**Can Large Language Models Understand Graphs like Humans? Evaluating GPT-4o as Simulated Participants in Data Visualization Research**

PSYC 6135

Jingyuan Sophie Li

Data visualization plays a key role in storytelling and effective communication. In recent years, Artificial Intelligence (AI) has been applied to various stages of data storytelling and visualization, reducing manual effort (Li et al., 2024). However, earlier AI tools had notable limitations, which is why human researchers often collaborated with AI. One major challenge was the need for researchers to manually design and train models for AI to understand data. The emergence of large language models (LLMs) such as Chatgpt, has helped alleviate this issue. Researchers no longer need to build and train custom models themselves. Instead, they can interact with LLMs through natural language conversations and guide their performance by chatting with LLMs, making the collaboration more accessible and efficient.

One question that arises is: Can LLMs understand graphs? Friendly and Claude (2025) explored this by interacting with the chatbot Claude. They found that Claude could grasp the basic meaning of the graph, however, it struggled with more complex tasks, for instance, recognizing the limitation of the graph initially and recommending how to best visualize relationships between variables. To build on this work and assess whether newer LLMs show improved comprehension, I replicated some of their procedures using GPT-4o by using the same graph and prompt that Friendly and Claude (2025) used – specifically, a graph depicting number of applicants and acceptance rates (see the chat content on <https://chatgpt.com/share/67fbac62-54ac-8008-b97d-3f4e868c2aa9>). Overall, GPT-4o showed a deeper level of comprehension. It not only interpreted the text content but also calculated year-by-year percentage increases and decreases. Furthermore, after being asked about whether the graph broke standard rules for data graphics, it identified misleading aspects and explained them in detail, without being further prompted with hints. It generated a table to summarize these aspects (see Figure 1). In addition, it was also able to provide R code to reproduce most of the key elements from the original graph (see Figure 2A) and to create alternative versions that better represent the data (see Figure 2B & 2C). The R code can be found through the link provided above. It seems that GPT-4o was able to interpret the reverse relationship between number of applicants and acceptance rate and aim to depict their relationship through graphs.

**Figure 1**.

*Output from GPT-4o: Evaluation on UNC Acceptance Graph*

A screenshot of a test

Description automatically generated

*Note*. The table summarizes GPT-4o’s assessment of various visual design aspects of the original graph, highlighting both strengths and potential issues.

**Figure 2**

*UNC Applicants and Acceptance Rate (2019-2024)*

A

A graph of a graph showing the rise and fall of a higher education

Description automatically generated with medium confidence

A graph showing the rate of the application rate

Description automatically generated with medium confidenceB C

A graph with numbers and points

Description automatically generated*Note.*Panel A shows GPT-4o’s reproduction of the original chart using the prompt from Friendly and Claude (2025). Panels B and C display alternative visualizations generated by GPT-4o to better represent the inverse relationship between total applicants and acceptance rate.

How GPT-4o interprets uncertainty in visualizations? I tested it using a classic example: the hurricane forecast cone. I first asked if GPT-4o could identity what the graph is about then asked whether there are misleading aspects. Impressively, GPT-4o identified key limitations of the cone, noting that it can falsely imply a high level of precision in storm path predictions and failing to communicate uncertainty in storm intensity (See the chat content on <https://chatgpt.com/share/6800f8bc-d288-8008-95a5-43dba03afc69>).

These results suggest that LLMs like GPT-4o are not only capable of interpreting visual data but can also critically assess the effectiveness and communication shortcomings of such visuals. They support researchers by generating visualizations, evaluating them, and integrating researchers’ feedback, functioning like knowledgeable collaborators or skilled graduate student assistants. But a question is: can LLMs do more than its role as a research assistant or collaborator – could they also become human participants? If LLMs can reliably mirror how laypeople perceive and interpret graphs, they could reduce the time, cost, and effort required for participant recruitment in data visualization research.

Communicating uncertainty is a persistent challenge in data design, and poor visual representations can lead to misinterpretation and poor decision-making (Padilla et al., 2020). For example, ensemble hurricane forecasts are often misunderstood as deterministic paths, despite being probabilistic by nature. Evaluating whether LLMs can not only interpret but also simulate human reasoning under uncertainty provides a test of their ability to mirror nuanced human cognitive processes. If LLMs can navigate this complexity, it would suggest they are capable of more than surface-level pattern recognition. They may be able to replicate cognitive strategies people use when reasoning about risk and ambiguity.

There is a growing body of research exploring whether LLMs simulate human participants across domains. For instance, studies have shown that LLMs align well with human responses in moral judgment tasks, voting, behaviours in economic games, and heuristic problem-solving (Dillon et al., 2023). In a recent study, Hewitt et al. (2024) tested GPT-4’s ability to simulate representative American respondents in social science experiments. When asked to predict participants' reactions to experimental stimuli, GPT-4’s simulated responses correlated strongly with actual human data (*r* = .85). However, this approach has yet to be extended to the field of data visualization, particularly in evaluating how different graph alternatives influence comprehension and decision-making.

**Present Research**

This study investigates whether GPT-4o can simulate how human participants interpret and respond to uncertainty in hurricane forecast visualizations, and whether its outcomes align with actual human judgements. To generate these simulated responses, we will prompt the LLM using demographic profiles drawn from real participants. Evaluating whether LLMs can serve as substitutes for human participants in this context is valuable for several reasons. It offers a more scalable and efficient way to test design alternatives, provides a cost-effective method for piloting visual materials, and contributes to broader questions about how machines can model human perception – an intersection of cognitive science, data visualization, and AI.

**Method**

**Design**

This study will adopt a mixed factorial design with a 4 (number of simulated ensemble members: 9, 17, 33, and 65) × 2 (collocation: on-line vs. off-line) × 2 (side of distribution: left vs. right) structure. The number of ensemble tracks will be manipulated as a between-participants factor, while collocation and side of distribution will be manipulated within participants. Therefore, each participant will evaluate four forecast maps, one for each combination of the within-subject conditions.

**Participants**

We will recruit 500 adult participants from an online platform (e.g., Prolific or MTurk). Eligibility criteria will include fluency in English and current residence in the United States. Participants will complete the study on their own devices. After completing the task, participants will report demographic information including age, gender, ethnicity, education level, political orientation, geographic region, household income. This demographic information will be used to construct prompts for LLMs, allowing simulation of participants responses.

**Stimuli**

Uncertainty is typically continuous, but visualizations often impose hard boundaries that can distort interpretation. Ensemble displays – where multiple predicted tracks are visualized – offer a more effective representation of uncertainty by showing the distribution of possible storm paths. However, previous research has demonstrated that people tend to overestimate risk when an ensemble track overlaps with a point of interest, a phenomenon known as the collocation effect (Padilla et al., 2020). Because ensemble tracks are randomly sampled from model runs, no single line is a deterministic prediction, yet viewers may treat them as such. To mitigate this bias, Padilla et al. (2020) explored design features, such as increasing the number of tracks, to reduce collocation effects.

We will adapt visualizations from Padilla et al. (2020) to examine how ensemble size and collocation influence damage perceptions in both actual and simulated participants. Stimuli will consist of hurricane forecast maps displaying 9, 17, 33, or 65 straight-line ensemble tracks (see Figure 3). Each map will include a black dot representing the location of an offshore oil platform. For each ensemble size, two versions will be created: one where a track overlaps the oil platform (“on-line”) and one where none do (“off-line”). The side of the distribution (left vs. right) on which the oil rig appears will also be manipulated (see Figure 4). All other visual elements, including map scale and layout, will be held constant across conditions.

**Figure 3**

*Forecast Maps Showing Varying Ensemble Sizes as Stimuli*

A B

A map of the world

Description automatically generated A map of the united states

Description automatically generated

C D

A map with a red line

Description automatically generated with medium confidence A map with red lines

Description automatically generated

*Note*. A: 9-track; B: 17-track; C: 33-track; D: 65-track. The black dot indicates the location of the offshore oil rig and none are collated with a hurricane track (Padilla et al., 2020).

**Figure 4**

*Collocation and Mirroring Conditions in the 33-track Display*

A B

A map with red threads

Description automatically generated A map with red threads

Description automatically generated

C D

A map of the world with red lines

Description automatically generated A map of the world with red lines

Description automatically generated

*Note*. Panel A shows collocation condition, where one ensemble track overlaps with the offshore oil rig; Panel C shows non-collocation condition, where none of the tracks do. Panels B and D are horizontally mirrored versions of A and C, respectively. These stimuli allow us to test whether perceived risk depends on spatial alignment of uncertainty with the point of interest (Padilla et al., 2020).

**Procedure**

Following Padilla et al. (2020), participants will be informed that they will view a series of hurricane forecast maps showing potential storm paths in the Gulf of Mexico. Their task will be to assess: “What is the level of damage that the oil platform will incur?” They will rate it on a 7-point Likert scale ranging from 1 (no damage) to 7 (severe damage). Participants will also rate their confidence in each judgment on a scale from 1 (not at all confident) to 7 (very confident). At the end of survey, they will answer three comprehension questions from Padilla et al. (2020) by choosing “Yes” or “No.”

After completing all trials, participants will be debriefed and compensated. Demographic profiles and trial assignments will then be used to simulate each individual participant using GPT-4o. One simulated response set will be generated for each of the 500 human participants. To increase efficiency, an automated pipeline will be used to construct prompts and batch trials. Each LLM prompt will include demographic information derived from the human participant. LLMs will be asked to respond to two questions per trial, damage rating and confidence, on a 7-point scale. Comprehension questions will be included at the end of the final prompt. The following is a template of prompt adapted from Hewitt et al. 2024:

You are a [Liberal/conservative], [Age], [Race/ethnicity], [Gender], American with [Education level], who identifies as [Party]. The first page of the survey says: [instruction from Padilla et al. 2020]. The next page of the survey gives of four maps, under each map, there are questions: 1) What is the level of damage that the oil platform will occur? 2) Please rate your confidence on in your judgement. Please choose a number from 1 – 7. You choose:

After LLM provides answer, the prompt to test the comprehension will be:

The last page of the survey has question: 1) The display indicates that the forecaster is less certain about the path of the hurricane as time passes. 2) Locations that are more likely to be hit by the storm than locations equidistant from the center of the forecast but not touching a hurricane track. 3) The hurricane forecast shows all possible paths the hurricane could take. Please choose either “Yes” or “No”. You choose:

**Data Analysis Plans**

Mixed-effects ANOVAs will be conducted to examine the effects of ensemble size, collocation, and side of distribution on damage and confidence ratings. Comprehension accuracy, based on three Yes/No questions, will be analyzed using logistic regression to assess whether the forecast design affects understanding. To investigate similarity between LLM and human responses, Pearson correlations will be conducted on the damage and confidence ratings across matched conditions. Additionally, mixed-effects models will be conducted with participant type (human vs. LLM) entered as a fixed effect to test whether the effects of design (e.g., collocation) differ by participant source. Comprehension responses will be compared between human and LLM participants using descriptive statistics and chi-square tests.

**Discussion**

This proposed study aimed to evaluate whether LLMs can simulate human responses to hurricane forecast visualizations and assess uncertainty in ways that align with actual human judgments. The results have the potential to offer support for the potential of LLMs in data visualization research, however, there are several limitations.

One limitation concerns sampling bias. LLMs are trained primarily on internet-based text corpora that overrepresent dominant groups. Previous research also found that LLM-generated texts aligned with traits associated with older, male, and politically liberal individuals (Sourati et al., 2025). As a result, their simulated responses may systematically reflect the cognitive and cultural patterns of these groups and failing to capture the perspectives of those underrepresented groups (Hewitt et al., 2024). Even efforts to prompt LLMs using detailed demographic profiles, such as age, gender, political orientation, and education level, may be insufficient to fully correct these underlying biases.

At the same time, it is important to recognize the unique advantages of LLMs as research participants. Unlike human participants, LLMs do not experience fatigue, boredom, or attention lapses. Tasks such as rating a large set of graphs or answering repetitive comprehension questions can be tedious for human participants, leading to reduced engagement and increased error rates. LLMs, on the other hand, can respond to hundreds of trials rapidly and consistently without fatigue. Emerging methods such as self-refinement techniques allow LLMs to review and revise their own responses, increasing the accuracy and well-reasoned answers through iterative improvement. This could make LLMs especially useful for early-phase piloting of visual materials, pretesting experimental manipulations, and comparing alternative design strategies. By reducing reliance on human participants in the initial stages of research, LLMs can save considerable cost and effort. This is particularly valuable in today’s research landscape, where funding is often constrained, and resources must be used efficiently. By simulating participant responses, researchers can rapidly pilot study materials, test hypotheses, and refine experimental designs before committing to full-scale data collection with human samples.

Reference

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Friendly, M., & Claude, H. (2025). What Does AI Understand About a Graph?, *Nightingale* [online]. https://nightingaledvs.com/what-does-ai-understand-about-a-graph/

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (2024, August 8). Predicting results of social science experiments using large language models. Stanford University and New York University. Retrieved from https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf

Li, H., Wang, Y., Qu, H., Sas, C., Dugas, P. T., Wilson, M. L., Williamson, J. R., Shklovski, I., Kyburz, P., & Mueller, F. F. (2024). Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration. *CHI 2024 - Proceedings of the 2024 CHI Conference on Human Factors in Computing Sytems*, 1–19. https://doi.org/10.1145/3613904.3642726

Padilla, L. M. K., Creem-Regehr, S. H., & Thompson, W. (2020). The Powerful Influence of Marks: Visual and Knowledge-Driven Processing in Hurricane Track Displays. *Journal of Experimental Psychology. Applied*, *26*(1), 1–15. https://doi.org/10.1037/xap0000245

Sourati, Z., Karimi-Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A., Trager, J., Tak, A., Chen, M., Morstatter, F., & Dehghani, M. (2025). *The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models*. https://doi.org/10.48550/arxiv.2502.11266