CausalLink: An Interactive Evaluation Framework for Causal Reasoning

Anonymous ACL submission

Abstract

We present CausalLink, an innovative evaluation framework that interactively assesses causal reasoning skills in conversational language models. Each CausalLink test case creates a hypothetical environment in which the language models are instructed to apply interventions to entities whose interactions follow predefined causal relations generated from controllable causal graphs. Our evaluation framework isolates causal capabilities from the confounding effects of world knowledge and semantic cues. We evaluate a series of LLMs in a scenario featuring movements of geometric shapes and discover that models start to exhibit reliable reasoning on two or three variables at the 14-billion-parameter scale. However, the performance of state-of-the-art models such as GPT4o degrades below random chance as the number of variables increases. We identify and analyze several key failure modes.

1 Introduction

Evaluating the causal reasoning abilities of generative AI has become a popular research area, especially given the advancements of modern LLMs(Zhang et al., 2023; Kıcıman et al., 2023; Cai et al., 2023; Liu et al., 2024). However, there are at least two major challenges with regards to the effectiveness of causal reasoning benchmarks: 1) clearly defining the targeted abilities and 2) disentangling reasoning processes from confounding factors such as data contamination and shortcuts. In this study, we tackle these two challenges and propose a dynamic evaluation framework in which the models are instructed to *discover* causal rules by interacting with hypothetical entities.

Although human causal reasoning has been systematically studied in various domains including computer science, psychology, and cognitive science(Goldvarg and Johnson-Laird, 2001; Gopnik et al., 2004; Pearl, 2009; Goddu and Gopnik, 2024), we still observe blurred lines among different facets of causal reasoning in the current AI literature, where the term “causal capabilities" may refer to a range of abilities from retrieving commonsense knowledge (Du et al., 2022; Frohberg and Binder, 2022; Srivastava et al., 2022) to multi-step structural inference (Jin et al., 2023).

We bifurcate causal reasoning skills into two general categories based on whether or not the reasoning process depends on existing world knowledge of causal facts (e.g., given a known causal relationship between X and Y such that X causally impacts Y). Fact-dependent abilities include effect retrieval, cause retrieval, and mechanism explanation. While all types of fact-dependent abilities fit under the general framework of causal reasoning, none of them requires a genuine understanding of causality. In other words, applying knowledge-based causal abilities requires no higher level of sophistication than superficial knowledge retrieval.

In contrast, fact-independent reasoning abilities represent the foundational mechanisms of causal reasoning, which do not rely on exhaustive knowledge of causal facts. These abilities enable humans to derive new causal insights, design experiments, and build the body of common knowledge. Humans develop foundational causal reasoning skills, such as reasoning about immediate context and actions, in the early stages of cognitive development prior to language acquisition (Goddu and Gopnik, 2024). As a results, LLMs may lack causal reasoning skills parallel to early-stage human causal reasoning through interactions, which serves as a foundation for more advanced reasoning processes(Goddu and Gopnik, 2024). In this paper, we define a causal capability named *action identification*, which entails identifying the correct intervention, observing the effects of its action, and reasoning about whether a causal relationship exists.

The key contributions of our work are threefold:

- We introduce a novel multi-round interactive evaluation framework for causal reasoning, designed to isolate the effects of grounded knowledge and semantic cues. This method is broadly applicable beyond the specific use case demonstrated in this paper. The framework and code is available at [REDACTED].
- We systematically evaluate and compare a series of LLMs, establishing our evaluation framework as a reliable benchmarking tool.
- We analyze failure modes in state-of-the-art models, identifying recurring "cognitive biases" that affect their causal reasoning.

2 Related Work

Current causal benchmarks usually evaluate models with cases developed from real-world causal scenarios (Du et al., 2022; Frohberg and Binder, 2022; Srivastava et al., 2022; Kıcıman et al., 2023; Jin et al., 2023; Wang, 2024). While real-world causal cases are undoubtedly effective in testing fact-dependent causal skills (Kıcıman et al., 2023), researchers must carefully mitigate the potential bias that causal claims can be made with knowledge recall rather than actual reasoning (Cai et al., 2023)). Previous mitigation included reversing the direction of causality to make causal relations counterfactual (Jin et al., 2023), using nonsensical descriptors to eliminate semantic cues (Jin et al., 2023), and questioning the model from multiple perspectives differentiating the directions of causality and the presence of interventions (Wang, 2024). Our solution is to construct a hypothetical world from scratch with underlying causal rules, which offers several distinctive advantages compared to previous work. First, the causal relations in the benchmark are systematically generated from causal graphs, allowing precise control over the difficulty of the task. Second, the causal entities are customizable and can be designed to be entirely free of linguistic or contextual clues that could otherwise create semantic shortcuts in reasoning. Finally, the interactive nature of the benchmark allows for the analysis of model strategies.

2.1 Causal Reasoning in LLMs

Chan et al. (2023) evaluated temporal, causal, and discourse relation tasks and found that ChatGPT achieved the best performance relative to fine-tuned SotA models in causal relation tasks specifically. This work again confirmed the models' ability to match commonsense knowledge patterns in causality-related tasks. By contrast, Jin et al. (2023) showed that even the most advanced GPT-4 model (OpenAI, 2023) struggles with the formal causal reasoning task, CLADDER. They proposed a tailored Chain-of-Thought prompt (Wei et al., 2022b) that marginally increases the overall accuracy from 64.28% to 66.64% (Jin et al., 2023). Liu et al. (2023) investigated causal reasoning abilities in code-based LLMs and reported that models leveraging code prompts — which explicitly encode conditional structures — exhibit superior performance in identifying causal relations. Jin et al. (2024) introduced the Corr2Cause benchmark to assess the ability of LLMs to infer causation from correlational data, showing that these models often perform near chance levels when faced with out-of-distribution examples. Finally, Chi et al. (2024) proposed the G2-Reasoner framework, which augments LLMs with external general knowledge and goal-driven prompts to elevate their reasoning from simple, fact-dependent associations (level-1) toward more robust, inference-driven capabilities (level-2). While LLMs show notable strengths in leveraging vast amounts of training data to recognize common causal patterns, significant gaps remain in achieving genuine, context-independent causal reasoning.

2.2 Interactive Evaluation

Advancements in conversational language models have paved the way for interactive evaluations, moving beyond the limitations of traditional static datasets. Prior to the era of LLMs, Kiela et al. (2021) identified the need for dynamic benchmarking to address the rapid saturation of model performance on static datasets. This need has become even more pronounced as models are increasingly exposed to vast amounts of training data. Platforms like Chatbot Arena (Chiang et al., 2024) have introduced effective evaluation methods by leveraging human preferences, where rankings emerge naturally through pairwise battles rather than relying on predefined ground-truth labels. Similarly, Hu et al. (2024) proposed GameArena, a framework that evaluates reasoning abilities through human-AI interactions constrained by gaming rules designed to test deductive and inductive reasoning.

Building on these approaches, we argue that dynamic benchmarks are inherently more effective and flexible; however, we aim to reduce reliance on human involvement. While some may view close-

form labels as a weakness (Chiang et al., 2024), we argue that they offer a clear and objective standard. Our framework introduces carefully designed programs as stand-ins for human evaluators, enabling robust interaction-driven assessments while maintaining scalability and consistency.

3 Framework Description

CausalLink evaluates causal reasoning in language models through interactive simulations grounded in formal causal graphs. The system comprises three integrated components: (1) a configurable causal graph generator that encodes ground-truth relationships, (2) a dynamic simulation environment where variables map to interactive entities, and (3) a language model interface that tests causal understanding through multi-step interventions. Each component is formalized as follows.

3.1 Causal Graph Construction

The foundational causal structure is implemented as a directed acyclic graph (DAG). We support three core structural paradigms: direct causation ($A \rightarrow B$), mediation ($A \rightarrow B \rightarrow C$), and confounder ($A \leftarrow B \rightarrow C$). Optional secondary edges ($A \rightarrow C$) are allowed in confounder structures, allowing exploration of both canonical and perturbed causal configurations. The perturbed confounder structure inherently contains the collider structure ($A \rightarrow B \leftarrow C$). We also allow randomly generated DAGs with any number of variables for test cases with varying difficulty. For random causal graphs, structural integrity is maintained through constrained edge generation. Causal connectivity between variables is algorithmically validated, which serves as ground truths in our evaluation. We implement graphical computation using the NetworkX (Hagberg et al., 2008) library.

3.2 Interactive Simulation Environment

In our experiments, we define a simulated environment called *ShapeWorld* where the abstract causal variables are represented as geometric shapes with dynamic states, moving or static. Given a causal graph $G = (E, V)$, for any edge $e \in E$ from $v_1 \in V$ to $v_2 \in V$, v_1 and v_2 represent two shapes s_1 and s_2 such that the movement of s_1 causes s_2 to move. Models can manipulate shapes by either moving them, thereby activating their causal descendants, or holding them in place, which prevents movement if no other causal factors remain. Causal effects propagate throughout the system according to the underlying graph structure, and deactivating an influence follows a backward-tracing process to verify dependencies and remove effects accordingly. Our implementation follows Markovian state transitions such that each intervention's effects depend solely on the current set of active elements and the causal graph structure.

While we use *ShapeWorld* as an example, our general framework can be extended to other simulated environments with different themes. We choose geometric shapes as causal entities because they have minimal semantic meaning, minimizing the risk of models relying on pretraining biases or external knowledge. This design choice allows us to isolate causal reasoning from knowledge grounding, ensuring that model performance reflects an understanding of causal relationships rather than memorized associations. To construct a new simulated environment within our framework, several key principles must be followed. First, each variable in the causal graph should correspond to a pair of an entity and its change. Despite our setup having only one type of change (movement), the system can incorporate multiple types of changes as long as the mapping between causal relationships and observed transformations is clear. Second, the environment must include a static or neutral state for entities, preserving the visibility of the underlying causal graph. In other words, it is necessary to maintain the possibility of removing the effects of a variable from the system. Finally, the system must define well-structured interventions that can reliably activate changes in an entity, ensuring that causal dependencies can be systematically tested.

3.3 Language Model Interaction Protocol

We evaluate LLMs through templated dialogues, requiring models to select shapes to intervene, interpret the feedback from the environment after each action, and conclude whether a specified causal relationship exists. The interaction process is illustrated in Figure 1. We present the general idea of the prompting process for the four phases shown in Figure 1 in this section and provide complete sets of prompts in the appendix.

Initialization: We present the settings of the hypothetical world of shapes and specify the rules of the task. We give the models the initial states of each shape in the system and propose a question that asks whether the movement of one shape causes the movement of another.

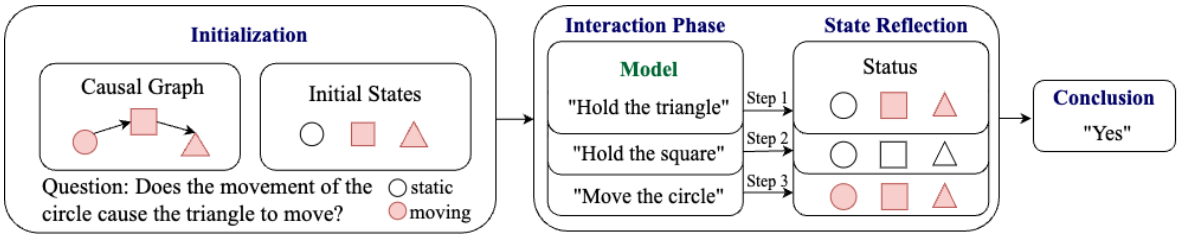

Figure 1: Illustration of the interaction process between our CausalLink system and a language model. This figure demonstrates an example test case where the model hypothetically correctly solves the task.

Model	Acc. (F)	Acc. (T)	Acc.	Avg. Steps	Err.
Llama 3.2 3B(AI@Meta, 2024b)	0.766	0.622	0.702	1.88	5
Llama 3.1 8B (AI@Meta, 2024b)	0.979	0.189	0.631	3.35	0
Mistral 7B(*) (Jiang et al., 2023)	0.957	0.216	0.631	1.02	3
Qwen2.5 3B (Yang et al., 2024)	0.894	0.297	0.631	1.00	0
GPT-4o-mini (OpenAI, 2024a)	1.000	0.243	0.666	1.54	0
Qwen2.5 14B (Yang et al., 2024)	0.936	0.622	0.797	2.23	1
DS-Distill Qwen 14B(DeepSeek-AI, 2025)	0.826	0.789	0.809	2.02	4
Qwen2.5 32B (*) (Yang et al., 2024)	0.979	0.784	0.893	1.68	1
DS-Distill Qwen 32B (*) (DeepSeek-AI, 2025)	0.979	0.865	0.929	1.63	0
Llama 3.1 Nemotron 70B(*) (Wang et al., 2024)	**1.000**	0.892	**0.952**	2.40	0
Gemini 2.0 Flash (Mallick and Kilparick, 2025)	0.787	**0.919**	0.845	1.40	0
GPT-4o(*) (Hurst et al., 2024)	0.915	0.892	0.904	1.57	0

Table 1: Model Performances on the core set. Acc. (F) and Acc. (T) refer to accuracy scores on test cases with False and True ground truth labels respectively. Err. refers to the number of errors due to invalid formats, invalid answers, or invalid actions. Models with (*) perform better with rate-limiting instruction and models without (*) perform better without rate-limiting instruction. DS-Distill is a shorthand form of "DeepSeek Distilled".

> You are in a world of shapes. The movements of shapes follow internal causal rules. You are required to interact with the shapes until you can answer a question about the causal rules. All changes in the world are deterministic and consistent. There is no hidden confounder. You can either 1) move a static shape or 2) hold a moving shape. A shape only stops moving when there are no other causes of its movement.
> Following are your current observations: (initial states of shapes)
> Please interact with the shapes to answer: Does triangle moving cause square to move?

Intervention Phase: We ask the model to propose JSON-formatted {shape, action} pairs, which we use to apply interventions to our system.

> Please propose your interaction. Please provide your response by filling the JSON:
> {"shape":"", "action":""}

State Reflection: We feed the model with post-intervention state updates and request the model to either continue interaction or answer the question.

> Following your last action, the current states of shapes are: (current states of shapes)
> Based on the results you observed so far, please decide to continue the interaction or answer the question.

Conclusion: We conclude a test case when the model is ready to answer the question. The model delivers a final yes/no judgment.

> You are ready to answer the question: (question)

3.4 Experiment Setup

Our experiments are divided into two primary components. The first, referred to as the core set, comprises only direct (two-variable), mediation (three-variable), and confounder (three-variable) causal structures. The second, the advanced set, features randomly generated causal graphs that include more than three variables. We use the core set as a comprehensive test of basic causal structures. For each causal structure in the core set, we systematically generate all possible initial configurations of active nodes and formulate pairwise

causal queries (i.e., $A \rightarrow B$ and $B \rightarrow A$) for every pair of shapes. To create a comprehensive set of initial setups, we enumerate all combinations of nodes, simulate the cascade of causal effects based on the underlying graph, and eliminate redundant configurations. This results in 84 test cases for the core set.

The advanced set is designed to simulate increasingly complex problems. We rely on randomized experiments to increase the likelihood of capturing the most challenging cases. Given the rapid growth of combinatorial possibilities, it is infeasible to exhaustively test all configurations of experimental setups and cause-effect pairs. Therefore, for the advanced set, we restrict our analysis to the "all-active" setup, where all shapes are in motion, and we randomly sample six pairs of variables for each generated graph. This method balances computational feasibility with sufficient complexity to evaluate model performance on more difficult causal inference tasks. We generate 50 random graphs with 50% connectivity for 4 to 7 variables, resulting in 1200 test cases for the advanced set.

For a test case with n variables, the model is allowed up to $2n$ intervention steps, after which it is considered to have failed due to timeouts. We assess model performance using three key metrics: **accuracy**, defined as the proportion of correct causal judgments relative to the ground-truth graph; **efficiency**, measured as the mean number of steps required to reach a final judgment; and **robustness**, evaluated based on the frequency of invalid actions, format errors, and timeouts. We run the experiments twice with two prompting strategies: the basic prompt and one that specifically instructs the models to reach the conclusion in the fewest steps possible. We limit prompt engineering to avoid conflating the evaluation of reasoning ability with instruction-following.

In our experiments, we capture 3 error modes: **invalid action** (where the model attempts to choose action-shape pairs outside of valid settings), **invalid format** (where the model fails to follow the instructed format), and **invalid answer** (where the model answers neither yes nor no).

4 Experiment Results

4.1 Core Set Performance

We run experiments on both locally deployed open-source models using the HuggingFace Transformer (Wolf et al., 2020) library and OpenAI GPT models and Gemini 2.0 Flash through API calls. We report the better performance out of the two prompting strategies in Table 1 and present the complete sets of results in Appendix E.

The results indicate that causal reasoning on our core set of test cases aligns with the pattern of emergent abilities (Wei et al., 2022a), with reasoning skills generally appearing at scales of 14 billion parameters and above. Smaller models except Llama 3.2 3B exhibit a strong bias toward concluding that no causal relationship exists, achieving a maximum of only 29.7% (GPT-4o-mini) accuracy rate of positive cases. Llama 3.2 3B generates relatively balanced outputs but still underperforms with 70.2% accuracy. Llama 3.1 Nemotron 70B outperforms other models, including GPT-4o, achieving 95.2% accuracy. Additionally, providing an instruction to reach the conclusion as quickly as possible generally benefits larger models but negatively impacts smaller ones. We observe that this instruction limits the generation of error-prone and sometimes contradicting rationales in larger models, allowing them to reason more accurately with fewer interactions. This phenomenon is particularly pronounced in the DeepSeek-distilled Qwen 2.5 32B model, which shows a remarkable 30.2% performance improvement when the instruction is applied. One special case is the best-performing Llama 3.1 Nemotron model whose accuracy and the average number of steps both increase with the step-limiting instruction. Figure 2 shows side-by-side comparisons of model performances across the three causal structures. Models achieving more than 80% overall accuracy can perfectly solve all of the two-variable cases, with which the smaller models struggle. As shown by the error bars in Figure 2, we also observe greater variability in performance across different initial setups in the mediation structure compared to the confounder structure.

Although large models achieve seemingly strong performance, we argue that the core set is intentionally designed to be fundamental and straightforward. Any failure on these tasks suggests gaps in the action identification skill we aim to evaluate. We will explore failure cases further in the following sections.

4.2 Advanced Set Performance

Due to practical constraints on computational resources, we select GPT4o and Gemini 2.0 Flash as the test models for the advanced set study because of their representative performance on the

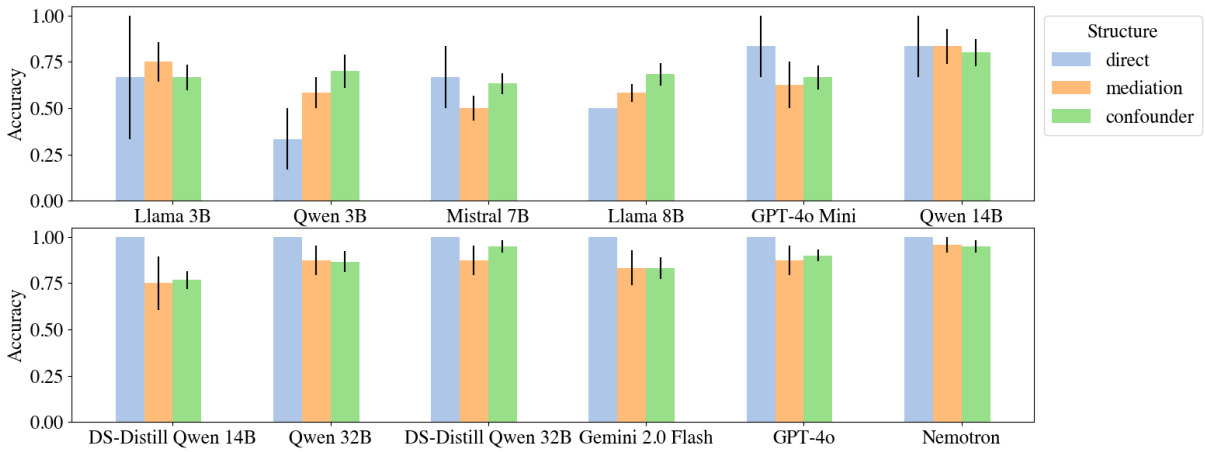

Figure 2: Comparison of model performances by causal structures. Similar to Table 1, this graph only showcases the better performance out of the two prompting strategies for each model. We recognize perfectly solving cases with direct causal structures as the indicator of basic causal capability. Error bars indicate the standard error of accuracy with respect to initial setups.

core set and their widespread popularity. Figure 3 shows the performance of the two models across the core and the advanced sets grouped by the number of variables. We observe a clear pattern that the model's performance degrades as the number of variables increases.

Although the number of variables increases, the fundamental reasoning processes required to solve the task remain unchanged. For humans, the increased difficulty may stem primarily from the cognitive demand of managing more information (Sweller, 2011) rather than requiring more sophisticated reasoning skills. The observed decline in models' performance as the number of variables increases may suggest a lack of genuine causal reasoning. This deficiency is less apparent in the simpler core set but becomes more evident when the complexity of the problem increases.

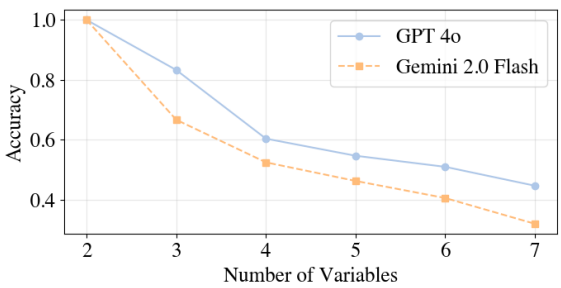

Figure 3: Performance of GPT4o and Gemini 2.0 Flash on *ShapeWorld* with increasing number of variables. We compare the average accuracy score among the "all-active" initial setups for fairness.

5 Discussion

5.1 Observed Abilities in Simple Settings

In our study, models at the 14B scale and above exhibit a basic understanding of causal intervention. When prompted with a question, the models reliably select the cause variable to intervene and observe the corresponding effect. They also demonstrate the capability of identifying potential confounding variables in simple settings; for example, when the effect variable is already moving in the initial setup, the models may attempt to halt its movement by holding a shape that is neither the cause nor the effect variable. Additionally, these models are generally efficient at solving the problem, rarely engaging in unnecessary or repeated interaction steps.

5.2 Observed Weaknesses in Complex Settings

Despite demonstrating basic causal reasoning abilities in simpler scenarios, state-of-the-art models do not scale well to more complex setups. Performance declines sharply as the number of variables increases, dropping below random chance when the variable count exceeds six, highlighting a clear gap in causal reasoning capabilities. While models are generally efficient, they sometimes fail by prematurely jumping to conclusions without sufficient evidence (see section 5.3 for examples). Additionally, even when all necessary observations are available, models can misinterpret causal relationships and arrive at incorrect conclusions. Notably, our experimental setup is already a highly distilled

simulation of real-world causal reasoning, reducing complex interactions to the movements of abstract shapes. Furthermore, we design interactions such that the underlying causal structure is fully observable through interventions, whereas real-world scenarios often present much greater ambiguity. The fact that models struggle even under these controlled conditions underscores the limitations of their causal reasoning abilities.

5.3 "Cognitive" Bias in Failure Cases

We now present case studies on model failure modes. To ensure the validity of our analysis, we focus only on recurring error patterns in models that perform well on the core set. Table 2 shows the prevalence of each type of failure mode and how we identify each patter.

Our findings suggest that these "cognitive biases" stem not from a lack of causality-related knowledge (e.g., confounding variables) but rather from its misapplication. While models often generate rationales that include correct principles—such as "to conclude causality, I need to isolate the effects"—their actual behaviour does not always align with their stated reasoning.

In the following examples, (m) indicates a shape is moving and (s) indicates a shape is static.

Root Cause Bias As discussed in Section 4.1, even the best-performing models may struggle with the simple mediator structure. Given the causal structure $A \rightarrow B \rightarrow C$, where B mediates the effect of A on C, it is important to recognize both A and B as the cause of C. This concept is crucial in front-door adjustment, a key technique in causal inference (Pearl, 2009). However, we observe a prevalent pattern that models incorrectly disregard the mediate as a potential cause as soon as they figure out the root cause. Models tend to attribute movements of the shapes to be systematically dependent on the root cause and do not attempt to further investigate other internal interactions. Despite explicit instructions in the prompt allowing for multiple causes, models remain biased toward the false assumption that only a single cause is responsible for an effect.

Does the square moving cause the circle to move?
Setup: triangle (m); square (m); circle (m)
[hold square]
triangle (m); square (m); circle (m)
[hold triangle]
triangle (s); square (s); circle (s)
Model (Nemotron) Answer: no

Correlation Bias Language models may struggle to differentiate correlation from causation. When two variables exhibit the same behaviour across multiple actions, models tend to infer a causal relationship as soon as their states change together. In such cases, models may disregard the direction of causality entirely or the existence of a confounder.

Does the octagon moving cause the triangle to move?
Setup: triangle (m); octagon (m);
rectangle (m); circle (m)
[hold octagon]
triangle (m); octagon (m);
rectangle (m); circle (m)
[hold rectangle]
triangle (m); octagon (m);
rectangle (m); circle (m)
[hold circle]
triangle (m); octagon (m);
rectangle (m); circle (m)
[hold triangle]
triangle (s); octagon (s);
rectangle (s); circle (s)
Model (GPT4o) Answer: yes

Another form of correlation bias occurs when the effect shape is already moving in the initial state. In such cases, the model moves the supposed cause shape and infers causality when it observes both shapes in motion.

Does the circle moving cause the triangle to move?
Setup: triangle (m); square (s); circle (s)
[move circle]
triangle (m); square (s); circle (m)
Model (GPT4o) Answer: yes

Interestingly, this type of correlation bias does not appear in the direct causal structure. One possible explanation is that the presence of a static third shape (e.g., the square) leads the model to assume that confounding factors are controlled. This assumption may then reinforce its incorrect inference of causality.

Illusive Confounder Bias The illusive confounder bias complements the correlation bias such that the model refuses to identify a positive causal relationship due to the potential existence of confounders even if there is evidence against it. In the following example, holding the triangle effectively eliminates the movement of the square as a potential confounder. However, the model mistakenly concludes that the square may be an intermediary factor, confusing mediators with confounders, and denies the causal relationship.

Failure Mode	Percentage	Result	Characteristic Pattern
Root Cause Bias	42.6%	False negative	all shapes become static
Correlation Bias	14.8%	False positive	two shapes in question act the same
Illusive Confounder Bias	11.1%	False negative	all necessary evidence present
Reverse Collider Bias	11.1%	False negative	two shapes in question not in sync

Table 2: Prevalence of failure modes by percentage of occurrences among 55 failed cases on the core set (using both prompting strategies) from the three best-performing models (Nemotron, DS-distill Qwen 32B, and GPT4o). Note that 20% of the failure cases are not categorized due to variations in failure patterns.

Does the triangle moving cause the circle to move?
Setup: triangle (m); square (m); circle (m)
[hold triangle]
triangle (s); square (s); circle (s)
Model (DeepSeek Distilled Qwen 2.5 32B) Answer: no

Reverse Collider Bias In causal inference, the collider structure is characterized by two cause variables (A and B) influencing the same effect variable (C) (Pearl, 2009). A collider bias refers to the false positive claim of causality where the cause and effect variables (A and B) in question both influence a third common variable (C) that is controlled due to problematic experimental design (Pearl, 2009; Holmberg and Andersen, 2022). We observe a related but different pattern in language models where the model controls A and concludes A does not cause C because there is another variable (unidentified B) that also causes C.

Does the square moving cause the hexagon to move?
Setup: square (m); ellipse (m); hexagon (m); circle (m)
[hold square]
Setup: square (s); ellipse (m); hexagon (m); circle (m)
Model (GPT4o) Answer: no

6 Conclusion

In this paper, we introduced CausalLink, a novel interactive evaluation framework that rigorously assesses a fact-independent causal reasoning skill that we term "action identification" in LLMs. By constructing a controlled, simulated environment with predefined causal relationships, we effectively isolate genuine reasoning from the confounding influences of world knowledge and semantic cues. This approach not only enables precise measurement of causal reasoning abilities but also offers a generalizable methodology for a wide range of experimental designs.

Our empirical evaluations reveal that, although larger models demonstrate foundational causal reasoning skills, their performance becomes increasingly fragile as the complexity of causal interactions grows. Importantly, we identify recurring cognitive biases—including single cause bias, correlation bias, and illusive confounder bias. These underscore a critical gap: models *misapply* their causal knowledge rather than lack it outright. These discrepancies between the models' articulated reasoning and their actual behaviour highlight the limitations of current approaches in achieving robust, context-independent causal reasoning.

By establishing a new benchmark for causal inference, our study underscores the need for improved methodologies that enhance both the reliability and generalizability of causal reasoning in AI systems. Future directions include mitigating model biases and extending the framework to evaluate more aspects of causal reasoning.

7 Limitations

7.1 Knowledge-agnostic causal reasoning

Disentangling grounded knowledge from the reasoning process remains a challenging and important task that helps assess whether models can generalize causal reasoning to novel scenarios without being biased by encoded knowledge. While we strive to achieve this, we acknowledge several limitations in our current approach.

First, our synthetic environment does not fully capture the complexity of real-world causal structures. The experimental setup employs symbolic representations for entities which, while effective in controlling for semantic cues, lacks inherent real-world meaning. While this design choice minimizes confounding factors related to knowledge recall, it may also alter model behaviour in unintended ways. Future research should further explore whether models rely on semantic information for causal reasoning and how best to introduce fine-grained controls to separate genuine reasoning

from implicit knowledge recall.

Second, our system enables fully automated interactions, requiring human effort only in the initial design of a hypothetical world, the naming of entities, and the identification of associated changes. While this allows for the efficient generation of large-scale test cases, the structured nature of these cases may lead to overly rigid evaluations. We believe that interactive benchmarks should become the standard for evaluating language models' causal reasoning abilities. However, further studies are needed to determine the optimal balance between efficiency and flexibility in such benchmarking systems.

7.2 Causal structures and test difficulty

Our experimental setup relies on randomly generated causal graphs, which entails statistical soundness but limits our ability to precisely control the causal structures that models encounter. Carefully designed complex causal graphs may yield new insights into model performance.

Additionally, test case difficulty does not always scale with the number of variables. For example, if the initial setup consists of entirely static shapes, the correct solution remains the same (acting on the cause shape and observing the effect shape) regardless of the total number of shapes present. To address this, we adopt an "all-active" setup, where all entities are subject to potential changes. While this effectively increases task difficulty as a function of the number of variables, it also reduces our ability to precisely manipulate test complexity.

Furthermore, our current design, which distinguishes only between movement and static states, represents a simplified model of causality. Introducing additional actions, changes, and interactions could enhance the challenge for models even in cases with a limited number of entities.

7.3 Model Performance

To ensure that our evaluation focuses on reasoning rather than instruction-following capabilities, we intentionally limit the level of prompt engineering in our experiments. This design choice allows us to assess models' ability to perform causal reasoning without relying on explicitly structured guidance. However, we acknowledge that models may perform better than reported if provided with detailed, step-by-step instructions on how to approach each test case. Future work could explore the extent to which prompt engineering influences causal reasoning performance.

Additionally, our observations and conclusions are restricted to the models tested in this study. Due to constraints on computational resources and access to proprietary models, we do not present exhaustive results across all available large language models. While our findings provide valuable insights, broader generalization to other models remains an open question. Future research should aim to expand coverage across a wider range of models and architectures to obtain a more comprehensive understanding of causal reasoning capabilities.

References

AI@Meta. 2024a. Llama 3 License. Accessed: 2025-02-10.

AI@Meta. 2024b. Llama 3 model card.

Apache. 2004. Apache License, Version 2.0. Accessed: 2025-02-10.

Hengrui Cai, Shengjie Liu, and Rui Song. 2023. Is knowledge all large language models needed for causal reasoning? *arXiv preprint arXiv:2401.00139*.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. ChatGPT evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Cloud@Alibaba. 2024. Qwen research license agreement. Accessed: 2025-02-10.

DeepSeek. 2023. DeepSeek MIT License. Accessed: 2025-02-10.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), page 432–446, Dublin, Ireland. Association for Computational Linguistics.

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France. European Language Resources Association.

Mariel K. Goddu and Alison Gopnik. 2024. The development of human causal learning and reasoning. *Nature Reviews Psychology*, 3(5):319–339.

Eugenia Goldvarg and Philip N Johnson-Laird. 2001. Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive science*, 25(4):565–610.

Google. 2021. Google APIs Terms of Service. Accessed: 2025-02-10.

Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. 2004. A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, 111(1):3.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Mathias J Holmberg and Lars W Andersen. 2022. Collider bias. *Jama*, 327(13):1282–1283.

Lanxiang Hu, Qiyu Li, Anze Xie, Nan Jiang, Ion Stoica, Haojian Jin, and Hao Zhang. 2024. GameArena: Evaluating LLM reasoning through live computer games. (arXiv:2412.06394). ArXiv:2412.06394.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLADDER: Assessing causal reasoning in language models.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. (arXiv:2305.00050). ArXiv:2305.00050 [cs, stat].

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023. The magic of IF: Investigating causal reasoning abilities in large language models of code. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9009–9022, Toronto, Canada. Association for Computational Linguistics.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. 2024. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*.

Shrestha Basu Mallick and Logan Kilparick. 2025. Gemini 2.0: Flash, Flash-Lite and Pro.

NVIDIA. 2024. NVIDIA AI Foundation Models Community License Agreement. Accessed: 2025-02-10.

OpenAI. 2023. GPT-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2024a. gpt-4o-mini-advancing-cost-efficient-intelligence.

OpenAI. 2024b. OpenAI Terms of Use. Accessed: 2025-02-10.

Judea Pearl. 2009. *Causality*. Cambridge University Press.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

John Sweller. 2011. Cognitive load theory.

Zeyu Wang. 2024. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, page 143–151, Bangkok, Thailand. Association for Computational Linguistics.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. HelpSteer2-Preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

A Prompts

We follow a step-by-step prompting structure (see A.1) where the prompts (see A.2) are categorized as **system**, **initial**, **choice**, **interaction**, and **answer**. At each step, prompts are combined and formatted with generated strings describing questions or current states of shapes.

We only apply the system role where applicable, otherwise all instructions are given through the user role.

A.1 Step-by-Step Prompting Structure

1. **Initial Setup: System** prompt + **Initial** prompt
2. **Choice:** Updates on shape status +**Choice** prompt
3. **Interaction: Interaction** prompt
4. **Answer: Answer** prompt

The program starts the conversation by presenting the initial setups and then loops between choice and interaction until the model chooses to answer the question.

A.2 Prompts in Experiments

System

You are in a world of shapes. The movements of shapes follow internal causal rules. You are required to interact with the shapes until you can answer a question about the causal rules. All changes in the world are deterministic and consistent. There is no hidden confounder. *Please reach the conclusion in the least number of steps possible(only for the step-limiting prompting strategy).*

You can either 1) move a static shape or 2) hold a moving shape. A shape only stops moving when there is no other causes of its movement.

Initial

Following are your current observations: {}

Please interact with the shapes to answer: {}

Please propose your first interaction. Please provide your response by filling the JSON below:

- The value to "shape" field must be one of the listed shapes: {}

- The value to "action" field must be one of the listed actions: {}

{"shape":"","action":""}

Choice

Based on the results you observe so far, please decide to continue interaction or answer the question: {}.

Please provide your response by filling JSON below:

- The value to "next" field must be either "continue interaction" or "answer the question"

{"next":""}

Interaction

Please propose your next interaction. Please provide your response by filling the JSON below:

- The value to "shape" field must be one of the listed shapes: {}

- The value to "action" field must be one of the listed actions: {}

{"shape":"","action":""}

Answer You are ready to answer the question: {}

Please answer the question by filling the JSON below.

- The value to "answer" field must be "yes" or "no"

{"answer":""}

B Experiment Details

For all of the models in our experiments, we use the original configuration of hyperparameters released

with the models. We stick to any recommended setting (for example, temperature $= 0.6$ for DeepSeek Distilled models) provided by the models' authors. Details of each model are linked in Table 3. The models are instructed by the prompts to answer in JSON formats, but we also allow an output length of up to 2048 characters to accommodate any reasoning processes models may generate.

We run experiments once using the setups described in Section 3.4. We implement the chat-style interface with HuggingFace's `text-generation` pipeline[1], OpenAI's `chat completions` APIs[2], and Google GenAI's `chat` APIs [3]. Runtimes of the experiment vary depending on the sizes of the model, ranging from approximately 1 hour to 4 hours. GPUs used in the experiments are specified in Table 3.

C The Use and Release of Scientific Artifacts

C.1 Models, Licenses, and Hardware

Model cards, licenses, and GPU hardware used to run each model are listed in table 3. The OpenAI models we use in the experiments are `gpt-4o-2024-08-06` and `gpt-4o-mini-2024-07-18`. Our use of the models is consistent with their intended uses as specified in the licenses and terms of use.

C.2 Release of Artifact

Code for CausalLink is released under the MIT License. Due to the interactive nature of our evaluation framework, we do not produce any datasets as an artifact.

D Use of AI Assistants

Generative AI assistants are used to polish original content and identify relevant literature. The authors check, review, and edit any generated content or suggested references to ensure accuracy. We do not use generative AI for new ideas.

For coding, we use AI assistants to help with non-novel components (including regular expressions, statistics computation, and plotting).

E Proximal Experiment Results

Experiment results using basic and step-limiting prompting strategies are listed in Table 4 and Table 5.

[1] HuggingFace Text Generation Pipeline
[2] OpenAI Text Generation
[3] Google genai text generation

Model Card (linked)	License	GPU used
Llama 3.2 3B	(AI@Meta, 2024a)	1 A40
Llama 3.1 8B	(AI@Meta, 2024a)	1 A40
Mistral 7B	(Apache, 2004)	1 A40
Qwen 2.5 3B	(Cloud@Alibaba, 2024)	1 A40
Qwen 2.5 14B	(Cloud@Alibaba, 2024)	1 A40
Qwen 2.5 32B	(Cloud@Alibaba, 2024)	1 A40
DS-Distill Qwen 14 B	(DeepSeek, 2023)	1 A40
DS-Distill Qwen 32B	(DeepSeek, 2023)	2 A40
Llama 3.1 Nemotron 70B	(NVIDIA, 2024)	4 A40
GPT4o	(OpenAI, 2024b)	-
GPT4o mini	(OpenAI, 2024b)	-
Gemini 2.0 Flash	(Google, 2021)	-

Table 3: Models and GPU hardware.

Model	Acc. (F)	Acc. (T)	Overall Acc.	Avg. Steps	Err. Count
Llama 3.2 3B	0.766	0.622	0.702	1.88	5
Llama 3.1 8B	0.979	0.189	0.631	3.35	0
Mistral 7B	0.915	0.189	0.595	1.27	7
Qwen2.5 3B	0.894	0.297	0.631	1.00	0
Qwen2.5 14B	0.936	0.622	0.797	2.23	1
DeepSeek Distill Qwen 14B	0.826	0.789	0.809	2.02	4
Qwen2.5 32B	0.957	0.703	0.845	2.75	3
DeepSeek Distill Qwen 32B	0.809	0.432	0.642	1.69	26
GPT-4o-mini	1.000	0.243	0.666	1.54	0
GPT-4o	0.915	0.838	0.881	1.59	0
Llama 3.1 Nemotron 70B	1.000	0.784	0.905	2.29	0
Gemini 2.0 Flash	0.787	0.919	0.845	1.40	0

Table 4: Model Performance (basic template)

Model	Acc. (F)	Acc. (T)	Overall Acc.	Avg. Steps	Err. Count
Llama 3.2 3B	0.766	0.622	0.702	1.88	5
Llama 3.1 8B	0.936	0.135	0.583	3.29	1
Mistral 7B	0.957	0.216	0.631	1.02	3
Qwen2.5 3B	0.894	0.297	0.631	1.00	0
Qwen2.5 14B	1.000	0.541	0.798	1.99	0
DeepSeek Distill Qwen 14B	0.851	0.703	0.786	2.01	7
Qwen2.5 32B	0.979	0.784	0.893	1.68	1
DeepSeek Distill Qwen 32B	0.979	0.865	0.929	1.63	0
GPT-4o-mini	0.979	0.189	0.631	1.58	0
GPT-4o	0.915	0.892	0.905	1.57	0
Llama 3.1 Nemotron 70B	1.000	0.892	0.952	2.40	0
Gemini 2.0 Flash	0.723	0.892	0.798	1.42	0

Table 5: Model Performance (Step-limiting template)