Can Large Language Models Build Causal Graphs?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Building causal graphs can be a laborious process. To ensure all relevant variables have been captured, researchers often have to discuss with clinicians and experts while also reviewing extensive relevant medical literature. By encoding common and medical knowledge, large language models (LLMs) represent an opportunity to ease this process by automatically scoring edges (i.e., connections between two variables) in potential graphs. LLMs however have been shown to be brittle to the choice of probing words, context, and prompt that the user employs. In this work, we evaluate if LLMs can be a useful tool in speeding up causal graph development.

1 Introduction

Advances in causal inference have had important implications in empirical research as most research questions asked in the health and medical context are not associational, but *causal* in nature. Examples of such research questions include: *What is the efficacy of a given drug in a given population? What is the effect of a given intervention on a given outcome?* Common amongst these research questions is the desire to uncover the cause-and-effect relationships amongst a set of variables i.e., treatments, interventions, and outcomes. Such *causal* questions cannot be answered without knowledge of the data-generating processes, directly from the data itself, or from the distributions that govern said data [17].

Causal diagrams play a central role in causal inference because they encode contextual knowledge of the observable and unobservable variables, and their causal dependencies. Causal inference pioneer Judea Pearl refers to the nodes (i.e., variables) in a causal diagram as a "*society of listening variables*" [18]. The term "*listening*" stresses the defining property of directed and acyclic relationships between the variables, i.e., listening being asymmetrical, variable A listening to variable B, does not imply variable B listening to variable A, motivating the commonly adapted nomenclature of Directed Acyclic Graphs (DAGs) [5, 4].

Typically the first step for developing a causal model to address causal questions is to draw a causal diagram e.g., a causal DAG. However, with the growing complexity and depth of health and medical knowledge being generated and increasing availability of new research articles daily, research databases are reaching dimensions that limit the possibility of parsing through the enormity of evidence needed to craft comprehensive DAGs [19]. On the other hand, the ongoing developments of Large Language Models (LLM) have promise to help overcome some of these challenges by leveraging existing text data that may express causal sentiments (e.g., "X causes Y").

This research aims to address the question, "*Can large language models help researchers build causal diagrams?*" Here we will conduct experiments to determine under what conditions (e.g., prompt engineering, use of alternative language) GPT-3 [1] is able to provide accurate answers regarding the relationship between variables and what are its limitations in doing so.

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

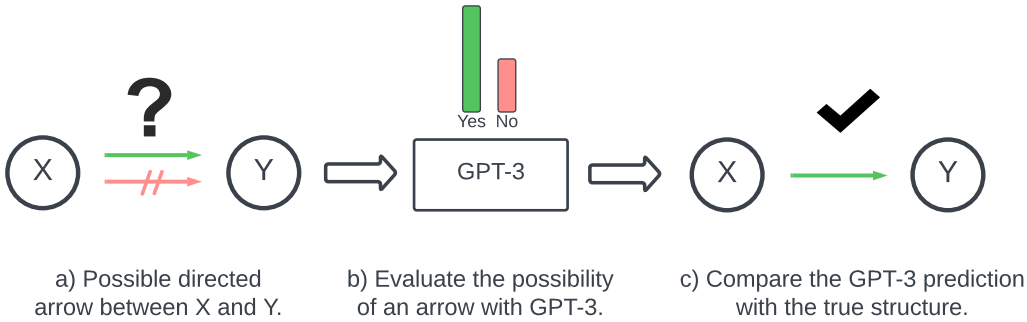

Figure 1: **Overview of the evaluation.** To predict the structure of a causal graph, for every ordered variable pair, we scored two statements using GPT-3, where the first statement implied the presence of an arrow and the second implied the absence of an arrow. GTP-3 was accurate if the correct statement had a higher accuracy score than the incorrect statement. For example, GPT-3 would be accurate if the statement implying the presence (or absence) of an arrow has a higher accuracy than the incorrect statement and the arrow is present (or absent) in the true DAG.

The main contributions of this paper are:

- Determining whether GPT-3 can signal the presence or absence of an edge between two variables in a DAG.
- Evaluating whether the use of certain language in prompts or linking verbs improves the classification accuracy of GPT-3.
- Exploring the limitations of GPT-3 in understanding the causal relationships between variables.

2 Background

2.1 Large language models

Large language models capture non-trivial relationships and knowledge about the datasets they have been trained upon. This knowledge has the possibility to unlock numerous applications in healthcare such as summarizing research papers, accessing patient risks from subjective symptoms, and diagnosing patients from clinical notes.

Although LLM perform well on general NLP tasks, its performance have been shown to be sensitive to its prompt [15, 7]. The advent of *prompt-based learning* introduced a possible solution to context sensitive text, by querying LLMs with a prompt that uses in-domain examples or task descriptions [14]. For example, chain-of-thought prompts such as *"Let's take this step by step"* has been shown to trigger multi-step reasoning in solving arithmetic problems [11]. Such prompts have also been shown to significantly improve performance in reasoning about medical questions [13].

Large language models are also sensitive to the type of text data they are trained on. For instance, GPT-3 [1] was trained on the corpus of text information on the internet. As one can imagine, the entirety of the internet would include a range of text data from lay and casual use of language on social media to more formal language in news articles. These differences in writing styles may influence the frequency of the use of causal language describing non-causal relationships. For instance, an individual writing a social media post use the word 'cause' more lightly than medical researchers in medical journals.

2.2 Causal diagram overview

Causal models are typically accompanied by graphical representations i.e., Directed Acyclic Graphs (DAGs) which are acyclic graphs that succinctly illustrate the qualitative assumptions made by the models, not captured by conventional statistical models or machine learning algorithms [3, 5].

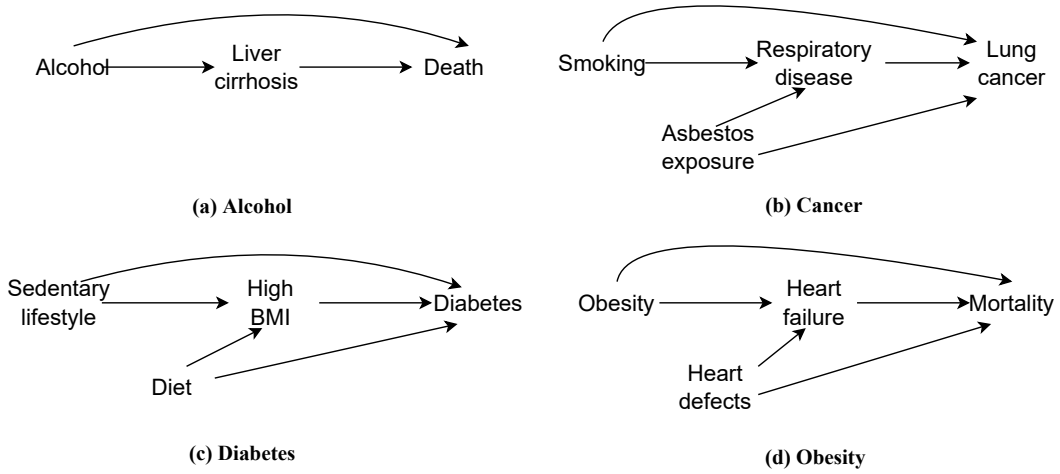

Figure 2: **Ground Truth DAGs.** Four DAGs illustrating well-known exposure-outcome effects in the medical literature. DAG (a) represents the simplest DAG evaluated by GPT-3. DAGs (B-D) represent more complex structures involving a collider variable (node with two arrows pointing into it e.g., 'respiratory disease', 'high BMI', and 'heart failure') with a common cause with the outcome.

In epidemiological research, DAGs have a variety of purposes including: (1) representing the causal relationships amongst variables [3, 4, 16]; (2) identifying the potential confounding variables which need to be controlled for in order to estimate causal effects [4, 16, 20, 10]; and more recently (3) as a means of classifying the types of causal relationships that may give rise to selection bias [9].

A DAG is composed of variables (nodes), both measured and unmeasured, and their connections are displayed via line segments (directed edges) [5, 9]. The *absence* of an arrow between variables indicates the lack of a direct relationship between the variables. If the edge has an arrowhead, variable at the tail is the parent node and the variable at the arrowhead is the child node [4]. An edge or arc is any line (with an arrowhead or not) that connects two variables [3]. The main characteristics of DAGs are that they are: (1) *directed* i.e., the edge has a defined direction (arrowhead), and (2) *acyclic* i.e., lack of cycles or loops within the graph.

A DAG is causal if: (1) the arrows between variables can be interpreted as direct causal effects, and (2) all common causes of any pair of variables are present [9]. The causal effects are 'direct' relative to certain degrees of abstraction in that the DAG does not include any variables that may mediate the effect [4]. As the name suggests, DAGs are acyclic because a variable cannot be the cause of itself, either directly or indirectly through another variable i.e., there are no feedback loops; as illustrated by each DAG in Figure 2 [9]. Additionally, in DAGs, causal pathways are represented with directed paths from the starting variable to the final variable; thus, a variable is the cause of its descendants and an effect of its ancestors [4].

3 Experiments

3.1 Experimental details

To empirically validate the effectiveness of LLMs in building DAGs, we used four DAGs representing well-known exposure-outcome relationships in the medical literature (Figure 2) as the Ground Truth. These DAGs are varied in complexity, amount of variables, and reflect different medical contexts. For a DAG of N variables, there are $\binom{N}{2}$ possible combinations of variables, and there are twice this amount of possible arrows since the arrows are directed. For example, a DAG of 4 variables has $2 \times \binom{4}{2} = 12$ possible arrows.

For each DAG, we looped through every ordered variable pair, and asked GPT-3 to score two statements per pair: (1) one implying the presence of a directed edge from variable 1 to variable 2, and (2) one implying the absence of a directed edge from variable 1 to variable 2. The presence

or absence of an edge between two variables is a binary decision (Yes / No), thus, we defined the prediction as accurate, if GPT-3 scored the correct statement higher than the incorrect one. We reported the *accuracy* or the proportion of correct predictions of our model.

3.2 Results

Q1: Does using prompt engineering lead to more accurate answers? We investigated if the prediction accuracy of GPT-3 could be improved by prompting the statements with a reference to a medical authority. For example,

"*According to X, var1 increases the risk of developing var2*",

instead of

"*var1 increases the risk of developing var2*" (baseline),

where X is an individual or entity with medical authority or expertise, e.g., medical doctors, medical studies, or "Big Pharma". These prompts were chosen as they vary in their credibility with the public. We found that in 2 cases (Diabetes and Obesity DAGs) prompt engineering did not help and baseline outperformed all other prompts. While in 2 other cases, the "According to medical doctors," prompting significantly improved the accuracy of GPT-3. See Table 1 for all results.

DAG name	Prompt	Accuracy
Alcohol	Baseline	0.33
	Big Pharma	0.50
	Medical doctors	**0.83**
	Medical studies	0.67
Cancer	Baseline	0.75
	Big Pharma	0.58
	Medical doctors	**1.00**
	Medical studies	**1.00**
Diabetes	**Baseline**	**0.67**
	Big Pharma	0.50
	Medical doctors	0.33
	Medical studies	0.42
Obesity	**Baseline**	**0.75**
	Big Pharma	0.58
	Medical doctors	0.75
	Medical studies	0.75

Table 1: **Prompt engineering:** The medical authority used to prompt the statement.

Q2: Does the verb used to denote the relationship between the variables have an impact on accuracy? For instance, "Variable 1 *X* Variable 2" where *X* represents the verb (or phrase) that denotes the relationship between the variables, e.g., "causes" or "increases the risk".

Our results demonstrated that while no verb consistently improved classification accuracy, the choice of verb linking the two variables of interest influenced accuracy. 'Increase risk' had the highest accuracy for three of the four DAGs. Though it did not achieve the highest accuracy in the Alcohol DAG. Overall, the use of 'cause' yielded decent results for all DAGs. Results are reported in Table 2.

Q3: Does specificity in language improve accuracy? We investigated if making our statements more specific or descriptive improved GPT-3's accuracy.

Unsurprisingly, rephrasing the "alcohol" variable to "excessive alcohol consumption" increased the accuracy of GPT-3 on the Alcohol DAG. However, being more specific about the number of cigarettes being smoked and using a clinical term to qualify obesity resulted in worse accuracy for the Cancer and Obesity DAGs. Overall, In this analysis, more specific statements did not increase the accuracy and often resulted in worse accuracy for different linking verbs. Results are reported in Table 3.

DAG name	Linking Verb	Accuracy
Alcohol	Cause	0.33
	Increase likelihood	**0.50**
	Increase risk	0.33
Cancer	Cause	0.58
	Increase likelihood	0.58
	Increase risk	**0.75**
Diabetes	Cause	0.58
	Increase likelihood	0.42
	Increase risk	**0.67**
Obesity	Cause	0.58
	Increase likelihood	0.42
	Increase risk	**0.75**

Table 2: **Linking verb:** The verb or phrase used to link the two variables of interest.

DAG name	Variable Name	Linking Verb	Accuracy
Alcohol	Alcohol	Cause	0.33
		Increase risk	0.50
	Excessive alcohol consumption	Cause	0.33
		Increase risk	**0.67**
Cancer	**Cigarette smoking**	Cause	0.58
		Increase risk	**0.67**
	Smoking 100 cigarettes a day	Cause	0.50
		Increase risk	0.58
Obesity	**Obesity**	Cause	0.58
		Increase risk	**0.67**
	Excessive fat accumulation	Cause	0.58
		Increase Risk	0.58

Table 3: **Specificity:** More extensive descriptions of variables/concepts.

4 Discussion

In this work, we explored if LLMs could be used to speed up the workflow of researchers by automatically scoring edges in potential DAGs. For the relatively simple and well-studied DAGs that we tested GPT-3 on, the results were overall encouraging as the performance reached much higher than 50% accuracy (random chance) on all DAGs for at least one of the tested settings (e.g., prompt or linking verb). In this analysis, we found that GPT-3's accuracy performance was influenced by different prompts and linking verbs between variables of interest.

To the best of our knowledge, this is the first study to examine using LLM for causal diagrams in the medical context. Though there is growing interest, to date, there are few studies exploring the utility of LLM in causal diagram development. A recent study by Willig et al., (2022) [22] compared the performance of three query LLMs in making causal graph predictions in a general context. There also has been some interesting works applying causal inference in the LLM context. For instance, Vig et al., (2020) [21] investigated gender bias present in LLM using causal mediation analysis. Feder et al., (2022) [2] released a preprint of a consolidated exploration of causal inference situated in NLP. These works suggest more focus is being devoted to researching how causal inference can be applied to LLMs and NLP.

Furthermore, there has been some research investigating LLM's ability to answer and reason with medical text data. Several recent studies [13, 6] showed promising results on LLMs ability to answer

medical exam questions. Others [15, 7] have shown that context-specific LLMs such as BioBert are able to outperform GPT-3 in medical domain NLP tasks.

Limitations This study has some limitations. First, it must be acknowledged that the updating of LLMs, themselves as well as the data they are trained upon, lags behind the availability of new medical literature, and, thus may not be useful for informing the building of DAGs for novel diseases. Furthermore, GPT-3 was trained upon text data uploaded to the internet. The language used on the broader internet is likely more casual with the use of causal language than the medical academic literature [8].

Future work Future work aims to use a medical language context-specific LLM such as BioBERT[12] to signal the presence or absence of edges in DAGs using medical terminology. Additionally, since our preliminary evaluations only examined the presence/absence of arrows and their direction, upcoming projects will be focused on controlling for acyclicity amongst variables, another important characteristic of DAGs.

5 Conclusion

Large Language Models represent an exciting opportunity to extract common knowledge from the medical literature to speed up DAG creation, but further research must be done to address the limitations reported above.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

[3] S. Greenland and B. Brumback. An overview of relations among causal modelling methods. *International journal of epidemiology*, 31(5):1030–1037, 2002.

[4] S. Greenland and J. Pearl. Causal diagrams. 2007.

[5] S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

[6] Q. Guo, S. Cao, and Z. Yi. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, 37(11):8548–8564, 2022.

[7] B. J. Gutiérrez, 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.

[8] N. Haber, S. Wieten, J. Rohrer, O. Arah, P. Tennant, E. Stuart, E. Murray, S. Pilleron, S. Lam, E. Riederer, et al. Causal and associational language in observational health research: a systematic evaluation. *American Journal of Epidemiology*, 2022.

[9] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

[10] M. A. Hernán, S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology*, 155(2):176–184, 2002.

[11] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

[12] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[13] V. Liévin, C. E. Hother, and O. Winther. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*, 2022.

[14] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[15] 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.

[16] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[17] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 2009.

[18] J. Pearl. The eight pillars of causal wisdom. *UCLA*, 2017.

[19] V. Raghupathi, W.; Raghupathi. Big data analytics in healthcare- promise and potential. *Health Information Science and Systems*, 2, 2014.

[20] J. M. Robins. Data, design, and background knowledge in etiologic inference. *Epidemiology*, pages 313–320, 2001.

[21] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401, 2020.

[22] M. Willig, M. Zečević, D. S. Dhami, and K. Kersting. Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*, 2022.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Figure 2: Ground Truth DAGs.

(b) Did you include complete proofs of all theoretical results? [N/A] This work was an exploration of LLM usage for causal diagram building and did not require any proofs for results.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The experiment section outlines how we performed our study in a way that is replicable by others.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]