Evaluating human and machine understanding of data visualizations

Arnav Verma[1]**, Kushin Mukherjee**[2]**, Christopher Potts**[1]**, Elisa Kreiss**[3]**, Judith E. Fan**[1]
Stanford University, Stanford, CA, United States
University of Wisconsin-Madison, Madison, WI, United States
University of California, Los Angeles, CA, United States

Abstract

Although data visualizations are a relatively recent invention, most people are expected to know how to read them. How do current machine learning systems compare with people when performing tasks involving data visualizations? Prior work evaluating machine data visualization understanding has relied upon weak benchmarks that do not resemble the tests used to assess these abilities in humans. We evaluated several state-of-the-art algorithms on data visualization literacy assessments designed for humans, and compared their responses to multiple cohorts of human participants with varying levels of experience with high school-level math. We found that these models systematically underperform all human cohorts and are highly sensitive to small changes in how they are prompted. Among the models we tested, GPT-4V most closely approximates human error patterns, but gaps remain between all models and humans. Our findings highlight the need for stronger benchmarks for data visualization understanding to advance artificial systems towards human-like reasoning about data visualizations.

Keywords: graph literacy; visual reasoning; quantitative reasoning; artificial intelligence; benchmarking

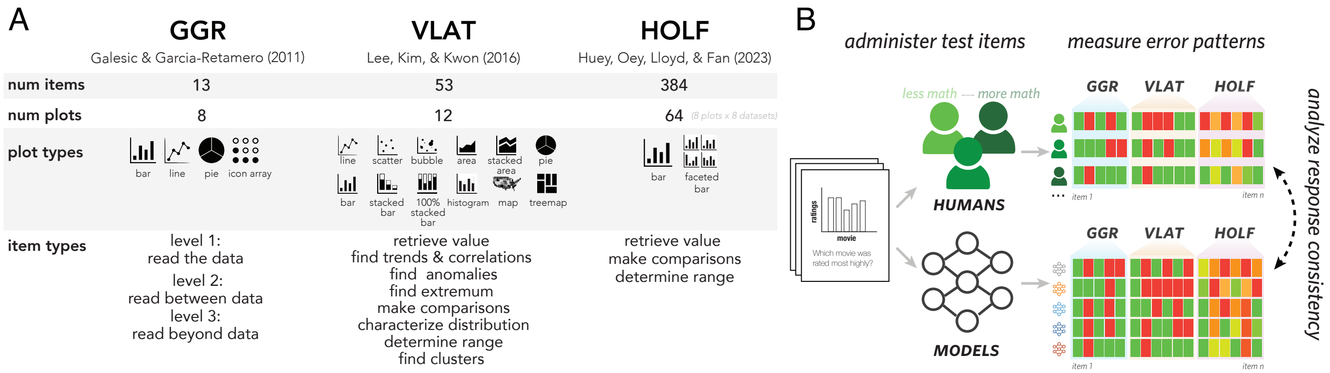
Introduction

Humans can engage with a wide range of visual input modalities, ranging from natural scenes and drawings to diagrams and data visualizations (Tversky, 2011; Franconeri et al., 2021; Fan et al., 2023). Data visualizations — also commonly known as *graphs*, *charts*, and/or *plots* — are especially important because they support reasoning about phenomena that might be too large in scale (or too slow or too uncertain) to be observed directly. They do so by leveraging color, shape, size, position, and other visual variables to encode and convey quantitative patterns and relationships in data (Bertin, 1981; Tufte, 1983; Wilkinson, 2012). As such, they are now indispensable in modern scientific workflows to support exploratory analysis and statistical reasoning (Tukey et al., 1977; Börner et al., 2019; Cumming & Finch, 2005). Moreover, the acquisition of data visualization literacy — a robust ability to parse data visualizations and derive insights from them (Fry, 1981; Curcio, 1987; Friel et al., 2001; Shah & Hoeffner, 2002; Boy et al., 2014; Börner et al., 2019; Firat et al., 2022) — has been a longstanding priority in STEM coursework throughout K-12 and beyond (Pellegrino et al., 2014).

Nevertheless, there are fundamental gaps in current knowledge of what cognitive operations underlie data visualization understanding. In part, these gaps reflect inherent challenges in operationalizing such a complex cognitive construct — the same dataset can be visualized in many different ways and a wide variety of tasks can be performed with any single data visualization (Brehmer & Munzner, 2013; Friel et al., 2001). Even among the most prevalent types of data visualizations (e.g., bar plots, line plots, scatter plots), a person might sometimes want only to search for a single value and other times to derive broader insights about complex trends (Boy et al., 2014; Lee et al., 2016; Kim & Heer, 2018; Börner et al., 2019; Lundgard & Satyanarayan, 2021). The ability to perform any of these tasks is thought to rely on the coordination of several mental processes (Hegarty, 2005), including: rapid perceptual computations (Cleveland & McGill, 1984) with respect to a known graph schema (Pinker, 1990); explicit numerical operations (Gillan & Lewis, 1994) constrained by finite working memory resources (Padilla et al., 2018); and interpretive processes that lead to more general insights (Carpenter & Shah, 1998), which may be influenced by prior content knowledge (Shah & Freedman, 2011).

In principle, computational modeling approaches could provide greater precision concerning the exact operations that support data visualization understanding. Recent advances in artificial intelligence (AI) have yielded a cohort of "multimodal" AI systems that can operate over a combination of visual and linguistic inputs to perform a wide variety of cognitive tasks. The complexity of these tasks has begun to approach that of tasks that humans routinely face in real-world settings, including at school and in the workplace (Chung et al., 2024; S. Zhang et al., 2022; OpenAI, 2023; Liu et al., 2023; Bommasani et al., 2021; Katz et al., 2023; Yue et al., 2023). While strong performance has been reported for some of these systems on data visualization understanding, these reports rely upon relatively weak benchmarks that do not resemble the tests used to assess the same abilities in humans Masry et al. (2022); Lu et al. (2023); OpenAI (2023); Yue et al. (2023). As such, it remains unclear to what degree state-of-the-art vision-language models achieve human-like understanding of data visualizations, for any cohort of humans.

In this paper, we aim to address this gap in three ways: *First*, we identify reliable tests of data visualization understanding that have been used in previous human



	GGR	VLAT	HOLF
	Galesic & Garcia-Retamero (2011)	Lee, Kim, & Kwon (2016)	Huey, Oey, Lloyd, & Fan (2023)
num items	13	53	384
num plots	8	12	64 (8 plots x 8 datasets)
plot types	bar line pie icon array	line scatter bubble area stacked area pie bar stacked bar 100% stacked bar histogram map treemap	bar faceted bar
item types	level 1: read the data level 2: read between data level 3: read beyond data	retrieve value find trends & correlations find anomalies find extremum make comparisons characterize distribution determine range find clusters	retrieve value make comparisons determine range

Figure 1: **(A)** The current experiments include three tests of data visualization understanding, which vary in their length and composition. **(B)** Each of these tests was administered to both human participants and a set of state-of-the-art AI models, enabling quantitative comparison between human and model error patterns.

studies. *Second*, we develop an evaluation protocol to assess data visualization performance in multimodal vision-language models, designed to enable direct comparison to human response patterns. *Third*, we benchmark the performance of several state-of-the-art vision-language models against existing human datasets on the same visualization understanding test suite, building on recent work employing similar approaches (Binder et al., 2023; Bear* et al., 2021; Mukherjee et al., 2023). Taken together, this work presents initial insights and evaluation methods that could be leveraged to make further progress towards computational models that expose the cognitive mechanisms that support real-world quantitative reasoning.

Methods

Progress towards AI systems that achieve human-like understanding of data visualizations requires meeting two key challenges: *first,* establishing common standards by which to assess understanding of data visualizations in humans and AI systems, and *second,* conducting controlled evaluations of human and AI understanding of data visualizations that support direct comparison between humans and models.

Common benchmarks for data visualization understanding

Meeting the first challenge requires identifying valid and reliable measures of data visualization understanding. Towards this end, we leverage prior work developing assessments of data visualization literacy (DelMas et al., 2005; Galesic & Garcia-Retamero, 2011; Lee et al., 2016; Börner et al., 2019; Boy et al., 2014; Ge et al., 2023). These assessments are generally structured in the same way, consisting of a set of test items, each presenting a data visualization and a question about it. Nevertheless, they also vary in how they are organized and what aspects of data visualization understanding they emphasize. Here, we include three tests of data visualization understanding with complementary attributes (Figure 1A).

GGR The first test, which we dub GGR, is a widely used 13-item assessment containing three bar plots, three line plots, as well as an icon array and pie chart (Galesic & Garcia-Retamero, 2011). The test was designed to probe a compact hierarchy of abstract abilities, progressing from "reading the data" to "reading between the data" to "reading beyond the data" (Friel et al., 2001). Nine of the test items require a numerical response and four of them were multiple choice.

VLAT The second test, known as the Visualization Literacy Assessment Test (VLAT), is an influential 53-item assessment containing 12 plots (Lee et al., 2016), each generated from unique real-world data sources: line chart, bar chart, stacked bar chart, normalized stacked bar chart, pie chart, histogram, scatter plot, bubble chart, area chart, stacked area chart, choropleth map, and tree map. VLAT also groups items into a broader suite of more concrete tasks than in GGR, including items that involve: retrieving values, finding extrema, finding anomalies, making comparisons, determining ranges, finding correlations and trends, and finding clusters. All of the test items are multiple choice (34 items with four options; 3 with three options; 16 were True/False).

HOLF The third test, which we dub HOLF, is a 384-item test containing 64 bar plots, consisting of 8 variants generated from each of 8 real-world datasets (Huey et al., 2023). While in VLAT and GGR each plot is paired with an uneven number and variety of types of questions, in HOLF each bar plot variant is paired with all questions pertaining to its corresponding dataset (i.e., retrieve value, make comparisons, determine range). This balanced set of question-plot combinations makes it possible to distinguish the impact of attributes of the plot from the impact of the dataset itself. All of the test items in HOLF require a numerical response.

Measuring data visualization understanding in humans and models

Human participants All behavioral data from human participants included in the current analyses were collected in two recent studies in accordance with the UC San Diego IRB. Data from $1,135$ U.S.-based participants recruited via a combination of Prolific and the UCSD study pool were included in analyses of human performance on GGR and VLAT (Lloyd et al., 2023). Data from 531 U.S.-based participants recruited via Prolific were included in analyses of human performance on HOLF (Huey et al., 2023).

Human evaluation procedure In Lloyd et al. (2023), every participant completed all items in both GGR and VLAT, with test order randomized across participants. In Huey et al. (2023), each participant was presented with eight items from HOLF, such that they answered a single question about one plot generated using each of the eight datasets. In both studies, participants completed a post-study survey wherein they were asked to indicate whether they had taken various high-school math courses: algebra, calculus, and/or statistics. To explore the relationship between the amount of formal math training human participants had received and performance on data visualization understanding tasks, participants were divided into two groups: **less math**, defined as having taken 1 or 2 of the 3 highlighted math courses (GGR: $N = 454$ participants; VLAT: $N = 454$ participants; HOLF: $N = 284$ participants); and **more math**, defined as having taken all 3 highlighted math courses (GGR: $N = 632$ participants; VLAT: $N = 632$ participants; HOLF: $N = 164$ participants).

Model suite To determine which models to include in our evaluation, we prioritized those that have been reported to achieve strong performance on other benchmarks that involve reasoning over visual and linguistic inputs (Li et al., 2023; Yue et al., 2023). Here, we include four vision-language models that vary along several dimensions (i.e., architecture, size, training objective, training data): **LLaVA-1.5-Vicuna-7b** (Zheng et al., 2024), **BLIP-2-FLAN-T5-XL** (Chung et al., 2024), **BLIP-2-FLAN-T5-XXL** (Chung et al., 2024), and **GPT-4V**[1](OpenAI, 2023)

Model evaluation procedure Each model was evaluated on all 450 test items from across GGR, VLAT, and HOLF. For each test item, the input to models consisted of two components: an image containing a data visualization and a corresponding question about the visualization supplied as a text prompt written in English. We recorded the full text response produced by each model and applied post-processing to extract the most relevant information.

Assessing impact of prompt. To quantify the degree to which model behavior was sensitive to the form of the text prompt, we tested all models on two variants of the prompt: the *raw* prompt contained the exact task instructions and question text provided to human participants; the *adapted* prompt was modified to more closely align with the format of the prompts provided to each model during its training (e.g., prepending the word *Question:* before each question). We thus evaluated all four models on 900 test items, 450 using the raw prompt and 450 using the adapted prompt. To improve the robustness of our findings, we presented every item 10 times to each model, yielding a total of 9000 responses per model. We sampled outputs from each model using nucleus sampling (temperature = 0.1; top-p = 0.4), a commonly used technique for improving the diversity and fluency of language model outputs (Holtzman et al., 2019; Gunjal et al., 2024). We report findings based on specific temperature and top-p values, but explored a wide range of values for these parameters to identify ones that were associated with higher model performance, and thus a stronger basis for comparison with human behavior.

Postprocessing model output. Several models produced verbose responses that did not conform to any of the required response formats (i.e., multiple choice, True/False, numerical response). In particular, **LLaVA-1.5-Vicuna-7b** often returned the full prompt as part of its response. As such, we applied further processing to excise the prompt from any responses that included them. Specifically, following prior work (Yue et al., 2023), we used GPT-4[2] to extract only the relevant information from the raw model output. A subset of these post-processed responses were then reviewed by a member of the research team to verify their validity.

Results

How often do models produce *non-empty* responses to questions? A minimum bar for any model to clear on these tasks is that it produces non-empty responses to all questions. We found that when using the raw prompt, **LLaVA-1.5-Vicuna-7b** failed to reliably produce non-empty strings in response (Figure 2A) and produced the lowest proportion of non-empty responses for raw prompts (24.4%). Using the adapted prompts produced substantial improvements in the rate of non-empty strings returned across models (adapted = 99.83%; raw = 80.28%). These findings suggest that, as capable as some of these systems might be on various tasks, their ability to produce outputs at all can be highly sensitive to relatively minor changes in the format of the prompt.

How much does prompting strategy impact *what* models say? When models did generate non-empty responses, to what degree was the text generated under the raw and adapted prompts different? To evaluate this question, we computed the Jaccard similarity index between responses to the same test item between different model pairs and prompt types. This index measures the ratio between the number of overlapping words in two responses and the total number of words appearing in both responses. If two responses were

[1] Evaluation done through Azure OpenAI services using model GPT-4V version `vision-preview` from April-May 2024.

[2] Evaluation done through Azure OpenAI services using model GPT-4 version `1106-preview` from April-May 2024.

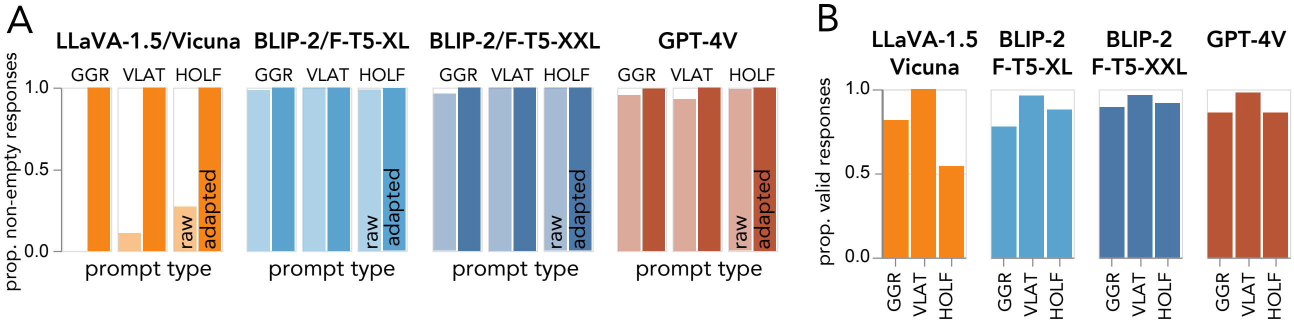


Figure 2: **(A)** Proportion of non-empty responses produced by each model on each test using both raw and adapted prompts. **(B)** Proportion of valid responses generated by each model on each test using only adapted prompts.

exactly the same, they would have a Jaccard similarity index of 1, and if they shared none of the same words, the Jaccard similarity would be 0. We computed the mean similarity between prompting method pairs by averaging similarity values between responses across all test items that produced non-empty outputs.

We found that across all three tests, the text generated by the same model differed under different prompting methods as measured using mean Jaccard similarities (GGR = 0.36, 95% CI = [0.20, 0.50]; VLAT = 0.76, 95% CI = [0.72, 0.80]; HOLF = 0.20, 95% CI = [0.20, 0.21]). These findings provide converging evidence that subtle variations in prompt formatting can systematically impact model outputs, even when only considering those cases where any text was generated at all.

How often do models produce *valid* responses? Having established that using the adapted prompts more reliably yielded non-empty responses, we sought to evaluate how often these outputs contained information that was actually relevant to answering the question.

Our first step was to determine whether the model responses to test items were in the 'valid' format required by the question. For multiple-choice questions in GGR and VLAT, a response was considered to be valid if the processed response was an exact match to one of the multiple choice options. For numerical-response questions in GGR and HOLF, a response was considered to be valid if it was possible to extract a numerical value from the response string. If multiple numerical values were given (e.g. "3.1 or 4.1"), the one closest to the correct answer was extracted.

Using these criteria, we computed the proportion of valid responses to each test item for each model (Figure 2B). Overall, we found that models did not always provide valid responses to numerical questions, with **LLaVA-1.5/Vicuna** producing the lowest proportion of valid responses (60.44%), followed by **GPT-4V** (85.76%), **BLIP-2/Flan-T5-XL** (88.40%), and **BLIP-2/Flan-T5-XXL**, which produced the highest proportion of valid responses (92.09%) (Figure 2B). These results demonstrate that even when models produce non-empty responses, the text generated may not contain a valid response to the question.

How often do models produce *accurate* answers? Having implemented a procedure for identifying only the valid responses produced by all four models (i.e., **LLaVA-1.5/Vicuna**, **BLIP-2/F-T5-XL**, **BLIP-2/F-T5-XXL**, **GPT-4V**), we next sought to compare their accuracy to that of two groups of human participants who differed in their prior experience with high school-level mathematics—**Humans (less math)**, **Humans (more math)**.

For GGR and VLAT, we measured human and model performance by computing the proportion of correct responses. Doing so was straightforward for the multiple choice items; for the items that required numerical responses, responses were only deemed correct if they *exactly* matched the true answer provided by the original test designers.

For HOLF, all of the items required a numerical response. We measured performance by computing the median normalized absolute deviation between human/model responses and the correct value. Specifically given an agent-type (model/human), for each item we obtained raw errors by taking the absolute difference between each response and the correct value. However, because plots in HOLF were generated from datasets containing variables measured in different units with highly disparate scales (i.e., some in the range 10^1-10^2 and others in the range 10^4-10^5), it is not clear based on these raw error values how accurate any given response was in the context of the variable it pertains to. As such, we obtained *normalized* errors by dividing the absolute value of raw errors by the interval spanned by the y-axis in the corresponding plot, providing a relative measure of how far off a response was from the correct value compared to how wide a range of values was observed for that variable.

Overall, we found that **GPT-4V** performed best among the four models, achieving higher accuracy on GGR (**GPT-4V** = 0.34, 95% CI = [0.18, 0.52]; **Other Models** = 0.05, 95% CI = [0.01, 0.07]) and VLAT (**GPT-4V** = 0.62, 95% CI = [0.55, 0.75]; **Other Models** = 0.32, 95% CI = [0.25, 0.39]), while achieving lower median error on HOLF (**GPT-4V** = 0.09, 95% CI = [0.08, 0.10]; **Other Models** = 0.3, 95% CI = [0.3, 0.33]). Nevertheless, **GPT-4V** did not perform as well as **Humans (less math)** on any of the tests (ΔGGR = 0.44, 95% CI = [0.26, 0.60]; ΔVLAT =

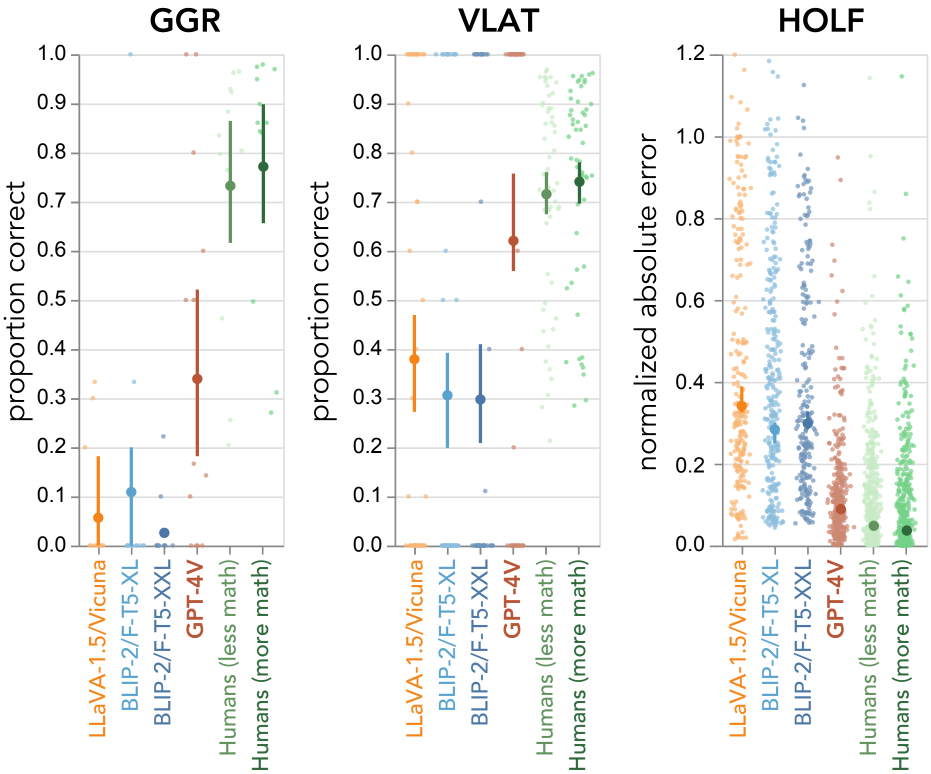


Figure 3: Human and model performance on each test. Each dot represents either the mean proportion correct (GGR & VLAT) or median normalized absolute error (VLAT) for a single test item. Error bars represent bootstrapped 95% confidence intervals.

0.13, 95% CI = [0.08, 0.27]; ΔHOLF = -0.04, 95% CI = [−0.06, −0.03]). These findings suggest that there remains a meaningful performance gap between current state-of-the-art vision-language models and humans, even when only considering individuals with relatively modest amounts of prior experience with high school-level math.

How similar were responses generated by models and humans? While there is only one way to answer all of the questions correctly, there are many possible ways to answer them incorrectly. As such, examining these response patterns can reveal correspondences between model and human behavior that might not be revealed by analyses of task performance alone. Towards this end, we computed similarities between the response patterns generated by different models and humans, for each pair of models (or human-model pair). For the multiple-choice items (i.e., in GGR and VLAT), we computed the proportion of matching responses. For items requiring a numerical response (i.e., in HOLF), we used normalized absolute error values, which provide a measure of the magnitude of deviations between responses.

Overall, we found that consistency between all models (taken together) and all humans (from both groups) was relatively low for all three tests: GGR (0.18; 95% CI = [0.10, 0.26]), VLAT (0.44, 95% CI = [0.39, 0.49]), and HOLF (0.24, 95% CI = [0.22, 0.25]), although GPT-4V produced responses that were most similar to those of humans among the four models (Fig. 4). By contrast, response consistency was relatively high between the two groups of human participants, **Humans (less math)** and **Humans (more math)**: GGR = 0.68, 95% CI = [0.61, 0.76]; VLAT = 0.68, 95% CI = [0.66, 0.71]; HOLF = 0.05, 95% CI = [0.04, 0.05], and higher than the consistency between different vision-language models (GGR = 0.17, 95% CI = [0.07, 0.25]; VLAT = 0.50, 95% CI = [0.46, 0.56]; HOLF = 0.33, 95% CI = [0.31, 0.39]). Taken together, these results suggest that there is a high degree of systematicity in human response patterns on these data visualization tasks, but none of the four models could reproduce these patterns.

Discussion

Recently developed vision-language models have been claimed to show competence at a wide variety of visual tasks, including data visualization understanding (Lu et al., 2023; OpenAI, 2023). However, existing results rely on weak benchmarks and and do not use the materials and standards by which human data visualization literacy is assessed (Galesic & Garcia-Retamero, 2011; Lee et al., 2016; Börner et al., 2019; Boy et al., 2014). We conducted controlled evaluations of human and model performance on three visualization literacy assessments previously used in human behavioral studies. Using these assessments, we evaluated a set of

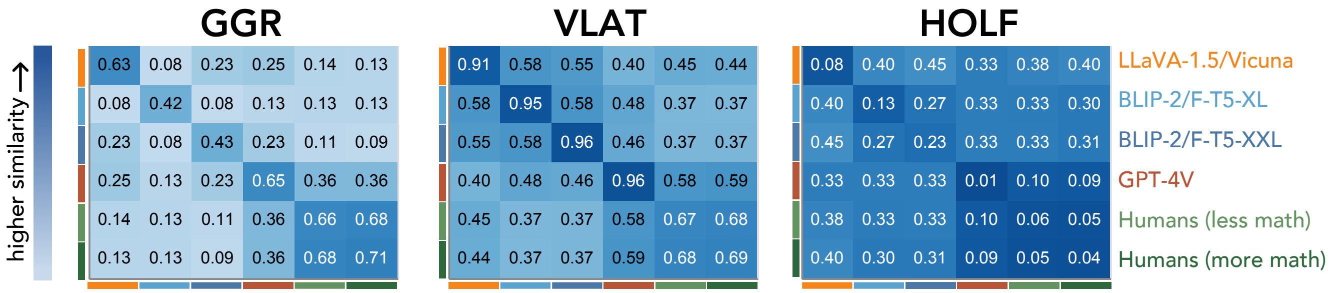


Figure 4: Pairwise comparison of interpretable responses between all vision-language models and humans. For GGR and VLAT, we compared responses using Jaccard similarity, with higher values indicating greater similarity (and identical responses achieving a value of 1). For HOLF, we computed differences between responses by computing the median normalized absolute error, with lower values indicating greater similarity (and identical responses achieving a value of 0).

four state-of-the-art vision-language models and compared their performance to that of human participants who varied in their amount of formal math training. We found that current models are sensitive to minor variations in the format of the prompt and often produce invalid outputs that are not responsive to the question at hand. Even when only considering valid responses from models, we found that they performed reliably worse than human participants including those who had lesser formal math training. Nevertheless, we found that the highest-performing model, **GPT-4V**, also currently produces responses that are most consistent with those made by humans.

Our paper contributes to a growing body of cognitive-AI benchmarking efforts that employ large-scale behavioral experimentation to rigorously evaluate both humans and AI systems on a common set of controlled tasks on relatively naturalistic stimuli (Binder et al., 2023; Bear* et al., 2021; Martinez et al., 2023; Mukherjee et al., 2019). These efforts not only advance human-AI alignment by identifying specific gaps in current AI systems but also generate fruitful hypotheses concerning the viability of current AI models as computational cognitive models of the human mind. We found that vision-language models that have achieved strong performance on other visual tasks (e.g., object recognition, image captioning) still fall short on tasks involving interpreting data visualizations, both in terms of performance and consistency with humans (Radford et al., 2021; Xie et al., 2021; J. Zhang et al., 2023).

Key outstanding questions concern where these gaps come from and how to close them. Data visualization literacy is acquired by humans through formal education and training. While modern vision-language models are trained on very large datasets that likely include data visualizations, they generally do not engage with these inputs or receive social feedback in the ways that human learners do (Gweon et al., 2023). An important future direction will thus be to uncover the aspects of human learning environments that are critical for observing robust acquisition of these skills in humans, and explore to what degree these insights can be leveraged to develop more robust and sample-efficient artificial learning systems.

More broadly, we envision the use of stimuli and tasks that approach the complexity of natural behavior in real-world environments being crucial for advancing theories of human perception, learning, and reasoning. Moreover, developing AI systems that display more human-like understanding of abstract visual inputs could be used to design both more effective STEM learning environments and visualizations for scientific communication.

Acknowledgments

We thank members of the Cognitive Tools Lab for their input and feedback. This work was supported by NSF CAREER #2047191, a seed grant from the School of Social Sciences at UC San Diego, and a seed grant from the Stanford Human-Centered AI Institute and Stanford Accelerator for Learning.

References

Bear*, D., Wang*, E., Mrowca*, D., Binder*, F., Tung, H.-Y., RT, P., … Fan**, J. (2021). Physion: Evaluating physical prediction from vision in humans and machines. *Advances in Neural Information Processing Systems*.

Bertin, J. (1981). *Graphics and graphic information processing*. Walter de Gruyter.

Binder, F. J., Cross, L. M., Friedman, Y., Hawkins, R., Yamins, D. L., & Fan, J. E. (2023). Advancing Cognitive Science and AI with Cognitive-AI Benchmarking. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, *116*(6), 1857–1864.

Boy, J., Rensink, R. A., Bertini, E., & Fekete, J.-D. (2014). A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics*, *20*(12), 1963–1972.

Brehmer, M., & Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, *19*(12), 2376–2385.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*(2), 75.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... others (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, *79*(387), 531–554.

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American psychologist*, *60*(2), 170.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for research in mathematics education*, *18*(5), 382–393.

DelMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In *Fourth forum on statistical reasoning, thinking, and literacy (srtl-4).*

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, *2*(9), 556–568.

Firat, E. E., Joshi, A., & Laramee, R. S. (2022). Interactive visualization literacy: The state-of-the-art. *Information Visualization*, *21*(3), 285–310.

Franconeri, S. L., Padilla, L., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, *22*(3), 110–161.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in mathematics Education*, *32*(2), 124–158.

Fry, E. (1981). Graphical literacy. *Journal of Reading*, *24*(5), 383–389.

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical decision making*, *31*(3), 444–457.

Ge, L. W., Cui, Y., & Kay, M. (2023). CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–18).

Gillan, D. J., & Lewis, R. (1994). A componential model of human interaction with graphs: 1. Linear regression modeling. *Human Factors*, *36*(3), 419–440.

Gunjal, A., Yin, J., & Bas, E. (2024). Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, pp. 18135–18143).

Gweon, H., Fan, J., & Kim, B. (2023). Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220048.

Hegarty, M. (2005). Multimedia learning about physical systems. *The Cambridge handbook of multimedia learning*, 447–465.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations.*

Huey, H., Oey, L. A., Lloyd, H., & Fan, J. E. (2023). How do communicative goals guide which data visualizations people think are effective? In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). Gpt-4 passes the bar exam. *Available at SSRN 4389233.*

Kim, Y., & Heer, J. (2018). Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum* (Vol. 37, pp. 157–167).

Lee, S., Kim, S.-H., & Kwon, B. C. (2016). Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, *23*(1), 551–560.

Li, B., Wang, R., Wang, G., Ge, Y. G. Y., & Shan, Y. (2023). SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125.*

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). Improved Baselines with Visual Instruction Tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following.*

Lloyd, H., Huey, H., Brockbank, E., Padilla, L., & Fan, J. E. (2023). What is graph comprehension and how do you measure it? In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., ... Gao, J. (2023). MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *The Twelfth International Conference on Learning Representations.*

Lundgard, A., & Satyanarayan, A. (2021). Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, *28*(1), 1073–1083.

Martinez, J., Binder, F. J., Wang, H., Haber, N., Fan, J. E., & Yamins, D. (2023). Measuring and Modeling Physical Intrinsic Motivation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).

Masry, A., Do, X. L., Tan, J. Q., Joty, S., & Hoque, E. (2022). ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 2263–2279).

Mukherjee, K., Hawkins, R., & Fan, J. (2019). Conveying semantic part information in drawings. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Mukherjee, K., Huey, H., Lu, X., Vinker, Y., Aguina-Kang, R., Shamir, A., & Fan, J. E. (2023). SEVA: Leveraging sketches to evaluate alignment between human and machine visual abstraction. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

OpenAI. (2023). *GPT-4 Technical Report.*

Padilla, L., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications*, *3*(1), 1–25.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). Developing assessments for the next generation science standards. *National Academies Press*.

Pinker, S. (1990). A theory of graph comprehension. *Artificial intelligence and the future of testing*, 73–126.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . others (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763).

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, *3*(3), 560–578.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational psychology review*, *14*(1), 47–69.

Tufte, E. R. (1983). *The visual display of quantitative information* (Vol. 2). Graphics press Cheshire, CT.

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499–535.

Wilkinson, L. (2012). *The grammar of graphics*. Springer.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., . . . others (2023). MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv e-prints*, arXiv–2311.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2023). Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., . . . others (2022). OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., . . . others (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*.