LLMs Are Not Reliable Human Proxies to Study Affordances in Data Visualizations

Kylie Lin
klin368@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Chase Stokes
cstokes@ischool.berkeley.edu
University of California, Berkeley
Berkeley, California, USA

Cindy Xiong Bearfield
cxiong@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

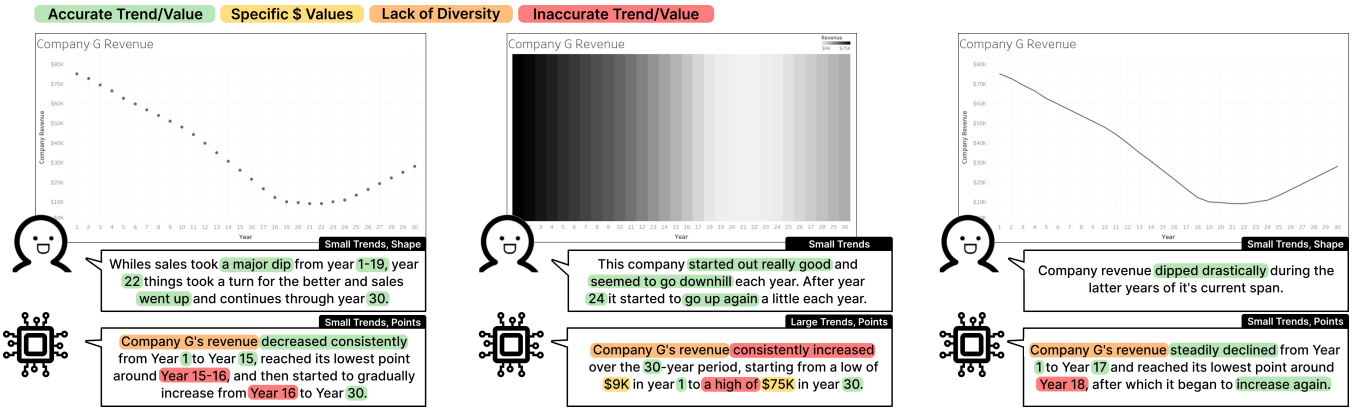


Figure 1: In our work, we identify key limitations of GPT-4o in predicting human takeaways from visualizations including inaccuracies, lack of semantic diversity, and failures to adequately capture visualization affordances. From these results, we conclude that LLMs are not yet reliable proxies for human responses.

Abstract

Identifying the relationship between data visualization design and human interpretation often requires time-consuming crowdsourced studies that generate large text corpora. Given recent academic exploration into the use of large language models (LLMs) as proxies for human study participants, we anticipate interest within the visualization community on LLM predictions as proxies for human chart interpretation. We present a case study on the effectiveness of OpenAI's GPT-4o model to predict human takeaways from charts. Using the lens of visualization affordances, we conduct a factor analysis on human chart takeaways, identifying five affordance factors. We then compare the affordances of different chart types between human readers and GPT-4o, revealing discrepancies in takeaway accuracy, semantic diversity, response length, and alignment with human interpretations. We caution against using LLMs as human proxies in empirical studies and outline critical directions for future research on LLM predictions of human reasoning with data visualizations.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; *Visualization design and evaluation methods*; *Information visualization.*

Keywords

Information visualization, affordance, language models, human subjects

2nd HEAL Workshop at CHI Conference on Human Factors in Computing Systems, Yokohama, Japan
2025.

1 Introduction

In recent years, academic research communities have debated the feasibility of using Large Language Models (LLMs) as proxies for human subjects in empirical studies [16, 27] and begun to explore strategies for generating synthetic research data [4, 22, 25, 30]. At the same time, the visualization research community has begun to more broadly examine the use of LLMs to assist in chart interpretation, identifying difficulties that LLMs have when emulating human behavior [7, 57]. Given increasing interest in how LLMs might perform as human proxies in visualization research studies, we present an investigation into the ability of a state-of-the-art LLM to match human responses to data visualizations.

In this work, we focus on the relationship between data visualization design and reader responses through the lens of *visualization affordances* [8, 15, 52]. Choices of visual encoding during the visualization design process, such as spatial arrangement, can influence what people conclude from data [6, 20]. Recently, visualization researchers have likened the visualization design space to a soundboard for audio mixing, where every design decision acts as a switch or knob that alters the resulting design [55]. Design adjustments result in distinct visual experiences that influence reader interpretations in unique ways. A visualization affordance, then, is the unique relationship between design decisions and a reader's interpretation of information communicated via the design [58].

To better understand the affordances of different visualization designs, visualization researchers and designers need a scalable model to predict human responses based on visualization design choices. Current efforts to develop these models involve labor-intensive empirical studies and collecting large corpora of qualitative data on human responses, which often have high degrees of lexical and semantic ambiguity [20, 59].

To examine the alignment between LLM outputs and human visualization interpretation, we compare the quality of human takeaways from visualizations to takeaways generated by OpenAI's GPT-4o model [1] and ask two primary research questions:

(1) **How well can a state-of-the-art LLM emulate human takeaways from a visualization?**
(2) **How can experimenters evaluate the capabilities of LLMs as proxies for human participants in visualization studies?**

To answer these questions, we investigate how three different chart types (dot plots, heatmaps, and line charts) influence people's main takeaways from the charts in a crowdsourced study, identifying five affordance factors: *points*, *small trends*, *shape*, *large trends*, and *clusters*. This preliminary study lays the groundwork for our main study, where we elicit human takeaways from an expanded set of experimental stimuli and then adapt the procedure into a prompt given to GPT-4o. Finally, we compare the responses from humans to those from GPT-4o to identify issues of accuracy, diversity, and failure in matching human takeaways.

We contribute: (1) Five affordance factors of human takeaways from visualizations, (2) A systematic comparison of human takeaways and GPT-4o-generated takeaways to assess the model's ability to emulate human insights from visualizations; (3) Considerations for using LLMs to emulate human responses in empirical visualization studies.

2 Related Works

Researchers have started testing the extent to which LLMs are capable of simulating human subjects research data, exploring the ability of models to match human moral judgments [16], economic decision-making behaviors [30], and text annotation abilities [22]. As a result, members of the HCI community have discussed the validity and ethical appropriateness of LLMs in empirical work [5], highlighting concerns such as the extent to which crowdworkers might leverage LLMs to complete studies [25]. In this work, we seek to extend the discussion on LLMs as proxies of human subjects in the context of visualization research space via an inital case study on the ability of an LLM to capture visualization affordances.

2.1 Visualization Affordances

Design choices (e.g. groupings of bars in a bar chart [6]) have important implications for how readers interpret presented data. While effective visualization designs can improve information processing [43, 59], poor designs can obscure or distort the intended messages of a chart [12, 51, 55]. To address the critical relationship between design and reader interpretation, visualization researchers have aggregated takeaways from empirical studies into a structured framework that optimizes data communication [29, 36, 39].

For example, bar charts encourage readers to make magnitude comparisons, such as "A is larger than B," while a line graph highlights trends and changes over time, such as "A is increasing at a higher rate than B" [6, 48, 60]. Visualizations that aggregate data points, including bar charts, can lead readers to infer causality, whereas those that display probabilistic outcomes, such as scatterplots, promote better understanding of uncertainty [28, 33].

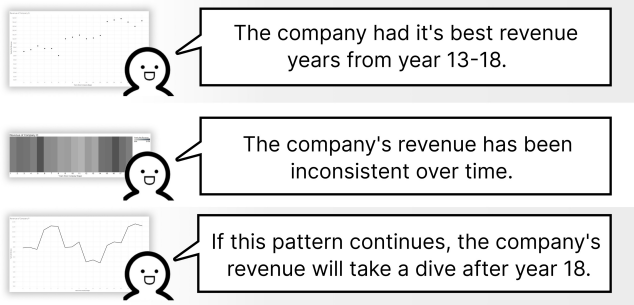
Design choices also influence decisions that readers make [42, 63, 64]. For example, in an investigation on representations of wildfire risk, researchers found that icon arrays with a small number of icons resulted in distinct decision making patterns compared to numerical representations and icon arrays with a large number of icons [38]. Providing explanation for these results, related work has found that people focus on the denominator of icon arrays, interpreting a larger number of icons as a 'less risky' scenario [46]. In contrast, numerical representations afford specific calculations, increasing the chance of data consumers using complex reasoning strategies to reach a conclusion, rather than taking a mental shortcut [58].

2.2 Evaluating LLMs on Visualization Tasks

Large Language Models are advanced statistical models pre-trained on vast corpora of natural language data that use the predicted likelihood of words in a sequence to generate response to natural language queries. Recently, these models have been fine-tuned for tasks related to visualizations and visual analytics [11, 14, 24, 35, 54]. Highly capable LLMs, such as OpenAI's GPT-4, are increasingly integrated into user workflows, with remarkable performance across natural language tasks related to visualization [1, 40].

Researchers have developed benchmarks for evaluating LLMs on visualization related tasks [26, 32, 61]. For example, VisEval [13] takes an automated approach in evaluating LLMs for issues related to the validity, legality, and readability of their output. Xu and Wall [62] have evaluated how well LLMs perform low-level visual analytic tasks as outlined by Amar et al. [3], using SVG-based visualizations as input. At this level, LLMs are able to adequately respond to visualizations and even identify deceptive visualizations [37]. To better account for inaccuracies and hallucinations in LLM output, Goswami et al. developed ChartCitor as a multi-agent framework for grounding LLM chart interpretations to the chart image [23].

At the same time, recent work has also begun to evaluate LLM chart interpretation capabilities by comparing output to human responses [57], finding that LLMs appear limited in their ability to emulate human reactions. For example, presented with varying spatial arrangements of the same dataset (e.g., horizontal vs. vertically aligned bar charts), viewers are likely to make different visual comparisons and draw unique conclusions [21, 59]. LLMs, on the other hand, are not sensitive to manipulations of spatial arrangement and are more influenced by the topic of the dataset [57]. Further, when recommending visual encoding specifications, suggestions generated by GPT models differ from best-practice guidelines established through human-subject experiments [56]. Drawing on these findings, we anticipate that LLMs are likely to also be limited in their ability to provide human-like perspectives on a visualization because they struggle to predict visualization affordances.



Intuitive Task	Unit Reason	Perceptual Task	Scope	Factor
Identify Cluster Value	X Value Series *Average Y Value Extreme*	Extreme	Local	**Points**
Identify Overall Shape	Motif *Complex Shape Exists*	Distribution	Global	**Shape**
Identify Overall Trend	All *Trend Exists/Extrapolation*	Prediction	Global	**Large Trends**

Figure 2: Example charts and conclusions generated from participants in our preliminary study, with category schemes and factors assigned to each conclusion.

In this work, we investigate the sensitivity of a state-of-the-art LLM to visualization designs to inform the use of LLMs as participant proxies in visualization studies and guide the technical development and usage of LLMs to better predict visualization affordances. While computational social sciences research has identified processes of conditioning models with human sociodemographic data [4], we present a case study with GPT-4o [1] out-of-the-box to determine a baseline of LLM capabilities.

3 Study Overview

We present a two-part study that (1) establishes a baseline set of visualization affordances characterizing canonical data visualization types and (2) compares chart takeaways generated by GPT-4o to takeaways provided by study participants.

Preliminary study We begin by assessing the diversity of reader percepts across three canonical visualization designs. We identify five factors characterizing readers' perceptions of visualizations, which we then used to inform our subsequent experiments and serve as visualization affordance factors.

Main Study Next, we analyze human free-response takeaways from data visualizations and compare human responses to takeaways provided by the GPT-4o model.

4 Preliminary Study: Affordance Factors

As a preliminary study, we crowdsourced takeaways from participants reading dot plots, line charts, and heat maps. We categorized the takeaways into two sets of category schemes: 'bottom-up data-driven schemes' and 'top-down theory-driven schemes.' From these schemes, we performed a factor analysis to identify five factors that people derive from reading visualizations. These factors serve as the visualization affordance factors in the main study of this work.

4.1 Procedure

We recruited 62 participants for this study through the online crowdsourcing platform Prolific [44]. They were compensated $7.13 for a 45-minute survey hosted using Qualtrics [50] survey software. Participants viewed and reported their main takeaways from six charts (two each of dot plots, line charts, and heatmaps). All charts depicted a neutral topic of *functional company revenue over 18 years*. Participants first completed two practice trials with unique practice data sets. For each chart, participants were asked to type the first takeaway they drew from the chart with the statement, *"Please write down the FIRST conclusion you drew from this data. Mention any relevant year(s) in your response."* After entering the takeaway, participants reported the range of years used in the takeaway. From there, they entered their second and third takeaways.

After collecting participant responses, we categorized takeaways in two steps. We first constructed a set of 49 codes that represented which sentences referred to highly similar percepts or data features (e.g., a single increasing trend). Next, we coded all takeaways multiple times, each using a different coding scheme referencing taxonomies in visualization and psychology research literature [10, 47, 49, 53], as shown in Figure 2.

The **Intuitive Task** coding scheme was derived from a data-driven thematic analysis that clustered into ten codes (e.g., overall trend, end point comparison, between group value). The **Unit Reason** scheme included information on the unit the participant selected (e.g., a *data point*, a *subset of data*), the property of the unit described (e.g., *trend*), and the operations performed with or across units (e.g., *comparison of similar units*). The other coding schemes were based on prior taxonomies. The **Perceptual Task** scheme was based on perceptual tasks from Amar et al. [2]. Some of these tasks required participants to identify specific values, so we adjusted those tasks to better characterize takeaways from our participants. This resulted in 11 tasks (e.g., extreme, cluster, trend). We also coded for the **Perceptual Scope** of the takeaways, referencing perceptual psychology literature [41]. When encountering a new scene or object, viewers tend to first process the broader *global* shape or feature statistic (e.g., the dataset as a whole), before diving down into *local* components (e.g., a subset of data, a specific point) [41].

4.2 Generating Affordance Factors

We collected 1,161 responses (390 from dot plots, 384 from heatmaps, and 387 from line charts). We conducted an exploratory factor analysis across coding schemes using the `Psych` R package [45]. We compared the empirical BIC and model complexity of factor

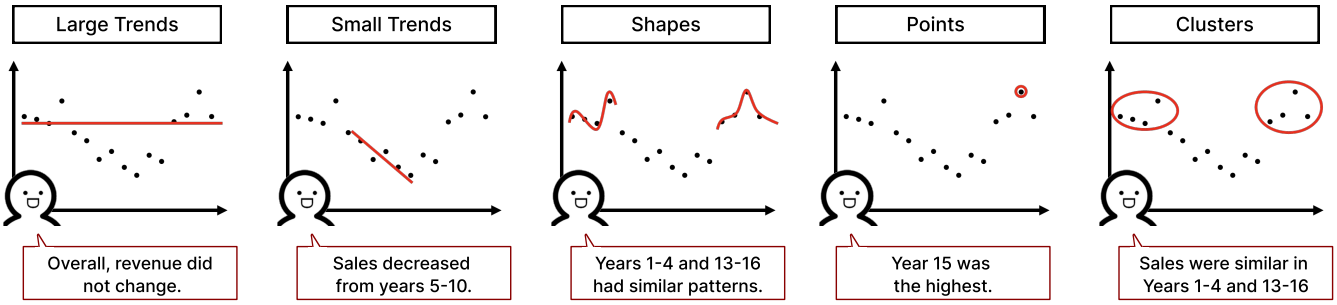


Figure 3: Example takeaways of the five factors selected from exploratory factor analysis.

models consisting of 1 to 9 factors as well as how factors within each model correlated with each other and within each factor. Based on the balance of attributes, we applied the five factor model to our codes. The five factors became our identified affordance factors: *point* selection and values, *small trends* and changes, overall *shape*, *large trends* and descriptions, and *clusters* of data points.

Points encompassed the selection of specific points and values, typically the maximum and minimum points. It also contained the specification of values and the range of the data. **Small Trends** included the selection of subsets of data to examine and the description of trends within these subsets. It also contained the description of large changes between adjacent points, such as 'spikes' or 'dips' between points. **Shapes** contained the description of overall shape of the distribution or patterns within subsets of data. **Large Trends** covered takeaways which examined and described the chart holistically and encompassed all the data points. Takeaways regarding predictions about future data points and calculations of overall averages were also included in this factor, since participants used the chart as a whole to determine these features. **Clusters** involved any grouping or clustering of the data with similar y-values.

5 Main Study: Human vs. GPT-4o Takeaways

Having established affordance factors that characterize human takeaways in dot plots, heatmaps, and line charts, the following study examines human and LLM-provided takeaways from each chart type. We recruited participants for a survey study, where they were shown charts and provided free-response takeaways. At the same time, we used the survey instructions and stimuli to create a prompt asking the GPT-4o model to provide takeaways.

5.1 Participants

Based on a power analysis using the G*Power software [19], 700 participants would provide approximately 85% power to detect an effect at alpha = 0.05. We recruited 770 participants from Prolific [44], filtering for participants whose primary language was English and with an approval rate above 98%. After excluding participants based on an attention check and the quality of their response, the final sample size was 716. Most participants (57%) were between the ages of 25 and 44. Many participants (41%) had a 4-year degree.

5.2 Experimental Stimuli

To improve generalizabilty of our results, we expanded the stimuli set to include a wider range of data patterns. We examined the data trends present in the MASSVIS "targets393" chart corpus [9] and used a card-sorting procedure to identify a set of distinct patterns that captured real-world variations while minimizing redundancy. We then refined the set to ensure balanced representation of overall data direction (i.e., net change over time). This process resulted in 45 total charts (3 chart types x 15 dataset patterns).

Some dataset patterns showed steady increases and decreases, with varying levels of consistency. Others followed distinctly nonlinear trends, such as exponential growth or logarithmic decay. Several patterns exhibited cyclical behavior with repeating peaks and troughs, while others depicted bimodal distributions. These patterns captured a broad range of data behaviors and aligned with existing work identifying key pattern types in data [17]. We used line charts, dot plots, and heatmaps [8] to examine a variety of visual encodings and emphasize different features of the datasets.

5.3 Procedure

5.3.1 Eliciting Human Takeaways. Participants gave informed consent and then completed a Qualtrics survey. First, they completed an attention check and a practice trial on writing natural language chart takeaways. Next, participants viewed a single chart randomly selected from our pool of 45 and reported their main takeaway from the data. To resolve potential ambiguities in their free-response takeaways, participants also provided the range of years they had considered when writing the takeaway and the overall unit of focus (e.g., point, subset, chart, etc.). Finally, participants reported demographic information (gender, age, and education). The survey took around 4 minutes, and participants were paid $0.80.

We checked participant takeaways for correctness and coded them according to the five affordance factors identified in the preliminary study. Takeaways were randomly distributed between authors such that sixty percent were coded by multiple authors to assess coder consistency. The average κ was 0.73, demonstrating consistent agreement. All authors coded the data individually and then met to discuss discrepancies to finalize the assigned codes.

5.3.2 Prompting GPT-4o. We adhered closely to the instructions from the online survey provided to participants to construct a prompt to provide to GPT-4o with its default parameter settings:

System Prompt: I am a visualization researcher and you are a participant in a research study on graphical perception for data visualizations. You should act as a research participant to provide 1) the MAIN conclusion you draw from a data visualization, 2) relevant years in your response, and 3) what your conclusion specifically focuses on.

User Prompt: I will provide you with an image of a data visualization. You will be asked to provide the MAIN conclusion you draw from the data. You should mention any relevant year(s) in your response. Based on the data visualization provided, please provide the MAIN conclusion you drew from this data and mention any relevant year(s) in your response. Provide your answer in three parts: 1) the MAIN conclusion you draw from the data with relevant year(s) in your response 2) What year(s) did you use in this conclusion? Enter only numbers or ranges of numbers separated by commas. For example, 3-8, 10. 3) Out of the following options, which does your conclusion most specifically focus on: individual singular data point(s), subset(s) or groupings of data, the entire dataset, chart axes or title, or other.

« Image of Chart »

We provided this prompt to GPT-4o 15 times for each of the 45 visualizations, resulting in 675 takeaways. The authors coded the takeaways provided by GPT-4o according to the five affordance factors, as in Section 5.3.1. The average κ was 0.71, demonstrating consistent agreement. After completing individual coding, all authors met to resolve discrepancies.

5.4 Results

Overall, humans produced more takeaways that accurately described information from the charts compared to GPT-4o, and while GPT-4o produced lengthier responses on average, variation in length was greater in human responses. In terms of affordances, we found minimal overlap between affordance factors in takeaways from humans and factors in takeaways from GPT-4o.

5.4.1 Humans Outperform GPT-4o In Terms of Accuracy. Almost all human responses (96.6%) accurately described the presented charts, with only 24 takeaways being incorrect. We categorized the inaccuracies as inaccurately describing the *trend* in the chart (9 total, including 5 people who reversed the trend and 4 people who omitted an essential part of a trend), listing *inaccurate values* (7), or containing *likely typos* (9). Of the five responses that reversed the depicted trend (e.g., increasing instead of decreasing), four described heatmaps and one described a line chart.

GPT-4o performed much worse than humans in providing accurate descriptions of the charts. We found that 428 (63.4%) takeaways were accurate while 247 (36.5%) contained inaccurate information. Of the inaccurate conclusions, 83 (33.6%) inaccurately described the trend in the chart, and most of these (n = 77) described the *opposite trend*, such as indicating an increase instead of a decrease. All of these errors occurred with heatmaps, suggesting a potential systematic weakness of GPT-4o. The other six takeaways in this category hallucinated other trends not present in the visualization. Moreover, 168 (68.0%) takeaways listed inaccurate values: 115 described an inaccurate *peak value* when describing maximum or minimum revenue, 27 described an *offset cycle of data* for charts with a sinusoidal pattern, 23 included an otherwise incorrect value, and 3 depicted a *cycle with inaccurate frequency* for sinusoidal patterns.

5.4.2 Human Takeaways are More Diverse than Takeaways from GPT-4o. While coding the two sets of takeaways, we noticed a qualitative difference in the wording of takeaways generated by GPT-4o and the wording of takeaways generated by humans. Responses generated by GPT-4o typically followed a similar format (e.g., many takeaways started with the phrasing "Company {A, B, C, etc.}'s revenue {increased, decreased, etc.}"; see Figure 1).

We conducted a two-sample t-test to compare text length across humans and GPT-4o and found that takeaways generated by GPT-4o were significantly longer than takeaways generated by humans in terms of the number of characters in each takeaway ($Mean_{GPT}$ = 145.86, $Mean_{Human}$ = 102.39; $t = 14.55$, $p < 0.001$). Examining the standard deviations (SD) of the number of characters between humans and GPT-4o, we found that the SD of responses of human takeaways was much larger (67.32) than that of GPT-4o (38.67), indicating greater variation in the length of human takeaways.

5.4.3 Human Takeaways Contain Fewer Factors than Those Produced by GPT-4o. We found that 34% of human takeaways mentioned more than one factor, with each takeaway having received 1.4 factor codes on average. On the other hand, over three-quarters (78%) of the takeaways produced by GPT-4o included more than one factor, with each takeaway receiving 1.8 factor codes on average.

As partial explanation for this phenomenon, only 6% of responses mentioned a specific dollar value from the visualization. For GPT elicited takeaways, however, 26% of responses mentioned a specific dollar value. Because the presence of specific values was used to code a takeaway with the *points* factor, more takeaways produced by GPT-4o were coded as *points* in addition to other factors.

5.4.4 Affordances Partially Align Between Humans and GPT-4o. For human responses, the most common factor was *small trends*, followed by *clusters*. On the other hand, the most common factor in GPT-4o takeaways was *points*, followed by *small trends*. Taking into account the fact that GPT-4o's tendency to include specific values increases the overall amount of takeaways categorized as *points*, we identify partial alignment between the most common takeaways from humans and GPT-4o.

For human responses, we found significant variations in takeaways across chart type via a chi-squared test ($\chi^2 = 46.3, df = 8, p < 0.001$). From examining the standardized residuals of the test, we found that takeaways with the *clusters* affordance factor were much more common in heatmaps than in other chart types ($R = 6.063$). Takeaways with the *shapes* factor were more common in dot plots ($R = 2.432$), and takeaways with *small trends* were more common in line charts than other chart types ($R = 2.089$). For takeaways generated by GPT-4o, we again found variations by chart type ($\chi^2 = 304.24$, df = 8, p<0.001). The *small trends* factor ($R = 6.057$) was most common in dot plots. *Large trends* ($R = 14.734$)

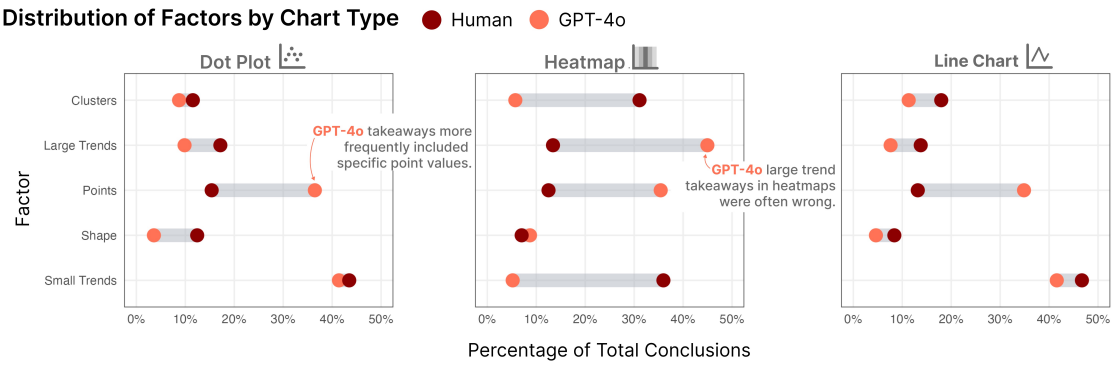


Figure 4: Comparison of human and GPT-4o takeaways. While there was alignment on some factors (e.g., small trends takeaways from dot plots), other factors displayed less consistent alignment within chart type (e.g., large trends takeaways from heatmaps).

and *shape* ($R = 3.28$) factors were more common in heatmaps than other chart types. *Clusters* ($R = 2.349$) and *small trends* ($R = 6.012$) factors were more common in line charts.

We compare the distribution of factors by chart type across human and GPT-4o takeaways in Figure 4. Across all chart types, GPT-4o responses included more takeaways containing the *points* factor. Dot plots elicited primarily *small trends* takeaways from GPT-4o, in line with the human responses. However, GPT-4o takeaways did not capture the *shapes* or *clusters* affordances at as high a frequency as human takeaways. The greatest discrepancy in human and GPT-4o takeaways was found for heatmaps, which resulted in primarily *large trends* and *points* takeaways from GPT-4o and primarily *small trends* and *clusters* takeaways from humans. We observed the greatest alignment between human and GPT-4o takeaways in line charts, which elicited primarily *small trends* takeaways from both humans and GPT-4o but lower relative amounts of *clusters*, *large trends*, and *shape* conclusions from GPT-4o compared to humans.

6 Discussion

Our findings reveal significant limitations in GPT-4o's ability to serve as a human proxy in visualization studies. While GPT-4o can generate structured responses that appear to interpret visualizations, its performance critically diverges from human interpretation in accuracy, diversity, and affordance alignment. Given these limitations, **we caution against the use of AI as a proxy for human participants in visualization studies**. However, we acknowledge that approaches involving alternative model preconditioning [18], strategic prompting methods [31, 35], and the use multiple LLMs [34] may enhance LLM performance to be closer to that of humans. Based on our work, we articulate the following considerations for researchers attempting to leverage LLMs for large-scale, human-like visualization interpretation studies.

Establish a human baseline before collecting LLM data. Prior to prompting LLMs in visualization studies, researchers should collect a small but representative set of human responses to establish baseline expectations for accuracy, diversity, and affordance alignment. This allows for direct comparisons between LLM outputs and human reasoning, ensuring that LLM responses are not grossly misrepresentative of human cognition. While this may increase the cost and time required to conduct the study, it offers the critical advantage of ensuring that accuracy rates and semantic diversity are similar between populations.

Evaluate accuracy with a structured error analysis. LLM responses can introduce systematic inaccuracies. By categorizing and comparing these errors to common errors in the human baseline, researchers can provide additional context in the LLM prompt. These errors may also vary in severity; reporting an approximate rather than exact value would be inconsequential in this study, but the complete misinterpretation of a dataset trend was significant.

Consider fine-tuning or k-shot prompting LLM with human data. Human baseline responses can be used to fine-tune an LLM to better capture the nuances of human responses. In this case, using human-generated takeaways as training data or as examples during prompting could create more parity between LLM and human responses. This may assist with both semantic diversity and the differences found between human and LLM results.

Limit use of LLM participants to very basic visualization studies. Our experiment with GPT-4o revealed difficulties with less common encoding mechanisms (i.e., color mapping in a heatmap). As such, it may be necessary to either exclude relevant visualizations from LLM-proxy studies, thus decreasing the complexity of the visualization study itself. Additionally, if the task required of the LLM proxy is relatively complex or nuanced, such as the extraction of a conclusion, LLMs may not be able to replicate the cognitive processes which support human performance on these tasks. For simpler or less cognitively demanding tasks, the similarity of LLM and human responses may be higher.

7 Conclusion

While GPT-4o exhibited some degree of overlap with human responses, such as the common identification of small trends in dot plots and line charts, its high error rate, lack of semantic diversity, and failure to completely align with human affordances result in an unreliable substitute for human participants in visualization studies. Researchers should exercise caution when considering LLMs as a proxy for human reasoning and should always validate LLM-generated interpretations against human baselines. Future LLMs or more model-intensive approaches may offer more reliable assistance in visualization interpretation, but this remains to be seen.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023), 100 pages.

[2] Robert Amar, James Eagan, and John Stasko. 2005. Low-Level Components of Analytic Activity in Information Visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS '05)*. IEEE Computer Society, USA, 15. doi:10.1109/INFOVIS.2005.24

[3] R. Amar and J. Stasko. 2004. A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations. In *IEEE Symposium on Information Visualization*. IEEE, Austin, TX, USA, 143–150. doi:10.1109/INFVIS.2004.10

[4] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[5] Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. LLMs as research tools: Applications and evaluations in HCI data work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[6] Cindy Xiong Bearfield, Chase Stokes, Andrew Lovett, and Steven Franconeri. 2024. What Does the Chart Say? Grouping Cues Guide Viewer Comparisons and Conclusions in Bar Charts. *IEEE Transactions on Visualization and Computer Graphics* 30, 8 (2024), 5097–5110. doi:10.1109/TVCG.2023.3289292

[7] Alexander Bendeck and John Stasko. 2024. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE VIS* (2024), 1–11. doi:10.1109/TVCG.2024.3456155 to appear.

[8] E. Bertini, M. Correll, and S. Franconeri. 2020. Why Shouldn't All Charts Be Scatter Plots? Beyond Precision-Driven Visualizations. In *2020 IEEE Visualization Conference (VIS)*. IEEE Computer Society, Los Alamitos, CA, USA, 206–210. doi:10.1109/VIS47514.2020.00048

[9] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.

[10] Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2376–2385.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[12] A. Burns, C. Xiong, S. Franconeri, A. Cairo, and N. Mahyar. 2020. How to evaluate data visualizations across different levels of understanding. In *2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization (BELIV)*. IEEE Computer Society, Los Alamitos, CA, USA, 19–28. doi:10.1109/BELIV51497.2020.00010

[13] Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2024. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. *IEEE VIS* (2024), 19 pages. to appear.

[14] Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, Nam Wook Kim, and Jinwook Seo. 2024. Enhancing Data Literacy On-demand: LLMs as Guides for Novices in Chart Interpretation. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–17. doi:10.1109/TVCG.2024.3413195

[15] R Jordon Crouser and Remco Chang. 2012. An affordance-based framework for human computation and human-computer collaboration. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2859–2868.

[16] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (2023), 597–600.

[17] Salomon Eisler and Joachim Meyer. 2025. The Classification of Patterns in Visualizations: Effects of Data Complexity and Individual Differences. *IEEE Transactions on Visualization and Computer Graphics* (2025).

[18] Wan-Cyuan Fan, Yen-Chun Chen, Mengchen Liu, Lu Yuan, and Leonid Sigal. 2024. On Pre-training of Multimodal Language Models Customized for Chart Understanding. arXiv:2407.14506 [cs.CV] https://arxiv.org/abs/2407.14506

[19] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[20] R. Fygenson, S. Franconeri, and E. Bertini. 2024. The Arrangement of Marks Impacts Afforded Messages: Ordering, Partitioning, Spacing, and Coloring in Bar Charts. *IEEE Transactions on Visualization and Computer Graphics* 30, 01 (jan 2024), 1008–1018. doi:10.1109/TVCG.2023.3326590

[21] Aimen Gaba, Vidya Setlur, Arjun Srinivasan, Jane Hoffswell, and Cindy Xiong. 2022. Comparison conundrum and the chamber of visualizations: An exploration of how language influences visual design. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 1211–1221.

[22] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[23] Kanika Goswami, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. 2025. ChartCitor: Multi-Agent Framework for Fine-Grained Chart Visual Attribution. arXiv:2502.00989 [cs.CL] https://arxiv.org/abs/2502.00989

[24] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022), 20 pages.

[25] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[26] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. ChartLlama: A Multimodal LLM for Chart Understanding and Generation. arXiv:2311.16483 [cs.CV] https://arxiv.org/abs/2311.16483

[27] Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. 2024. AI language models cannot replace human research participants. *Ai & Society* 39, 5 (2024), 2603–2605.

[28] Eli Holder and Cindy Xiong. 2022. Dispersion vs Disparity: Hiding Variability Can Encourage Stereotyping When Visualizing Social Outcomes. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 624–634.

[29] Md Naimul Hoque, Darius Coelho, and Klaus Mueller. 2019. Examining the visualization practices of data scientists on Kaggle. In *IEEE VIS*. IEEE, 20–25.

[30] John J Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.

[31] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R. Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024. Do LVLMs Understand Charts? Analyzing and Correcting Factual Errors in Chart Captioning. arXiv:2312.10160 [cs.CL] https://arxiv.org/abs/2312.10160

[32] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486* (2022).

[33] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5092–5103. doi:10.1145/2858036.2858558

[34] Jaeyoung Kim, Jongho Lee, Hong-Jun Choi, Ting-Yao Hsu, Chieh-Yang Huang, Sungchul Kim, Ryan Rossi, Tong Yu, Clyde Lee Giles, Ting-Hao 'Kenneth' Huang, and Sungchul Choi. 2025. Multi-LLM Collaborative Caption Generation in Scientific Documents. arXiv:2501.02552 [cs.CL] https://arxiv.org/abs/2501.02552

[35] Ashley Liew and Klaus Mueller. 2022. Using Large Language Models to Generate Engaging Captions for Data Visualizations. arXiv:2212.14047 [cs.CL] https://arxiv.org/abs/2212.14047

[36] Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. 2014. A survey on information visualization: recent advances and challenges. *The Visual Computer* 30 (2014), 1373–1393.

[37] Leo Yu-Ho Lo and Huamin Qu. 2024. How Good (Or Bad) Are LLMs at Detecting Misleading Visualizations? *arXiv preprint arXiv:2407.17291* (2024), 10 pages.

[38] Laura E. Matzen, Breannan C. Howell, Michael C. S. Trumbo, and Kristin M. Divis. 2023. Numerical and Visual Representations of Uncertainty Lead to Different Patterns of Decision Making. *IEEE Computer Graphics and Applications* 43, 5 (2023), 72–82. doi:10.1109/MCG.2023.3299875

[39] Tamara Munzner. 2014. *Visualization analysis and design*. CRC press.

[40] Arpit Narechania, Arjun Srinivasan, and John Stasko. 2020. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 369–379.

[41] David Navon. 1977. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology* 9, 3 (1977), 353–383.

[42] Lace Padilla, Racquel Fygenson, Spencer C Castro, and Enrico Bertini. 2022. Multiple forecast visualizations (mfvs): Trade-offs in trust and performance in multiple covid-19 forecast visualizations. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 12–22.

[43] Lace M Padilla, Sarah H Creem-Regehr, Mary Hegarty, and Jeanine K Stefanucci. 2018. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications* 3, 1 (2018), 29.

[44] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[45] William Revelle and Maintainer William Revelle. 2015. Package 'psych'. *The comprehensive R archive network* 337 (2015), 338.

[46] Marilyn M Schapira, Ann B Nattinger, and Colleen A McHorney. 2001. Frequency or probability? A qualitative study of risk communication formats used in health care. *Medical Decision Making* 21, 6 (2001), 459–467.

[47] Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2013. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2366–2375.

[48] Priti Shah and Eric G. Freedman. 2011. Bar and Line Graph Comprehension: An Interaction of Top-Down and Bottom-Up Processes. *Topics in Cognitive Science* 3, 3 (2011), 560–578. doi:10.1111/j.1756-8765.2009.01066.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-8765.2009.01066.x

[49] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*. Elsevier, 364–371.

[50] Jonathan Snow and M Mann. 2013. Qualtrics survey software: handbook for research professionals. *Provo, UT: Qualtrics Labs* (2013), 209 pages.

[51] Brian G Southwell, J Scott Babwah Brennen, Ryan Paquin, Vanessa Boudewyns, and Jing Zeng. 2022. Defining and measuring scientific misinformation. *The ANNALS of the American Academy of Political and Social Science* 700, 1 (2022), 98–111.

[52] Chase Stokes, Chelsea Sanker, Bridget Cogley, and Vidya Setlur. 2024. From Delays to Densities: Exploring Data Uncertainty through Speech, Text, and Visualization. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15100. Issue 3. doi:10.1111/cgf.15100

[53] Melanie Tory and Torsten Moller. 2004. Rethinking visualization: A high-level taxonomy. In *IEEE Symposium on Information Visualization*. IEEE, Austin, TX, USA, 151–158.

[54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023), 77 pages.

[55] Emily Wall, Laura Matzen, Mennatallah El-Assady, Peta Masters, Helia Hosseinpour, Alex Endert, Rita Borgo, Polo Chau, Adam Perer, Harald Schupp, et al. 2024. Trust Junk and Evil Knobs: Calibrating Trust in AI Visualization. In *2024 IEEE 17th Pacific Visualization Conference (PacificVis)*. IEEE, 22–31.

[56] Huichen Will Wang, Mitchell Gordon, Leilani Battle, and Jeffrey Heer. 2024. DracoGPT: Extracting Visualization Design Preferences from Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* (2024), 11 pages.

[57] Huichen Will Wang, Jane Hoffswell, Sao Myat Thazin Thane, Victor S Bursztyn, and Cindy Xiong Bearfield. 2024. How Aligned are Human Chart Takeaways and LLM Predictions? A Case Study on Bar Charts with Varying Layouts. *IEEE VIS* (2024), 11 pages. to appear.

[58] Cindy Xiong, Elsie Lee-Robbins, Icy Zhang, Aimen Gaba, and Steven Franconeri. 2024. Reasoning Affordances With Tables and Bar Charts. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (2024), 3487–3502. doi:10.1109/TVCG.2022.3232959

[59] Cindy Xiong, Vidya Setlur, Benjamin Bach, Eunyee Koh, Kylie Lin, and Steven Franconeri. 2021. Visual arrangements of bar charts influence comparisons in viewer takeaways. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 955–965.

[60] Cindy Xiong, Joel Shapiro, Jessica Hullman, and Steven Franconeri. 2019. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 853–862.

[61] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2024. ChartBench: A Benchmark for Complex Visual Reasoning in Charts. arXiv:2312.15915 [cs.CV] https://arxiv.org/abs/2312.15915

[62] Zhongzheng Xu and Emily Wall. 2024. Exploring the Capability of LLMs in Performing Low-Level Visual Analytic Tasks on SVG Data Visualizations. *IEEE VIS* (2024), 5 pages.

[63] F. Yang, M. Cai, C. Mortenson, H. Fakhari, A. D. Lokmanoglu, J. Hullman, S. Franconeri, N. Diakopoulos, E. C. Nisbet, and M. Kay. 2024. Swaying the Public? Impacts of Election Forecast Visualizations on Emotion, Trust, and Intention in the 2022 U.S. Midterms. *IEEE Transactions on Visualization and Computer Graphics* 30, 01 (jan 2024), 23–33. doi:10.1109/TVCG.2023.3327356

[64] Elmira Zohrevandi, Carl AL Westin, Jonas Lundberg, and Anders Ynnerman. 2022. Design and evaluation study of visual analytics decision support tools in air traffic control. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 230–242. Issue 1.