Do Large Language Models Understand Cause and Effect?

The nature of intelligence in both humans and machines is a long-standing question in cognitive science. While there is no universally accepted definition, the ability to reason causally is often regarded as a pivotal aspect of intelligence (Lake et al., 2017). Large Language Models (LLMs), trained on vast text volumes, have demonstrated impressive capabilities in generating human-like text given a prompt. This has sparked a debate about whether these models have the capacity for "understanding" (Mitchell et al., 2023) or if they rely primarily on statistical patterns in human language (Willig et al., 2023). While evaluating how well LLMs perform causal inference is an active field of research (Kıcıman et al., 2023; Jin et al., 2023), their performance has rarely been contrasted to those of humans on the same tasks. This is a key question, which provides insights into whether LLMs reproduce human biases and whether their performance lags or even exceeds human capabilities to reason causally.

Contribution This work investigates the causal reasoning abilities of language models by benchmarking them on custom, well-defined causal inference tasks against the performance of human subjects. We algorithmically generated a dataset of natural language causal inference question based on an underlying graph. In contrast to Keshmirian et al. (2024), we focus on the collider graph, rather than the chain or common-cause graph, allowing us to investigate phenomena like "explaining away", for which human causal reasoning is known to be biased (Rehder, 2014; Cruz et al., 2020). Following Jin et al. (2023), to assess reliance on statistical patterns rather than causal inference, our dataset consists of "commonsensical" scenarios, meaning the variables and relationships align with the physical world, and "anti-commonsensical" scenarios, which are physically implausible but constitute a valid causal inference task. We find that humans outperform LLMs on our custom dataset, but the difference between ChatGPT4 and humans is not as pronounced as one might expect, while ChatGPT-3.5 performs no better than chance.

Dataset Generation We algorithmically generated the causal inference dataset in two steps: First, we choose a causal model of the "world", given by a Causal Bayesian Network (CBN) (Pearl, 2009), and observations of the causal variables C_1, C_2 and observed effect E, as well as a query type (e.g. $P(C_1 \mid C_2, E) < P(C_1 \mid E)$?). Next, we select the empirical alignment (commonsensical or anti-commonsensical), the "story" describing the causal variables in natural language (see Fig. 1 for illustrations), and finally synthesize the full causal inference task into a prompt. Our focus is on the collider graph, linking two causes to a common effect via `AND`, `OR`, or `XOR`. This enables comparison with human studies on how observing C_2 affects the inferred probability of C_1. Specifically, whether the underlying causal structure implies *explaining away* ($P(C_1 \mid C_2, E) < P(C_1 \mid E)$), *augmentation* ($P(C_1 \mid C_2, E) > P(C_1 \mid E)$), or *conditional independence* ($P(C_1 \mid C_2, E) = P(C_1 \mid E)$). Humans exhibit systematic biases in these causal inferences, diverging from the ground truth (Rehder, 2014; Cruz et al., 2020), which prompts the question: Do LLMs replicate these biases?

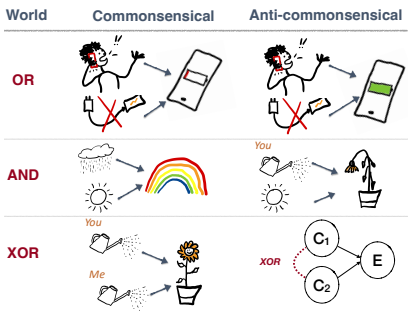
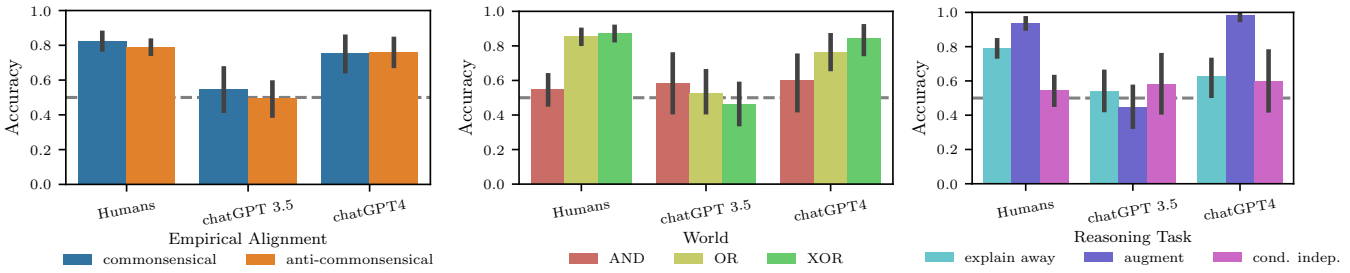


Figure 1: Illustrations of prompts given a causal "world" model and an empirical alignment type.



(a) Empirical alignment. (b) Causal world model. (c) Reasoning task.

Figure 2: Accuracy of recorded answers compared to normative ground truth answer with 95% CI computed via bootstrapping.

Results We compared ChatGPT-3.5, ChatGPT-4, and four human subjects on 180 causal inference questions. As Fig. 2 shows, on average humans outperformed LLMs with 0.80 accuracy, followed by ChatGPT-4 (0.76), while ChatGPT-3.5 performed no better than chance. Stratified by empirical alignment, humans performed slightly better in the commonsensical scenario (0.83) than the anti-commonsensical one (0.79), while ChatGPT-4 performed equally in both (0.76), and ChatGPT-3.5 showed no significant difference (Fig. 2(a)). Broken down by "world" (`AND`, `OR`, `XOR`), humans struggled most with the `AND` relationship (0.55) and performed best at `XOR` (0.88) (Fig. 2(b)). ChatGPT-4 followed a similar trend but with lower accuracy, except for `AND`, while ChatGPT-3.5 showed opposite patterns. Humans' struggles with the `AND` world suggest difficulty grasping conditional independence. ChatGPT-4 shares this bias, confirmed by Figure 2(c), where both perform best on augmentation and worst on conditional independence tasks. In summary, humans outperform LLMs on causal inference tasks, though the gap is smaller than expected, suggesting LLMs have some causal reasoning ability. Humans' reliance on pattern matching is hinted at by stronger performance on commonsensical questions. LLMs have been previously demonstrated to also only act as "causal parrots" (Willig et al., 2023). Our results so far do not clearly support this observation for ChatGPT4. Future work should investigate the extent to which LLMs can generalize causal ideas beyond their training distribution and compare it to human performance on the same tasks, as piloted in this work.