Investigating the Ability of Large Language Models to Explain Causal Relationships in Time Series Data

Elizabeth Healey[1] **and Isaac Kohane**[2]
[1]Massachusetts Institute of Technology, [2]Harvard Medical School
Cambridge, MA 02138
ehealey@mit.edu

Abstract

Large language models (LLMs) can enhance how individuals interact with and process information from large amounts of data. In many settings, the ability to explain the causal reasons behind observations in data is important. In this work, we investigate the ability of LLMs to provide accurate explanations about causal relationships in time series data. We generated synthetic datasets based on three distinct directed acyclic graphs (DAGs) representing causal relationships between multiple time series variables, and we evaluated how state-of-the-art LLMs answer questions related to causal effects within the observed data. Initially, we used abstract variable names in the analysis and later assigned real-world meanings to these variables to align with the DAG structures. We tested how accurately the LLMs identified the variables that caused specific observations in an outcome variable and found shortcomings with state-of-the-art models. We highlight challenges and opportunities for research in this space.

1 Introduction

Large language models (LLMs) have garnered widespread attention for their ability to expedite information extraction and answer questions about data. There has been recent interest in exploring how LLMs can be used to analyze and provide interpretations of time series data [9, 14]. For example, an individual may want to use an LLM as a tool to answer questions about wearable data related to health status. Although there has been work on building LLM agents to interface and analyze data [9, 11], there has been less focus on assessing the reliability of interpretations of causal patterns in the data.

Recent work has focused on investigating how causal relationships are embedded in LLMs [6, 8, 13]. In a recent study, [12] suggested that though LLMs may perform well on causal tasks, it does not necessarily mean they are doing causal reasoning. A line of work has investigated how LLMs reason [1, 10], with a focus on assessing if causal reasoning is performed by large models. In [7], authors investigated how LLMs can reason specifically about causal interventions and tried to differentiate causal reasoning from causal memorization through their experiments. Others have proposed solutions to incorporate causal knowledge into LLMs [2].

Reliable explanations of causal relationships in data from LLMs is an important area of research. For example, consider an individual with diabetes who uses multiple wearable devices to manage their disease. They may want to ask questions about their data, such as asking why their blood glucose was out of range at a given time. An ideal LLM for diabetes management 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 known causal relationships in time series data.

Preprint. Under review.

In this work, we evaluate the ability of LLMs to answer questions about multivariate time series data regarding known causal relationships. Specifically, we assess whether an LLM can determine what caused an outcome variable based on available data and known relationships. This is done by asking the LLM to give the direct and indirect causes of an effect on an outcome variable. We evaluate the causal reasoning abilities of state-of-the-art LLMs in scenarios where variables have abstract names and compare that to the performance when the variables have real names with known causal relationships. Lastly, we investigate how the performance varies when data are presented in an alternative format to the LLM.

2 Methods

2.1 Experimental Setup Overview

We evaluated four 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 focused on a setting where there was an outcome variable, Y, 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 the 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, the variables were presented with abstract names. In a follow-up experiment, the variables were reassigned to real-world meanings to align with the DAG structures. We also tested different ways to present the 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 four different time series variables across 30 time units. In our setup, we denoted the variable of interest as Y. We named 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. The data generation process was based on the three DAGs and is detailed in Appendix A.1. In each synthetic dataset, we included the 4 variables and generated sequences of length 30.

For each DAG, we generated datasets based on six different scenarios. All variables were either zero for the duration of the time period or they had a randomly occurring spike at time r equal to 1. We tested a baseline case where $U1$, $U2$, and $U3$ were all zero and Y had an spike. We then tested 3 scenarios, Scenario 1, Scenario 2, and Scenario 3, in which an spike response occurred in $U1$, $U2$, and $U3$, respectively. Scenario 4 and Scenario 5 were designed as adversarial scenarios to try to trick the LLM. In these scenarios, Y had an impulse that was not explained by the other variables according to the definitions, and $U2$ had an spike. In Scenario 4, the spike in $U2$ occurred more than 2 time units before Y, and in Scenario 5, the spike in $U2$ occurred after Y. Each scenario type was generated five times with the spike occurring at a random time between 0 and 25.

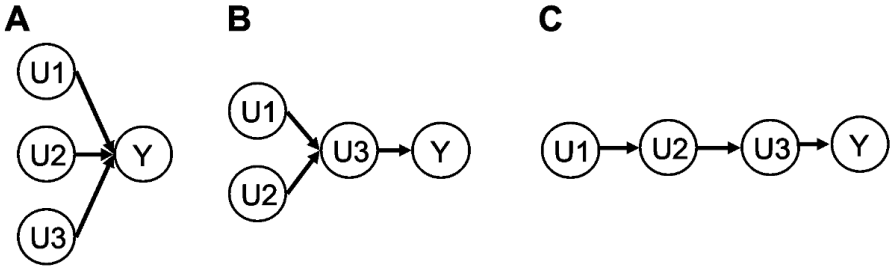

Figure 1: DAGs of causal relationships used to construct synthetic time series data.

2.3 Tasks and Prompts

The goal of this work was to understand if LLMs can provide causal explanations in time-series data. As our main evaluation metric, we measured how well the model answered 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 direct and indirect causes 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 components: instructions, descriptions of relationships, tasks, and data. Appendix A.2 details the components of the prompts.

2.4 Experiment Workflow

We evaluated 4 different open source LLMs. We used two widely used open source models, gpt-4 [4] and llama3.1-405b [5], and two smaller open source models, gpt-4o-mini [4] and gemma2-27b [3]. Each model was tested using the same scenarios.

2.5 Additional Experiments

Assessment using realistic variables: In an additional experiment, we changed the variable names to words with real-world meanings. We did this to understand whether the known associations embedded in each LLM affected performance positively or negatively. This approach was similar to what was done in [7], where the 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 as heart rate, U1 as exercise, US as caffeine, U3 as stress.
2. DAG B: We rename Y as the risk of cardiovascular events, U1 as heart rate, U2 as caffeine, and U3 as exercise.
3. DAG C: We rename Y as the risk of cardiovascular events, U1 as heart rate, U2 as caffeine, and U3 as coffee consumption.

Assessment of alternative data formats: We also investigated how the data format of the time series data affected the results. We tested an alternative format of the data in which the data was transformed into a textual summary stating the times at which each variable was non-zero with the non-zero value. This is detailed in Appendix A.2.

3 Results

Figure 2 shows the performance of the base case model in which the variables had synthetic names. Figure 3 shows the difference in performance when real-world variable names are used. In Appendix A.3, the results from changing the data format are shown. The variable naming affected performance in a differential way. For some scenarios, namely scenario 2, changing variable names to words with real meaning caused performance to decrease for most models. However, for the adversarial scenarios, scenario 4 and 5, the performance often increased when including a real variable name.

Overall, performance was worse for the tasks for all models for Scenarios 4 and 5. In these scenarios, the LLM explanations often noted that because $U2$ was observed to have a non-zero value, it had a causal effect on Y, even though the timing difference violated the known causal relationship. Notably, performance was better across scenarios and tasks for the larger models. Llama3.1-405b and gpt-4 had consistently good performance across scenarios. Performance was generally better for tasks asking for direct causes of the observed variable Y, compared to indirect causes. On examination of the explanations, it was observed that sometimes the model hallucinated values for the indirect variables that were zero. On other occasions, explanations indicated that the model noted variables did not have non-zero entries, yet incorrectly listed them as having an indirect effect.

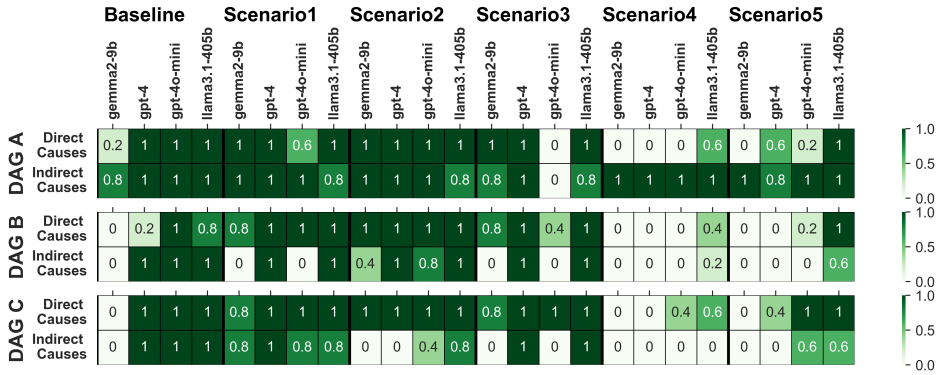

Figure 2: Heatmap showing the questions answered correctly using variables with abstract names and raw data as input. Numbers show percentage of correct answers across 5 random scenarios.

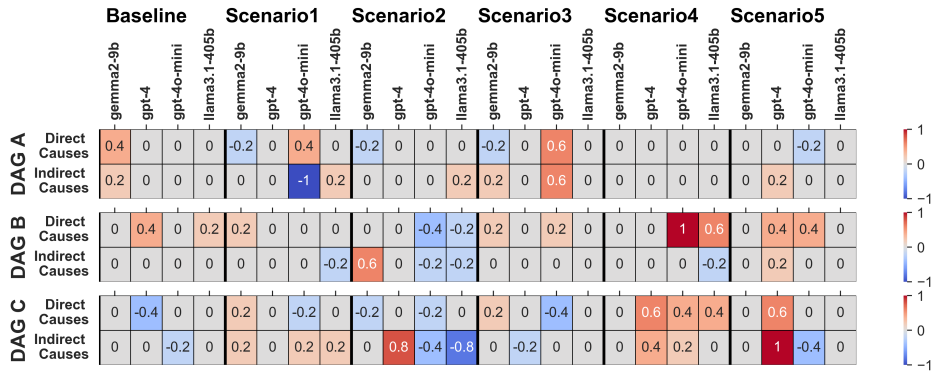

Figure 3: Heatmap showing the difference in the correct answers of simulations from changing the names of the variables from abstract names to names with real-world meanings.

4 Discussion

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

There are a few main findings from this work. We observed that changing the variable names from arbitrary names to real names had a differential effect on performance. In the adversarial scenarios where the other variables did not impact the outcome variable, the real variable names improved performance across most models. It is possible that for these adversarial cases, the causal knowledge about the variables used such as "coffee" and "heart rate", enhanced the ability of the LLM to correctly assess that the observed variables did not satisfy the causal relationship. However, in scenarios 1, 2, and 3, performance often decreased when using variables with real names. In our framework, 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 causal. Changing the way data were presented to LLM affected performance and often improved performance in scenarios 4 and 5. This suggests that textual representations of the data, as compared to raw values, may be important.

Future work should investigate how to best optimize the formatting of time series data for analysis by LLMs, and investigate models specifically designed for time series analysis. Lastly, in our analysis, we kept the temperature constant. In future analyses, we will explore performance at different

temperatures and with different ways of formatting the data. This work serves as a preliminary investigation into the ability of LLMs to explain causal relationships in time series data.

5 Acknowledgments

This work was supported by the National Science Foundation Graduate Research Fellowship Program under grant number 2141064.

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 et al. Gemma: Open models based on gemini research and technology. *arXiv [cs.CL]*, March 2024.

[4] OpenAI et al. GPT-4 technical report. *arXiv [cs.CL]*, March 2023.

[5] Aaron Grattafiori and et al. The llama 3 herd of models. *arXiv [cs.AI]*, July 2024.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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] 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.

[13] 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.

[14] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K Gupta, and Jingbo Shang. Large language models for time series: A survey. *arXiv [cs.LG]*, February 2024.

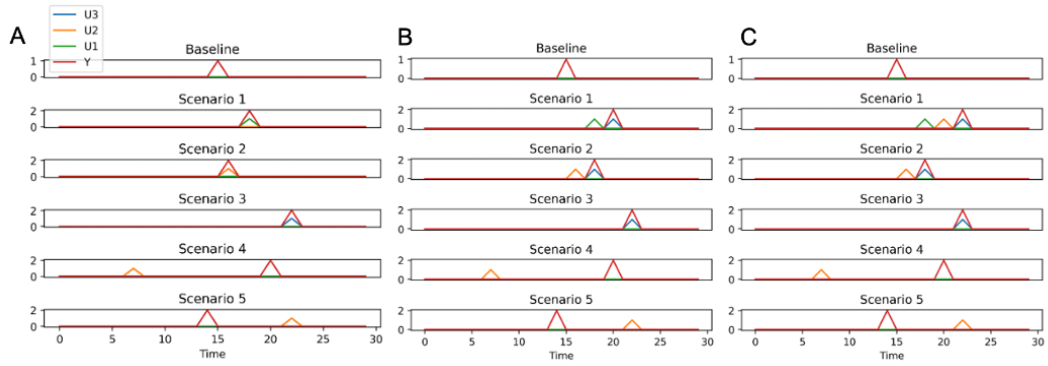

Figure 4: Synthetic Time Series Data.Examples Time Series Data: Baseline represents scenario where Y has step response without any cause from any input variables. Scenario 1 occurs when $U1$ has a step response. Scenario 2 occurs when $U2$ has a step response and Scenario 2 occurs when $U3$ has a step response.

A Appendix

A.1 Data Generation

The synthetic data was generated using the equations below and is plotted in Figure 4.

A

$$Y[t] = U1[t] * 2 + U2[t] * 2 + U3[t] * 2 \tag{1}$$

B

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

C

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

A.2 Prompt and Technical Details

For each experiment, a temperature of .2 was used. For the OpenAI models, we used GPT-4 version "gpt-4-0613" and GPT-4o-mini version "gpt-4o-mini-2024-07-18".

Prompts:

Table 1 shows an example of the components of the prompts. The data is formatted using raw values. Below, we show the alternative data format using text to represent the data.

> 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.

A.3 Change in data formatting

Figures 5 and 6 show the results when the data format was changed to be a textual representation of the time series data.

Prompt Components	
Instructions	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.
Relationships	Below are the descriptions of relationships between the data: U1 causes U2 to increase value by 1 after 2 time units. U2 causes U3 to increase value by 1 after 2 time units. U3 causes Y to increase value by 2 units at the same time. The effects are additive. Other factors may affect Y.
Data	U1: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, U2: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, U3: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, Y: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
Directions	Answer the following questions. Using the data, which of the input signals caused an observable effect on Y? 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? Format your answer as a list numbered with 1. List of direct variables or "none", 2. List of indirect variables or "none", and 3. An explanation.

Table 1: Prompt components example for the baseline scenario for DAG C. Prompts included all four of these components in the input.

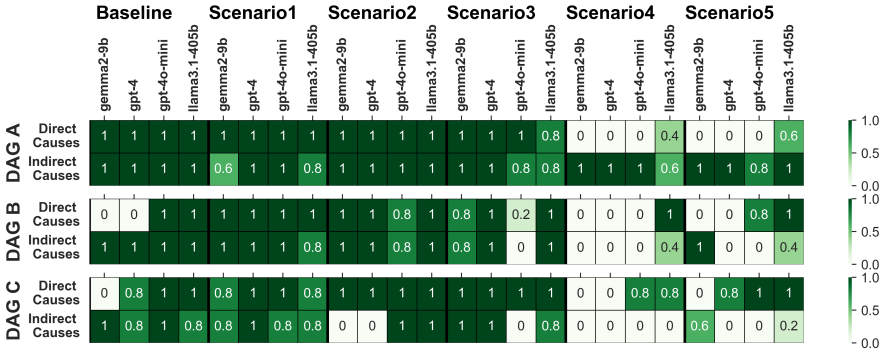

Figure 5: Heatmap showing tasks answered correctly using variables with abstract names and data formatted as text summary. Numbers show percentage of 5 random scenarios.

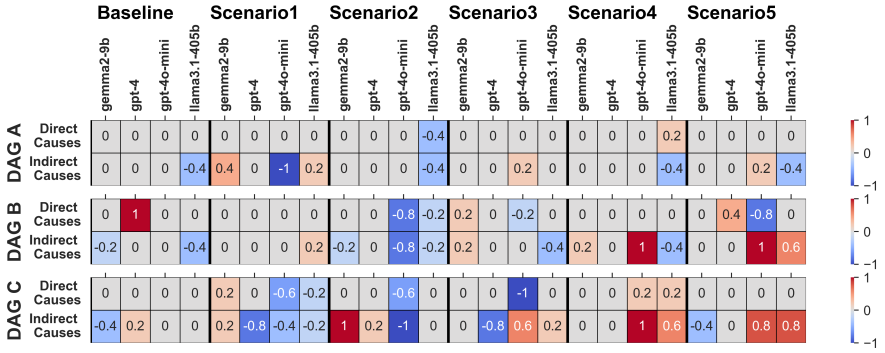

Figure 6: Heatmap showing the difference in the correct answers of simulations from changing the names of the variables from abstract names to names with real-world meanings with data formatted as text summary.