Leveraging LLMs to Infer Causality from Visualized Data: Alignments and Deviations from Human Judgments

Arran Zeyu Wang[*]
UNC-Chapel Hill

David Borland[†]
RENCI, UNC-Chapel Hill

David Gotz[‡]
UNC-Chapel Hill

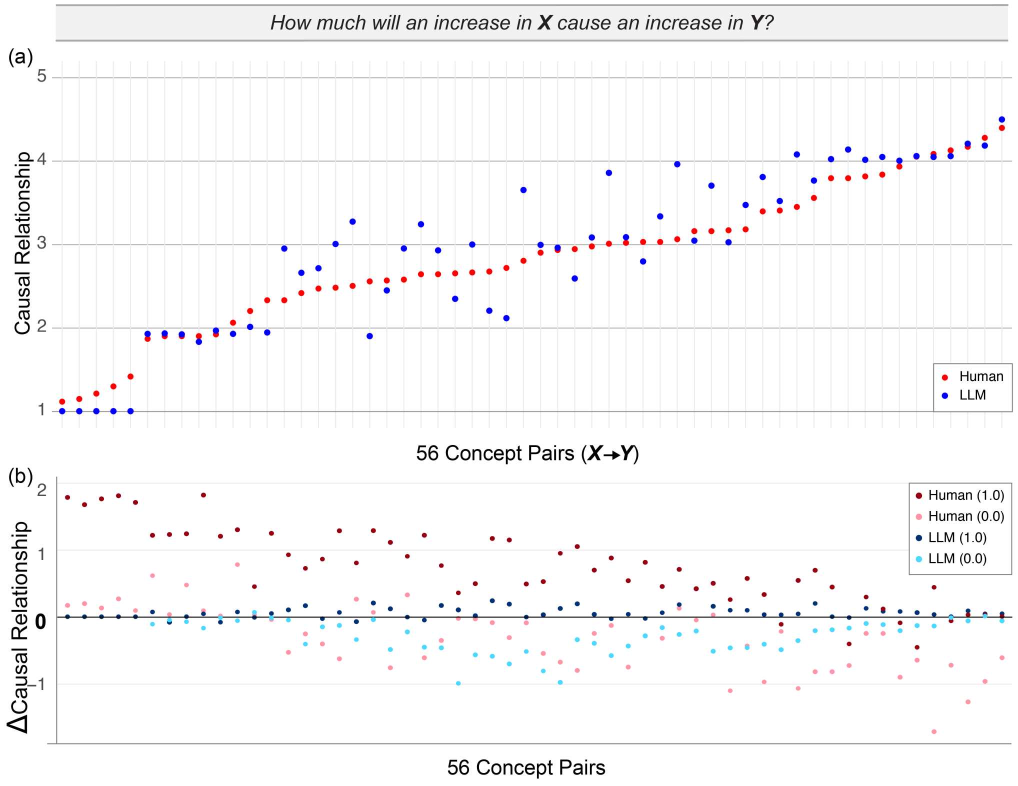

Figure 1: (a) Compares human-rated and LLM-rated causal ratings between 56 concept pairs collected from open-source datasets, as rated without viewing any visualizations. (b) Shows differences in human-rated and LLM-rated casual ratings after seeing charts with two different visualized association levels, 0.0 and 1.0. An association of 0.0 depicts no clear trend, and 1.0 a clear increasing trend. See Fig. 2 for examples. In both (a) and (b) concept pairs are ordered along the x-axis by increasing human rating.

Abstract

Data visualizations are commonly employed to convey relationships between variables from complex datasets in exploratory data analysis. Recent advancements in Large Language Models (LLMs) have shown surprising performance in assisting data analysis and visualization. In this poster, we investigate the capabilities of LLMs for reasoning about causality between concept pairs in visualized data using line charts, bar charts, and scatterplots. By using LLMs to replicate two human-subject empirical studies about causality judgments, we how their inferences about causality between concept pairs compare to those of humans, both with and without accompanying visualizations showing varying association levels. Our findings indicate that LLMs' causality inferences are more likely to align with human results without visualizations at very high or very low causal ratings, but LLMs are more influenced by low visualized associations and relatively unaffected by high visualized associations.

1 Introduction

Visualization is commonly used to infer causal relationships between depicted variables [1]. Although most visualizations depict correlations or other associations between variables, users often interpret these visualized associations—combined with their underlying preconceptions of causal links between variables—as causal relationships [7].

Recently, the rapid development of large language models (LLMs) has made great advances in assisting users in performing daily tasks. In the visualization community, such generative AI techniques have been widely applied to support visualization recommendation and generation [8]. With respect to interpreting data visualizations, a recent study reported that LLMs showed varying capabilities in low-level tasks and achieved extremely high performance in certain specific tasks such as retrieving values and finding anomalies [9].

However, it is still not clear how well LLMs can interpret visualizations at a higher level, such as reasoning about visual causal inference tasks. In this poster, we preliminarily explore the performance of LLMs for visual causal inference by replicating two human-subject studies [7]: (1) reporting the preconceived causal strength between concept pairs and (2) inferring the causal strength between the same concept pairs while presented with visualizations showing varying association levels between the concept pairs. We

[*]e-mail: zeyuwang@cs.unc.edu
[†]e-mail: borland@renci.org
[‡]e-mail: gotz@unc.edu

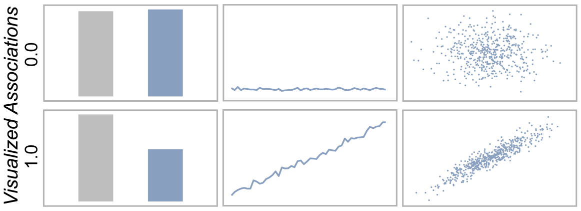

Figure 2: Example visualization stimuli for association levels 0.0 (top row) and 1.0 (bottom row).

report on and discuss the similarities and differences between LLMs and humans with respect to these tasks.

2 Background

The poster is related to visual causal inference and the usage of LLMs in visualizations. Causal inference has become an increasingly important topic in visualization research. Researchers have employed many techniques to support visual causal inference such as graphical causal models [2] or counterfactuals [6]. Most relevant, an empirical study found that users' perception of causality in visualized data is significantly impacted by both depicted associations in visualizations and their underlying prior related to variable pairs [7]. We replicate this study using LLMs and compare with their human-subject results.

LLMs have been widely used in visualization research, with most studies focused on automation, such as generating stylized charts [8]. In a recent study, Xu and Wall explored how LLMs can perform low-level statistical tasks in data visualizations, with LLMs performing well or struggling depending on task type [9]. However, they focused on analyzing low-level statistics, with limited attention to more complex visual comprehension tasks. In this work we begin exploring the use of LLMs for high-level reasoning by examining how LLMs respond to visual cues when making causal inferences.

3 Methodology

The study consisted of two tasks, following the procedure from [7], but replacing human feedback with feedback from an LLM.

Tasks: The tasks required users to estimate the causality between two concepts on a 5-point scale, e.g., "How much will an increase in X cause an increase in Y?" Task 1 asked users to report the causal strength of concept pairs with no visual stimuli, indicating a "causal prior." Task 2 asked a different group of users to judge causal strengths of the same concept pairs while presented with charts showing different visualized associations.

Stimuli: The concept pairs were collected from variables in widely used machine learning datasets. For Taks 2, three chart types were used, line charts, bar charts, and scatterplots, each with two association levels: 0.0 and 1.0 (Fig. 2). These association levels are indicated by the average trend or differences shown in the visualizations. Each association level was shown for each variable pair, such that each variable pair appeared twice in Task 2.

Please see [7] for more detailed task and stimuli settings, noting that in that study five different association levels were used.

LLM settings: We employed OpenAI's GPT-4 model [4] to complete the tasks. We provided the task descriptions and original questions from the human-subject study to GPT, and for Task 2 we also provided images showing the corresponding visualizations. As with the human study, and aligning with previous suggestions [9], we explicitly instructed the answer to be an integer in the range [1, 5], and for the second task, we explicitly instructed GPT to consider the provided chart when answering the question. We asked each question fifty times and calculated the average score as the causal strength rating. As the original study focused on the general public using MTurk, we instructed GPT to rate the causal strength based on the general public's knowledge and not professional research papers.

4 Results and Discussion

Fig. 1 (a) shows the average causal strength ratings between concept pairs for both humans and the LLM. These results indicate that the LLM's causal strength ratings align with human judgments when the average causal relationships are very low (i.e., on the left-hand side roughly smaller than 2) or very high (i.e., on the right-hand side roughly larger than 4). However, when the concept pairs' causal strengths are toward the middle of the range (i.e., from 2 to 4) for human ratings, we see a larger deviation between the LLM and human ratings.

Fig. 1 (b) shows differences between causal strength ratings for both humans and the LLM after seeing visualizations showing the two association levels. We see that overall, the LLM results are more stable than human results, indicating that the LLM relies more on preconceived causal relationships between the terms, and that humans respond more to the visualized associations. While human results can be significantly impacted by visualized association levels, we see that when the average causal strength rating between concept pairs is at a very low or very high level, the LLM tends to maintain a stable rating, no matter what associations are shown in the visualizations. For concept pairs with less extreme causal strengths, however, we see that the LLM is more likely to be impacted by lower visualized associations compared to higher ones. In fact, we do not notice any obvious impact of high association levels in visualizations on the LLMs causality strength ratings across all concept pairs.

The results indicate that LLMs do align well with human judgments of causality to some extent, agreeing with results from previous studies [3]. However, the observed deviations of ratings when shown visual stimuli reveal that LLMs differ from humans when performing high-level comprehension tasks from visualizations. This suggests the need for LLMs that can better understand human perception in relation to high-level comprehension from visualizations [5].

However, the investigation is still preliminary and limited in some ways. For example, we only studied the GPT-4 model, focused on three chart types, and lacked explorations of more complex prompt settings. Those aspects, as well as a deeper dive into the reasons why larger deviations occur in the concept pairs with middle causal priors, should be studied in future research.

Acknowledgments

This research is made possible in part by NSF Award #2211845.

References

[1] D. Borland, A. Z. Wang, and D. Gotz. Using counterfactuals to improve causal inferences from visualizations. *IEEE CG&A*, 44(1):95–104, 2024.

[2] G. Guo, E. Karavani, A. Endert, and B. C. Kwon. Causalvis: Visualizations for causal inference. In *ACM CHI*, pp. 1–20, 2023.

[3] C. R. Jones and B. K. Bergen. People cannot distinguish gpt-4 from a human in a turing test. *arXiv preprint arXiv:2405.08007*, 2024.

[4] OpenAI. Chatgpt-4: Optimizing language models for dialogue., 2023. Accessed: 2024-05-15.

[5] G. J. Quadri, A. Z. Wang, Z. Wang, J. Adorno, P. Rosen, and D. A. Szafir. Do You See What I See? A Qualitative Study Eliciting High-Level Visualization Comprehension. In *ACM CHI*, pp. 1–26, 2024.

[6] A. Z. Wang, D. Borland, and D. Gotz. An empirical study of counterfactual visualization to support visual causal inference. *Information Visualization*, 23(2):197–214, 2024.

[7] A. Z. Wang, D. Borland, T. Peck, W. Wang, and D. Gotz. Causal priors and their influence on judgements of causality in visualized data. *IEEE TVCG (Proc. IEEE VIS 2024)*, 2025.

[8] J. Wu, J. J. Y. Chung, and E. Adar. viz2viz: Prompt-driven stylized visualization generation using a diffusion model. *arXiv preprint arXiv:2304.01919*, 2023.

[9] Z. Xu and E. Wall. Exploring the capability of llms in performing low-level visual analytic tasks on svg data visualizations. *IEEE VIS Short Papers*, 2024.