Evaluating Interventional Reasoning Capabilities of Large Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Numerous decision-making tasks require estimating causal effects under interventions on different parts of a system. As practitioners consider using large language models (LLMs) to automate decisions, studying their causal reasoning capabilities becomes crucial. A recent line of work evaluates LLMs ability to retrieve commonsense causal facts, but these evaluations do not sufficiently assess how LLMs reason about interventions. Motivated by the role that interventions play in causal inference, in this paper, we conduct empirical analyses to evaluate whether LLMs can accurately update their knowledge of a data-generating process in response to an intervention. We create benchmarks that span diverse causal graphs (e.g., confounding, mediation) and variable types, and enable a study of intervention-based reasoning. These benchmarks allow us to isolate the ability of LLMs to accurately predict changes resulting from their ability to memorize facts or find other shortcuts. We evaluate six LLMs on the benchmarks, finding that GPT models show promising accuracy at predicting the intervention effects.

1 Introduction

Large language models (LLMs) have achieved impressive performance on a variety of human-relevant tasks, from summarizing web-based information and answering complex questions, to carrying out tasks as web-based agents [Chen et al., 2021, Brown et al., 2020, Li et al., 2022, Katz et al., 2024, Drouin et al., 2024]. As LLMs become increasingly used to make decisions, which fundamentally require understanding the causal impact various actions can have, there has been a recent push to evaluate whether LLMs demonstrate causal reasoning ability [Cai et al., 2023, Jin et al., 2023, 2024, Kiciman et al., 2024, Liu et al., 2024]. The challenge for this emerging field is to operationalize notions of causal reasoning that can be verbalized as text-based questions for LLMs. Kiciman et al. [2024] address this question by evaluating the ability of LLMs to retrieve known cause-and-effect relations. In contrast, Jin et al. [2023, 2024], Liu et al. [2024], Cai et al. [2023] focus on defining various queries that could only be solved with knowledge of causality and graphical models, testing abstract causal reasoning that could generalize to unseen contexts. In this paper, we contribute to the growing body of work that evaluates the causal reasoning capabilities of LLMs by focusing on an aspect of causality that is both intuitive for humans [Waldmann and Hagmayer, 2005] and crucial for decision-making [Binz and Schulz, 2023]: the ability to adapt our model of the world in response to *interventions*.

The notion of intervening on a variable is at the core of causality. In this paper, we focus on perfect interventions, where a variable in a system is manipulated and set to a particular new value. Interventions modify the causal graphical model [Pearl, 2009] of a system by deleting all incoming edges to the intervened variable. Experiments in the field of cognitive psychology [Waldmann and Hagmayer, 2005] suggest that humans instinctively recognize that an action (i.e., an

Submitted to NeurIPS 2024 Workshop on Causality and Large Models (CaLM). Do not distribute.

intervention) changes existing causal relationships, correctly making different inferences before and after interventions are performed on various variables.

Besides being natural for humans, reasoning about interventions is crucial for causal inference. Consider the example of an LLM agent that acts as a "science assistant." It knows about some causal relationships and must infer new ones given observational evidence to suggest candidate experiments. For ease, consider three variables, A, B and C, where A is known to cause B and C, but where the effect that B has on C is unknown. Initially, the agent cannot conclude anything about the effect of B on C due to the confounding effect of A, but if it encounters new information that B was intervened on, and B was observed to be correlated with C, the agent should correctly conclude that the intervention severs the confounding effect of A, allowing it to infer that B must actually have a causal effect on C. The ability to understand how interventions affect known causal relationships is central to drawing new causal conclusions when presented with new evidence.

The key contribution of this paper is to evaluate intervention reasoning in LLMs by introducing *intervention effects*, binary classification tasks that evaluate which causal relations in a graph are modified by an observed intervention. We provide a framework for verbalizing a wide range of intervention effects to LLMs, varying the choice of causal graphs and names of variables. With this framework, we develop a benchmark of intervention effect tasks that help us disentangle the intervention reasoning capabilities of LLMs from other factors that contribute to performance such as memorization of facts and the ability to extract graphs from text.

Related work. This paper relates most closely to recent papers that develop benchmarks to evaluate LLMs on various causal reasoning tasks. Kiciman et al. [2024] introduced multiple causal reasoning benchmarks for LLMs, including evaluating the ability of LLMs to recover the bivariate causal DAGs introduced in the Tübingen pairs dataset [Mooij et al., 2016]. Kiciman et al. [2024] found that GPT models recovered known causal relationships with up to 96% accuracy when experimenting with various prompting strategies such as including system prompts. However, evaluating LLMs on their ability to retrieve causal knowledge about known variables constitutes *commonsense causal reasoning*. In contrast, this paper contributes to work that evaluates abstract causal reasoning [Binz and Schulz, 2023, Jin et al., 2024, 2023, Cai et al., 2023, Liu et al., 2024], assessing the ability of LLMs to use axioms of causality to solve tasks involving general or even new variables.

In the vein of causal reasoning, Jin et al. [2023] studied whether LLMs could correctly infer some causal relationships based on conditional independence statements, comparing LLM predictions to those made by an oracle causal discovery algorithm for observational data, where all causal relations cannot be resolved. In contrast to observational causal discovery, this paper focuses on reasoning about interventions. Also focusing on interventions, Jin et al. [2024] introduced CLadder, a comprehensive benchmark that includes the estimation of causal effects from quantitative data. Causal effect estimation is a complex task that requires solving multiple sub-tasks such as: (i) parsing the prompt to extract a causal DAG, (ii) inferring a function that estimates the effect given the DAG, and (iii) applying that function to the given quantitative data. Concurrently, Cai et al. [2023] introduced a task that asks LLMs to output only causal relationships given a tabular dataset that includes variable names. They focus on disentangling the impacts that prior knowledge (e.g., variables names) and quantitative data have on LLM performance. The empirical study we conduct to assess whether LLMs are sensitive to the presence of plausibly memorized causal relations is similar to experiments conducted by Cai et al. [2023]. In contrast to these benchmarks, intervention effects target a narrower question than the general estimation of causal effects, since intervention effects involve binary classification only (i.e., the absence/presence of causal relations in DAGs). We argue that the evaluations we design better isolate causal reasoning from sub-tasks like drawing statistical inferences from quantitative data provided in-context, which both CLadder and the work of Cai et al. [2023] require.

In focusing on intervention effects, we build on the work of Binz and Schulz [2023], who were motivated by prior work in psychology [Waldmann and Hagmayer, 2005] that shows humans weight collected observational evidence and experimental evidence differently when drawing causal conclusions. Binz and Schulz [2023] adapted this psychology study for LLMs, creating prompts that describe observational and post-interventional findings to LLMs to see if they update their beliefs about a system after interventions. They found that GPT-3, unlike human subjects, fared poorly at understanding the implications of interventions. Motivated by their focus on intervention-based reasoning, we significantly expand on the evaluations designed by Binz and Schulz [2023], systemati-

cally generating intervention effects with varying degrees of difficulty to further explore the effects of plausible memorization and shortcuts like relation retrieval.

2 Understanding changes in causal relationships via Intervention Effects

We begin by summarizing causal directed acyclic graphical models (DAGs) and perfect interventions, the two concepts central to this work. After reviewing these concepts, we introduce intervention effects, the binary prediction tasks that serve as key contributions of this paper.

2.1 Background

A causal DAG $G = (V = \{V_1, \ldots, V_n\}, E)$ is defined by a vertex set V that consists of random variables $\{V_1, \ldots, V_n\}$ and directed edges in the set E that represent causal relationships between the variables. A DAG defines a joint distribution over the random variables that factorizes according to the graph,

$$P(V_1, \ldots, V_n) = \prod_{i=1}^{n} P(V_i | \mathrm{Pa}_i^G), \tag{1}$$

where Pa_i^G denotes the parents of V_i in the DAG G. In general, we say that a causal relation exists between two variables u and v when there is a path from u to v in G.

Figure 2 in the appendix illustrates the three causal DAGs that we focus on in this paper: bivariate, confounding, and mediation. In the confounding graph, the variable A causes both B and C, thus confounding our ability to infer the causal effect that B has on C from observational data alone. In the mediation graph, the variable A only has an indirect effect on C via B.

Causal DAGs also entail distributions after interventions to these variables, distinguishing them from standard DAG models. In this paper, we focus on a class of interventions called *perfect interventions*, represented using the operation $\mathrm{do}(V_i = v)$, which means that the variable V_i is set to the value v. We define the distribution over variables post-intervention by modifying G to delete all incoming edges to V_i, severing the links between V_i and its parents.

2.2 Intervention effects

To formalize intervention effects, the key binary classification task that we consider in this work, we begin by defining causal relations in DAGs precisely.

Definition 1 (Causal relation) *Given a causal DAG $G = (V, E)$, we say that there exists a causal relation between V_u and V_v in DAG G if there exists a directed path $e_{ua} \to \ldots \to e_{bv}$ in G from V_u to V_v (where each edge e_{ij} along the path is in G). We define an indicator variable $C_{uv}(G) \in \{0, 1\}$, such that $C_{uv}(G) = 1$ if and only if there is a causal relation between V_v and V_u.*

In words, a causal relation captures whether or not a variable exerts an indirect or direct causal influence on another variable in a particular DAG G that contains these variables.

Definition 2 (Intervention effect) *Given a causal DAG $G = (V, E)$, a variable V_i on which we perform a perfect intervention captured by $\mathrm{do}(V_i = *)$ (where we use "$*$" to indicate that we do not care about the value that V_i is set to), and a query causal relation C_{uv}, an **intervention effect** (IE) defines a binary classification task as follows,*

$$\mathrm{IE}_i^G(C_{uv}) = C_{uv}(G) - C_{uv}(G^i), \tag{2}$$

where the DAG G^i is the modification of DAG G under an intervention to the variable V_i.

An $\mathrm{IE}_i^G(C_{uv})$ is defined with respect to a DAG G and intervention target V_i, and assigns each causal relation C_{uv} to a binary label of 1 or 0. When a causal relation C_{uv} in G is modified by a perfect intervention on V_i, i.e., the relation C_{uv} is different in the modified DAG G^i, $\mathrm{IE}_i^G(C_{uv})$ is 1 (and 0 otherwise). Intuitively, since interventions sever edges between a variable and its parents, $\mathrm{IE}_i^G(C_{uv})$ is 1 when the intervention target is a variable along the path from u to v (including v itself).

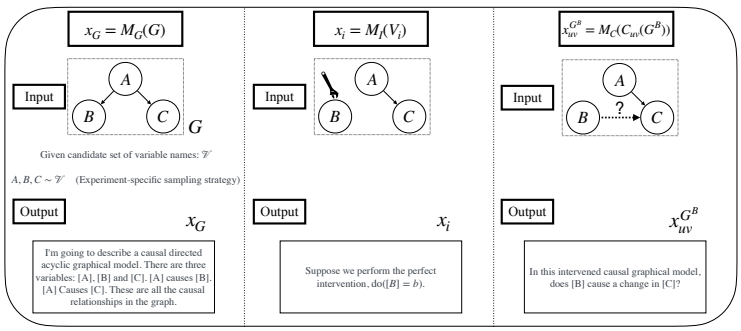

Figure 1: An illustration of the mapping functions that verbalize the information for a single intervention effect estimation task into a prompt.

Given a causal DAG G, each possible intervention that we can perform defines a different classification task over causal relations, defined by a particular $\mathrm{IE}_i^G(\cdot)$, which serves as a labeling function for that task. We focus on the three causal DAGs in Figure 2 since they sufficiently capture many real scenarios without introducing unnecessary complexity that could render the empirical findings ambiguous. The key goal of this paper, addressed in the next section, is to evaluate LLMs on these IE classification tasks in a zero-shot way (i.e., without training examples) by verbalizing the tasks as prompts.

3 Evaluating LLMs on Intervention Effects

To evaluate an LLM on a binary classification problem defined by $\mathrm{IE}_i^G(C_{uv})$, we need to verbalize the causal DAG G, the concept of an intervention as well as the intervention target V_i, and the causal relation C_{uv} of interest. Figure 1 illustrates how we verbalize these three steps to generate a complete prompt.

Step 1: Generating variable names. To verbalize an input causal DAG $G = (V, E)$, we select names for the variables V and describe each edge $E_{ij} \in E$ using the format "[i] causes [j]." We further specify that these are all known causal relationships in the graph to avoid ambiguities, e.g., around the presence of unobserved confounding variables. In the empirical studies, we assess multiple ways of choosing variable names, designing studies to tease apart how well LLMs generalize to novel contexts instead of relying on variable names and facts that were potentially encountered during training.

Step 2: Describing an intervention. In the second step, we specify that a perfect intervention is performed, during the notation $\mathrm{do}(V_i = v)$ since the do-operator could have been encountered during training. To further study the robustness of LLMs to memorizing facts from training, the empirical studies include experiments that use random strings to describe do-operator.

Step 3: Verbalizing the binary classification problem. In the final step, we verbalize the target $\mathrm{IE}_i^G(C_{uv})$, using the phrase "does [u] cause a change in [v]?" to query the presence or absence of a causal relation between u and v.

Evaluation metric. After verbalizing an intervention effect as a prompt to an LLM, we receive a yes/no response that we refer to as $\hat{\mathrm{IE}}_i^G(C_{uv})$. Although prediction accuracy is a natural evaluation metric in this setting, it can misleading in this setting. To see this, we begin by noting that the LLM forms some belief about the presence or absence of a causal relation in a DAG G, denoted $\hat{C}_{uv}(G)$. Consider the scenario where,

$$C_{uv}(G) = 1,\ C_{uv}(G^i) = 1,$$
$$\hat{C}_{uv}(G) = 0,\ \hat{C}_{uv}(G^i) = 0.$$

In this example, a target causal relation C_{uv} is true in both the base causal DAG G and its modified, post-intervention counterpart G^i, but the LLM incorrectly predicts that these causal relations are false under both graphical scenarios. The accuracy metric misleads us when the LLM correctly predicts that the causal relation does not vary, but does not parse the causal relation correctly in either graphical scenario. Thus, to ensure that the accuracy is 0 in such cases, we slightly modify the

Table 1: **IE prediction accuracy on the Random Char benchmark**. GPT-4 variants are the best performing models, while LLaMA-2 appears to struggle with interventional reasoning. Performances that are significantly($\alpha = 0.05$) worse than 80% accuracy are shown in red and top performances are indicated by bolded figures.

Graph Type	**Bivariate**		**Confounding**			**Mediation**		
Intervened Variable	A	B	A	B	C	A	B	C
GPT-3.5	0.83 ± 0.08	0.87 ± 0.06	0.80 ± 0.09	0.69 ± 0.12	0.36 ± 0.09	0.58 ± 0.11	0.36 ± 0.12	0.67 ± 0.12
GPT-4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.78 ± 0.09	0.82 ± 0.08	0.96 ± 0.03
GPT-4-turbo	$\mathbf{1.0 \pm 0.0}$	$\mathbf{0.97 \pm 0.03}$	$\mathbf{0.96 \pm 0.03}$	$\mathbf{0.93 \pm 0.05}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{0.98 \pm 0.02}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$
GPT-4o	1.0 ± 0.0	1.0 ± 0.0	0.98 ± 0.02	0.91 ± 0.04	1.0 ± 0.0	0.93 ± 0.05	1.0 ± 0.0	0.87 ± 0.06
LLaMA-2	0.50 ± 0.12	0.40 ± 0.12	0.56 ± 0.12	0.53 ± 0.11	0.16 ± 0.06	0.69 ± 0.09	0.56 ± 0.12	0.64 ± 0.12
LLaMA-3	0.80 ± 0.09	0.83 ± 0.06	0.51 ± 0.1	0.47 ± 0.08	0.96 ± 0.03	0.38 ± 0.1	0.53 ± 0.08	0.4 ± 0.1

accuracy metric to be,

$$\mathbf{1}[\mathrm{IE}_i^G(C_{uv}) = \hat{\mathrm{IE}}_i^G(C_{uv})] \cdot \mathbf{1}[C_{uv}(G) = \hat{C}_{uv}(G)] \tag{3}$$

Now, an LLM's prediction is only considered correct when it understands the DAG G correctly and accurately solves an IE problem. To infer the belief $\hat{C}_{uv}(G)$ that an LLM has about a causal relation, we make a minor modification to the prompt in Figure 1, omitting the second step that verbalizes an intervention, and asking the LLM about a causal relation but in the "observed causal graphical model" instead of the intervened one. These additional prompts allow us to evaluate LLMs on IEs.

In the next section, we investigate several research questions about LLMs and intervention reasoning ability using the proposed framework to define a suite of IE problems.

4 Empirical Analysis

To study the ability of LLMs to understand interventions, we use the introduced framework to develop three benchmarks that differ in how variable names are selected to verbalize IEs:

1. **Random Char**: Variable names are randomly chosen English characters.
2. **Tübingen** : Variables names are chosen from entities that appear in the Tübingen pairs (TP) dataset of causal relations [Mooij et al., 2016] such that some causal relations exist in the TP data.
3. **Random Tübingen** : Variable names are chosen from entities that appear in the TP dataset so that no causal relations exist in the TP data.

See Appendix A.2 for an illustration of the full details about the datasets. For each benchmark and IE, we sample variable names fifteen times to report significant differences.

We investigate four research questions (RQs) on the proposed benchmarks, studying four LLMs: GPT-3.5, GPT-4, GPT-4-turbo [OpenAI, 2023], and LLaMA-2 [Touvron et al., 2023]. Unless otherwise specified, for each IE, we aggregate results over the enumerated causal relations. We present the findings for the first two questions in the main paper and discuss the remaining in Appendix B

RQ1: How accurate are at LLMs at IEs generated with random characters as variables? To study this first question, we evaluate the performance of the four LLMs on the Random benchmark. Table 1 summarize these results. We see that GPT-based models perform notably better than the LLaMA models, with GPT-4-turbo demonstrating near-perfect accuracy across all effects. LLaMA-2 and LLaMA-3's performance suggests that they do not reliably model interventions. In what follows in the main paper, we focus on results for GPT models, deferring LLaMA results to Appendix B.2.

RQ2: To what extent is LLM performance affected by possibly memorized causal relations? While we might be tempted to conclude that GPT-4 reliably predicts changes to models after interventions are performed, we consider spurious factors that can affect model performance. In particular, Kiciman et al. [2024] found that GPT models reliably retrieved information about TP causal relations, suggesting that these relations could have been included in the training data for LLMs. (We reproduce their findings in Appendix Table 4.) This leads to a worrying possibility:

Table 2: For causal relations that appear in TP, and for interventions that modify these relations, worse performance on the Tübingen benchmark compared to performance on Random Char or Random Tübingen provides evidence that LLMs might be relying on memorized facts to achieve high accuracy on IEs. Bolded figures indicate performances that are significantly ($\alpha = 0.05$) worse than the corresponding performance on Tübingen .

Model	Graph	Intervened Variable	Random Char	Tübingen	Random Tübingen
GPT-3.5	Bivariate	B	0.8 ± 0.1	0.8 ± 0.1	0.53 ± 0.13
	Confounding	C	0.47 ± 0.13	0.4 ± 0.13	0.47 ± 0.13
	Mediation	B	0.4 ± 0.13	0.4 ± 0.13	0.4 ± 0.13
		C	0.67 ± 0.12	0.67 ± 0.12	0.6 ± 0.13
GPT-4	Bivariate	B	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Confounding	C	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Mediation	B	0.8 ± 0.1	$\mathbf{0.73 \pm 0.11}$	0.93 ± 0.06
		C	0.93 ± 0.06	1.0 ± 0.0	0.93 ± 0.06
GPT-4-turbo	Bivariate	B	0.93 ± 0.06	$\mathbf{0.8 \pm 0.1}$	0.73 ± 0.11
	Confounding	C	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Mediation	B	1.0 ± 0.0	1.0 ± 0.0	0.93 ± 0.06
		C	1.0 ± 0.0	1.0 ± 0.0	0.93 ± 0.06
GPT-4o	Bivariate	B	1.0 ± 0.0	0.93 ± 0.06	0.73 ± 0.11
	Confounding	C	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Mediation	B	1.0 ± 0.0	1.0 ± 0.0	0.87 ± 0.09
		C	0.93 ± 0.06	0.87 ± 0.09	0.8 ± 0.1

after interventions, could LLMs fail to update their beliefs about causal relations that they have potentially memorized? To study this, we consider only the subset of causal relations that appear in TP and the interventions that *sever these relations*. If an LLM achieves good performance in general by having memorized some known causal relations, then it would achieve worse performance on Tübingen for the selected IEs, which sever known causal relations and go against the LLM's learned biases, compared its performance on Random or Random Tübingen , where post-interventional causal relations do not directly contradict known facts. In Appendix B.2, we specify which causal relations appear in TP for each causal DAG we study, and which interventions modify these relations. Table 2 summarizes the results. Bolded figures indicate that the performance on Tübingen drops significantly (with $\alpha = 0.05$) compared to performance on either of the other benchmarks. Interestingly, GPT do not show evidence of relying purely on memorized causal relations, since performance is overall comparable across the benchmarks. Further, the models seem to struggle with the Random Tübingen benchmark more than they do with the other benchmarks, leading to a question about whether they are in fact sensitive to variations on the TP dataset.

5 Discussion and Limitations

The goal of this paper is to introduce a causal reasoning benchmark that stress-tests the ability of LLMs to accurately predict how knowledge should be updated after interventions are performed, without conflating other aspects of reasoning such as statistical inference on quantitative data. Our findings are optimistic, but we believe that nevertheless, they underscore the continued need for benchmarks and studies that evaluate varied aspects of abstract causal reasoning in LLMs, especially if practitioners wish to use LLMs to generate candidate decisions.

While the intervention effect prediction task we define in this paper has the benefit of being easy to evaluate, since it requires binary responses, the findings that this task can suggest are also limited. For example, IE prediction cannot help us assess how accurately LLMs perform *causal identification*, the process of deciding which causal inferences can be made given a causal DAG. Moreover, we focus on evaluation in this paper and do not propose methods for improving causal reasoning in LLMs via few-shot learning or fine-tuning. Both of these limitations point to future research directions that we think are worth exploring.

References

[1] Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2218523120.

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

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

[4] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[5] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? In *Forty-first International Conference on Machine Learning*, 2024.

[6] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? *arXiv:2306.05836*, 2023.

[7] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[9] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=mqoxLkX210. Featured Certification.

[10] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[11] Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *arXiv:2402.17644*, 2024.

[12] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

[13] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023.

[14] Arkil Patel, Satwik Bhattamishra, Siva Reddy, and Dzmitry Bahdanau. Magnifico: Evaluating the in-context learning ability of large language models to generalize to novel interpretations. *arXiv preprint arXiv:2310.11634*, 2023.

[15] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

[17] Michael R Waldmann and York Hagmayer. Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2):216, 2005.

A Experiment Setup Details

A.1 Causal DAGs

We consider 3 basic causal DAGs: Bivariate, Confounding and Mediation as shown in Figure 2

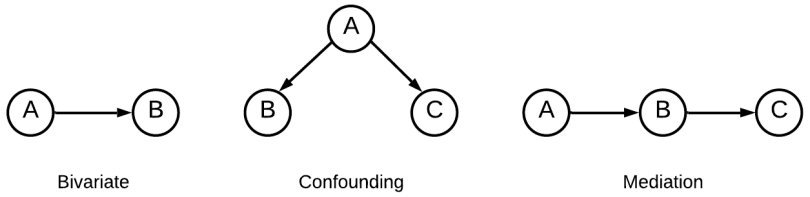

Figure 2: **Causal DAGs.** In the empirical studies, we define intervention effects based on three causal DAGs: bivariate, confounding and mediation graphs.

A.2 Dataset Generation

Random Char (R). In the causal graphs used in this benchmark, the nodes are labeled with arbitrary English letters (E.g. 'z', 'b') that are independently generated for each graph.

Tübingen (T). For the Tübingen benchmark, we select the graph nodes from the Tübingen pairs (TP) dataset in the following manner:

- **Bivariate.** We sample cause-effect pairs from the TP dataset and each cause-effect pair is used to generate a graph.
  Example - Altitude (A) - Temperature (B) is a cause-effect pair in the TP dataset, and the input graph states: *Altitude causes temperature*.
- **Confounding.** For variable B and C, we first sample cause-effect pairs and then for sampled pair we randomly select a variable (for A) from the TP dataset that it is different from the variables in the corresponding cause-effect pair.
  Example - Altitude (B) - Temperature (C) is the sampled cause-effect pair and Age (A) is the randomly selected variable from the TP dataset. The relationships in the input graph are as follows: *Age causes altitude ; Age causes temperature*.
- **Mediation.** For variable A and C, we first randomly sample cause effect pairs and then for sampled pair we randomly select a variable (for B) from the TP dataset that it is different from the variables in the corresponding cause-effect pair.
  Example - Altitude (A) - Temperature (C) is the sampled cause-effect pair and Age (B) is the randomly selected variable from the TP dataset. The relationships in the input graph are as follows: *Altitude causes age ; Age causes temperature*.

The rationale behind defining the Tübingen dataset in this manner is elaborated in Appendix B.2.

Random Tübingen (RT). Similar to the Tübingen case, we select the graph nodes from the set of variables present in the TP dataset. However, none of the causal relations defined in the graphs are present in the TP dataset. Instead the causal relations defined are between randomly selected variables, as follows:

- **Bivariate.** We randomly sample two unrelated variables from the TP dataset and define a cause-effect relationship in the input graph.
  Example - Age (A) - Temperature (B) are variables without a cause-effect relation in the TP dataset, and the input graph states: *Age causes Temperature*.
- **Confounding.** We randomly sample 3 variables from the TP dataset that no two variables selected have a causal relation in the TP dataset and we define the graph.
  Example - Altitude (A) - Horsepower (B) - Cement (C) are randomly selected variables from the TP dataset. No two variables have a causal relationship in the TP dataset. The

relationships in the input graph is as follows: *Altitude causes Horsepower ; Altitude causes Cement*.

- **Mediation.** We select the 3 variables exactly as we did for the confounding case and define the graph.
  Example - Altitude (A) - Horsepower (B) - Cement (C) are randomly selected variables from the TP dataset. The relationsihips in the input graph is as follows: *Altitude causes Horsepower ; Horspower causes Cement*.

A.3 Details regarding LLMs

A.3.1 Querying LLMs.

To query the GPT models, we used the OpenAI API [1] through the Langchain interface [2]. Meanwhile, VLLM library [3] was used for fast inference from LLaMA models.

Regarding compute resources, since the GPT models are hosted remotely by OpenAI, we were able to make API calls locally with minimal CPU usage. Conversely, for LLaMA models, we first had to load the model onto a cluster equipped with a GPU (A100/80 GB RAM) before submitting our queries.

A.3.2 LLM Output.

To ensure that LLMs respond with a yes or no to queries, we formalized a response format requiring LLMs to encapsulate their yes or no answers within an `<answer></answer>` tag. If the LLM does not adhere to this format, we initiate a retry, instructing it to comply with the required format. After the first retry, we relax the response format requirements, expecting only a yes or no answer. If after 10 retries the LLM fails to meet this criterion, we mark the attempt as a failure and attribute zero accuracy on the intervention effect prediction task.

A.3.3 Model Ids.

The specific model-ids of the LLMs are:

- GPT-3.5: `gpt-3-turbo-16k`
- GPT-4: `gpt-4`
- GPT-4-turbo: `gpt-4-turbo`
- GPT-4o: `gpt-4o`
- LLaMA-2: `llama-2-7b`
- LLaMA-3: `llama-3.1-8b`

A.4 Substitution Task

For this task, we first describe the concept of interventions to LLM and refer to them as some arbitrary string of characters (E.g. 'xyz'). We then follow the same prompt template (Figure 1) to test the intervention effect performance, except we never use the word intervention again and replace it with the chosen substitution word. The prompt verbalization for this task is illustrated in Figure 3.

A.5 Ground Truth Intervention Effects

We provide the ground truth value for every intervention effect task in our analysis in Table 3, i.e., we provide the true change in the causal relations after interventions.

This should help the reader understand clearly what queries we considered for the different research questions in our analysis.

[1] `https://openai.com/blog/openai-api`

[2] `https://python.langchain.com/docs/expression_language/interface`

[3] `hhttps://docs.vllm.ai/en/latest/`

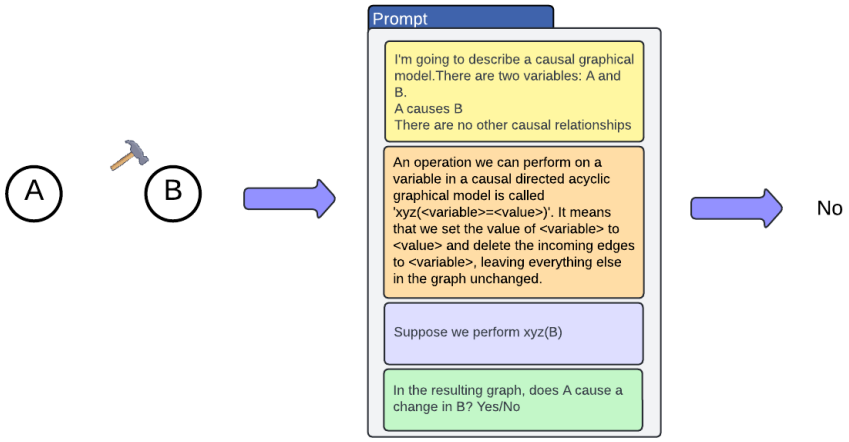

Figure 3: **Prompt design for intervention reasoning in the substitution task.** Instead of using the word intervention directly in the prompt, we illustrate the concept of interventions and define it with a random word, for instance, operation 'xyz' in the prompt template above.

Table 3: Ground truth intervention effects for all the scenarios (causal graph, intervention variable, causal relation) considered in our benchmark.

Graph	Intervention	Questions	Ground Truth IE
Bivariate	A	$A \to B$	0
		$B \to A$	0
	B	$A \to B$	1
		$B \to A$	0
Confounding	A	$A \to B$	0
		$A \to C$	0
		$B \to C$	0
	B	$A \to B$	1
		$A \to C$	0
		$B \to C$	0
	C	$A \to B$	0
		$A \to C$	1
		$B \to C$	0
Mediation	A	$A \to B$	0
		$A \to C$	0
		$B \to C$	0
	B	$A \to B$	1
		$A \to C$	1
		$B \to C$	0
	C	$A \to B$	0
		$A \to C$	1
		$B \to C$	1

B Additional Results

We now go over the empirical studies in detail and provide the results for the remaining research questions.

B.1 Reproducing results on the Tübingen dataset

We reproduced the results of Kiciman et al. [2024] on the Tübingen dataset as shown in Table 4. Since their task did not involve the LLM to reason about the effect of interventions, we term this task of inferring relations from the input graph in the prompt as relation retrieval.

Table 4: Relation Retrieval accuracy of GPT models on the **Tübingen** dataset.

Model	Accuracy
GPT-3 (*text-davinci-003*)	0.80
GPT-3.5	0.89
GPT-4	0.96

Table 5: **Relation Retrieval accuracy of models on Random benchmark.** The GPT models show good performance but LLaMA-2 performs poorly. Hence, as per our criteria we drop the LLaMA-2 model for analysis in RQ3.

Model	Bivariate	Confounding	Mediation
GPT-3.5	1.0 ± 0.0	1.0 ± 0.0	0.98 ± 0.02
GPT-4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
GPT-4-turbo	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
GPT-4o	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
LLaMA-2	0.73 ± 0.11	0.59 ± 0.13	0.84 ± 0.09
LLaMA-3	1.0 ± 0.0	1.0 ± 0.0	0.94 ± 0.06

B.2 Research Questions

RQ1: How accurate are LLMs at predicting the effects of interventions?

To understand how effective LLMs are at IE prediction, we focused on the **Random (R)** benchmark in the main paper (Table 1), as we wanted to remove any distracting effect of semantically meaningful entities as graph nodes. We now present the same results on the **Tübingen (T)** and **Random Tübingen (RT)** benchmarks as well in Table 6.

We find that the results in both the cases are very similar to the case with the **Random** benchmark; GPT-4-turbo performs the best and GPT models outperform LLaMA models by a big margin.

RQ2: To what extent is LLM performance affected by memorized causal relations? In **RQ1**, we computed the IE performace by considering all the causal relationship queries for each of the interventions. In the case of **RQ2**, we only consider specific causal relations that could be potentially memorized by the LLMs and interventions that sever these relations. For:

1. Bivariate DAG, we consider $A \to B$ with the intervention on B.
2. Confounding DAG, we consider $B \to C$ with the intervention on C.
3. Mediation DAG, we consider $A \to C$ with an intervention on B and C separately.

The **Tübingen** benchmark (Appendix A.2) is defined in such a way that, all of the causal relations under consideration above are cause-effect pairs present in the TP dataset that could be potentially memorized by the LLM.

We provide results for role of memorization in LLaMA models in Table 7. Given their overall poor IE performance in the random benchmark, it is difficult to draw conclusions about the impact of memorization but they have similar performance across the benchmarks.

Table 6: **IE prediction accuracy on all three benchmarks**. GPT-4 variants are the best performing models, while LLaMA models struggle with interventional reasoning.

Graph Type	Benchmark	**Bivariate**		**Confounding**			**Mediation**		
Intervened Variable		A	B	A	B	C	A	B	C
GPT-3.5	R	0.83 ± 0.08	0.87 ± 0.06	0.8 ± 0.09	0.69 ± 0.12	0.36 ± 0.09	0.58 ± 0.11	0.36 ± 0.12	0.67 ± 0.12
	T	0.87 ± 0.06	0.83 ± 0.06	0.82 ± 0.05	0.67 ± 0.11	0.31 ± 0.07	0.64 ± 0.08	0.42 ± 0.1	0.67 ± 0.12
	RT	0.83 ± 0.1	0.7 ± 0.12	0.87 ± 0.09	0.67 ± 0.12	0.47 ± 0.13	0.64 ± 0.12	0.33 ± 0.12	0.63 ± 0.12
GPT-4	R	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.78 ± 0.09	0.82 ± 0.08	0.96 ± 0.03
	T	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.84 ± 0.07	1.0 ± 0.0
	RT	1.0 ± 0.0	1.0 ± 0.0	0.93 ± 0.06	0.87 ± 0.09	1.0 ± 0.0	0.98 ± 0.04	0.97 ± 0.05	0.97 ± 0.05
GPT-4-turbo	R	1.0 ± 0.0	0.97 ± 0.03	0.96 ± 0.03	0.93 ± 0.05	1.0 ± 0.0	0.98 ± 0.02	1.0 ± 0.0	1.0 ± 0.0
	T	0.93 ± 0.04	0.9 ± 0.05	0.62 ± 0.09	0.78 ± 0.08	0.96 ± 0.03	0.69 ± 0.09	0.96 ± 0.04	0.96 ± 0.04
	RT	0.8 ± 0.1	0.8 ± 0.1	0.73 ± 0.11	0.6 ± 0.13	1.0 ± 0.0	0.8 ± 0.1	0.97 ± 0.05	0.97 ± 0.05
GPT-4o	R	1.0 ± 0.0	1.0 ± 0.0	0.98 ± 0.02	0.91 ± 0.04	1.0 ± 0.0	0.93 ± 0.05	1.0 ± 0.0	0.87 ± 0.06
	T	0.87 ± 0.06	0.97 ± 0.03	0.89 ± 0.04	0.93 ± 0.03	1.0 ± 0.0	0.84 ± 0.05	0.96 ± 0.04	0.87 ± 0.06
	RT	0.87 ± 0.09	0.83 ± 0.1	0.9 ± 0.08	0.8 ± 0.1	1.0 ± 0.0	0.93 ± 0.06	0.93 ± 0.06	0.87 ± 0.09
LLaMA-2	R	0.5 ± 0.12	0.4 ± 0.12	0.56 ± 0.11	0.53 ± 0.11	0.16 ± 0.06	0.69 ± 0.09	0.56 ± 0.12	0.64 ± 0.12
	T	0.4 ± 0.12	0.2 ± 0.08	0.42 ± 0.11	0.4 ± 0.11	0.09 ± 0.04	0.47 ± 0.12	0.31 ± 0.12	0.44 ± 0.1
	RT	0.43 ± 0.13	0.23 ± 0.11	0.47 ± 0.13	0.47 ± 0.13	0.2 ± 0.1	0.62 ± 0.13	0.4 ± 0.13	0.7 ± 0.12
LLaMA-3	R	0.8 ± 0.1	0.83 ± 0.1	0.47 ± 0.13	0.6 ± 0.13	0.87 ± 0.09	0.38 ± 0.13	0.5 ± 0.13	0.47 ± 0.13
	T	0.73 ± 0.11	0.87 ± 0.09	0.53 ± 0.13	0.53 ± 0.13	1.0 ± 0.0	0.38 ± 0.13	0.6 ± 0.13	0.27 ± 0.11
	RT	0.67 ± 0.12	0.87 ± 0.09	0.47 ± 0.13	0.4 ± 0.13	1.0 ± 0.0	0.4 ± 0.13	0.53 ± 0.13	0.43 ± 0.13

Table 7: **IE prediction performance of LLaMA models on specific scenarios for all the benchmarks to understand the role memorization.** LLaMA models demonstrate relatively similar performance across the three benchmarks.

Model	**Graph**	**Intervened Variable**	**Random**	**Tübingen**	**Random Tübingen**
LlaMA-2	Bivariate	B	0.4 ± 0.13	0.4 ± 0.13	0.27 ± 0.11
	Confounding	C	0.0 ± 0.0	0.07 ± 0.06	0.33 ± 0.12
	Mediation	B	0.33 ± 0.12	0.33 ± 0.12	0.4 ± 0.13
		C	0.53 ± 0.13	0.53 ± 0.13	0.6 ± 0.13
LlaMA-3	Bivariate	B	0.67 ± 0.12	0.73 ± 0.11	0.8 ± 0.1
	Confounding	C	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Mediation	B	0.67 ± 0.12	0.53 ± 0.13	0.67 ± 0.12
		C	0.53 ± 0.13	0.33 ± 0.12	0.4 ± 0.13

RQ3: Could LLMs be learning a shortcut to predict intervention effects? Consider the confounding DAG (Figure 2) and the causal relation $C_{BC}(G)$ which does not change under intervention on the variable A. In these examples, LLMs that accurately parse causal relations from text descriptions of DAGs would also obtain accurate IE estimates. Hence, predicting causal relations from the input graph in text – call this task relation retrieval – offers a *shortcut*: an LLM can attend to tokens in the context to solve relation retrieval and still perform good at IE estimation, thereby confounding the conclusions that can be drawn with this benchmark.

However, the intervention effects we defined to study RQ2 offer an insight into how we can disentangle relation retrieval and accurate IE prediction. Notice that the IEs we defined to study RQ2 characterize scenarios where causal relations differ between the base DAG G and the post-intervention DAG. Thus, LLMs that rely only on relation retrieval cannot accurately estimate IEs. Building on this insight, we divide all the IEs $\kappa^i_G(C_{uv})$ into two groups:

1. $\mathbf{IE = 0}$: Graph doesn't change as a result of intervention; $C_{uv}(G) = C_{uv}(G^i)$.
2. $\mathbf{IE = 1}$: Graph changes as a result of intervention; $C_{uv}(G) \neq C_{uv}(G^i)$

Table 3 clearly categorizes the causal relations under intervention into these two groups.

Note that drop in performance on group ($IE = 1$) as compared to the group ($IE = 0$) can indicate reliance on shortcuts based on relation retrieval. We consider the average performance of the LLMs on the effects in each group, selecting only those LLMs which achieve an accuracy ≥ 0.95 on relation retrieval (which we report in Appendix Table 5). We focus on the **Random** benchmark to exclude any impacts due to memorized variable names. Table 8 summarizes this study. We find that the general trend does not show strong reliance on shortcuts across LLMs; only GPT-3.5 has a significant drop

Table 8: **IE prediction performance across sub-cases to isolate the effect of relation retrieval on Random benchmark**. LLMs do not significantly rely on shortcuts related to relation retrieval from the input prompt. Since intervening on variable A never changes the causal graph in any scenario, we don't consider them for this analysis.

Graph Type		**Bivariate**	**Confounding**		**Mediation**	
Intervened Variable		B	B	C	B	C
GPT-3.5	IE = 0	0.93 ± 0.06	0.6 ± 0.13	0.47 ± 0.13	0.33 ± 0.12	0.67 ± 0.12
	IE = 1	0.8 ± 0.1	0.67 ± 0.12	0.2 ± 0.1	0.37 ± 0.12	0.67 ± 0.12
GPT-4	IE = 0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.87 ± 0.09	1.0 ± 0.0
	IE = 1	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.8 ± 0.1	0.93 ± 0.06
GPT-4-turbo	IE = 0	1.0 ± 0.0	0.96 ± 0.05	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	IE = 1	0.93 ± 0.06	0.93 ± 0.06	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
GPT-4o	IE = 0	1.0 ± 0.0	0.87 ± 0.09	1.0 ± 0.0	1.0 ± 0.0	0.87 ± 0.09
	IE = 1	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.87 ± 0.09
LlaMA-2	IE = 0	0.4 ± 0.13	0.5 ± 0.13	0.0 ± 0.0	0.33 ± 0.12	0.67 ± 0.12
	IE = 1	0.4 ± 0.13	0.53 ± 0.13	0.13 ± 0.09	0.33 ± 0.12	0.6 ± 0.13
LlaMA-3	IE = 0	1.0 ± 0.0	0.4 ± 0.13	1.0 ± 0.0	0.6 ± 0.13	0.27 ± 0.11
	IE = 1	0.67 ± 0.12	0.6 ± 0.13	0.87 ± 0.09	0.5 ± 0.13	0.47 ± 0.13

Table 9: **IE prediction accuracy on the Random benchmark for the substitution task**. The performance of GPT-3.5 & GPT-4 is worse under substitution compared to the non-substitution case (Table 1), while GPT-4-turbo does not show significant change.

Graph Type	**Bivariate**		**Confounding**			**Mediation**		
Intervened Variable	A	B	A	B	C	A	B	C
GPT-3.5	0.67 ± 0.12	0.83 ± 0.06	0.56 ± 0.11	0.58 ± 0.11	1.0 ± 0.0	0.62 ± 0.10	0.44 ± 0.10	0.4 ± 0.12
GPT-4	0.97 ± 0.03	1.0 ± 0.0	0.89 ± 0.05	0.82 ± 0.07	1.0 ± 0.0	0.96 ± 0.03	0.76 ± 0.10	1.0 ± 0.0
GPT-4-turbo	0.9 ± 0.07	0.97 ± 0.03	0.96 ± 0.03	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
GPT-4o	0.67 ± 0.12	0.83 ± 0.1	0.67 ± 0.12	0.73 ± 0.11	1.0 ± 0.0	0.76 ± 0.11	0.83 ± 0.1	0.73 ± 0.11
LLaMA-2	0.57 ± 0.13	0.37 ± 0.12	0.6 ± 0.13	0.6 ± 0.13	0.73 ± 0.11	0.62 ± 0.13	0.6 ± 0.13	0.43 ± 0.13
LLaMA-3	0.67 ± 0.12	0.83 ± 0.1	0.6 ± 0.13	0.4 ± 0.13	1.0 ± 0.0	0.4 ± 0.13	0.67 ± 0.12	0.33 ± 0.12

in its relative performance on group ($IE = 1$) vs group ($IE = 0$) for the confounding DAG case. Since, the LLaMA-2 model does not satisfy our constraint of high relation retrieval accuracy, we don't consider it for further analysis.

RQ4: Are LLMs robust to descriptions of interventions in-context? Consider the verbalization of an intervention in Figure 1. It mentions the $\mathrm{do}(\cdot)$ operator and refers to a "perfect intervention" to prompt an LLM to rely on facts about causal reasoning that could have appeared in the training dataset. We ask whether LLMs can achieve the same performance on the intervention effects in the random benchmark if we varied the verbalization to instead "teach" an LLM [Patel et al., 2023] about a new graphical operation that behaves identically to an intervention. We randomly generate strings to instantiate this operation. Figure 3 in the appendix illustrates how we generate prompts for this task which we refer to as the substitution task. Table 9 summarizes IE prediction accuracy for the substitution task on the **Random** benchmark. Contrasting these results against Table 1, we see that for GPT-3.5, GPT-4 and GPT-4o, the performance generally suffers. However, GPT-4-turbo appears to be robust to changes in the way interventions are described.