

Assessing LLM Reasoning Through Implicit Causal Chain Discovery in Climate Discourse

Appendix

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

[†] Department of Computer Science, KU Leuven, Belgium

[‡] Institute of Argumentation, Linguistics, and Semiotics (IALS),
Università della Svizzera italiana, Switzerland

Abstract

The Appendix provides the implementation details and outline of the post-processing steps to facilitate the reproduction of the causal chain discovery experiments and evaluation setups A1 - A4. Information related to the human evaluation setup (A5) is presented, including the sets of causal chains and screenshots of the evaluation form.

1. Causal Chain Discovery

1.1. Implementation Details

Checkpoints and Model Size All models were called via the OpenAI API client, with the OpenAI proprietary models (i.e., O1, O1-mini, GPT4o) deployed through the OpenAI Platform¹ and the other, open-source models through NVIDIA NIM and the NVIDIA Developer Program². An overview of all model checkpoints is given in Table 1.

Name	Checkpoint	Size
o1	o1-preview-2024-09-12	/
o1-mini	o1-mini-2024-09-12	/
GPT4o	gpt-4o-2024-11-20	/
Deepseek R1	deepseek-r1	671B
Llama 3.1 Nemotron	llama-3.1-nemotron-ultra-253b-v1	253B
Llama 3 70b	llama3-70b-instruct	70B
Mistral Nemo	mistral-nemo-12b-instruct	12B
Mixtral	mixtral-8x22b-instruct-v0.1	141B
Phi 4-mini	phi-4-mini-instruct	3.8B

Table 1: Checkpoints and number of parameters (= size) of the implemented models.

Hyperparameter Settings We implemented the hyperparameter settings suggested by the model architects, see Table 2.

1.2. Post-Processing of Chains

Post-processing steps for transforming the output in a consistently structured set of causal chains include:

* Current affiliation: Department of Electronics and Information Systems, Ghent University, Belgium. Contact: liesbeth.allein@ugent.be.

¹<https://platform.openai.com/docs/overview>

²<https://build.nvidia.com>

Hyperparameter	Setting
Temperature	0.6
Top p	0.95
Max tokens	4096
Frequency penalty	0
Presence penalty	0
Seed	256

Table 2: Hyperparameter Settings.

- Remove detailed reasoning steps captured between tokens `<think>` and `</think>` prior to chain output.
- Detect use of `<chain>` in the output;
 - If `<chain>` is not in output, detect patterns such as “*causal chain:*”, “*chain 2:*”, and “*causal chain 5*” in the output using regular expression: `r" (?i) causal chain:|chain:|chain \d+:|causal chain \d+:|causal chain \d+|chain \d+";`
 - Remove unwanted asterisks `*` and hyphens `-`;
 - Split in causal chains.
- Detect improper use of `<step>` in the generated chains, i.e., only one occurrence at the end of the chain;
 - If only one `<step>` in chains, extract text before `<step>`.
- Obtain intermediate events in chains by splitting at the `<step>` token.
- Remove unwanted tokens from the events:
 - Token `<\step>`;
 - Arrows `→`;
 - Parentheses `()` using regex pattern matching `r"\ (.*?) "`;

- Numbering at the beginning of the chain using regex pattern matching `r"^\d+\.\s*";`
- “leads to” introducing the event.
- Remove redundant intro (e.g., “*The causal chains are:*” and outro (e.g., “*The chains describe ...*”).
- Remove empty chains, often the result of poor formatting in the output.
- Remove newline `\n`.

2. Human Evaluation Setup (A5)

2.1. Sample Selection Procedure

The selection of the *maintained* and *violated* chains in terms of *chain integrity* follows a waterfall filtering system.

Maintained We first retrieve the CE relations from PolarIs4CAUS that have at least one maintained chain according to at least one LM. A maintained chain is a chain for which the LM answered “yes” in A1 and “no” in A2 for each of its intermediate chains. We then filter the chains with at least four intermediate CE pairs and retain chains considered maintained by at least three different LMs. If two or more chains are retained for the same CE relation, we select the chain with the highest number of LMs considering its integrity maintained. If multiple chains are retained, we select the one that comes first in the generation of o1. This results in 18 maintained chains for 18 CE relations.

Violated We retrieve for those 18 CE relations the invalid chains and select those that have been considered violated by the highest number of LMs. If two or more chains are retained, we select the chain that has a length that is approximately equal to that of the maintained chain to ensure consistency. If multiple chains are retained, we select the one that comes first in the generation of o1.

2.2. Causal Chain Pairs for Human Evaluation

The three sets of causal chain pairs presented to the experts during human evaluation can be found in Table 3, 4, and 5.

2.3. Screenshots of Form Used in Human Evaluation

Screenshots of the Google Form used in the human evaluation in A5 can be found in Figure 1 (instructions given to evaluators), Figure 2 (questions on causal chains), and Figure 3 (final survey).

Evaluating causal chains

Task:

In each question, you will be given two **causal chains**. Each chain starts with an **initial cause** and ends with a **final effect**. A causal chain is a sequence of **at least three distinct events**, where **each event directly causes the next**.

Examples of Causal Chains:

Heavy Rainfall → River Overflows → Flooding of Homes → Evacuation of Residents

Weakened immune system in the winter → Higher chance of viral infections → Pneumonia → Respiratory failure → Death

Your Job:

You will evaluate how **valid** and **logically coherent** each causal chain is.

✅ **A valid** causal chain means:

- Each link in the chain shows a **potential cause-and-effect relationship**. A chain is **invalid** if at least one link does not present a cause-and-effect relationship.
- For example, in the link *Heavy Rainfall* → *River Overflows*, heavy rainfall can actually cause the river to overflow.
- The direction of causality must make sense: the cause leads to the effect, not the other way around. There also exists a temporal flow: *Heavy Rainfall* comes before *River overflows*, not the other way around.

✅ **A logically coherent** chain means:

- The sequence of events is **plausible** and **makes sense in the real world**.
- Each step should follow naturally from the one before it, without missing steps or unlikely jumps.

Figure 1: Human evaluation form: Instructions given to evaluators.

3. On The Use of Existing Artifacts

In terms of data; PolarIs3CAUS and PolarIs4CAUS fall under the following licence: NCCR Evolving Language - Restricted access: for research only. The use of the datasets in this work is consistent with their intended use as they were only used in research contexts.

In terms of pretrained models; we made sure to select pretrained models that have a license that allows for research use.

4. On the Use of AI Assistants in Research or Writing

AI assistants were used in this research 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.

#	Chains	Integrity Main- tained	Integrity Violated
Set #1			
1	Chain A: climate change → earlier seasonal changes → mismatched reproductive cycles → population decline → species at risk of extinction Chain B: climate change → ocean acidification → death of shell-forming organisms → disruption of marine food webs → collapse of fish populations → species at risk of extinction	A	B
2	Chain A: cows → grazing land expansion → deforestation → carbon sink reduction → atmospheric co2 increase → climate change Chain B: cows → feed crop cultivation → fertilizer application → nitrous oxide emissions → greenhouse gas increase → climate change	B	A
3	Chain A: cold → decreased immunity → infection → sepsis → death Chain B: cold → vasoconstriction → increased blood pressure → stroke → death	B	A
4	Chain A: deforestation → increased soil exposure to sunlight → higher soil temperatures → enhanced nutrient volatilization → nutrient loss in the soils Chain B: deforestation → diminished leaf litter → decreased organic matter input → lower soil fertility → nutrient loss in the soils	A	B
5	Chain A: rich countries → industrialization → fossil fuel combustion → enhanced greenhouse effect → rich countries → overconsumption → resource depletion → environmental degradation → climate change Chain B: rich countries → high waste generation → methane emissions from landfills → greenhouse gas accumulation → climate change	B	A
6	Chain A: climate change → warmer sea surface temperatures → stronger hurricanes and cyclones → storm surges → flooding Chain B: climate change → intensified droughts → widespread wildfires → deforestation → soil degradation → reduced water absorption → flooding	A	B

Table 3: Causal Chains – Set #1.

#	Chains	Integrity Main- tained	Integrity Violated
Set #2			
1	Chain A: stoppage of natural winds → reduced evaporation rates → decreased cloud formation → increased solar radiation reaching earth → climate change Chain B: stoppage of natural winds → disruption of ocean currents → reduced heat distribution → global temperature anomalies → climate change	A	B
2	Chain A: fumigation → use of fumigants → production of fumigants → co2 emissions in production → increase in co2 Chain B: fumigation → death of insects → decomposition of insects → release of co2 → increase in co2	B	A
3	Chain A: natural greenhouse effect → warming of earth's surface → increased sea surface temperature → increased phytoplankton activity → release of dimethyl sulfide → increased cloud condensation nuclei → reflection of solar radiation back into space Chain B: natural greenhouse effect → warming of earth's surface → increased evaporation → increased cloud formation → reflection of solar radiation back into space	B	A
4	Chain A: human activity → construction activities → use of heavy equipment → fuel combustion → co2 emissions Chain B: human activity → water production → desalination plants → energy consumption → co2 emissions	A	B
5	Chain A: capitalism → weak environmental regulations → high pollution levels → atmospheric degradation → climate change Chain B: capitalism → energy demand increase → coal power plant proliferation → elevated carbon emissions → climate change	B	A
6	Chain A: climate change → increased occurrence of floods → crop damage → lower harvest yields → reduced availability of fresh foods → escalating prices for scarce fresh foods Chain B: climate change → increased energy costs → higher agricultural production costs → increased food prices → reduced affordability of fresh foods → escalating prices for scarce fresh foods	A	B

Table 4: Causal Chains – Set #2.

#	Chains	Integrity Main- tained	Integrity Violated
Set #3			
1	Chain A: overpopulation → higher water usage → depletion of water sources → reduction in hydroelectric power → increased reliance on fossil fuels → global warming → climate catastrophe Chain B: overpopulation → increased air travel → higher aviation emissions → introduction of pollutants at high altitudes → enhanced greenhouse effect → global warming → climate catastrophe	A	B
2	Chain A: coal-fired plants → emission of greenhouse gases → global warming → increase in environmental heat → dissipation of heat Chain B: coal-fired plants → generation of electricity → operation of electrical devices → production of heat → dissipation of heat	B	A
3	Chain A: excess of co2 → reduced stomatal conductance in plants → improved water use efficiency → increased plant growth → increase in biomass Chain B: excess of co2 → global warming → longer growing seasons → increased plant growth → increase in biomass	B	A
4	Chain A: co2 → ocean acidification → marine life decline → implementation of marine protection policies → improvement in the health of ocean life Chain B: co2 → climate change awareness → environmental policies → reduced ocean pollution → improvement in the health of ocean life	A	B
5	Chain A: climate change → increased co2 levels → ocean acidification → loss of marine life → food scarcity → malnutrition Chain B: climate change → thermal expansion of water → sea level rise → flooding → water contamination → waterborne diseases	B	A
6	Chain A: deforestation → increased carbon emissions → climate change → increased heavy rainfall events → land sliding Chain B: deforestation → loss of vegetation cover → increased surface runoff → soil erosion → deforestation → habitat loss → decreased soil fauna activity → soil degradation → land sliding	A	B

Table 5: Causal Chains – Set #3.

Chain #1

Chain A: overpopulation → higher water usage → depletion of water sources → reduction in hydroelectric power → increased reliance on fossil fuels → global warming → climate catastrophe

Chain B: overpopulation → increased air travel → higher aviation emissions → introduction of pollutants at high altitudes → enhanced greenhouse effect → global warming → climate catastrophe

Chain #1 Select the answer that fits best *

	Yes	No	Can't say
Is Chain A valid?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is Chain B valid?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is Chain A logically coherent?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is Chain B logically coherent?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Chain #1 Which causal chain is most valid and logically coherent? *

☐ Chain A
 ☐ Chain B
 ☐ They are equally (in)valid and (in)coherent
 ☐ Can't say

Figure 2: Human evaluation form: Causal chain evaluation.

Final questions

How difficult/easy was it to **judge** the **validity** of the chains? *

12345

Very difficult☐☐☐☐☐Very easy

How difficult/easy was it to **judge** the **logical coherence** of the chains? *

12345

Very difficult☐☐☐☐☐Very easy

How difficult/easy was it to **compare** the chains? *

12345

Very difficult☐☐☐☐☐Very easy

If given one cause and one effect about climate change, would you be able to infer a valid and logically coherent causal chain that explains the causal mechanism connecting them? *

☐ Yes
 ☐ No
 ☐ Not sure

If 'Yes', how would your chains compare to the chains you have evaluated?

☐ They would be shorter and less detailed
 ☐ They would have a similar length and level of detail
 ☐ They would be longer and more detailed

If 'No' or 'Maybe', briefly motivate why

Jouw antwoord

(a) Survey Form 1/2

(b) Survey Form 2/2

Figure 3: Human evaluation form: Survey.