Can large language models reason about causal relationships in multimodal time series data?

Abstract

Large Language Models (LLMs) have demonstrated promise in transforming the ways that individuals synthesis and interact with large amounts of information. However, current LLMs are limited in their ability to provide explanations about causal relationships in data. In this paper, we investigate the ability of LLMs to answer queries related to causal relationships within time series data. We generate synthetic datasets based on three distinct directed acyclic graphs (DAGs) representing causal relationships among time series variables. Initially, we use abstract variable names in the analysis and later assign real-world meanings to these variables to align with the DAG structures. Using in-context learning, we present the relationships of these variables to the LLM in the prompt and evaluate how effectively the LLMs identify the variables that caused specific observations in an outcome variable.

1 Introduction

Large language models (LLMs) have demonstrated immense promise in summarization and distillation. There has been recent interest in exploring how LLMs can be used to analyze and provide interpretations of time series data [5, 8]. For example, a person may want to use an LLM as a tool to answer questions about their wearable data. While there has been work building LLMs agents to interface with time series data [8, 11], an open question is how reliable LLM interpretations of causal patterns in the data are.

For example, consider a individual with diabetes who uses multiple wearable devices to manage their disease. They may want to ask questions about their data, such as inquiring about why their blood glucose was out of range at a given time point. An ideal LLM chatbot for diabetes would be equipped to answer questions about causal relationships in the data and provide a response that integrates the patient's data from their activity monitor and insulin pump. However, there is currently limited work exploring the limits and best practices for eliciting LLM explanations of causal relationships in time series data. In this paper, we explore the ability of LLMs to respond to queries about causal relationships in time series data.

Our work has four primary objectives:

1. Investigate how LLMs can answer questions regarding known causal relationships in time series data.
2. Determine whether an LLM can explain what caused an observed change in an output variable of interest using in-context learning.
3. Evaluate the causal reasoning abilities of state-of-the-art LLMs in scenarios where variables have abstract names versus real names.
4. Evaluate the performance of different ways of presenting data to an LLM.

Preprint. Under review.

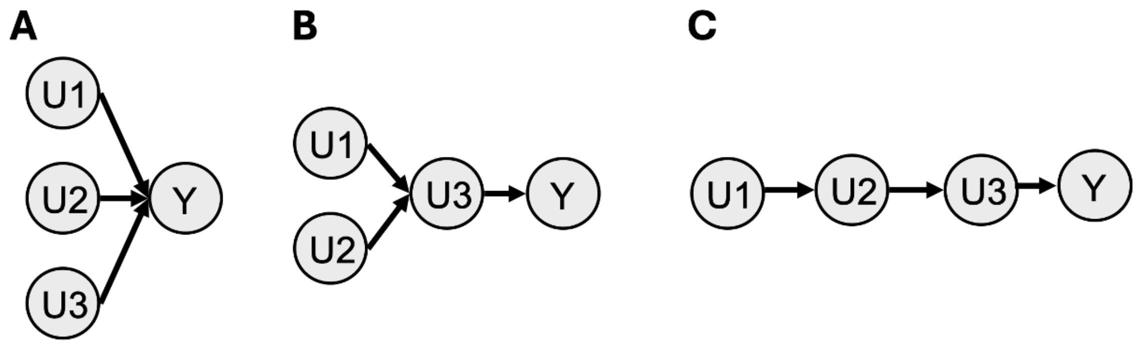

Figure 1: DAGs of causal relationships used to build synthetic datasets.

The goal of this work is to assess the causal reasoning abilities of common state-of-the-art LLMs to explain causal relationships between multi-modal time series data. We also make our code and synthetic datasets public, to serve as a benchmark for the community.

Recent work on causality and LLMS: There has been recent interest in investigating how causal relationships are embedded in or learned through language models [4, 7, 14]. Of recent, [13] suggested that while LLMs may perform well on causal tasks, they are not inherently causal. Another line of work has investigated how LLMs reason [1, 9]. Other work has investigated how LLMs can reason about causal interventions [6]. Here, the authors evaluated the performance when variables were named with different underlying meanings, attempting to separate causal reasoning from causal memorization. Other have been proposing solutions to incorporate causal knowledge into LLMs [2].

2 Methods

2.1 Experimental Setup Overview

In this paper, we evaluated 4 widely-used LLMs on a variety of tasks. We first created synthetic time series datasets based on known causal relationships from three directed acyclic graphs (DAGs). We imagined a scenario where there is an outcome variable, Y, that is of particular interest to an individual. The task was designed to have the LLM analyze a dataset and provide an explanation of why Y was observed, given known relationships between Y and other variables. We presented different datasets from different scenarios to an LLM where it was prompted to identify the direct cause and indirect of changes in an outcome variable. We evaluated how well LLMs can identify these variables.

Initially, variables were be presented with abstract names. In a followup experiment, the variables were reassigned to real-world meanings to align with DAG structures. We also tested different ways to present that data to the LLM and how the performance changed when the relationships were not given to the LLM in the prompt.

2.2 Data Generation

We generated synthetic datasets with 4 different time series variables across 30 time unites. In our setup, we denoted the variable of interest as Y. We called the three other variables $U1$, $U2$, $U3$. We derived the datasets based on 3 different causal configurations. These are shown in Figure 1. We explore 3 different frameworks. Below, we describe the data generation process based on the three DAGs. In each synthetic dataset, we included the 4 variables and generate sequences of length 30.

A

$$Y[t] = U1[t] * 2 + U2[t] * 2 + U3[t] * 2 \quad (1)$$

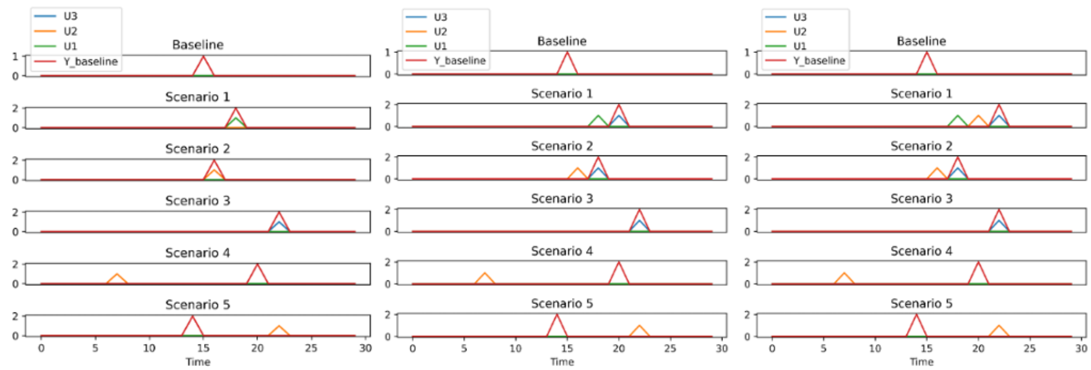

Figure 2: Examples Time Series Data: Baseline represents scenario where Y has step response without any cause from any input variables. Scenario 1 occurs when $U1$ has a step response. Scenario 2 occurs when $U2$ has a step response and Scenario 2 occurs when $U3$ has a step response.

B

$$U3[t] = U1[t-2] * 2 + U2[t-2] * 2 \tag{2}$$
$$Y[t] = U3[t-2] * 2 \tag{3}$$

C

$$U2[t] = U1[t-2] * 2 \tag{4}$$
$$U3[t] = U2[t-2] * 2 \tag{5}$$
$$Y[t] = U3[t] * 2 \tag{6}$$

Scenario Descriptions

For each DAG, we generated datasets based on 6 different scenarios. All variables were either zero for the duration of the time period unless or had a randomly occurring impulse at time r. We tested a baseline case, where none of the input variables have an observable effect on Y, but Y has an effect. We then tested 3 scenarios, Scenario 1, Scenario 2, and Scenario 3, in which an impulse response occurred, as given in Eq. 7, for $U1$, $U2$, and $U3$, respectively. Scenario 4 and Scenario 5 were designed as an edge case to try to trick the LLM. In these scenarios, $U2$ had an impulse that did not affect Y by the definitions given. In Scenario 4, the impulse occurred more than 2 time units before Y, and in Scenario 5, the impulse occurred after Y.

$$\delta[t-r] - \delta[t-(r+1)], r \sim Uniform(0, 25) \tag{7}$$

Tasks and Prompts

The goal of our work was simply to understand if LLMs can decribe causal explanations in time series data. Accordingly, the primary outcome we measure is how well the model answers the following two questions:

1. **Direct Cause** Which of U1, U2, U3 directly caused Y to be non-zero?
2. **Indirect Cause** Which of U1, U2, U3 indirectly caused Y to be non-zero?

We separated the direct and indirect cause to better understand whether the LLM was incorporating information from the DAG. The direct causes of Y were only the variables that shared an edge with Y in the DAG. The baseline prompt had four main componts: instructions, descriptions of the relationships, the task, and the data. Figure 3 shows the components of the prompts.

Instruction	Relationships	Task	Data
I am going to give you data on a variable of interest, Y. I am trying to understand why Y was non-zero, given other time series data. Please analyze the time series data to determine why Y behaved in a certain way based on the time series from Y, and other time series signals, U1, U2, and U3. Time and magnitude units have been normalized. """	Below are the descriptions of relationships between the data: U1 causes U2 to increase value by 1 after 5 units. U2 causes U3 to increase value by 1 after 5 units. U3 causes Y to increase value by 2 units. The effects are additive. Other factors may affect Y.	Answer the following questions. Using the data, for this time-series data, which of the input signals caused Y to be non-zero. 1. Which of U1, U2, U3 directly caused Y to be non-zero? 2. Which of U1, U2, U3 indirectly caused Y to be non-zero?	U1: 0, 0 ... U2: 0, 0 ... U3: 0, 0 ... Y: 0, 0 ...

Figure 3: Baseline Prompt Components

2.3 Experiment Workflow

We evaluate 4 different open-source large language models. We evaluate two widelyused open-source models, gpt-4 [10] and llama3.1-405b [12], and two open-source efficient models gpt-4o-mini [10] and gemma2-27b [3]. Each model was tested on the same scenarios.

2.4 Additional Experiments

Assessment using realistic variables: In a subsequent analysis, we changed the variable names to represent things with real meanings. We did this to understand if known associations embedded in each LLM affect performance, positively or negatively. This approach was similar to the one done in [6], where authors evaluated performance differences with variables named differently. Below we describe how the variables were changed in the prompt.

1. DAG A: We rename Y to be risk of cardiovascular event, U1 to be exercise, US to be caffeine, U3 to be stress.
2. DAG B: We rename Y to be risk of cardiovascular event, U1 to be heart rate, U2 to be caffeine, U3 to be exercise.
3. DAG C: We rename Y to be risk of cardiovascular event, U1 to be heart rate, U2 to be caffeine, U3 to be coffee consumption.

Assessment of alternative data formats: We also investigated how the data format of the time series data affected the results. The baseline case as shown in Figure 3 included the raw time series data in the prompt. We tested an alternative format of the data where the data was transformed into a linguistic summary.

> *Linguistic Data Representation Example:*
> U1 observed all zero values, U2 observed Value: 1.0 at Time: 7, U3 observed all zero values, and Y observed Value: 2.0 at Time: 20.

Assessment of inclusion of descriptions: Lastly, we investigated how the removal of the Relationships component of the prompt Figure 3 affected the results.

3 Results

Figure 4 shows the model performance of base case where the variables had synthetic names and the prompt was structured as in Figure 3. Figure 5 shows the difference in performance when real-world variable names are used. In Appendix A.1 and A.3, the results from changing the data format and DAG descriptions are shown.

Variable naming effected the performance in a differential manner. For some scenarios, changing the variable names to words with real meaning caused worse performance. In other scenarios, it caused improved performance. We show some example responses and explanations in Figure 6 from the experiments ran with and without the DAG explanations. In this example, without the DAG explanation of the relationships, the LLM still answers correctly when the variables have known meanings.

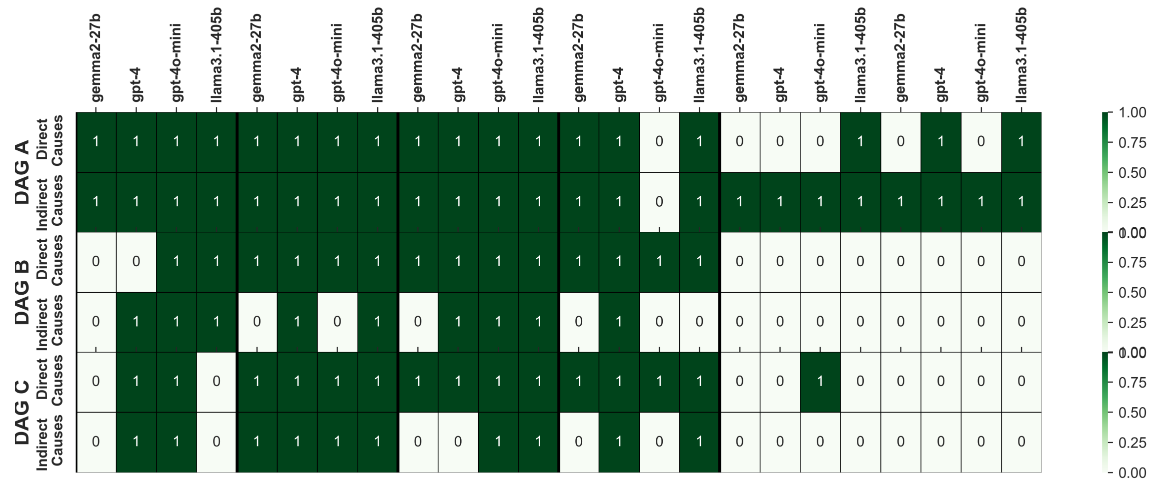

Figure 4: Synthetic Variable Names: Heatmap showing tasks answered correctly.

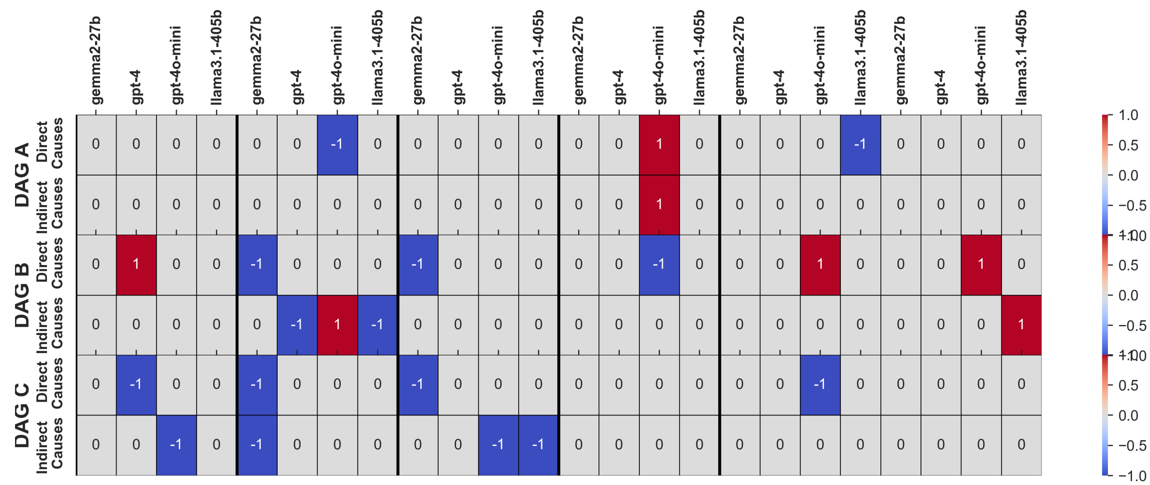

Figure 5: Real-World Variable Names: Heatmap showing the difference in the correct answers of simulations from the synthetic variables to the real-world meaning variables

Overall, the performance was poor for the tasks for scenario 4 and 5. In these scenarios, the LLM explanations often noted that since U2 was observed to have a non-zero value, it had a cuasal effect on Y, even though the timing difference violated the rule. Below is an example of an incorrect reasoning:

> *From the data provided, we can see that the Risk of cardiac event became non-zero at the 21st time unit. Looking at the other time series, we can see that there was no exercise during this period, so it did not directly contribute to the Risk of cardiac event. However, there was a consumption of caffeine at the 8th time unit. According to the relationships provided, caffeine causes the heart rate to increase by 1 after 2 time units. This means that the heart rate would have increased at the 10th time unit. The heart rate, in turn, causes the Risk of cardiac event to increase by 2 units at the same time. However, there is a delay of 11 time units between the increase in heart rate and the increase in Risk of cardiac event.*

	Without DAG Description	With DAG Description
Arbitrary Variable Names	'1. U2, U3', '2. None', ' ` Explanation: Looking at the time series data, we can see that the variable Y becomes non-zero at the same time as U2 and U3. Specifically, U2 becomes non-zero at the 17th time point and U3 becomes non-zero at the 19th time point, while Y becomes non-zero at the 19th time point. This suggests that both U2 and U3 have a direct impact on Y…	'1. U3', '2. U2', '3. Explanation: From the given data, we can see that U3 directly caused Y to be non-zero. This is because U3 increases by 1 unit at time 19, and Y also increases by 2 units at the same time, which is consistent with the given relationship that U3 causes Y to increase value by 2 units at the same time. U2 indirectly caused Y to be non-zero. This is because U2 increases by 1 unit at time 17, and after 2 time units (at time 19), U3 increases by 1 unit. This increase in U3 then causes Y to increase by 2 units, as per the given relationships.
Real variable names	1. Heart rate', '2. Caffeine', '3. The Risk of cardiac event became non-zero at the same time when the Heart rate became non-zero, indicating a direct cause-effect relationship. The Caffeine intake was observed to increase one time step before the increase in Heart rate, suggesting an indirect effect on the Risk of cardiac event.	'1. Heart rate', '2. Caffeine', '3. Explanation: According to the data, the only time the Risk of cardiac event is non-zero is at time unit 19, where it is 2. Looking at the Heart rate data, we can see that it increases by 1 at time unit 18, which directly causes the Risk of cardiac event to increase by 2 at the same time unit. Therefore, Heart rate directly caused the Risk of cardiac event to be non-zero. On the other hand, the Caffeine data shows an increase at time unit 16, which according to the given relationships, causes the Heart rate to increase by 1 after 2 time units (at time unit 18).

Figure 6: Example Responses and Explanations: Above shows the output of gpt-4 for scenario 2 from DAG B. The green shaded answers are correct.

4 Conclusion

In this work we assessed the ability of LLMs to answer questions about causal relationships in time series data. We systematically evaluate 4 widely-used LLMs on synthetic data to understand the shortcomings of causal reasoning of time series analysis.

Our main findings can be summarized below:

- Changing the variable names from arbitrary variable names to real variable names had a differential effect on performance. In some cases, it improved performance. In other cases, it led to the the LLM answering questions about the causal variables instead of getting answer from the data.
- In the current framework, the LLMs often failed to properly account for the time stamps on the data when determining the causal relationships. In Scenario 4 and Scenario 5, the LLM often mistakenly noted a variable as being causal even when it came after the outcome variable.
- Changing the way the data was presented to the LLM affected performance. Future work should investigate how to best optimize how time series data is formatted for analysis by LLMs.

This work serves as a preliminary investigation in the ability of LLMs to describe causal relationships in time series data. Our code and synthetic data is public for the community to use as a benchmark.

References

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

[2] Abdolmahdi Bagheri, Matin Alinejad, Kevin Bello, and Alireza Akhondi-Asl. C2P: Featuring large language models with causal reasoning. *arXiv [cs.LO]*, July 2024.

[3] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex

Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-Hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology. *arXiv [cs.CL]*, March 2024.

[4] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: A benchmark to assess causal reasoning capabilities of language models. *arXiv [cs.CL]*, December 2023.

[5] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Others. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Adv. Neural Inf. Process. Syst.*, 36, 2024.

[6] Tejas Kasetty, Divyat Mahajan, Gintare Karolina Dziugaite, Alexandre Drouin, and Dhanya Sridhar. Evaluating interventional reasoning capabilities of large language models. *arXiv [cs.LG]*, April 2024.

[7] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv [cs.AI]*, April 2023.

[8] Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. Transforming wearable data into health insights using large language model agents. *arXiv [cs.AI]*, June 2024.

[9] Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. *arXiv [cs.AI]*, November 2023.

[10] OpenAI1. OpenAI1.

[11] Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, Wang Xiaofeng, Anh Nguyen Duc, and Pekka Abrahamsson. Can large language models serve as data analysts? a multi-agent assisted approach for qualitative data analysis. *arXiv [cs.SE]*, February 2024.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv [cs.CL]*, February 2023.

[13] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *arXiv [cs.AI]*, August 2023.

[14] Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. Understanding causality with large language models: Feasibility and opportunities. *arXiv [cs.LG]*, April 2023.

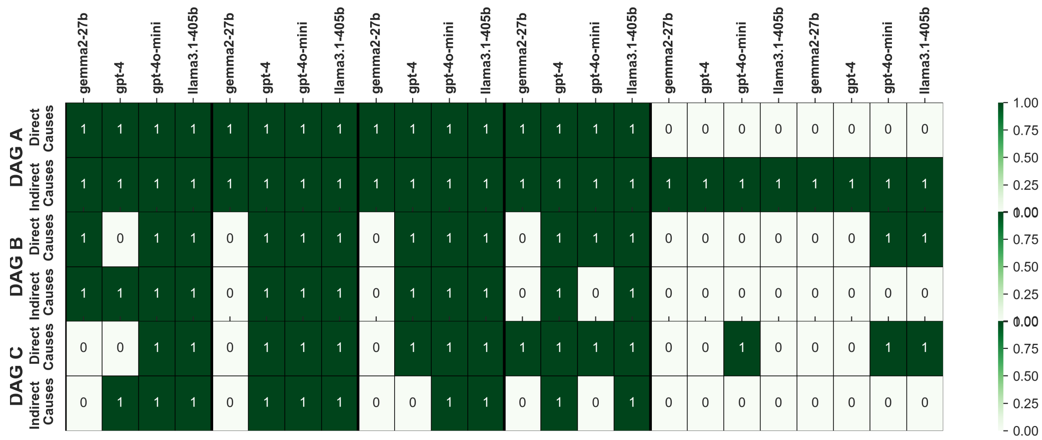

Figure 7: Synthetic Variable Names: Heatmap showing tasks answered correctly with data formatted as linguistic summaries

A Appendix

A.1 Change in data formatting

Figures 7 and 8 show the results when the data format was changed.

A.2 Real

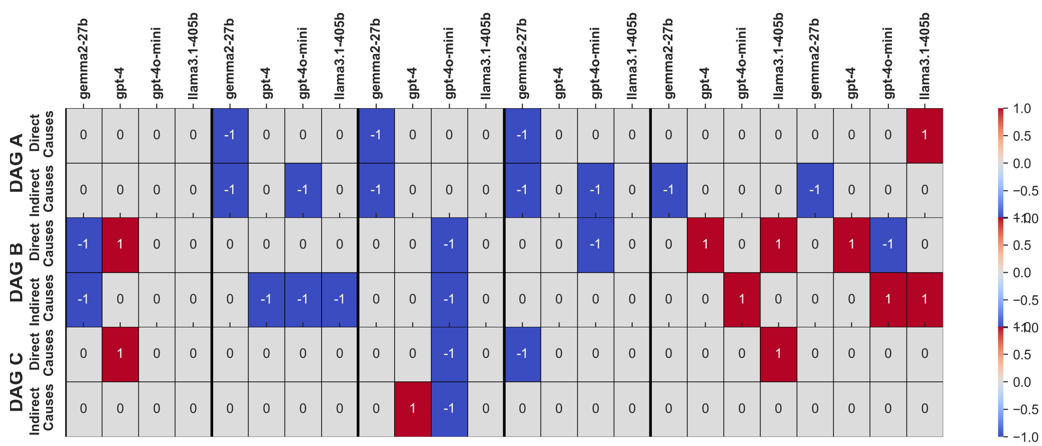

Figure 8: Real-World Variable Names: Heatmap showing the difference in the correct answers of simulations from the synthetic variables to the real-world meaning variables with data formatted as linguistic summaries

A.3 Ablation of Description of DAG

Figures 9 and 9 show the results when the DAG description was omitted from the prompt.

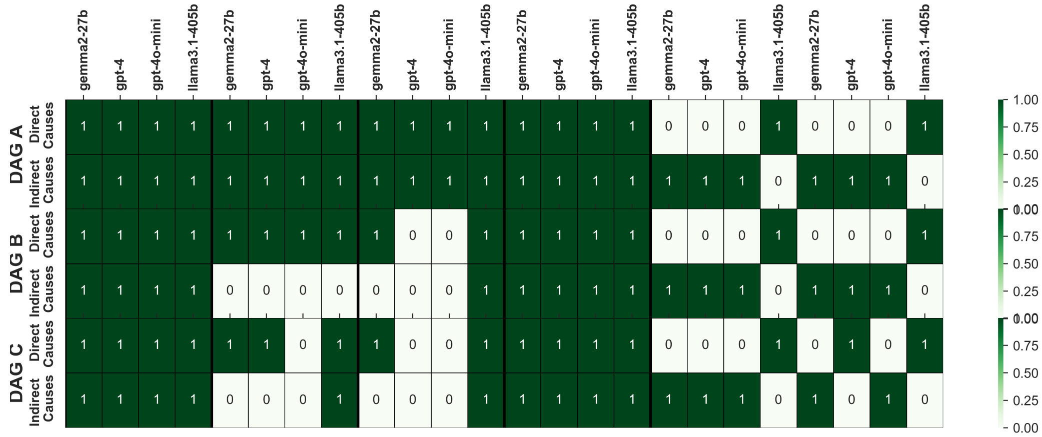

Figure 9: Synthetic Variable Names: Heatmap showing tasks answered correctly with DAG explanations omitted

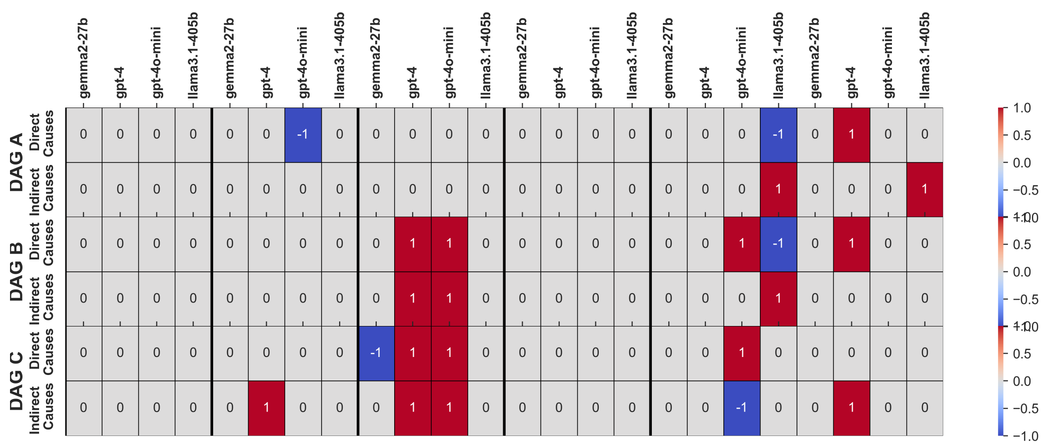

Figure 10: Real-World Variable Names: Heatmap showing the difference in the correct answers of simulations from the synthetic variables to the real-world meaning variables with DAG explanations omitted