Does Reasoning Emerge? Examining the Probabilities of Causation in Large Language Models

Javier González
Gonzalez.Javier@microsoft.com
Microsoft Research, Cambridge

Aditya V. Nori
Aditya.Nori@microsoft.com
Microsoft Research, Cambridge

Abstract

Recent advances in AI have been significantly driven by the capabilities of large language models (LLMs) to solve complex problems in ways that resemble human thinking. However, there is an ongoing debate about the extent to which LLMs are capable of actual *reasoning*. Central to this debate are two key probabilistic concepts that are essential for connecting causes to their effects: the probability of necessity (PN) and the probability of sufficiency (PS). This paper introduces a framework that is both theoretical and practical, aimed at assessing how effectively LLMs are able to replicate real-world reasoning mechanisms using these probabilistic measures. By viewing LLMs as abstract machines that process information through a natural language interface, we examine the conditions under which it is possible to compute suitable approximations of PN and PS. Our research marks an important step towards gaining a deeper understanding of when LLMs are capable of reasoning, as illustrated by a series of math examples.

1 Introduction

Large language models (LLMs) have revolutionized the way we interact with technology, enabling more natural and intuitive communication between humans and computers in applications like writing assistants [8], sentiment analysis in social media [29], healthcare [10, 35] and many others. Despite the surge of interest and recent breakthroughs [5], the ability of LLMs to *reason* about real-world problems as humans do continues to be a topic of intense research [1, 14].

Reasoning is a cognitive process that involves drawing conclusions, making judgments, or forming inferences based on facts or premises. This process has been explored from various perspectives. *Symbolic reasoning* [17] involves the manipulation of symbols that represent ideas or objects and it is often used in mathematics and logic to represent numerical values or logical propositions. *Causal reasoning* [26] focuses on discerning the relationship between a cause and its effect, aiming to understand how certain events can impact other. Other forms of reasoning include *inductive reasoning* [7] (making broad generalisations from specific observations), *deductive reasoning* [27] (applying general principles to specific cases), and *abductive reasoning* [2] (forming the best hypothesis based on incomplete information).

In the realm of LLMs, reasoning is typically understood to be the ability of these models to demonstrate *emergent* capabilities that surpass mere statistical pattern recognition in the training set. It entails systematically breaking down problems into a logical sequence of smaller, manageable steps and then processing these steps internally to arrive at accurate conclusions that are grounded in reality. This concept is the foundation for techniques such as *chain of thoughts prompting* [34], which aim to

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Direct prompt

Does six divide {number}? Use the factor method to answer this question. Be as concise as possible.

Counterfactual prompt

Imagine that {number} has three as a prime factor while retaining all its other prime factors. With this assumption does six divide {number}? Use the factor method to answer this question. Be as concise as possible.

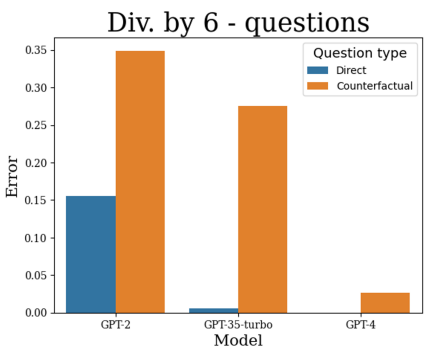


Figure 1: Illustration of the actual vs. perceived reasoning abilities of GPT-2, GPT-35-turbo and GPT-4 for a simple arithmetic problem. We posed two distinct types of questions (direct and counterfactual) to the models, each repeated 10 times, for every `{number}` from 1 to 50. All three models showed an inflated sense of reasoning capability when answering the direct questions. The discrepancy is especially pronounced in GPT-35-turbo, which performed nearly flawlessly on direct questions, but experienced a surge in error rate, exceeding 25%, when handling counterfactual questions.

teach LLMs how to reason by providing examples where problems are solved through a sequence of smaller steps.

Assessing the reasoning abilities of LLMs involves distinguishing between two aspects: the accuracy with which an LLM solves a problem, and its capacity to understand and process the fundamental elements that lead to that solution. Judea Pearl, in his hierarchy of causality [25], asserts: *"Only machines that can correctly perform correlations, interventions and counterfactuals will have reasoning abilities comparable to human"*. As demonstrated in [16], while LLMs are remarkable in using learnt patterns from their training data to generate correct answers (correlations), they falter when faced with hypothetical/imaginary scenarios that were not part of their training (counterfactuals). This is depicted in Figure 5, which presents a straightforward arithmetic problem (this is the **direct prompt** in the figure). Both GPT-35-turbo and GPT-4 can accurately determine the divisibility of numbers by 6, suggesting at first glance that they can reason about divisibility. However, when the questions are framed in a counterfactual manner (this is the **counterfactual prompt** in the figure), only GPT-4 maintains a low error rate, indicating its superior ability to handle such reasoning tasks.

In this paper, we introduce a systematic method to assess the reasoning capabilities of LLMs by examining the concepts *necessity* and *sufficiency*, which are key elements of logical reasoning and have been studied in multiple fields such as logic, probability, and causality [22, 19, 12]. In propositional logic, a sufficient condition is defined as $X \implies Y$, indicating that the presence of X ensures the occurrence of Y. On the other hand, a necessary condition is defined as $Y \implies X$, signifying that the occurrence of Y necessitates the prior occurrence of X. We focus on the probabilistic interpretations of necessity and sufficiency [24]. The *probability of necessity* (PN) between two boolean variables X and Y is defined as $\mathrm{PN}(x, y) := \mathbb{P}(y'_{x'}|x, y)$. Here, $y'_{x'}$ represents the counterfactual value of $Y = y'$, had X been set to a different value x'. By conditioning on both $X = x$ and $Y = y$, this measure captures probability of observing a different outcome in the absence of the event $X = x$. The *probability of sufficiency* (PS), on the other hand, is defined as $\mathrm{PS}(x, y) := \mathbb{P}(y_x|x', y')$ and measures the probability that $X = x$ results in $Y = y$, for cases where both originally had different values.

We show that when a problem can be solved via a reasoning graph of boolean conditions, denoted by $\mathcal{G}$, the PN and PS can be computed using a causal model underlying $\mathcal{G}$. As described in [24], the exact computation of PN and PS requires samples from the (causal) data generative model, counterfactual data (experiments) as well as other monotonicity assumptions. As a *reasoning test*, we statistically compare the true PN and PS measures (computed by sampling from the original and the intervened graph) with those simulated via factual and counterfactual datasets generated by an LLM. Figure 2 presents an informal illustration of the reasoning test advocated in this paper, focusing on the specific problem of determining whether a number N is divisible by 6. The test relies on the reasoning principle that: "*A natural number N that is divisible by both 2 and 3 is also divisible by 6*". This logic is represented in a reasoning graph $\mathcal{G}$ that links the conditions C_2 (divisibility by 2) and C_3

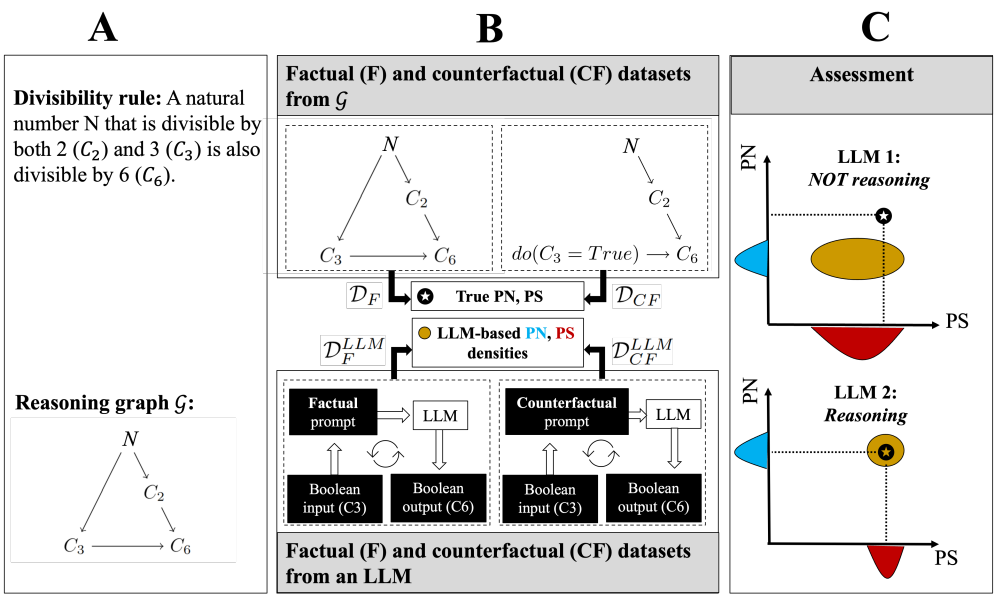


Figure 2: Reasoning test for assessing an LLM's reasoning abilities. **A)** Divisibility rule and the corresponding reasoning graph. **B)** Dataset generation for computing PN and PS. **C)** Analysis comparing actual values of PN and PS with PN and PS estimates for the LLM-generated data.

(divisibility by 3) to the conclusion C_6 (divisibility by 6). We test the reasoning abilities of an LLM using natural numbers N from 1 to 400. This is shown in Figure 2(A).

As indicated in Figure 2(B), we create two sets of data based on $\mathcal{G}$. The first is a factual dataset ($\mathcal{D}_F$) which captures whether each number N satisfies conditions C_2 and C_3. The second is a counterfactual dataset, ($\mathcal{D}_{CF}$), which assumes condition C_3 is always true and then records whether each number X would satisfy C_6 under this assumption/intervention (realised by $do(C_3 = True)$ in the figure). For the LLM being evaluated, we also produce two datasets. The first, $\mathcal{D}_F^{LLM}$, documents the LLM's response for C_6 for each number N, when the prompt is based on the reasoning graph $\mathcal{G}$. The second, $\mathcal{D}_{CF}^{LLM}$, involves a hypothetical scenario where we assume C_3 is true and then record the LLM's prediction for C_6 given this "counterfactual prompt". This process is repeated multiple times (several answers from the LLMs are collected from each prompt). We assess the LLM's reasoning capability by comparing the estimated (distribution of) PN and PS from the $\mathcal{D}_F^{LLM}$ and $\mathcal{D}_{CF}^{LLM}$ datasets with the actual values derived from $\mathcal{D}_F$ and $\mathcal{D}_{CF}$ datasets. Figure 2(C) displays these comparisons, plotting PN vs. PS. The closer the estimated PN/PS values to the actual PN/PS values, the better it is at reasoning. In this case, LLM 2 demonstrates better reasoning abilities than LLM 1.

Related work: Reasoning in LLMs has been studied from multiple perspectives. [15] presents an overview paper that elucidates key reasoning concepts utilised by LLMs. [13] examines the similarity between reasoning with a language model and planning with a world model, proposing a novel reasoning framework that redefines the LLM as both a world model and a reasoning agent. Various studies [28, 4] have focused on assessing the reasoning and problem-solving abilities of LLMs, yet none have used the probabilities of causation as the primary objects of computation as done in our research. [32] carries out a series of experiments to show that LLMs can indeed derive benefits from reasoning errors, offering potentially cost-effective strategies by using mistakes to bolster reasoning capabilities. Recent research indicates that LLMs like GPT-3.5 and GPT-4 are effective at causal reasoning tasks, including pairwise causal discovery [18]. These models have achieved state-of-the-art performance on multiple causal benchmarks, outperforming existing algorithms. Nevertheless, LLMs also exhibit unpredictable failure modes, and currently, they are not capable of discovering new knowledge or making high-stakes decisions with a high level of precision [20, 36, 21, 6, 3].

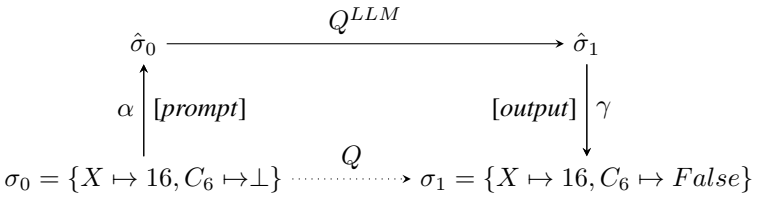


Figure 3: The HEX diagram depicts two approaches for solving the problem (Q, σ_0) outlined in Example 1. The dotted path corresponds to the actual process of solving the problem, while the solid path represents the one taken by the LLM.

Contributions: This paper presents two main contributions.

i) A novel, theoretical and practical, framework to evaluate the reasoning capabilities of language models using the probabilistic of causation, specifically the probability of necessity and the probability of sufficiency. Our approach is unique in the sense that it allows to differentiate generalization by reasoning from merely replicating statistical patterns in the training data.

ii) Empirical tests on various reasoning problems as well as several insights about the reasoning abilities of language models in the GPT family.

2 LLMs as abstract machines

As described by the HEX framework [9], an LLM functions as an abstract machine that uses natural language as an interface. In this section, we introduce the core elements of this framework, which will subsequently enable us to define an LLM's internal representation of PN and PS.

We define a *problem* as a query-state pair (Q, σ). The state σ is a mapping defined by $\sigma : \mathcal{V} \to \mathcal{C}$, which assigns values from a specified domain $\mathcal{C}$ to a set of variables $\mathcal{V} = \{V_1, \ldots, V_n\}$. The query $Q : 2^{\mathcal{V} \to \mathcal{C}} \to 2^{\mathcal{V} \to \mathcal{C}}$ is a mapping that transforms an input state σ to a well-defined output state. To *solve a problem* is to calculate $\sigma_1 = Q(\sigma_0)$, where σ_0 and σ_1 represent the states before and after the query Q is applied. To clarify this, we consider the following example:

Example 1. *"Given that a natural number divisible by both 2 and 3 is also divisible by 6, determine whether the number 10 is divisible by 6."*

To solve Example 1, we apply the query Q to the state $\sigma_0 = \{N \mapsto 10, C_6 \mapsto \bot\}$, where $Q = \lambda\sigma\,.\,(\sigma(N) \pmod 2 \equiv 0) \wedge (\sigma(N) \pmod 3 \equiv 0)$[1]. This results in a final state $\sigma_1 = \{N \mapsto 10, C_6 \mapsto False\}$, thereby resolving the problem with $\sigma_1(C_6) = Q(\sigma_0) = False$.

We now turn to the question of how an LLM solves a problem defined by a query-state pair (Q, σ_0). This process involves three essential steps as illustrated by Figure 3:

1. First, an abstraction mapping translates the initial state σ_0 into a latent state $\hat{\sigma}_0$ via a *prompt*.
2. Next, the LLM processes (via the query Q^{LLM}) this latent state $\hat{\sigma}_0$.
3. Finally, the output mapping transforms the LLM output latent state $\hat{\sigma}_1$ back into a concrete state, producing the final *output* σ_1.

Formally, solving a problem (Q, σ_0) with an LLM can be described as a sequence of function applications resulting in the output $\sigma_1 = (\gamma \circ Q^{LLM} \circ \alpha)(\sigma_0)$. To illustrate this, the problem statement is given as a prompt input to GPT-4 [23]. The response from GPT-4 is "False", which matches the result obtained by applying the query Q directly to the input state σ_0. When both the direct application of Q and the LLM computation yield the same answer, we say that the diagram, as shown in Figure 3, is commutative–meaning that following either the dotted line or the solid lines lead to the same result. For a more in-depth explanation of this framework, please refer to [9].

[1] See https://en.wikipedia.org/wiki/Lambda_calculus for a quick introduction to Lambda calculus.

3 Probabilities of causation for an LLM

To assess the reasoning abilities of an LLM, we must link its generated responses to the actual reasoning processes that produced those responses. For a problem (Q, σ), we postulate the existence of a causal model $\mathcal{M}_\mathcal{V}$ defined over variables in $\mathcal{V}$, and by a set of structural equations and endogenous variables. For a detailed introduction to causal models, refer to Appendix A. Additionally, the seminal work by Pearl [25] on causality provides foundational insights on this area. Here, we are particularly interested in causal models that represent the logical steps involved in problem-solving. However, it is important to note that the concept of a causal model is broadly applicable beyond this specific application.

We assume that $\mathcal{V} = \{X, Y, Z\}$, which includes X and Y as boolean variables, and Z as a variable (which may be multivariate) that encompasses all necessary factors that are required to understand how an intervention on X would affect Y. In the context of causality, this means that the distribution $\mathbb{P}(Y|do(X = x'))$, where do denotes the intervention operator defined in [25], is identifiable. This means we can predict the outcome for Y, and that the counterfactual $Y_{X=x'}$, that can be read as *"the value of Y had X been x'"*, is well-defined. For further details, please refer to Appendix A. For ease of exposition in the following text, we will simplify our notation by omitting the explicit reference to Z. Therefore, we will denote $Y_{X=x}(Z = z)$ more succinctly as $Y_{X=x}$.

As studied in [31], if Y is monotonic with respect to X, then PN and PS can be computed as follows:

$$\texttt{PN}(x, y) = \frac{\mathbb{P}(y) - \mathbb{P}(y|do(x'))}{\mathbb{P}(x, y)} \text{ and } \texttt{PS}(x, y) = \frac{\mathbb{P}(y|do(x)) - \mathbb{P}(y)}{\mathbb{P}(x', y')}. \tag{1}$$

To estimate PN and PS, we need two different types of datasets. The first is a *factual* dataset $\mathcal{D}_F = \{x_i, y_i, z_i\}_{i=1}^n$, which is used to infer $\mathbb{P}(y)$, $\mathbb{P}(x, y)$ and $\mathbb{P}(x', y')$. The second dataset $\mathcal{D}_{CF} = \{x_i, Y_{X=x_i}, z_i\}_{i=1}^n$ is a *counterfactual* one, and is necessary to determine $\mathbb{P}(y|do(x))$ and $\mathbb{P}(y|do(x'))$.

There are various methods to acquire the datasets $\mathcal{D}_F$ (factual) and $\mathcal{D}_{CF}$ (counterfactual). For a physical process, the usual method would be through observation and experimentation. However, in this paper, we presume access to a comprehensive reasoning graph that is equivalent to a causal model $\mathcal{M}_\mathcal{V}$. This allows us to simulate and generate the $\mathcal{D}_F$ and $\mathcal{D}_{CF}$ datasets. Both $\mathcal{M}_\mathcal{V}$ and the sub-model $\mathcal{M}_{\mathcal{V},do(X=x)}$ define two distinct joint probability distributions $\mathbb{P}_{\mathcal{M}_\mathcal{V}}$ and $\mathbb{P}_{\mathcal{M}_{\mathcal{V},do(X=x)}}$ over X, Y and Z. We obtain the datasets $\mathcal{D}_F$ and $\mathcal{D}_{CF}$ by sampling from these respective probability distributions. These datasets are then used to calculate PS and PN using Eq. (1).

3.1 LLM-based counterfactuals

Can an LLM reason in a manner that is consistent with $\mathbb{P}_{\mathcal{M}_\mathcal{V}}$? In Example 1, we obtained consistent answers (that is, the corresponding HEX diagram commutes) for a direct divisibility question. However, to evaluate the reasoning abilities of the LLM, it is crucial that this consistency is also observed when the queries are framed in a counterfactual manner. This is necessary to ensure that the LLM can apply its reasoning to imaginary situations that are unlikely to be present in the training set, demonstrating its ability to generalise based on a correct internal representation of the reasoning logic of the problem. Practically, this means employing the LLM as a "counterfactual data simulator", where the data it generates under these hypothetical conditions are used to estimate PN and PS.

Definition 1 (Counterfactual query)**.** *Consider a problem (Q, σ_0), with $\sigma_0 = \{X \mapsto x, Y \mapsto y, Z \mapsto z\}$ being an initial state. Let $\mathcal{M}_\mathcal{V}$ be a causal model over the variables $\mathcal{V}$. We can then define a counterfactual query Q' as follows: $Q'(\sigma_0) = \{X \mapsto x', Y \mapsto Y_{X=x'}, Z \mapsto z\}$.*

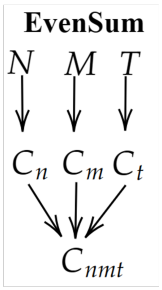
In other words, a counterfactual query updates two variables of the state: it sets X to its new value x', and Y to the counterfactual $Y_{X=x'}$. An LLM-based counterfactual $Y_{X=x'}^{LLM}$ is computed as follows:

$$Y_{X=x'}^{LLM} = (\gamma \circ Q'^{LLM} \circ \alpha)(\sigma_0)(Y)$$

where $\sigma_0 = \{X \mapsto x, Y \mapsto y, Z \mapsto z\}$, and Q'^{LLM} is a counterfactual query. This entire process simulates counterfactual reasoning within the LLM, and is facilitated through textual prompts that are structured to elicit the desired counterfactual outcome.

Definition 2 (Counterfactual prompt)**.** *A counterfactual prompt is a textual encoding of a counterfactual query for some initial state σ_0.*

$\mathcal{D}_F$	$C_6=0$	$C_6=1$
$C_3=0$	267	0
$C_3=1$	67	66
$\mathcal{D}_{CF}$	$C_6=0$	$C_6=1$
$C_3=0$	133	0
$C_3=1$	133	134
$\mathcal{D}_F^{GPT-4}$	$C_6=0$	$C_6=1$
$C_3=0$	267	0
$C_3=1$	67	66
$\mathcal{D}_{CF}^{GPT-4}$	$C_6=0$	$C_6=1$
$C_3=0$	132	1
$C_3=1$	132	135



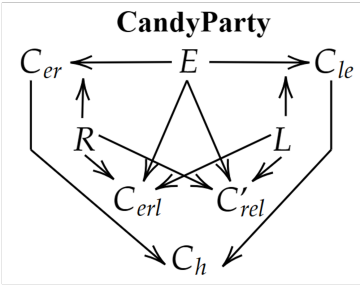


Figure 4: *Left*: Contingency tables for $\mathcal{D}_F$, $\mathcal{D}_{CF}$ and $\mathcal{D}_{CF}^{\text{GPT}-4}$ in Example 1. *Right*: Reasoning graphs for the other math problems in this paper. C-type nodes in the graph represent boolean conditions. See Appendix C for details.

Figure 1 shows an example of a counterfactual prompt. To create a comprehensive dataset $\mathcal{D}_{CF}^{LLM}$ of counterfactuals based on an LLM, we start with the factual dataset $\mathcal{D}_F^{LLM}$. From this dataset, we generate a set of initial states $\sigma_{0,i} = \{X \mapsto x_i, Y \mapsto y_i, Z \mapsto z_i\}$, which serve as the basis for deriving counterfactuals using the LLM. To compute PN and PS, we substitute $\mathcal{D}_F$ with $\mathcal{D}_F^{LLM}$ and $\mathcal{D}_{CF}$ with $\mathcal{D}_{CF}^{LLM}$ in Eq. (1).

Example 1 revisited. We construct four distinct datasets using every integer in $[1, 400]$: the factual dataset $\mathcal{D}_F$, the counterfactual dataset $\mathcal{D}_{CF}$, the LLM-based factual dataset $\mathcal{D}_F^{LLM}$, and the LLM-based counterfactual dataset $\mathcal{D}_{CF}^{LLM}$. These datasets, shown in Figure 4 (*Left*) are generated following the causal model shown in Figure 2, its modified version with interventions, and the LLM prompting methods mentioned previously. We obtain $\texttt{PN} = 1$ and $\texttt{PS} = 0.50$ for the datasets $\mathcal{D}_F$ and $\mathcal{D}_{CF}$. On the other hand, $\texttt{PN}^{\text{GPT}-4} = 0.984$ and $\texttt{PS}^{\text{GPT}-4} = 0.505$, when we use the factual $\mathcal{D}_F^{LLM}$ and counterfactual $\mathcal{D}_{CF}^{LLM}$ datasets generated by GPT-4.

3.2 Counterfactual consistency in LLMs

Definition 3 (β-counterfactual consistency)**.** *Consider a structural causal model $\mathcal{M}_\mathcal{V}$ with variables $\mathcal{V} = \{X, Y, Z\}$. Let $\mathcal{A}_{X=x}(Z)$ be a function that generates counterfactuals for Y. We say that $\mathcal{A}$ is β-counterfactual consistent with $\mathcal{M}_\mathcal{V}$ if the following condition is satisfied: $\mathbb{E}_{\mathbb{P}(X,Y,Z)}\left[\mathcal{A}_{X=x}(Z=z) \neq Y_{X=x}(Z=z)\right] \leq \beta$, where $\beta \leq 0$.*

β-counterfactual consistency defines the limit error rate for counterfactuals produced by $\mathcal{A}_{X=x}(Z = z)$. This error rate should ideally be zero for an LLM that exhibits flawless reasoning abilities. The following lemma specifies the conditions necessary for this property to hold (the proof can be found in Appendix D).

Lemma 1. *Let $\mathcal{M}_\mathcal{V}$, with variables $\mathcal{V} = \{X, Y, Z\}$, be a structural causal model for a problem (Q, σ_0), and let M be an LLM that generates counterfactuals for Y. Then M is β-counterfactual consistent with $\mathcal{M}_\mathcal{V}$ if and only if its associated* HEX *diagram for the problem (Q', σ_0), where Q' is the counterfactual version of Q, is commutative for all admissible values of X, Y and Z.*

Lemma 1, provides a theoretical criterion to test the consistency of an LLM with some reasoning graph described by $\mathcal{M}_\mathcal{V}$. As we show next in the experimental section, the commutability of the diagram is tested in expectation by sampling over the variables in $\mathcal{V}$.

4 Empirical illustrations

We focus on three math problems, each with a progressively higher level of difficulty.

Divisibility by 6 (`Div6`): We compute the PN and PS to determine the impact that an integer N's divisibility by 3 (denoted as C_3) has on its divisibility by 6 (denoted as C_6). For our analysis, we consider $N \in [1, 400]$.

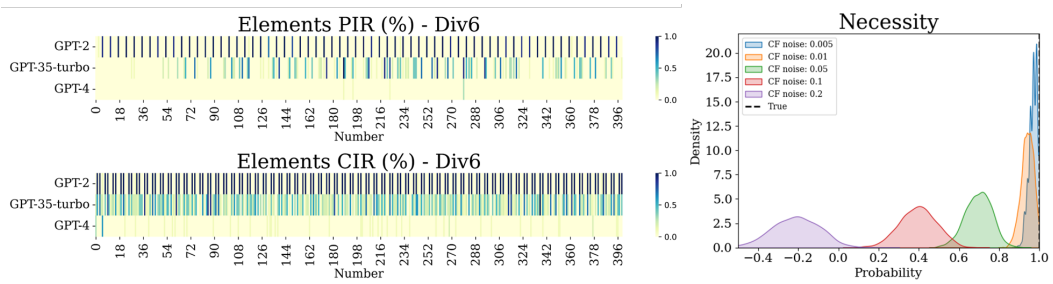


Figure 5: *Left*: Heatmaps comparing the consistency of data generated by GPT-2, GPT-3.5-turbo, and GPT-4 for the `Div6` problem. Each heatmap cell represents the error rate of the corresponding model for each element of the problem across 10 replicated tests. *Right*: Sensitivity of the simulated PN relative to varying levels of random noise introduced in the true counterfactuals.

Even sum of integers (`EvenSum`): We examine scenarios where the sum of three integers M, N and T is even. This can occur under two conditions: when all three integers are even, or when one is even and the other two are odd. We evaluate PN and PS for impact that M being odd or even (C_m) has on the resulting sum being odd or even (C_{mnt}). For our analysis, we consider all possible values for M, N and T, with each integer ranging from 1 and 8.

Candy party (`CandyParty`): In this hypothetical scenario, Rafa is having his birthday party with two guests, Lara and Emma. They have 20 candies to distribute among themselves. The party will be considered 'happy' if the candy distribution satisfies at least one of the following conditions: (i) Each person gets the same number of candies, or (ii) Rafa gets more candies than both Lara and Emma, but Lara and Emma each receive an equal number of candies, with both receiving at least one candy each. We compute the PN and PS for the impact that Lara and Emma receiving an equal number of candies (denoted as C_{lm}) has on the party being 'happy' (denoted as C_h).

A fourth problem (`ConPref`) is included in Appendix B. The reasoning graphs for the problems `EvenSum` and `CandyParty` are shown in Figure 4 (*Right*). The structural equations corresponding to each of these graphs can be found in the Appendix C. We estimate the PN and PS for each of these problems using three difference language models: GPT-2, GPT-3.5-turbo and GPT-4 [23]. Our objective is to investigate whether the ability to reason, as conceptualised in this paper, *emerges* as the complexity and size of the models grow. While similar evaluations could be conducted using other families of LLMs, such as Llama [33], Gemini [30], Phi [11], etc., we have chosen to limit our analysis to the GPT series here for the sake of a clearer and more straightforward exposition, but more results are available in Appendix J.

To assess the reasoning abilities of various models, we use the following metrics:

1. *Factual Inconsistency Rate* (FIR): This measures the rate of inconsistencies when the models respond to factual queries.
2. *Counterfactual Inconsistency Rate* (CIR): Similar to FIR, but this metric measures inconsistencies in responses to counterfactual queries.

For a detailed explanation of these metrics, please refer to Appendix H. We estimate the standard errors of FIR and CIR by examining the variations in outputs across multiple model responses. We take this aspect into account by collecting multiple answers from the models and propagating the stochasticity of the answers to the computation of PN and PS. Additionally, we capitalise on this variability to construct the densities over the inferred PN and PS. This process involves generating 500 bootstrap samples from the model's factual and counterfactual responses[2]. From these densities, we calculate γ-PN-overlap, which measures the concentration of the probability distribution within a radius γ around the actual PN, and γ-PS-overlap does the same for PS[3].

[2]Note that while generating a larger number of model answers could potentially increase accuracy, the computational costs are prohibitive. Therefore, the bootstrap approach serves as a reasonable compromise.

[3]The code to reproduce the analyses and figures can be provided upon request, and will be made open source if this work is accepted for publication.

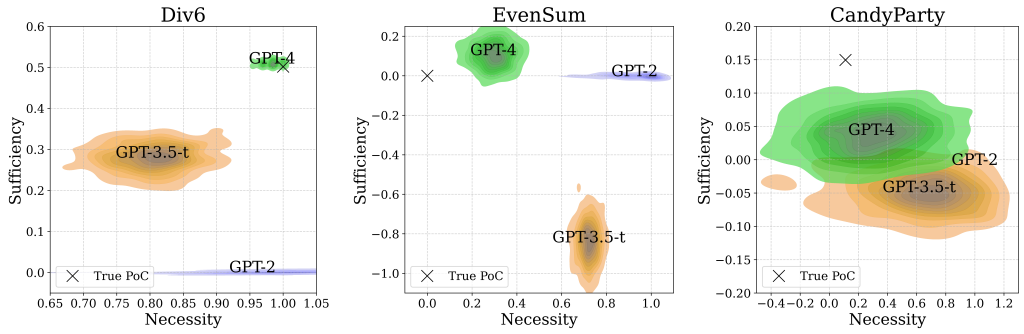


Figure 6: True PN and PS vs. inferred PN and PS using GPT-2, GPT-35-turbo and GPT-4. The densities of the estimated probabilities capture the uncertainty associated with the responses by each model.

4.1 Factual vs. counterfactual predictions

Figure 5 (*Left*) illustrates the alignment between the outputs of GPT-2, GPT-3.5-turbo, and GPT-4, and the factual predictions and counterfactuals for the `Div6` problem. The shading within each cell of the heatmap indicates the degree of mismatch between model-generated outputs and the true information, with the colour intensity reflecting the level of disagreement based on the 10 answers from the models. As highlighted in Figure1—where the average disagreement across the first 100 columns of these heatmaps informs the results—more sophisticated models like GPT-4 demonstrate a closer match with the counterfactuals derived from the true reasoning graph. For similar comparisons involving other problems, please refer to Appendix I.

One might wonder if the evaluation of reasoning truly requires PN and PS, or if it could be sufficiently assessed by examining only the inconsistency rates in factual/counterfactual data. Figure 5 (*Right*) underscores the importance of PN and PS. It presents the estimated distributions of PN for the `Div6` problem, based on 500 replicates under five scenarios where true counterfactuals are randomly altered with probabilities 0.005, 0.001, 0.05, 0.1 and 0.2. As we might anticipate, the greater the deviation from a dataset free of counterfactual errors, the more significant the discrepancy from the actual PN $= 1$ for this example. Notably, even minor perturbations can lead to substantial shifts in the estimated PN. For example, with a 0.05 probability of counterfactual perturbation, the estimated PN varies between 0.5 and 0.9. This suggests that relying solely on counterfactual errors could lead to an overestimation of the models' reasoning abilities, particularly their understanding of the necessary and sufficient conditions within a problem. Furthermore, a counterfactual error rate of 0.2 in this example results in entirely inconsistent (negative) probabilities due to the mismatch between the conditional and interventional distributions, as defined in Eq. 1.

4.2 Evaluation of LLMs reasoning

We computed the CIR, FIR, γ-PN-overlap, and γ-PS-overlap for the problems `Div6`, `EvenSum` and `CandyParty` using GPT-2, GPT-3.5-turbo, and GPT-4.

Figure 6 illustrates the estimated PN and PS for each problem, obtained through bootstrap resampling. Each density is labeled with the model that was used used to generate the data for those results. The true values of the PS and PN in each problem is marked with a cross. A model is considered capable of reasoning if the PN-PS density estimates overlap with the true probabilities of causation. Such an overlap was only achieved by GPT-4 for `Div6` problem. Other results varied, indicating generally weak reasoning abilities. Negative values of PN and PS in several instances, are due to inconsistencies in $\mathcal{D}_F^{LLM}$ and $\mathcal{D}_{CF}^{LLM}$ as detailed in Section 4.1.

Figure 7 (*Left*, *Centre*) features the γ-PN-overlap and γ-PN-overlap curves for all models and problems, where ideal reasoning corresponds to the metrics equalling one for any value of γ. GPT-4 shows this level of reasoning for the `Div6` problem. However, GPT-2 had an accurate PN for `Even-Sum`, but the PS estimates were notably less accurate.

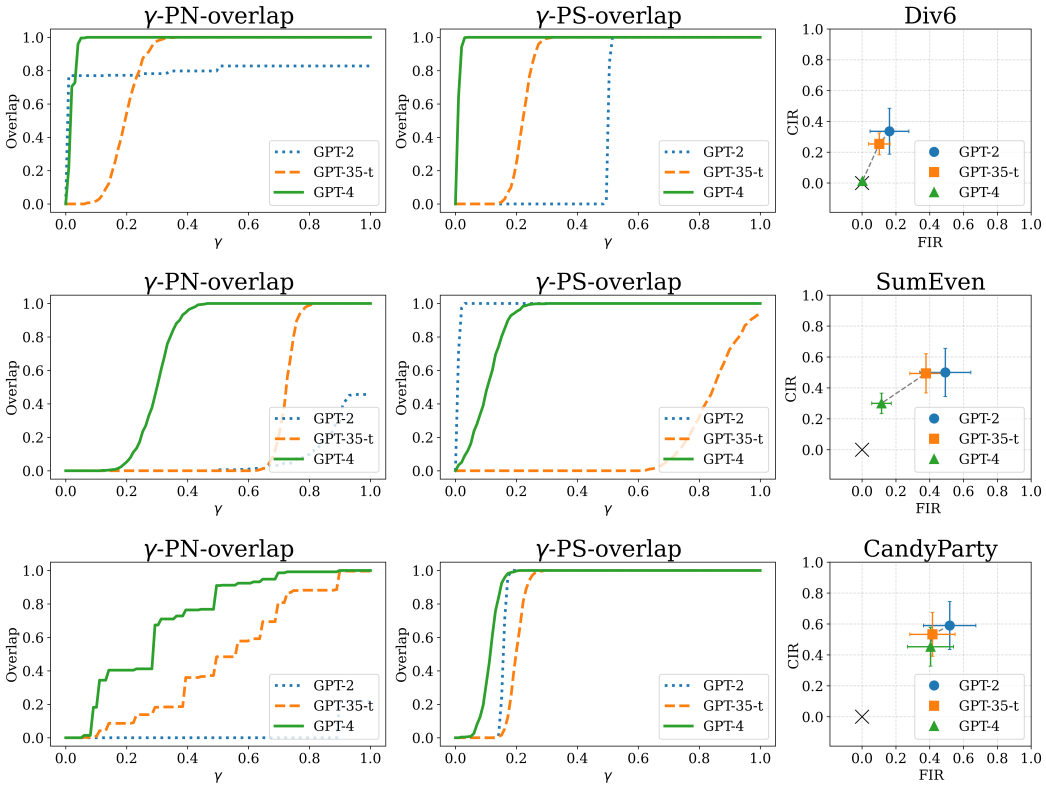


Figure 7: *Left*, *Centre*: Reconstruction of the γ-PN-overlap and γ-PS-overlap curves for GPT-2, GPT-35-turbo and GPT-4. Ideal reasoning is achieved when the overlap is one for all values of γ. *Right*: Visualization of FIR and PIR. Ideal reasoning is attained when both metrics are zero (denoted by a $\times$).

Figure 7 (*Right*) presents the values of CIR and PIR (with the standard deviations included brackets). An emerging trend towards reasoning is observed in the GPT family of models, particularly seen with GPT-4 for the `Div6` problem. An intriguing question is whether future versions of these models will similarly approach the true PN and PS for other problems as well.

5 Discussion

The primary objective of this paper was to explore and understand the reasoning abilities of LLMs, which is essential for their successful deployment in a range of applications. Given the growing dependence on LLMs for complex reasoning tasks, such as mathematics, programming, or strategic planning, understanding this is crucial. To evaluate these reasoning abilities, we introduced a novel framework that employs probabilistic measures of necessity and sufficiency, and find that while various models (GPT-2, GPT-3.5-turbo, and GPT-4) can replicate aspects of reasoning to some degree, they often falter when it comes to counterfactual reasoning. What makes our approach unique is that we test the models with scenarios where generalization by reasoning, rather than replicating statistical patterns in the training data, is required to provide correct answers. Notably, the ability to reason, as defined in this paper, does improve with more complex models, yet it is still far from flawless. This observation leads to the question of whether future versions of these models will achieve perfect reasoning. Our results are significant, as they reveal the limitations of LLMs, and emphasize the need for further research to enhance their reasoning capabilities.

In general, reasoning goes beyond the math examples that we have included in the paper. We believe that the same theory and tools can be used in other domains. The Hex framework, that serves as a mathematical framework to formalize our ideas, requires the definition of a query, state, and an abstract execution machine that is used to predict how the state is modified given the query. In this

framework, for instance, one could think about problems in vision, where the elements of the state are the objects in an image, and the query corresponds to a counterfactual query that describes an intervention in the environment. The concepts of necessity and sufficiency still apply in this scenario. In an image where an object is removed or altered, the framework can help determine the impact of this change on a property of the overall scene. We believe that this approach and will be key in other fields such as robotics, and or social sciences, where understanding the necessity and sufficiency between different elements is crucial for accurate reasoning and decision-making.

Limitations: Our approach has several limitations that we acknowledge, but did not address within the scope of this research.

1. *Dependence on reasoning graphs*: our method requires access to reasoning graphs. This requirement may hinder our ability to fully understand the reasoning abilities of LLMs in situations where it is challenging to derive relationships, including causal ones.
2. *Boolean variable restriction*: our method is designed to work with boolean valued variables, which is restrictive, particularly for cases involving multiple states or conditions occurring at the same time. However, we believe that this issue can be addresss with further research.
3. *Prompt-dependent results*: The findings we report are based on an LLM's reasoning abilities as determined by two specific types (factual/counterfacutal) of prompts that we used. Of course, other techniques like chain-of-thought prompting could be used, but our focus remains on establishing a consistent baseline. Future work could explore the impact of different prompting strategies on reasoning performance, potentially leading to more refined and effective methods for evaluating and enhancing reasoning capabilities in LLMs. Our experiments did not aim to fine-tune these prompts or to 'optimise reasoning'—a separate area of ongoing research. Instead, our goal was to offer valuable insights that can aid the community in developing new benchmarks and employing LLMs responsibly.

Broader impact: Evaluating the reasoning capabilities of LLMs is essential as it significantly influences their effectiveness in various domains. In education and research, it is important for the model to be able to provide accurate explanations and to formulate meaningful hypotheses. In the commercial sector, the effectiveness of automated processes/systems relies heavily on how well the model can reason. When it comes to accessibility, the model must be able to understand and meet diverse user needs, which hinges on its reasoning ability. Moreover, identifying and mitigating biases in AI systems—a key aspect of ethical and equitable AI—requires a detailed examination of the models' reasoning processes. Therefore, while LLMs hold immense promise, ensuring their responsible and beneficial use is predicated on a thorough appraisal of their reasoning abilities. We believe that our research is an important step in this direction.

Acknowledgments

We would like to thank Ted Meeds, Alicia Curth and Sushrut Karmalkar for their invaluable discussions and feedback throughout the process of writing this paper.

References

[1] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin. Large language models for mathematical reasoning: Progresses and challenges, 2024.

[2] A. Aliseda. *Abductive Reasoning: Logical Investigations into Discovery and Explanation*. Spring Publications, 2006.

[3] S. Ashwani, K. Hegde, N. R. Mannuru, M. Jindal, D. S. Sengar, K. C. R. Kathala, D. Banga, V. Jain, and A. Chadha. Cause and effect: Can large language models truly understand causality?, 2024.

[4] C. Bowen, R. Sætre, and Y. Miyao. A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, page 323–339, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. A. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y.-F. Li, S. M. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *ArXiv*, abs/2303.12712, 2023.

[6] Z. Chen, Q. Gao, A. Bosselut, A. Sabharwal, and K. Richardson. DISCO: Distilling counterfactuals with large language models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] A. Feeney and E. Heit, editors. *Inductive Reasoning: Experimental, Developmental, and Computational Approaches*. Cambridge University Press, 2007.

[8] W. Gan, Z. Qi, J. Wu, and J. C.-W. Lin. Large language models in education: Vision and opportunities. *arXiv preprint arXiv:2311.13160*, 2023.

[9] J. González and A. V. Nori. Beyond words: A mathematical framework for interpreting large language models. *arXiv preprint arXiv:2311.03033*, 2023.

[10] J. González, C. Wong, Z. Gero, J. Bagga, R. Ueno, I. Chien, E. Oravkin, E. Kiciman, A. Nori, R. Weerasinghe, R. S. Leidner, B. Piening, T. Naumann, C. Bifulco, and H. Poon. Trialscope: A unifying causal framework for scaling real-world evidence generation with biomedical language models, 2023.

[11] S. Gunasekar, Y. Zhang, J. Aneja, C. Cesar, T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. Singh Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, and Y. Li. Textbooks are all you need, June 2023.

[12] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. part i: Causes. *British Journal of Philosophy of Science*, 56(4):843–887, 2005.

[13] S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore, December 2023. Association for Computational Linguistics.

[14] J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey, 2023.

[15] J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics.

[16] Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. Can large language models infer causation from correlation?, 2024.

[17] D. Kelley. *The Art of Reasoning (with symbolic logic)*. Unknown Publisher, 1990.

[18] E. Kıcıman, R. Ness, A. Sharma, and C. Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2023.

[19] D. Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973.

[20] J. Li, L. Yu, and A. Ettinger. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 804–815, Toronto, Canada, July 2023. Association for Computational Linguistics.

[21] Y. Li, M. Xu, X. Miao, S. Zhou, and T. Qian. Large language models as counterfactual generator: Strengths and weaknesses, 2023.

[22] J. L. Mackie. Causes and conditions. *American Philosophical Quarterly*, 2(4):245–264, 1965.

[23] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[24] J. Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121(1-2):93–149, 1999.

[25] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[26] J. Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[27] L. J. Rips. *The Psychology of Proof: Deductive Reasoning in Human Thinking*. The MIT Press, 1994.

[28] S. M. Seals and V. L. Shalin. Evaluating the deductive competence of large language models. *arXiv preprint arXiv:2309.05452*, 2023.

[29] P. F. Simmering and P. Huoviala. Large language models for aspect-based sentiment analysis. *arXiv preprint arXiv:2310.18025*, 2023.

[30] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T. Lillicrap, A. Lazaridou, O. Firat, J. Molloy, M. Isard, P. R. Barham, T. Hennigan, B. Lee, F. Viola, M. Reynolds, Y. Xu, R. Doherty, E. Collins, C. Meyer, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, J. Krawczyk, C. Du, E. Chi, H.-T. Cheng, E. Ni, P. Shah, P. Kane, B. Chan, M. Faruqui, A. Severyn, H. Lin, Y. Li, Y. Cheng, A. Ittycheriah, M. Mahdieh, M. Chen, P. Sun, D. Tran, S. Bagri, B. Lakshminarayanan, J. Liu, A. Orban, F. Güra, H. Zhou, X. Song, A. Boffy, H. Ganapathy, S. Zheng, H. Choe, Ágoston Weisz, T. Zhu, Y. Lu, S. Gopal, J. Kahn, M. Kula, J. Pitman, R. Shah, E. Taropa, M. A. Merey, M. Baeuml, Z. Chen, L. E. Shafey, Y. Zhang, O. Sercinoglu, G. Tucker, E. Piqueras, M. Krikun, I. Barr, N. Savinov, I. Danihelka, B. Roelofs, A. White, A. Andreassen, T. von Glehn, L. Yagati, M. Kazemi, L. Gonzalez, M. Khalman, J. Sygnowski, A. Frechette, C. Smith, L. Culp, L. Proleev, Y. Luan, X. Chen, J. Lottes, N. Schucher, F. Lebron, A. Rrustemi, N. Clay, P. Crone, T. Kocisky, J. Zhao, B. Perz, D. Yu, H. Howard, A. Bloniarz, J. W. Rae, H. Lu, L. Sifre, M. Maggioni, F. Alcober, D. Garrette, M. Barnes, S. Thakoor, J. Austin, G. Barth-Maron, W. Wong, R. Joshi, R. Chaabouni, D. Fatiha, A. Ahuja, G. S. Tomar, E. Senter, M. Chadwick, I. Kornakov, N. Attaluri, I. Iturrate, R. Liu, Y. Li, S. Cogan, J. Chen, C. Jia, C. Gu, Q. Zhang, J. Grimstad, A. J. Hartman, X. Garcia, T. S. Pillai, J. Devlin, M. Laskin, D. de Las Casas, D. Valter, C. Tao, L. Blanco, A. P. Badia, D. Reitter, M. Chen, J. Brennan, C. Rivera, S. Brin, S. Iqbal, G. Surita, J. Labanowski, A. Rao, S. Winkler, E. Parisotto, Y. Gu, K. Olszewska, R. Addanki, A. Miech, A. Louis, D. Teplyashin, G. Brown, E. Catt, J. Balaguer, J. Xiang, P. Wang, Z. Ashwood, A. Briukhov, A. Webson, S. Ganapathy, S. Sanghavi, A. Kannan, M.-W. Chang, A. Stjerngren, J. Djolonga, Y. Sun, A. Bapna, M. Aitchison, P. Pejman, H. Michalewski, T. Yu, C. Wang, J. Love, J. Ahn, D. Bloxwich, K. Han, P. Humphreys, T. Sellam, J. Bradbury, V. Godbole, S. Samangooei, B. Damoc, A. Kaskasoli, S. M. R. Arnold, V. Vasudevan, S. Agrawal, J. Riesa, D. Lepikhin, R. Tanburn, S. Srinivasan, H. Lim, S. Hodkinson, P. Shyam, J. Ferret, S. Hand, A. Garg, T. L. Paine, J. Li, Y. Li, M. Giang, A. Neitz, Z. Abbas, S. York, M. Reid, E. Cole, A. Chowdhery, D. Das, D. Rogozińska, V. Nikolaev, P. Sprechmann, Z. Nado, L. Zilka, F. Prost, L. He, M. Monteiro, G. Mishra, C. Welty, J. Newlan, D. Jia, M. Allamanis, C. H. Hu, R. de Liedekerke, J. Gilmer, C. Saroufim, S. Rijhwani, S. Hou, D. Shrivastava, A. Baddepudi, A. Goldin, A. Ozturel, A. Cassirer, Y. Xu, D. Sohn, D. Sachan, R. K. Amplayo, C. Swanson, D. Petrova, S. Narayan, A. Guez, S. Brahma, J. Landon, M. Patel, R. Zhao, K. Villela, L. Wang, W. Jia, M. Rahtz, M. Giménez, L. Yeung, J. Keeling, P. Georgiev, D. Mincu, B. Wu, S. Haykal, R. Saputro, K. Vodrahalli, J. Qin, Z. Cankara, A. Sharma, N. Fernando, W. Hawkins, B. Neyshabur, S. Kim, A. Hutter, P. Agrawal, A. Castro-Ros, G. van den Driessche, T. Wang, F. Yang, S. yiin Chang, P. Komarek, R. McIlroy, M. Lučić, G. Zhang, W. Farhan, M. Sharman, P. Natsev, P. Michel, Y. Bansal, S. Qiao, K. Cao, S. Shakeri, C. Butterfield, J. Chung, P. K. Rubenstein, S. Agrawal, A. Mensch, K. Soparkar, K. Lenc, T. Chung, A. Pope, L. Maggiore, J. Kay, P. Jhakra, S. Wang, J. Maynez, M. Phuong, T. Tobin, A. Tacchetti, M. Trebacz, K. Robinson, Y. Katariya, S. Riedel, P. Bailey, K. Xiao, N. Ghelani, L. Aroyo, A. Slone, N. Houlsby, X. Xiong, Z. Yang, E. Gribovskaya, J. Adler, M. Wirth, L. Lee, M. Li, T. Kagohara, J. Pavagadhi, S. Bridgers, A. Bortsova, S. Ghemawat, Z. Ahmed, T. Liu, R. Powell, V. Bolina, M. Iinuma, P. Zablotskaia, J. Besley, D.-W. Chung, T. Dozat, R. Comanescu, X. Si, J. Greer, G. Su, M. Polacek, R. L. Kaufman, S. Tokumine, H. Hu, E. Buchatskaya, Y. Miao, M. Elhawaty, A. Siddhant, N. Tomasev, J. Xing, C. Greer, H. Miller, S. Ashraf, A. Roy, Z. Zhang, A. Ma, A. Filos, M. Besta, R. Blevins, T. Klimenko, C.-K. Yeh, S. Changpinyo, J. Mu, O. Chang, M. Pajarskas, C. Muir, V. Cohen, C. L. Lan, K. Haridasan, A. Marathe, S. Hansen, S. Douglas, R. Samuel, M. Wang, S. Austin, C. Lan, J. Jiang, J. Chiu, J. A. Lorenzo, L. L. Sjösund, S. Cevey, Z. Gleicher, T. Avrahami, A. Boral, H. Srinivasan, V. Selo, R. May, K. Aisopos, L. Hussenot, L. B. Soares, K. Baumli, M. B. Chang, A. Recasens, B. Caine, A. Pritzel, F. Pavetic, F. Pardo, A. Gergely, J. Frye, V. Ramasesh, D. Horgan, K. Badola, N. Kassner, S. Roy, E. Dyer, V. C. Campos, A. Tomala, Y. Tang, D. E. Badawy, E. White, B. Mustafa, O. Lang, A. Jindal, S. Vikram, Z. Gong, S. Caelles, R. Hemsley, G. Thornton, F. Feng, W. Stokowiec, C. Zheng, P. Thacker, Çağlar Ünlü, Z. Zhang, M. Saleh, J. Svensson, M. Bileschi, P. Patil, A. Anand, R. Ring, K. Tsihlas, A. Vezer, M. Selvi, T. Shevlane,

M. Rodriguez, T. Kwiatkowski, S. Daruki, K. Rong, A. Dafoe, N. FitzGerald, K. Gu-Lemberg, M. Khan, L. A. Hendricks, M. Pellat, V. Feinberg, J. Cobon-Kerr, T. Sainath, M. Rauh, S. H. Hashemi, R. Ives, Y. Hasson, E. Noland, Y. Cao, N. Byrd, L. Hou, Q. Wang, T. Sottiaux, M. Paganini, J.-B. Lespiau, A. Moufarek, S. Hassan, K. Shivakumar, J. van Amersfoort, A. Mandhane, P. Joshi, A. Goyal, M. Tung, A. Brock, H. Sheahan, V. Misra, C. Li, N. Rakićević, M. Dehghani, F. Liu, S. Mittal, J. Oh, S. Noury, E. Sezener, F. Huot, M. Lamm, N. D. Cao, C. Chen, S. Mudgal, R. Stella, K. Brooks, G. Vasudevan, C. Liu, M. Chain, N. Melinkeri, A. Cohen, V. Wang, K. Seymore, S. Zubkov, R. Goel, S. Yue, S. Krishnakumaran, B. Albert, N. Hurley, M. Sano, A. Mohananey, J. Joughin, E. Filonov, T. Kępa, Y. Eldawy, J. Lim, R. Rishi, S. Badiezadegan, T. Bos, J. Chang, S. Jain, S. G. S. Padmanabhan, S. Puttagunta, K. Krishna, L. Baker, N. Kalb, V. Bedapudi, A. Kurzrok, S. Lei, A. Yu, O. Litvin, X. Zhou, Z. Wu, S. Sobell, A. Siciliano, A. Papir, R. Neale, J. Bragagnolo, T. Toor, T. Chen, V. Anklin, F. Wang, R. Feng, M. Gholami, K. Ling, L. Liu, J. Walter, H. Moghaddam, A. Kishore, J. Adamek, T. Mercado, J. Mallinson, S. Wandekar, S. Cagle, E. Ofek, G. Garrido, C. Lombriser, M. Mukha, B. Sun, H. R. Mohammad, J. Matak, Y. Qian, V. Peswani, P. Janus, Q. Yuan, L. Schelin, O. David, A. Garg, Y. He, O. Duzhyi, A. Älgmyr, T. Lottaz, Q. Li, V. Yadav, L. Xu, A. Chinien, R. Shivanna, A. Chuklin, J. Li, C. Spadine, T. Wolfe, K. Mohamed, S. Das, Z. Dai, K. He, D. von Dincklage, S. Upadhyay, A. Maurya, L. Chi, S. Krause, K. Salama, P. G. Rabinovitch, P. K. R. M, A. Selvan, M. Dektiarev, G. Ghiasi, E. Guven, H. Gupta, B. Liu, D. Sharma, I. H. Shtacher, S. Paul, O. Akerlund, F.-X. Aubet, T. Huang, C. Zhu, E. Zhu, E. Teixeira, M. Fritze, F. Bertolini, L.-E. Marinescu, M. Bölle, D. Paulus, K. Gupta, T. Latkar, M. Chang, J. Sanders, R. Wilson, X. Wu, Y.-X. Tan, L. N. Thiet, T. Doshi, S. Lall, S. Mishra, W. Chen, T. Luong, S. Benjamin, J. Lee, E. Andrejczuk, D. Rabiej, V. Ranjan, K. Styrc, P. Yin, J. Simon, M. R. Harriott, M. Bansal, A. Robsky, G. Bacon, D. Greene, D. Mirylenka, C. Zhou, O. Sarvana, A. Goyal, S. Andermatt, P. Siegler, B. Horn, A. Israel, F. Pongetti, C.-W. L. Chen, M. Selvatici, P. Silva, K. Wang, J. Tolins, K. Guu, R. Yogev, X. Cai, A. Agostini, M. Shah, H. Nguyen, N. O. Donnaile, S. Pereira, L. Friso, A. Stambler, A. Kurzrok, C. Kuang, Y. Romanikhin, M. Geller, Z. Yan, K. Jang, C.-C. Lee, W. Fica, E. Malmi, Q. Tan, D. Banica, D. Balle, R. Pham, Y. Huang, D. Avram, H. Shi, J. Singh, C. Hidey, N. Ahuja, P. Saxena, D. Dooley, S. P. Potharaju, E. O'Neill, A. Gokulchandran, R. Foley, K. Zhao, M. Dusenberry, Y. Liu, P. Mehta, R. Kotikalapudi, C. Safranek-Shrader, A. Goodman, J. Kessinger, E. Globen, P. Kolhar, C. Gorgolewski, A. Ibrahim, Y. Song, A. Eichenbaum, T. Brovelli, S. Potluri, P. Lahoti, C. Baetu, A. Ghorbani, C. Chen, A. Crawford, S. Pal, M. Sridhar, P. Gurita, A. Mujika, I. Petrovski, P.-L. Cedoz, C. Li, S. Chen, N. D. Santo, S. Goyal, J. Punjabi, K. Kappaganthu, C. Kwak, P. LV, S. Velury, H. Choudhury, J. Hall, P. Shah, R. Figueira, M. Thomas, M. Lu, T. Zhou, C. Kumar, T. Jurdi, S. Chikkerur, Y. Ma, A. Yu, S. Kwak, V. Ähdel, S. Rajayogam, T. Choma, F. Liu, A. Barua, C. Ji, J. H. Park, V. Hellendoorn, A. Bailey, T. Bilal, H. Zhou, M. Khatir, C. Sutton, W. Rzadkowski, F. Macintosh, K. Shagin, P. Medina, C. Liang, J. Zhou, P. Shah, Y. Bi, A. Dankovics, S. Banga, S. Lehmann, M. Bredesen, Z. Lin, J. E. Hoffmann, J. Lai, R. Chung, K. Yang, N. Balani, A. Bražinskas, A. Sozanschi, M. Hayes, H. F. Alcalde, P. Makarov, W. Chen, A. Stella, L. Snijders, M. Mandl, A. Kärrman, P. Nowak, X. Wu, A. Dyck, K. Vaidyanathan, R. R, J. Mallet, M. Rudominer, E. Johnston, S. Mittal, A. Udathu, J. Christensen, V. Verma, Z. Irving, A. Santucci, G. Elsayed, E. Davoodi, M. Georgiev, I. Tenney, N. Hua, G. Cideron, E. Leurent, M. Alnahlawi, I. Georgescu, N. Wei, I. Zheng, D. Scandinaro, H. Jiang, J. Snoek, M. Sundararajan, X. Wang, Z. Ontiveros, I. Karo, J. Cole, V. Rajashekhar, L. Tumeh, E. Ben-David, R. Jain, J. Uesato, R. Datta, O. Bunyan, S. Wu, J. Zhang, P. Stanczyk, Y. Zhang, D. Steiner, S. Naskar, M. Azzam, M. Johnson, A. Paszke, C.-C. Chiu, J. S. Elias, A. Mohiuddin, F. Muhammad, J. Miao, A. Lee, N. Vieillard, J. Park, J. Zhang, J. Stanway, D. Garmon, A. Karmarkar, Z. Dong, J. Lee, A. Kumar, L. Zhou, J. Evens, W. Isaac, G. Irving, E. Loper, M. Fink, I. Arkatkar, N. Chen, I. Shafran, I. Petrychenko, Z. Chen, J. Jia, A. Levskaya, Z. Zhu, P. Grabowski, Y. Mao, A. Magni, K. Yao, J. Snaider, N. Casagrande, E. Palmer, P. Suganthan, A. Castaño, I. Giannoumis, W. Kim, M. Rybiński, A. Sreevatsa, J. Prendki, D. Soergel, A. Goedeckemeyer, W. Gierke, M. Jafari, M. Gaba, J. Wiesner, D. G. Wright, Y. Wei, H. Vashisht, Y. Kulizhskaya, J. Hoover, M. Le, L. Li, C. Iwuanyanwu, L. Liu, K. Ramirez, A. Khorlin, A. Cui, T. LIN, M. Wu, R. Aguilar, K. Pallo, A. Chakladar, G. Perng, E. A. Abellan, M. Zhang, I. Dasgupta, N. Kushman, I. Penchev, A. Repina, X. Wu, T. van der Weide, P. Ponnapalli, C. Kaplan, J. Simsa, S. Li, O. Dousse, F. Yang, J. Piper, N. Ie, R. Pasumarthi, N. Lintz, A. Vijayakumar, D. Andor, P. Valenzuela, M. Lui, C. Paduraru, D. Peng, K. Lee, S. Zhang, S. Greene, D. D. Nguyen, P. Kurylowicz, C. Hardin, L. Dixon, L. Janzer, K. Choo, Z. Feng, B. Zhang, A. Singhal, D. Du,

D. McKinnon, N. Antropova, T. Bolukbasi, O. Keller, D. Reid, D. Finchelstein, M. A. Raad, R. Crocker, P. Hawkins, R. Dadashi, C. Gaffney, K. Franko, A. Bulanova, R. Leblond, S. Chung, H. Askham, L. C. Cobo, K. Xu, F. Fischer, J. Xu, C. Sorokin, C. Alberti, C.-C. Lin, C. Evans, A. Dimitriev, H. Forbes, D. Banarse, Z. Tung, M. Omernick, C. Bishop, R. Sterneck, R. Jain, J. Xia, E. Amid, F. Piccinno, X. Wang, P. Banzal, D. J. Mankowitz, A. Polozov, V. Krakovna, S. Brown, M. Bateni, D. Duan, V. Firoiu, M. Thotakuri, T. Natan, M. Geist, S. tan Girgin, H. Li, J. Ye, O. Roval, R. Tojo, M. Kwong, J. Lee-Thorp, C. Yew, D. Sinopalnikov, S. Ramos, J. Mellor, A. Sharma, K. Wu, D. Miller, N. Sonnerat, D. Vnukov, R. Greig, J. Beattie, E. Caveness, L. Bai, J. Eisenschlos, A. Korchemniy, T. Tsai, M. Jasarevic, W. Kong, P. Dao, Z. Zheng, F. Liu, F. Yang, R. Zhu, T. H. Teh, J. Sanmiya, E. Gladchenko, N. Trdin, D. Toyama, E. Rosen, S. Tavakkol, L. Xue, C. Elkind, O. Woodman, J. Carpenter, G. Papamakarios, R. Kemp, S. Kafle, T. Grunina, R. Sinha, A. Talbert, D. Wu, D. Owusu-Afriyie, C. Du, C. Thornton, J. Pont-Tuset, P. Narayana, J. Li, S. Fatehi, J. Wieting, O. Ajmeri, B. Uria, Y. Ko, L. Knight, A. Héliou, N. Niu, S. Gu, C. Pang, Y. Li, N. Levine, A. Stolovich, R. Santamaria-Fernandez, S. Goenka, W. Yustalim, R. Strudel, A. Elqursh, C. Deck, H. Lee, Z. Li, K. Levin, R. Hoffmann, D. Holtmann-Rice, O. Bachem, S. Arora, C. Koh, S. H. Yeganeh, S. Põder, M. Tariq, Y. Sun, L. Ionita, M. Seyedhosseini, P. Tafti, Z. Liu, A. Gulati, J. Liu, X. Ye, B. Chrzaszcz, L. Wang, N. Sethi, T. Li, B. Brown, S. Singh, W. Fan, A. Parisi, J. Stanton, V. Koverkathu, C. A. Choquette-Choo, Y. Li, T. Lu, A. Ittycheriah, P. Shroff, M. Varadarajan, S. Bahargam, R. Willoughby, D. Gaddy, G. Desjardins, M. Cornero, B. Robenek, B. Mittal, B. Albrecht, A. Shenoy, F. Moiseev, H. Jacobsson, A. Ghaffarkhah, M. Rivière, A. Walton, C. Crepy, A. Parrish, Z. Zhou, C. Farabet, C. Radebaugh, P. Srinivasan, C. van der Salm, A. Fidjeland, S. Scellato, E. Latorre-Chimoto, H. Klimczak-Plucińska, D. Bridson, D. de Cesare, T. Hudson, P. Mendolicchio, L. Walker, A. Morris, M. Mauger, A. Guseynov, A. Reid, S. Odoom, L. Loher, V. Cotruta, M. Yenugula, D. Grewe, A. Petrushkina, T. Duerig, A. Sanchez, S. Yadlowsky, A. Shen, A. Globerson, L. Webb, S. Dua, D. Li, S. Bhupatiraju, D. Hurt, H. Qureshi, A. Agarwal, T. Shani, M. Eyal, A. Khare, S. R. Belle, L. Wang, C. Tekur, M. S. Kale, J. Wei, R. Sang, B. Saeta, T. Liechty, Y. Sun, Y. Zhao, S. Lee, P. Nayak, D. Fritz, M. R. Vuyyuru, J. Aslanides, N. Vyas, M. Wicke, X. Ma, E. Eltyshev, N. Martin, H. Cate, J. Manyika, K. Amiri, Y. Kim, X. Xiong, K. Kang, F. Luisier, N. Tripuraneni, D. Madras, M. Guo, A. Waters, O. Wang, J. Ainslie, J. Baldridge, H. Zhang, G. Pruthi, J. Bauer, F. Yang, R. Mansour, J. Gelman, Y. Xu, G. Polovets, J. Liu, H. Cai, W. Chen, X. Sheng, E. Xue, S. Ozair, C. Angermueller, X. Li, A. Sinha, W. Wang, J. Wiesinger, E. Koukoumidis, Y. Tian, A. Iyer, M. Gurumurthy, M. Goldenson, P. Shah, M. Blake, H. Yu, A. Urbanowicz, J. Palomaki, C. Fernando, K. Durden, H. Mehta, N. Momchev, E. Rahimtoroghi, M. Georgaki, A. Raul, S. Ruder, M. Redshaw, J. Lee, D. Zhou, K. Jalan, D. Li, B. Hechtman, P. Schuh, M. Nasr, K. Milan, V. Mikulik, J. Franco, T. Green, N. Nguyen, J. Kelley, A. Mahendru, A. Hu, J. Howland, B. Vargas, J. Hui, K. Bansal, V. Rao, R. Ghiya, E. Wang, K. Ye, J. M. Sarr, M. M. Preston, M. Elish, S. Li, A. Kaku, J. Gupta, I. Pasupat, D.-C. Juan, M. Someswar, T. M., X. Chen, A. Amini, A. Fabrikant, E. Chu, X. Dong, A. Muthal, S. Buthpitiya, S. Jauhari, N. Hua, U. Khandelwal, A. Hitron, J. Ren, L. Rinaldi, S. Drath, A. Dabush, N.-J. Jiang, H. Godhia, U. Sachs, A. Chen, Y. Fan, H. Taitelbaum, H. Noga, Z. Dai, J. Wang, C. Liang, J. Hamer, C.-S. Ferng, C. Elkind, A. Atias, P. Lee, V. Listík, M. Carlen, J. van de Kerkhof, M. Pikus, K. Zaher, P. Müller, S. Zykova, R. Stefanec, V. Gatsko, C. Hirnschall, A. Sethi, X. F. Xu, C. Ahuja, B. Tsai, A. Stefanoiu, B. Feng, K. Dhandhania, M. Katyal, A. Gupta, A. Parulekar, D. Pitta, J. Zhao, V. Bhatia, Y. Bhavnani, O. Alhadlaq, X. Li, P. Danenberg, D. Tu, A. Pine, V. Filippova, A. Ghosh, B. Limonchik, B. Urala, C. K. Lanka, D. Clive, Y. Sun, E. Li, H. Wu, K. Hongtongsak, I. Li, K. Thakkar, K. Omarov, K. Majmundar, M. Alverson, M. Kucharski, M. Patel, M. Jain, M. Zabelin, P. Pelagatti, R. Kohli, S. Kumar, J. Kim, S. Sankar, V. Shah, L. Ramachandruni, X. Zeng, B. Bariach, L. Weidinger, A. Subramanya, S. Hsiao, D. Hassabis, K. Kavukcuoglu, A. Sadovsky, Q. Le, T. Strohman, Y. Wu, S. Petrov, J. Dean, and O. Vinyals. Gemini: A family of highly capable multimodal models, 2024.

[31] J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

[32] Y. Tong, D. Li, S. Wang, Y. Wang, F. Teng, and J. Shang. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*, 2024.

[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[34] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[35] C. Wong, S. Zhang, Y. Gu, C. Moung, J. Abel, N. Usuyama, R. Weerasinghe, B. Piening, T. Naumann, C. Bifulco, and H. Poon. Scaling clinical trial matching using large language models: A case study in oncology, 2023.

[36] C. Zhang, S. Bauer, P. Bennett, J. Gao, W. Gong, A. Hilmkil, J. Jennings, C. Ma, T. Minka, N. Pawlowski, and J. Vaughan. Understanding causality with large language models: Feasibility and opportunities, 2023.

Appendix for: 'Does Reasoning Emerge? Examining the Probabilities of Causation in Large Language Models'

A Causal models, counterfactuals and probabilities of causation

This section provides technical background on structural causal models, counterfactuals and probabilities of causation [25].

Definition 4 (Causal Model)**.** *A causal model $\mathcal{M}$ is a triple $\langle \mathcal{V}, \mathcal{F}, \varepsilon \rangle$ where:*

1. *$\mathcal{V} = \{V_1, \ldots, V_n\}$ is a set of endogenous variables.*
2. *$\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_n\}$ is a set of exogenous variables. The exogenous variables ε are assumed to be independent of each other and represent the unobserved factors that influence the values of V.*
3. *$\mathcal{F} = \{f_1, \ldots, f_n\}$ is a set of functions. Each function f_i determines the value of V_i as a function of its parents $\text{PA}_i \subseteq \mathcal{V} \cup \varepsilon$, where PA_i are the variables that directly cause V_i.*

Any causal model can be represented by a directed acyclic graph (DAG) $\mathcal{G}$, where the nodes represent the variables $\mathcal{V}$, and the edges are the direct causal relationships between these variables. Let X be a subset of variables in $\mathcal{V}$, and x be a specific realization of the values these variables can take. We define a submodel $\mathcal{M}_{X=x}$ to be a causal model $\langle \mathcal{V}, \mathcal{F}_t, \epsilon \rangle$, where $\mathcal{F}_t = \{f_i : V_i \notin T\} \cup \{X = x\}$.

Definition 5 (Intervention, do operator)**.** *Consider a causal model $\mathcal{M} = \langle \mathcal{V}, \mathcal{F}, \varepsilon \rangle$, with X being a subset of variables in $\mathcal{X}$, and x a particular realization of T. The effect of the intervention $do(X = x)$ in $\mathcal{M}$ is given by the submodel $\mathcal{M}_{X=x}$.*

Definition 6 (Potential outcome and counterfactual)**.** *Let Y be a variable in $\mathcal{V}$, and let X be a subset of $\mathcal{V}$. The potential outcome of Y resulting from the intervention $do(X = x)$, denoted by $Y_{X=x}(\epsilon) = y$, is the solution for Y in the set of equations $\mathcal{F}_t$. A counterfactual is defined as the potential outcome $Y_{X=x}(\epsilon)$ for the hypothetical scenario "what would the value that Y have been if X had been set to x".*

A distribution $\mathbb{P}$ over the exogenous variables ε establishes a corresponding a probability distribution over the endogenous variables $\mathcal{X}$ as well as the potential outcomes. In practical applications, $\mathbb{P}(\epsilon)$ characterizes the target population of the study. The probability of a counterfactual $Y_{X=x'}$ induced by the submodel $\mathcal{M}_{X=x}$ is:

$$\mathbb{P}(Y_{X=x} = y) = \sum_{\{\epsilon | Y_{X=x}(\epsilon) = y\}} \mathbb{P}(\epsilon)$$

In addition, probabilities of the type $\mathbb{P}(Y_{X=x'} | X = x, Y = y)$ can be computed as

$$\mathbb{P}(Y_{X=x'} = y' | X = x, Y = y) = \sum_{\epsilon} \mathbb{P}(Y_{X=x'}(\epsilon) = y') \mathbb{P}(\epsilon \mid X = x, Y = y)$$

By conditioning on $X = x$ and $Y = y$, the counterfactual outcome y' under the intervention $do(X = x')$ is the expectation of the index function $Y_{X=x'}(\epsilon) = y'$ with respect to the updated probability distribution $\mathbb{P}(\epsilon | X = x, Y = y)$. Three special cases of distributions of this type are of special interest for us.

Definition 7 (Probability of necessity, [24])**.** *Let X and Y be two binary variables in a causal model $\mathcal{M} = \langle \mathcal{X}, \mathcal{F}, \varepsilon \rangle$. The probability of necessity (PN) is defined as:*

$$PN := \mathbb{P}(Y_{X=x'} = y' | X = x, Y = y)$$

Definition 8 (Probability of sufficiency, [24])**.** *Let X and and Y be two binary variables in a causal model $\mathcal{M} = \langle \mathcal{X}, \mathcal{F}, \varepsilon \rangle$. The probability of sufficiency (PS) is defined as:*

$$PS := \mathbb{P}(Y_{X=x} = y | X = x', Y = y')$$

The PN is the probability of observing a different outcome in the absence of the event $X = x$. The PS is the probability of x to generate y in cases where both had different values (x' and y').

Definition 9 (Probability of necessity and sufficiency, [24])**.** *Let X and and Y be two binary variables in a causal model $\mathcal{M} = \langle \mathcal{X}, \mathcal{F}, \varepsilon \rangle$. The probability of necessity and sufficiency is defined as:*

$$PNS := \mathbb{P}(y_x, y'_{x'}) = \mathbb{P}(x, y)PN + P\left(x', y'\right) PS$$

The PNS computes the probability that $X = x$ is the only way of obtaining $Y = y$, in other words, the probability that $X = x$ is both necessary and sufficient to observe $Y = y$. The probabilities PN, PS and PNS are not identifiable with observational or experimental data unless Y is monotonic with respect to X, and both observational and experimental data are available [31]. If this condition is satisfied, then they are identifiable and can be computed as follows:

$$PN = \frac{\mathbb{P}(y) - \mathbb{P}(y|do(x'))}{\mathbb{P}(x, y)} \text{ and } PS = \frac{\mathbb{P}(y|do(x)) - \mathbb{P}(y)}{\mathbb{P}(x', y')}. \tag{2}$$

Note that PN and PS require the knowledge of $do(X = x)$ and $do(X = x')$. These quantities are generally unobserved for the whole population since observed individuals are only subject to one of the two conditions, unless experimental data is available.

B ConPref problem

Congruent preferences (`ConPref`): Consider three real numbers M, N and T. If $M \leq N$ and $N \leq T$, then $M \leq T$. We compute PN and PS for the condition $M \leq N$ (C_{mn}) to having enough evidence to know if $M \leq T$ (C_{mnt}). If $M \leq N$ or $N \leq T$ are false then C_{mnt} is false. For our evaluation, we consider all combinations of values for M, N and T, for numbers between 1 and 8.

Reasoning graph

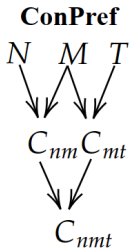


Figure 8: Reasoning graph for the `ConPref` problem.

Reasoning results

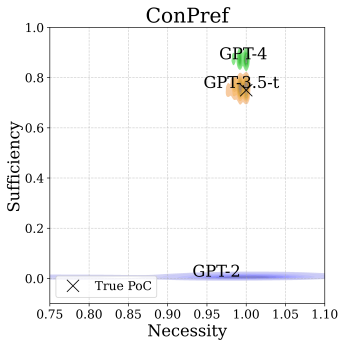


Figure 9: True PN and PS vs. inferred PN and PS using GPT-2, GPT-35-turbo and GPT-4 for the `ConPref` problem. .

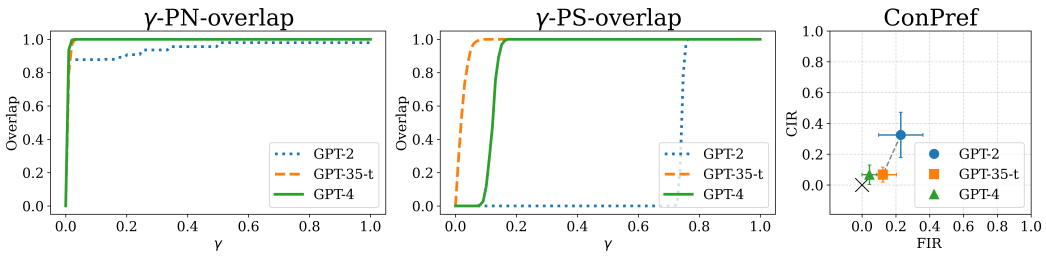


Figure 10: *Left*, *Centre*: Reconstruction of the γ-PN-overlap and γ-PS-overlap curves for GPT-2, GPT-35-turbo and GPT-4 for the `ConPref` problem. Ideal reasoning is achieved when the overlap is one for all values of γ. *Right*: Visualization of FIR and PIR. Ideal reasoning is attained when both metrics are zero (denoted by a $\times$).

C Problems: structural equations

C.1 `Div6`

$$\begin{aligned} N &\sim \mathbb{P}_N \\ C_2 &= N \pmod 3 \equiv 0 \\ C_3 &= N \pmod 2 \equiv 0 \\ C_6 &= C_2 \wedge C_3 \end{aligned}$$

where C_2, C_3 and C_6 represent boolean values that indicate whether the number is divisible by 2, 3 and 6 respectively. $\mathbb{P}_N$ is the mechanism for generating the original numbers (1 to 400 in our examples).

C.2 `EvenSum`

$$\begin{aligned} N &\sim \mathbb{P}_N \\ M &\sim \mathbb{P}_M \\ T &\sim \mathbb{P}_T \\ C_n &= N \pmod 2 \\ C_m &= M \pmod 2 \\ C_t &= T \pmod 2 \\ C_{nmt} &= (C_n + C_m + C_t = 1) \ \wedge \ (C_n + C_m + C_t = 3) \end{aligned}$$

where C_n, C_m, C_t and C_{nmt} represent boolean values and $\mathbb{P}_N$, $\mathbb{P}_M$, $\mathbb{P}_T$ are the mechanisms to generate the original numbers (1-8 in our examples).

C.3 `ConPref`

$$\begin{aligned} N &\sim \mathbb{P}_N \\ M &\sim \mathbb{P}_M \\ T &\sim \mathbb{P}_T \\ C_{nm} &= N \leq M \\ C_{mt} &= M \leq T \\ C_{nmt} &= C_{nm} \wedge C_{mt} \end{aligned}$$

where C_{nm}, C_{mt} and C_{nmt} represent boolean values and $\mathbb{P}_N$, $\mathbb{P}_M$, $\mathbb{P}_T$ are the mechanisms to generate the original numbers (1-8 in our examples).

C.4 *Candy Party*

$$
\begin{aligned}
R &\sim \mathbb{P}_R \\
L &\sim \mathbb{P}_L \\
E &\sim \mathbb{P}_E \\
C_{r>0} &= R > 0 \\
C_{l>0} &= L > 0 \\
C_{e>0} &= E > 0 \\
C_{r\geq 2} &= R \geq 2 \\
C_{l\geq 2} &= L \geq 2 \\
C_{e\geq 2} &= E \geq 2 \\
C_{rl} &= R > L \\
C_{re} &= R > E \\
C_{l=e} &= L = E \\
C_{r\geq 0, l\geq 0, e\geq 0} &= C_{r\geq 0} \wedge C_{l>0} \wedge C_{e\geq 0} \\
C_{r>l, r>e} &= C_{rl} \wedge C_{re} \\
C_{r\geq 2, l\geq 2, e\geq 2} &= C_{r\geq 2} \wedge C_{l\geq 2} \wedge C_{e\geq 2} \\
C_h &= (C_{r\geq 2, l\geq 2, e\geq 2} = 1) \wedge (C_{l=e} \vee C_{r\geq 0, l\geq 0, e\geq 0})
\end{aligned}
$$

where $\mathbb{P}_R, \mathbb{P}_L, \mathbb{P}_E$ are the mechanisms to generate the original numbers (all combinations in which 20 candies can be shared in our example).

D Proof of LLMs Zero-counterfactual consistency

Proof. Commutability of the Hex diagram implies that all paths from σ_0 to σ_1 result in the same outcome. This holds for all counterfactuals, which implies that $Y^{LLM}_{X=x} = Y_{X=x}$ for any value of X and Z. Therefore:

$$\mathbb{E}_{\mathbb{P}(X,Y,Z)}\left[Y^{LLM}_{X=x} \neq Y_{X=x}\right] = 0$$

for any $\mathbb{P}(X, Y, Z)$. □

E Direct and counterfactual prompts

E.1 `Div6` problem

Direct Prompt:

```
"Does 6 divide {‘X’}?  Use the factor method to answer this question.  Be
as concise as possible."
```

Counterfactual Prompt:

```
"Imagine that {‘X’} {‘has’/‘has not’} 3 as prime factor while retaining all
its other prime factors.  With this assumption does {self.divisor} divide
{‘X’}?  Use the factor method to answer this question.  Be as concise as
possible."
```

E.2 `EvenSum` problem

Counterfactual Prompt:

```
"Let N, M and T be three integers.  Then N+M+T is even if the three numbers
are even or if only one is even and the remaining two are odd.  Consider
the numbers N={N}, M={M} and T={T}.  Is N+M+T even?  Be as concise as
possible."
```

Direct Prompt:

Let N, M and T be three integers. Then N+M+T is even if the three numbers are even or if only one is even and the remaining two are odd. Consider the numbers N={N}, M={M} and T={T} and imagine that N {is/is not} even. With this assumption, is N+M+T even? Be as concise as possible."

E.3 ConPref problem

Direct Prompt:

"Let N, M and T be three integers. We know that if N is smaller or equal that M and M is smaller or equal than T then N is smaller or equal than T.Consider the numbers N={N}, M={M} and T={T}. By only looking at the relationships (N={N} vs. M={M}) and (M={M} vs. T={T}), can we know if N is smaller or equal that T? Be as concise as possible."

Counterfactual Prompt:

"Let N, M and T be three integers. We know that if N is smaller or equal that M and M is smaller or equal than T then N is smaller or equal than T. Consider the numbers N={N}, M={M} and T={T}. Now imagine that the number N {‘is smaller or equal ’/ ‘is not smaller or equal’} than M. Even if this contradict the values of the numbers X and Y, use this assumption and the relationships between and M={M} and T={T}, to decide if can we tell if N is smaller or equal that T? Don’t make any conclusion or comment based on the values, just based on the assumption and the relationships. Be as concise as possible."

E.4 CandyParty problem

Direct Prompt:

"Rafa has invited Lara and Emma to his birthday party. He has {num candies} to distribute among them. They all will be happy in the party in one of the following cases: 1) Each of them gets at least 2 candies or 2) Lara and Emma get the same number of candies, but at least one candy each, and Rafa gets more than them. After distributing the candies, Lara gets {L}, Emma gets {E} and Raphael gets {R} candies. With this candies distribution, will they all be happy in the party? Be as concise as possible."

Counterfactual Prompt:

Rafa has invited Lara and Emma to his birthday party. He has {num candies} candies to distribute among them. They all will be happy in the party in one of the following cases: 1) Each of them gets at least 2 candies or 2) Lara and Emma get the same number of candies, but at least one candy each, and Rafa gets more than them After distributing the candies. After distributing the candies, Lara gets {L}, Emma gets {E} and Rafa gets {R} candies. Consider the number of candies distributed to each of them and imagine that they think that {‘Lara and Emma have the same number of candies’/‘Lara and Emma have different number of candies’}. With this assumption, will they all be happy in the party? Be as concise as possible."

F GPT-4 concretization

Prompt: `"You are an entity extractor expert. I am going to give you a question-answer pair. I want you to say if the meaning of answer is positive or negative. If the answer has words like 'Yes' this will make it positive. If the answer contains words like 'No' this will make it negative. Always answer with only one word (Positive or Negative). For the question '{question}' and the answer '{answer}' the meaning is"`

G HEX for counterfactual query

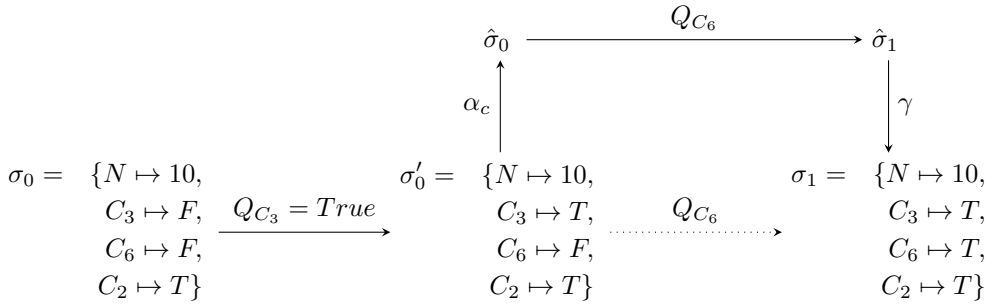


Figure 11: The HEX diagram for a counterfactual query in the `Div6` problem. We split the query in two sub-queries $Q_{C_3=True}$ and Q_{C_6} that performs the two operations need to compute the counterfactual state. $Q_{C_3=True}$ only sets the value of C_3 to True. Q_{C_6} replaces the value of C_6 by its counterfactual. This operation can be executed via the concrete path (using the structural causal model of the problem) of by using an LLM.

H Evaluation metrics

Let n be the number of instances of each problem. For example, $n = 400$ for the `Div6` problem because we use the first 400 integers to test reasoning. For the intervention node X and the outcome note Y, we distinguish between factual predictions $Y|X$ (simulated from the original reasoning graph) and counterfactual predictions $Y_{X=x}$ (simulated from the intervened graph). We respectively denote the LLM versions of this quantities as $Y^{LLM}|X = x$ and $Y^{LLM}_{X=x}$, which are computed via factual and counterfactual prompts.

$$\texttt{FIR} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[(Y^{LLM}|X = x) \neq (Y|X = x)\right]. \tag{3}$$

$$\texttt{CIR} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[Y^{LLM}_{X=x} \neq Y_{X=x}\right], \tag{4}$$

Let m be the number of bootstrap samples used from the binary answers of the LLM. $\hat{PN}_i$ and $\hat{PS}_i$ are estimation of PN and PS for the ith bootstrap sample. Then

$$\gamma - \texttt{PNO} := \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\left[|\hat{PN}_i^{LLM} - PN| \leq \gamma\right] \tag{5}$$

$$\gamma - \texttt{PNS} := \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\left[|\hat{PS}_i^{LLM} - PS| \leq \gamma\right] \tag{6}$$

I Element-wise PIR and FIR

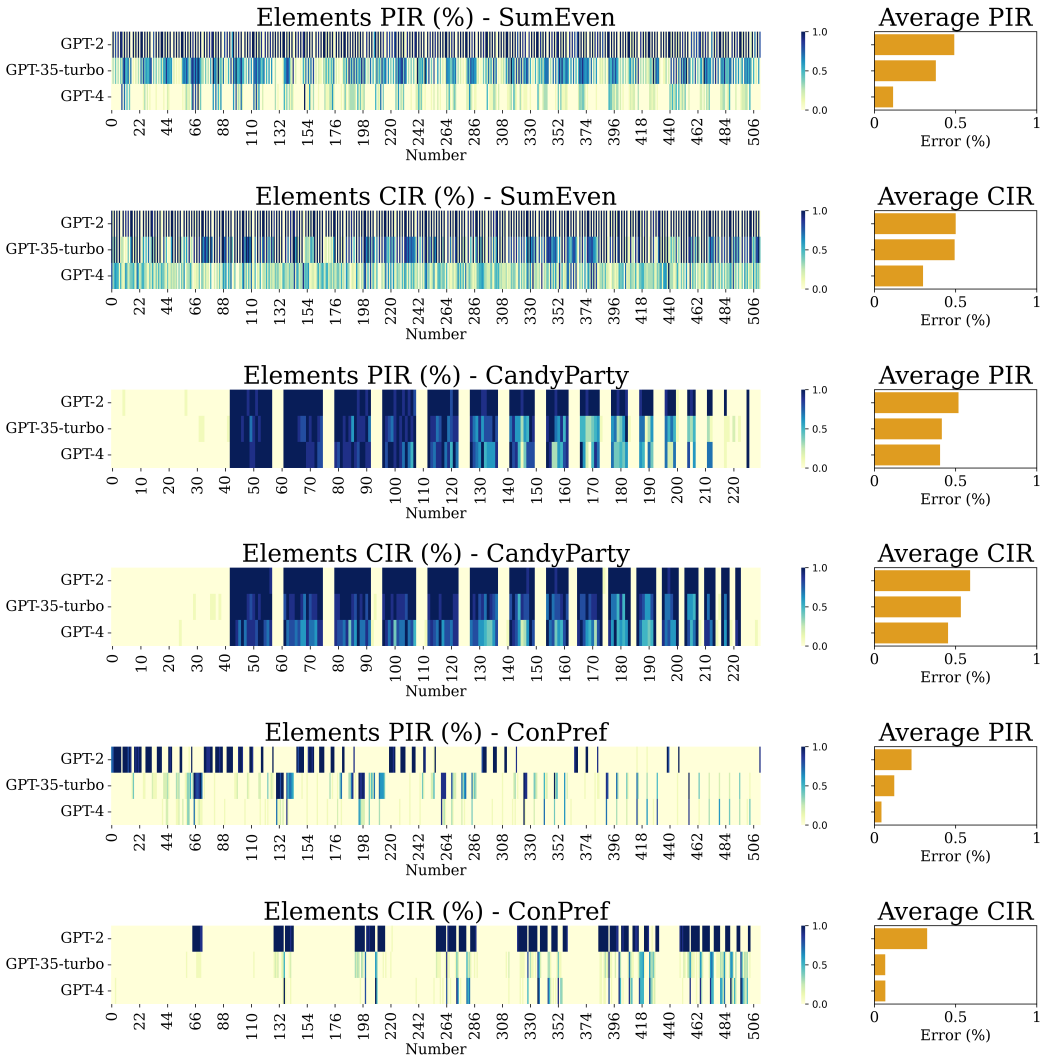


Figure 12: Element and aggregated CIR and FiR for the SumEven and CandyParty problems.

J Experiments with other model families

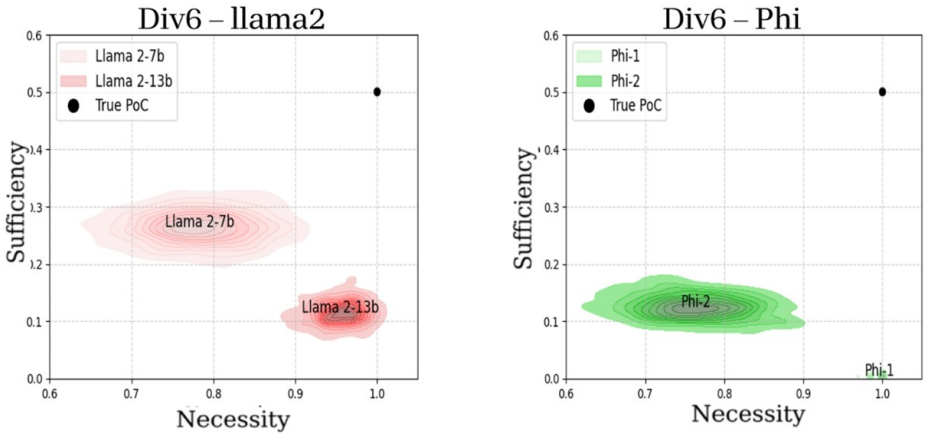


Figure 13: in other families of models. We re-run the Div6 problem with the same setup used in the paper with two new families of models: Llama (7-7b and 13b) and Phi (1,2). The results are consistent with the findings discussed in the main body of the paper

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims in abstract and title match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: We have included a list of limitations in the conclusions of the work.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All proofs to the claims in the paper are available in the supplementary materials.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The methods and the experimental setup are detailed in the main body of the paper and are enough to reproduce the evaluation metrics and the experimental results of this work. No new model or achitecture is proposed.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not provided, but all the prompts used to extract the results of the language models are detailed in the supplementary materials and can be used to reproduce the results of this work.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the details needed to reproduce the experimental section of this work are included in the paper.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are included in the results of the paper as well as metrics that capture the statistical significance of the experiments.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: Experiments do not require extensive compute resources (like large memory of GPUs) since they only require calls to GPT API models and some simple local computation. The experiments of this work can be reproduced in any personal laptop.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed the NeurIPS Code of Ethics.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: An impact statement has been included after the conclusions.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No models are trained or datasets released in this work. All the experiments are based in simulated data that can be reproduced following the indications described in the work.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly reference all authors and assets used in this work.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No assets were generated in the development of these research. The metrics proposed in this work can be easily implemented are used by anyone interested.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Only simulated data was used in the research of this work.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Only simulated data was used in the research of this work.