CLEAR: Can Language Models Really Understand Causal Graphs?

Sirui Chen[*1], **Mengying Xu**[2], **Kun Wang**[3],
Xingyu Zeng[3], **Rui Zhao**[3], **Shengjie Zhao**[1], **Chaochao Lu**[†4]

[1]Tongji University, [2]Independent Researcher,
[3]The Chinese University of Hong Kong, [4]Shanghai Artificial Intelligence Laboratory
chensirui@pjlab.org.cn, xumy13@tsinghua.org.cn,
{wangkun, 1155023462, rzhao}@link.cuhk.edu.hk, luchaochao@pjlab.org.cn

Abstract

Causal reasoning is a cornerstone of how humans interpret the world. To model and reason about causality, causal graphs offer a concise yet effective solution. Given the impressive advancements in language models, a crucial question arises: can they really understand causal graphs? To this end, we pioneer an investigation into language models' understanding of causal graphs. Specifically, we develop a framework to define causal graph understanding, by assessing language models' behaviors through four practical criteria derived from diverse disciplines (e.g., philosophy and psychology). We then develop CLEAR, a novel benchmark that defines three complexity levels and encompasses 20 causal graph-based tasks across these levels. Finally, based on our framework and benchmark, we conduct extensive experiments on six leading language models and summarize five empirical findings. Our results indicate that while language models demonstrate a preliminary understanding of causal graphs, significant potential for improvement remains. Our project website is at https://github.com/OpenCausaLab/CLEAR.

1 Introduction

Causal reasoning is fundamental to how humans understand the world and solve challenges (Sloman and Sloman, 2009). The ability to reason causally allows us to explain phenomenon and predict the future (Woodward, 2005; Pearl, 2009; Bunge, 2017). There are various causal models used to investigate and represent causation, including mathematical equations, logical statements, and causal graphs (Pearl and Mackenzie, 2018). Among them, causal graph gains widespread adoption due to its intuitive and concise representation of complex causal relationships (Pearl, 1995, 1998).

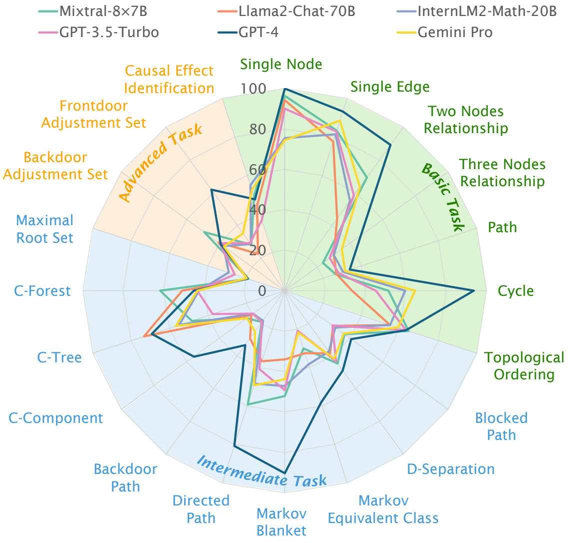


Figure 1: **Performance of six leading language models across 20 diverse tasks in CLEAR.** Further details on the experimental results can be found in Section 4. We use distinct colors to represent different levels.

A causal graph is essentially a bayesian network where each node represents a variable, and the directed edges denote definite or possible causal relationships between variables (Helmert, 2004). Understanding causal graphs is essential, as it enables us to grasp the relationships between variables (Kocaoglu et al., 2017). Furthermore, causal graphs can be leveraged for probability calculation (Kleinberg, 2013), providing solutions to problems across all three rungs of the *ladder of causation* (i.e., association, intervention, and counterfactuals) (Pearl and Mackenzie, 2018). With the rapid advancement of language models, there has been a surge in research exploring their ability to solve graph-related problems (Zhang et al., 2023b; Chai et al., 2023; Fatemi et al., 2023; Ye et al., 2023; Zhang et al., 2023a; Besta et al., 2024; Chen et al., 2024c; Wang et al., 2024a; Luo et al., 2024). In contrast to the abundant research on general graph prob-

[*] Work done when interning at Shanghai AI Laboratory.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 6247–6265
November 12-16, 2024 ©2024 Association for Computational Linguistics

lems, the ability of language models to understand causal graphs is yet to be investigated. Therefore, this paper aims to shed light on the question: *Can language models really understand causal graphs?*

Addressing this question poses three major challenges: (1) What does it mean for a model to understand causal graphs? (2) How to design a causal graph-based benchmark that can measure a model's understanding? (3) How to quantify a model's understanding when presented with causal graphs?

In this work, we first propose a framework to evaluate language models' *understanding* of causal graphs, by establishing four criteria: performance exceeding random guesses, robustness against question types, correct utilization of causal definitions, and performance constrained by task dependence. These criteria draw on insights from machine learning, philosophy, and psychology, providing a multidisciplinary approach to evaluating the comprehension of causal graphs by language models. Next, we construct the CLEAR, a novel benchmark created specifically for evaluating how well language models understand causal graphs. Finally, guided by our proposed framework of *understanding* in causal graphs, we systematically evaluate models' performance on CLEAR across all four criteria. To ensure a diverse evaluation, we select six leading models and utilize four prompts (e.g., in-context learning (IcL) (Brown et al., 2020)). Our extensive experiments yield the following key findings:

1. The model's ability to handle different causal graph-based tasks is uneven, exhibiting notable weaknesses in specific areas (Figure 1).
2. Language models have a preliminary understanding of causal graphs (Figure 5), and are observed to focus on key information required to deduce the correct answer (Figure 10).
3. Model performance is sensitive to the question type. A model's understanding of causal graphs might be artificially inflated if evaluation relies on limited types (Figure 6).
4. Models exhibit a capacity for utilizing both explicit and implicit concepts related to causal graphs, and their proficiency with these concepts varies considerably (Figure 7).
5. The performance of most models is not constrained by task dependency (i.e., although Task B depends on Task A, performance on Task B often exceeds that on Task A), showcasing a notable divergence in their performance trends. This might suggest heterogeneity in knowledge representation and application across different models (Figure 8).

Overall, we make four main contributions:

- We make, to the best of our knowledge, the first-ever attempt to evaluate language models' capacity for understanding causal graphs.
- We propose a framework for measuring a model's understanding of causal graphs by defining four specific criteria.
- We construct CLEAR, the first benchmark designed specifically to assess language models' understanding of causal graphs. CLEAR features three levels, encompasses 20 causal tasks, and considers six question types.
- Extensive experiments with six leading language models yield insightful findings and valuable observations about their capacity for understanding causal graphs.

2 What Do We Mean by *Understanding* in Language Models?

2.1 Multiple Facets of *Understanding*

Unlocking the mysteries of human social behavior (Adler et al., 2006), decision-making (Frensch and Funke, 2014), and personality development (Lapsley et al., 2004) hinges on our ability of *understanding*. Our investigation into *understanding* begins with a brief summary of the existing definitions across various disciplines.

From the philosophical and psychological perspectives, *understanding* means: (1) More than just knowing isolated facts. It involves recognizing and grasping the relationships that weave together the various elements of a subject (Kvanvig, 2003; Carter and Gordon, 2014; Grimm, 2021). (2) Beyond the formula or definition. It encompasses the ability to not only grasp concepts or formulas but also to adeptly apply them in practical contexts (Rumelhart, 1991; De Regt, 2004). (3) Variation in degree. Understanding is not binary, its completeness depends on the individual's conceptual context and background knowledge (Nickerson, 1985).

Considering recent machine learning endeavors, Choudhury et al. (2022) propose three criteria to assess if a reading comprehension model reaches human-level ability. They focus on whether a model could solve problems correctly, whether it uses information that humans would deem relevant, and whether its performance is consistently robust. Although the three conditions provided in Choud-

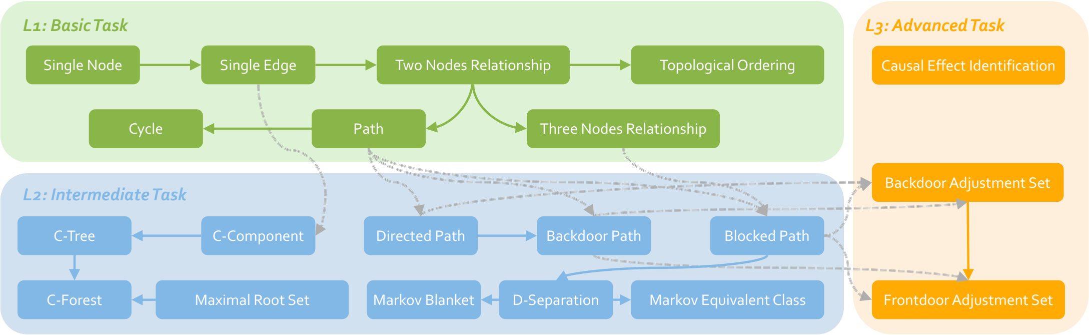


Figure 2: **Hierarchy and dependent relationships of tasks in CLEAR.** We define three complexity levels, the three-level definition is novel and tailored specifically for this benchmark. (1) Level 1: *Basic Task*. Mastering these concepts is a prerequisite for understanding any general graph. (2) Level 2: *Intermediate Task*. These tasks represent the most common characteristics in causal graphs. Causal graph-based reasoning relies heavily on understanding these fundamental problems. (3) Level 3: *Advanced Task*. These tasks present complex, high-level challenges that are central to causal graph understanding. Solid arrows indicate the dependencies between tasks within the same level, while dashed arrows represent the tasks' dependencies across different levels. Task dependency design draws on established research (Shpitser and Pearl, 2006; Pearl, 2009; Bareinboim and Pearl, 2012; Pearl et al., 2016; Pearl and Mackenzie, 2018; Jaber et al., 2019).

hury et al. (2022) sufficiently define a model's *understanding*, there is still room for improvement. For instance, these conditions fail to offer precise quantitative criteria and lack explicit clarification on what type of information is considered relevant.

2.2 Exploring Language Models' *Understanding* of Causal Graphs

Numerous studies have identified understanding as a key factor in the pursuit of human-level artificial intelligence (McCarthy, 2007; Adams et al., 2012; McClelland et al., 2020). However, arriving at a definition of *understanding* within language models is an ongoing challenge. Evaluating models' *understanding* based on accuracy is currently the dominant approach and certainly essential (Ashwani et al., 2024; He et al., 2024; Xu et al., 2024), but this method suffers from inherent limitations. Real-world problems are complex, and data often contains noise (Gupta and Gupta, 2019; Moran et al., 2020; Bansal et al., 2022). These make it practically impossible for any model to be perfectly accurate all the time (even humans rarely achieve this) (Valverde-Albacete and Peláez-Moreno, 2014). While it is clear that understanding varies in degree (Nickerson, 1985), pinning down a specific threshold is difficult. This difficulty is compounded by the variability in task complexity and the subjective nature of interpreting "levels of understanding". Consequently, rather than define *"what constitutes understanding of causal graphs in a language model"*, we think it might be more productive to consider *"if a language model understands causal graphs, how should it behave?"*

2.3 Seeking *Understanding* of Causal Graphs in Model Behavior

To measure how well language models understand causal graphs, we develop a three-level evaluation hierarchy comprising 20 meticulously crafted causal graph-based tasks (as Figure 2 illustrates). These tasks include graphs' basic tasks (e.g., cycle), intermediate tasks (e.g., markov equivalent class), and advanced tasks (e.g., causal effect identification). Proficiency in these 20 tasks serves as a valid measure of a model's understanding of causal graphs. Therefore, combining the analyses from Section 2.1 and Section 2.2, we propose that a language model that exhibits understanding would demonstrate the following four behaviors in our tasks.[1] The performance of a model is denoted by P, random guess by P_r, the original response of a model by R, and the ground truth by GT.

B1: Performance exceeding random guesses. Existing work suggests that random guess implies a lack of extensive understanding of the given problem (Capraro et al., 2012). Moreover, using random guess as baseline is a common and reasonable practice in evaluating model performance (Chen et al., 2023; Wang et al., 2024a; Chen et al., 2024b). This behavior can be formulated as $P > P_r$.

[1] More thoughts about our framework are in Appendix A.1.

B2: Robustness against question types. Numerous studies highlight that altering the question type or description of a graph, while preserving the original meaning of the problem, can significantly impact model performance (Fatemi et al., 2023; Hu et al., 2023; Luo et al., 2024). Therefore, we suppose that if a model's *understanding* of a causal graph and its related tasks is genuine, its performance should not be sensitive to superficial changes in the causal graph's question type.

B3: Correct utilization of causal definitions. As De Regt (2004) emphasizes, understanding implies the ability to utilize given definitions to solve problems. This behavior indicates that the model not only recognizes terms but also understands their meanings and how they relate to the given context. This behavior can be defined as: $R \leftarrow def. = GT$, where $R \leftarrow def.$ means a model's response after adding a causal definition to the prompt. The definition can be conveyed either explicitly within the prompt or implicitly through the provision of examples (e.g., IcL) (Li et al., 2022; Zheng et al., 2023; Richens and Everitt, 2024).

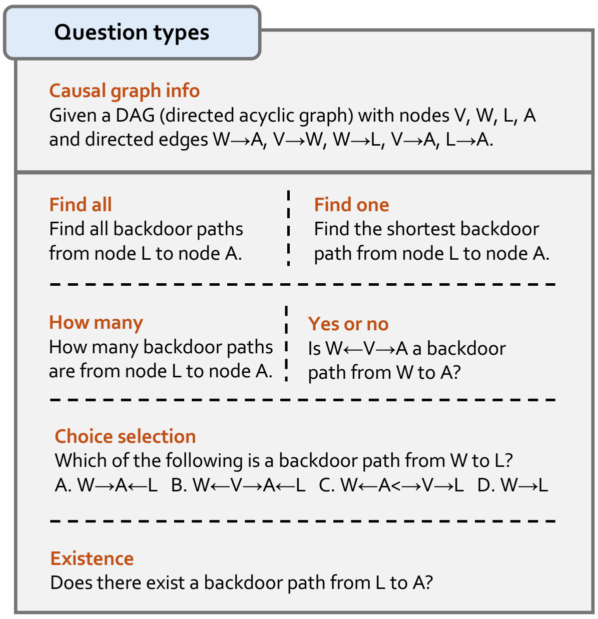
B4: Performance constrained by task dependence. Task dependence consistently emerges as a crucial factor in studies focused on understanding (Kvanvig, 2003; Carter and Gordon, 2014; Grimm, 2021). As shown in Figure 2, we determine that Task B is dependent on Task A if it requires knowledge acquired from Task A for resolution, whereas solving Task A does not necessitate knowledge from Task B. Mastery of the foundational task is thus deemed essential for succeeding in the dependent task. This performance constraint due to task dependence serves as a critical metric for assessing a model's depth of understanding.

3 The CLEAR Benchmark

To explore the question: ***C**an **L**anguage mod**E**ls re**A**lly unde**R**stand causal graphs*? we propose CLEAR, the first benchmark dedicated to causal graph understanding. We ensure dataset diversity by accounting for various factors: the size, type, and density of causal graphs, as well as the richness of tasks and question types.

3.1 Benchmark Construction

Generating random graphs. We begin by randomly creating a set of graphs. A graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ represent set of nodes and edges. To ensure diversity, we cover both general and causal graphs, differentiated by structure into four types: undirected graph, directed graph, directed acyclic graph (DAG), and acyclic directed mixed graph (ADMG) (Peters et al., 2017). The undirected and directed graphs, typical of general graph types, are employed primarily in basic tasks. Conversely, DAGs and ADMGs, which are causal in nature, are utilized in intermediate and advanced tasks. To control the complexity, we vary the number of nodes (n_v) from 4 to 9 and adjust the number of edges (n_e) from $n_v - 1$ to 10 for each n_v. These graph types involve three types of edges: undirected edge, directed edge, and bi-directed edge. The undirected edges symbolize reciprocal relationships, while the bi-directed edges suggest the presence of confounding between nodes. For ADMGs that contain both directed and bi-directed edges, we maintain the ratio of bi-directed to directed edges at or below 0.5 to prevent excessive complexity. We denote nodes using letters, and to ensure neutrality and mitigate bias from the model's potential prior knowledge, the alphabetical order of $\mathcal{V}$ is randomized.



Question types

Causal graph info
Given a DAG (directed acyclic graph) with nodes V, W, L, A and directed edges W→A, V→W, W→L, V→A, L→A.

Find all
Find all backdoor paths from node L to node A.

Find one
Find the shortest backdoor path from node L to node A.

How many
How many backdoor paths are from node L to node A.

Yes or no
Is W←V→A a backdoor path from W to A?

Choice selection
Which of the following is a backdoor path from W to L?
A. W→A←L B. W←V→A←L C. W←A<→V→L D. W→L

Existence
Does there exist a backdoor path from L to A?

Figure 3: **Six question types.** Taking the backdoor path as an example, we design six question types in CLEAR. A complete question is formulated by combining the causal graph info with a specific question type.

Generating causal reasoning questions. Based on the causal graphs, we generate questions with corresponding ground truth for various causal tasks and question types. The questions are produced

Table 1: **Concise statistics of the CLEAR benchmark.** We tally the number of different causal tasks, organizing them by various levels. Type indicates question type.

Causal task	**# Type**	**# Sample**
Basic Task		
Single node (SN)	4	192
Single edge (SE)	4	192
Two nodes relationship (2NR)	5	120
Three nodes relationship (3NR)	5	120
Path (PT)	5	168
Cycle (CL)	4	144
Topological ordering (TO)	3	144
Intermediate Task		
Blocked path (BLP)	3	144
D-separation (DS)	3	120
Markov equivalent class (MEC)	2	120
Markov blanket (MB)	3	144
Directed path (DP)	5	120
Backdoor path (BKP)	5	144
C-component (CC)	3	108
C-tree (CT)	1	120
C-forest (CF)	1	120
Maximal root set (MRS)	4	192
Advanced Task		
Backdoor adjustment set (BAS)	4	132
Frontdoor adjustment set (FAS)	4	144
Causal effect identification (CEI)	1	120
Total	6	2808

using predefined templates. Specifically, we design 20 causal tasks and six question types.[2] And as Figure 3 demonstrates, these question types can be divided into two types of subjective questions (i.e., *find all* and *find one*) and four types of objective questions (i.e., *how many*, *yes or no*, *choice selection*, and *existence*), providing an in-depth evaluation of models' *understanding*. The objective questions have a single, clearly verifiable answer based on the question and causal graph. Regarding subjective questions, we take *find all* as an example. *Find all* requires listing all answers meeting specific criteria (e.g., "Find all paths from node X to node Y"). The subjectivity arises primarily from the answer format, as there can be multiple correct ways to express the answer (e.g., both "X→Y" and "a path from X to Y" are acceptable).

3.2 Data Statistics

Our CLEAR benchmark includes 20 causal tasks, spanning all three complexity levels. We generate 2808 questions in total. For each causal task, we ensure that the number of questions exceeds 100 to support the validity of our experimental conclusions. Table 1 presents the overview of the CLEAR.

4 Experiments

4.1 Setups

Models. Our evaluation encompasses six models. This selection includes both open-access models (InternLM2-Math-20B (Ying et al., 2024), Mixtral-8×7B (Jiang et al., 2024), and Llama2-Chat-70B (Touvron et al., 2023)), and limited-access models (GPT-3.5-Turbo (OpenAI, 2022), GPT-4 (Achiam et al., 2023), and Gemini Pro (Team et al., 2023)). They originate from various creators and exhibit a spectrum of model scales. We use the default hyper-parameter settings for all models.

Prompts. In Section 4.2, 4.3 and 4.5, we employ the basic prompt (i.e., <question>). In Section 4.4, we adopt basic prompt, 1/3-shot IcL (Brown et al., 2020), and definition-guided prompt (i.e., <instruction, definition, question>).[3]

Metrics. The evaluation metric is accuracy. Objective questions are assessed via answer extraction using GPT-4 and exact-match scoring.[4] Subjective questions are evaluated manually. For a more efficient and accurate human evaluation, we develop a dedicated HTML-based tool. This tool not only facilitates a more intuitive visualization of the model outputs but also enhances the overall review process. More details about our tool and human evaluation are provided in Appendix C.2.

4.2 Comparison with Random Guess

Figure 4 illustrates the models' performances on all causal tasks. Each cell in the figure represents a model's accuracy. Moreover, to ensure the soundness of our benchmark design and the trustworthiness of our results, we conduct a multi-turn evaluation in Appendix C.3. From Figure 4, we can conclude that: (1) Although limited (i.e., approximately 40% to 60% room for improvement), language models do exhibit preliminary *understanding* (i.e., exceed random guess) of causal graphs. The rightmost column of the figure indicates the models' average accuracies, demonstrating that all models outperform their random guesses. This suggests that they possess basic *understanding* of the causal graphs. Despite exceeding random guesses, there remains substantial room for improvement. Even the top-performing model, GPT-

[2] Appendix B provides more information on the dataset.

[3] See Appendix C for details on these prompts.

[4] Prior studies have shown that strong language models (e.g., GPT-4) can be effective judges (Lu et al., 2023; Zheng et al., 2024b), demonstrating the validity of this approach.

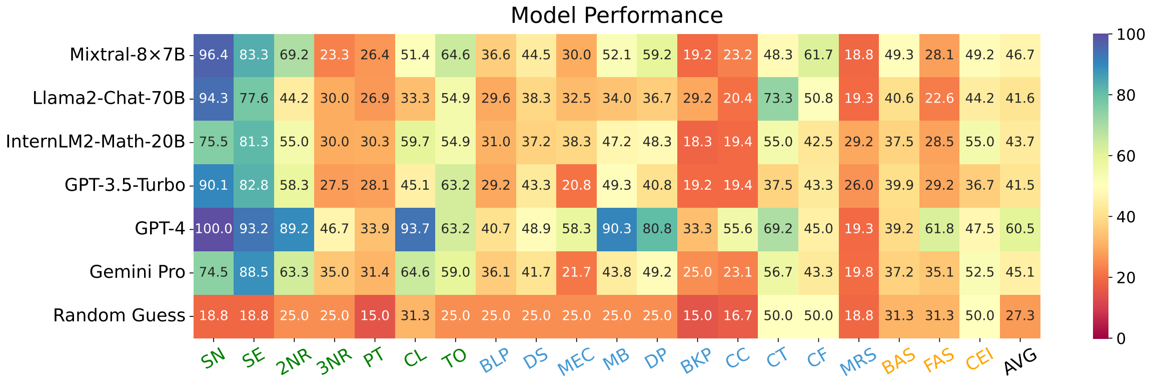


Figure 4: **Overall model performance.** Each cell corresponds to the model's accuracy on that specific task.

4, only reaches an accuracy of 60.5%, while the remaining models hover around 40.0%. (2) Language models demonstrate a good grasp of the fundamental elements that constitute a causal graph. All models achieve over 70.0% accuracy on the single node and single edge tasks, with GPT-4 even reaching 100.0% on the single node. These results provide valuable insights for designing future tasks involving causal graphs. (3) The model's error response is the dominant factor contributing to its subpar performance compared with random guess. We adopt the error types defined in Chen et al. (2024b) and observe the model exhibit errors such as causal hallucination, contradiction, and misunderstanding. For instance, when the model's response exhibits contradiction, it might simultaneously answer "yes" and "no". This ambiguity makes it challenging to extract answers using GPT-4. GPT-4 would output "unknown" in this scenario, rendering the response invalid.[5]

Figure 5 presents the models' average accuracies across three levels. We find that: (1) Language models excel at the *basic task* level. All models achieve an accuracy exceeding 50.0%, with the highest reaching 74.3%. Conversely, most average accuracies attained on the remaining two levels fail to surpass 40.0%. Our three-level structure highlights the limitations of current models and offers potential guidance for the construction of future benchmarks. (2) The five models, excluding GPT-4, demonstrate similar performance.

4.3 Is the Model Robust?

To evaluate the models' robustness, we consider all six different question types. Different question types within a specific causal task, when presented with the same causal graph, aiming to probe the same core concept of causality. Importantly, we acknowledge the potential impact of question type on both the probabilities of random guesses and the phrasing of questions. Our objective is to conduct a preliminary investigation into how question types influence model robustness.

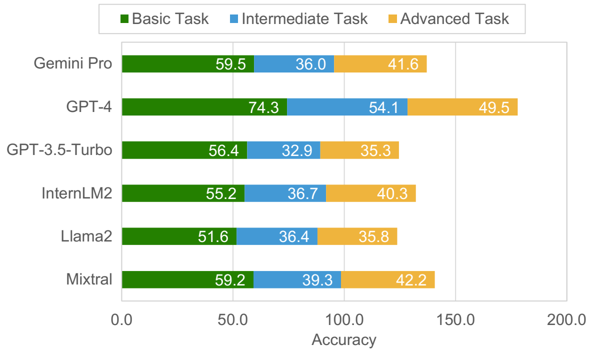


Figure 5: **Model performance across the three levels of CLEAR.** The term Mixtral refers to Mixtral-8×7B, Llama2 to Llama2-Chat-70B, and InternLM2 to InternLM2-Math-20B.

Figure 6 shows the average accuracy of the models across different question types. We draw the following conclusions: (1) Model performance is sensitive to question type. All models excel in YN and EX question types but struggle with FA, FO, and HM. Wherein, Llama2-Chat-70B, InternLM2-Math-20B, and Gemini Pro exhibit performance discrepancies exceeding 35.0% across different question types. Although GPT-3.5-Turbo is not the top performer, it demonstrates the minimal performance difference, measuring at 22.8%. (2) A model's understanding of causal graphs might be artificially inflated if evaluation relies on limited question types. The selection bias inherent in language models raises concerns about their robust-

[5] The qualitative analysis is provided in Appendix C.4.

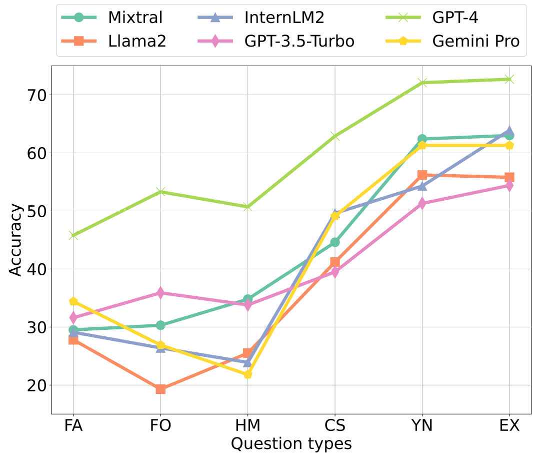


Figure 6: **How the question types affect model robustness.** We compare the models' accuracies across different question types. FA stands for find all, FO for find one, HM for how many, CS for choice selection, YN for yes or no, and EX for existence.

ness (Zheng et al., 2024a; Chen et al., 2024a). If we only evaluate language models on CS, YN and EX, we risk overestimating their true capabilities. It is the diversity of question types that reveals the actual understanding capability of a model.[6]

4.4 Definition Proficiency of the Model

To investigate whether the models could effectively utilize the provided definitions related to a causal graph, we further conduct experiments on seven tasks (i.e., 3NR, PT, BLP, BKP, CC, MRS, and FAS).[7] For these tasks, the average accuracies across all models on the objective questions are below 40%. Moreover, the seven tasks span all levels in Figure 2, which can fully demonstrate the effectiveness of the experiments. For prompts, we select the basic prompt, 1/3-shot IcL, and definition-guided prompt. There is ample work validating the effectiveness of IcL (Wu et al., 2023; Wang et al., 2023). Therefore, to assess a model's ability to correctly apply or abstract a causal definition, IcL serves as an ideal reference.

Figure 7 shows the overall accuracy difference of each model across seven causal tasks using different prompts.[8] The baseline for comparison is the average accuracy of each model under the basic prompt. By analyzing this figure, we can draw the following conclusions: (1) The models exhibit notable differences in their *understanding* of definitions related to a causal graph. Providing the causal definition significantly enhances the performance of GPT-4, GPT-3.5-Turbo and Mixtral-8×7B. Notably, the improvement is most pronounced for GPT-4, which even surpasses both 1-shot IcL and 3-shot IcL. The improvements on GPT-3.5-Turbo and Mixtral-8×7B are also remarkable, both outperforming 1-shot IcL. However, the remaining three models do not benefit from the provided definition. Specifically, InternLM2-Math-20B exhibits the most prominent accuracy decline. (2) Models capable of (explicitly) utilizing definitions correctly are often observed performance improvements (implicitly) through IcL. However, even if a model's performance can be considerably promoted by IcL, it does not necessarily mean the model can successfully apply (explicit) definitions. Despite the potential for accuracy gains (over 60% cumulatively) of Gemini Pro using 3-shot IcL, it struggles to correctly apply (explicitly) provided definitions, resulting in diminished performance.

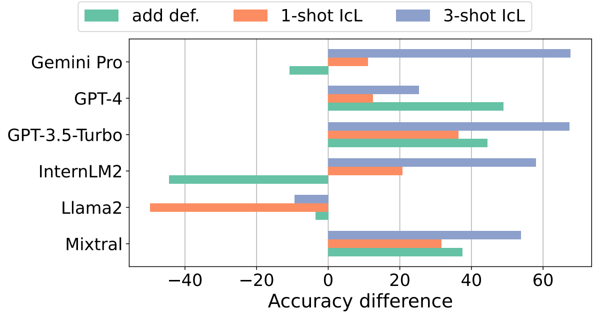


Figure 7: **Explicit and implicit definition proficiency.** We compare how well the model could utilize definitions, examining both explicitly and implicitly. Add def. indicates the definition-guided prompt.

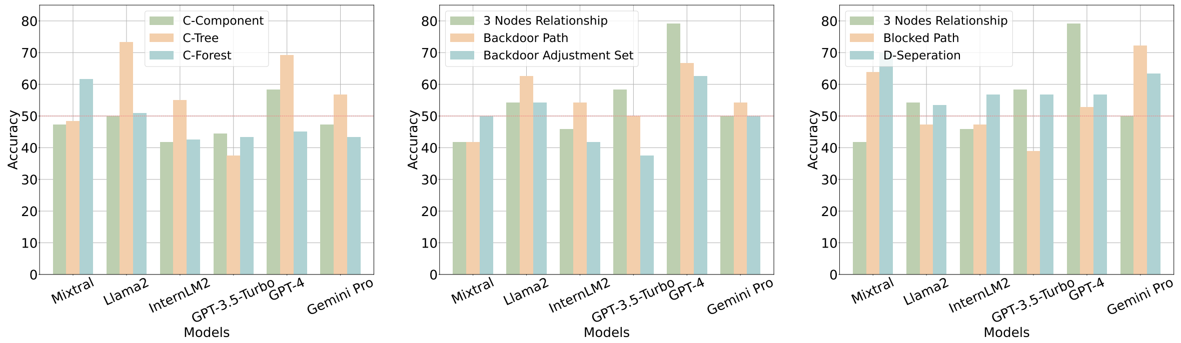
4.5 How Task Dependence Shapes Model Performance

Based on Figure 2, we select three representative sets of dependent causal tasks and consider the YN question type. (1) *Tasks within the same level*: we choose CC→CT→CF, all located at *intermediate task*. (2) *Tasks across distinct levels*: we choose a sequence spanning different levels: 3NR (*basic task*)→BKP (*intermediate task*)→BAS (*advanced task*) are considered. (3) *Tasks with partial level overlap*: we focus on a combination where some tasks share the same level: 3NR (*basic task*)→BKP

[6] There is a growing trend in benchmark design towards incorporating a wider variety of question types or providing more choices for models (Chen et al., 2024b; Wang et al., 2024b; Röttger et al., 2024).

[7] These abbreviations are given in Table 1. Detailed definitions of all seven tasks are in Table 5 (Appendix C.5).

[8] We provide the detailed data in Table 6 (Appendix C.5).



(a) CC→CT→CF

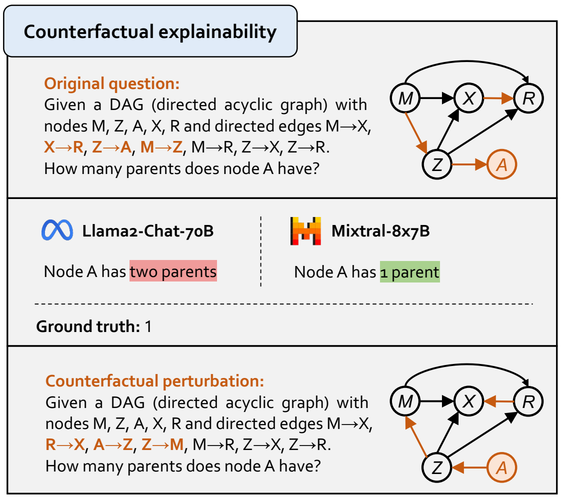
(b) 3NR→BKP→BAS

(c) 3NR→BLP→DS

Figure 8: **Task dependence's impact on model performance.** We evaluate model performance across three groups of causal tasks categorized by their correlations. The orange dashed line represents the accuracy of random guess.

(*intermediate task*)→DS (*intermediate task*).

Upon meticulous examination of Figure 8, we have the following observations: (1) The performances of most models are not constrained by dependent causal tasks. Out of all models, only GPT-3.5-Turbo and GPT-4 in Figure 8(b) exhibit the expected accuracy trend (i.e., 3NR≥BKP≥BAS). These results suggest that the models might not truly *understand* the causal relationships between tasks, but rather rely on other spurious correlations. It is also possible that not all models possess the capacity for human-level causal reasoning and knowledge transfer ability. (2) Different models exhibit varying performance trends when tackling the same group of dependent causal tasks. This highlights the heterogeneity of knowledge representation and application among different models.



Counterfactual explainability

Original question:
Given a DAG (directed acyclic graph) with nodes M, Z, A, X, R and directed edges M→X, X→R, Z→A, M→Z, M→R, Z→X, Z→R.
How many parents does node A have?

Llama2-Chat-70B

Node A has two parents

Mixtral-8x7B

Node A has 1 parent

Ground truth: 1

Counterfactual perturbation:
Given a DAG (directed acyclic graph) with nodes M, Z, A, X, R and directed edges M→X, R→X, A→Z, Z→M, M→R, Z→X, Z→R.
How many parents does node A have?

Figure 9: **Counterfactual perturbation used in this case.** Starting with the original question, we obtain answers for both models. Next, we establish the baseline using counterfactual perturbation. Finally, we calculate the token attribution of key information to understand its influence on the model's output.

4.6 Counterfactual Explainability

The analyses from Section 4.2 to Section 4.5 are based on directly calculating the accuracy of the models' outputs. To extend beyond mere accuracy, we leverage `Captum` (Kokhlikyan et al., 2020; Miglani et al., 2023), a `Python` library for model interpretability, to explore language models' *understanding* of causal graphs from a counterfactual perspective. We primarily use the *perturbation based methods* provided by `Captum`.[9] As depicted in Figure 9, we first query both Llama2-Chat-70B and Mixtral-8×7B, which are of comparable scale and have been widely adopted, using the original question to obtain their respective answers. Our main focus is the impact of "Z→A" on the model's response. We suspect "X→R" and "M→Z", which are located near the "Z→A", could also potentially impact the model's response. Consequently, we use counterfactual perturbations to analyze the influences of these three statements on the model. We set the counterfactual perturbations as baseline (i.e., perturbation-based algorithm uses it as reference value), and the model's response as target string. Finally, using the target function in `Captum`, we calculate the log probability of the model generating its answer given the question.

Figure 10 displays the token attributions of the models' responses. We find that "Z→A" is the most positive factor in getting the right answer "1" for Mixtral-8×7B, with "X→R" and "M→Z" also contributing positively. This confirms that Mixtral-8×7B correctly identifies and utilizes the relevant information in its reasoning process. In contrast,

[9] For further guidance, refer to the tutorial at: `https://captum.ai/tutorials/Llama2_LLM_Attribution`.

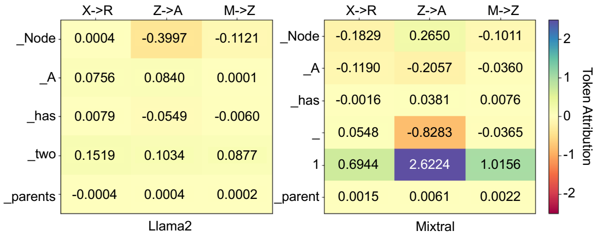


Figure 10: **Token attribution.** On the y-axis, underscores mark the tokenizer's divisions of each target string. The x-axis displays key information of the question.

Llama2-Chat-70B produces a wrong answer. Neither "Z→A", "X→R" nor "M→Z" exhibit a significant positive impact on its answer, suggesting that Llama2-Chat-70B fails to identify key information. The results support the claim in Section 4.2 that models have a preliminary understanding of causal graphs. More importantly, the results demonstrate a strong link between a model's *understanding* of the causal graph and its ability to focus on the essential information within the graph.

5 Related Work

Language models' *understanding* ability. Language models' understanding is being probed through various perspectives, such as causality (Hobbhahn et al., 2022; Kim et al., 2023; Ashwani et al., 2024), real-world problems (Choi et al., 2023; He et al., 2024; Xu et al., 2024), disciplines (Castro Nascimento and Pimentel, 2023; Guo et al., 2024). A common approach in these studies is to establish a benchmark, and then evaluate a model's performance. A more rigorous exploration of what means *understanding* in models is still needed.

Graph-based benchmarks. The capacity of language models to solve graph-based problems is attracting growing attention. Wang et al. (2024a) propose the NLGraph, concentrating primarily on essential graph tasks. Luo et al. (2024) introduce the GraphInstruct benchmark. Fatemi et al. (2023) propose the GraphQA to explore the impact of different graph encoding methods. LLM4DyG (Zhang et al., 2023a) addresses the dynamic graphs. Despite progress in applying models to graph tasks, their ability to reason about causality within graphs still requires further investigation.

Causal evaluation on language model. The quest to understand causality in language models is heating up. Jin et al. (2023) propose CLADDER, a dataset encompassing over 10K diverse questions. Liu et al. (2024) investigate the capabilities of language models in handling data-based problems. Chen et al. (2024b) develop CaLM, a 120,000+ bilingual dataset for in-depth evaluation of language models' causal reasoning ability. Numerous other efforts further enrich this area (Nie et al., 2023; Jin et al., 2024; Zhou et al., 2024; Kiciman et al., 2024). However, a dedicated benchmark specifically from the perspective of causal graphs is still lacking.

6 Conclusion

This paper provides a comprehensive and in-depth exploration on the question: *Can language models really understand causal graphs?* We define a practical framework for accessing a model's understanding. We introduce `CLEAR`, a novel benchmark designed to evaluate a model's understanding of causal graphs, filling a significant gap in existing research. We validate our framework through extensive experiments and conclude five insightful findings.

7 Limitations

Despite our best efforts to design a framework for causal graph understanding, construct a benchmark, and conduct thorough experiments on six models, we acknowledge that our work still has limitations. The language of `CLEAR` is relatively limited. Due to time and budget constraints, our benchmark only considers English. As language models are increasingly used worldwide, we acknowledge that a multilingual dataset could provide more meaningful findings. Moreover, the definition of understanding still requires further exploration. For instance, how to extend the concept of robustness to broader scenarios. Additionally, evaluating the understanding of large vision language models (LVLMs) will likely require considering a wider set of factors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sam Adams, Itmar Arel, Joscha Bach, Robert Coop, Rod Furlan, Ben Goertzel, J Storrs Hall, Alexei Samsonovich, Matthias Scheutz, Matthew Schlesinger, et al. 2012. Mapping the landscape of human-level artificial general intelligence. *AI magazine*, 33(1):25–42.

Ronald Brian Adler, George R Rodman, and Alexandre Sévigny. 2006. *Understanding human communication*, volume 10. Oxford University Press Oxford.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality? *arXiv preprint arXiv:2402.18139*.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pages 1466–1482. PMLR.

Elias Bareinboim and Judea Pearl. 2012. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mario Bunge. 2017. *Causality and modern science*. Routledge.

Mary Margaret Capraro, Song A An, Tingting Ma, A Fabiola Rangel-Chavez, and Adam Harbaugh. 2012. An investigation of preservice teachers' use of guess and check in solving a semi open-ended mathematics problem. *The Journal of Mathematical Behavior*, 31(1):105–116.

J Adam Carter and Emma C Gordon. 2014. Objectual understanding and the value problem. *American Philosophical Quarterly*, 51(1):1–13.

Cayque Monteiro Castro Nascimento and André Silva Pimentel. 2023. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. 2023. Graphllm: Boosting graph reasoning ability of large language model. *arXiv preprint arXiv:2310.05845*.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024a. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *arXiv preprint arXiv:2403.18346*.

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024b. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2024c. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Henk W De Regt. 2004. Discussion note: Making sense of understanding. *Philosophy of Science*, 71(1):98–109.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2023. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*.

Peter A Frensch and Joachim Funke. 2014. Definitions, traditions, and a general framework for understanding complex problem solving. In *Complex problem solving*, pages 3–25. Psychology Press.

Stephen Grimm. 2021. Understanding. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Siyuan Guo, Aniket Didolkar, Nan Rosemary Ke, Anirudh Goyal, Ferenc Huszár, and Bernhard Schölkopf. 2024. Learning beyond pattern matching? assaying mathematical understanding in llms. *arXiv preprint arXiv:2405.15485*.

Shivani Gupta and Atul Gupta. 2019. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, 161:466–474.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Malte Helmert. 2004. A planning heuristic based on causal graph analysis. In *ICAPS*, volume 16, pages 161–170.

Marius Hobbhahn, Tom Lieberum, and David Seiler. 2022. Investigating causal understanding in llms. In *NeurIPS ML Safety Workshop*.

Yuntong Hu, Zheng Zhang, and Liang Zhao. 2023. Beyond text: A deep dive into large language models' ability on understanding graph data. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*.

Amin Jaber, Jiji Zhang, and Elias Bareinboim. 2019. Causal identification under markov equivalence: Completeness results. In *International Conference on Machine Learning*, pages 2981–2989. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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*.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Yuheun Kim, Lu Guo, Bei Yu, and Yingya Li. 2023. Can chatgpt understand causal language in science claims? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 379–389.

Samantha Kleinberg. 2013. *Causality, probability, and time*. Cambridge University Press.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. 2017. Experimental design for learning causal graphs with latent variables. *Advances in Neural Information Processing Systems*, 30.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Jonathan L Kvanvig. 2003. *The value of knowledge and the pursuit of understanding*. Cambridge university press.

Daniel K Lapsley et al. 2004. Moral functioning: Moral understanding and personality. In *Moral development, self, and identity*, pages 347–360. Psychology Press.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9215–9235. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Zihan Luo, Xiran Song, Hong Huang, Jianxun Lian, Chenhao Zhang, Jinqi Jiang, Xing Xie, and Hai Jin. 2024. Graphinstruct: Empowering large language models with graph understanding and reasoning capability. *arXiv preprint arXiv:2403.04483*.

John McCarthy. 2007. From here to human-level ai. *Artificial Intelligence*, 171(18):1174–1182.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974.

Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using captum to explain generative language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173.

Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. 2020. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072.

Raymond S Nickerson. 1985. Understanding understanding. *American Journal of Education*, 93(2):201–239.

Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36:78360–78393.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Blog post.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 1998. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Jonathan Richens and Tom Everitt. 2024. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

David E Rumelhart. 1991. Understanding understanding. *Memories, thoughts and emotions: Essays in honor of George Mandler*, 257:275.

Ilya Shpitser and Judea Pearl. 2006. Identification of joint interventional distributions in recursive semi-markovian causal models. In *AAAI*, pages 1219–1226.

Steven Sloman and Steven A Sloman. 2009. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Francisco J Valverde-Albacete and Carmen Peláez-Moreno. 2014. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PloS one*, 9(1):e84217.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2024a. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems*, 36.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

James Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. " a good pun is its own reword": Can large language models understand puns? *arXiv preprint arXiv:2404.13599*.

Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. 2023. Natural language is all a graph needs. *arXiv preprint arXiv:2308.07134*.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*.

Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Yijian Qin, Simin Wu, and Wenwu Zhu. 2023a. Llm4dyg: Can large language models solve problems on dynamic graphs? *arXiv preprint arXiv:2310.17110*.

Ziwei Zhang, Haoyang Li, Zeyang Zhang, Yijian Qin, Xin Wang, and Wenwu Zhu. 2023b. Graph meets llms: Towards large graph models. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024a. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Step-back prompting enables reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. Causalbench: A comprehensive benchmark for causal learning capability of large language models. *arXiv preprint arXiv:2404.06349*.

A Considerations in Understanding Framework Design

A.1 Two-pronged Approach

In Section 2.2, we propose that we should consider *"if a language model understands causal graphs, how should it behave?"* We believe the question needs to be considered from two aspects: (1) Human-centric perspective. To properly assess language models, we must first define understanding in a way that aligns with human cognition. This is crucial for ensuring language models truly achieve human capabilities. (2) Model-centric perspective. While human-centric definitions provide a starting point, there exist foundational differences in information processing between human brains and language models (Caucheteux et al., 2023). Therefore, we need to explore practical definitions that are suitable to the characteristics of models.

To this end, we carefully design four criteria in Section 2.3. And from a model-centric perspective, we define understanding by examining behaviors related to B1: performance exceeding random guesses and B2: robustness against question types. From a human-centric perspective, we consider B3: correct utilization of causal definitions and B4: performance is constrained by task dependence.

B Design Details of CLEAR

B.1 Overall Statistics

The detailed statistics of CLEAR are in Table 2.

B.2 Question Templates

Templates for questions of CLEAR are listed in Table 3. While potentially impacting diversity (Cobbe et al., 2021), this method enables efficient data scaling and accesses whether a model can recognize subtle distinctions within the templates (Chen et al., 2024b).

C Details for Experiments

C.1 Prompt Settings

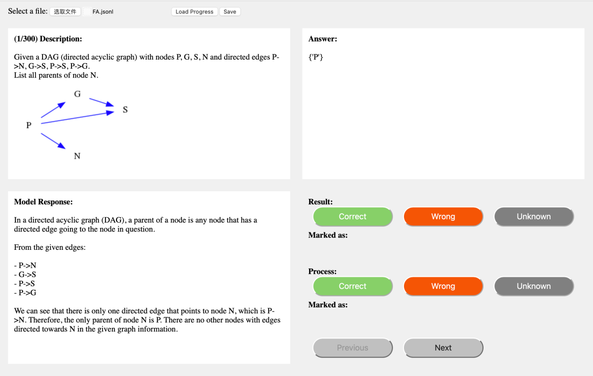
Basic prompt. Our basic prompt aligns with the definition in Chen et al. (2024b), referring to providing only the question requiring an answer.

Definition-guided prompt. Taking BKP as an example, Figure 11 illustrates how to incorporate the definition of this causal task into the prompt (i.e., definition-guided prompt).

Backdoor path with definition

Prompt:
You will be provided with a definition of a concept. Using this definition, answer the following question.
Definition:
Given an ordered pair of variables (X, Y), a backdoor path is any path from X to Y that starts with an arrow pointing into X. This backdoor path is a non-causal path from X to Y.
Question:
Given a DAG (directed acyclic graph) with nodes V, W, L, A and directed edges W⟶A, V⟶W, W⟶L, V⟶A, L⟶A.
Is W⟵V⟶A a backdoor path from W to A?

Figure 11: **Definition-guided prompt.** We explicitly provide the model with definitions relevant to the questions.



(1/300) Description:

Given a DAG (directed acyclic graph) with nodes P, G, S, N and directed edges P->N, G->S, P->S, P->G.
List all parents of node N.

Answer:

{'P'}

Model Response:

In a directed acyclic graph (DAG), a parent of a node is any node that has a directed edge going to the node in question.

From the given edges:

- P->N
- G->S
- P->S
- P->G

We can see that there is only one directed edge that points to node N, which is P->N. Therefore, the only parent of node N is P. There are no other nodes with edges directed towards N in the given graph information.

Result:

Marked as:

Process:

Marked as:

Figure 12: **Human evaluation tool.** We develop a dedicated tool for a more efficient and accurate evaluation.

C.2 Human Evaluation

We provided a comprehensive guideline to the annotators, covering key aspects such as: (1) Relevance: Assessing whether the model's response is relevant to the question. (2) Accuracy: Evaluating the factual correctness of the response.

To ensure the quality of the evaluation, we implemented the following measures: (1) Expertise of annotators: Both annotators have over six years of experience in computer science. (2) Specialized evaluation tool: We develop an HTML-based tool for intuitive visualization and streamlined evaluation of model outputs. Figure 12 demonstrates the interface of our human evaluation tool, and it is publicly available at https://github.com/OpenCausaLab/CLEAR. (3) Training session: Both experts underwent a training session to align their understanding and application of the guidelines.

As for our inter-agreement quality, the annotators work collaboratively to resolve discrepancies and ensure consistency. This collaborative approach helps maintain a high level of agreement and reliability in the evaluations.

Table 2: **Detailed statistics of the CLEAR benchmark.** We tally the number of different question types within each causal task, organizing them by various levels. YN indicates yes or no.

Causal task	**Find all**	**Find one**	**How many**	**Choice selection**	**YN**	**Existence**	**Total**
Basic Task							
Single node (SN)	48	-	48	48	48	-	192
Single edge (SE)	48	-	48	48	48	-	192
Two nodes relationship (2NR)	24	-	24	24	24	24	120
Three nodes relationship (3NR)	24	-	24	24	24	24	120
Path (PT)	24	72	24	24	24	-	168
Cycle (CL)	-	36	-	36	36	36	144
Topological ordering (TO)	-	48	-	48	48	-	144
Intermediate Task							
Blocked path (BLP)	-	72	-	36	36	-	144
D-separation (DS)	-	60	-	30	30	-	120
Markov equivalent class (MEC)	-	60	-	-	60	-	120
Markov blanket (MB)	-	48	-	48	48	-	144
Directed path (DP)	24	-	24	24	24	24	120
Backdoor path (BKP)	24	48	24	24	24	-	144
C-component (CC)	36	-	36	-	36	-	108
C-tree (CT)	-	-	-	-	120	-	120
C-forest (CF)	-	-	-	-	120	-	120
Maximal root set (MRS)	48	-	48	48	48	-	192
Advanced Task							
Backdoor adjustment set (BAS)	-	72	-	24	24	12	132
Frontdoor adjustment set (FAS)	-	72	-	24	24	24	144
Causal effect identification (CEI)	-	-	-	-	120	-	120
Total	300	588	300	510	966	144	2808

Table 3: **Question template for CLEAR.**

Causal task	Type	Template
Basic Task		
SN	FA	`List all nodes of this graph.`
	HM	`How many nodes does this graph have?`
	CS	`Which of the following is/is NOT a node of this graph?`
	YN	`Is {variable} a node of this graph?`
SE	FA	`List all edges of this graph.`
	HM	`How many edges does this graph have?`
	CS	`Which of the following is/is NOT an edge of this graph?`
	YN	`Is {variable} a edge of this graph?`
2NR	FA	`List all parents/descendants/children/ancestors of {variable}.`
	HM	`How many parents/descendants/children/ancestors does {variable} have?`
	CS	`Which of the following is one of parents/descendants/children/ancestors of {variable}?`
	YN	`Is {variable} one of parents/descendants/children/ancestors of {variable}?`
	EX	`Does {variable} have any parents/descendants/children/ancestors?`
3NR	FA	`List all chains/forks/v-structures of this graph.`
	HM	`How many chains/forks/v-structures does this graph have?`
	CS	`Which of the following is a chain/fork/v-structure of this graph?`
	YN	`Does {variables} form a chain/fork/v-structure in this graph?`
	EX	`Are there any chain/fork/v-structure of this graph?`
PT	FA	`Find all path from {variable} to {variable}.`
	FO	`Find one/the shortest/the longest path from {variable} to {variable}.`
	HM	`How many paths are there from {variable} to {variable}.`
	CS	`Which of the following is a path from {variable} to {variable}?`
	YN	`Is {variable} a path from {variable} to {variable}?`
CL	FO	`Find one cycle in this graph.`
	CS	`Which of the following is a cycle in this graph?`
	YN	`Is {variable} a cycle in this graph?`
	EX	`Are there any cycle in this graph?`
TO	FO	`Find one valid topological ordering in this graph.`
	CS	`Which of the following is a valid topological ordering of this graph?`
	YN	`Is {variable} a valid topological ordering of this graph?`
Intermediate Task		
BLP	FO	`Find one valid/the minimal nodeset that can block {variable}.`
	CS	`Which of the following nodesets can block {variable}?`
	YN	`Can {variable} be blocked by {variable}?`
DS	FO	`Find one valid/the minimal nodeset that can d-separate {variable} and {variable}.`
	CS	`Which of the following nodesets can d-separate {variable} and {variable}?`
	YN	`Are {variable} and {variable} d-separated by {variable}?`
MEC	FO	`Find another graph that belongs to the same markov equivalent class of the given graph.`
	YN	`Given another DAG with nodes {variable} and directed edges {variable}, do these two graphs belong to the same markov equivalent class?`
MB	FO	`What is the markov blanket of {variable}.`
	CS	`Which of the following is the markov blanket of {variable}?`
	YN	`Is {variable} the markov blanket of {variable}?`
DP	FA	`Find all directed paths from {variable} to {variable}.`
	HM	`How many directed paths are there from {variable} to {variable}?`
	CS	`Which of the following is a directed path from {variable} to {variable}?`
	YN	`Is {variable} a directed path from {variable} to {variable}?`
	EX	`Is there a directed path from {variable} to {variable}?`
BKP	FA	`Find all backdoor paths from {variable} to {variable}.`
	FO	`Find the shortest/the longest backdoor path from {variable} to {variable}.`
	HM	`How many backdoor paths are there from {variable} to {variable}.`
	CS	`Which of the following is a backdoor path from {variable} to {variable}?`
	YN	`Is {variable} a backdoor path {variable} to {variable}?`
CC	FA	`It can be uniquely partitioned into a set C(G) of subgraphs, each a maximal C-component. Write down such partition of the given graph.`
	HM	`It can be uniquely partitioned into a set C(G) of subgraphs, each a maximal C-component. How many subgraphs are there in C(G)?`
	YN	`Is it a C-component??`
CT	YN	`Is it a C-tree?`
CF	YN	`Is it a C-forest?`
MRS	FA	`Find the maximal root set of this graph.`
	HM	`How many nodes are there in the maximal root set of this graph?`
	CS	`Which of the following options is the maximal root set of this graph?`
	YN	`Is {variable} the maximal root set of this graph?`
Advanced Task		
BAS	FO	`Find one valid/one minimal/one maximal backdoor adjustment set for {variable} and {variable}.`
	CS	`Which of the following sets is a valid backdoor adjustment set for {variable} and {variable}?`
	YN	`Is {variable} a valid backdoor adjustment set for {variable} and {variable}?`
	EX	`Does there exist a valid backdoor adjustment set for {variable} and {variable}?`
FAS	FO	`Find one valid/one minimal/one maximal frontdoor adjustment set for {variable} and {variable}.`
	CS	`Which of the following sets is a valid frontdoor adjustment set for {variable} and {variable}?`
	YN	`Is {variable} a valid frontdoor adjustment set for {variable} and {variable}?`
	EX	`Does there exist a valid frontdoor adjustment set for {variable} and {variable}?`
CEI	YN	`Can the causal effect of {variable} on {variable} be identified or not?`

C.3 Multi-turn Evaluation

We substantiate the robustness of our benchmark design and the reliability of our results through a multi-turn evaluation. We choose our four types of objective questions (i.e., *how many*, *yes or no*, *choice selection*, and *existence*) to make the multi-turn evaluation more efficient. As mentioned in Section 4.1, our objective questions can be automatically evaluated at scale using GPT-4 extraction and exact-match scoring.

Table 4: **Concise statistics of out multi-turn evaluation.**

Model	Round 1	Round 2	Round 3
Mixtral-8×7B	51.9	46.9	47.0
Llama2-Chat-70B	47.6	45.9	46.0
InternLM2-Math-20B	48.9	47.5	47.6
GPT-3.5-Turbo	43.6	49.9	48.3
GPT-4	63.4	65.9	65.4
Gemini Pro	49.8	50.5	-

We track the average accuracy across 20 tasks over three experimental rounds, with an overall comparison presented in Table 4. The later two rounds of experiments occurred roughly three months after the first. We further provide detailed results for the first and second rounds of experiments across all 20 tasks in Figure 13 and Figure 14, respectively. The consistent results across three rounds of experiments, as visualized in the table and figures, support the soundness of our benchmark design and the reliability of the conclusions in Section 4. The results from the limited-access models (e.g., GPT-3.5-Turbo) exhibit minor fluctuations, which we believe is reasonable. Given the three-month gap between the first and subsequent rounds of experiments, it is plausible that even the same API might have undergone model updates or improvements.

C.4 Qualitative Analysis of Error Response

Figure 15 provides cases for models' error responses.

C.5 Results of Definition Proficiency

Definitions for the seven selected tasks are provided in Table 5. The complete results of the four prompts are shown in Table 6.

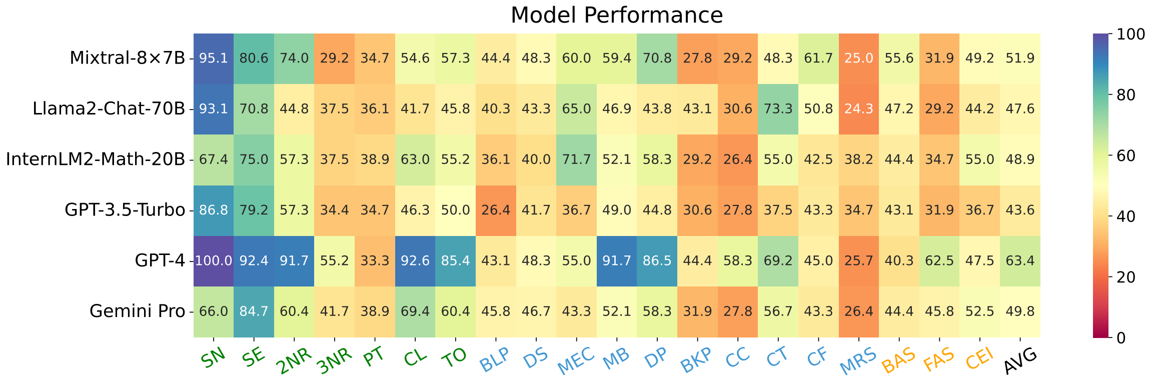


Figure 13: **Model performance of Round 1.** Each cell corresponds to the model's accuracy on that specific task.

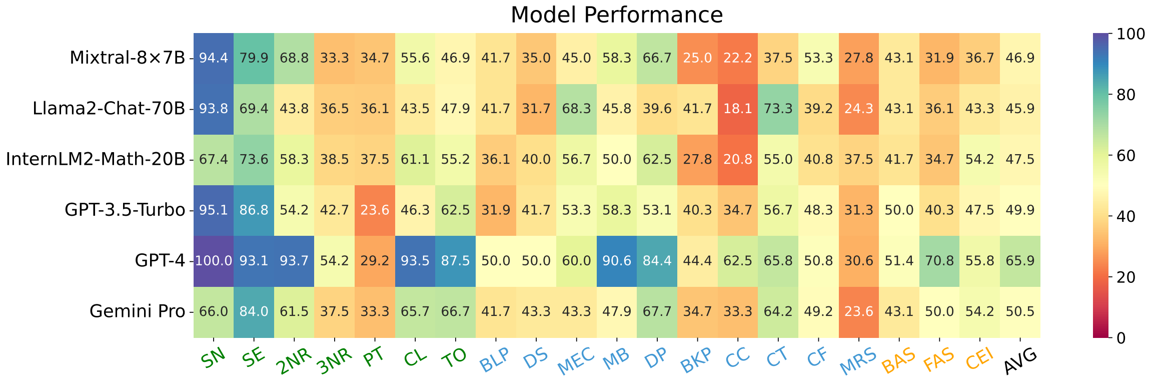
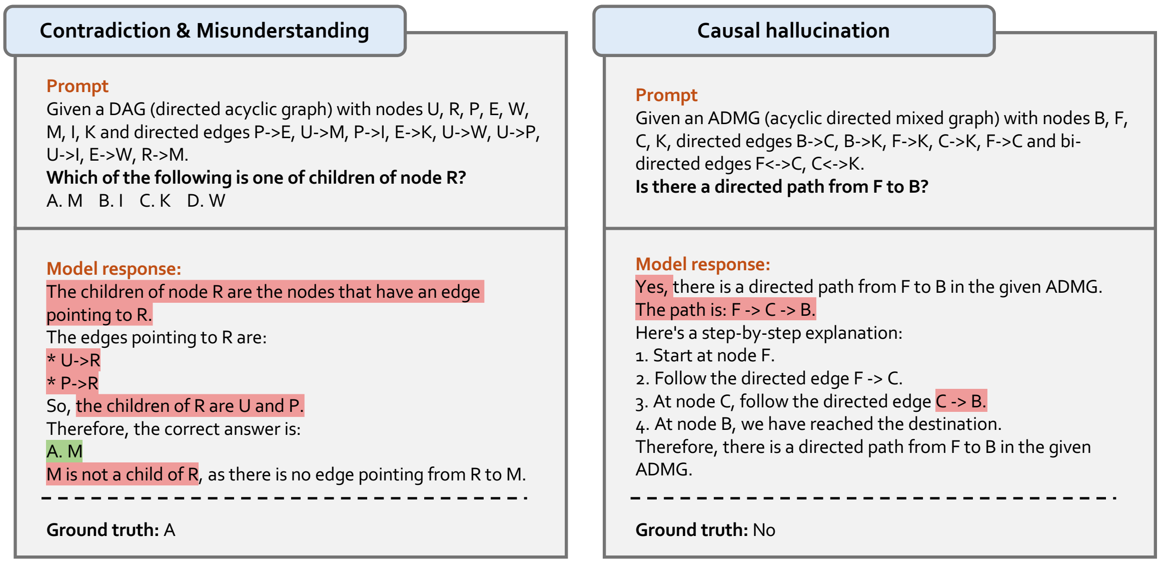


Figure 14: **Model performance of Round 2.** Each cell corresponds to the model's accuracy on that specific task.



Contradiction & Misunderstanding

Prompt

Given a DAG (directed acyclic graph) with nodes U, R, P, E, W, M, I, K and directed edges P->E, U->M, P->I, E->K, U->W, U->P, U->I, E->W, R->M.
Which of the following is one of children of node R?
A. M B. I C. K D. W

Model response:

The children of node R are the nodes that have an edge pointing to R.
The edges pointing to R are:
* U->R
* P->R
So, the children of R are U and P.
Therefore, the correct answer is:
A. M
M is not a child of R, as there is no edge pointing from R to M.

Ground truth: A

Causal hallucination

Prompt

Given an ADMG (acyclic directed mixed graph) with nodes B, F, C, K, directed edges B->C, B->K, F->K, C->K, F->C and bi-directed edges F<->C, C<->K.
Is there a directed path from F to B?

Model response:

Yes, there is a directed path from F to B in the given ADMG.
The path is: F -> C -> B.
Here's a step-by-step explanation:
1. Start at node F.
2. Follow the directed edge F -> C.
3. At node C, follow the directed edge C -> B.
4. At node B, we have reached the destination.
Therefore, there is a directed path from F to B in the given ADMG.

Ground truth: No

(a) Example of contradiction and misunderstanding.

(b) Example of causal hallucination.

Figure 15: **Error response.** We adopt the error types defined in Chen et al. (2024b). The red text means the wrong response. The green text means the right response.

Table 5: **Definitions of the selected seven causal tasks.**

Causal task	**Definition**
Basic Task	
Three nodes relationship	Given a DAG with three nodes X, Y, Z. (1) A "chain" is a sequence of nodes connected by edges where each node has only one predecessor and one successor (except for the first and last nodes in the chain). The simplest chain in a causal graph can be illustrated as "X->Y->Z". (2) A "fork" refers to a situation where one node has multiple outgoing edges leading to different successor nodes. The simplest fork in a causal graph can be illustrated as "X<-Y->Z". (3) A "v-structure" means one node is a child of the two others that themselves are not adjacent. The simplest v-structure in a causal graph can be illustrated as "X->Y<-Z".
Path	A path in a DAG is a sequence of (at least two) distinct nodes $i_1, \ldots, i_m$ such that there is an edge between i_k and i_{k+1} for all $k = 1, \ldots, m$.
Intermediate Task	
Blocked path	In a DAG, a path p is said to be blocked by a set of nodes Z if and only if: (1) p contains a chain i->m->j or a fork i<-m->j such that the middle node m is in Z, or (2) p contains an inverted fork (or collider) i->m<-j such that the middle node m is not in Z and such that no descendant of m is in Z.
Backdoor path	Given an ordered pair of variables (X, Y), a backdoor path is any path from X to Y that starts with an arrow pointing into X. This backdoor path is a non-causal path from X to Y.
C-component	Let G be a causal graph such that a subset of its bidirected arcs forms a spanning tree over all nodes in G. Then G is a C-component.
Maximal root set	Let G be a causal graph and X is one node that belongs to G. If X does not have any descendant, then we call X a root set of G. Maximal root set contains all the root sets of G.
Advanced Task	
Frontdoor adjustment set	If a set of variables Z satisfies the front-door criterion relative to an ordered pair of variables (X, Y): (1) Z intercepts all directed paths from X to Y; (2) there is no unblocked back-door path from X to Z; and (3) all back-door paths from Z to Y are blocked by X. Then we call Z a frontdoor adjustment set, this set allows us to accurately estimate the causal effect of X on Y.

Table 6: **Model performance on seven selected causal tasks.**

Causal task	**Prompt**	**Mixtral**	**Llama2**	**InternLM2**	**GPT-3.5-Turbo**	**GPT-4**	**Gemini Pro**
Basic Task							
Three nodes relationship	Basic	29.2	37.5	37.5	34.4	55.2	41.7
	add def.	35.4	33.3	29.2	40.6	60.4	42.7
	1-shot IcL	34.4	32.3	35.4	40.6	57.3	41.7
	3-shot IcL	42.7	38.5	51.0	44.8	54.2	42.7
Path	Basic	34.7	36.1	38.9	34.7	33.3	38.9
	add def.	34.7	30.6	33.3	30.6	23.6	34.7
	1-shot IcL	36.1	29.2	37.5	26.7	31.9	38.9
	3-shot IcL	43.1	31.9	48.6	44.4	26.4	50.0
Intermediate Task							
Blocked path	Basic	44.4	40.3	36.1	26.4	43.1	45.8
	add def.	40.3	37.5	18.1	44.4	56.9	36.1
	1-shot IcL	50.0	31.9	37.5	44.4	44.4	36.1
	3-shot IcL	47.2	43.1	40.3	43.1	48.6	40.3
Backdoor path	Basic	27.8	43.1	29.2	30.6	44.4	31.9
	add def.	48.6	37.5	18.1	31.9	62.5	31.9
	1-shot IcL	40.3	19.4	38.9	36.1	52.8	40.3
	3-shot IcL	44.4	30.6	37.5	40.3	55.6	52.8
C-component	Basic	29.2	30.6	26.4	27.8	58.3	27.8
	add def.	30.6	37.5	26.4	43.1	59.7	31.9
	1-shot IcL	18.1	26.4	29.2	31.9	48.6	34.7
	3-shot IcL	22.2	30.6	27.8	34.7	65.3	48.6
Maximal root set	Basic	25.0	24.3	38.2	34.7	25.7	26.4
	add def.	29.9	22.2	29.9	27.1	43.1	27.1
	1-shot IcL	29.2	18.7	38.9	32.6	30.6	31.9
	3-shot IcL	31.9	26.4	45.1	34.7	34.0	40.3
Advanced Task							
Frontdoor adjustment set	Basic	31.9	29.2	34.7	31.9	62.5	45.8
	add def.	40.3	38.9	41.7	47.2	65.3	43.1
	1-shot IcL	45.8	33.3	44.4	44.4	69.4	45.8
	3-shot IcL	44.4	30.6	48.6	45.8	63.9	51.4