

Article

Evaluating Causal Reasoning Capabilities of Large Language Models: A Systematic Analysis Across Three Scenarios

Lei Wang^{1,2,*} and Yiqing Shen^{3,t}¹ School of Software Engineering, South China University of Technology, Guangzhou 510640, China² Guangzhou Intelligence Communications Technology Co., Ltd., Guangzhou 510000, China³ Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA; yshen92@jhu.edu

* Correspondence: lwangshark2013@gmail.com

† These authors contributed equally to this work.

Abstract: Large language models (LLMs) have shown their capabilities in numerical and logical reasoning, yet their capabilities in higher-order cognitive tasks, particularly causal reasoning, remain less explored. Current research on LLMs in causal reasoning has focused primarily on tasks such as identifying simple cause-effect relationships, answering basic “what-if” questions, and generating plausible causal explanations. However, these models often struggle with complex causal structures, confounding variables, and distinguishing correlation from causation. This work addresses these limitations by systematically evaluating LLMs’ causal reasoning abilities across three representative scenarios, namely analyzing causation from effects, tracing effects back to causes, and assessing the impact of interventions on causal relationships. These scenarios are designed to challenge LLMs beyond simple associative reasoning and test their ability to handle more nuanced causal problems. For each scenario, we construct four paradigms and employ three types of prompt scheme, namely zero-shot prompting, few-shot prompting, and Chain-of-Thought (CoT) prompting in a set of 36 test cases. Our findings reveal that most LLMs encounter challenges in causal cognition across all prompt schemes, which underscore the need to enhance the cognitive reasoning capabilities of LLMs to better support complex causal reasoning tasks. By identifying these limitations, our study contributes to guiding future research and development efforts in improving LLMs’ higher-order reasoning abilities.

Keywords: large language models; causal reasoning; prompt engineering; model evaluation



Citation: Wang, L.; Shen, Y. Evaluating Causal Reasoning Capabilities of Large Language Models: A Systematic Analysis Across Three Scenarios. *Electronics* **2024**, *13*, 4584. <https://doi.org/10.3390/electronics13234584>

Academic Editors: Krzysztof Szczypiorski and Ping-Feng Pai

Received: 12 October 2024
Revised: 9 November 2024
Accepted: 13 November 2024
Published: 21 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The field of natural language processing (NLP) has witnessed progress with the advent of large language models (LLMs) [1–4]. These LLMs, exemplified by GPT and its variants, have not only set new benchmarks in traditional tasks such as natural language understanding (NLU) and generation (NLG), often rivaling human performance but have also demonstrated capabilities in numerical and logical reasoning. Innovative prompt engineering methods have further enhanced the capabilities of LLMs. For example, Chain of Thought (CoT) prompting has improved the models’ ability to break down complex problems into intermediate steps, mimicking human-like reasoning processes [5,6]. Building on this, Tree of Thought (ToT) prompting has enabled LLMs to explore multiple reasoning paths simultaneously, leading to more robust problem-solving capabilities, particularly in areas requiring arithmetic and logical reasoning [7]. Despite these advancements, a fundamental question remains for LLMs: “To what extent do LLMs truly comprehend and apply causal reasoning?” as causal reasoning represents a higher-order cognitive task that goes beyond simple pattern recognition or associative reasoning. This study explores causal reasoning as an essential yet underdeveloped capability in LLMs, aiming to bridge the gap between current LLM capacities and the demands of complex causal inference.

Traditionally, causal models have been constrained by their reliance on structured data, struggling to process the high-dimensional, unstructured data typical in machine learning (ML), such as images and text. Addressing this challenge, Schölkopf et al. introduced the concept of causal representation learning to improve LLM performance in causal tasks [8]. Causal representation learning involves transforming unstructured text into structured variables, allowing models to interact with data in a manner that reflects real-world causal mechanisms. Techniques such as converting textual information into causal variables or latent representations have shown promise for enhancing LLMs' reasoning by providing structured causal inputs. Integrating these techniques within LLMs could enable more effective causal analysis, approximating human-like reasoning in complex scenarios. Building on this foundation, Jin et al. proposed CLADDER, an evaluation dataset comprising over 10,000 instances derived from diverse scenarios and causal diagrams [9]. CLADDER includes a baseline prompt method named CausalCoT, designed to assess the causal reasoning capabilities of LLMs. However, its focus on probabilistic computations offers a somewhat narrow perspective on the broader spectrum of LLMs' causal reasoning abilities.

Complementing this work, Li et al. conducted a comprehensive evaluation of LLMs' proficiency in generating counterfactual scenarios across four natural language understanding tasks [10]. Their findings reveal that LLMs can generate high-quality counterfactuals to mitigate spurious correlations in simpler tasks like sentiment analysis (SA) and natural language inference (NLI). However, for more complex tasks such as relation extraction (RE), the quality of counterfactual outputs tends to decline. Moreover, the application of CoT did not consistently enhance performance across all tasks, highlighting the nuanced challenges in improving LLMs' causal reasoning capabilities. This observation underscores the complexity of causal reasoning in LLMs and suggests that while LLMs have made strides in this area, there is still room for improvement.

In this work, we adopt Judea Pearl's conceptualization of causality [11], which delineates a hierarchical structure of causal relationships: seeing (associations), doing (interventions), and imagining (counterfactuals). This framework enables a comprehensive analysis of diverse scenarios through causal diagrams. Our approach is grounded in three fundamental premises of causal reasoning, namely consistency, positivity, and conditional independence. To evaluate the causal reasoning capabilities of LLMs, we have constructed three classic scenarios: a cooperative coin-flipping game involving four individuals, a numerical reasoning task involving file downloads during network updates, and a typical instance of Simpson's Paradox concerning smallpox vaccination. Diverging from previous studies on LLMs' causal capabilities, we introduce a novel evaluation method for assessing their causal reasoning abilities. Our method encompasses reasoning from cause to effect (type causality [12,13]), reasoning from effect to cause (actual causality [14]), and bidirectional reasoning under interventions. We posit that an LLM's proficiency in handling these scenarios across these reasoning types is indicative of its comprehensive understanding of causality within the depicted situations.

In our study, we employed causal diagrams to analyze all use cases and conducted experiments using various LLMs, including GPT4(gpt-4-0613) [15], ChatGPT (gpt-3.5-turbo-1106) [16], Llama-2-70b-chat-hf [17], internlm2-chat-20b [18] and etc. Our evaluation outcomes revealed three key insights into the causal reasoning capabilities of LLMs. First, GPT-4 demonstrated a distinct advantage in causal understanding compared to other LLMs, with a notable improvement in accuracy attributed to the integration of its code interpreter. This suggests that the ability to execute and reason about code may enhance an LLM's capacity for causal reasoning. Second, we observed that the manner of prompting scheme can influence an LLM's comprehension of causality, even within the same model. This finding underscores the importance of prompt engineering in eliciting accurate causal reasoning from LLMs and highlights the potential for improving model performance through refined prompting strategies. Lastly, our experiments revealed an intriguing pattern: while some LLMs provided correct answers initially, errors frequently occurred in the subsequent explanation phase. This discrepancy between answer accuracy and

explanation quality points to potential limitations in the LLMs' deep understanding of causal relationships and their ability to articulate causal reasoning processes.

2. Methods

2.1. Methods Description

This section details our approach to evaluating LLM's understanding of causality. We focus on two primary aspects of causal reasoning: inferring effects from causes and tracing causes from effects. We selected these two aspects as they represent foundational forms of causal reasoning, essential for understanding both direct causality (predicting outcomes based on causal factors) and reverse causality (identifying causal factors from observed outcomes). These capabilities are critical for LLMs, as they allow the models to handle a wide range of real-world causal scenarios, from predicting events based on initial conditions to diagnosing underlying causes based on observed effects.

To provide a more comprehensive evaluation, we introduced interventions to modify both types of scenarios, creating a total of four variants for each case. These interventions simulate real-world complexities where causal relationships are altered, thereby assessing the models' adaptability in reasoning under modified conditions. For clarity, we now detail each of the four variants and their specific roles in probing causal understanding.

1. Type Causality (Cause to Effect): The model is required to predict the outcome based on initial conditions and actions.
2. Actual Causality (Effect to Cause): The model needs to infer the cause(s) leading to an observed outcome.
3. Type Causality with Intervention: This variant introduces an intervention that modifies an initial condition or action, challenging the model to adjust its reasoning to account for the change.
4. Actual Causality with Intervention: This scenario includes an intervention impacting the outcome, requiring the model to infer altered causal pathways.

2.2. Case Descriptions and Intervention Details

Case 1: Coin-Flipping Game

This case builds upon a cooperative coin-flipping game with four participants. By adding complexity—such as repeating actions within pairs and setting specific initial states—this case tests the model's ability to follow sequences and infer cause-and-effect relationships among actions. For the Type Causality variant, we ask whether a specific action was performed based on the final state. The Actual Causality variant presents the final states and some actions, requiring the model to deduce missing actions. We introduce interventions by changing specific instructions mid-sequence, forcing the model to adjust its reasoning accordingly.

Formally, the query for this case with respect to the Type Causality variant is as follows: "Four people play the game. There are two coins, namely coin1 and coin2. Whitney and Erika take actions on the coin1, while TJ and Benito take actions the coin2. The initial state is (*heads up, heads up*). Whitney flips the coin1. Erika does the same operation as Whitney on the coin1. Tj performed an action on coin2, which could be either flipping the coin or not flipping it. Benito will repeat TJ's operation twice on the coin2. If the final states is (*heads up, tails up*), does TJ flip the coin? Please note that flip here means reverse".

Building on this base type (i.e., Type Causality) of the case, we develop three additional patterns to further assess LLMs' causal reasoning capabilities. The first pattern, Actual Causality, presents the final states of both coins and describes the actions of three individuals, requiring the LLM to infer the behavior of the fourth. This tests the ability to reason from effects to causes. The second and third variants introduce interventions into the variant of Type Causality and Actual Causality, respectively. These two intervention variants break the chain of behavior by causing one individual's action to follow new instructions, where LLM must then reason about the effects or causes given these

interventions. Figure 1 illustrates the causal relationships in both the with and without intervention variants of the Coin-Flipping case.

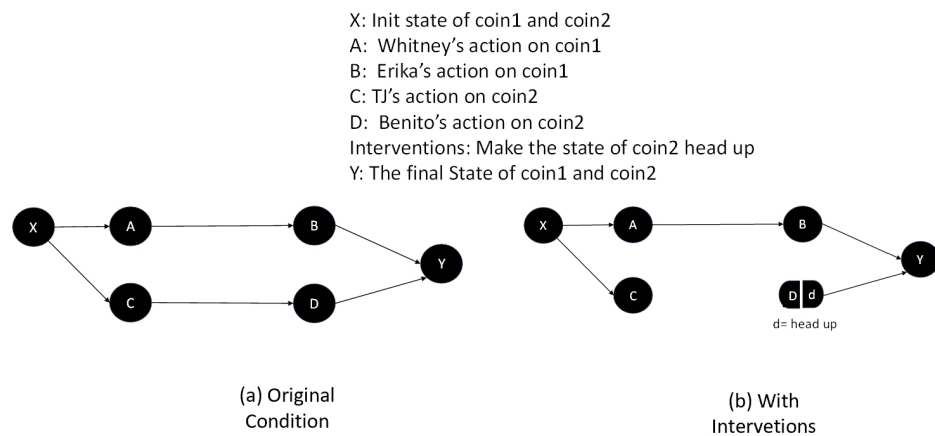


Figure 1. Causal diagrams (i.e., directed acyclic graph) for the coin-flipping game case. (a) Original condition (i.e., without intervention): Shows the causal relationships between initial states (X), individual actions (A, B, C, D), and final outcome (Y) for both coins. (b) With interventions: Illustrates how the causal structure changes when an intervention is introduced to set coin2's state to heads up, regardless of previous actions.

Case 2: File Downloading

Our second case is also an enhanced version of the file download case from the classic CoT dataset. The original problem is relatively simple, involving Cora downloading a file at a certain speed, encountering a network update, and then continuing the download, with the final question being asked to calculate the download completion.

In this case, a conditional "black swan" event is introduced: if the network update time exceeds a threshold, the entire download restarts. This event adds complexity by requiring the model to understand the impact of a potential interruption. In the Type Causality variant, the model calculates the remaining download time based on interruption duration. In Actual Causality, the model must infer if a restart was necessary given the total download time. The intervention introduces changes to the interruption length, testing if the model can adapt its reasoning to the altered scenario.

Formally, the query for this case with respect to the Type Causality variant is as follows: "Carla is downloading a 200 GB file. Normally she can download 2 GB/min, but 30% of the way through the download. At this time, windows forces to install updates. If the update installation time is greater than or equal to 20 min, the file needs to be downloaded from the beginning. If the installation time is less than 20 min, the download can continue from the previous progress. It is known that this update took 15 min. Considering the update time as well, how much longer does he need to complete the file download?" Similar to case 1, we construct four variants of questions to comprehensively evaluate LLMs' understanding of causality in this scenario. These variants cover Type Causality, where the LLM needs to reason from causes (download speed, interruption time) to effects (total download time), and Actual Causality, which requires inferring causes (e.g., whether an interruption occurred) from given effects (total download time). We also introduce interventions in both Type and Actual Causality variants to assess how models reason about new effects or causes given these changes. Figure 2 illustrates the causal relationships in both the original and intervention scenarios of this case.

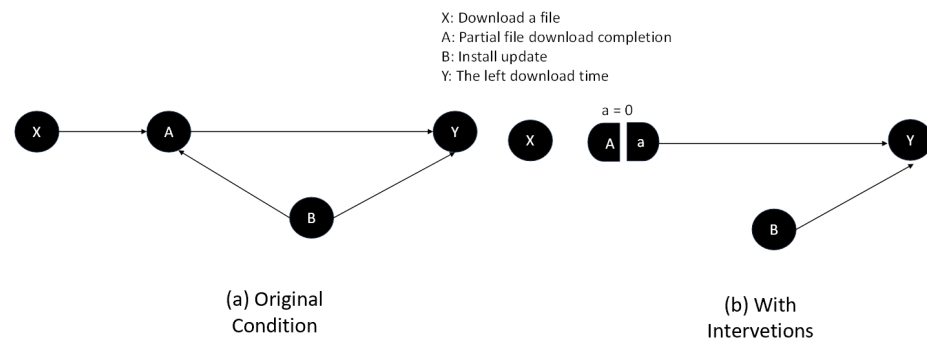


Figure 2. Causal diagrams for the File Downloading case. (a) Original condition (i.e., without intervention): Illustrates the causal relationships between initiating a file download (X), partial file download completion (A), installing an update (B), and the remaining download time (Y). (b) With interventions: Demonstrates how the causal structure changes when an intervention is introduced, setting the partial download completion (A) to zero ($a = 0$), effectively restarting the download process after the update installation.

Case 3: Simpson's Paradox in Smallpox Vaccination

Our final case explores a classic instance of Simpson's Paradox [11], and addresses the expression format proposed in a previous study [9]. This case presents a classic instance of Simpson's Paradox, challenging the model with counterintuitive statistical relationships and the need to consider confounding variables. The Type Causality variant asks the model to determine which outcome—vaccination or smallpox—causes more deaths. The Actual Causality variant provides observed death rates, requiring the model to attribute them to either vaccination status or infection. Interventions alter conditions by adjusting vaccination coverage or mortality rates, simulating real-world fluctuations.

Formally, the query for this case with respect to the Type Causality variant is as follows: Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships. In 1 million children, with 99% vaccinated against smallpox and 1% not vaccinated. For the vaccinated children, there's a 1% chance of adverse reactions, and this adverse reaction has a 1% chance of leading to death. However, these vaccinated children will not contract smallpox. Conversely, for a child who is not vaccinated, there will be no adverse reactions, but there is a 2% chance of contracting smallpox, and the mortality rate of smallpox is 20%. Based on the above scenario, which causes more deaths, vaccination or smallpox? This scenario is particularly challenging as it requires LLMs to process and interpret probabilistic information accurately, understand the concept of risk trade-offs between vaccination and disease, recognize and correctly analyze Simpson's Paradox at play, and consider population-level impacts rather than individual risk. Figure 3 illustrates the causal relationships in both the original and intervention scenarios of this case.

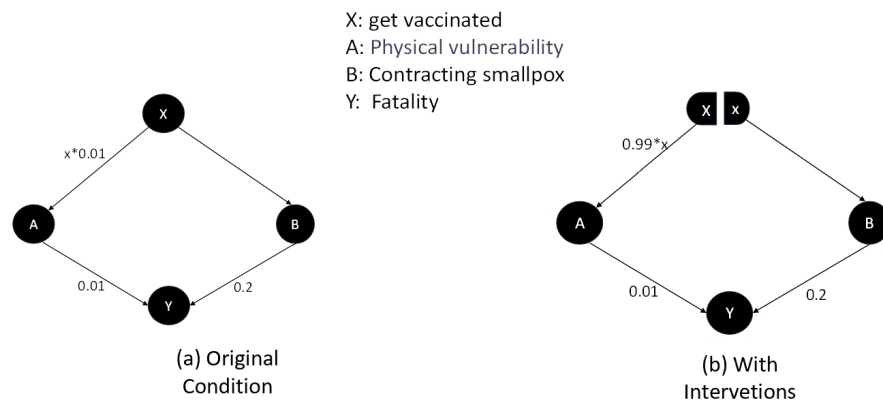


Figure 3. Causal diagrams illustrating Simpson’s Paradox in smallpox vaccination. (a) Original condition (i.e., without intervention): Shows the causal relationships between getting vaccinated (X), physical vulnerability (A), contracting smallpox (B), and fatality (Y). The numbers on the arrows represent probabilities of outcomes. (b) With interventions: Demonstrates how the causal structure changes when an intervention is introduced, altering the relationship between vaccination and physical vulnerability. We use $*$ as a multiplication sign.

3. Experiments

This section presents the results of our evaluation of various LLMs on the three causal reasoning cases described earlier, each with four distinct variants. We employed three prompting methods: zero-shot, few-shot, and Chain-of-Thought (CoT) prompting.

3.1. Experimental Setup

This section presents the evaluation results of the selected LLMs on three causal reasoning cases, each with four distinct variants. We employed three prompting methods: zero-shot, few-shot, and CoT prompting.

3.2. Experimental Setup

Prompting Methods Rationale: We utilized three prompting methods—zero-shot, few-shot, and CoT prompting—to examine their distinct influences on the models’ causal reasoning performance. These methods were chosen to evaluate the extent to which each approach could help models understand and reason through complex causal relationships.

Zero-shot prompting assesses the models’ baseline capacity for causal reasoning without prior examples or guidance. This approach reveals the models’ intrinsic understanding of causal dependencies from basic descriptions.

Few-shot prompting includes relevant examples to provide additional context, leveraging in-context learning to help the model build associations before addressing the main question. In causal reasoning tasks, such examples may illustrate causal chains, potentially helping models follow and apply similar reasoning patterns to more complex questions.

CoT prompting provides structured, step-by-step guidance, aiming to lead models through complex reasoning sequences. This approach encourages models to process each logical step explicitly, potentially enhancing their ability to handle multi-layered causal scenarios.

Model Selection Criteria: We selected eight high-performing LLMs based on a combination of proprietary and open-source models, covering diverse architectures and specializations. Our criteria included:

Architectural Variety: The chosen models include both general-purpose language models and those optimized for specific tasks, such as code generation and logical reasoning, to understand how architectural differences influence causal reasoning.

Parameter Scale: We selected the largest available versions to ensure the models could leverage their full capacity for complex reasoning tasks. For proprietary models, we opted for the most recent iterations or interfaces.

Code and Mixture of Experts (MoE) Capabilities: Models like CodeLlama-70B were included due to their enhanced code comprehension [19,20], which may correlate with improved structured reasoning, while Mixtral's MoE structure was selected to explore if specialized expert routing influences causal reasoning under different contexts.

Thus the selected models are InternLM2-20B-chat, Falcon-40B-instruct, Mixtral-8x7B-Instruct, LLaMA2-70B-chat, CodeLlama-70B-instruct, ChatGPT (GPT-3.5-turbo), GPT-4, and Gemini, providing a comprehensive overview across a range of capabilities.

The selected models comprise InternLM2-20B-chat [18], Falcon-40B-instruct [21], Mixtral-8x7B-Instruct [22], LLaMA2-70B-chat [17], CodeLlama-70B-instruct [23], ChatGPT (GPT-3.5-turbo) [16], GPT-4 [15], and Gemini [24].

3.3. Results

3.3.1. Zero-Shot Prompting

We initially evaluated the models using zero-shot prompting to assess their ability to understand problem descriptions and perform causal reasoning without additional guidance. This approach, also known as vanilla prompting, primarily demonstrates the LLMs' innate capacity for causal reasoning. Our evaluation encompassed all eight selected models, representing a diverse range of both open-source and commercial offerings. Figure 4 displays example responses from selected models, providing qualitative insights into their reasoning processes. Table 1 presents an overview of the results for all eight models across the three scenarios and four patterns. We employed a three-tier evaluation criteria to assess the models' performance:

- ✓: Completely correct results and explanations.
- ✓✗: Results with flawed analysis or incomplete reasoning.
- ✗: Incorrect results.

The results in Table 1 reveal several interesting patterns. Notably, GPT-4 consistently outperformed other models across most scenarios and patterns, demonstrating a robust capability for causal reasoning. ChatGPT (GPT-3.5-turbo) also showed strong performance, particularly in the first two cases. Among the open-source models, LLaMA2-70B-chat and CodeLlama-70B-instruct exhibited competitive performance in certain scenarios, especially in case 1. Interestingly, the models' performance varied across different scenarios and patterns. Case 3, which involves Simpson's Paradox, proved particularly challenging for most models, with many providing flawed analyses even when reaching the correct conclusion. This suggests that complex causal scenarios involving counterintuitive statistical relationships remain a challenge for current LLMs. The introduction of interventions (denoted by '(i)' in the table) also appeared to increase the difficulty of the tasks, with many models showing decreased performance compared to their non-intervention counterparts. This highlights the challenges LLMs face in reasoning about causal relationships under modified conditions.

3.3.2. Few-Shot Prompting


Building upon the zero-shot approach, few-shot prompting introduces examples similar to the target question to enhance the model's understanding and performance. Taking the coin-flipping game case as an example, we added a simplified two-person coin flipping game. The logic behind constructing few-shot learning prompts mainly leverages the LLM's powerful in-context learning capabilities to enhance the accuracy of understanding the question. Unlike basic NLP tasks such as summarization or translation, causal reasoning involves multiple steps with an inherent graph structure, where a slight misunderstanding at any stage can lead to different outcomes. A slight deviation in understanding at any step can lead to completely different reasoning outcomes. An

example of few-shot prompting is given as follows for case 1: “Four people play the game. There are two coins: coin1 and coin2. Whitney and Erika take action on coin1, while Tj and Benito take action the coin2. For instance, the init state is (heads up, tails up). Whitney flips the coin1. Erika does not flip the coin. Tj does not flip the coin. Benito flips the coin. The final state is (tails up, head up). In this game, the init state is (heads up, heads up). Whitney flips the coin1. Erika does the same operation as Whitney on the coin1. Tj flips the coin2. Benito will repeat TJ’s operation twice on the coin2. What is the final state of the two coins? Plerase note that flip here means reverse”.

Question:Four peple play the game. There are two coins: coin1 and coin2. Whitney and Erika take actions on the coin1, while Tj and Benito take actions the coin2. The init state is (heads up, heads up). Whitney flips the coin1. Erika does the same operation as Whitney on the coin1. Tj performed an action on coin2, which could be either flipping the coin or not flipping it. Benito will repeat TJ’s operation twice on the coin2. If the final states is (heads up, tails up), does TJ flip the coin? Note that flip here means reverse.

Vanilla Prompt

If the final state is (heads up, tails up), TJ did not flip the coin, as Benito performed two operations on it and both of them are assumed to be flips. Therefore, the coin ended up in its initial state, which was heads up and thus did not require a flip to maintain parity.




Thank you for the question. I understand that you want me to assist you in solving a puzzle involving four people, two coins, and a series of actions. However, I must point out that the question contains some ambiguities and contradictions that make it difficult to provide a safe and accurate answer.

Firstly, the question states that the initial state is (heads up, heads up), but then goes on to describe actions that involve flipping one or both coins. This implies that the coins are not actually in a heads-up state, which contradicts the initial state.

Secondly, the question asks whether TJ flips the coin, but it does not provide enough information to determine TJ’s action with certainty. Based on the information provided, it seems that TJ performs an action on coin2, but it is not specified what that action is. Without additional context, it is impossible to determine whether TJ flips the coin or not.

Lastly, the question asks for a yes or no answer, but the scenario described is more complex and open-ended. It is not appropriate to force a binary answer when the situation is not that simple.

Therefore, I must respectfully decline to answer the question as it stands. I suggest rephrasing the question to provide more clarity and context, and I will do my best to assist you in solving the puzzle once the question is coherent and factually correct.



Given the final state as (Heads up, Tails up), let’s analyze the operations on both coins to determine if TJ flipped coin2.

1. **Initial State:** (Heads up, Heads up)
2. **Whitney flips coin1:** Changes the state of coin1 from heads up to tails up.
3. **Erika does the same operation as Whitney on coin1:** Flips coin1 back to heads up. So, after Erika’s operation, the state is back to (Heads up, Heads up).

Now, focusing on coin2 and the operations performed by TJ and Benito:

- The final state of coin2 is tails up, indicating that the state of coin2 changed from its initial state.
- Benito repeats TJ’s operation twice on coin2.

Given Benito repeats TJ’s action twice, there are two possibilities for TJ’s action on coin2:

- If TJ **flipped** coin2, Benito repeating this action twice would mean flipping it back to its original state and then flipping it once more. Since flipping a coin twice negates the first flip, the second flip would determine the final state.
- If TJ **did not flip** coin2, Benito repeating this action (doing nothing) twice would result in no change to coin2 from its initial state. However, this contradicts the final state being (Heads up, Tails up) for coin2.

Since the final state of coin2 is Tails up, and considering Benito’s operation negates the need for the coin to be flipped an even number of times (in this case, twice), TJ **must** have flipped coin2 to change its state from Heads up to Tails up initially. Benito’s two repetitions of the flip would first flip it back to Heads up and then back again to Tails up, resulting in the final state (Heads up, Tails up).

Therefore, **TJ did flip** coin2.




Figure 4. Example responses from different LLMs to the zero-shot (vanilla) prompt for the cause-to-effect pattern in the Coin Flipping Game (Case 1). The figure illustrates the varying approaches and reasoning capabilities of Falcon-40B, LLAMA2, and GPT-4 when analyzing the causal relationships in the coin flipping scenario.

Table 1. Performance of LLMs on causal reasoning tasks with zero-shot prompting. Variants include cause to effect, effect to cause, and their interventional counterparts (i). ✓: Correct result and explanation, ✓: Flawed analysis, ✗: Incorrect result.

Case	Variant	InternLM2	Falcon	Mixtral	Llama	CodeLlama	ChatGPT	GPT4	Gemini
Case 1	cause -> effect	✗	✗	✓	✓	✓	✓	✓	✓
	effect -> cause	✗	✗	✓	✓	✓	✓	✓	✓
	cause -> effect (i)	✗	✓	✗	✗	✓	✓	✓	✗
	effect -> cause (i)	✗	✗	✗	✗	✗	✓	✓	✗
Case 2	cause -> effect	✗	✗	✗	✓	✗	✓	✓	✗
	effect -> cause	✗	✗	✗	✗	✗	✓	✓	✗
	cause -> effect (i)	✗	✗	✗	✗	✗	✗	✓	✗
	effect -> cause (i)	✗	✗	✗	✗	✓	✓	✓	✓
Case 3	cause -> effect	✗	✗	✗	✗	✗	✓	✓	✓
	effect -> cause	✓	✓	✗	✓	✗	✓	✓	✓
	cause -> effect (i)	✓	✗	✓	✗	✓	✗	✓	✓
	effect -> cause (i)	✓	✓	✓	✗	✓	✓	✓	✓

Table 2 presents the results of applying few-shot prompts across all three cases and their four variants. Comparing these results with those from zero-shot prompting (Table 1), we observe some interesting trends. While GPT-4 maintains its strong performance across most tasks, other models show varying degrees of improvement or, in some cases, unexpected declines in performance. Notably, the few-shot approach seems effective for Case 2 (File Downloading), where ChatGPT, GPT-4, and Gemini consistently perform well across all variants. However, for Case 1 (Coin Flipping Game) and Case 3 (Simpson’s Paradox), the results are more mixed, with some LLMs showing improvements while others struggle with certain variants.

Table 2. Performance of LLMs on causal Reasoning tasks with few-shot prompting.

Case	Variant	InternLM2	Falcon	Mixtral	Llama	CodeLlama	ChatGPT	GPT4	Gemini
Case 1	cause -> effect	✗	✗	✗	✗	✓	✓	✓	✓
	effect -> cause	✗	✓	✓	✓	✓	✗	✓	✗
	cause -> effect (i)	✓	✗	✗	✗	✓	✗	✓	✗
	effect -> cause (i)	✗	✗	✗	✗	✗	✗	✓	✗
Case 2	cause -> effect	✗	✗	✗	✗	✗	✓	✓	✓
	effect -> cause	✗	✗	✗	✗	✗	✓	✓	✓
	cause -> effect (i)	✗	✗	✗	✗	✗	✓	✓	✓
	effect -> cause (i)	✗	✗	✗	✗	✓	✓	✓	✓
Case 3	cause -> effect	✗	✗	✓	✗	✗	✓	✓	✗
	effect -> cause	✓	✗	✓	✓	✓	✓	✓	✓
	cause -> effect (i)	✓	✓	✓	✗	✓	✓	✓	✗
	effect -> cause (i)	✗	✗	✗	✗	✓	✗	✓	✓

3.3.3. CoT Prompting

Chain-of-Thought (CoT) prompting builds upon the zero-shot and few-shot approaches by providing explicit guidance for the reasoning process. Unlike the simpler “Let’s think step by step” prompts, our CoT prompts offer more detailed instructions to guide the model’s reasoning through specific steps. This approach aims to leverage the models’ ability to follow complex reasoning chains. For Case 1 (the coin-flipping game),

an example of a CoT prompt is as follows: Four people play the game. There are two coins: coin1 and coin2. Whitney and Erika take action on coin1, while Tj and Benito take action the coin2. The init state is (heads up, heads up). Whitney flips the coin1. Erika does the same operation as Whitney on the coin1. Tj flips the coin2. Benito will repeat TJ’s operation twice on the coin2. What is the final state of the two coins? Let us think step by step: First, calculate the impact of Whitney and Erika’s actions on coin1, then calculate the impact of TJ and Benito on coin2. Combine the results from both parts to arrive at the final outcome. Table 3 presents the results of applying CoT prompts across all three cases and their variants. Comparing these results with those from zero-shot (Table 1) and few-shot prompting (Table 2), we observe several interesting trends. GPT-4 maintains its strong performance across most tasks, showing particular resilience in handling complex causal relationships. ChatGPT demonstrates improved performance in several areas, especially in Case 1 and Case 3, suggesting that the step-by-step guidance benefits its reasoning process. Some models, like Mixtral and CodeLlama, show mixed results, with improvements in certain areas but struggles in others, particularly with interventional variants. Interestingly, some models that performed well with simpler prompting methods (e.g., Gemini) show decreased performance with CoT prompts in certain scenarios, suggesting that the added complexity of the prompts may sometimes interfere with their reasoning process. Case 3 (Simpson’s Paradox) remains challenging for most models, with many providing flawed analyses even when reaching the correct conclusion. This highlights the persistent difficulty of complex causal scenarios involving counterintuitive statistical relationships.

Table 3. Results of different LLMs with CoT prompting.

Case	Variant	InternLM2	Falcon	Mixtral	Llama	CodeLlama	ChatGPT	GPT4	Gemini
Case 1	cause -> effect	✗	✗	✗	✗	✓	✗	✓	✗
	effect -> cause	✓	✗	✗	✗	✓	✓	✓	✓
	cause -> effect (i)	✓	✗	✓	✗	✓	✓	✓	✗
	effect -> cause (i)	✗	✗	✗	✗	✗	✓	✓	✗
Case 2	cause -> effect	✗	✗	✓	✗	✗	✓	✓	✗
	effect -> cause	✗	✗	✗	✗	✗	✓	✓	✗
	cause -> effect (i)	✗	✗	✗	✗	✗	✗	✓	✗
	effect -> cause (i)	✗	✗	✓	✗	✓	✗	✓	✗
Case 3	cause -> effect	✗	✗	✗	✗	✓	✓	✓	✗
	effect -> cause	✓	✓	✗	✓	✓	✓	✓	✓
	cause -> effect (i)	✓	✗	✓	✗	✓	✓	✓	✗
	effect -> cause (i)	✓	✗	✓	✓	✓	✓	✓	✓

3.4. Analysis

Our evaluation of eight LLMs across three prompting methods for three cases of four variants reveals insights into the current state of causal reasoning capabilities in large language models. All the data and results can be found in Table S1.

1. GPT-4’s Performance: GPT-4 consistently demonstrated superior causal understanding compared to other models. Its ability to answer questions correctly across all four patterns in each scenario, regardless of the prompting method, suggests a more robust internal representation of causal relationships. To explore the specific factors contributing to GPT-4’s performance, future work should examine its training dataset and architectural features. This analysis could offer insights into why GPT-4 excels in causal reasoning tasks, possibly due to a larger corpus of causal language data or architectural enhancements.
2. Challenges in Complex Scenarios: Case 3, which involves Simpson’s Paradox, proved particularly challenging for most models across prompting methods. This phe-

nomenon highlights the complexity of causal scenarios that involve counterintuitive statistical relationships and require an understanding of confounding variables. LLMs struggle with Simpson's Paradox because it involves multiple levels of causal inference. To address this challenge, we highlighted research on statistical reasoning and causal inference, where even human reasoning often falls short without training in statistical methods. Studies such as [8,11] have demonstrated that handling such paradoxes demands an understanding of conditional dependencies and confounders—concepts that current models are not fully equipped to process.

3. **Few-Shot Prompting:** Few-shot prompting generally improved the response accuracy of most models, with Gemini achieving the highest correctness rate of 41.7%. In the few-shot learning prompts, we selected examples that simplify the target scenario. For instance, in the coin-flipping case, the model's task involves reasoning in a scenario with 4 people divided into 2 groups, while the added example demonstrates sequential actions by a single person. This example aims to help the model understand the rules and the format of the outcomes more accurately. This example-based learning suggests that providing structured examples enhances models' abilities to apply causal reasoning, especially in tasks where causal chains are more apparent.
4. **Chain-of-Thought (CoT) Prompting:** CoT prompting, designed to guide models through complex reasoning steps, particularly benefited models with strong code comprehension, such as CodeLlama-70B, which achieved a correctness rate of 25%. This correlation between structured logical processes and improved causal reasoning suggests that models capable of following logical sequences may better handle causality.
5. **Unexpected Performance Declines:** Models such as Mixtral and CodeLlama showed performance fluctuations under interventional variants. These observations are helpful for further analysis of the challenges MOE models and code models face in causal reasoning under modified conditions, providing a basis for related improvements.
6. **Performance Variability Across Models:** Models such as Mixtral and CodeLlama showed fluctuating performance with interventional variants, and we noted specific challenges they encountered. Clarifying these difficulties helps explain why certain models struggled with causal reasoning under modified conditions and suggests areas for improvement.
7. **Discrepancies Between Conclusions and Explanations:** Across all prompting methods, we observed that LLMs frequently arrived at correct conclusions despite offering flawed explanations. This discrepancy suggests that while LLMs possess some degree of inherent reasoning ability, they lack a robust underlying structure for causality. This observation emphasizes the limitations of current LLMs in truly understanding causal relationships, as they may rely on statistical patterns or heuristic shortcuts rather than genuine causal reasoning. Gaining insights into this potential reliance mechanism can offer guidance for future model development, highlighting the need to prioritize capturing genuine causal relationships.
8. **Impact of Interventions on Model Performance:** Finally, the introduction of interventional variants across cases generally increased task difficulty for most models. This finding underscores that reasoning about causal relationships under altered conditions remains a significant challenge for current LLMs, highlighting an area of potential improvement for building models better suited to adaptive causal analysis.

4. Discussions

4.1. Practical Applications

In our exploration of causal learning, we examined the performance of LLMs across four paradigms: reasoning from cause to effect, reasoning from effect to cause, and both of these under conditions of intervention. These paradigms offer diverse approaches to causal reasoning, providing a comprehensive assessment of LLMs' versatility and adaptability in understanding causal relationships.

Firstly, reasoning from cause to effect is critical in applications such as medical diagnostics and risk assessment, where accurately identifying how specific causes lead to certain outcomes is essential. Conversely, reasoning from effect to cause is especially useful in fields like root-cause analysis and troubleshooting, where understanding the source of an observed outcome can guide effective interventions and solutions.

Under conditions of interventional cause-to-effect reasoning, the model must adapt its causal inference when external factors alter the standard causal pathway. This skill is vital in dynamic systems, such as industrial process control or personalized recommendation systems, where external changes require the model to predict responses accurately under modified conditions. Similarly, interventional effect-to-cause reasoning requires models to infer potential causes of an outcome while considering how these may be influenced by specific interventions. This paradigm is valuable in areas like policy analysis and regulatory oversight, where understanding the upstream causes under external constraints can support better strategic decisions.

We found that case-based learning is crucial for enhancing the causal reasoning abilities of models within these paradigms. Guiding models through specific examples helps them navigate complex causal chains and hidden dependencies, improving adaptability in real-world applications. This insight indicates that case-based causal learning can enhance model robustness and accuracy, especially in practical causal analysis, helping decision-makers achieve reliable causal insights in diverse scenarios.

4.2. Limitations

Our study on causal reasoning capabilities in LLMs has advanced understanding in this field, but several important limitations must be acknowledged. The research focused on a select group of prominent and accessible LLMs, including GPT-4 and LLaMA-70B, which, while representing advancements in language understanding, do not encompass the full spectrum of available models. Future studies should expand this scope to include emerging models and proprietary systems with limited public access, providing a more comprehensive evaluation of causal reasoning capabilities across diverse architectures. The study's scope was limited to three scenarios and expanded to three cases through three prompting methods. This limited scale may not fully capture the breadth of causal reasoning scenarios encountered in real-world applications. To address this, future research should aim to construct a large-scale dataset that systematically evaluates LLMs' causal understanding capabilities. Such a benchmark dataset should incorporate various causal graph structures described through diverse narrative contexts, covering logical reasoning, numerical computation, probability calculations, and common-sense reasoning. Our analysis primarily relied on qualitative assessments of model responses, which may introduce subjective biases in interpreting model capabilities. Developing more nuanced and objective metrics for evaluating causal reasoning in LLMs remains a critical area for future research. This could involve creating standardized benchmarks or developing automated evaluation techniques that can more accurately gauge the depth and accuracy of causal reasoning. While we attempted to simulate interventions within our scenarios, these simulations were inherently limited by the scenarios' design. The ability to generalize these findings to scenarios involving complex, real-world interventions remains constrained. Future work should explore more sophisticated interventional scenarios that better reflect the intricacies of real-world causal relationships. Finally, this study primarily focused on English-language models and general-purpose scenarios, limiting the generalizability of findings to models trained in other languages or specialized in specific domains. Cross-lingual and domain-specific evaluations of causal reasoning capabilities could provide valuable insights into the robustness and transferability of these skills across different contexts.

4.3. Future Works

Our future work will primarily consist of two main parts. The first part involves constructing a dataset for causal reasoning in large models based on the four paradigms we propose. We plan to extend the existing CoT and Cladder datasets to incorporate these paradigms and enrich the dataset with diverse sample types, particularly beyond mathematical computation and number games. We will use emerging models and proprietary systems to evaluate the dataset, aiming to further understand and gain insights into the causal reasoning capabilities of LLMs.

The second area of work focuses on enhancing LLM' causal reasoning capabilities. Two technical approaches are worth exploring. One approach is to build a causal reasoning agent by decomposing complex problems into causal analysis segments. By fine-tuning the LLM to deepen its understanding of common causal patterns, such as Confounding, Mediation, Collision, and Diamond structures, we aim to strengthen its causal analysis skills.

The other approach involves incorporating causal reasoning at the model encoding level. Causal graphs can essentially be linearized based on methods like topological sorting and transformed into conditional probability compositions using mathematical techniques. Given that conditional probabilities align fundamentally with the LLM framework, it is theoretically possible to introduce causal reasoning at the model's decoding layer, thereby enhancing the model's capacity for causal understanding and reasoning.

5. Conclusions

This study provides an evaluation of causal reasoning capabilities in LLMs through a series of carefully designed case studies and prompting methods. Our findings reveal variations in performance across different LLMs and prompting strategies, with GPT-4 consistently demonstrating superior causal understanding compared to other models. The research highlights that most LLMs struggle with complex causal reasoning tasks, particularly those involving interventions or counterintuitive statistical relationships like Simpson's Paradox. We observed that the effectiveness of different prompting methods (zero-shot, few-shot, and Chain of Thought) varies across LLMs, with some benefiting more from structured guidance than others. A concerning pattern emerged where many LLMs arrived at correct conclusions despite flawed reasoning processes, suggesting limitations in their underlying causal understanding. Interestingly, LLMs with stronger code comprehension capabilities tended to perform better in causal reasoning tasks, indicating a potential link between structured logical thinking and causal reasoning. These results underscore the need for continued research and development in enhancing LLMs' causal reasoning abilities.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/electronics13234584/s1>, Table S1: The Datasets of Casual Reasoning.

Author Contributions: Conceptualization, L.W. and Y.S.; methodology, L.W. and Y.S.; software, L.W.; validation, L.W.; formal analysis, L.W. and Y.S.; investigation, Y.S.; resources, L.W.; writing—original draft preparation, L.W.; writing—review and editing, Y.S.; visualization, L.W.; project administration, L.W.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program No. 301, the National Postdoctoral Science Foundation under grant No. 2021M690777, and the funding for postdoctoral research projects in Guangzhou.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it only involves research on the reasoning mechanisms of large language models and open-source datasets, with no involvement of humans or animals.

Data Availability Statement: In our work, we modified the COT dataset, transforming single reasoning into four types of causal reasoning patterns. The original dataset can be downloaded

from the following link https://github.com/kojima-takeshi188/zero_shot_cot/tree/main/dataset (accessed on 28 March 2023).

Conflicts of Interest: Author Lei Wang was employed by the company Guangzhou Intelligence Communications Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
2. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *OpenAI Blog* **2018**, 1–12.
3. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, 1, 9.
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
5. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
6. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv* **2022**, arXiv:2203.11171.
7. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 1–14.
8. Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N.R.; Kalchbrenner, N.; Goyal, A.; Bengio, Y. Toward causal representation learning. *Proc. IEEE* **2021**, *109*, 612–634. [[CrossRef](#)]
9. Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Lyu, Z.; Blin, K.; Gonzalez Adauto, F.; Kleiman-Weiner, M.; Sachan, M.; et al. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Adv. Neural Inf. Process. Syst.* **2024**, *36*. [[CrossRef](#)]
10. Li, Y.; Xu, M.; Miao, X.; Zhou, S.; Qian, T. Large language models as counterfactual generator: Strengths and weaknesses. *arXiv* **2023**, arXiv:2305.14791.
11. Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*; Basic Books: New York City, NY, USA, 2018.
12. Illari, P.; Russo, F. *Causality: Philosophical Theory Meets Scientific Practice*; OUP: Oxford, UK, 2014.
13. Feltz, B.; Missal, M.; Sims, A. *Free Will, Causality, and Neuroscience*; Brill: Leiden, The Netherlands, 2019.
14. Halpern, J.Y. *Actual Causality*; MIT Press: Cambridge, MA, USA, 2016.
15. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
16. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
17. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
18. Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; et al. InternLM2 Technical Report. *arXiv* **2024**, arXiv:2403.17297.
19. Madaan, A.; Zhou, S.; Alon, U.; Yang, Y.; Neubig, G. Language models of code are few-shot commonsense learners. *arXiv* **2022**, arXiv:2210.07128.
20. Zhang, L.; Dugan, L.; Xu, H.; Callison-Burch, C. Exploring the curious case of code prompts. *arXiv* **2023**, arXiv:2304.13250.
21. Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Alobeidli, H.; Cappelli, A.; Pannier, B.; Almazrouei, E.; Launay, J. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 79155–79172.
22. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; Casas, D.D.L.; Hanna, E.B.; Bressand, F.; et al. Mixtral of experts. *arXiv* **2024**, arXiv:2401.04088.
23. Roziere, B.; Gehring, J.; Gloeckle, F.; Sootla, S.; Gat, I.; Tan, X.E.; Adi, Y.; Liu, J.; Remez, T.; Rapin, J.; et al. Code llama: Open foundation models for code. *arXiv* **2023**, arXiv:2308.12950.
24. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open models based on gemini research and technology. *arXiv* **2024**, arXiv:2403.08295.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.