Investigating Causal Reasoning in Large Language Models

Atul Rawal,
Towson University

Adrienne Raglin
DEVCOM Army Research Laboratory

Qianlong Wang
Old Dominion University

Ziying Tang
Towson University

Abstract

With the widespread utilization of LLMs for a plethora of applications, challenges associated with trust, safety, and fairness need to be addressed. One of these challenges for LLMs relates to their capability for causal reasoning. This can be critical for successfully utilizing LLMs in sensitive applications such as biomedical, healthcare, technology, law, and government. To address this, we investigate LLMs for whether they can identify cause and effect relations using a combination of benchmarked causal datasets (Tuebingen dataset), image datasets (Animals with Attributes 2), and LLM benchmark dataset (CRASS). For causal discovery, we present LLMs' ability to identify causal relations and generate causal graphs given an observational dataset. For causal inference, we investigate whether they can generate counterfactual reasoning on natural language questions. Using multi-modal data, experimental results demonstrate the capability of LLMs to complement and contribute to the growing field of causal reasoning for AI systems by aiding in causal discovery and treatment effect estimation methods based on traditional techniques. However, we also highlight the limitations of LLMs to generate causal reasoning as the data complexity is increased.

1 Introduction

Commercial and open-source Large Language Models (LLMs) such as ChatGPT, Gemini, or LlaMA have pushed the envelope for what is possible via GenAI (1; 2; 3; 4). Unlike traditional AI/ML models that are limited in scope, LLMs can have a wider range of functionality. This flexibility in their utility for an array of applications, ranging from general natural language processing (NLP) to more domain-specific tasks, has made them wildly popular in the past few years. However, there are still challenges associated with the use of LLMs, such as safety and trust. LLMs have been shown to give erroneous responses to basic questions and "hallucinate" (5). Even though advanced LLMs are able to perform well on simple NLP tasks, their performance has been proven to be heavily dependent on the prompt entered by the user. So much so, that it gave rise to a new field of research termed "prompt engineering" (6; 7; 8). Finally, the capability of LLMs to generate responses based on both implicit and explicit causal reasoning has been a challenge.

Causal knowledge must be incorporated as a vital component for LLMs as it highlights their capacity to process data and generate responses based on causal reasoning and not simple correlation. LLMs capable of causal reasoning can bridge the gap for artificial reasoning systems to achieve human like thinking. Recent studies have investigated whether LLMs are truly capable of causal reasoning or whether they are simply acting as retrieval augment generator models (9; 10; 11). However, there is a lack of a comprehensive study that investigates the causal capabilities of LLMs for multiple data modalities and complexities. In this paper, we investigate the capabilities of LLMs to generate causal

NeurIPS 2024 Workshop on Causality and Large Models (CaLM).

reasoning from various data modalities and increasing data complexity. The main contributions of the paper are as follows: 1) We adopt ChatGPT as the LLM of choice to evaluate its capability of determining the causal directions, including pairwise causal discovery between two variables and causal discovery between larger number of variables. 2) We evaluate ChatGPT's capability to generate causal reasoning from different data modality, i.e., images from AWA-2 dataset. 3) We evaluate ChatGPTs counterfactual capabilities using the CRASS dataset.

Our simulation results demonstrate that the LLM model is able to generate causal reasoning to a certain extent and even understand the limitations of itself, which would be crucial for LLMs to achieve robust and trustworthy use for the end-users. The paper is organized as follows: Section 2 provides the data and methods used for this study, while Section 3 provides the results and discussion. Section 4 includes concluding remarks.

2 Data & Methods

The data for this study were derived from multiple sources. To investigate the causal reasoning capabilities with tabular datasets, we used the benchmarked causal datasets from the Tuebingen dataset, including temperature-altitude, auto-mpg, and abalone. For investigating the causal capabilities with image data, we utilized the Animals with Attributes 2. Finally, we used the CRASS benchmark dataset to evaluate the causal reasoning capabilities when asked prompt questions.

The Tuebingen dataset contains numerous smaller datasets for causal research. The dataset contains the benchmarked CauseEffectPairs that consists of data for 100 different cause-effect pairs selected from 37 data sets from many domains to identify the ground truth causal directions of all pairs (12). The animals with attributes (AWA) dataset consists of 37322 images of 50 different classes of animals with pre-extracted feature representations for each class (13). All the animal classes are characterized by 85 different attributes, and attributes that are shared across the different classes allow for transferring information between the classes. We used images from the Tiger and Zebra classes and the attributes table to investigate causal reasoning with LLMs. The last dataset used for this study is the CRASS dataset (14). It consists of Premise-Counterfactual Tuples (PCTs) with a hypothetical situation that uses a counterfactual condition against a base premise. The dataset consists of 274 such hypothetical situations with questions to be asked as prompts and multiple counterfactual options, with one being the correct response.

For evaluating LLMs causal reasoning capabilities, we utilized OpenAI's ChatGPT 4o as the LLM of choice. This was done mainly as the authors believe it to be the most advanced LLM currently available on the market. ChatGPT 4o was used to evaluate its causal reasoning capabilities for causal discovery and inference based on a range of causal tasks as described below:

1. Determining causal direction – We evaluated the capability of ChatGPT to determine a) pairwise causal discovery between two variables based on observational data and b) causal discovery between a larger number of variables in an observational dataset. This was done for the DWD data, auto-mpg data, and abalone data from the Tuebingen dataset . With increased data complexity we input the entire auto-mpg, and DWD datasets into ChatGPT and asked it to generate the causal graphs.
2. Counterfactual reasoning based on questions – We utilized the CRASS dataset to evaluate ChatGPTs counterfactual capabilities. We first input a premise into the prompt, and then asked a counterfactual question. We then asked the same counterfactual questions but gave multiple-choice answers and asked it to pick one.
3. Causal reasoning for image datasets – We also evaluated ChatGPTs' capability to generate causal reasoning from images. To do this, we used the benchmarked Animals with Attributes 2 dataset. We input the attribute data directly into ChatGPT and asked it to generate a causal graph based on the attributes for animal identification. We then input images and asked it to identify the animal and give causal reasoning for why it was identified a certain way.

While other studies have shown and highlighted quantitative metrics for the accuracy of LLMs for different aspects of causal reasoning (15; 11; 16), these studies have also highlighted the importance of prompt engineering/tuning. Long et. al. presented mixed results for LLMs' ability to generate DAGs based on different prompts that were engineered to mimic different levels of expertise in the medical field (16). Kiciman et. al. highlighted the difference in output from LLMs for the same question prompts when the prompt is "prepended" with a helpful message. The authors presented a significant increase in output accuracy when the prompt asked a single direct question of the more likely causal direction between two variables while explaining its reasoning (11). For this study, we approached the metrics with a qualitative lens rather than a quantitative one, as the output responses were affected by prompt engineering. So, we believe that instead of a strictly quantitative metric, a more holistic view of LLMs' causal capabilities can be achieved by combining both quantitative and qualitative comparison. To evaluate ChatGPT's capability for causal reasoning, we compared the responses generated for different causal tasks.

For pairwise causal tasks, we compared the output responses to the ground truth obtained from the datasets. With the increased data complexity of combining the auto-mpg and DWD datasets, we compared the output to the ground truth. For the image datasets, we did a qualitative assessment of whether the generated responses were causally logical. With the attribute dataset, we checked to see whether the features highlighted in the generated causal graphs matched the generated response output. Finally, for the counterfactual tasks with multiple choice responses, we compared the generated responses to the correct response from the dataset. We then compared the output responses from prompts without the multiple-choice responses. For the different modalities, the holistic evaluation approach utilized in this study highlights the need for new evaluation frameworks for LLMs, where simple accuracy metrics fail to address the associated challenges such as prompt engineering are taken into consideration.

3 Results & Discussion

The first result we would like to highlight in this chapter, which aims to investigate the capabilities of LLMs to generate causal reasoning, is the simple prompt of asking ChatGPT 4o if it is capable of causal reasoning.

> ***Prompt****: Are you capable of causal reasoning?*
>
> ***Response****: As an AI language model, I can perform various forms of reasoning, including causal reasoning to some extent. However, it's essential to note that my capabilities are limited compared to human reasoning. While I can understand and generate responses based on causal relationships present in the data I've been trained on, my understanding is based on patterns in language rather than true comprehension of causality in the way humans understand it. Therefore, while I can provide insights and explanations based on causal relationships presented in the data, my responses should be taken as suggestions rather than definitive conclusions in critical contexts.*

The response "my understanding is based on patterns in language rather than true comprehension of causality in the way humans understand it" perfectly sums up the capabilities of GenAI/LLMs models to generate causal reasoning based on causal knowledge instead of statistical correlation. These models are trained mostly on large datasets that are generated from web scraping, and not necessarily on specific algorithms that are capable of differentiating between causation and correlation.

For the pairwise causal direction task, Table 1 provides the variables, along with the ground truth and ChatGPT's response. When given the data for two variables, ChatGPT is able to accurately predict the causal direction between them, as shown in Table 1. However, ChatGPT is also able to highlight its own limitations regarding its capabilities for causal reasoning. For example, for the altitude-temperature dataset, when asked to generate a causal graph for the two variables it stated its limitations stating that "*we can make some reasonable assumptions based on general knowledge*" rather than causal knowledge, even when the generated causal graph was correct. This highlights that advanced LLMs such as ChatGPT can derive causal relations between pairwise variables from a given observational dataset, with the understanding that there are limitations. The limitations here are that these causal relations are not being identified by LLMs based on causal knowledge, but rather correlation and learned representations from the vast amount of data it is trained upon.

The next task with the cause-effect datasets was to investigate ChatGPTs' capability to generate causal graphs for datasets with multiple variables. For this, we input the entire auto-mpg, DWD and sachs-protein datasets. These dataset have been utilized in multiple studies and as examples for multiple open-source causal discovery tools such as the causal discovery toolbox, the visual causality analyst, and gCastle (17; 18; 19). For our study we adopted the causal graph from the ground truth provided with the dataset. Table 1 presents the results from the generated causal graphs. Here we evaluate the graphs based on the correct number of nodes and their direction. For each graph the number of nodes is compared to the ground truth along with it's direction between the variables.

Table 1: Accuracy of ChatGPT4o on the different causal tasks.

Causal Task	Dataset	Variables	Ground truth	Output
Pairwise Causal Direction	DWD	Altitude-Temperature	Altitude ->Temperature	Altitude ->Temperature
		Altitude-Precipitation	Altitude ->Precipitation	Altitude ->Precipitation
		Altitude-Sunshine	Altitude ->Sunshine	Altitude ->Sunshine
	Abalone	Rings - Length	Rings ->Length	Rings ->Length
		Rings – Shell weight	Rings ->Shell weight	Rings ->Shell weight
		Rings – Diameter	Rings ->Diameter	Rings ->Diameter
	Auto-Mpg	Displacement- Mpg	Displacement ->Mpg	Displacement ->Mpg
		Horsepower-Mpg	Horsepower ->Mpg	Horsepower ->Mpg
		Weight-Mpg	Weight ->Mpg	Weight ->Mpg
Causal Discovery		**Node Accuracy**	**Direction Accuracy**	**Total Accuracy**
	Auto-MPG	100%	80%	90%
	DWD	66.66%	66.66%	66.66%
	Sachs Protein	21.05%	15.78%	18.41%
Counterfactual	CRASS	**Counterfactual Questions**	**Multiple-Choice Accuracy**	**Non-Multiple-Choice Accuracy**
		20	100%	0%

Figure 1 shows the causal graphs from the ground-truth and the ones generated by ChatGPT for the auto-mpg and DWD datasets. Here we can see that ChatGPT is able to generate causal graphs, but the accuracy of the graphs based on the number of nodes and the direction of the nodes varies for each dataset, ranging from 18% to 90%. The study by Long et. al., highlighted their top performing gpt-3 model was able to achieve 100% accuracy for creating causal graphs for their cancer dataset. However, when using other datasets, their study did not achieve the same accuracy with rates ranging from 33% to 58% for the lowest and 67% to 83% for the highest accuracy. This shows the remarkable ability (albeit limited) of LLMs to create entire causal graphs from raw observational datasets (albeit with limitations on accuracy), helping alleviate some of the burden from researchers who might not have the technical expertise to utilize open-source tools such as DoWhy or CDT. We also asked ChatGPT directly: *"can you create a causal graph for the tuebingen auto-mpg dataset?"* For both the prompts: entering the data manually, and asking it to perform data-scraping, the generated causal graphs were the same. However, there was a difference in the reasoning behind the causal graphs. This once again highlights the effect of prompt engineering for LLMs as two separate prompts with the same dataset gave two separate explanations, even though the final causal graphs were the same.

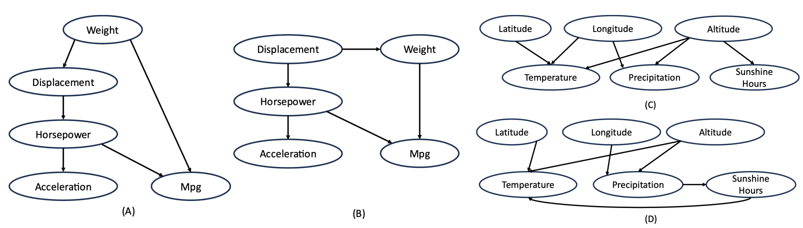


Figure 1: Causal graphs for the (A) auto-mpg dataset based on the ground-truth, (B) created by ChatGPT, (C) DWD dataset based on the ground-truth, and (D) created by ChatGPT.

For the second causal task of counterfactual reasoning for questions, we used the CRASS dataset to ask counterfactual questions and investigate the generated responses. We then asked the same question with the multiple-choice in the response. We repeated the same process with other prompts from the dataset. Here ChatGPT is able to generate multiple scenarios based on the original prompt, all of them reasonable and mimicking human-like thinking. When given the option of multiple-choice answers, it is able to pick the correct one. We repeated this task for 20 counterfactual tasks and the results were the same, with ChatGPT picking the correct answer out of the given choices every time. An examples of prompt and response is provided below:

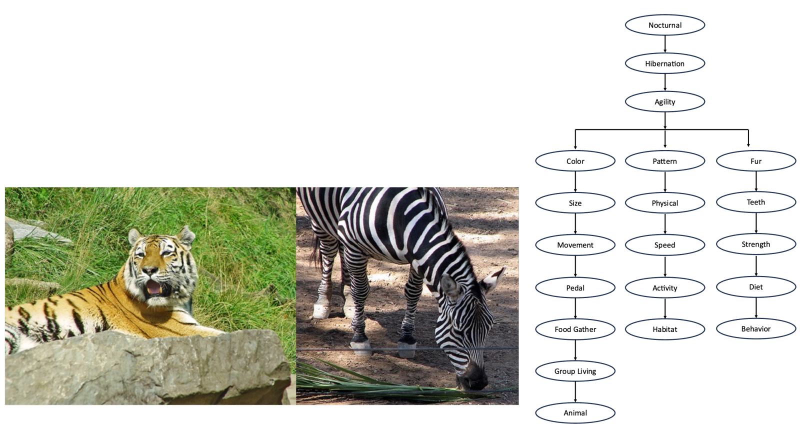
> ***Prompt:*** *A girl kisses a boy. What would've happened if the girl slapped the boy?*
> ***Response:*** *If the girl had slapped the boy instead of kissing him, the scene would have been very different regarding emotions and physical expressions. Here's a potential scenario:*

1. *Emotional Reaction:*
    - *The boy might look shocked, hurt, or confused, displaying a range of emotions from surprise to anger.*
    - *The girl might appear angry, upset, or frustrated, depending on the reason for the slap.*
2. *Physical Expression:*
    - *The boy could be recoiling or holding his cheek, showing the immediate aftermath of the slap.*
    - *The girl's hand might still be in the air from the slap, or she might have a stern or upset expression.*
3. Contextual Implications:
    - The atmosphere would be tense and charged, contrasting the peaceful and affectionate setting of a kiss.
    - Any bystanders or the environment around might reflect this tension, with less serene or harmonious elements.

Prompt: *A girl kisses a boy. What would've happened if the girl slapped the boy? a) The girl would have been angry. b) The girl would have been happy. c) That is not possible. d) Everything would have been fine.*

Response: *If the girl had slapped the boy, the most likely outcome would be: a) The girl would have been angry.*

For the final causal task of causal reasoning based on the input images, we asked ChatGPT to create causal graphs from image data. For the attributes, we input the attribute table for the AWA dataset and asked it to generate a causal graph based on the animal label. Fig. 2(B) displays the causal graph generated by ChatGPT 4o. The explanation given for the generated causal graph is also presented in the Appendix. Because we do not have ground truth for the AWA dataset, we cannot compare it to the causal graph generated by ChatGPT. However, even without ground truth, the generated explanation provides causal reasoning for the graph at a human-like level [Prompt and Response for explanation provided in the appendix]. As the explanation states, the features on the left are considered causal factors that influence the features on the right. So, the color of the animal has an influence on the pattern and the fur type of the animal. However, it also stated the importance of all the features for the animal label "as they collectively contribute to determining the animal species". For the next step in investigating the causal capabilities of LLMs with images, we input images directly into the prompt and asked it to generate causal reasoning from images. We input different images of Zebras and Tigers and asked the causal reasoning for labeling the animal as tiger or zebra [Fig. 2 (A)]. For both the images of Tiger and Zebra, ChatGPT is able to generate causal reasoning to mimic human-like thinking [Prompt and Response for explanation provided in the appendix]. By highlighting the distinctive features of both the animals, such as the stripes, color, and facial features. While we can't measure this type of causal reasoning with quantitative metrics, qualitative analysis of being able to highlight the most prominent features of the two animals mimics human-like thinking.



(A) (B)

Figure 2: (A)Prompt input for images of a Tiger and Zebra. (B)Causal graph generated for AWA dataset based on the attributes.

Prompt*: Can you generate causal reasoning from images?*

Response*: Yes, I can help interpret images for causal reasoning if they contain visual data relevant to causal relationships. If you provide images depicting causal scenarios, I can assist in explaining or generating causal reasoning based on the content of the images.*

For the first causal task of identifying causal direction, ChatGPT was able to successfully identify the correct causal direction between all the pairwise variables tested from the Teubingen dataset. However, when asked to generate a complete causal graph from a given dataset, it had mixed results for the different datasets. Even with the inaccurate graphs, it provided reasoning. Here the importance of prompt engineering for LLMs was highlighted as for the same dataset, when asked to perform web-scraping and gather the dataset vs being fed the dataset, the reasoning behind the generated causal graph was different. For the second task of generating causal reasoning from images, ChatGPT could mimic human-like thinking when providing reasoning for labeling specific animals. By specifying the color, patterns and fur of the animal, it was able to identify the most prominent features of the animals. Finally, for the causal tasks of answering counterfactual questions, ChatGPT once again provided varying results based on the prompts. When given the option of multiple-choice answers it was able to correctly predict the answer all 20 times. However, when asked the counterfactual question without being given multiple choice options it provided a variety of responses which were all logically and causally sound.

With all three types of causal tasks and data types, the major finding of this study is that while LLMs are capable of generating causal reasoning to a certain extent, they understand their own limitations. One such limitation is the need for prompt engineering/prompt tuning. This, while obvious for GenAI/LLMs due to the rise in research for the field, will be crucial for LLMs to achieve robustness and trust with end-users. Even though LLMs are trained on large observational data in the form of texts scraped from the web and not any causal frameworks or algorithms, they are somewhat capable of generating causal relations from observational data to a certain extent when they should not be able to. This phenomenon of exactly why LLMs can demonstrate causal capabilities merits further investigation.

LLMs capable of robust causal reasoning can supplement human domain knowledge/expertise for both commercial and research/development applications. With increased improvement in their causal capabilities, opportunities are abundant for implementations of end-to-end LLM solutions that are capable of importing data and providing causal analyses to the end users to help aid in decision-making. The ease and flexibility of LLM APIs such as ChatGPT provide an excellent example of versatile human-centered tools that do not require advanced technical expertise to perform simple causal tasks such as generating causal directions. While causal analysis has traditionally required a level of expertise and understanding of causality, computer science, and statistics, LLMs can once again supplement the need for a technical end-user by aiding in causal analysis tasks such as generating causal graphs from observational datasets. Traditional causal analysis tools such as DoWhy, CausalLearn, EconML, and CDT have been successful in investigating causal reasoning for observational datasets. However, as Kiciman et. al. highlighted these tools are limited when it comes to reasoning with necessity, sufficiency, normality, and responsibility. LLMs can alleviate these limitations as they can reason with these elements while aiming for actual causal reasoning.

4 Conclusion

In this paper, we investigated the capabilities of GenAI/LLMs to generate causal reasoning for three distinct causal tasks and found that LLMs can generate causal reasoning to a certain degree. These models providing provide human-like thinking in some instances while completely failing in others. This can be attributed to their black-box nature and complex training. We have shown that LLMs can generate complete causal graphs, competing with available state-of-the-art causal tools. However, we also highlight the limitations of LLMs for causal reasoning, mainly their dependency on the prompts that are entered. Prompt engineering/tuning plays a major role in the generated responses from the LLMs as they can either produce robust causal reasoning or fall short of causal reasoning and produce erroneous irrelevant responses. Future studies are planned to compare how different LLM models such as Claude or LLaMa perform on similar causal tasks.

As LLMs get more robust and advanced, their popularity and utilization will grow and embed into more applications, resulting in their use for causal analysis on a more regular basis. Especially to help alleviate the need for human-domain knowledge in the forms of SMEs for simple tasks such as generating causal graphs or answering counterfactual questions. Finally, because LLMs provide a human-centered interface for answering causal queries, they can serve to provide logical causal analysis to human queries.

References

[1] I. Ahmed and R. Islam, "Gemini-the most powerful llm: Myth or truth," *Authorea Preprints*, 2024.

[2] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, vol. 99, p. 101861, 2023.

[3] M. Masalkhi, J. Ong, E. Waisberg, N. Zaman, P. Sarker, A. G. Lee, and A. Tavakkoli, "A side-by-side evaluation of llama 2 by meta with chatgpt and its application in ophthalmology," *Eye*, pp. 1–4, 2024.

[4] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[5] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[6] B. J. Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, and Y. Su, "Thinking about gpt-3 in-context learning for biomedical ie? think again," *arXiv preprint arXiv:2203.08410*, 2022.

[7] M. Moradi, K. Blagec, F. Haberl, and M. Samwald, "Gpt-3 models are poor few-shot learners in the biomedical domain," *arXiv preprint arXiv:2109.02555*, 2021.

[8] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. C. Comeau *et al.*, "Opportunities and challenges for chatgpt and large language models in biomedicine and health," *Briefings in Bioinformatics*, vol. 25, no. 1, p. bbad493, 2024.

[9] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[10] M. Chevalley, Y. Roohani, A. Mehrjou, J. Leskovec, and P. Schwab, "Causalbench: A large-scale benchmark for network inference from single-cell perturbation data," *arXiv preprint arXiv:2210.17283*, 2022.

[11] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," *arXiv preprint arXiv:2305.00050*, 2023.

[12] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: methods and benchmarks," *Journal of Machine Learning Research*, vol. 17, no. 32, pp. 1–102, 2016.

[13] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[14] J. Frohberg and F. Binder, "Crass: A novel data set and benchmark to test counterfactual reasoning of large language models," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 2126–2140.

[15] M. Hobbhahn, T. Lieberum, and D. Seiler, "Investigating causal understanding in llms," in *NeurIPS ML Safety Workshop*, 2022.

[16] S. Long, T. Schuster, and A. Piché, "Can large language models build causal graphs?" *arXiv preprint arXiv:2303.05279*, 2023.

[17] D. Kalainathan and O. Goudet, "Causal discovery toolbox: Uncover causal relationships in python," *arXiv preprint arXiv:1903.02278*, 2019.

[18] J. Wang and K. Mueller, "The visual causality analyst: An interactive interface for causal reasoning," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 230–239, 2015.

[19] K. Zhang, S. Zhu, M. Kalander, I. Ng, J. Ye, Z. Chen, and L. Pan, "gcastle: A python toolbox for causal discovery," *arXiv preprint arXiv:2111.15155*, 2021.