleading graph. The events may be wrongly seen as independent from each other. We mark such a subordinate relation between two events by including the overarching event in **[*Cause–Belongs_to*] (*string*)**, when the subordinate event is the cause, or in **[*Effect–Belongs_to*] (*string*)**, when the subordinate event is the effect. We also indicate that the causal relation does not arise from a combination of events **[*Combined*] (*yes / no*)** to distinguish it from a combination of events.

(C.2) Combination of events A causal relation may only arise from the combination of events. However, determining this faithfully would require extensive domain knowledge. Given the linguistic expertise of the annotators, we therefore focus on linguistic signals that indicate whether multiple causes produce a single effect or one cause leads to multiple effects. A clear example of such a signal is *"the combined effects of"* in:

Statement: *"Nearly 50% of coastal wetlands have been lost over the last 100 years, as a result of the combined effects of localised human pressures, sea level rise, warming and extreme climate events."*

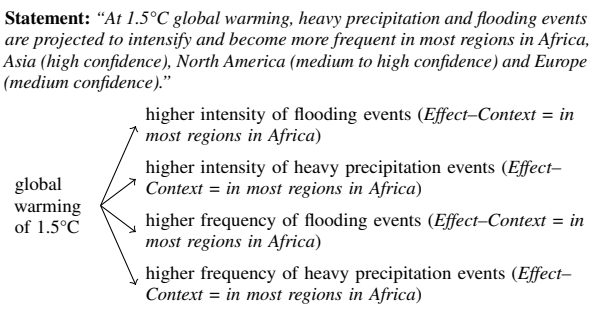


We include the full noun phrase covering all events in the combination in [*Cause–Belongs_to*] (*string*) or [*Effect–Belongs_to*] (*string*) and mark that the causal relation is explicitly reported to arise from the combination of events [*Combined*] (*yes / no*).

(D) Nested causality Causal relations that are nested within another cause or effect are included as standalone causal relations. A striking example is the term *CO2-FFI* (i.e., *"carbon dioxide emissions from Fossil Fuel combustion and Industrial processes"*), which encapsulates *fossil fuel combustion → carbon dioxide emissions* and *industrial processes → carbon dioxide emissions*. We include the overarching effect in [*Effect–Belongs_to*] and mark the relation as **[*Nested*] (*yes/no*)**.

(E) Spatiotemporal contextualization The report often describes cause-effect pairs that occur in specific spatiotemporal contexts. We include such spatiotemporal information in **[*Cause–Context*] (*string*)** and **[*Effect–Context*] (*string*)**. The importance of contextualization is clear in the following statement, where different confidence

levels have been assigned to the causality between events across geographical locations:



These four causal relations are repeated four times, where the effect is linked to a different spatial context in [*Effect–Context*], i.e., *"in most regions in Asia"*, *"in most regions in North America"*, and *"in most regions in Europe"*. The relations are paired with their designated confidence level. We consider spatiotemporal information as contextual when it is not an essential aspect of the meaning of the cause or effect, which is in line with *circumstances* in FrameNet (Baker et al., 1998). We include it in [*Cause–NP*] or [*Effect–NP*] when it is essential (e.g. *substantial losses in cryospheric ecosystems*).

4.1.4 Relation Type

[*Relation Type*] (*positive/negative*) describes the type of relation between cause and effect, i.e., *positive* (cause/production/facilitation) or *negative* (prevention) (Goldvarg and Johnson-Laird, 2001; Sloman et al., 2009)[6]. A positive relation indicates that the existence of the cause *leads to* the existence of the effect. A negative relation suggests that the existence of the cause *prevents* the existence of the effect. For instance, a negative relation between *well-implemented land-based mitigation options* → *trade-offs in terms of employment* can be triggered by the target *"to avoid"*.

4.1.5 Correlation

[*Correlation*] (*positive/negative*) marks the direction of the association between cause and effect. A correlation is classified as *positive* when a change in the cause produces a change in the effect in the *same* direction. It is *negative* when the change occurs in the *opposite* direction. Since cause and effect expressions in [*Cause–NP*] and [*Effect–NP*] frequently contain quantifiers or directional cues (e.g., *"reductions of"* and *"greater"*), the annotators manually remove these lexical signals to isolate the underlying concepts. These reformulations are stored in **[*Cause–No_Quantifier*] (*string*)** and **[*Effect–No_Quantifier*] (*string*)**. The value of [*Correlation*] is then determined based on the relationship between the quantifier-free forms.

[6]Two other relations exist (*allow* and *allow not*), yet they are considered weak (Goldvarg and Johnson-Laird, 2001). Hence, our focus on stronger relations *cause* and *prevention*.

4.2 Annotation Procedure

We recruit two expert annotators with an academic background in linguistics and argumentation, who have been involved in a previous annotation campaign on causal discovery from text (more details in Appendix D).

4.2.1 Annotation Round 1

The first annotation round starts with an oral discussion of the annotation guidelines, after which the two annotators independently annotate a subset of 11 statements (14.67% of the dataset). Their annotations are compared, and disagreements are resolved by a third annotator.

There is perfect agreement on the presence and absence of causality (§4.1.1): Cohen's $\kappa = 1.0$; 10 *yes*, 1 *no*. This suggests that causality is clearly signaled in the statements and/or that the annotators share a consistent understanding of causality. In contrast, agreement on target identification (§4.1.2) is substantially lower, with $k = -.075$, indicating less-than-chance agreement. This low level of agreement extends to the other causal annotations. Implicit causal relations and nested causal structures, especially those involving report-specific abbreviations such as *CO_2-FFI* and *CO_2-LULUCF*, were frequently identified by one annotator only. Violations against multi-event decomposition were also common. In the subsequent feedback session with the expert annotators, it became clear that disagreements were most often due to gaps in domain knowledge (e.g., unfamiliarity with report-specific abbreviations) and vagueness in the annotation guidelines. Annotators typically accepted the causal relations identified by the other once clarified. We revised the annotation guidelines based on these observations.

4.2.2 Annotation Round 2

The second round adopts a validation-based approach to reduce the annotation time and cost. The more expressive annotator is asked to annotate all remaining statements. Here, *expressiveness* is treated as a recall metric over the annotated causal relations in the first round; i.e., the expressiveness

of annotator A is defined as the number of causal relations identified by annotator A and B, divided by the total number of causal relations annotated by B. The less expressive annotator then performs one of the following four actions on each causal relation identified and annotated by the first annotator, motivating the action in a free-text comment:

- *Valid*: causal relation appears in the statement and all annotations are correct. The annotations are kept as such.
- *Remove*: causal relation does not appear in the statement. The annotations are therefore invalid and should be removed entirely.
- *Add*: causal relation in the statement has not been identified by the first annotator. The second annotator provides the annotations.
- *Change*: causal relation appears in the statement, but not all annotations are correct. The second annotator changes the annotations and indicates what has been changed.

By adopting this kind of correction-based annotation setup, we aimed to reduce the cognitive load for the second annotator and specifically target significant errors (*validate/change*), omissions (*add*), and redundancies (*remove*).

The second annotator deemed all causal relation annotations valid for 43 statements and proposed actions for the remaining statements. The majority involved changes to the effect reformulations in [*Effect–NP*], motivated by the redundant inclusion of the target word in the reformulation. In the end, we decided to keep the target words in the formulation of cause and/or effect when inclusion resulted in a more truthful representation of the event.

5 Dataset Analysis

5.1 General Insights

Table 3 summarizes the dataset statistics. Causation seems to be a profound framing device for discussing climate change processes and mitigation strategies, with 63 out of 75 statements each reporting 14.06 causal relations on average. IPCC authors express causality primarily through positive relation types. Cause-effect associations are typically positive. Chi-squared tests indicate a significant association ($N = 874, p < 0.01$) between correlation type and explicitness ($\chi^2 = 61.31$), relation type ($\chi^2 = 86.14$), and nested causality ($\chi^2 = 26.53$). Explicit relations occur more frequently than implicit ones, especially for negative correlations. Negative relation types are strongly linked to negative correlations, whereas nested relations exclusively exhibit positive correlations.

Feature	Count
Sections / paragraphs	10 / 19
Statements	75
[*Causation*]: Yes | No	63 | 12
Causal relations ; –NP	874
[*Explicitness*]: Explicit | Implicit	593 | 281
[*Relation type*]: Positive | Negative	827 | 47
[*Correlation*]: Positive | Negative	581 | 293
[*Nested*]: No | Yes	828 | 46
Overarching structures: No | Elaboration of examples | Combination of events	517 | 254 | 103
Unique relations; –No_Quantifier	653
Unique target words	95

Table 3: General dataset statistics.

5.2 Readability and Semantic Complexity of Reported Causality

The annotators reported a high cognitive load during causal relationship identification. They attributed this to the low readability of the statements in *ClimateCause*. Readability is indeed low: most statements require college-level reading; Flesch Reading Ease $\in [0, 30]$ (Flesch, 1948)[7]. However, traditional readability metrics, which are typically based on lexical complexity or sentence length, do not capture how easily readers can understand and retrieve reported causal relationships (Kincaid et al., 1975; Coleman and Liau, 1975; Chall and Dale, 1995).

We therefore propose metrics to measure readability through the *semantic complexity of the underlying causal structures*. Here, high complexity implies low readability. While structural properties such as graph depth and breadth mainly contribute to the structural complexity of a causal graph, we base our metrics on annotated causal properties such as explicitness and correlation. Relying on such semantic aspects go beyond existing metrics based on counts of causal connectives, verbs, and particles (Follmer et al., 2021).

We assume that IPCC authors generally follow Grice's maxim of manner, i.e., avoid unnecessary complexity (Grice, 1975), as they want to inform non-experts about the current scientific consensus

[7]More extensive readability analyses in Appendix E.

on climate change. Under this assumption, less frequent causal patterns are treated as more complex. The metrics therefore penalize causal graphs with common cause/effect structures, elaborations of examples, nested causality, and causal relations with negative correlation or relation type. We include additional penalties when these properties occur frequently and/or sequentially and encompass a large number of relations.

5.2.1 Readability Metrics

Elaboration of examples and common cause/effect structures Both structures include a set of events that have a subordinate relation to a shared overarching event, where the combination of events in the set is either facultative (i.e., *elaboration of examples*) or mandatory (i.e., *common cause/effect*) in the representation of causality.

We identify the subordinating relations, i.e., event $\rightarrow$ overarching event, from [*Cause–No_Quantifier*] to [*Cause–Belongs_to*] and [*Effect–No_Quantifier*] to [*Effect–Belongs_to*] in s. Let R^{com} be the set of relations with [*Combined*] (*yes*), and R^{ex} those with [*Combined*] (*no*) and [*Nested*] (*no*). We construct graphs G^{com} and G^{ex} from those two sets, where nodes represent events and edges subordinating relations. We then decompose G^{com} and G^{ex} into connected subgraphs: $\hat{G}^{com} = \{\hat{G}^{com}_1, \ldots, \hat{G}^{com}_M\}$ and $\hat{G}^{ex} = \{\hat{G}^{ex}_1, \ldots, \hat{G}^{ex}_N\}$, where a subgraph $\hat{G}^{com}_m$ captures the m^{th} overarching event with all its subordinating relations. Finally, complexity is computed as:

$$C^{com}(s) = \sum_{i=1}^{M} \text{count\_edges}(\hat{G}^{com}_i) + M; \quad (1)$$

$$C^{ex}(s) = \sum_{i=1}^{N} \text{count\_edges}(\hat{G}^{ex}_i) + N; \quad (2)$$

where the additive terms M and N penalize statements with many common cause/effect relations and elaboration of example structures.

Nested causality These structures embed one or more causal relations in another event or concept.

We associate all causal relations in s for which [*Nested*] (*yes*) with its overarching "nesting" event in [*Effect–Belongs_to*]. Let $L^{nest} = \{l_1, \ldots, l_K\}$ be the set of nesting events and T_k the number of causal relations that are nested within nesting event l_k. Complexity is computed as:

$$C^{nest}(s) = \sum_{i=1}^{K} (T_i + T_i \log T_i); \quad (3)$$

where $T_i \log T_i$ penalizes events nesting a large number of relations, e.g., *human-caused climate change* (1 relation) vs *CO_2-LULUCF: carbon dioxide emissions from land use, land-use change, and forestry* (3 relations).

Correlation Negatively correlated causal events imply a change in the opposite direction when intervening on the cause. We construct a directed graph $G = (V, E)$ using all cause-effect pairs in s, where nodes V represent events, edges E represent causal relations, and each edge is labeled by the correlation between connected events. We compute $|E^-|$, the total number of negatively labeled edges; $|E^-_{path}|$, the number of paths containing two or more consecutive negative edges; $|V^{mix}_{in}|$, the number of nodes with both positive and negative *incoming* edges; and $|V^{mix}_{out}|$, the number of nodes with both positive and negative *outgoing* edges. Complexity is scored as:

$$C^{corr}(s) = |E^-| + |E^-_{path}| + |V^{mix}_{in}| + |V^{mix}_{out}|; \quad (4)$$

penalizing graphs with a high frequency and long sequences of negatively correlated causal events, and a mix of positive and negative correlations.

Relation type Complexity is computed analogously to $C^{corr}(s)$ with relation type annotations as edge labels:

$$C^{pol}(s) = |E^-| + |E^-_{path}| + |V^{mix}_{in}| + |V^{mix}_{out}|; \quad (5)$$

penalizing graphs with a high frequency and long sequences of negative relation types, and a mix of positive and negative relation types.

Total semantic complexity of reported causality The complexity metrics show distinct ranges in *ClimateCause*: $C^{com} = [0, 12]$, $C^{ex} = [0, 16]$, $C^{nest} = [0, 20.93]$, $C^{corr} = [0, 70]$, and $C^{pol} = [0, 40]$. Metrics with larger ranges may bias the overall complexity, C, when simply summed. We therefore normalize each metric $C^i(s)$, with $i \in \{com, ex, nest, corr, pol\}$ through min-max normalization based on its observed range across all statements, such that $\tilde{C}^i(s) \in [0, 1]$:

$$\tilde{C}^i(s) = \frac{C^i(s) - \min(C^i)}{\max(C^i) - \min(C^i)} \quad (6)$$

$$C(s) = \sum_{i=1}^{5} w_i \tilde{C}^i(s); \quad (7)$$

keeping w_i constant ($= 1$) in this work.

5.2.2 Discussion

In *ClimateCause*, 43 statements (57.33%) exhibit semantically complex causal structures, indicated by $C(s) > 0$ (max. $C(s) = 1.821$). Most graphs are complex in one (27 statements) or two (10 statements) metrics. None involves all five. We also observe a significant positive Pearson correlation between statement length (token count) and total complexity; $n = 43, r = .590, p < 0.01$. This is expected since longer statements may include more causal relations, which potentially increases the semantic complexity of a causal graph. More analysis are included in Appendix F.

While our metrics are frequency-driven and tailored to the semantic annotations in *ClimateCause*, they lay the groundwork for topic-agnostic readability metrics that estimate cognitive complexity in interpreting causality in text. Future work could explore cognitive validation through user studies, adaptive weighting schemes, and integration with visualization tools to dynamically simplify complex causal structures for diverse audiences.

6 Benchmarking Causal Reasoning

6.1 Problem Formalization

Causal reasoning is the ability to discern cause-and-effect relationships between variables from available data and draw causal inferences (Jin et al., 2023; Wang, 2024; Yu et al., 2025). We argue that *ClimateCause* is valuable for evaluating causal reasoning through multiple causality-specific tasks, such as causal event extraction, including event detection and event argument extraction (Simon et al., 2024), causal DAG inference (Kiciman et al., 2023), and implicit causal chain discovery (Allein et al., 2025), but also through more general tasks like reading comprehension (RC). Here, we focus on correlation inference and causal chain reasoning.

Correlation Inference (CorrI) CorrI is the ability to identify the direction of association between two variables, where positive correlation means that the variables change in the *same* direction and negative the *opposite* direction.

- **CorrI**: Given a set of causal pairs R;
- **CorrI+RC**: Given a statement s and a set of causal pairs R.

For each $(e_i, e_j) \in R$, determine whether the correlation between e_i and e_j is positive or negative. Label distribution is {*positive*: 581, *negative*: 293}.

Causal Chain Reasoning (CCR) CCR is the ability to identify and analyze causal chains, where a causal chain is defined as a directed path of at least three nodes in a causal graph (Pearl, 2009).

- **CCR**: Given a causal graph where nodes V represent events, and edges causal relations.
- **CCR+ECI+RC**: Given a statement s and the set of all causal events in s, V.

In CCR+ECI+RC, the causal relations between the events in V need to be inferred from the text (i.e., event causality inference (ECI)) before a model can reason over their membership or position in a causal graph. ***Membership***: Determine for $e \in V$ whether it is part of a causal chain structure; {*yes*: 115, *no*: 397}. ***Position***: Determine for $e \in V$ which position it holds in a causal chain; {*start* (= source node): 32, *middle* (= mediator node): 48, *end* (= sink node): 35, *none*: 397}.

6.2 Methodology

We assess GPT5.1 through in-context learning. CorrI, CorrI+RC, and CCR+ECI+RC are evaluated through (i) zero-shot, (ii) few-shot, and (iii) chain-of-thought prompting. For CCR, we select three prompting strategies defined in Sheth et al. (2025), where the encoding of the given causal graph (i) lists all edges (adjacency), (ii) textually presents each node with its direct effects (single node), or (iii) uses GraphML format. Each prompting strategy is tested using three prompt variations to enforce robustness of the results. Events are represented using [*Cause-No_Quantifier*] and [*Effect-No_Quantifier*]. Table 4 shows the results, with per-class performance for ***position*** in Appendix G.

6.3 Analysis

Correlation inference The results for CorrI and CorrI+RC are both high and comparable. This suggests that the LLM effectively captures the correlation between events and that the presence of the original statement, which often verbalizes the correlation through adjectives and verbs, has little impact on prediction performance. We verify this by comparing the predictions with (CorrI+RC) and without (CorrI) the statement using the McNemar test (McNemar, 1947), which focuses on discordant results. A Bonferroni correction ($\alpha = 0.05/9 \approx 0.0056$) was applied to control Type I

Prompt	Precision	Recall	F1
CorrI			
0-shot	$0.8204_{\pm 0.1204}$	$0.8003_{\pm 0.2032}$	$0.7887_{\pm 0.0738}$
F-shot	$0.8874_{\pm 0.1017}$	$0.9486_{\pm 0.0211}$	$0.9156_{\pm 0.0652}$
CoT	$0.8556_{\pm 0.0971}$	$0.9369_{\pm 0.0364}$	$0.8934_{\pm 0.0683}$
Avg	$0.8544_{\pm 0.0970}$	$0.8953_{\pm 0.1259}$	$0.8659_{\pm 0.0839}$
CorrI+RC			
0-shot	$0.8294_{\pm 0.1330}$	$0.7194_{\pm 0.1294}$	$0.7574_{\pm 0.0391}$
F-shot	$0.9426_{\pm 0.0644}$	$0.9641_{\pm 0.0209}$	$0.9529_{\pm 0.0430}$
CoT	$0.8621_{\pm 0.0858}$	$0.8789_{\pm 0.0820}$	$0.8678_{\pm 0.0635}$
Avg	$0.8780_{\pm 0.0993}$	$0.8542_{\pm 0.1325}$	$0.8594_{\pm 0.0952}$

(a) Correlation inference.

Prompt	Precision	Recall	F1
CCR *membership*			
Adjacency	$0.5382_{\pm 0.3205}$	$0.8754_{\pm 0.0329}$	$0.6308_{\pm 0.2192}$
Single N.	$0.5023_{\pm 0.2619}$	$0.9739_{\pm 0.0151}$	$0.6371_{\pm 0.2087}$
GraphML	$0.6421_{\pm 0.3130}$	$0.9362_{\pm 0.1030}$	$0.7205_{\pm 0.1601}$
Avg	$0.5609_{\pm 0.2670}$	$0.9285_{\pm 0.0695}$	$0.6628_{\pm 0.1766}$
CCR *position*			
Adjacency	$0.2927_{\pm 0.0757}$	$0.9253_{\pm 0.0686}$	$0.4417_{\pm 0.0893}$
Single N.	$0.2914_{\pm 0.0878}$	$0.9615_{\pm 0.0666}$	$0.4432_{\pm 0.1043}$
GraphML	$0.2667_{\pm 0.0646}$	$0.9885_{\pm 0.0199}$	$0.4175_{\pm 0.0791}$
Avg	$0.2836_{\pm 0.0676}$	$0.9585_{\pm 0.0560}$	$0.4341_{\pm 0.0802}$
CCR+ECI+RC *membership*			
0-shot	$0.2860_{\pm 0.0080}$	$0.9014_{\pm 0.0411}$	$0.4339_{\pm 0.0047}$
F-shot	$0.3661_{\pm 0.1218}$	$0.6899_{\pm 0.2016}$	$0.4514_{\pm 0.0482}$
CoT	$0.3532_{\pm 0.1132}$	$0.7855_{\pm 0.1267}$	$0.4718_{\pm 0.0728}$
Avg	$0.3351_{\pm 0.0912}$	$0.7923_{\pm 0.1517}$	$0.4524_{\pm 0.0467}$
CCR+ECI+RC *position*			
0-shot	$0.1851_{\pm 0.0160}$	$0.6706_{\pm 0.1336}$	$0.2896_{\pm 0.0324}$
F-shot	$0.1980_{\pm 0.0823}$	$0.5545_{\pm 0.1170}$	$0.2781_{\pm 0.0650}$
CoT	$0.2600_{\pm 0.0641}$	$0.5087_{\pm 0.2241}$	$0.3405_{\pm 0.1094}$
Avg	$0.2144_{\pm 0.0631}$	$0.5780_{\pm 0.1602}$	$0.3027_{\pm 0.0717}$

(b) Causal chain reasoning.

Table 4: Mean and standard deviation results for each prompting strategy individually (3 runs) and all strategies together (9 runs; *Avg*).

errors (Bonferroni, 1936). The results show significant discordance in five prompt variations. This indicates that statement availability does affect predictions, though not improving the performance.

Causal chain inference Despite promising recall performance, the LLM does not seem to fully understand what causal chains are or consistently infer them as part of its reasoning. For example, one main feature of a chain is that it involves at least three events. A manual error analysis of the membership predictions shows that for 45 statements without chains, the LLM (in at least one of its prompt settings) predicts only one or two events as chain members in 37 cases (CCR+ECI+RC). This may be due to the abstract causal graph-oriented definition of a causal chain in the prompt, which does not align well with text-only input. However, this behavior persists in 23 statements when the input includes an explicit causal graph (CCR).

Chain membership identification is an inherent step prior to position identification. If the LLM takes this step during position identification, events outside a chain (*membership* = *no*) should have *position* = *none*, and those within (*membership* = *yes*) should have *position* = *start*, *middle*, or *end*. However, McNemar tests reveal significant discordance in 8 (CCR) and 7 (CCR+ECI+RC) of 9 prompt pairs. The LLM also frequently over-predicts start, middle, and end positions, with low mean precision scores: $0.2687/0.5008/0.3123$ (CCR) and $0.1934/0.2788/0.2283$ (CCR+ECI+RC). Confusion matrices show that most false positives belong to the 'none' class, followed by the middle class for start and end instances. Start and end are rarely confused.

Last, our causality-focused readability metric, $C(s)$, shows meaningful variation across predicted membership and position labels in CCR+ECI+RC. Kruskal-Wallis H tests (Kruskal and Wallis, 1952) indicate significant differences in $C(s)$ across the predicted labels ($p < 0.05$). For ***membership***, all nine runs are significant: $H(1) \in [9.55, 84.70]$, with small to large effect sizes ($\epsilon^2 \in [0.02, 0.16]$). For ***position***, five runs are significant: $H(3) \in [7.92, 47.01], p < 0.05$, with small to moderate effect sizes, $\epsilon^2 \in [0.01, 0.09]$. Dunn's post-hoc comparisons (Dunn, 1964) with Holm correction reveal no significant pairwise differences, suggesting subtle distributed effects rather than strong contrasts between labels.

7 Conclusion

This work introduced a manually annotated dataset designed to uncover complex causal structures in climate change reports, with unique annotations for correlation, nested causality, and spatiotemporal contextualization. Beyond its value for benchmarking correlation inference and causal chain reasoning in large language models, we demonstrated its potential for broader applications such as causality-focused readability metrics. Looking ahead, we consider spatiotemporal contextualization and scientific confidence assessments of causal relations particularly interesting avenues for advancing confounder and causal strength inference.

Limitations

Properties of causal relations not specifically addressed in this annotation setup but that are potentially relevant in the climate change discussion include causal strength and causal uncertainty. Parallels could be drawn with the confidence level labels that the report writers assigned to the statement. However, it is unclear whether the confidence level applies to every causal relation in that statement as the confidence evaluation does not focus solely on causality. Molina and Abadal (2021), for example, discovered a shift over time towards higher certainty levels in IPCC reports, implying a "call to action". These properties were excluded due to the difficulty of reliably inferring them from text without additional context or expert judgment.

The readability metrics are based on the semantic complexity of reported causal structures, where they penalize properties that occur less frequent in the *ClimateCause* dataset, adding additional penalities for structures that include a high number of such properties. Aligning these metrics with findings in the cognitive science and examining the frequencies of the observed causal patterns in *ClimateCause* would provide stronger, dataset-agnostic motivations for the metrics. Moreover, while the weights in complexity metric $C(s)$ are kept constant, one can arguably look into learning these weights empirically or normalizing them based on the observed ranges.

ClimateCause is moderate in size and heavily focused on a single topic. The benchmarking results on correlation inference and causal chain reasoning should not be seen as a generalizable evaluation of LLM capabilities. Larger and topic-diverse resources are necessary to make generalizable conclusions. Since *ClimateCause* is a newly constructed and does not modify or adapt an existing causal reasoning benchmark, it avoids the issues and limitations related to benchmarking on reused datasets, such as data contamination in the pre-training data of the LLM (Bean et al., 2025).

Ethical Considerations

Our use of the data from the Semantic Climate Wikibase is consistent with the intended use described in the Apache License 2.0. All annotations were made voluntarily and without remuneration. Given the nature of the IPCC data, the annotated content is free of any material that could negatively affect the integrity of the annotators or their personal well-being.

Acknowledgements

This work was funded by the Research Foundation - Flanders (FWO) under grant G0L0822N and the Swiss National Science Foundation (SNSF) under grant 209674 through the CHIST-ERA project "iTRUST Interventions against Polarisation in Society for Trustworthy Social Media. From Diagnosis to Therapy". Liesbeth Allein is further supported by a junior postdoctoral fellowship from the FWO under grant 12AGW26N.

References

Liesbeth Allein, Nataly Pineda-Castañeda, Andrea Rocci, and Marie-Francine Moens. 2025. Assessing LLM Reasoning Through Implicit Causal Chain Discovery in Climate Discourse. *arXiv preprint arXiv:2510.13417*.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley Framenet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Ralf Barkemeyer, Suraje Dessai, Beatriz Monge-Sanz, Barbara Gabriella Renzi, and Giulio Napolitano. 2016. Linguistic analysis of IPCC summaries for policymakers and associated coverage. *Nature Climate Change*, 6(3):311–316.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

CE Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3.

Wändi Bruine de Bruin, Lila Rabinovich, Kate Weber, Marianna Babboni, Monica Dean, and Lance Ignon. 2021. Public understanding of climate change terminology. *Climatic Change*, 167(3):37.

Roberto Ceraolo, Dmitrii Kharlapenko, Amélie Reymond, Rada Mihalcea, Bernhard Schölkopf, Mrinmaya Sachan, and Zhijing Jin. 2024. Analyzing Human Questioning Behavior and Causal Curiosity through Natural Queries. In *Causality and Large Models@ NeurIPS 2024*.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, Cambridge, Massachusetts.

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.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Shaobo Cui, Luca Mouchel, and Boi Faltings. 2025. Uncertainty in Causality: A New Frontier. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8022–8044, Vienna, Austria. 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.

Olive Jean Dunn. 1964. Multiple Comparisons Using Rank Sums. *Technometrics*, 6(3):241–252.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

D Jake Follmer, Ping Li, and Roy Clariana. 2021. Predicting Expository Text Processing: Causal Content Density as a Critical Expository Text Metric. *Reading Psychology*, 42(6):625–662.

Eugenia Goldvarg and Philip N Johnson-Laird. 2001. Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, 25(4):565–610.

Herbert Paul Grice. 1975. *Speech Acts*, volume 3 of *Syntax and Semantics*, chapter Logic and conversation.

Tim Hagen, Niklas Deckers, Felix Wolter, Harrisen Scells, and Martin Potthast. 2025. Investigating Counterclaims in Causality Extraction from Text. *arXiv preprint arXiv:2510.08224*.

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.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. *Proceedings of the AAAI conference on artificial intelligence*, 35(7):6384–6392.

Utkarshani Jaimini and Amit Sheth. 2022. CausalKG: Causal Knowledge Graph Explainability Using Interventional and Counterfactual Reasoning. *IEEE Internet Computing*, 26(1):43–50.

Dale Jamieson. 2015. Responsibility and Climate Change. *Global Justice: Theory Practice Rhetoric*, 8(2).

S Mo Jang. 2013. Framing responsibility in climate change discourse: Ethnocentric attribution bias, perceived causes, and policy attitudes. *Journal of environmental psychology*, 36:27–36.

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, 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*.

Anja Kalch, Helena Bilandzic, Andrea Sappler, and Sarah Stellinger. 2021. Am I responsible? The joint effect of individual responsibility attributions and descriptive normative climate messages on climate mitigation intentions. *Journal of Environmental Psychology*, 78:101711.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *Transactions on Machine Learning Research*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval

Technical Training Command Millington TN Research Branch.

William H Kruskal and W Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

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.

Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. 2025. VidEvent: A Large Dataset for Understanding Dynamic Evolution of Events in Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5128–5136.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(2):1–18.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. 2025. ExpliCa: Evaluating Explicit Causal Reasoning in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17335–17355, Vienna, Austria. Association for Computational Linguistics.

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.

Tomas Molina and Ernest Abadal. 2021. The Evolution of Communicating the Uncertainty of Climate Change to Policymakers: A Study of IPCC Synthesis Reports. *Sustainability*, 13(5):2466.

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.

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].

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.

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.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI conference on artificial intelligence*, 33(01):3027–3035.

ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. 2025. Exploring multi-modal data with tool-augmented LLM agents for precise causal discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 636–660, Vienna, Austria. Association for Computational Linguistics.

Ivaxi Sheth, Bahare Fatemi, and Mario Fritz. 2025. CausalGraph2LLM: Evaluating LLMs for Causal Queries. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2076–2098, Albuquerque, New Mexico. Association for Computational Linguistics.

Étienne Simon, Helene Olsen, Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2024. Generative Approaches to Event Extraction: Survey and Outlook. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 73–86, Miami, Florida, USA. Association for Computational Linguistics.

Steven Sloman, Aron K Barbey, and Jared M Hotaling. 2009. A Causal Model Theory of the Meaning of Cause, Enable, and Prevent. *Cognitive Science*, 33(1):21–50.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Fiona Anting Tan, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, Tiancheng Hu, and 1 others. 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.

Vy Vo, Lizhen Qu, Tao Feng, Yuncheng Hua, Xiaoxi Kang, Songhai Fan, Tim Dwyer, Lay-Ki Soon, and Gholamreza Haffari. 2025. ACCESS : A Benchmark for Abstract Causal Event Discovery and Reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1049–1074, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Daniel Yarlett and Michael Ramscar. 2019. Uncertainty in Causal and Counterfactual Inference. In *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society*, pages 956–961. Routledge.

Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2025. CausalEval: Towards Better Causal Reasoning in Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12512–12540, Albuquerque, New Mexico. Association for Computational Linguistics.

A IPCC Sections

Table 5 displays a breakdown of all sections of the IPCC Climate Change 2023 Report from which the statements in *ClimateCause* are taken.

Section	Title
2	Current Status and Trends
2.1	Observed Changes, Impacts and Attribution
2.1.1	*Observed Warming and its Causes*
2.1.2	*Observed Climate System Changes and Impacts to Date*
2.3	Current Mitigation and Adaptation Actions and Policies are not Sufficient
2.3.2	*Adaptation Gaps and Barriers*
3	Long-Term Climate Change and Development Futures
3.1	Long-Term Climate Change, Impacts and Related Risks
3.1.1	*Long-term Climate Change*
3.1.3	*The Likelihood and Risks of Abrupt and Irreversible Change*
4	Near-Term Responses in a Changing Climate
4.1	*The Timing and Urgency of Climate Action*
4.8	Strengthening the Response: Finance, International Cooperation and Technology
4.8.1	*Finance for Mitigation and Adaptation Actions*
4.9	*Integration of Near-Term Actions Across Sectors and Systems*

Table 5: Sections of the IPCC Climate Change 2023 Report from which the texts in *ClimateCause* have been drawn.

B Annotation Guidelines

We suggested to the annotators to follow a three-step annotation flow when identifying and annotating causal relations for a given statement:

1. **Causal relation extraction:** Decide whether the statement expresses at least one causal relation. If so, identify all the cause-effect pairs, specify their context mentioned in the statement, highlight the terms in the statement that trigger the causal relation between the events or actions that make up the cause and effect, and indicate whether the causality is explicitly or implicitly conveyed;

2. **Standardization and characterization:** Standardize the phrasing of the cause and effect by formulating the events into noun phrases and characterize the relation type and correlation of the causal relation;

3. **Complex causal structures:** Identify and label causal structures present in the statement, such as common cause/effect.

B.1 Annotation Setting

The annotators were presented with statements from the IPCC reports, where each statement (ranging from single sentences to full paragraphs) was shown on a line in an excel file. The features they had to annotate were assigned their own column, which they had to fill out.

B.2 Annotation Features

Table 6 gives an overview of the features annotated in *ClimateCause*.

B.3 Presence/Absence of Causal Relations [*Causation*]

Given a statement, indicate whether or not it contains at least one causal relation [*Causation*].

- If no causal relation is present: [*Causation*] = No.

In this case, no further annotation action needs to be taken for the statement.

Feature	Description
Statement link	URL to the statement in Wikibase (no action needed)
Section	Section number from which the statement is taken (no action needed)
Paragraph	Paragraph number from which the statement is taken (no action needed)
Series ordinal	Position in the paragraph from which the statement is taken (no action needed)
Confidence level	Confidence level of the statement (no action needed)
Statement	The statement (no action needed)
Causation	Binary indicator (yes/no) whether the statement reports a causal relation (§B.3)
Target	Target word(s) (string) that evokes the causal relation (§B.4)
Cause – NP	Noun phrase reformulation of the cause (string) (§B.5.2)
Cause – Context	Spatiotemporal context of the cause (string) (§B.5.4)
Cause – No_quantifier	Reformulation of the cause without quantifiers (string) (§B.5.5)
Cause – Belongs_to	Event to which the cause belongs (string) (§B.5.3)
Effect – NP	Noun phrase reformulation of the effect (string) (§B.5.2)
Effect – Context	Spatiotemporal context of the effect (string) (§B.5.4)
Effect – No_quantifier	Reformulation of the effect without quantifiers (string) (§B.5.5)
Effect – Belongs_to	Event to which the effect belongs (string) (§B.5.3)
Combined	Binary indicator (yes/no) whether the connection between the cause/effect in *–NP* and the overarching event in *–Belongs_to* is binding
Nested causality	Binary indicator (yes/no) whether the causal relation is the nested part in a nesting construction
Explicitness	Binary label (E/I) whether the causal relation is conveyed explicitly or implicitly (§B.6)
Relation type	Binary label (positive/negative) whether the relation type is positive (CAUSES) or negative (PREVENTS) (§B.8)
Correlation	Binary label (positive/negative) whether correlation is positive (increase → increase) or negative (increase → decrease) (§B.7)
Abbreviations	Set of abbreviations used in the statement resolved to their full meaning (string) (§B.9)

Table 6: Overview of all features with their description.

- If at least one causal relation is present: [*Causation*] = Yes.

B.4 Target Identification [*Target*]

Write down the [*Target*] that evokes the causal relation.

Span of the target When the target is a **verb**, include:

- All 'verb' parts of the verb (e.g., has caused, is causing, can cause, will continue to increase, cannot avoid).
- Prepositions in case of preposition-combined verbs (e.g., arise from, contribute to, lead to, account for).
- Words that would otherwise break the span (e.g., driven in part by, have adversely affected, can also cause).
- 'By' in passive constructions (e.g., can be avoided by, exacerbated by).
- 'To' for infinitives (e.g., to minimize, to address).

If target is a **noun** (rather rare), include:

- Determiner/article 'a' and 'the' (e.g., a cause of, the effect of).
- Adjectives and quantifiers (e.g., the direct cause of, other causes of, historical and unequal contributions arising from).

If target is **other**, include:

- If causality embedded in word, abbreviation, or adjective: give that word, abbreviation, or adjective as target (e.g., AFOLU, anthropogenic).
- Discourse markers (e.g., due to, from).
- Special case: "human-caused climate change": here, the target is "caused" with causal relation: humans (cause) → climate change (effect).
- If the target is not explicit, which means that the causal relation is indirect without explicit target, use "/" to mark the target.

B.5 Identification of Causal Relations

B.5.1 General Rules

Each entry presents one causal relation Each causal relation in the statement is identified and assigned to a new annotation entry, copying the statement and all other non-modifiable metadata (i.e., statement link, section, paragraph, series ordinal, and confidence level) to a new row in the annotation file. This way, one row in the file points to one causal relation.

The cause is contained in [*Cause–NP*] and the effect in [*Effect–NP*], with *–NP* referring to noun phrase reformulation (noun phrase reformulation is described in §B.5.2).

Note on confidence level A statement may mention multiple confidence levels. Make sure to assign each confidence level to the appropriate causal relation.

Reference resolution References such as *"it"* are replaced with their referent, for which the annotators have access to the full text of the IPCC reports.

B.5.2 Noun Phrase Reformulation *(–NP)*

General rule The cause and effect should be formulated as noun phrases in such a way that one can read the causal relation as *"[Cause–NP] causes/leads to [Effect–NP]"*, and the causal relation can be understood as outside its communicative context.

Reformulation guidelines Make minimal alterations to the semantics of the cause and effect so that the reformulated version stays as close as possible to the semantics of the original phrasing. Rely on words used in the statement for reformulation as much as possible.

Example: *"With further global warming, every region is projected to increasingly experience concurrent and multiple changes in climatic impact-drivers."*

The two causal relations are (7):

Cause–NP	Effect–NP
further global warming	increase in concurrent changes in climatic impact-drivers
further global warming	increase in multiple changes in climatic impact-drivers

Table 7: Example of [Cause–NP] and [Effect–NP].

Here, *"increase in"* is inspired by *"increasingly experience"*. Moreover, the formulation of the effect as *"increase in concurrent changes in climatic impact-drivers"* is more precise and closer to the effect report in the statement than *"concurrent changes in climatic impact-drivers"*.

Example: *"At 2°C or above, [...] more frequent and/or severe agricultural and ecological droughts are projected in Europe, Africa, Australasia and North, Central and South America."*

Four causal relations can be formulated as (8):

Here, the cause *"2°C or above"* is reformulated as *"global warming of 2°C or above"* such that the cause can be understood outside of its context. To resolve this, annotators have access to the original texs of the IPCC reports.

B.5.3 Multi-Component Events *(–Belongs_to)* [*Combined)*]

Causal relations with the multi-component cause or effect are broken down into their individual causal relations, and each causal relation is assigned to its own line in the excel file.

Cause–NP	Effect–NP
global warming of 2°C or above	higher frequency of agricultural droughts
global warming of 2°C or above	higher severity of agricultural droughts
global warming of 2°C or above	higher frequency of ecological droughts
global warming of 2°C or above	higher severity of ecological droughts

Table 8: Example of [Cause–NP] and [Effect–NP].

In general Causes and effects are split into separate events.

Example: *"With further global warming, every region is projected to increasingly experience concurrent and multiple changes in climatic impact-drivers."*

The two causal relations are (9):

Cause–NP	Effect–NP
further global warming	increase in concurrent changes in climatic impact-drivers
further global warming	increase in multiple changes in climatic impact-drivers

Table 9: Example of [Cause–NP] and [Effect–NP].

When multi-component event presents examples of an overarching event The components can be presented as examples of an overarching event or concept. If the examples act as *cause*, then the overarching event to which they belong is marked in [*Cause–Belongs_to*]. If the examples act as *effect*, then the overarching event to which they belong is marked in [*Effect–Belongs_to*]

Example: *"Impacts in ecosystems from slow-onset processes such as ocean acidification, sea level rise or regional decreases in precipitation have also been attributed to human-caused climate change."*

In the above case, there should be four lines in the excel sheet for each causal relation (10):

On [*Combined*] In this case, [*Combined*] = *No* to indicate that the combination of the events in cause or effect is not a necessary condition for the validity of the causal relation.

Cause–NP	Effect–NP	Cause–Belongs_To	Combined
slow-onset processes	impacts in ecosystems	–	–
ocean acidification	impacts in ecosystems	slow-onset processes	No
sea level rise	impacts in ecosystems	slow-onset processes	No
regional decreases in precipitation	impacts in ecosystems	slow-onset processes	No

Table 10: Example of multi-component event with elaboration of examples of an overarching event.

When multi-component event presents a combination of events The next example statement shows an instance where the causal relation is only established through the combination of the events in the cause. If the events in the combination act as *cause*, then the overarching event to which they belong is marked in [*Cause–Belongs_to*]. If the events in the combination act as *effect*, then the overarching event to which they belong is marked in [*Effect–Belongs_to*]

Example: *"Nearly 50% of coastal wetlands have been lost over the last 100 years, as a result of the combined effects of localised human pressures, sea level rise, warming and extreme climate events."*

In the above case, there should be five lines in the excel sheet for each causal relation (11):

On [*Combined*] Here, [*Combined*] = *Yes* to indicate that the causal relation is reported to be only established through the combination of events.

Cause–NP	Effect–NP	Cause–Belongs_To	Combined
combined effects of localised human pressures, sea level rise, warming and extreme climate events	loss of nearly 50% of coastal wetlands	–	–
localised human pressures	loss of nearly 50% of coastal wetlands	combined effects of localised human pressures, sea level rise, warming and extreme climate events	Yes
sea level rise	loss of nearly 50% of coastal wetlands	combined effects of localised human pressures, sea level rise, warming and extreme climate events	Yes
warming	loss of nearly 50% of coastal wetlands	combined effects of localised human pressures, sea level rise, warming and extreme climate events	Yes
extreme climate events	loss of nearly 50% of coastal wetlands	combined effects of localised human pressures, sea level rise, warming and extreme climate events	Yes

Table 11: Example of multi-component event with combination of events.

B.5.4 Modifiers of Space and Time *(–Context)*

Modifiers regarding space and time are often included in cause and effect. This spatiotemporal context of the cause and the context of the effect are annotated as [*Cause–Context*] and [*Effect–Context*], respectively.

The basic rule is:

→ **If the cause or effect is clear on itself and the modifier is NOT necessary to understand the event.**

The cause/effect should be written down **without** the modifier and the modifier should be included in either [*Cause–Context*] (when the event is a cause) or [*Effect–Context*] (when the event is an effect).

Example: *"Ocean warming and ocean acidification have adversely affected food production from shellfish aquaculture and fisheries in some oceanic regions (high confidence)."*:

- *"in some oceanic regions"* is not necessary to understand the effect *decrease in food production from shellfish aquaculture* or the effect *decrease in food production from fisheries*;
- Therefore, [*Effect–Context*] = *"in some oceanic regions"*.

→ **If the cause or effect is NOT clear on itself**

and the modifier is necessary to understand the event:

The cause/effect should be written down **together with** the modifier.

Example: *"Climate change has caused substantial damages, and increasingly irreversible losses, in terrestrial, freshwater, cryospheric and coastal and open ocean ecosystems."*:

- *"in terrestrial ecosystems* is necessary to capture the event more exactly;

- Therefore, [*Effect–NP*] = *"substantial damages in terrestrial ecosystems"*, [*Effect–Context*] is empty.

In case of multiple modifiers Some statements mention more than one spatiotemporal modifier for one causal relation. In this case, a distinction between *necessary* and *contextual* modifiers should be made. As with cause and effect, the modifiers should be split into single locations and times. The causal relation should then be repeated multiple times, each time annotated with one modifier, either as part of the cause/effect or as context in [*Cause–Context*] or [*Effect–Context*].

Example: *"Climate change has caused substantial damages, and increasingly irreversible losses, in terrestrial, freshwater, cryospheric and coastal and open ocean ecosystems."*

In the above case, the modifiers are **necessary**. There should be ten lines in the excel sheet reflecting each modifier (12):

Example: *"Climate change has contributed to desertification and exacerbated land degradation, particularly in low lying coastal areas, river deltas, drylands and in permafrost areas."*

In the above case, the modifiers are **contextual**. There should be ten lines in the excel sheet (13):

The first two causal relations also are valid because of the word *"particularly"*. It signals that the causal relation is particularly noticeable in the areas mentioned after it, but it does not exclude other areas.

Cause–NP	**Effect–NP**
climate change	substantial damages in terrestrial ecosystems
climate change	substantial damages in freshwater ecosystems
climate change	substantial damages in cryospheric ecosystems
climate change	substantial damages in coastal ecosystems
climate change	substantial damages in open ocean ecosystems
climate change	increasingly irreversible losses in terrestrial ecosystems
climate change	increasingly irreversible losses in freshwater ecosystems
climate change	increasingly irreversible losses in cryospheric ecosystems
climate change	increasingly irreversible losses in coastal ecosystems
climate change	increasingly irreversible losses in open ocean ecosystems

Table 12: Example of spatiotemporal contextualization - **necessary** modifiers.

Cause–NP	**Effect–NP**	**Effect–Context**
climate change	desertification	
climate change	exacerbated land degradation	
climate change	desertification	in low lying coastal areas
climate change	exacerbated land degradation	in low lying coastal areas
climate change	desertification	in river deltas
climate change	exacerbated land degradation	in river deltas
climate change	desertification	in drylands
climate change	exacerbated land degradation	in drylands
climate change	desertification	in permafrost areas
climate change	exacerbated land degradation	in permafrost areas

Table 13: Example of spatiotemporal contextualization - **contextual** modifiers.

Alignment with confidence label Contextual modifiers can also be linked to different confidence labels. In this case, make sure that the confidence label, which is given beforehand, aligns with the annotated causal relation.

Example: *"At 1.5°C global warming, heavy precipitation and flooding events are projected to intensify and become more frequent in most regions in Africa, Asia (high confidence), North America (medium to high confidence) and Europe*

(medium confidence)."

Three of the sixteen causal relations that should be annotated look as follows (14):

Cause–NP	Effect–NP	Effect–Context	Confidence Level
global warming of 1.5°C	higher intensity of heavy precipitation events	in most regions in Africa	high confidence
global warming of 1.5°C	higher intensity of heavy precipitation events	in most regions in North America	medium to high confidence
global warming of 1.5°C	higher intensity of heavy precipitation events	in most regions in Europe	medium confidence

Table 14: Example of spatiotemporal contextualization - alignment with confidence label.

B.5.5 No Quantifiers *(–No_Quantifier)*

General rule You should be able to read the causal relation between [*Cause–No_Quantifier*] and [*Effect–No_Quantifier*] as either:

- *"An increase in* [*Cause–No_Quantifier*] *causes an increase in* [*Effect–No_Quantifier*]*"*;
- or, *"An increase in* [*Cause–No_Quantifier*] *causes a decrease in* [*Effect–No_Quantifier*]*"*.

Reformulation Make minor alterations to the original phrasing of the cause and effect in [*Cause–NP*] and [*Effect–NP*], which are then included as reformulations in [*Cause–No_Quantifier*] and [*Effect–No_Quantifier*].

- Remove adjectives, adverbs, and nouns that signal the **direction** of the effect:
  - Increasingly irreversible losses in cryospheric ecosystems [*Effect–NP*]; irreversible losses in cryospheric ecosystems [*Effect–No_Quantifier*].
  - Reduced food security [*Effect–NP*]; food security [*Effect–No_Quantifier*].
  - Reduction of cryospheric elements [*Effect–NP*]; cryospheric elements [*Effect–No_Quantifier*].
- Remove adjectives, adverbs, and nouns that signal an **aspect** of the effect:
  - Affected water security [*Effect–NP*]; water security [*Effect–No_Quantifier*].
  - Hindered efforts to meet Sustainable Development Goals [*Effect–NP*]; efforts to meet Sustainable Development Goals [*Effect–No_Quantifier*].
  - Severe water scarcity [*Effect–NP*]; water scarcity [*Effect–No_Quantifier*].

"Anthropogenic" is reformulated as *"human activities"* [*Cause–NP*].

B.6 Explicitness

Write "E" or "I" in the [*Explicitness*] column. The distinction between explicit and implicit is as follows:

- Implicit (causal relation is not obvious from the target):
  - The [*Target*] is "/".
  - The [*Target*] is not a "standard" causation-invoking verb like *"cause"*, *"affect"*, and *"increase"*. Examples are *"expose"*, *"cope with"*, *"encourage"*.
  - Causality is embedded in a word, abbreviation, or adjective (e.g., AFOLU, anthropogenic).
- Explicit otherwise.

B.7 Correlation

[*Correlation*] reflects the direction of the correlation between [*Cause–No_Quantifier*] and [*Effect–No_Quantifier*].

- The [*Correlation*] = *Positive*, when:
  - *"An increase in* [*Cause–No_Quantifier*] *causes an increase in* [*Effect–No_Quantifier*]*"*;
  - *"A decrease in* [*Cause–No_Quantifier*] *causes a decrease in* [*Effect–No_Quantifier*]*"*.
- The [*Correlation*] = *Negative*, when:
  - *"An increase in* [*Cause–No_Quantifier*] *causes a decrease in* [*Effect–No_Quantifier*]*"*;
  - *"A decrease in* [*Cause–No_Quantifier*] *causes an increase in* [*Effect–No_Quantifier*]*"*.

B.8 Relation Type

[*Relation type*] reflects the type of relation between [*Cause–No_Quantifier*] and [*Effect–No_Quantifier*].

- The [*Relation type*] = *Positive*, when *"The existence of* [*Cause–No_Quantifier*] *leads to the existence of* [*Effect–No_Quantifier*]*"*.
- The [*Correlation*] = *Negative*, when *"The existence of* [*Cause–No_Quantifier*] *prevents the existence of* [*Effect–No_Quantifier*]*"*.

A popular target that signals a negative relation type is *"avoid"*.

B.9 Abbreviations

If the statement mentions an abbreviation, e.g., CO_2, then this should be resolved in all columns except [*Target*] and included as a set in [*Abbreviations*]. [*Abbreviations*] should be structured as follows, for example, for GHG: [*Abbreviations*] = *GHG = greenhouse gases*. In case two or more abbreviations are used (e.g., GHG and AFOLU), then [*Abbreviations*] = {*GHG = greenhouse gases; AFOLU = Land Use, Land Use Change, and Forestry*} (mention them in the order of appearance in the statement).

Note for [*Target*] Abbreviations should be resolved in all columns EXCEPT [*Target*]; in that column you can use the abbreviation as is.

C Overview Resolved Abbreviations

Table 15 shows an overview of all resolved abbreviations.

Abbr.	Meaning
AFOLU	Agriculture, Forestry, and Other Land Use
CH4	Methane
CO2	Carbon dioxode
CO2-FFI	Carbon dioxide from fossil fuels and industrial processes
CO2-LULUCF	Carbon dioxide emissions from land use, land-use change and forestry
GHG	Greenhouse gases
GDP	Gross domestic product
LDC	Least Developed Countries
N2O	Nitrous oxide
O3	Tropospheric ozone
SIDS	Small Island Developing States
SLCF	Short-Lived Climate Forcers

Table 15: Twelve abbreviations used in IPCC that occur in *ClimateCause*.

D Annotation Description and Disagreement Resolution

D.1 Annotator Description

Both annotators are within the [25-30 years] age range, female, non-native speakers of English (C1 proficiency), and highly educated. They had been involved in another causality annotation campaign prior to this study. Annotator A is European, has a PhD, and holds a master's degree in linguistics (English). Annotator B is South American, pursues a PhD in the field of linguistics and communication, and has a master's degree in the humanities. The third annotator consulted during disagreement resolution is within the [30-40 years] age range, non-binary, non-native speaker of English (C1/C2 proficiency), and highly educated (PhD). No one has an academic background in environmental science.

D.2 Disagreement Resolution

We followed the following procedure to resolve disagreement in the annotations during the first annotation round:

1. Manually identify the causal relations that both annotators retrieved.
2. Compare annotations for these relations:
   (a) Mark and resolve incorrect spans.
   (b) Mark and correct violations against annotation guidelines.
   (c) Mark disagreement between annotators.
3. Look at the causal relations that one annotator annotated but the second did not:
   (a) Include relations that the second annotator did not include but that are very similar to other relation they annotated before, e.g., anthropogenic. They possibly missed these due to a high level of cognitive load of the task.
   (b) Include relations that the missed annotator did not include, but indicated in the comments that they were not sure about this relation.
   (c) Include relations that are implied in abbreviations. The lack of annotating the relations is presumed to be caused by a lack of knowledge. The abbreviations are checked.

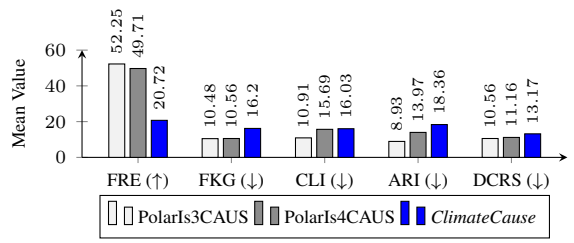


Figure 2: Mean readability scores of statements in PolarIs3CAUS, and PolarIs4CAUS, and *ClimateCause* (ours). Arrows indicate direction of higher readability.

(d) Include causal relations that are not annotated due to violations to the annotation guidelines, e.g., *reductions in CH4 and other ozone precursors* not split in two events.

4. For remaining unresolved relations:

   (a) Consult third annotator, then majority vote.

E Readability of IPCC Reports

IPCC reports are known for low readability (Barkemeyer et al., 2016). We examine whether statements in the ClimateCause dataset are low in readability and whether they are less readable than those in related climate causality datasets (Pineda and Allein, 2025a,b). We measure readability using five established metrics: Flesch Reading Ease (FRE) (Flesch, 1948), Flesch-Kincaid Grade Level (FKG) (Kincaid et al., 1975), Automated Readability Index (ARI) (Kincaid et al., 1975), Coleman-Liau Index (CLI) (Coleman and Liau, 1975), and Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995). Formulae are in Table 16.

Results confirm low readability in *ClimateCause*. Most statements require college-level reading, with half rated very difficult (FRE $\in [10, 30]$) or extremely difficult (FRE $\in [0, 10]$). They are also less readable than those in climate causality datasets from social media (Pineda and Allein, 2025a,b) (Figure 2; Table 17).

F Readability of Reported Causality: Examples

F.1 Co-occurrence between Metrics

The coincidence matrix illustrates the co-occurrence between complexity metrics in *ClimateCause*:

com	*10*				
ex	2	*17*			
nest	2	5	*12*		
corr	4	8	4	*24*	
pol	1	2	-	3	*3*
	com	*ex*	*nest*	*corr*	*pol*

F.2 Statements with High Complexity Scores

Overarching cause/effect $C^{com}(s) = 12$. Statement: *"Similarly, integrated transport and energy infrastructure planning and operations can together reduce the environmental, social, and economic impacts of decarbonising the transport and energy sectors."*

Examples $C^{ex}(s) = 16$. Statement: *"Accelerated support from developed countries and multilateral institutions is a critical enabler to enhance mitigation and adaptation action and can address inequities in finance, including its costs, terms and conditions, and economic vulnerability to climate change. Scaled-up public grants for mitigation and adaptation funding for vulnerable regions, e.g., in Sub-Saharan Africa, would be cost-effective and have high social returns in terms of access to basic energy. Options for scaling up mitigation and adaptation in developing regions include: increased levels of public finance and publicly mobilised private finance flows from developed to developing countries in the context of the USD 100 billion-a-year goal of the Paris Agreement; increase the use of public guarantees to reduce risks and leverage private flows at lower cost; local capital markets development; and building greater trust in international cooperation processes. A coordinated effort to make the post-pandemic recovery sustainable over the long term through increased flows of financing over this decade can accelerate climate action, including in developing regions facing high debt costs, debt distress and macroeconomic uncertainty."*

Nested causality $C^{nest}(s) = 20.93$. Statement: *"Human-caused climate change is a consequence of more than a century of net GHG emissions from energy use, land-use and land use change, lifestyle and patterns of consumption, and production. Emissions reductions in CO2 from fossil fuels and industrial processes (CO2-FFI), due to improvements in energy intensity of GDP and carbon intensity of energy, have been less than emissions increases from rising global activity levels in industry, energy supply, transport, agriculture and buildings. The 10%*

Metric	Formula
Flesch Reading Ease	$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$
Flesch-Kincaid Grade Level	$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$
Automated Readability Index	$4.71\left(\frac{\text{characters}}{\text{words}}\right) + 0.5\left(\frac{\text{words}}{\text{sentences}}\right) - 21.43$
Coleman-Liau Index	$0.0588\left(\frac{\text{letters}}{\text{words}}\right) \times 100 - 0.296\left(\frac{\text{sentences}}{\text{words}}\right) \times 100 - 15.8$
Dale-Chall Readability	$0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$

Table 16: Readability metrics and their formula.

Metric	Median	Mean	Std	Min	Max
ClimateCause (Science-for-policy reports)					
FRE (↑)	**26.95**	**20.72**	**30.39**	**−67.15**	**93.26**
FKG (↓)	**16.06**	**16.20**	**6.05**	**1.80**	**30.27**
CLI (↓)	**15.97**	**16.03**	**5.01**	**1.71**	**25.72**
ARI (↓)	**18.76**	**18.36**	**6.97**	**2.20**	**34.84**
DCRS (↓)	**13.21**	**13.17**	**1.34**	**9.92**	**15.62**
PolarIs3CAUS (Reddit)					
FRE (↑)	53.64	52.25	21.66	−1.28	103.54
FKG (↓)	10.45	10.48	4.28	0.80	20.55
CLI (↓)	10.36	10.91	4.35	1.53	23.17
ARI (↓)	8.77	8.93	5.21	−3.30	21.74
DCRS (↓)	10.43	10.56	2.01	6.01	16.52
PolarIs4CAUS (Twitter/X)					
FRE (↑)	47.79	49.71	23.70	−9.48	103.54
FKG (↓)	10.32	10.56	4.82	0.52	26.10
CLI (↓)	15.20	15.69	5.39	2.98	31.34
ARI (↓)	13.33	13.97	5.73	1.41	34.89
DCRS (↓)	11.47	11.16	2.07	5.50	15.26

Table 17: Readability scores of statements in ClimateCause (ours) from science-for-policy reports, PolarIs3CAUS (Pineda and Allein, 2025a) from Reddit discussions, and PolarIs4CAUS (Pineda and Allein, 2025b) from Twitter/X posts. Arrows indicate direction for higher readability (↑: higher score means higher readability; ↓: lower score means higher readability).

of households with the highest per capita emissions contribute 34–45% of global consumption-based household GHG emissions, while the middle 40% contribute 40–53%, and the bottom 50% contribute 13–15%. An increasing share of emissions can be attributed to urban areas (a rise from about 62% to 67–72% of the global share between 2015 and 2020). The drivers of urban GHG emissions are complex and include population size, income, state of urbanisation and urban form.

Correlation $C^{corr}(s) = 70$. Statement: *"Human-caused climate change is already affecting many weather and climate extremes in every region across the globe. This has led to widespread adverse impacts on food and water security, human health and on economies and society and related losses and damages to nature and people."*

Relation type $C^{pol}(s) = 40$. Statement: *"Trade-offs in terms of employment, water use, land-use competition and biodiversity, as well as access to, and the affordability of, energy, food, and water can be avoided by well-implemented land-based mitigation options, especially those that do not threaten existing sustainable land uses and land rights, with frameworks for integrated policy implementation."*

G Benchmarking Causal Reasoning

G.1 Data and Code Availability

Data We released *ClimateCause* as a CSV-formatted file in a dedicated GitHub repository under the CC-BY 4.0 license, which is intended for use in academic and research settings: https://github.com/laallein/ClimateCause.

Code for data retrieval We released the Python scripts for retrieving the IPCC statements from the Wikibase (SQL).

Code for readability/benchmarking We released the Python scripts for running the readability analyses with the proposed complexity metrics regarding the readability of reported causality in text and the benchmarking experiments for correlation inference and causal chain reasoning under an open-source license in the GitHub repository.

G.2 Implementation Details

We evaluate causal reasoning abilities in GPT5.1; version `gpt-5.1-2025-11-13`, which we access through the OpenAI API. We therefore rely on the OpenAI hardware facilities for running the experiments. We use the default hyperparameter settings. The model's context window size is 400,000; its maximum number of output tokens is 128,000,

and its knowledge cutoff date is 30 September 2024. The experiments were sent to the Batch API, where in total 34,303 requests were made, such that the model handled 9,880,620 completion tokens (Costs: $5.052 input tokens; $11.4 output tokens).

G.3 Evaluation Metrics

Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

F1-score

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

G.4 Breakdown of Results

Per-class performance for chain position identification is reported in Table 18. A full breakdown of performance with each individual prompt is given in Table 19 for correlation inference and in Table 20 for causal chain reasoning.

Class	Precision	Recall	F1
CCR *position*			
none	$0.9915_{\pm 0.0082}$	$0.5357_{\pm 0.1814}$	$0.6800_{\pm 0.1496}$
start	$0.2687_{\pm 0.0551}$	$0.8438_{\pm 0.0950}$	$0.4055_{\pm 0.0680}$
middle	$0.5008_{\pm 0.1714}$	$0.8565_{\pm 0.0887}$	$0.6126_{\pm 0.1191}$
end	$0.3123_{\pm 0.1877}$	$0.7619_{\pm 0.1400}$	$0.4291_{\pm 0.2027}$
CCR+ECI+RC *position*			
none	$0.8842_{\pm 0.0167}$	$0.5679_{\pm 0.1402}$	$0.6818_{\pm 0.1065}$
start	$0.1934_{\pm 0.0632}$	$0.4931_{\pm 0.1752}$	$0.2665_{\pm 0.0818}$
middle	$0.2788_{\pm 0.0631}$	$0.4769_{\pm 0.1726}$	$0.3317_{\pm 0.0604}$
end	$0.2283_{\pm 0.0690}$	$0.4762_{\pm 0.1340}$	$0.3001_{\pm 0.0813}$

Table 18: Per-class performance for chain position identification.

G.5 Prompts

We provide all prompts for correlation inference in Table 23 (CorrI) and 24 (CorrI+RC), and for causal chain reasoning in Table 25 (CCR, membership) 26 (CCR, position), 27 (CCR+ECI+RC, membership), 28 and 29 (CCR+ECI+RC, position).

G.5.1 Examplar Selection for Few-Shot Setting

The following examples are taken from the IPCC report from which the statements in *ClimateCause* are drawn. We made sure to select them from different sections and/or paragraphs and with causal relations that are not included in the dataset.

Prompting Strategy	Precision	Recall	F1
CorrI			
CorrI_0_1	0.7877	0.9449	0.8592
CorrI_0_2	0.7197	0.8881	0.7951
CorrI_0_3	0.9538	0.5680	0.7120
CorrI_F_1	0.9284	0.9593	0.9436
CorrI_F_2	0.7716	0.9243	0.8410
CorrI_F_3	0.9621	0.9621	0.9621
CorrI_CoT_1	0.8565	0.9552	0.9032
CorrI_CoT_2	0.7580	0.8950	0.8208
CorrI_CoT_3	0.9522	0.9604	0.9563
CorrI+RC			
CorrI_RC_0_1	0.7491	0.7091	0.7286
CorrI_RC_0_2	0.7561	0.8537	0.8019
CorrI_RC_0_3	0.9830	0.5955	0.7417
CorrI_RC_F_1	0.9803	0.9681	0.9742
CorrI_RC_F_2	0.8683	0.9415	0.9034
CorrI_RC_F_3	0.9794	0.9828	0.9811
CorrI_RC_CoT_1	0.8986	0.7935	0.8428
CorrI_RC_CoT_2	0.7641	0.8864	0.8207
CorrI_RC_CoT_3	0.9236	0.9570	0.9400

Table 19: Performance results for CorrI strategies.

Example 1 Statement: *"Climate resilient development is enabled by increased international cooperation including mobilising and enhancing access to finance, particularly for developing countries, vulnerable regions, sectors and groups and aligning finance flows for climate action to be consistent with ambition levels and funding needs."* Causal relation: *access to finance → climate resilient development*. Correlation: *positive* [CorrI-F, CorrI+RC-F]. Motivation: Switched order of cause and effect in text; cause and effect are not next to each other; and cause is an example of overarching event in the statement (i.e., increased international cooperation). Position in IPCC report: 4.8.2. International Cooperation and Coordination, p 112.

Example 2 Statement: *"Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options."* Causal relation: *persistent barriers → political feasibility of deploying AFOLU mitigation options*. Correlation: *negative* [CorrI-F, CorrI+RC-F]. List of events: [*persistent barriers*, *region-specific barriers*, *economic feasibility of deploying AFOLU mitigation options*, *political feasibility of deploying AFOLU mitigation options*] [CCR_ECI_member_F, CCR_ECI_position_F]. Motivation: Split of NPs; mention of AFOLU; long effect formulation; no causal chains. Position in IPCC report: 2.3.1. The Gap Between Mitigation Policies, Pledges and

Prompting Strategy	Precision	Recall	F1
CCR *member*			
CCR_member_A_4	0.9083	0.8609	0.8839
CCR_member_A_5	0.3551	0.8522	0.5013
CCR_member_A_6	0.3512	0.9130	0.5072
CCR_member_SN_4	0.8043	0.9652	0.8775
CCR_member_SN_5	0.3631	0.9913	0.5315
CCR_member_SN_6	0.3394	0.9652	0.5023
CCR_member_ML_4	1.0000	0.8174	0.8995
CCR_member_ML_5	0.5067	0.9913	0.6706
CCR_member_ML_6	0.4197	1.0000	0.5913
CCR *position*			
CCR_position_A_4	0.3784	0.9655	0.5437
CCR_position_A_5	0.2348	0.9643	0.3776
CCR_position_A_6	0.2651	0.8462	0.4037
CCR_position_SN_4	0.3919	1.0000	0.5631
CCR_position_SN_5	0.2295	1.0000	0.3733
CCR_position_SN_6	0.2527	0.8846	0.3932
CCR_position_ML_4	0.3404	1.0000	0.5079
CCR_position_ML_5	0.2203	1.0000	0.3611
CCR_position_ML_6	0.2393	0.9655	0.3836
CCR+ECI+RC *member*			
CCR_ECI_member_0_4	0.2774	0.9478	0.4291
CCR_ECI_member_0_5	0.2873	0.8870	0.4340
CCR_ECI_member_0_6	0.2933	0.8696	0.4386
CCR_ECI_member_F_4	0.3363	0.6609	0.4457
CCR_ECI_member_F_5	0.2620	0.9043	0.4063
CCR_ECI_member_F_6	0.5000	0.5043	0.5022
CCR_ECI_member_CoT_4	0.4839	0.6522	0.5556
CCR_ECI_member_CoT_5	0.2875	0.8000	0.4230
CCR_ECI_member_CoT_6	0.2881	0.9043	0.4370
CCR+ECI+RC *position*			
CCR_ECI_position_0_4	0.1942	0.7692	0.3101
CCR_ECI_position_0_5	0.1944	0.7241	0.3066
CCR_ECI_position_0_6	0.1667	0.5185	0.2523
CCR_ECI_position_F_4	0.1429	0.6774	0.2360
CCR_ECI_position_F_5	0.1585	0.5417	0.2453
CCR_ECI_position_F_6	0.2927	0.4444	0.3529
CCR_ECI_position_CoT_4	0.3167	0.6333	0.4222
CCR_ECI_position_CoT_5	0.2727	0.6429	0.3830
CCR_ECI_position_CoT_6	0.1905	0.2500	0.2162

Table 20: Performance results for CCR strategies.

CorrI	CorrI+RC	χ^2	p-value	
0_1	0_1	75.3216	3.9997×10^{-18}	*
0_2	0_2	13.8996	1.9284×10^{-4}	*
0_3	0_3	0.1147	7.3488×10^{-1}	
F_1	F_1	8.0152	4.6388×10^{-3}	*
F_2	F_2	20.3125	6.5770×10^{-6}	*
F_3	F_3	0.0000	1.0000	
CoT_1	CoT_1	84.3005	4.2501×10^{-20}	*
CoT_2	CoT_2	0.5042	4.7768×10^{-1}	
CoT_3	CoT_3	2.9605	8.5320×10^{-2}	

(a) CorrI vs CorrI+RC.

CCR member	CCR position	χ^2	p-value	
A_4	A_4	41.4902	1.1846×10^{-10}	*
A_5	A_5	27.8409	1.3171×10^{-7}	*
A_6	A_6	11.5056	6.9386×10^{-4}	*
SN_4	SN_4	20.0000	7.7442×10^{-6}	*
SN_5	SN_5	13.7931	2.0408×10^{-4}	*
SN_6	SN_6	19.5556	9.7716×10^{-6}	*
ML_4	ML_4	65.1268	7.0232×10^{-16}	*
ML_5	ML_5	45.1765	1.8006×10^{-11}	*
ML_6	ML_6	0.1023	7.4912×10^{-1}	

(b) CCR membership vs CCR position.

CCR+ECI+RC member	CCR+ECI+RC position	χ^2	p-value	
0_4	0_4	99.3103	2.1588×10^{-23}	*
0_5	0_5	59.5044	1.2202×10^{-14}	*
0_6	0_6	101.1500	8.5275×10^{-24}	*
F_4	F_4	27.0096	2.0245×10^{-7}	*
F_5	F_5	37.8125	7.7881×10^{-10}	*
F_6	F_6	0.1552	6.9364×10^{-1}	
CoT_4	CoT_4	1.1228	2.8931×10^{-1}	
CoT_5	CoT_5	104.6639	1.4471×10^{-24}	*
CoT_6	CoT_6	129.0076	6.7558×10^{-30}	*

(c) CCR+ECI+RC membership vs position.

Table 21: McNemar tests across different label set comparisons. Each subtable reports χ^2 and p-values; * indicates significance.

Pathways that Limit Warming to 1.5°C or Below 2°C, p 61.

Example 3 Statement: *"Adaptation options can become maladaptive due to their environmental impacts that constrain ecosystem services and decrease biodiversity and ecosystem resilience to climate change or by causing adverse outcomes for different groups, exacerbating inequity."* List of events: [*adaptation options*, *environmental impacts*, *ecosystem services*, *biodiversity*, *ecosystem resilience to climate change*, *adverse outcomes for different groups*, *inequity*, *maladaptive adaptation options*] [CCR_ECI_member_F, CCR_ECI_position_F]. Motivation: Low readability statement, many causal relations, two chains, start and end of chain not in their relative position in the statement (i.e., end of chain is mentioned before start). Position in IPCC report: 3.2 Long-term Adaptation Options and Limits, p 78.

H Statistical Testing

The results of the McNemar tests between CorrI and CorrI+RC are in Table 21a, between CCR member and position in Table 21b, and between CCR+ECI+RC member and position in Table 21c.

Prompt strategy	H	p-value	ε^2	Sig.
CCR + ECI + RC *member* ($k = 2, n = 512$)				
0_4	47.9251	4.428×10^{-12}	0.0920	*
0_5	73.6743	9.213×10^{-18}	0.1425	*
0_6	84.7026	3.468×10^{-20}	0.1641	*
F_4	50.4575	1.218×10^{-12}	0.0970	*
F_5	9.5520	0.001997	0.0168	*
F_6	46.1151	1.115×10^{-11}	0.0885	*
CoT_4	12.0326	0.0005228	0.0216	*
CoT_5	13.2954	0.0002661	0.0241	*
CoT_6	52.4552	4.402×10^{-13}	0.1009	*
CCR + ECI + RC *position* ($k = 4, n = 512$)				
0_4	15.7083	0.001301	0.0250	*
0_5	4.6300	0.201	0.0032	
0_6	6.2238	0.1012	0.0063	
F_4	6.7837	0.07912	0.0074	
F_5	47.0067	3.464×10^{-10}	0.0866	*
F_6	29.7173	1.583×10^{-6}	0.0526	*
CoT_4	7.9218	0.04765	0.0097	*
CoT_5	7.5165	0.05714	0.0089	
CoT_6	28.1288	3.413×10^{-6}	0.0495	*

Table 22: Kruskal–Wallis test results for total complexity $C(s)$ across CCR+ECI+RC member and position label conditions. Reported are the test statistic (H), p-value, effect size (ε^2), and significance indicator (* for $p < 0.05$).

I On the Use of AI Assistants in Coding and Writing

In this research, artificial intelligence assistants were used to assist in coding (Copilot) and writing (ChatGPT). After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Variant	Prompt	Expected output
CorrI-0-1	"[MASK] in {positive, negative}. There is a [MASK] correlation between e_i and e_j."	{positive, negative}
CorrI-0-2	"[MASK] in {same, opposite}. e_i impact(s) e_j. When we would intervene in e_i, e_i and e_j change in the [MASK] direction."	{same, opposite}
CorrI-0-3	"[MASK] in {increase, decrease}. If e_i (the cause) were to increase, e_j (the effect) would [MASK]."	{increase, decrease}
CorrI-F-1	"There is a positive correlation between access to finance and climate resilient development. There is a negative correlation between persistent barriers and political feasibility of deploying AFOLU mitigation options. There is a [MASK] correlation between e_i and e_j."	{positive, negative}
CorrI-F-2	"Access to finance impacts climate resilient development. When we would intervene in access to finance, access to finance and climate resilient development change in the same direction. Persistent barriers impact political feasibility of deploying AFOLU mitigation options. When we would intervene in persistent barriers, persistent barriers and political feasibility of deploying AFOLU mitigation options change in the opposite direction. e_i impact(s) e_j. When we would intervene in e_i, e_i and e_j change in the [MASK] direction."	{same, opposite}
CorrI-F-3	"If access to finance (the cause) were to increase, climate resilient development (the effect) would increase. If persistent barriers (the cause) were to increase, political feasibility of deploying AFOLU mitigation options (the effect) would decrease. If e_i (the cause) were to increase, e_j (the effect) would [MASK]."	{increase, decrease}
CorrI-CoT-1	"Let's think step by step. [MASK] in {positive, negative}. There is a [MASK] correlation between e_i and e_j."	{positive, negative}
CorrI-CoT-2	"Let's think step by step. [MASK] in {same, opposite}. e_i impact(s) e_j. When we would intervene in e_i, e_i and e_j change in the [MASK] direction."	{same, opposite}
CorrI-CoT-3	"Let's think step by step. [MASK] in {increase, decrease}. If e_i (the cause) were to increase, e_j (the effect) would [MASK]."	{increase, decrease}

Table 23: Prompts for correlation inference without explicit statement context (CorrI).

Variant	Prompt	Expected output
CorrI+RC-0-1	"[MASK] in {positive, negative}. Statement: s. The statement reports a [MASK] correlation between e_i and e_j."	{positive, negative}
CorrI+RC-0-2	"[MASK] in {same, opposite}. Statement: s. e_i impact(s) e_j. Based on the statement, when we would intervene in e_i, e_i and e_j change in the [MASK] direction."	{same, opposite}
CorrI+RC-0-3	"[MASK] in {increase, decrease}. Given the statement: s. If e_i (the cause) were to increase, e_j (the effect) would [MASK]."	{increase, decrease}
CorrI+RC-F-1	"Statement: "Climate resilient development is enabled by increased international cooperation including mobilising and enhancing access to finance, particularly for developing countries, vulnerable regions, sectors and groups and aligning finance flows for climate action to be consistent with ambition levels and funding needs.". The statement reports a positive correlation between access to finance and climate resilient development. Statement: "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options.". The statement reports a negative correlation between persistent barriers and political feasibility of deploying AFOLU mitigation options. Statement: s. The statement reports a [MASK] correlation between e_i and e_j."	{positive, negative}
CorrI+RC-F-2	"Statement: "Climate resilient development is enabled by increased international cooperation including mobilising and enhancing access to finance, particularly for developing countries, vulnerable regions, sectors and groups and aligning finance flows for climate action to be consistent with ambition levels and funding needs.". Access to finance impacts climate resilient development. Based on the statement, when we would intervene in access to finance, access to finance and climate resilient development change in the same direction. Statement: "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options.". Persistent barriers impact political feasibility of deploying AFOLU mitigation options. Based on the statement, when we would intervene in persistent barriers, persistent barriers and political feasibility of deploying AFOLU mitigation options change in the opposite direction. Statement: s. e_i impact(s) e_j. Based on the statement, when we would intervene in e_i, e_i and e_j change in the [MASK] direction."	{same, opposite}
CorrI+RC-F-3	"Given the statement: "Climate resilient development is enabled by increased international cooperation including mobilising and enhancing access to finance, particularly for developing countries, vulnerable regions, sectors and groups and aligning finance flows for climate action to be consistent with ambition levels and funding needs.". If access to finance (the cause) were to increase, climate resilient development (the effect) would increase. Given the statement: "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options.". If persistent barriers (the cause) were to increase, political feasibility of deploying AFOLU mitigation options (the effect) would decrease. Given the statement: s. If e_i (the cause) were to increase, e_j (the effect) would [MASK]."	{increase, decrease}
CorrI+RC-CoT-1	"Let's think step by step. [MASK] in {positive, negative}. Statement: s. The statement reports a [MASK] correlation between e_i and e_j."	{positive, negative}
CorrI+RC-CoT-2	"Let's think step by step. [MASK] in {same, opposite}. Statement: s. e_i impact(s) e_j. Based on the statement, when we would intervene in e_i, e_i and e_j change in the [MASK] direction."	{same, opposite}
CorrI+RC-CoT-3	"Let's think step by step. Statement: s. [MASK] in {increase, decrease}. If e_i (the cause) were to increase, e_j (the effect) would [MASK]."	{increase, decrease}

Table 24: Prompts for correlation inference with explicit statement context.

Variant	Prompt	Expected output
CCR_member_A_4	"You will be given a causal graph. The causal relationships in this causal graph are- G_A. Now answer using this causal graph only, determine whether e_i is part of a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{Yes, No} </Answer>."	{Yes, No}
CCR_member_A_5	"The causal relationships in a causal graph are- G_A. Based on this graph, determine whether e_i belongs to a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Answer with {Yes, No} only."	{Yes, No}
CCR_member_A_6	"The given causal graph includes the following causal relations: G_A. Study this graph carefully, and decide whether the graph contains a causal chain structure that includes e_i. Answer with {Yes, No} only."	{Yes, No}
CCR_member_SN_4	"You will be given a causal graph. The causal relationships in this causal graph are- G_{SN}. Now answer using this causal graph only, determine whether e_i is part of a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{Yes, No} </Answer>."	{Yes, No}
CCR_member_SN_5	"The causal relationships in a causal graph are- G_{SN}. Based on this graph, determine whether e_i belongs to a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Answer with {Yes, No} only."	{Yes, No}
CCR_member_SN_6	"The given causal graph includes the following causal relations: G_{SN}. Study this graph carefully, and decide whether the graph contains a causal chain structure that includes e_i. Answer with {Yes, No} only."	{Yes, No}
CCR_member_ML_4	"You will be given a causal graph. The causal relationships in this causal graph are- G_{ML}. Now answer using this causal graph only, determine whether e_i is part of a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{Yes, No} </Answer>."	{Yes, No}
CCR_member_ML_5	"The causal relationships in a causal graph are- G_{ML}. Based on this graph, determine whether e_i belongs to a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Answer with {Yes, No} only."	{Yes, No}
CCR_member_ML_6	"The given causal graph includes the following causal relations: G_{ML}. Study this graph carefully, and decide whether the graph contains a causal chain structure that includes e_i. Answer with {Yes, No} only."	{Yes, No}

Table 25: Prompts for causal chain reasoning (CCR) benchmarking (***membership***), where e_i is an entity and G_A, G_{SN}, G_{ML} represent causal graphs.

Variant	Prompt	Expected output
CCR_position_A_4	"You will be given a causal graph. The causal relationships in this causal graph are- G_A. Now answer using this causal graph only, determine whether e_i is part of a causal chain and, if yes, which position it holds in that chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{start, middle, end, none} </Answer>."	{start, middle, end, none}
CCR_position_A_5	"The causal relationships in a causal graph are- G_A. Based on this graph, determine whether e_i belongs to a causal chain and, if yes, which position in the chain that event can be found (start, middle, or end). A causal chain is a directed path of at least three nodes in a causal graph. Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_position_A_6	"The given causal graph includes the following causal relations: G_A. Study this graph carefully, and decide whether the graph contains a causal chain structure that includes e_i and which position e_i holds in that chain. Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_position_SN_4	"You will be given a causal graph. The causal relationships in this causal graph are- G_{SN}. Now answer using this causal graph only, determine whether e_i is part of a causal chain and, if yes, which position it holds in that chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{start, middle, end, none} </Answer>."	{start, middle, end, none}
CCR_position_SN_5	"The causal relationships in a causal graph are- G_{SN}. Based on this graph, determine whether e_i belongs to a causal chain and, if yes, which position in the chain that event can be found (start, middle, or end). A causal chain is a directed path of at least three nodes in a causal graph. Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_position_SN_6	"The given causal graph includes the following causal relations: G_{SN}. Study this graph carefully, and decide whether the graph contains a causal chain structure that includes e_i and which position e_i holds in that chain. Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_position_ML_4	"You will be given a causal graph. The causal relationships in this causal graph are- G_{ML}. Now answer using this causal graph only, determine whether e_i is part of a causal chain and, if yes, which position it holds in that chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{start, middle, end, none} </Answer>."	{start, middle, end, none}
CCR_position_ML_5	"The causal relationships in a causal graph are- G_{ML}. Based on this graph, determine whether e_i belongs to a causal chain and, if yes, which position in the chain that event can be found (start, middle, or end). A causal chain is a directed path of at least three nodes in a causal graph. Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_position_ML_6	"The given causal graph includes the following causal relations: G_{ML}. Study this graph carefully, and decide whether the graph contains a causal chain structure that includes e_i and which position e_i holds in that chain. Answer with {start, middle, end, none} only."	{start, middle, end, none}

Table 26: Prompts for causal chain reasoning (CCR) benchmarking (***position***), where e_i is an entity and G_A, G_{SN}, G_{ML} represent causal graphs.

Variant	Prompt	Expected output
CCR_ECI_member_0_4	"Statement: "s". All causal events in the statement: $event\_list$. Based on the statement and the provided list of causal events, determine whether e_i is reported as part of a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Give an answer within <Answer>{Yes, No} </Answer>."	{Yes, No}
CCR_ECI_member_0_5	"Given a statement: "s" and a list of all causal events: $event\_list$. Determine whether e_i belongs to a causal chain reported, either explicitly or implicitly, in the statement. A causal chain is a directed path of at least three nodes in a causal graph. Answer with {Yes, No} only."	{Yes, No}
CCR_ECI_member_0_6	"The following list of causal events $event\_list$ can be found in statement "s". Does the statement report a causal chain structure that includes e_i? A causal chain is a directed path of at least three nodes in a causal graph. Answer with {Yes, No} only."	{Yes, No}
CCR_ECI_member_F_4	"A causal chain is a directed path of at least three nodes in a causal graph. Statement: "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options.". All causal events in the statement: [persistent barriers, region-specific barriers, economic feasibility of deploying AFOLU mitigation options, political feasibility of deploying AFOLU mitigation options]. Based on the statement and the provided list of causal events, is region-specific barriers reported as part of a causal chain? No. Statement: "Adaptation options can become maladaptive due to their environmental impacts that constrain ecosystem services and decrease biodiversity and ecosystem resilience to climate change or by causing adverse outcomes for different groups, exacerbating inequity.". All causal events in the statement: [adaptation options, environmental impacts, ecosystem services, biodiversity, ecosystem resilience to climate change, adverse outcomes for different groups, inequity, maladaptive adaptation options]. Based on the statement and the provided list of causal events, is inequity reported as part of a causal chain? Yes. Statement: "s". All causal events in the statement: $event\_list$. Based on the statement and the provided list of causal events, is e_i reported as part of a causal chain?"	{Yes, No}
CCR_ECI_member_F_5	"A causal chain is a directed path of at least three nodes in a causal graph. Given a statement: "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options." and a list of all causal events: [persistent barriers, region-specific barriers, economic feasibility of deploying AFOLU mitigation options, political feasibility of deploying AFOLU mitigation options]. Determine whether region-specific barriers belongs to a causal chain reported, either explicitly or implicitly, in the statement. Answer: No. Given a statement: "Adaptation options can become maladaptive due to their environmental impacts that constrain ecosystem services and decrease biodiversity and ecosystem resilience to climate change or by causing adverse outcomes for different groups, exacerbating inequity." and a list of all causal events: [adaptation options, environmental impacts, ecosystem services, biodiversity, ecosystem resilience to climate change, adverse outcomes for different groups, inequity, maladaptive adaptation options]. Determine whether inequity belongs to a causal chain reported, either explicitly or implicitly, in the statement. Answer: Yes. Given a statement: "s" and a list of all causal events: $event\_list$. Determine whether e_i belongs to a causal chain reported, either explicitly or implicitly, in the statement. Answer:"	{Yes, No}
CCR_ECI_member_F_6	"A causal chain is a directed path of at least three nodes in a causal graph. The following list of causal events [persistent barriers, region-specific barriers, economic feasibility of deploying AFOLU mitigation options, political feasibility of deploying AFOLU mitigation options] can be found in statement "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options.". Does the statement report a causal chain structure that includes region-specific barriers? No. The following list of causal events [adaptation options, environmental impacts, ecosystem services, biodiversity, ecosystem resilience to climate change, adverse outcomes for different groups, inequity, maladaptive adaptation options] can be found in statement "Adaptation options can become maladaptive due to their environmental impacts that constrain ecosystem services and decrease biodiversity and ecosystem resilience to climate change or by causing adverse outcomes for different groups, exacerbating inequity.". Does the statement report a causal chain structure that includes inequity? Yes. The following list of causal events $event\_list$ can be found in statement "s". Does the statement report a causal chain structure that includes e_i?"	{Yes, No}
CCR_ECI_member_CoT_4	"Statement: "s". All causal events in the statement: $event\_list$. Based on the statement and the provided list of causal events, determine whether e_i is reported as part of a causal chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{Yes, No} </Answer>."	{Yes, No}
CCR_ECI_member_CoT_5	"Given a statement: "s" and a list of all causal events: $event\_list$. Determine whether e_i belongs to a causal chain reported, either explicitly or implicitly, in the statement. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Finally, answer with {Yes, No}."	{Yes, No}
CCR_ECI_member_CoT_6	"The following list of causal events $event\_list$ can be found in statement "s". Does the statement report a causal chain structure that includes e_i? A causal chain is a directed path of at least three nodes in a causal graph. Let's think step by step and answer with {Yes, No}."	{Yes, No}

Table 27: Prompts for causal reasoning benchmarking (CCR+ECI+RC) (***membership***), where s is a statement and $event\_list$ is the list of causal events extracted from it

Variant	Prompt	Expected output
CCR_ECI_ position_0_4	"Statement: "s". All causal events in the statement: $event\_list$. Based on the statement and the provided list of causal events, determine whether e_i is reported as part of a causal chain and, if yes, which position it holds in that chain. A causal chain is a directed path of at least three nodes in a causal graph. Give an answer within <Answer>{start, middle, end, none} </Answer>."	{start, middle, end, none}
CCR_ECI_ position_0_5	"Given a statement: "s" and a list of all causal events: $event\_list$. Determine whether e_i belongs to a causal chain reported, either explicitly or implicitly, in the statement and which position it holds in that chain. A causal chain is a directed path of at least three nodes in a causal graph. Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_ECI_ position_0_6	"The following list of causal events $event\_list$ can be found in statement "s". Does the statement report a causal chain structure that includes e_i and at which position in the chain can the event be found? A causal chain is a directed path of at least three nodes in a causal graph. Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_ECI_ position_F_4	"A causal chain is a directed path of at least three nodes in a causal graph. Statement: "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options.". All causal events in the statement: [persistent barriers, region-specific barriers, economic feasibility of deploying AFOLU mitigation options, political feasibility of deploying AFOLU mitigation options]. Based on the statement and the provided list of causal events, is region-specific barriers reported as part of a causal chain, and, if yes, which position it holds in that chain? none. Statement: "Adaptation options can become maladaptive due to their environmental impacts that constrain ecosystem services and decrease biodiversity and ecosystem resilience to climate change or by causing adverse outcomes for different groups, exacerbating inequity.". All causal events in the statement: [adaptation options, environmental impacts, ecosystem services, biodiversity, ecosystem resilience to climate change, adverse outcomes for different groups, inequity, maladaptive adaptation options]. Based on the statement and the provided list of causal events, is inequity reported as part of a causal chain, and, if yes, which position it holds in that chain? middle. Statement: "s". All causal events in the statement: $event\_list$. Based on the statement and the provided list of causal events, is e_i reported as part of a causal chain, and, if yes, which position it holds in that chain? Answer with {start, middle, end, none} only."	{start, middle, end, none}
CCR_ECI_ position_F_5	"A causal chain is a directed path of at least three nodes in a causal graph. Given a statement: "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options." and a list of all causal events: [persistent barriers, region-specific barriers, economic feasibility of deploying AFOLU mitigation options, political feasibility of deploying AFOLU mitigation options]. Determine whether region-specific barriers belongs to a causal chain reported, either explicitly or implicitly, in the statement and answer with the position the event holds in that chain. Answer: none. Given a statement: "Adaptation options can become maladaptive due to their environmental impacts that constrain ecosystem services and decrease biodiversity and ecosystem resilience to climate change or by causing adverse outcomes for different groups, exacerbating inequity." and a list of all causal events: [adaptation options, environmental impacts, ecosystem services, biodiversity, ecosystem resilience to climate change, adverse outcomes for different groups, inequity, maladaptive adaptation options]. Determine whether inequity belongs to a causal chain reported, either explicitly or implicitly, in the statement and answer with the position the event holds in that chain. Answer: middle. Given a statement: "s" and a list of all causal events: $event\_list$. Determine whether e_i belongs to a causal chain reported, either explicitly or implicitly, in the statement and answer with the position the event holds in that chain. {start, middle, end, none} Answer:"	{start, middle, end, none}
CCR_ECI_ position_F_6	"A causal chain is a directed path of at least three nodes in a causal graph. Answer with {start, middle, end, none} only. The following list of causal events [persistent barriers, region-specific barriers, economic feasibility of deploying AFOLU mitigation options, political feasibility of deploying AFOLU mitigation options] can be found in statement "Persistent and region-specific barriers also continue to hamper the economic and political feasibility of deploying AFOLU mitigation options.". Does the statement report a causal chain structure that includes region-specific barriers and what position does region-specific barriers hold? none. The following list of causal events [adaptation options, environmental impacts, ecosystem services, biodiversity, ecosystem resilience to climate change, adverse outcomes for different groups, inequity, maladaptive adaptation options] can be found in statement "Adaptation options can become maladaptive due to their environmental impacts that constrain ecosystem services and decrease biodiversity and ecosystem resilience to climate change or by causing adverse outcomes for different groups, exacerbating inequity.". Does the statement report a causal chain structure that includes inequity and what position does inequity hold? middle. The following list of causal events $event\_list$ can be found in statement "s". Does the statement report a causal chain structure that includes e_i and what position does e_i hold?"	{start, middle, end, none}

Table 28: Prompts for causal reasoning benchmarking (CCR+ECI+RC) (***position***) – examples and zero-shot variants.

Variant	Prompt	Expected output
CCR_ECI_ position_CoT_4	"Statement: "s". All causal events in the statement: $event\_list$. Based on the statement and the provided list of causal events, determine whether e_i is reported as part of a causal chain and, if yes, which position it holds in that chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Give reasoning and then give an answer within <Answer>{start, middle, end, none} </Answer>."	{start, middle, end, none}
CCR_ECI_ position_CoT_5	"Given a statement: "s" and a list of all causal events: $event\_list$. Determine whether e_i belongs to a causal chain reported, either explicitly or implicitly, in the statement and which position it holds in that chain. A causal chain is a directed path of at least three nodes in a causal graph. Think step by step. Finally, answer with {start, middle, end, none}."	{start, middle, end, none}
CCR_ECI_ position_CoT_6	"The following list of causal events $event\_list$ can be found in statement "s". Does the statement report a causal chain structure that includes e_i and at which position in the chain can the event be found? A causal chain is a directed path of at least three nodes in a causal graph. Let's think step by step and answer with {start, middle, end, none}."	{start, middle, end, none}

Table 29: Prompts for causal reasoning benchmarking (CCR+ECI+RC) (***position***) – CoT variants.