Do Large Language Models Reason Causally Like Us? Even Better?

Hanna M. Dettki
New York University
hmd8142@nyu.edu

Brenden M. Lake
New York University
brenden@nyu.edu

Charley M. Wu
University of Tübingen
charley.wu@uni-tuebingen.de

Bob Rehder
New York University
bob.rehder@nyu.edu

Abstract

Causal reasoning is a core component of intelligence. Large language models (LLMs) have shown impressive capabilities in generating human-like text, raising questions about whether their responses reflect true understanding or statistical patterns. We compared causal reasoning in humans and four LLMs using tasks based on collider graphs, rating the likelihood of a query variable occurring given evidence from other variables. LLMs' causal inferences ranged from often nonsensical (GPT-3.5) to human-like to often more normatively aligned than those of humans (GPT-4o, Gemini-Pro, and Claude). Computational model fitting showed that one reason for GPT-4o, Gemini-Pro, and Claude's superior performance is they didn't exhibit the "associative bias" that plagues human causal reasoning. Nevertheless, even these LLMs did not fully capture subtler reasoning patterns associated with collider graphs, such as "explaining away". These findings underscore the need to assess AI biases as they increasingly assist human decision-making.

Keywords: Large Language Models; Causal Inference; Human and Machine Reasoning

Introduction

Large Language Models (LLMs) have proven to be highly capable across a range of domains, including natural language understanding, answering questions, and engaging in creative tasks (Bubeck et al., 2023; Abdin et al., 2024; Gunter et al., 2024). In light of these recent advancements in LLMs, many believe that we are now truly entering an era of Artificial Intelligence (AI; Bottou & Schölkopf, 2023). The degree to which machines genuinely comprehend our environment carries significant implications for their reliability in various domains (Mitchell & Krakauer, 2023), including the automatic generation of news content, policy recommendations (Kekić et al., 2023), knowledge discovery, disease diagnosis (Nori, King, McKinney, Carignan, & Horvitz, 2023), and autonomous driving. The impressive capability of LLMs to produce text resembling human language raises the question of whether these models possess some form of world understanding, and if they reason similarly to humans.

Causal reasoning is widely regarded as a core aspect of intelligence (Lake, Ullman, Tenenbaum, & Gershman, 2017). It involves recognizing and inferring the causal relationships between variables, moving beyond mere correlations to uncover underlying mechanisms. Such capabilities are essential in practical applications, including the development of pharmaceutical drugs or the planning of public health strategies. Therefore, causal reasoning is considered an important milestone in the pursuit of Artificial General Intelligence (AGI; Obaid, 2023). Causal reasoning can be formalized using *causal Bayes nets* (CBNs) providing a probabilistic calculus for reasoning about the probability of some variables given others that are causally related (Pearl, 1995). By comparing human reasoners to CBNs, CBNs can serve as a normative benchmark (Glymour, 2003; Waldmann, Hagmayer, & Blaisdell, 2006) and help reveal human biases that deviate from ideal causal reasoning (Rehder & Waldmann, 2017; Bramley, Lagnado, & Speekenbrink, 2015). For instance, when reasoning about a simple collider graph $C_1 \rightarrow E \leftarrow C_2$, people exhibit biases such as *weak explaining away* and *Markov violations* (explained later; Rehder & Waldmann, 2017). These systematic deviations highlight the interplay between normative principles and cognitive heuristics in human causal reasoning.

A plethora of recent studies have assessed the capabilities of LLMs (e.g., Kıcıman, Ness, Sharma, & Tan, 2023), and concerns have been raised regarding their reliance on learned patterns rather than genuine causal relationships (Willig, Zecevic, Dhami, & Kersting, 2023; Jiang et al., 2024). For example, Shi et al. (2023) and Mirzadeh et al. (2024) demonstrated that introducing irrelevant context can drastically alter the outputs of LLMs. That even minor distractions influence their responses raises questions about the robustness of LLMs in high-stakes scenarios.

Indeed, a growing number of researchers have proposed that current LLMs are unable to generalize causal ideas beyond their training distribution and/or without strong user-induced guidance (e.g., chain-of-thought prompting; Jin et al., 2023; Kıcıman et al., 2023). Thus, understanding the extent to which LLMs reason causally, and whether they show similar biases to people when they deviate from normative principles has practical importance in deploying AI systems.

To this end, Jin et al. (2023) introduced the CLADDER dataset, comprising 10,000 causal reasoning questions designed to evaluate the formal causal reasoning abilities of LLMs. While they tested colliders in their dataset, they didn't contrast LLMs with humans. In addition, although the dataset serves as a valuable benchmark for assessing whether LLMs honor probabilistic rules, solving its tasks requires substantial background knowledge (college-level statistics and pen and paper), making it less suitable for direct human compari-

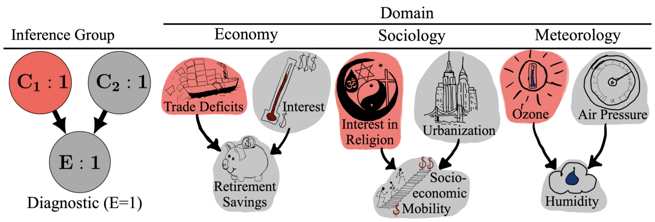


Figure 1: **Visualization of Causal Mechanism per Domain.** The left most graph represents task VI from the diagnostic inference group. The nodes are colored according to: ● $\rightarrow$ latent (query node); ● $\rightarrow$ observed $\in \{0,1\}$.

son. Keshmirian et al. (2024) directly compared humans and LLMs by asking them to judge the strength of a causal relationship $C \rightarrow B$ as a function of context. Human strength judgments were highest when $C \rightarrow B$ appeared in a chain ($A \rightarrow C \rightarrow B$) versus a fork ($A \leftarrow C \rightarrow B$) or in isolation – a pattern LLMs matched with a sufficiently high temperature. In contrast, the present work compares human and LLM causal inferences rather than their strength judgments.

Goals and Scope. As we increasingly rely on AI-supported decision making, our work aims to contribute to the investigation of biases in causal reasoning and compares those between LLMs and humans using human data previously collected in Rehder and Waldmann (2017). We assess a collider graph where two independent causes influence a shared effect ($C_1 \rightarrow E \leftarrow C_2$). A collider gives rise to four inference types: predictive inference (see Figure 2b), unconditional independence (Figure 2c), diagnostic inference with both effect present (Figure 2d) and absent (Figure 2e), from which more specific causal reasoning patterns emerge, such as explaining away. Using behavioral analyses and modeling with CBNs, we ask if LLMs reason like humans, if they reason normatively, and if their inferences reflect the use of domain knowledge that inheres in their training data.

Methods

Participants. We compare the human behavioral data collected in Rehder and Waldmann (2017) (Experiment 1, Model-Only condition, $N = 48$) with judgments gathered from four LLMs — GPT-3.5 (■), GPT-4o (■), Claude-3-Opus (■), and Gemini-Pro-1.5 (■) — which were prompted with the same inference tasks as humans over their respective APIs. We report results for temperature 0.0 as this ensures consistent and reproducible outputs.

Materials. The collider causal structure $C_1 \rightarrow E \leftarrow C_2$ was embedded in one of three cover stories from three different knowledge domains (meteorology, economics, and sociology), allowing for a natural language description of the causal structure. The three domains were chosen because the undergraduate subjects were expected to be relatively unfamiliar, such that their causal inferences would reflect the causal structure given to them and not idiosyncratic prior knowledge. Nevertheless, as an additional safeguard, the adjective describing each variable was counterbalanced (e.g., in the domain of sociology, some subjects were told that *high* urbanization causes *high* socio-economic mobility, others that it causes *low* socio-economic mobility, etc). In fact, Rehder and Waldmann (2017) did not find significant effects of domain or the counterbalancing factor, suggesting that subjects' inferences were not strongly influenced by domain knowledge. An important question we ask here is whether this also holds for the LLMs. Given a set of observations (a subset of the states of C_1, C_2, and E), both humans and LLMs were asked to provide a likelihood judgment on a continuous scale (0-100) for a specific *query variable* ●.

Below is an *example prompt* from the sociology domain, matching the visualization in Figure 1 and diagnostic task X in Figure 2e, where the query node (●) is $C_1 = 1$ and C_2 and the effect E are known to be absent. Note that only the *italicized text* following ":" was presented to LLMs in one piece.

- **Domain introduction:** *Sociologists seek to describe and predict the regular patterns of societal interactions. To do this, they study some important variables or attributes of societies. They also study how these attributes are responsible for producing or causing one another.*
- **Variables:** *Here are some variables: Urbanization is the degree to which the members of a society live in urban environments (i.e., cities) versus rural environments. Some societies have high urbanization. Others have normal urbanization. Interest in religion is the degree to which the members of a society show a curiosity in religion issues or participate in organized religions. Some societies have low interest in religion. Others have normal interest in religion. Socioeconomic mobility is the degree to which the members of a society are able to improve their social and economic status. Some societies have low socio-economic mobility. Others have normal socio-economic mobility.*
- **Causal mechanism:** *Assume you live in a world that works like this:*
  - $C_1 = 1 \rightarrow E = 1$: *High urbanization causes high socio-economic mobility.*
  - $C_2 = 1 \rightarrow E = 1$: *Also, low interest in religion causes high socio-economic mobility.*
- **Observation:** *Suppose that the society you live in currently exhibits the following: normal socio-economic mobility.*
- **Inference task, here** X ($p(C_1 = 1|E = 0)$): *Given the observations and the causal mechanism, how likely on a scale from 0 to 100 is high urbanization? 0 means definitely not likely and 100 means definitely likely. Please provide only a numeric response and no additional information.*

To summarize how humans reason with colliders, the empirical findings reported by Rehder and Waldmann (2017) are presented in Figure 2 (■) alongside the inferences drawn by the LLMs, which are discussed later. The eleven inference tasks (I-XI) are grouped into four types:

Predictive inferences in a collider network involve inferring the state of the effect given information about one or more of the causes. Reasoners should judge, for example, that $p(E = 1 \mid C_1 = 0, C_2 = 0) < p(E = 1 \mid C_1 = 0, C_2 = 1) < p(E = 1 \mid C_1 = 1, C_2 = 1)$. Figure 2b reveals that human reasoners in fact exhibit this pattern, indicated by a monotonically increasing slope, confirming that they made use of the causal knowledge on which they were instructed.

Independence of causes is another property of colliders. Because in CBNs exogenous causes are stipulated to be uncorrelated, reasoners should judge that the presence of one cause should not affect the likelihood of the other: $p(C_1 = 1 \mid$

$C_2 = 1) = p(C_1 = 1 \mid C_2 = 0)$, which would be reflected as a flat line in Figure 2c. Instead, humans judged that $p(C_1 = 1|C_2 = 1) > p(C_1 = 1|C_2 = 0)$. This is an instance of the well-known *Markov violations* that characterize how humans reason with numerous causal network topologies involving generative relations (Davis & Rehder, 2020). Markov violations have been characterized as an *associative bias* (or what Rehder & Waldmann, 2017, referred to as a *rich-get-richer* bias), where the presence of one causal variable makes another supposedly independent variable more likely. Markov violations with collider graphs have been documented in multiple studies (see Davis & Rehder, 2020, for a review).

Diagnostic inferences involve inferring the state of one cause given the effect and possibly the other cause. In collider structures with independent causes and the effect present, this gives rise to *explaining away*, where observing that one cause is present/absent should lower/raise the probability of the other cause. This phenomenon stipulates two conditions. (i) Explaining away proper is when observing one cause reduces the likelihood of the other, e.g., $p(C_1 = 1 \mid E = 1, C_2 = 1) < p(C_1 = 1 \mid E = 1)$. (ii) *Augmentation* arises when observing the absence of a cause increases the likelihood of the other, e.g., $p(C_1 = 1 \mid E = 1, C_2 = 0) > p(C_1 = 1 \mid E = 1)$. Figure 2d demonstrates that humans exhibited the overall explaining away pattern, consistent with the expected monotonically increasing slope under conditions (i) and (ii). However, the effect is weak (i.e., the slope is shallow), aligning with theoretical work showing that explaining away is often attenuated relative to normative expectations (Davis & Rehder, 2020; Rehder, 2024). If the causal relations in the experiment are assumed to be deterministically sufficient and necessary, then the absence of the effect should imply a zero probability for the presence of its causes. Yet Figure 2e revealed human likelihood judgments for $C_1 = 1$ well above zero, suggesting they did not fully endorse this deterministic framing.

Procedure. A key contribution of this work is the creation of a causal inference task dataset enabling direct comparisons between human causal inference judgments collected in Rehder and Waldmann (2017) and LLMs. The dataset closely replicates the experimental conditions of Rehder and Waldmann (2017) (Experiment 1, Model-Only condition) with some notable differences: The human procedure consisted of two phases. In the learning phase, subjects were presented and tested on domain knowledge, including causal mechanisms. In the testing phase, they completed each inference task in random order. A graphical representation of the collider structure remained visible during testing. In contrast, each LLM prompt included all domain knowledge and a single inference task. Whereas humans provided probability judgments using a 0–100 slider (default = 50.0), LLMs were instructed to provide a numerical answer $\in 0.0, 100.0$.

Results

Comparison of LLMs and Humans. As an initial assessment of LLM-human reasoning alignment, we computed the Spearman correlation between their inferences and those of humans in each domain. Table 1 reveals correlations that are positive and substantial in magnitude, indicating the LLMs are exhibiting a degree of human-like performance on the causal reasoning tasks. The highest average correlations were displayed by Gemini ■ ($r_s = .763$), followed by Claude ■ ($r_s = .677$) and GPT-4o ■ (.658). Least aligned was GPT-3.5 ■ ($r_s = .390$). This pattern was observed in all three domains.

Table 1: Spearman correlations r_s between human and LLM inferences in each domain / across domains (pooled).

	Domain			
Model	Economy (r_s)	Sociology (r_s)	Weather (r_s)	Pooled
Claude ■	.641	.739	.755	.677
GPT-4o ■	.618	.506	.767	.658
GPT-3.5 ■	.390	.473	.313	.390
Gemini ■	**.713**	**.743**	**.855**	**.763**

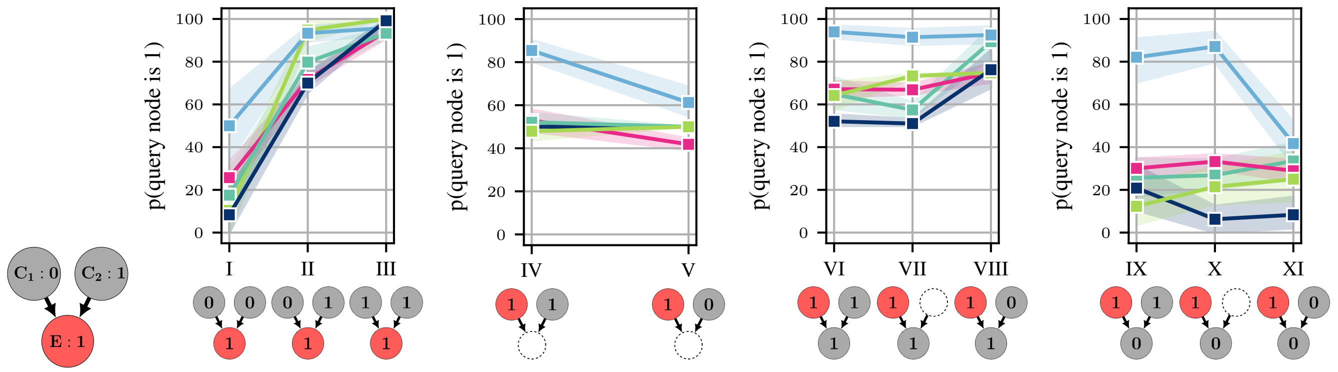
Figure 2 presents the LLMs' responses to the four inference tasks averaged over conditions. The main finding is that all LLMs except GPT-3.5 ■ provided sensible judgments for all inference tasks. Each task reveals distinct reasoning patterns across agents.

Predictive inferences (Fig. 2b, I-III) for the LLMs were a monotonic increasing function of the number of causes present, similar to the human judgments. This indicates that the LLMs were sensitive to the most rudimentary aspect of the task, namely, that causes make their effects more likely. Predictive inference is the only inference type where GPT-3.5 ■ provided sensible judgments.

Independence of causes (Fig. 2c), IV-V) means that the state of one cause should not affect the likelihood of the other (i.e., a flat line). GPT-3.5 ■ violated this principle by judging $p(C_1 = 1|C_2 = 1) > p(C_1 = 1|C_2 = 0)$ even more egregiously than humans. Conversely, Claude ■, GPT-4o ■, and Gemini ■ reasoned normatively, by respecting the independence of causes, indicated by a flat line.

Effect-Present Diagnostic Inference (Fig. 2d, VI–VIII) assessed explaining away via the slope of inferred probabilities reflected by a positive slope between tasks VI and VII if (i) holds. Gemini-Pro ■ showed the strongest effect, followed by both humans ■ and GPT-4o ■ with weak (i). Claude ■ and GPT-3.5 ■ violated explaining away, indicated by negative slope for (i). Conversely, GPT-4o and Claude showed strong augmentation (ii), assigning higher likelihood to $C_1 = 1$ when the alternative cause was absent, indicated by a positive slope between tasks VII and VIII. Gemini and GPT-3.5 exhibited numerically weak augmentation (< 2 points), and no model fully satisfied both conditions.

Effect-Absent Diagnostic Inference (Fig. 2e, IX-XI) has all agents produce lower ratings for the cause, with GPT-4o ■ and Claude ■ producing the lowest ratings across all conditions and Gemini ■ seeming to be closest aligned with humans ■. While humans and Gemini ■ are more likely to assign ratings in the middle of the scale, GPT-4o ■ is most inclined to assign



(a) Reference Graph: task II (b) Predictive Inference (c) Independence of C_1, C_2 (d) Diagnostic Inference ($E=1$) (e) Diagnostic Inference ($E=0$)

Figure 2: **Aggregated across all domains:** Likelihood judgments that query node ● has value $1 \in \{0, 100\}$ with bootstrapped 95% confidence intervals of humans ■ and LLMs (GPT-3.5 ■, GPT-4o ■, Claude ■, and Gemini ■) for each inference task (I-XI), aggregated across counterbalancing conditions and domains for temperature value 0.0 (most deterministic). Graphs on the x-axis visualize the conditional probability of the inference tasks (I-XI) where the nodes are colored according to: ● $\rightarrow$ query node that the question is asked about; ● $\rightarrow$ observed $\in \{0, 1\}$; and ◌ $\rightarrow$ no information.

a rating of 0 and treated the causal relations as closer to necessary and sufficient than any other agent. This interpretation is supported by the model fitting that follows, which yielded especially large estimates of the strengths of the causal relations for GPT-4o ■ (see Figure 3).

Note that the responses of three of the four LLMs in Figure 2 exhibited a greater range than the humans. The difference between the highest and lowest judgment was 95.0, 91.7, and 75.8 for GPT-4o, Gemini, and Claude, respectively, as compared to 66.0 for the humans. This tendency might stem from the experimental setup. Whereas LLMs were prompted to generate a single numeric value, humans responded using an interactive slider that defaulted to 50. This default could have introduced a motor bias that encouraged responses near the middle of the scale. The responses of GPT-3.5 exhibited the narrowest range (54.2). These inference patterns suggest LLMs capture core causal reasoning principles and are aligned with human responses to a considerable degree. Some LLMs' reasoning patterns in Figure 2 reveal that causal relations were treated as close to necessary and sufficient (e.g., GPT-4o ■), which is also supported later when we fit CBNs (see Figure 3).

CBN Model Fitting. Next, we evaluate LLMs and humans against normative inferences from a causal Bayes net (CBN). Since agents received only verbal descriptions, the CBN's parameters θ_M were treated as free parameters and fit to the data. These parameters were the causes' prior probabilities w_C, representing $p(C_1)$ and $p(C_2)$, the causal strength parameters $w_{C_1,E}$ and $w_{C_2,E}$, representing the strength of $C_1 \rightarrow E$ and $C_2 \rightarrow E$, and w_E, representing the influence of any exogenous causal influence on E.

The CBN is used to derive a joint probability distribution which was then used to derive the conditional probability appropriate for that task. For a collider causal graph $C_1 \rightarrow E \leftarrow C_2$, the joint distribution was derived assuming that $p(C_1, C_2, E) = p(E|C_1, C_2)p(C_1)p(C_2)$ and that $p(E = 1|C_1, C_2) = 1/(1+\exp(-(C_1 w_{C_1,E} + C_2 w_{C_2,E} + w_E)))$, where C_1 and C_2 are each coded as 1 when present and -1 when absent.[1] The CBNs were fit to each agent's set of causal judgments by identifying parameters that minimized squared error. Fits were carried out via an initial grid search followed by optimization.

We fit two variants of the basic collider CBN. The first assumed that the two causal strengths were equal, that is, $w_{C_1,E} = w_{C_2,E} = w_{C,E}$. Thus, the parameters of this model were w_C, $w_{C,E}$, and w_E. A 4-parameter variant allowed the strength of the causal relations to differ by fitting $w_{C_1,E}$ and $w_{C_2,E}$ separately. w_C was constrained to the range [0, 1] and the causal strength parameters were constrained to $[-3, 3]$. For the human data, these CBNs were fit to each subject. For the LLMs, they were separately fit to the judgments in each of the 3 domains $\times$ 4 counterbalancing $= 12$ conditions.

Table 2 presents the CBNs' best fitting parameters averaged over conditions for each agent. Several trends emerge. The correlations between the observed judgments and those predicted by the fitted CBNs were substantial for all the LLMs, ranging from 0.503 to 0.879. Notably, for Gemini-Pro, GPT-4o, and Claude-3 those correlations were all greater than 0.82 and so greater than those observed for the humans (0.77). They also exhibited more favorable model losses, defined as the average absolute prediction error on the 0–100 scale, than the humans. That is, if CBNs are accepted as the normative standard, these LLMs exhibited more accurate causal reasoning than the humans. In contrast, GPT-3.5 per-

[1] Note in this literature it is common to assume "noisy logical" generating functions, such as the noisy-OR function introduced in the PowerPC theory of causal learning by Cheng (1997). We report fits using the logistic generating function as it consistently yielded better fits to these data sets.

Table 2: Fits of causal Bayes nets (CBN) by agent.

Agent	NP	Average Parameter Estimates w_C	$w_{C,E}$	$w_{C_1,E}$	$w_{C_2,E}$	w_E	Measures of Fit R	AIC	Loss
Humans	**3**	**.528**	**1.06**			**0.91**	**.770**	**114.4**	**11.8**
	4	.529		1.09	1.04	0.92	.783	115.4	11.5
Gemini-Pro-1.5	**3**	**.553**	**1.55**			**1.87**	**.877**	**109.6**	**10.2**
	4	.556		1.57	1.59	1.96	.877	110.5	10.2
GPT-3.5	3	.843	0.60			1.76	.503	123.4	15.4
	4	**.845**		**0.61**	**0.59**	**1.83**	**.558**	**123.3**	**14.6**
GPT-4o	**3**	**.438**	**1.66**			**1.31**	**.879**	**107.1**	**9.88**
	4	.436		1.74	1.53	1.32	.881	107.6	9.88
Claude-3-Opus	**3**	**.555**	**1.26**			**1.16**	**.829**	**110.6**	**11.2**
	4	.554		1.32	1.21	1.17	.839	111.7	10.8

Note: NP = Number of model parameters. AIC = Akaike Information Criterion, used to choose winning CBN in bold.

formed worse than both humans and other LLMs, with correlations below 0.560 and model losses above 14.

Regarding the contrast between the 3- and 4-parameter CBNs, the human data did not benefit from the extra causal strength parameter. This result is consistent with past analyses of these data showing that neither domain nor the counterbalancing factor had a significant effect on subjects' judgments (Rehder & Waldmann, 2017). Turning to the LLMs, only GPT-3.5 yielded a better fit with two causal strength parameters. Although we expected that the LLMs might be more likely to assume causal relations of different strength by using the knowledge they have about economics, meteorology, and sociology, the fitted parameter values in Table 2 indicate that they were no more likely to do so than the humans.[2] A detailed investigation of the effect of domain on the parameter estimates would offer further insight into agents' sensitivity to contextual and linguistic variation in the causal cover stories but is beyond the scope of the current work.

To provide a more granular view of agent behavior than the averaged results in Table 2, Figure 3 shows fitted parameter distributions from the 4-parameter CBN. Each violin plot summarizes agent-specific parameter values across domains and counterbalancing. The figure also includes the absolute difference between the two causal strength parameters, $|w_{C_1,E} - w_{C_2,E}|$. Parameter w_C, representing the prior over causes $p(C_1)$ and $p(C_2)$, clustered around 0.5, with agent means ranging from 0.43 (GPT-4o) to 0.85 (GPT-3.5), and showed the least variation across agents. The causal strength parameters $w_{C_1,E}$ and $w_{C_2,E}$ showed greater variability. GPT-4o was most often best fit by high values, with median estimates of 1.59 and 1.30, suggesting strong deterministic assumptions. Although Claude and GPT-4o had a broader $w_{C_1,E}$, $w_{C_2,E}$ range than Gemini (0.206–2.70), Gemini had the highest median (1.65). Humans were the only group occasionally best fit by negative values for $w_{C_2,E}$ (range: -0.57 to 3.00), suggesting possible inhibitory interpretations. GPT-3.5 showed the least variability of any agent and favored smaller values (range: .05 to 1.73, median: .62 and .43). The absolute difference between causal strengths $|w_{C_1,E} - w_{C_2,E}|$ was generally small, ranging from 0.0 to 1.03, suggesting that, within a domain, $C_1 \to E$ and $C_2 \to E$ were treated as about equally strong. Like the causal strengths, parameter w_E, reflecting sensitivity to exogenous causes, also exhibited substantial variability over domains (range: $-.39$ to 3.0).

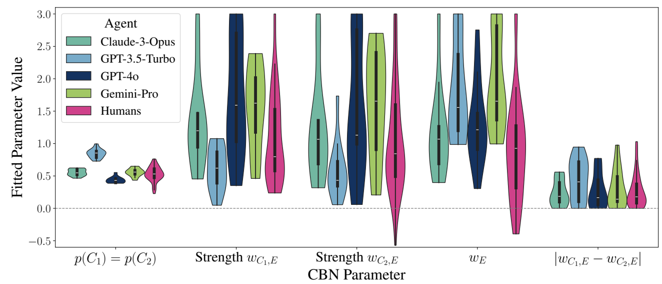


Figure 3: Fitted parameter distributions for each agent under the 4-parameter CBN model. Violin plots reflect aggregated fits across domains and counterbalancing. The white bar denotes the median; the box spans the 25th to 75th percentiles. The rightmost plot quantifies asymmetry in inferred causal strengths via the absolute difference $|w_{C_1,E} - w_{C_2,E}|$.

Fitting a Psychological Model. We also fit the LLM inferences with a model proposed as an account human causal reasoning, the *mutation sampler* (Davis & Rehder, 2020). The mutation sampler is an example of a *rational process model*, an algorithm that yields normative responses when cognitive resources are unlimited but that produces errors when they are not (Johnson & Busemeyer, 2016; Lieder, Griffiths, & Goodman, 2012; Vul, Goodman, Griffiths, & Tenenbaum, 2014). The mutation sampler carries out MCMC sampling over a causal graph's *state space* and draws inferences on the basis of samples. But because sampling begins at one of the graph's *prototypes states* (when causal relations are all generative, the states where variables are all present or all absent), errors are introduced when the number of samples drawn is limited. Davis and Rehder (2020) showed that the associative bias induced by the prototypes allowed the mutation sampler to account for the independence violations that arise in a wide variety of network topologies and the weak explaining away that arises when reasoning about collider graphs.

The mutation sampler was fit to the human data and the four LLMs. Table 3 presents the improvement for each data set relative to the 3-parameter CBN in Table 2 realized by adding the mutation sampler's *chain length* free parameter λ representing the number of MCMC samples. Replicating past findings, the mutation sampler yielded a better fit to the human data (according to AIC) compared to the 3-parameter CBN (Davis & Rehder, 2020). In contrast, Table 2 shows that it generally did *not* yield a better fit for the LLMs (GPT-3.5 was the only exception). Apparently, the LLMs were less susceptible to the associative reasoning processes that influence people's causal inferences, a conclusion supported by

[2] We also fit CBNs in which the two causes C_1 and C_2 each had their own parameter representing their prior probability. Generally, these models did not yield a better fit than the models with a single w_C parameter. The one exception was GPT-3.5, but as this model yielded relatively poor fits, we do not discuss this result further.

Table 3: Fits of the 4-parameter mutation sampler.

		Average Parameter Estimates				Measures of Fit		
Agent	*NP*	w_C	$w_{C,E}$	w_E	λ	*R*	*AIC*	Loss
Humans ■	3	.528	1.06	0.91		.770	114.4	11.8
	4	**.523**	**0.94**	**0.83**	**3.7**	**.810**	**113.0**	**10.8**
Gemini-Pro-1.5 ■	**3**	**.553**	**1.55**	**1.87**		**.877**	**109.6**	**10.2**
	4	.576	1.43	1.97	38.6	.881	109.7	10.0
GPT-3.5 ■	3	.843	0.60	1.76		.503	123.4	15.4
	4	.897	**-0.24**	**2.46**	**37.9**	**.776**	**113.1**	**16.4**
GPT-4o ■	**3**	**.438**	**1.66**	**1.31**		**.879**	**107.1**	**9.88**
	4	.398	1.50	1.09	26.0	.881	108.4	9.81
Claude-3-Opus ■	**3**	**.555**	**1.26**	**1.16**		**.829**	**110.6**	**11.2**
	4	.569	1.21	1.20	43.9	.826	114.1	11.5

Note: The 3-parameter CBN is included for comparison.

the fitted chain length parameters λ shown in Table 3. Because the impact of the starting point diminishes as the chain length grows, that the LLM fits exhibited relatively large chain lengths indicates that the associative influence induced by the prototypes had little impact on the LLMs' judgments.

Discussion

We compared the causal reasoning abilities of large language models to those of people. In Rehder and Waldmann (2017) undergraduates were taught hypothetical causal knowledge consisting of three variables that formed a collider causal graph and then were asked to draw simple causal inferences. Our first main finding is that given the same information[3], most LLMs tested can do the task. That is, after being told that the presence of one variable C causes the presence of another E, LLMs will judge the effect E is more likely when a cause C is present versus absent (and vice versa). Indeed, across all domains and tasks, the Spearman correlation r_s between LLM and human inferences ranged from .313 to .855.

Collider structures imply that causes are independent, yet human judgments often violate independence, reflected as a perceived positive correlation, consistent with associative reasoning (Davis & Rehder, 2020). Figure 2c shows that GPT-3.5 exhibited a similar but stronger violation. By contrast, Claude, Gemini, and GPT-4o adhered closely to the independence assumption, assigning uniform likelihoods (≈ 50) regardless of the status of the alternative cause.

The LLMs varied in how they exhibited explaining away, with no model fully capturing both defining conditions (i) and (ii) (Fig. 2d). Gemini-Pro was the only model to show a clear instance of condition (i), reducing the likelihood of one cause when the alternative was present compared to when there was no information ◌ about the alternative cause (Fig. 2d, VI, VII). Conversely, GPT-4o and Claude demonstrated strong augmentation (condition (ii)), increasing the likelihood of one cause when the other was absent (Fig. 2d, VII, VIII). Thus, whereas the LLMs generally exhibited impressive correlations with humans and their fitted CBNs, they did not capture some of the more subtle reasoning patterns implied by a collider graph.

[3]Humans underwent a learning phase with extensive exposure to background information, whereas the inference task was limited to a single screen displaying only essential details. In contrast, LLMs were presented with both the learning and testing phases simultaneously in one long prompt. What constitutes an equivalent input for LLMs remains an open question.

A collider structure also supports diagnostic reasoning in the absence of the effect (see Figure 2e). Note that, if the causal relations are interpreted as deterministically sufficient (the cause always produces the effect), then the likelihood of the causes should be zero when the effect is absent. GPT-3.5 deviated sharply from this prediction, providing judgments of between 40 and 90 (and consistent with its relatively weak fitted causal strength parameters in Table 2). In contrast, the other LLMs provided lower judgments for these inferences (consistent with their larger causal strength parameters). Of all the LLMs, the responses of GPT-4o were most consistent with deterministically sufficient causal relations.

In addition to humans, we compared LLMs to the normative inferences of fitted CBNs. Correlations between LLM and normative inferences ranged from .503 to .881, versus ≈ .77 for humans. GPT-3.5 showed the weakest correlations – lower than those of humans – whereas GPT-4o, Gemini, and Claude showed the highest, *exceeding* human correlations. Computational model fitting revealed that one reason for the better performance of the latter models is that they didn't exhibit the associative bias that plagues human casual reasoning.

Future Work. There are numerous avenues for future research. Here we compared human and LLM inferences on only one simple causal structure, whereas humans have been tested on causal networks with different topologies (e.g., forks, chain, etc.), causal relations (inhibitory vs. generative), integration functions (e.g., causes that combine conjunctively rather than independently), with more than three variables, and with continuous variables rather than binary ones. Besides the simple causal inferences examined here, there is a wealth of data on how humans intervene on causal systems, make causal attributions in cases of actual causation, and learn causal systems from observed data. Regarding LLMs, a deeper analysis of the effects of domain knowledge on their inferences is warranted as such knowledge can affect both independence (via inferred causal connections between the collider's causes) and explaining away (via treating the two causal relations as interactive rather than independent; Cruz, Hahn, Fenton, & Lagnado, 2020; Morris & Larrick, 1995). It is also important to better understand how their inferences are affected by factors such as the temperature parameter.

Overall, the tested LLMs largely engaged appropriately with the same complex prompts used in research on human causal reasoning. GPT-4o's responses aligned most closely with normative inferences, with Gemini exhibiting similar performance. Claude, while slightly less normatively aligned than the former two, more closely mirrored human reasoning patterns than GPT-4o. Notably, Gemini achieved both high normative consistency and the highest correlation with humans ($r_s = .763$). GPT-3.5 deviated markedly from both with the exception of the predictive inference tasks.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., ... others (2024). Phi-3 technical report: A highly capable language model locally on your phone, 2024. *URL https://arxiv. org/abs/2404.14219*.

Bottou, L., & Schölkopf, B. (2023). Borges and AI. *arXiv preprint arXiv:2310.01425*.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review*, *104*(2), 367.

Cruz, N., Hahn, U., Fenton, N., & Lagnado, D. (2020). Explaining away, augmentation, and the assumption of independence. *Frontiers in Psychology*, *11*(502751).

Davis, Z. J., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*, e12839.

Glymour, C. (2003). Learning, prediction and causal bayes nets. *Trends in cognitive sciences*, *7*(1), 43–48.

Gunter, T., Wang, Z., Wang, C., Pang, R., Narayanan, A., Zhang, A., ... others (2024). Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*.

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., ... Roth, D. (2024). A peek into token bias: Large language models are not yet genuine reasoners. *arXiv preprint arXiv:2406.11050*.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., ... others (2023). Cladder: A benchmark to assess causal reasoning capabilities of language models. *arXiv preprint arXiv:2312.04350*.

Johnson, J. G., & Busemeyer, J. R. (2016). A computational model of the attention process in risky choice. *Decision*, *3*(4), 254–280.

Kekić, A., Dehning, J., Gresele, L., von Kügelgen, J., Priesemann, V., & Schölkopf, B. (2023). Evaluating vaccine allocation strategies using simulation-assisted causal modeling. *Patterns*.

Keshmirian, A., Willig, M., Hemmatian, B., Hahn, U., Kersting, K., & Gerstenberg, T. (2024). Biased causal strength judgments in humans and large language models. In *ICLR 2024 Workshop on Representational Alignment.* Retrieved from `https://openreview.net/forum?id=544P6YidFk`

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems* (Vol. 25, pp. 2690–2798).

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331–355.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Obaid, O. I. (2023). From machine learning to artificial general intelligence: A roadmap and implications. *Mesopotamian Journal of Big Data*, *2023*, 81–91.

Pearl, J. (1995). From bayesian networks to causal networks. In *Mathematical models for handling partial knowledge in artificial intelligence* (pp. 157–182). Springer.

Rehder, B. (2024). Extending a rational process model of causal reasoning: Assessing markov violations and explaining away with inhibitory causal relations. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *50*, 1463–1488.

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition*, *45*, 245–260.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., ... Zhou, D. (2023, 23–29 Jul). Large language models can be easily distracted by irrelevant context. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 31210–31227). PMLR. Retrieved from `https://proceedings.mlr.press/v202/shi23a.html`

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*(6), 307–311.

Willig, M., Zecevic, M., Dhami, D. S., & Kersting, K. (2023). Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*, *8*.