DO LLMS SHARE HUMAN-LIKE BIASES? CAUSAL REASONING UNDER PRIOR KNOWLEDGE, IRRELEVANT CONTEXT, AND VARYING COMPUTE BUDGETS

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

Charley M. Wu
Technical University Darmstadt

Bob Rehder
New York University

ABSTRACT

Large language models (LLMs) are increasingly used in domains where causal reasoning matters, yet it remains unclear whether their judgments reflect normative causal computation, human-like shortcuts, or brittle pattern matching. We benchmark 20+ LLMs against a matched human baseline on 11 causal judgment tasks formalized by a collider structure ($C_1 \rightarrow E \leftarrow C_2$). We find that a small interpretable model compresses LLMs' causal judgments well and that most LLMs exhibit more rule-like reasoning strategies than humans who seem to account for unmentioned latent factors in their probability judgments. Furthermore, most LLMs do not mirror the characteristic human collider biases of weak explaining away and Markov violations. We probe LLMs' causal judgment robustness under (i) semantic abstraction and (ii) prompt overloading (injecting irrelevant text), and find that chain-of-thought (CoT) increases robustness for many LLMs. Together, this divergence suggests LLMs can complement humans when known biases are undesirable, but their rule-like reasoning may break down when uncertainty is intrinsic, as is in many real-world scenarios. This highlights the need to characterize LLM reasoning strategies for safe, effective deployment.

1 INTRODUCTION

Human causal judgment is powerful but brittle: people routinely deviate from normative inference under limited attention, fatigue, and contextual framing (Danziger et al., 2011; Venkatraman et al., 2007; Barnes et al., 2015; Wittbrodt & Millard-Stafford, 2018). This motivates AI-assisted decision-making in high-stakes settings where consistent causal analysis could complement human limitations. This is especially important since large language models (LLMs) are increasingly deployed in domains where causal understanding matters (e.g., legal and medical decision support) (Dahl et al., 2024; Riedemann et al., 2024). Successful deployment therefore depends not only upon whether LLMs can generate correct output in known contexts via *pattern matching / associative reasoning*, but crucially on whether they can apply *causal reasoning strategies* allowing them to generalize to new contexts.

A core challenge is that in the real world, many causal judgment settings are ill-defined (e.g., no exact base rates of events are known) and hence humans rely on heuristics and prior assumptions, often introducing biases. Before deploying LLMs in high-stakes settings, it is hence paramount to not only compare LLM judgments to humans and investigate whether they reproduce certain human biases, provided that LLMs have been trained on largely human generated data, but also whether LLMs provide sensible causal judgments (e.g., that the probability of a variable increases as more evidence arises).

Related Work. Recent work evaluates LLMs on causal benchmarks with fully specified generative models, such as CLADDER (Jin et al., 2023), where queries have a single numeric ground truth given explicit base rates and link strengths. Such settings primarily test whether models can *apply* statistical rules, but they are less informative about an agent's implicit world model (e.g., assumed priors or causal strengths) when these quantities are underspecified, as is often the case in real-world

settings. We therefore adopt a classic paradigm from human causal cognition (Rehder, 2003; 2014; Bender, 2020; Waldmann et al., 2006) that deliberately leaves causal tasks underspecified, enabling a direct probe of the assumptions agents bring to the task; it also avoids requiring formal training in causal statistics, making human baselines more straightforward to collect.

Our study connects to the small body of work directly comparing humans and LLMs on the *same* reasoning tasks (Gandhi et al., 2024; Lampinen et al., 2024; Keshmirian et al., 2024). It also relates to evidence that LLM judgments can shift under irrelevant context or superficial content changes (Shi et al., 2023; Mirzadeh et al., 2024), and that Chain-of-Thought (CoT) can sometimes mitigate such fragility—addressing concerns that apparent causal reasoning may reflect learned correlations rather than causal understanding (Willig et al., 2023). Similarly, Zhu & Griffiths (2024) study probabilistic *coherence* in LLM judgments via repeated elicitation and violations of probability identities, using Bayesian cognitive models to interpret systematic deviations; our work instead examines the coherence and robustness of qualitative reasoning signatures across different causal queries and content manipulations.

Within collider structures, humans reliably exhibit systematic departures from normative qualitative patterns—including *too little* explaining away (Fernbach & Rehder, 2013; Rottman & Hastie, 2014) and Markov violations in which beliefs about one cause depend on information about an alternative cause (Ali et al., 2011; Mayrhofer & Waldmann, 2016; Park & Sloman, 2013; Rehder & Waldmann, 2017). Whether and when LLMs reproduce these human-like biases, particularly given training on human-generated text, remains unclear. Finally, while Dettki et al. (2025) tested against the same human baseline but evaluated only four LLMs, substantially broader evaluations and systematic robustness analyses remain limited, motivating the present work.

Contributions. We pursue the question of how LLMs' causal reasoning compares to that of humans by evaluating 20+ LLMs on a causal inference benchmark derived from classic collider-graph judgments with a human baseline (Rehder & Waldmann, 2017). We are interested in three core questions: *(Human-comparison)* do models reproduce human-like response patterns and biases? *(Mechanism)* can their responses be compactly explained by a normative causal Bayesian network (CBN) model? *(Robustness)* do these behaviors persist under content manipulations that reduce reliance on memorized world knowledge as well as in the face of inserted irrelevant information? We test the above both with direct prompting ● where we ask for a single probability judgment as a response and with chain-of-thought (CoT) ● (Wei et al., 2022) where we ask LLMs to first reason step-by-step. Together, this yields a cognitively-grounded account of *how* different LLMs reason causally—often more rule-based than humans and mostly not replicating human biases, with systematic shifts under CoT prompting. We additionally release an LLM-friendly version of the causal inference benchmark (including the human baseline) and a software package CAUSAIIGN that supports structure-matched custom prompts and content manipulations.

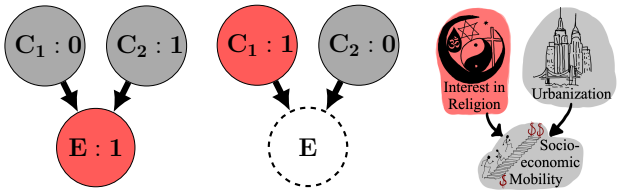
2 METHODS

Benchmark tasks and human baseline. We build on causal inference tasks from Rehder & Waldmann (2017) (RW17; Experiment 1, Model-Only condition; N=48 NYU undergraduates), which elicit probability judgments in a collider graph $C_1 \rightarrow E \leftarrow C_2$. Variables are binary, causes are asserted as independent, and participants report probabilities (scale 0–100) for a target event ("off"/0 or "on"/1) given observed values of the other variables. Rehder & Waldmann (2017) test 11 inference tasks (I–XI) corresponding to distinct conditional-probability queries on this collider—e.g., predictive queries such as $p(E{=}1 \mid C_1{=}0, C_2{=}1)$ in Figure 1a and Figure 2a—embedded in three different cover stories (sociology, weather, economy). Critically, because priors over variables and probabilities such as $p(E \mid C_1, C_2)$ are not specified, there is no ground-truth "correct" probability; the benchmark is therefore designed to assess *reasoning strategies* of causal judgment (e.g., how consistent agents reason across tasks, or how rule-following they are) and *qualitative patterns*, e.g., assessing the level of explaining away and Markov-(non) compliance and compare them across agents and to the human baseline.

LLM prompting and experimental conditions. For LLMs, we convert the RW17 training and test-phase information given to humans into a single textual prompt that includes (i) a domain introduction, (ii) causal mechanism statements for both causes, (iii) the observed evidence, and (iv) an inference question requesting a single numeric probability judgment $\in [0, 100]$. We evaluate two prompting

strategies: *direct* prompting (single numeric answer) and *chain-of-thought* (CoT) prompting ("think step by step" before the numeric answer). We set temperature to 0.0 to minimize sampling variance; for models that expose an explicit reasoning-budget control, we use it as provided by the API.

We test robustness via two content manipulations crossed with prompting, yielding $2 \times 2 \times 2 = 8$ experimental conditions: (i) *Prior knowledge*: original RW17 cover stories vs. *abstract* variants in which RW17 semantically meaningful variable names are replaced by randomly generated strings, reducing reliance on real-world knowledge; (ii) *Information load*: original prompts vs. *overloaded* prompts that append irrelevant text to reduce the signal-to-noise ratio (distraction); (iii) *Prompting*: direct vs. chain-of-thought (CoT). Across all conditions, the dependent variable is the agent's probability judgment (0–100) for each of the 11 conditional probability queries that we refer to as inference tasks (cf. Figure 1).



(a) Task 2 (b) Task 5 (c) Domain Sociology

Figure 1: Figure 1a and Figure 1b represent two out of 11 conditional probability (inference) queries. The colors of the nodes represent the three possible states the variables that the nodes of the graph can be in: ●→observed node∈{0, 1}, ●→latent (query node), ◌→no information. Figure 1c shows an example of one of the three domains (Sociology) used in the RW17 conditions that "verbalize" the graph, where each node is represented by a variable in the domain that takes on binary states, e.g., C_2 being on (1) / off (0) translating to "high /low urbanization" in the sociology domain.

Causal stories and prompt structure Below is an example prompt including its prompt scaffolding that conveys the causal story and prompt structure in a condensed form which all tasks regardless of their experimental condition shared. Here we present the ABSTRACT version where initially meaningful variables e.g., from the economy domain in Rehder & Waldmann (2017) are replaced with meaningless placeholders consisting of a random sequence of ten characters. The *blue text* illustrates injected noise that was part of the *overloaded* experimental conditions abbreviated with the suffix "-OVER". The injected noise was either neutral coming from the lorem ipsum vocabulary or meaningful but meaningless to the task coming from the original prompts in Rehder & Waldmann (2017) such that for example in the RW17-OVER condition a prompt from the weather domain would contain a distracting sentence from the sociology domain.[1]

- **Domain introduction:** *In abstract reasoning studies, researchers examine relationships between symbolic variables* `u8jzPde0Ig`*,* `xLd6GncfBA`*, and* `epfJBd0Kh8`*.*
- **Variable descriptions (*overloaded*):**
- C_1 (**`u8jzPde0Ig`**)**.** *Some systems have high* `u8jzPde0Ig`*. Others have low* `u8jzPde0Ig`*. Interest rates are the rates banks charge to loan money.*
- C_2 (**`xLd6GncfBA`**)**.** *Some systems have weak* `xLd6GncfBA`*. Others have strong* `xLd6GncfBA`*.*
- E (**`epfJBd0Kh8`**)**.** *Some systems have weak* `epfJBd0Kh8`*. Others have powerful* `epfJBd0Kh8`*.*
- **Causal relationships** *Here are the causal relationships:*

  - *High* `u8jzPde0Ig` *causes weak* `epfJBd0Kh8`*. The good economic times produced by . . . .*
  - *Weak* `xLd6GncfBA` *causes weak* `epfJBd0Kh8`*. Lorem ipsum dolor sit amet . . . .*

Both high `u8jzPde0Ig` *and weak* `xLd6GncfBA` *can independently bring about weak* `epfJBd0Kh8`*.*
- **Observations:** *You are currently observing: weak* `epfJBd0Kh8` *and high* `u8jzPde0Ig`*.*
- **Task instruction** *Your task is to estimate how likely it is that weak* `xLd6GncfBA` *is present on a scale from 0 to 100, given the observations and causal relationships described. . . . .*

3 RESULTS

We organize results by the following questions: (Q1) *Are Humans and LLMs aligned* in their responses and do agents provide sensible judgments; (Q2) *Do agents reason normatively* as measured via small,

[1]The **bold** text is there to help the reader understand the structure of the prompt. Only the italicized text is part of the actual prompt given to LLMs.

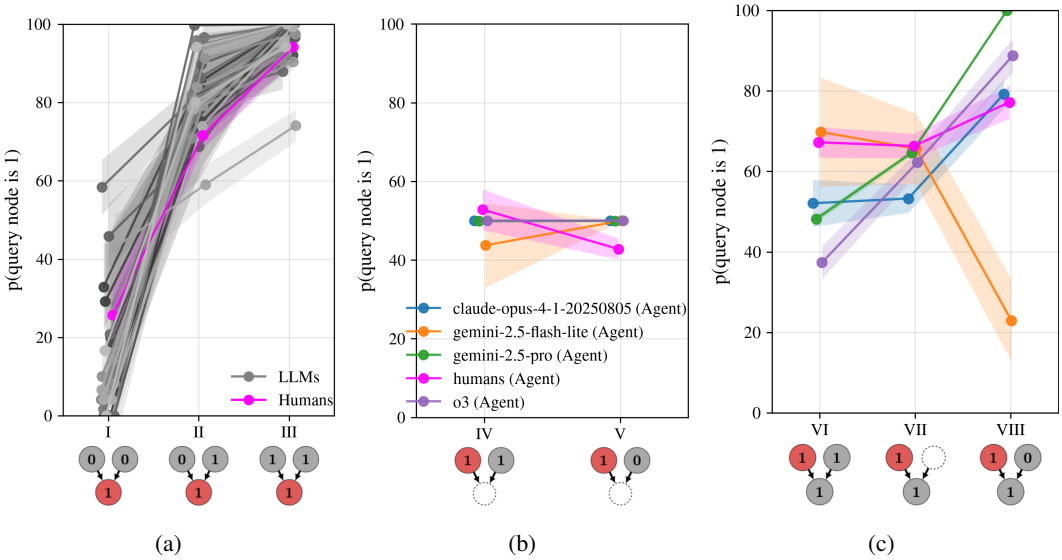


Figure 2: Like humans, LLMs judge the effect as more likely in the presence of more causes (Figure 2a). Unlike humans, most LLMs (a) do not reproduce the Markov violation bias and instead respect the independence of causes (Figure 2b) and (b) show strong explaining away where humans show only a weak effect (Figure 2c).

interpretable causal models. (Q3) *how rule-following are agents*, (Q4) *are LLMs biased like humans*; and (Q5) *whether LLMs reason robustly* across content manipulations.

No significant domain differences. Within each agent, probability-judgment distributions did not differ across domains (Kruskal–Wallis, k=3, df=2; Benjamini–Hochberg across agents: GPT-3.5-turbo $p_{\text{FDR-BH}}$=.146; all others $p_{\text{FDR-BH}} \geq .742$), so we pool the domains in all remaining analyses.

3.1 Q1: Are Humans and LLMs aligned and do agents provide sensible judgements on our tasks?

> Both LLMs and humans provide sensible judgments on our tasks and judge the effect as more probable the more causes are present. For less aligned models in the direct prompting condition, chain-of-thought prompting improves alignment with human judgments up to a ceiling effects at 0.85 (Spearman ρ).

Both LLMs and humans judge the effect as more probable the more causes are present. We first look at the three predictive inference tasks in our dataset where given one of the causes (known), the probability of the effect (to be predicted) is judged. Figure 2a shows the results for these tasks, where we find that both LLMs (shades of gray) and humans (pink) follow the same pattern: the more causes are present, the higher the judged probability of the effect, which we take as evidence for agents understanding the causal mechanism and providing sensible judgments.

Alignment Increases under Chain-of-Thought Prompting Figure 3 shows human–LLM alignment measured via Spearman correlation (ρ) between the probability judgments of humans and each LLM (95% CIs via 2000 bootstrap resamples), under both Direct and CoT prompting. CoT prompting generally increases alignment for less aligned LLMs under the direct prompting condition ●.

3.2 Q2: Can small interpretable (cognitively-grounded) models capture / predict LLM reasoning?

> LLM causal judgments are compressible by small interpretable causal Bayes net (CBN)models. Predicting causal judgements on held out tasks during fitting, we find that *chain-of-thought*

> *(CoT)* increases the alignment of CBN and LLM probability judgments, especially pronounced in noisy conditions.

Rationale. All tasks share the same underlying collider structure ($C_1 \rightarrow E \leftarrow C_2$). Since RW17 does not specify causal strengths or base rates, there is no unique ground-truth probability judgment to score against Rehder & Waldmann (2017). We therefore evaluate *causal-model consistency*: whether an agent's probability judgments are well predicted by a compact causal model with the causal graph structure that formalizes the tasks.

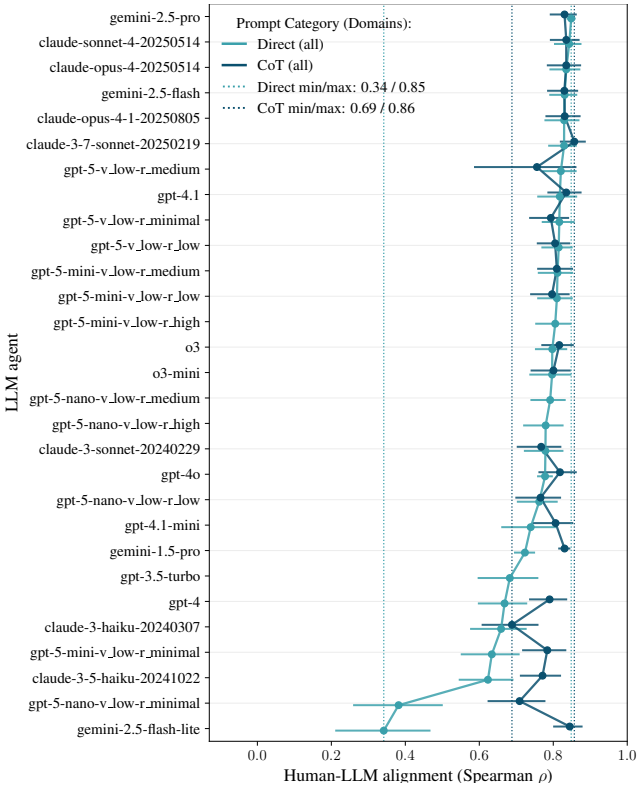


Figure 3: Human–LLM alignment: CoT ● boosts alignment.

Causal Bayes nets as small interpretable causal judgment baseline. For each agent, prompt condition, and 4 experimental conditions, we fit a causal Bayesian network (CBN) with a leaky noisy-OR parameterization Cheng (1997), pooling judgments across all 11 collider tasks t and domains, where the 11 tasks refer to different conditional-probability queries on the collider graph (see Figure 1, Figure 2). For binary causes $C_1, C_2 \in \{0, 1\}$, effect E, parameters θ background leak $b \in [0, 1]$ and causal strengths $m_1, m_2 \in [0, 1]$, the conditional probability is $\Pr(E{=}1 \mid C_1, C_2) = 1 - (1-b)\,(1-m_1)^{C_1}\,(1-m_2)^{C_2}$. This captures that each active cause independently increases the probability of the effect, while the leak b accounts for background activations in the absence of any active causes. We fit parameters by minimizing squared error between model predictions $\hat{y}_t(\theta)$ and an agent's normalized probability judgments y_t over tasks t.

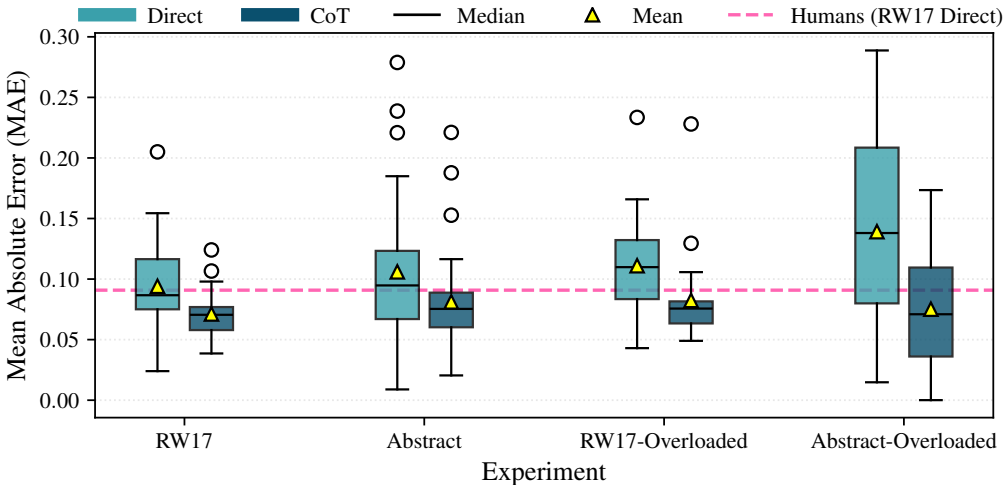


Figure 4: LLMs' probability judgments are well captured by an interpretable causal model, indicated by good causal Bayes net fits; CoT improves fit.

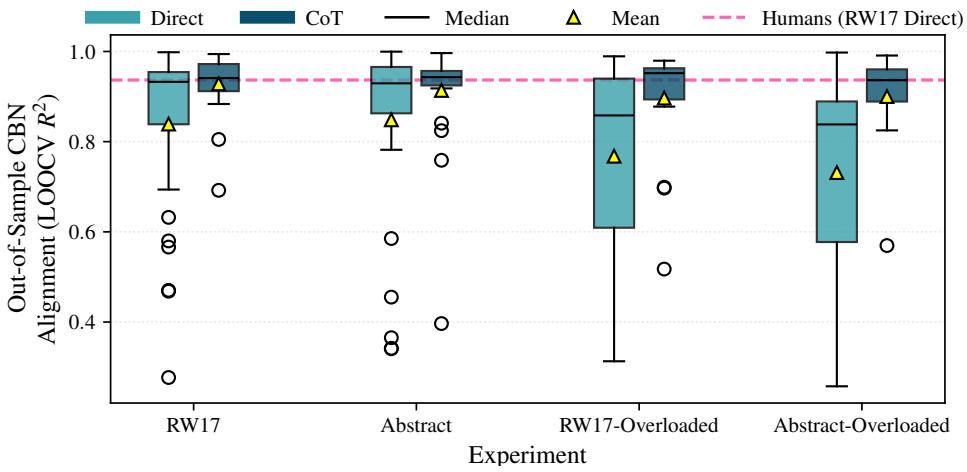


Figure 5: Out-of-sample CBN generalization (LOOCV R^2); higher is better.

LLMs' probability judgments are well captured by an interpretable causal model Figure 4 reports the mean absolute error (MAE) $\in [0, 1]$ (lower is better) for the fitted CBNs. Across experiments, most agents have a small MAE and are hence well-described by a single CBN. CoT ● prompting reduces MAE and dispersion relative to direct ● prompting. Overloaded prompts increase error, especially in the abstract setting. This indicates that both the direct prompting style ● and irrelevant content has agents apply more varied causal reasoning strategies.

CoT improves out-of-sample CBN generalization. To test whether an agent's judgments are captured by a *single* fitted CBN in a way that generalizes *beyond* the tasks used for fitting (rather than reflecting task-specific idiosyncrasies), we compute *out-of-sample CBN alignment* as task-level leave-one-out cross-validation (LOOCV) R^2 (higher is better; $R^2 \in (-\infty, 1]$). Concretely, we fit the CBN on 10 of 11 collider tasks, predict the held-out task, pool held-out predictions across folds, and compute R^2 on held-out judgments (Figure 5). CoT increases LOOCV R^2 and tightens dispersion across agents, with the largest gains under overloaded prompts, indicating that CoT improves the cross-task generalization of CBN-consistent causal reasoning under distraction. Several models approach or exceed the *human benchmark* (pink dashed line) and more so under CoT ●.

3.3 Q3: How much do agents attribute to external, unmentioned factors?

> Most LLMs act as stricter rule-followers than humans, showing high reliance on the stated causal links and low account for latent, unmentioned factors. Chain-of-Thought can shift models toward tighter causal rule-following in some cases, with mixed effects overall.

Rule-Following vs. Background Attribution: When agents interpret an explicitly stated causal model, they can differ in how much they treat the stated causes as *sufficient* versus how much they attribute outcomes to *external, unmentioned factors*. We operationalize this spectrum in a noisy-OR model via *Background-Adjusted Causal Strength* (BACS) $\in [-1, 1]$: $\mathrm{BACS}_{\text{agent}} = \overline{m}_{\text{agent}} - b_{\text{agent}}$, where $\overline{m}_{\text{agent}}$ is the mean of the causal strength parameters $m_{1,2}$ (how sufficient causes are perceived to be, $m = 1$ means completely sufficient) and b_{agent} is the leak/background parameter (baseline prevalence of the effect in the *absence* of the stated causes, capturing the influence of latent or unmentioned factors on the causal system). High $\mathrm{BACS} \to 1$ indicates *high causal rule-fidelity*: agents largely treat the provided causal relationships as sufficient (high m, low b). Low $\mathrm{BACS} \to -1$ reflects stronger *latent-factor accommodation* (elevated $b > 0$) and causes not being sufficient ($m < 1$). We therefore treat BACS as a summary index of reliance on the stated rule versus accounting for non-specified latent factors.

Most LLMs show higher causal rule-fidelity than humans Figure 6 shows Background-Adjusted Causal Strength (BACS) per experiment and prompt category by LLM.Most agents lie above the human benchmark (pink line) across experimental conditions marked by shape and size of scatters in Figure 6 indicating that LLMs typically follow the stated causal rule more tightly than humans,

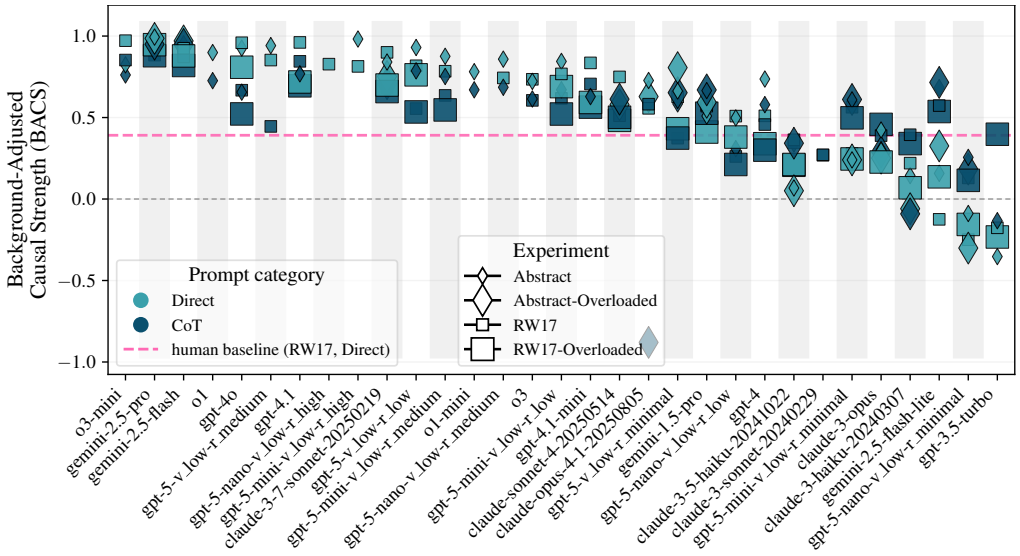


Figure 6: Background-Adjusted Causal Strength (BACS) per experiment and prompt category by LLM. Most LLMs have much higher rule-fidelity than humans. Higher BACS indicates tighter causal rule-following (stronger reliance on stated links relative to background leak), whereas lower BACS indicates greater background attribution and weaker adherence to the stated causal strengths.

whereas humans seem to be more prone to account for latent factors and treat the stated causes as insufficient. Some agents (e.g., `gemini-2.5-pro` and `gemini-2.5-flash`) approach near-maximal rule-fidelity (BACS$\approx$1.0) and are agnostic to experimental conditions as indicated by tight clusters. A smaller set of agents exhibits lower BACS than humans in the RW17-Direct condition (right side in Figure 6), treating the stated causal links as comparatively weak and expecting the effect to occur even in the absence of the stated causes. With CoT, these LLMs approach or slightly exceed the human baseline, whereas already highly rule-following models remain largely unaffected by prompting strategy or even become slightly less rule following with CoT.

3.4 Q4: Are LLMs biased like humans?

> Humans often show two biases in collider graphs: *weak* explaining away (EA) and *associative* dependence between causes (Markov violations (MV)). Across experiments, most LLMs show stronger explaining away than humans and are largely Markov-compliant; a minority exhibit human-like associative Markov violations. This suggests that most LLMs *do not* mirror the characteristic human collider biases (weak EA + MV).

Bias signatures in collider reasoning: no / weak explaining away and Markov violations In a collider graph, two qualitative signatures sharply distinguish normative causal inference from common human shortcuts. First, *explaining away* (EA) occurs when evidence for one cause reduces the belief in the alternative cause in the presence of the effect. We quantify the level of explaining away (EA) directly

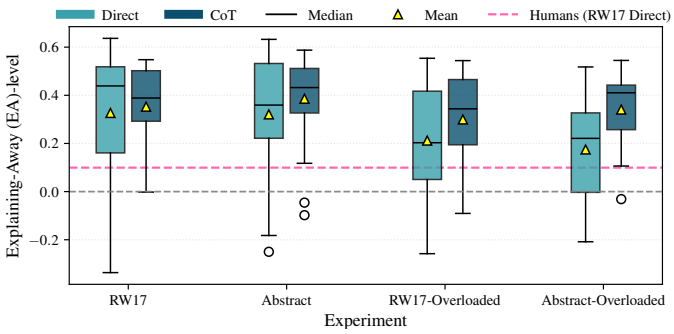


Figure 7: LLMs exhibit stronger explaining away (EA) than humans, indicated by EA-level $\gg 0$; Chain-of-Thought prompting enhances this effect in many models.

from agents' raw normalized probability judgments $\mathrm{EA} = \Pr(C_1{=}1 \mid E{=}1, C_2{=}0) - \Pr(C_1{=}1 \mid E{=}1, C_2{=}1)$, where $EA \leq 0$ represents absense of explaining away and the greater EA is, the more explaining away occurs. Qualitatively, explaining away corresponds to a positive slope in Figure 2c. Second, *Markov violation* captures the violation of the independence of causes in the absence of evidence about the effect: $\mathrm{MV} = \left|\Pr(C_1{=}1 \mid C_2{=}1) - \Pr(C_1{=}1 \mid C_2{=}0)\right|$, with Markov compliance when $\mathrm{MV} \approx 0$. Qualitatively Markov compliance corresponds to a flat line in Figure 2b. Humans frequently deviate on both: they often show *too little or no* explaining away Fernbach & Rehder (2013); Rottman & Hastie (2014) meaning evidence for one cause fails to reduce belief in an alternative cause. Humans also have been shown to *violate* cause-independence (Markov violation), meaning their judgment of one cause is influenced by the presence or absence of an alternative cause Ali et al. (2011); Mayrhofer & Waldmann (2016); Park & Sloman (2013); Rehder & Waldmann (2017). Whether LLMs show similar human biases is an open question we address in this work.

LLMs rarely reproduce human collider biases Across agents, explaining away is common and typically much stronger than in humans indicated by the mean and median EA-levels for LLMs in Figure 7 falling above the human baseline (pink, EA $\approx$ 0.1). Under overload, many LLMs show reduced EA relative to baseline, with CoT largely restoring EA toward baseline levels.

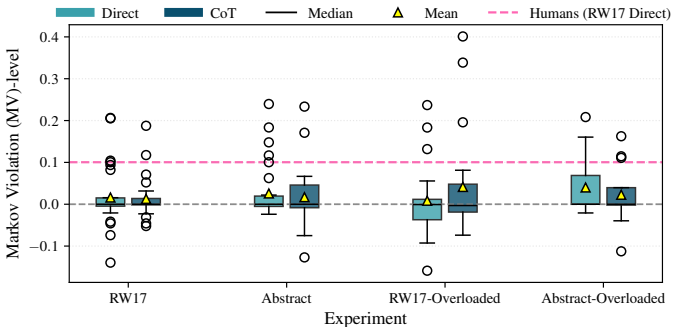


Figure 8: LLMs mostly don't exhibit Markov violation (MV), indicated by MV-level$\approx$0, while humans do.

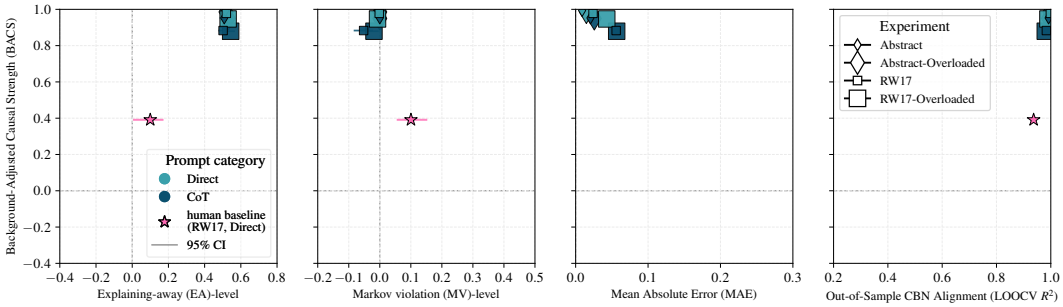
Most LLMs are Markov compliant; a minority show associative bias like humans. For the Markov condition, the modal pattern is *compliance*: most LLMs cluster tightly around $\mathrm{MV} = 0$ in Figure 8, indicating that one cause does not spuriously change belief in the other whereas for humans this is true. Under overload, Markov violations occur more frequently, albeit still less than in humans.

3.5 Q5: Do LLMs Reason Robustly Under Content Manipulations?

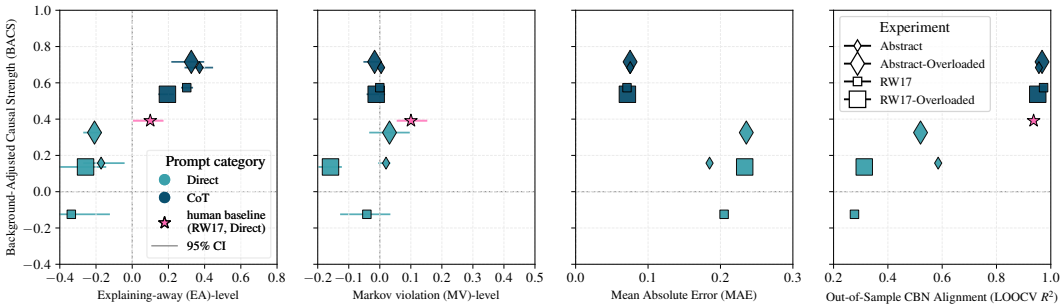
> Robustness remains highly model-dependent; Gemini-2.5-pro is near-invariant across all 8 experimental conditions. CoT typically increases robustness.

We evaluate whether agents preserve their causal-reasoning signatures under all 8 experimental conditions involving (i) content abstraction (RW17$\rightarrow$Abstract), (ii) noise injection/overload, and (iii) prompting strategy (Direct ● vs. CoT ●). Figure 9 summarizes robustness for a select number of LLMs at a glance across metrics considered in this work involving Background-Adjusted Causal Strength (BACS), Causal-Bayes-Net Alignment (LOOCV R^2), explaining away level (EA), and Markov Violation level (MV): each agent forms a cluster (robust) across all eight experimental conditions or separates (not robust).

Robustness is strongly model-dependent and CoT improves it for many models. *Gemini-2.5-pro* (Figure 9a) shows tight clustering across all 8 conditions (high BACS, high R^2, strong EA, MV$\approx$ 0), indicating near-invariant collider reasoning under all 8 manipulations. In contrast, *Gemini-2.5-flash-lite* (Figure 9b) shows clear prompting effects: For Direct ● prompting, clusters are more dispersed, while for CoT ● prompting, clusters are much tighter signaling higher content invariance when doing causal reasoning. CoT often increasing robustness by reducing the cluster size like for *Gemini-2.5-flash-lite* (Figure 9b) was e observed for several other models as well. CoT often increased robustness reducing the cluster size like for *Gemini-2.5-flash-lite* (Figure 9b).



(a) *Gemini-2.5-pro* reasons most robustly indicated by tight clusters among all 8 experimental conditions, showing content and prompting invariance.



(b) *Gemini-2.5-flash-lite* shows clear prompting effects indicated by color-separated clusters (Direct ●, CoT ●) with CoT ● increasing content invariance drastically.

Figure 9: Causal reasoning robustness across experimental conditions for metrics discussed in this work. Tight clustering indicates greater robustness to content manipulations and prompt variations.

4 Discussion

Our results suggest that many contemporary LLMs apply causal reasoning in collider graphs in a comparatively rule-like manner. Humans, by contrast, allocate more mass to unmentioned background factors (lower Background-adjusted causal strength; Figure 6), consistent with more open-world interpretations of the prompt. Yet LLMs' strong rule-following reasoning style risks failure in real-world settings where uncertainty is intrinsic, underscoring the need to better characterize LLM reasoning strategies to guide their safe and effective application. Despite being trained largely on human-generated text, most LLMs do not reproduce human collider biases: they show strong explaining away (EA) and are largely Markov-compliant, whereas humans exhibit low EA and frequent Markov violations. Robustness is strongly model-dependent and shaped by prompting: abstraction and overload shift causal signatures for many models, while CoT often restores more stable behavior; newer, larger models (e.g., Gemini-2.5-pro) are near-invariant across content manipulations, whereas smaller/older models are markedly more sensitive. Our results further show that both human and machine causal judgments can be predicted by—and compressed into—small interpretable causal Bayes nets. More broadly, this illustrates how Bayesian models can provide computational-level targets that help interpret and diagnose the behavior of multi-billion paramter neural networks, exemplifying that Bayesian and neural network models are complementary in understanding intelligence (Griffiths et al., 2024).

A key limitation is our focus on the common-effect (collider) graph. We chose this structure because colliders elicit well-studied signatures such as explaining away and because decades of work shows systematic human biases in precisely these settings; however, this limits the generality of the present conclusions. Additionally, our prompts do not explicitly specify whether unmentioned causes should be ignored or treated as plausible background factors, which may differentially affect humans and models and could partly explain the observed closed-world versus open-world gap. More broadly, the degree to which these laboratory-style tasks capture causal reasoning under real-world uncertainty remains an open question. Lastly, we don't have human responses for the overloaded and abstract conditions.

Future work should address the limitations above. Furthermore, human-LLM tandems could be explored to better understand how these models might complement human reasoning.

REFERENCES

Nilufa Ali, Nick Chater, and Mike Oaksford. The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, 119(3):403–418, 2011.

Christopher M. Barnes, Brian C. Gunia, and David T. Wagner. Sleep and moral awareness. *Journal of Sleep Research*, 24(2):181–188, 2015. doi: 10.1111/jsr.12231.

Andrea Bender. What is causal cognition? *Frontiers in psychology*, 11:3, 2020.

Patricia W. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367–405, 1997. doi: 10.1037/0033-295X.104.2.367.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1), January 2024. ISSN 2161-7201, 1946-5319. doi: 10.1093/jla/laae003. URL http://arxiv.org/abs/2401.01301.

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011. doi: 10.1073/pnas.1018033108.

Hanna M Dettki, Brenden M Lake, Charley M Wu, and Bob Rehder. Do large language models reason causally like us? even better? *Annual Conference of the Cognitive Science Society*, 2025.

Philip M Fernbach and Bob Rehder. Cognitive shortcuts in causal inference. *Argument & Computation*, 4(1):64–88, 2013.

Kanishk Gandhi et al. Human-like affective cognition in foundation models. *arXiv preprint arXiv:2409.11733*, 2024.

Thomas L Griffiths, Jian-Qiao Zhu, Erin Grant, and R Thomas McCoy. Bayes in the age of intelligent machines. *Current Directions in Psychological Science*, 33(5):283–291, 2024.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: A benchmark to assess causal reasoning capabilities of language models. *arXiv preprint arXiv:2312.04350*, 2023.

Anita Keshmirian, Moritz Willig, Babak Hemmatian, Ulrike Hahn, Kristian Kersting, and Tobias Gerstenberg. Biased causal strength judgments in humans and large language models. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL https://openreview.net/forum?id=544P6YidFk.

Andrew K Lampinen et al. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, July 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae233. URL https://doi.org/10.1093/pnasnexus/pgae233.

Ralf Mayrhofer and Michael R Waldmann. Sufficiency and necessity assumptions in causal structure induction. *Cognitive science*, 40(8):2137–2150, 2016.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Juhwa Park and Steven A Sloman. Mechanistic beliefs determine adherence to the markov property in causal reasoning. *Cognitive psychology*, 67(4):186–216, 2013.

Bob Rehder. A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1141, 2003.

Bob Rehder. Independence and dependence in human causal reasoning. *Cognitive Psychology*, 72: 54–107, 2014. doi: 10.1016/j.cogpsych.2014.02.002.

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

Lars Riedemann, Maxime Labonne, and Stephen Gilbert. The path forward for large language models in medicine is open. *npj Digital Medicine*, 7(1):339, 2024.

Benjamin Margolin Rottman and Reid Hastie. Reasoning about causal relationships: Inferences on causal networks. *Psychological bulletin*, 140(1):109, 2014.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Vinod Venkatraman, Y. M. Lisa Chuah, Scott A. Huettel, and Michael W. L. Chee. Sleep deprivation elevates expectation of gains and attenuates response to losses following risky decisions. *Sleep*, 30 (5):603–609, 2007. doi: 10.1093/sleep/30.5.603.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Moritz Willig, Matej Zecevic, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*, 8, 2023.

Matthew T. Wittbrodt and Melinda Millard-Stafford. Dehydration impairs cognitive performance: A meta-analysis. *Medicine & Science in Sports & Exercise*, 50(11):2360–2368, 2018. doi: 10.1249/MSS.0000000000001682.

Jian-Qiao Zhu and Thomas L Griffiths. Incoherent probability judgments in large language models. *arXiv preprint arXiv:2401.16646*, 2024.