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

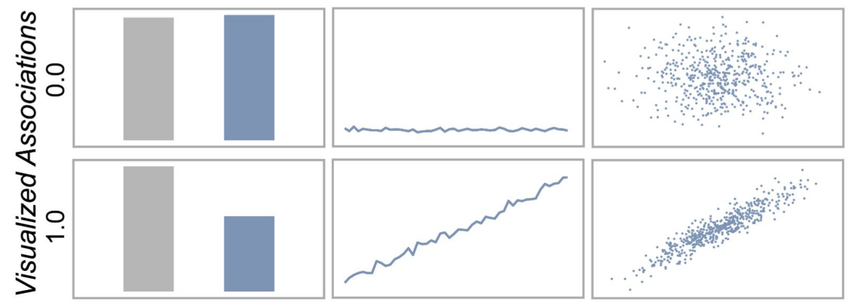
Arran Zeyu Wang, David Borland, *and* ***David Gotz***

University of North Carolina at Chapel Hill

Introduction

- LLMs have demonstrated extremely high performance in certain low-level tasks such as *retrieving values* and *finding anomalies*[2].
- It is still not known how well LLMs can interpret visualizations at a higher level, such as ***reasoning about causality from visualizations***.

Causality Judgment Tasks[1]



How much will an increase in X cause an increase in Y? (on a scale from 1 to 5)

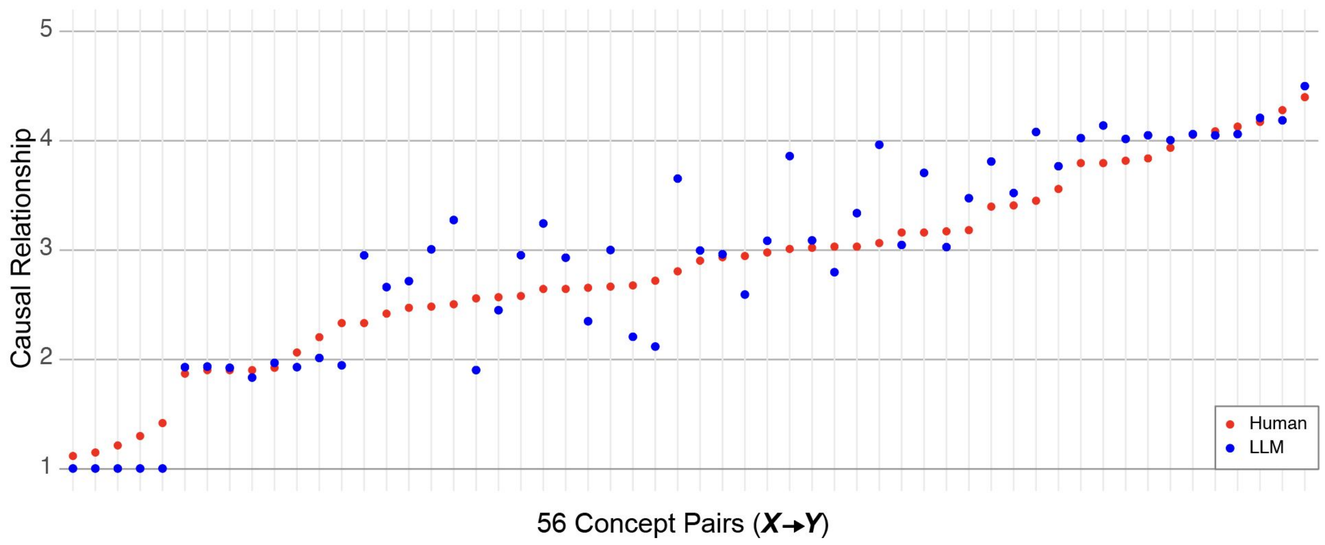
- Inferring the *causal strength between concept pairs* **with** and **without** *visualizations showing varying association levels*

Settings

Basic Settings:
- OpenAI's GPT-4 Model
- Same concept pairs and questions as [1]
- 50x for each concept pair

Prompts:
- Provide task descriptions first
- Ask same questions w/ and w/o charts
- Answer based on the general public's knowledge
- Do not search for professional research papers

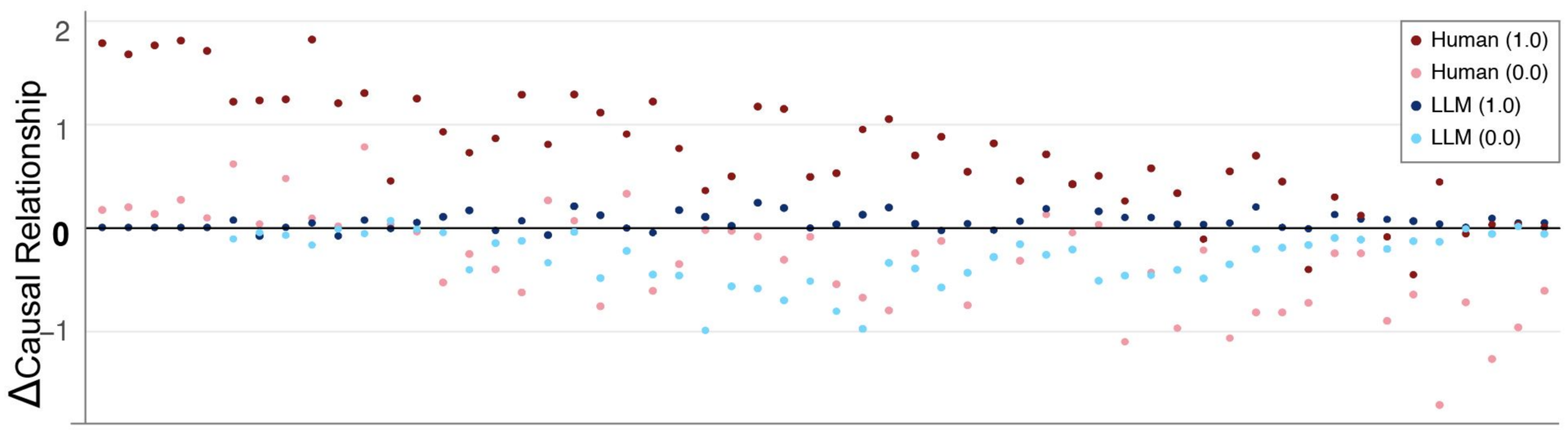
Judging without Visualizations



Comparing **human-rated** and **LLM-rated** causal ratings between 56 concept pairs collected from open-source datasets, as rated without viewing any visualizations.

1. LLM's causal strength ratings show better ***alignment*** with human ratings when:
   a. the average causal relationships are **very low** (i.e., on the left-hand side roughly smaller than 2)
   b. or **very high** (i.e., on the right-hand side roughly larger than 4).
2. The ratings show ***deviations*** when the concept pairs' causal strengths are toward **the middle of the range** (i.e., from 2 to 4)

Judging with Visualizations



Differences in casual ratings for both humans and LLMs after seeing charts with two different visualized association levels, 0.0 and 1.0.

1. The **LLM ratings have overall smaller differences than the human ratings**, indicating that the LLM relies more on preconceived causal relationships between the terms, whereas humans respond more to the visualized associations.
2. While human results can be significantly impacted by visualized association levels, **the LLM tends to maintain a stable rating with a very low or very high level causal priors**, no matter what associations are shown in the visualizations.
3. 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.

[1] Wang et al. Causal priors and their influence on judgements of causality in visualized data. IEEE TVCG (Proc. IEEE VIS 2024), 2025.
[2] Xu and Wall. Exploring the capability of llms in performing low-level visual analytic tasks on svg data visualizations. IEEE VIS Short Papers, 2024.