Nuance Matters: Probing Epistemic Consistency in Causal Reasoning

Shaobo Cui[1], **Junyou Li**[2], **Luca Mouchel**[1], **Yiyang Feng**[1], **Boi Faltings**[1]

[1]EPFL, Switzerland
[2]University of Waterloo, Canada
shaobo.cui@epfl.ch, j2626li@uwaterloo.ca, luca.mouchel@epfl.ch, yiyang.feng@epfl.ch, boi.faltings@epfl.ch

Abstract

Previous research on causal reasoning often overlooks the subtleties crucial to understanding causal reasoning. To address this gap, our study introduces the concept of *causal epistemic consistency*, which focuses on the self-consistency of Large Language Models (LLMs) in differentiating intermediates with nuanced differences in causal reasoning. We propose a suite of novel metrics – intensity ranking concordance, cross-group position agreement, and intra-group clustering – to evaluate LLMs on this front. Through extensive empirical studies on 21 high-profile LLMs, including GPT-4, Claude3, and LLaMA3-70B, we have favoring evidence that current models struggle to maintain epistemic consistency in identifying the polarity and intensity of intermediates in causal reasoning. Additionally, we explore the potential of using internal token probabilities as an auxiliary tool to maintain causal epistemic consistency. In summary, our study bridges a critical gap in AI research by investigating the self-consistency over fine-grained intermediates involved in causal reasoning.

Code — https://github.com/cui-shaobo/causal-consistency
Extended version — https://arxiv.org/abs/2409.00103

1 Introduction

Previous studies in causal reasoning have primarily focused on discovering or determining the existence of a causal relationship between two variables (Roemmele, Bejan, and Gordon 2011; Cui et al. 2024b). However, these causal relationships are not always absolute. They can be heavily influenced by additional intermediate factors, which may vary in both polarity and intensity (Fitzgerald and Howcroft 1998; Bauman et al. 2002). The polarity of these intermediates indicates whether they support or defeat (oppose) the original causal relationship, while their intensity determines the strength of this supporting or defeating influence.

Forming fine-grained differentiation is essential for precise causal modeling (Iwasaki and Simon 1994); however, it is insufficient for LLMs to merely generate these intermediates. It is as equally important to ensure that these intermediates are reliable and credible (Shi et al. 2023). One method to verify this is through assessing the consistency of LLMs' perception of the intermediates. We posit that if LLMs can correctly differentiate their generated intermediates based on varying polarities and intensities, these intermediates are self-consistent and thus, more reliable for making predictions and decisions. Drawing from this insight, our study proposes the concept of "**causal epistemic consistency**":

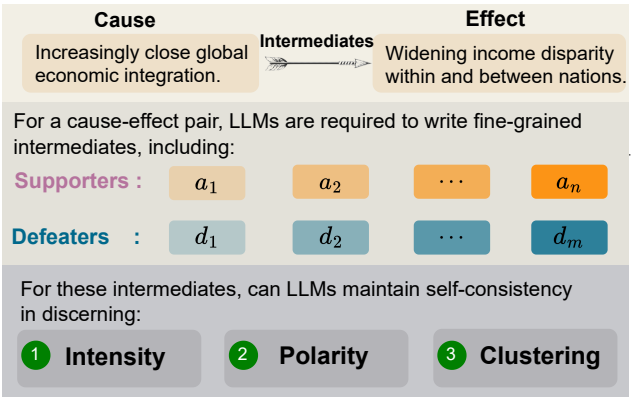

Figure 1: Overview of the evaluation framework for causal epistemic consistency. The first step involves instructing LLMs to generate fine-grained intermediates that influence a given causal relationship differently. The second step requires LLMs to rank their own generations based on their causal nuance. Finally, the proposed metrics are used to assess the self-consistency between ranking and generation, i.e., the LLMs' causal epistemic consistency.

Definition 1 (Causal epistemic consistency) *Causal epistemic consistency refers to an LLM's ability to maintain self-consistency in differentiating its generated intermediates in three aspects: (i) discerning **intensity**: accurately assessing the intensity nuance in their causal impact. (ii) differentiating **polarity**: effectively distinguishing between supporting and defeating intermediates, and (iii) forming cohesive **clusters**: creating well-separated clusters of intermediates based on their polarity and intensity.*

To quantify LLMs' ability to maintain causal epistemic consistency in the aforementioned aspects, we introduce a suite of novel metrics. These metrics include (i) Inten-

sity ranking concordance, which measures the models' self-consistency in ranking self-generated intermediates with varying intensity; (ii) Cross-group position (CGP) agreement, which indicates the models' consistency in determining the polarity of intermediates, specifically whether they support or defeat the original causal relationship; and (iii) Intra-group clustering (IGC), which assesses models' consistency to rank its generated intermediates of the same type closely together. We illustrate the evaluation framework of causal epistemic consistency in Figure 1.

To unravel the causal epistemic consistency of current LLMs, our empirical study evaluates 21 high-profile LLMs, including the renowned closed-source GPT, Claude, and Gemini series, alongside various scales of cutting-edge open-source alternatives such as Gemma (2B and 7B) (Mesnard et al. 2024), LLaMA2 (7B, 13B, and 70B) (Touvron et al. 2023), Phi-3 (3.8B, 7B, and 14B) (Abdin et al. 2024), and LLaMA3 (8B and 70B) (Meta 2024). Our findings reveal their striking incompetence in keeping causal epistemic consistency. Remarkably, even the advanced GPT-4 model performs unsatisfactorily. This underscores the complexities and challenges these models face in maintaining causal consistency and capturing causal nuances.

Furthermore, we explore whether internal token probability can serve as a useful signal for LLMs to maintain causal epistemic consistency. Our comprehensive empirical study highlights the application scope of internal token probability for LLMs to maintain causal epistemic consistency.

To summarize, our contributions are fourfold:

1. **Introduction of Causal Epistemic Consistency:** We propose the novel concept of causal epistemic consistency over fine-grained intermediates in causal reasoning, emphasizing self-consistency in differentiating the nuances hidden in fine-grained intermediates.
2. **Development of Evaluation Metrics:** We introduce a comprehensive suite of metrics designed to assess LLMs' causal epistemic consistency, covering aspects of intensity ranking concordance, cross-group position agreement, and intra-group clustering.
3. **Extensive Empirical Evaluation:** We assess the performance of 21 LLMs on their causal epistemic consistency, highlighting their deficiencies in maintaining causal epistemic consistency.
4. **Internal Token Probability Exploration:** We investigate the potential of using internal token probabilities as an auxiliary tool to help LLMs maintain causal epistemic consistency and highlight its application scope.

2 Task Definition

2.1 Problem Formulations

Causal epistemic consistency measures an LLM's self-consistency between generating fine-grained intermediates and subsequently ranking those fine-grained intermediates.

Specifically, in the generation phase, for a defeasible cause-effect pair (C, E), an LLM is tasked with generating an ordered sequence $\mathcal{I}$ of fine-grained intermediates, consisting of a subsequence $\mathcal{D} = (\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m)$ as the defeater group and a subsequence $\mathcal{A} = (\mathcal{I}_{m+1}, \mathcal{I}_{m+2}, \cdots, \mathcal{I}_{m+n})$ as the supporter group. Each individual intermediate changes the causal strength of (C, E) differently. Specifically, the causal influence of these intermediates is expected in the following order:

$$\begin{aligned} \mathcal{CS}(E|C \oplus \mathcal{I}_1) \leq \cdots \leq \mathcal{CS}(E|C \oplus \mathcal{I}_m) \\ \leq \mathcal{CS}(E|C) \\ \mathcal{CS}(E|C \oplus \mathcal{I}_{m+1}) \leq \cdots \leq \mathcal{CS}(E|C \oplus \mathcal{I}_{m+n}) \end{aligned} \tag{1}$$

where $\mathcal{CS}(E|C)$ measures the causal strength (Luo et al. 2016; Zhang et al. 2022), quantifying the likelihood that the cause event C would lead to the occurrence of the effect event E. [1] The $\oplus$ means the combination of two events.

Subsequently, in the ranking phase, the same LLM is asked again to rank its own generated intermediates $\mathcal{I}$, obtaining $\mathcal{I}'$, a permutation of $\mathcal{I}$. Ideally, an LLM with perfect causal epistemic consistency should have $\mathcal{I} = \mathcal{I}'$, satisfying the requirements of intensity, polarity, and clustering perfectly.

2.2 Key Research Questions

The study addresses three primary research questions:

- **RQ I**: How can we comprehensively measure the ability of LLMs to maintain the epistemic consistency over fine-grained intermediates in causal reasoning?
- **RQ II**: How well do current LLMs, with varying architectures and scales, maintain their causal epistemic consistency?
- **RQ III**: Are there any alternatives to prompting for LLMs to maintain causal epistemic consistency?

To answer **RQ I**, we propose novel metrics introduced in Section 3, which not only serve our specific study but also have broader applications across various tasks. In Section 4, we dive into the performance of twenty-one leading LLMs, exploring their ability to maintain epistemic consistency, thereby addressing **RQ II**. Lastly, in Section 5 , we assess whether internal token probability offers a more effective—or perhaps less effective—alternative to prompting for preserving causal epistemic consistency in LLMs, answering **RQ III**.

3 Metrics for Measuring Causal Epistemic Consistency

To evaluate the causal epistemic consistency of LLMs from the aspects of intensity, polarity, and clustering, we propose three types of automatic metrics: intensity ranking concordance, cross-group position agreement, and intra-group clustering. A graphical illustration of these metrics is shown in Figure 2. The mathematical notations below are consistent with Section 2.1.

[1] In this context, we assume that only one fine-grained intermediate is active for a cause-effect pair at a time. This design choice reflects the reality that a single argument is more often responsible for influencing the causal relationship than multiple arguments acting simultaneously.

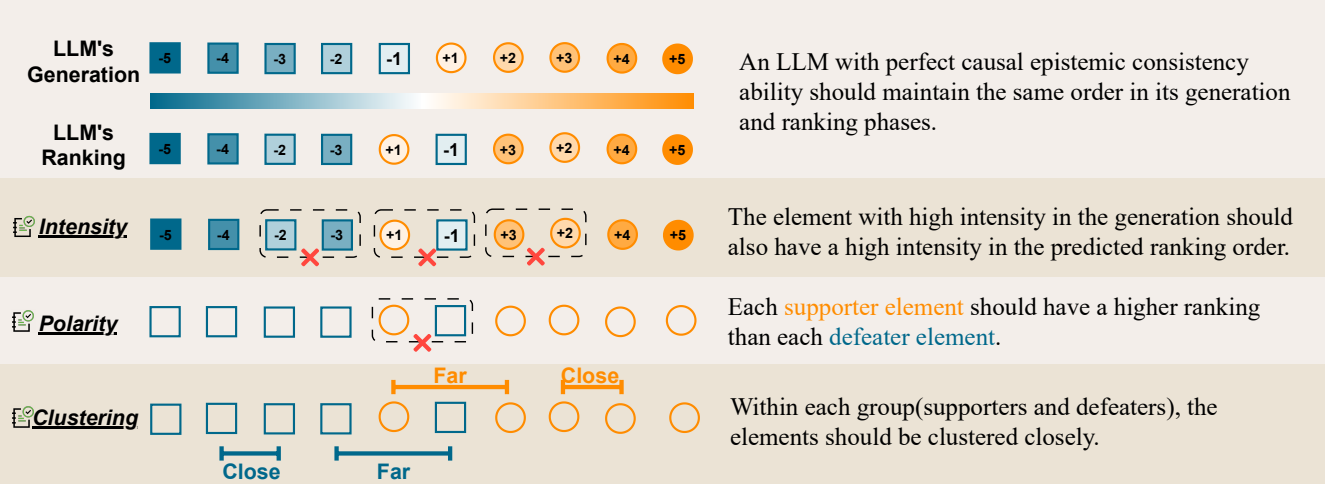

Figure 2: Illustration of the proposed metrics from three aspects: intensity (Section 3.1), polarity (Section 3.2), and clustering (Section 3.3). These metrics measure the self-consistency of LLMs in generating and ranking supporting (◯) and defeating (□) intermediates with varying intensities. Numbers -5, -4, ..., +4, +5 indicate the intensity of the generated intermediates, with the lowest value (-5) being the strongest generated defeater and the highest value (+5) the strongest generated supporter.

3.1 Intensity: Intensity Ranking Concordance

To assess the concordance between the order from the generation phase and the order from the ranking phase of these fine-grained intermediates, we leverage the Kendall Tau distance (Kendall 1938). This metric quantifies the similarity between two orders by counting the number of pairwise agreements and disagreements. For a sequence $\mathcal{I}$ of LLM-generated intermediates and its permutation $\mathcal{I}'$ ranked by the same LLM, a pair of elements from $\mathcal{I}$ is called *concordant* if they appear in the same order in both $\mathcal{I}$ and $\mathcal{I}'$. Conversely, the pair is called *discordant* if their order is reversed in $\mathcal{I}'$ compared to $\mathcal{I}$. The Kendall Tau τ is calculated as:

$$\tau = \frac{(\#\text{ concordant pairs}) - (\#\text{ discordant pairs})}{k(k-1)/2} \quad (2)$$

where k is the number of elements in the list, and $k(k-1)/2$ is the total number of pairs. The metric ranges from -1 to 1, where 1 indicates that these two lists are identical; -1 indicates completely reversed rankings; and values close to 0 indicate no association between the two lists. For our task, we have three intensity ranking concordance metrics: τ-$\mathcal{A}$, τ-$\mathcal{D}$, and τ-all, which evaluate the intensity ranking concordance within the supporter group, the defeater group, and the entire sequence of intermediates, respectively.

3.2 Polarity: Cross-Group Position (CGP)

To assess the relative positioning of elements between these two polarities–the defeater group $\mathcal{D}$ and the supporter group $\mathcal{A}$–we propose the Cross-Group Position (CGP) metric. This metric penalizes instances where elements from $\mathcal{A}$ are ranked lower than those from $\mathcal{D}$ [2]. Specifically, CGP is defined as:

$$\text{CGP}(\mathcal{I}', \mathcal{A}, \mathcal{D}) = 1 - \frac{\sum_{a\in\mathcal{A}}\sum_{d\in\mathcal{D}} \mathbb{1}[\text{index}(a)<\text{index}(d)]}{|\mathcal{A}|\times|\mathcal{D}|}$$

where $\text{index}(x)$ denotes the index of element x in the ranked sequence $\mathcal{I}'$. $\mathbb{1}[\cdot]$ denotes the indicator function that is set to 1 if the condition is true and 0 otherwise. CGP measures how often elements from $\mathcal{A}$ precede the elements of $\mathcal{D}$ in the ranked sequence $\mathcal{I}'$. It is normalized to the range [0, 1] by dividing with the maximum possible violations, i.e., $|\mathcal{A}|\times|\mathcal{D}|$. Higher values indicate better differentiation between groups $\mathcal{A}$ and $\mathcal{D}$.

3.3 Clustering: Intra-Group Clustering (IGC)

In this subsection, we introduce Intra-Group Clustering (IGC), a metric for LLMs' causal epistemic consistency by assessing the clustering degree of supporting and defeating intermediates. The intuition behind IGC is that all defeaters and all supporters should form cohesive clusters, with a minimal number of polarity changes (from supporting to defeating, or vice versa) when iterating the sequence.

Clustering Distance Based on Polarity Change. Given the LLM-ranked intermediates $\mathcal{I}'$, we define L_i to to be a binary polarity that indicates whether $\mathcal{I}'$ is in $\mathcal{A}$ or $\mathcal{D}$, while the cardinality $|L_i|$ refers to the number of intermediates sharing the same polarity as $\mathcal{I}'_i$. $d(i,j)$ is the sequence clustering distance between $\mathcal{I}'_i$ and $\mathcal{I}'_j$, calculated as follows:

$$d(i,j) = \sum_{k=i}^{j-1} \mathbb{1}[L_k \neq L_{k+1} \wedge L_{k+1} \neq L_i] \quad (3)$$

where $i < j$. The distance is based on the number of polarity changes, excluding reversions to the initial polarity.

IGC: A Measure of Clustering Quality in Sequence. With the distance based on polarity change, we use the silhouette score (Rousseeuw 1987; Shahapure and Nicholas

[2] We define the index of the strongest defeater to be the lowest and the strongest supporter to be the highest, consistent with Section 2.1.

Aspect	*Intensity Ranking Concordance*			*Cross-Group Position*	*Intra-Group Clustering*
	τ-$\mathcal{A}$ ↑	τ-$\mathcal{D}$ ↑	τ-all ↑	CGP ↑	IGC ↑
Closed-source LLMs					
GPT-3.5 Turbo	0.074 ± 0.429	0.045 ± 0.407	0.304 ± 0.409	0.750 ± 0.329	0.762 ± 0.244
GPT-4	0.384 ± 0.413	0.203 ± 0.440	0.587 ± 0.347	0.911 ± 0.235	0.916 ± 0.176
GPT-4 Turbo	0.397 ± 0.541	0.226 ± 0.459	0.526 ± 0.510	0.849 ± 0.330	0.942 ± 0.151
GPT-4o mini	0.142 ± 0.444	0.154 ± 0.418	0.472 ± 0.375	0.865 ± 0.281	0.889 ± 0.196
GPT-4o	0.317 ± 0.466	0.229 ± 0.426	0.637 ± 0.266	**0.964 ± 0.164**	**0.978 ± 0.099**
Claude 3 Haiku	0.120 ± 0.429	0.069 ± 0.388	0.406 ± 0.344	0.828 ± 0.270	0.809 ± 0.234
Claude 3 Sonnet	0.272 ± 0.429	0.046 ± 0.423	0.533 ± 0.290	0.916 ± 0.204	0.893 ± 0.195
Claude 3 Opus	0.509 ± 0.457	0.381 ± 0.451	**0.688 ± 0.342**	0.941 ± 0.204	0.957 ± 0.131
Claude 3.5 Sonnet	**0.610 ± 0.507**	**0.440 ± 0.501**	0.662 ± 0.492	0.885 ± 0.286	0.932 ± 0.159
Gemini 1.5 Flash	0.108 ± 0.451	0.115 ± 0.412	0.429 ± 0.362	0.842 ± 0.274	0.838 ± 0.225
Gemini 1.5 Pro	0.475 ± 0.435	0.165 ± 0.463	0.587 ± 0.326	0.900 ± 0.212	0.875 ± 0.205
Open-source LLMs					
Gemma-2B	-0.021 ± 0.412	0.001 ± 0.410	-0.002 ± 0.245	0.502 ± 0.190	0.468 ± 0.083
Gemma-7B	-0.006 ± 0.392	0.016 ± 0.389	0.085 ± 0.256	0.575 ± 0.203	0.484 ± 0.122
LLaMA2-7B	-0.018 ± 0.406	0.001 ± 0.412	-0.029 ± 0.261	0.477 ± 0.200	0.475 ± 0.092
LLaMA2-13B	-0.000 ± 0.411	0.026 ± 0.417	0.072 ± 0.256	0.560 ± 0.197	0.480 ± 0.109
LLaMA2-70B	0.012 ± 0.409	0.010 ± 0.434	0.234 ± 0.349	0.707 ± 0.271	0.629 ± 0.215
Phi-3 Mini (3.8B)	0.135 ± 0.431	0.012 ± 0.393	0.300 ± 0.336	0.740 ± 0.275	0.659 ± 0.222
Phi-3-Small (7.4B)	0.092 ± 0.443	0.204 ± 0.422	0.347 ± 0.348	0.753 ± 0.254	0.672 ± 0.220
Phi-3 Medium (14B)	-0.056 ± 0.441	0.154 ± 0.406	0.356 ± 0.367	0.801 ± 0.286	0.801 ± 0.230
LLaMA3-8B	0.030 ± 0.444	0.139 ± 0.436	0.273 ± 0.387	0.712 ± 0.285	0.639 ± 0.217
LLaMA3-70B	**0.357 ± 0.469**	**0.343 ± 0.419**	**0.586 ± 0.415**	**0.887 ± 0.274**	**0.923 ± 0.177**
Random					
Random	-0.003 ± 0.409	0.005 ± 0.406	-0.008 ± 0.249	0.496 ± 0.192	0.467 ± 0.077

Table 1: Empirical study of LLMs on the proposed metrics for causal epistemic consistency.

2020) to measure how similar an element is to its own cluster compared to other clusters in sequence:

$$s(i) = \frac{d_{nc}(i) - d_{ic}(i)}{\max(d_{ic}(i), d_{nc}(i))} \tag{4}$$

where $d_{ic}(i)$ and $d_{nc}(i)$ are the intra-cluster distance and nearest cluster distance for each intermediate $\mathcal{I}'_i$.

1. The intra-cluster distance $d_{ic}(i)$ captures the mean distance between $\mathcal{I}'_i$ and all other intermediates belonging to the same group, reflecting *internal cohesion*. It is calculated as:

$$d_{ic}(i) = \frac{1}{|L_i| - 1} \sum_{L_j = L_i, \mathcal{I}'_j \neq \mathcal{I}'_i} d(i, j). \tag{5}$$

2. The nearest cluster distance $d_{nc}(i)$ captures the mean distance between $\mathcal{I}'_i$ and all other points belonging to a different group, demonstrating the level of *separation* from other clusters. It is calculated as:

$$d_{nc}(i) = \frac{1}{|\mathcal{I}'| - |L_i|} \sum_{L_j \neq L_i} d(i, j). \tag{6}$$

The final Intra-Group Clustering (IGC) metric is computed as the average clustering of all elements:

$$\text{IGC} = \frac{1}{|\mathcal{I}'|} \sum_{i=1}^{|\mathcal{I}'|} s(i). \tag{7}$$

Range and Implications of IGC. The range of $s(i)$ is $[-1, 1]$: (i) Close to 1: The element is near its own group and far from the neighboring groups; (ii) Close to 0: The element is on the border between its cluster and a neighboring cluster. (iii) Close to -1: The element is in the wrong cluster. IGC quantifies the quality of cluster assignments, with a high score indicating well-clustered sequences. It is a general metric applicable to various contexts related to sequence clustering. Further details are in Appendix C.1.

4 Causal Epistemic Consistency of LLMs

4.1 Experimental Setup

Foundational Dataset. To ensure the defeasibility of causal pairs, allowing models to generate intermediates with varying polarity and intensity, we utilize the test dataset of

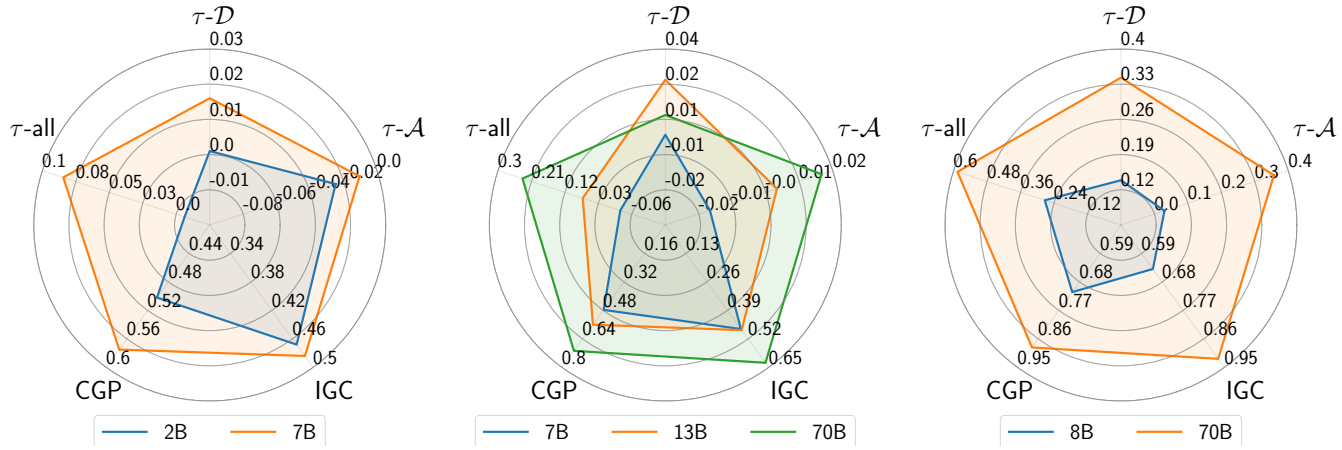

Figure 3: Radar charts comparing performance of various LLM architectures and sizes: Gemma (left), LLaMA2 (middle), and LLaMA3 (right) in maintaining causal epistemic consistency. Each colored line denotes a distinct model size.

δ-CAUSAL (Cui et al. 2024b) as our foundational dataset, which comprises 1,970 defeasible cause-effect pairs.

Three-Phase Assessment for LLMs' Causal Epistemic Consistency. There are three main phases in our experiments: (i) Intermediate generation: We provide LLMs with a single cause-effect pair and two preliminary intermediates: one supporting and one defeating. For each supporter and defeater, we instruct the LLMs to generate two weaker and two stronger intermediates. As a result, we compile a total of 10 intermediates as sequence $\mathcal{I}$, divided into two subsequences: subsequence $\mathcal{D}$ comprised of $m = 5$ intermediates that challenge the cause-effect relationship with differing intensities; and subsequence $\mathcal{A}$ consisting of $n = 5$ supporting intermediates that reinforce the cause-effect pair, also with varying intensities. The prompt for generating these fine-grained intermediates is presented in Figure 6; (ii) Intermediate ranking: From these generated intermediates, we use the same LLM to rank the intermediates to identify their polarities (supporting or defeating) and intensity. The prompt for ranking these fine-grained intermediates is presented in Figure 7; and (iii) Evaluation: Based on the actual order of generated intermediates in the first phase and the predicted ranking order in the second phase, we evaluate the causal epistemic consistency from the perspectives of Intensity Ranking Concordance (τ-$\mathcal{A}$, τ-$\mathcal{D}$, τ-all), Cross-Group Position (CGP) agreement, and Intra-Group Clustering (IGC).

Backbone Models. We assess a comprehensive suite of LLMs for causal epistemic consistency. Our evaluation includes: (i) 11 Closed-source models: GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, GPT-4o, GPT-4o mini, Claude 3 (Haiku, Sonnet, and Opus), Claude 3.5 (Sonnet) (Anthropic 2024), Gemini 1.5 (Flash and Pro) (Gemini-Team 2024); (ii) 10 Open-source models: Gemma (2B and 7B) (Mesnard et al. 2024), LLaMA2 (7B, 13B, and 70B) (Touvron et al. 2023), Phi-3 (mini, small, and medium) (Abdin et al. 2024), and LLaMA3 (8B and 70B) (Meta 2024).

4.2 Experimental Results

Table 1 presents a quantitative comparison of different models on causal epistemic consistency.

- **Closed-source models generally outperform open-source models**: For instance, GPT-4o achieves a τ-all score of 0.632, a CGP score of 0.962, and an IGC score of 0.973, whereas LLaMA3-70B, the best-performing open-source model, only achieves a τ-all score of 0.586, a CGP score of 0.887, and an IGC score of 0.923.
- **Maintaining consistency in intensity is more challenging than achieving consistency in polarity and clustering**: The patterns across different metrics are consistent among different models, suggesting that while LLMs can effectively maintain consistency over differentiating between supporting and defeating intermediates and clustering intermediates of the same polarity together, they find it more challenging to maintain consistent intensity rankings. Namely, achieving consistency over the nuances of causal intensity remains difficult.

4.3 Does a Larger Model Scale Mean Better Causal Epistemic Consistency?

Previous works (Kaplan et al. 2020; Hoffmann et al. 2024) have shown that with the increase in model scale, the improvement in performance follows a power-law relationship. However, the effectiveness of 'just scaling' for general causal understanding, especially in the context of causality, has become a subject of intense debate (Zečević et al. 2023).

Inspired by this question, we investigate whether increasing the model scale improves the causal epistemic consistency of LLMs. Since this model scale study is only possible for models available in multiple sizes, we conduct experiments with: (i) Gemma at sizes of 2B and 7B; (ii) LLaMA2 at sizes of 7B, 13B, and 70B; and (iii) LLaMA3 at sizes of 8B and 70B. The experimental results are presented in Figure 3. From these results, we clearly observe that **an increase in model size generally enhances causal epistemic consistency**. For instance, LLaMA2 and LLaMA3 demon-

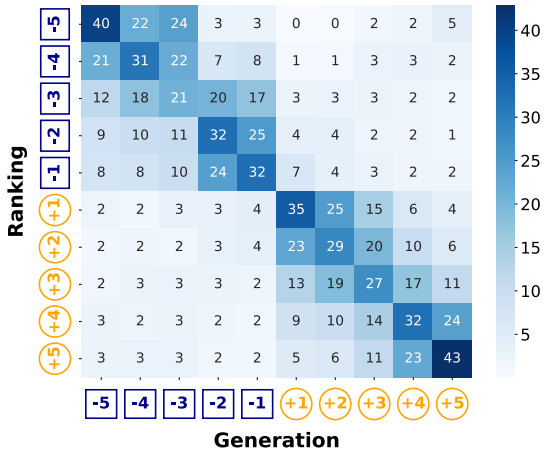

Figure 4: Visualization of LLaMA3-70B's alignment of intermediates' predicted ranking versus their generation phase ranking, indicating the models' self-consistency in intensity, polarity, and clustering. Each matrix element (i, j) indicates the percentage of instances where an intermediate ranked at position i during the generation phase was ranked at position j during the ranking phase. For example, (-3 , +4) indicates the percentage of instances with a label of defeater with an intensity of 3 in the generation phase that was ranked as the supporter with an intensity of 4 during the ranking phase.

strate significant improvements at larger scales, particularly at 70B, where the causal epistemic consistency scores are notably higher compared to their smaller-scale counterparts.

4.4 Visualization of Causal Epistemic Consistency

We plot the causal epistemic consistency matrices of LLaMA3-70B in Figure 4. In these matrices, the x-axis from left to right and the y-axis from top to bottom correspond to -5 -4 -3 -2 -1 +1 +2 +3 +4 +5 , where the square symbol -* represents defeaters while the circle symbol +* represents supporters. The numbers inside the symbols indicate the supporting or defeating intensity, with larger absolute values signifying stronger intensity (i.e., -5 is the strongest defeater and +5 is the strongest supporter). These matrices visualize how well the models maintain causal epistemic consistency by comparing the labels of intermediates of the generation phase with the predicted labels in the ranking phase.

The confusion matrices of other models are presented in Appendix D. From the results of the best closed-source and open-source models, we have the following observations:

- **Diagonal Dominance**: Higher values along the diagonal indicate better causal epistemic consistency. This dominance shows that the model often maintains consistency in both polarity and intensity by correctly matching the labels of intermediates from the generation phase to the ranking phase.
- **Off-Diagonal Elements**: These off-diagonal elements represent the number of instances where the predicted labels during the ranking phase diverge from the labels during the generation phase. Higher values in these cells suggest cases where the model struggles to maintain consistency. For instance, a higher value far from the diagonal indicates a more significant discrepancy between the ranking and generation phases, reflecting lower causal epistemic consistency due to overestimation or underestimation of the intensity of the generated intermediates.
- **Cluster Separation**: These matrices also indicate that these two models cluster supporting and defeating intermediates well, as shown by lower values in the lower left and upper right corners.

5 Beyond Prompting: Leveraging Internal Token Probability

This section explores using internal token probability as an alternative to the prompting method in Section 4 for maintaining causal epistemic consistency.

5.1 Internal Token Probability

Internal token probability has proven to be a reliable indicator for sequence correlation estimation (Malinin and Gales 2021; Farquhar et al. 2024; Cui et al. 2024b). For each cause-effect pair (C, E) and any supporting or defeating intermediate $\mathcal{I}_j$, we utilize the token probabilities p to estimate the causal strength $\mathcal{CS}(E|C \oplus \mathcal{I}_j)$ in Section 2.1:

$$\mathcal{CS}(E|C \oplus \mathcal{I}_j) = \prod_i p(E_i|C \oplus \mathcal{I}_j, w, E_{<i}) \quad (8)$$

where E_i is the i_{th} token of E and $E_{<i}$ is the first $i-1$ tokens of E. $p(E_i|C \oplus \mathcal{I}_j, w, E_{<i})$ is the internal (conditional) token probability. The conjunction word w connects the combination of the cause and the intermediate to the effect, and explicitly indicates the causation such as "because" and "therefore".

5.2 Experimental Setup

Models and Datasets. As closed-source models often do not provide a `logprob` API usage [3], our investigation resorts to open-source LLMs including Gemma (2B and 7B) (Mesnard et al. 2024), LLaMA2 (7B, 13B, and 70B) (Touvron et al. 2023), Phi-3 (3.8B, 7B, and 14B), and LLaMA3 (8B and 70B). We use the same foundation dataset described in Section 4.1.

Three-Phase Assessment. The experiment in this section involves three phases: (i) Intermediate generation: This phase involves generating a sequence of intermediates, $\mathcal{I}$, following the same procedure described in Section 4.1; (ii) Intermediate ranking based on conditional token probability: In this phase, we calculate the causal strength based on the conditional token probability using $\{\mathcal{CS}(E|C \oplus \mathcal{I}_j)|\mathcal{I}_j \in \mathcal{I}\}$. (iii) Evaluation: We assess the models' causal epistemic consistency using rankings from the generation phase and conditional probability values, based on the proposed metrics in Section 3.

[3] Even though `logprob` is provided, users cannot compute the probability of an arbitrary token given an input. The potential reason might be to avoid model distillation.

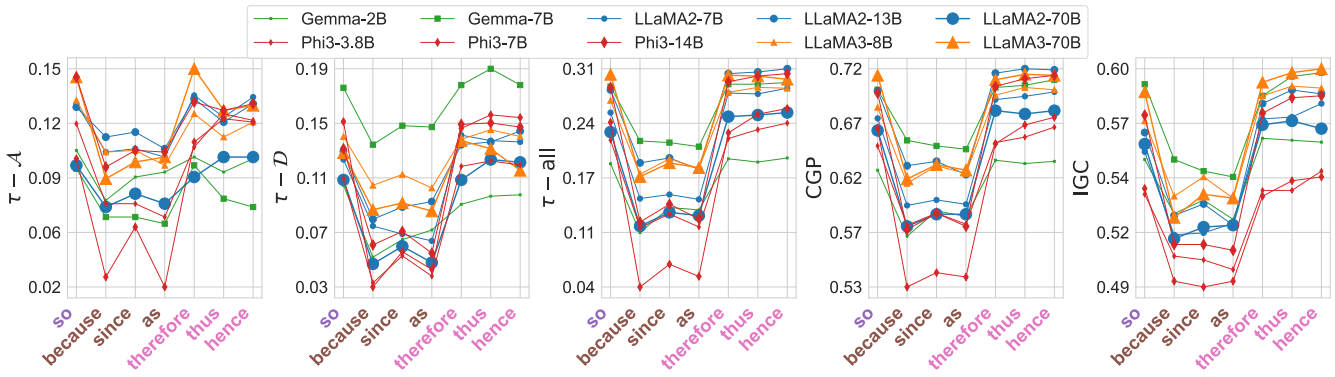

Figure 5: Impact of various conjunction words on the causal epistemic consistency across different LLMs. The x-axes categorize conjunction words into coordinating conjunctions, subordinate conjunctions, and conjunctive adverbs. The y-axes display values for causal epistemic consistency metrics. The analysis encompasses diverse model types (distinguished by marker color and shape) at different scales (represented by line thickness and marker size).

Conjunction Word Choices. We study multiple conjunction words, including (i) coordinating conjunctions (Grammarly 2024): "so"; (ii) subordinate conjunctions (Traffis 2020): "because", "since", and "as"; and (iii) conjunctive adverbs (Ellis 2023): "therefore", "thus", and "hence".

5.3 Results and Discussion

We analyze the results from two aspects: (i) the impact of conjunction words on models' causal epistemic consistency; and (ii) the efficacy of internal token probability against the prompting strategy.

Comparison of Different Conjunction Words. We present the impact of different conjunction words on models' causal epistemic consistency, with distinctions highlighted by varying colors on the x-axis labels in Figure 5. A consistent trend is observed across different models and causal epistemic consistency metrics. Specifically, coordinating conjunctions ("so") and conjunctive adverbs ("therefore", "thus", "hence") yield better results, while subordinate conjunctions ("because", "since", "as") underperform. We posit that placing subordinate conjunctions at the beginning of sentences aligns poorly with the natural language patterns seen by LLMs, potentially degrading performance.

We compare the effectiveness of internal token probability versus prompting methods in maintaining causal epistemic consistency, with a detailed comparison across models and metrics in Appendix D.1.

6 Related Work

LLMs and Causality. The investigation of LLMs in understanding and generating causal relations has garnered increasing attention. Previous studies often criticize LLMs for their propensity to inaccurately identify and comprehend the complex causal patterns among these facts (Jin et al. 2024; Li et al. 2024; Zečević et al. 2023; Cui et al. 2024a). Our study further contributes to this discourse by evaluating LLMs' self-consistency in reasoning about fine-grained intermediates in causality and by providing metrics and empirical evidence for LLMs' causal epistemic consistency.

Defeasibility in Causal Reasoning. Our study of fine-grained intermediates in causality extends the research initiated by δ-CAUSAL (Cui et al. 2024b), which introduced the concepts of defeaters and supporters in causal analysis. While δ-CAUSAL provided a foundational framework for understanding causal defeasibility, it did not delve into the granularity necessary for nuanced causal reasoning. Our research advances this field by moving beyond the binary classification of intermediates as simply supporting or opposing. We refine the categorization of intermediates by considering both their polarity stance (supporting or opposing) and the intensity of their influence.

Hallucination of LLMs. LLMs suffer from generating nonsensical and fallacious content, known as hallucinations (Huang et al. 2023; Mouchel et al. 2024). The most pertinent hallucination to causal epistemic consistency is the self-contradictory hallucination (Mündler et al. 2024), which means that LLMs generate two contradictory sentences given the same context. Specifically, our study on causal epistemic consistency investigates whether the causal intermediates generated by an LLM at various intensities contradict the ones ranked by the same LLM. However, our study is distinctive in that we focus on the discrepancies between the causal intermediate generation and differentiating behaviors of LLMs, rather than the inconsistencies within the generated text.

7 Conclusion

In conclusion, this study introduces causal epistemic consistency as a crucial framework for assessing the self-consistency of LLMs in distinguishing fine-grained causal intermediates. Supported by a novel suite of evaluation metrics, our comprehensive empirical analysis of 21 LLMs reveals significant limitations in their ability to maintain this consistency. This research addresses a critical gap in the understanding of complex causal reasoning and lays the foundation for the development of more self-consistent models capable of handling intricate causal relationships.

Acknowledgements

We are grateful for the IT and financial support from EPFL, Switzerland. Shaobo gratefully acknowledges the support from the LIA lab and the IC department of EPFL.

References

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Bauman, A. E.; Sallis, J. F.; Dzewaltowski, D. A.; and Owen, N. 2002. Toward a better understanding of the influences on physical activity: the role of determinants, correlates, causal variables, mediators, moderators, and confounders. *American journal of preventive medicine*, 23(2): 5–14.

Bird, S.; and Loper, E. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Barcelona, Spain: Association for Computational Linguistics.

Cui, S.; Jin, Z.; Schölkopf, B.; and Faltings, B. 2024a. The Odyssey of Commonsense Causality: From Foundational Benchmarks to Cutting-Edge Reasoning. *CoRR*, abs/2406.19307.

Cui, S.; Milikic, L.; Feng, Y.; Ismayilzada, M.; Paul, D.; Bosselut, A.; and Faltings, B. 2024b. Exploring Defeasibility in Causal Reasoning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 6433–6452. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Ellis, M. 2023. How to Use Conjunctive Adverbs — grammarly.com. https://www.grammarly.com/blog/conjunctive-adverbs/. [Accessed 25-06-2024].

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Fitzgerald, B.; and Howcroft, D. 1998. Towards dissolution of the IS research debate: from polarization to polarity. *Journal of Information technology*, 13(4): 313–326.

Gemini-Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Grammarly. 2024. FANBOYS: Coordinating Conjunctions — grammarly.com. https://www.grammarly.com/blog/coordinating-conjunctions/. [Accessed 25-06-2024].

Gugger, S.; Debut, L.; Wolf, T.; Schmid, P.; Mueller, Z.; Mangrulkar, S.; Sun, M.; and Bossan, B. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; and Oliphant, T. E. 2020. Array programming with NumPy. *Nature*, 585(7825): 357–362.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; Hennigan, T.; Noland, E.; Millican, K.; van den Driessche, G.; Damoc, B.; Guy, A.; Osindero, S.; Simonyan, K.; Elsen, E.; Vinyals, O.; Rae, J. W.; and Sifre, L. 2024. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Holtzman, A.; West, P.; Shwartz, V.; Choi, Y.; and Zettlemoyer, L. 2021. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7038–7051. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *CoRR*, abs/2311.05232.

Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.*, 9(3): 90–95.

Iwasaki, Y.; and Simon, H. A. 1994. Causality and model abstraction. *Artificial intelligence*, 67(1): 143–194.

Jin, Z.; Liu, J.; LYU, Z.; Poff, S.; Sachan, M.; Mihalcea, R.; Diab, M. T.; and Schölkopf, B. 2024. Can Large Language Models Infer Causation from Correlation? In *The Twelfth International Conference on Learning Representations*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *CoRR*, abs/2001.08361.

Kendall, M. G. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1/2): 81–93.

Li, H.; Chi, H.; Liu, M.; and Yang, W. 2024. Look Within, Why LLMs Hallucinate: A Causal Perspective. arXiv:2407.10153.

Luo, Z.; Sha, Y.; Zhu, K. Q.; won Hwang, S.; and Wang, Z. 2016. Commonsense Causal Reasoning between Short Texts. In *International Conference on Principles of Knowledge Representation and Reasoning*.

Malinin, A.; and Gales, M. J. F. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; Chowdhery, A.; Roberts, A.; Barua, A.; Botev, A.; Castro-Ros, A.; Slone, A.; Héliou, A.; Tacchetti, A.; Bulanova, A.; Paterson, A.; Tsai, B.; Shahriari, B.; Lan,

C. L.; Choquette-Choo, C. A.; Crepy, C.; Cer, D.; Ippolito, D.; Reid, D.; Buchatskaya, E.; Ni, E.; Noland, E.; Yan, G.; Tucker, G.; Muraru, G.; Rozhdestvenskiy, G.; Michalewski, H.; Tenney, I.; Grishchenko, I.; Austin, J.; Keeling, J.; Labanowski, J.; Lespiau, J.; Stanway, J.; Brennan, J.; Chen, J.; Ferret, J.; Chiu, J.; and et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *CoRR*, abs/2403.08295.

Meta, A. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Mouchel, L.; Paul, D.; Cui, S.; West, R.; Bosselut, A.; and Faltings, B. 2024. A Logical Fallacy-Informed Framework for Argument Generation. *arXiv preprint arXiv:2408.03618*.

Mündler, N.; He, J.; Jenko, S.; and Vechev, M. 2024. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 8024–8035.

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Shahapure, K. R.; and Nicholas, C. 2020. Cluster Quality Analysis Using Silhouette Score. In Webb, G. I.; Zhang, Z.; Tseng, V. S.; Williams, G.; Vlachos, M.; and Cao, L., eds., *7th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2020, Sydney, Australia, October 6-9, 2020*, 747–748. IEEE.

Shi, X.; Liu, J.; Liu, Y.; Cheng, Q.; and Lu, W. 2023. Know where to go: Make LLM a relevant, responsible, and trustworthy searcher. *arXiv preprint arXiv:2310.12443*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Traffis, C. 2020. What Is a Subordinating Conjunction? — grammarly.com. https://www.grammarly.com/blog/subordinating-conjunctions/. [Accessed 25-06-2024].

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Zečević, M.; Willig, M.; Dhami, D. S.; and Kersting, K. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. *Transactions on Machine Learning Research*.

Zhang, J. J.; Zhang, H.; Su, W. J.; and Roth, D. 2022. ROCK: Causal Inference Principles for Reasoning about Commonsense Causality. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 26750–26771. PMLR.