

CHART-6: Human-Centered Evaluation of Data Visualization Understanding in Vision-Language Models

Arnav Verma, Kushin Mukherjee, Christopher Potts, Elisa Kreiss, and Judith E. Fan

Abstract— Data visualizations are powerful tools for communicating patterns in quantitative data. Yet understanding any data visualization is no small feat — succeeding requires jointly making sense of visual, numerical, and linguistic inputs arranged in a conventionalized format one has previously learned to parse. Recently developed vision-language models are, in principle, promising candidates for developing computational models of these cognitive operations. However, it is currently unclear to what degree these models emulate human behavior on tasks that involve reasoning about data visualizations. This gap reflects limitations in prior work that has evaluated data visualization understanding in artificial systems using measures that differ from those typically used to assess these abilities in humans. Here we evaluated eight vision-language models on six data visualization literacy assessments designed for humans and compared model responses to those of human participants. We found that these models performed worse than human participants on average, and this performance gap persisted even when using relatively lenient criteria to assess model performance. Moreover, while relative performance across items was somewhat correlated between models and humans, all models produced patterns of errors that were reliably distinct from those produced by human participants. Taken together, these findings suggest significant opportunities for further development of artificial systems that might serve as useful models of how humans reason about data visualizations. All code and data needed to reproduce these results are available at: https://osf.io/e25mu/?view_only=399daff5a14d4b16b09473cf19043f18.

Index Terms—chart understanding, graph comprehension, artificial intelligence, visualization literacy, cognitive-AI benchmarking

1 INTRODUCTION

Humans can engage with a wide range of visual input modalities, ranging from natural scenes and drawings to diagrams and data visualizations [22, 26, 79]. Data visualizations — also commonly known as *graphs*, *charts*, and/or *plots* — are indispensable tools for supporting exploratory analysis and statistical reasoning [8, 20, 78]. They do so by leveraging a combination of visual features (e.g., color, shape, size, position) and text-based annotations (e.g., axis labels, legends) to efficiently convey patterns in quantitative data [5, 61, 77, 81]. As such, interpreting any data visualization relies upon the ability to correctly combine visual, linguistic, and quantitative information to answer some question at hand (Figure 1). Moreover, the acquisition of data visualization literacy — a robust ability to parse data visualizations and derive insights from them [8, 10, 21, 24, 27, 28, 73] — has been a longstanding priority in STEM education [16].

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 to facilitate understanding of different quantitative phenomena (e.g., a person might sometimes want only to search for a single value and other times to derive broader insights about complex trends) [3, 11, 27, 66]. The primary strategy for enhancing understanding of data using visualization is to encode the underlying data using different visual channels (size, shape, color, etc.) in order to produce different types of data visualizations (bar plots, line plots, scatter plots, etc.) [8, 10, 43, 46, 54]. The ability to perform visualization understanding tasks is thought to rely on the coordination of several mental processes [36], including: rapid perceptual computations

[15] with respect to a known graph schema [65]; explicit numerical operations [33] constrained by finite working memory resources [63]; and interpretive processes that lead to more general insights [13], which may be influenced by prior content knowledge [71].

Classical accounts of these processes are limited in that they either are not specified in computationally explicit terms or are derived based on a limited variety of data visualizations, thus limiting their generalizability [25, 65, 72, 75]. To more precisely describe the operations that enable visualization understanding, as well as developmental changes accompanying the acquisition of data visualization literacy, there is a need for computational models that can contend with the diversity of real-world visualizations and are adaptable to common visualization understanding tasks. Recently developed “multimodal” AI systems are promising candidate models because they can operate over a combination of visual and textual inputs to perform a wide variety of tasks that require integrating information from both these channels [1, 44, 62]. The complexity of tasks that these systems have been reported to perform well has begun to approach that of tasks that humans routinely face in real-world settings, including at school and in the workplace [6, 14, 42, 51, 62, 85, 87]. This progress has fueled the promise that such ‘vision-language models’ could serve as a robust foundation for developing scientific models of human reasoning over multiple information modalities.

However, for such AI systems to provide a useful basis for developing cognitive models of human visualization understanding, it is critical to evaluate to what degree they generate patterns of behavior on data visualization understanding tasks that approximate those generated by humans. While strong performance has been reported for some of these systems on data visualization understanding tasks, these reports rely upon different measures from those typically used to assess the same abilities in humans and generally do not directly compare model behavior to that of humans [53, 56, 57, 62, 80, 82, 85]. As such, it remains unclear to what degree any current systems approach human-level abilities or engage in human-like reasoning about data visualizations, thereby limiting any insights that can be drawn about the operations involved in human visualization understanding from such models.

Our paper addresses this gap in three ways: *First*, we present CHART-6 (*Comparative Human-AI Graphical Reasoning Tests*), a human-centered suite of data visualization understanding assessments from the psychology and visualization literatures. *Second*, we develop an evaluation protocol to rigorously assess the performance of vision-

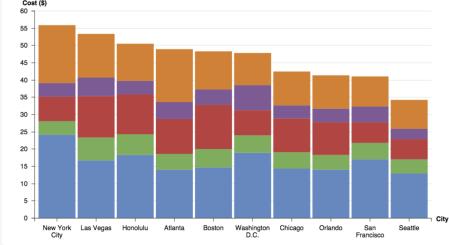
• Arnav Verma, Kushin Mukherjee, and Judith E. Fan are with the Department of Psychology at Stanford University.
E-mail: {arnavv | kushinm | jefan }@stanford.edu
• Christopher Potts is with the Department of Linguistics at Stanford University. E-mail: cgpotts@stanford.edu.
• Elisa Kreiss is with the Department of Communication at University of California, Los Angeles. E-mail: ekreiss@ucla.edu.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

About what is the ratio of the cost of a sandwich to the total cost of room service in Seattle?

1 to 10 2 to 10 **4 to 10** 6 to 10

Hotel Costs of Room Service



1 to 10



1 to 10



USER: (...) ASSISTANT: **2 to 10**.

USER: (...) ASSISTANT: The total cost of room service in Seattle is \$10, and the cost of a sandwich is \$6. So, the ratio of the cost of a sandwich to the total cost of room service in Seattle is **6 to 10**.



2 to 10



63



2.71875



The cost of a sandwich in (...) Therefore, the correct answer is **2 to 10**.



4 to 10



Fig. 1: Sample response from all evaluated models for a multiple-choice item. Responses after processing are shown in bold and are used for comparison against human and model responses. Responses without bold characters indicate invalid responses.

language models on question-answering tasks germane to visualization understanding, designed to enable direct comparison to human behavior. *Third*, we use this protocol to evaluate the performance of several state-of-the-art vision-language models against human behavior on CHART-6, with respect to both how well these models perform and how well they emulate *human-like* behavior on these tests. We found that many of these models often failed to produce valid responses when administered these tests. Even when focusing on items for which models did produce valid responses, we found that they still achieved reliably lower performance than did the adult human participants represented in this work. Direct comparison of human and model performance revealed that humans generally outperformed models, and that the items which humans found difficult were not necessarily those on which models also displayed worse performance, though there were some categories of items where human and model performance was comparable. Nevertheless, we found that no model produces patterns of responses that approach the human noise ceiling, suggesting that further innovation is needed to develop models that can form the basis of cognitive models of human visualization understanding.

2 METHOD

Progress towards computational models that emulate human understanding of data visualizations requires meeting two key methodological challenges: (1) establishing common standards by which to assess understanding of data visualizations in humans and AI systems, and (2) conducting controlled evaluations of human and AI understanding of data visualizations that support direct comparison between these two systems. This effort follows in the tradition of recent human and AI benchmarking work in the cognitive sciences [4, 7, 23, 32, 55, 58, 60].

2.1 Test suite

Leveraging prior work on developing tests of data visualization literacy in the psychology and visualization literature [30, 31, 40, 41, 46] we developed a diverse test suite that provides broad coverage of the skills that are considered to be important when assessing data visualization literacy in humans (Figure 2).

All of these tests consist of a series of test items, each presenting an image of a data visualization paired with a question posed in natural

language. Three tests consist primarily of multiple-choice questions, requiring a response that matches one of several provided options. The remaining three tests consist of questions requiring a numerical response. Since many tests had multiple questions paired with a given visualization, we refer to each unique visualization-question pair as a test ‘item’ in each of the tests. Below, we provided a brief description of each of the six tests included in CHART-6.

GGR GGR is a 13-item test containing three bar plots, three line plots, an icon array, and a pie chart [30]. 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” [27]. Nine of the test items require a numerical response and four of them were multiple choice. While the answers for several items are numeric, since the designers of the test assessed performance by computing the proportion of responses that were exact matches to the true answer, we also treated this test as one whose answers were ‘multiple-choice’.

VLAT The Visualization Literacy Assessment Test (VLAT) is a 53-item test containing 12 graph types [46] — 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 — each generated using data obtained from news articles. VLAT groups items into more concrete tasks than in GGR, including questions that involve: retrieving values, finding extrema, finding anomalies, making comparisons, determining ranges, finding correlations & 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).

CALVI The Critical Thinking Assessment for Literacy in Visualizations (CALVI) is a 60-item test focusing on the ability to compensate for misleadingly constructed data visualizations, such as the use of inappropriate scale ranges or unconventional scale directions [31]. It is composed of 45 items which feature such misleading visualizations, enabling direct comparison between human and model behavior in cases where many humans are expected to fail. All of the test items in CALVI require multiple-choice responses.

HOLF HOLF is a 384-item test containing 64 bar plots procedurally generated from eight real-world datasets. Each chart was paired with six different questions measuring the ability to retrieve values, make comparisons, and determine ranges, yielding 48 unique questions in

	GGR Galesic & Garcia-Retamero (2011)	VLAT Lee, Kim, & Kwon (2016)	CALVI Ge, Cui, & Kay (2023)	HOLF Huey, Oey, Lloyd, & Fan (2023)	HOLF-Multi Huey (2024)	ChartQA-Human Masry et al. (2022)
num items	13	53	60	384	216	125
num plots	8	12	60 (45 misleading + 15 standard)	64 (8 plots x 8 datasets)	72 (9 plots x 8 datasets)	125
plot types						
item types	read the data read between data read beyond data	retrieve value find trends & correlations find anomalies find extremum make comparisons characterize distribution determine range find clusters	retrieve value find trends & correlations make comparisons make predictions aggregate values	retrieve value make comparisons determine range	retrieve value make comparisons predict trend	undefined
response type	multiple choice	multiple choice	multiple choice	numerical	numerical	numerical

Fig. 2: We present **CHART-6** (*Comparative Human-AI Graphical Reasoning Tests*), a human-centered suite of data visualization understanding benchmarks, to assess how close state-of-the-art vision-language models are to achieving both *human-level* performance and *human-like* behavior on reasoning tasks involving data visualizations. This test suite spans a wide array of different approaches to designing such assessments, ensuring broad coverage of the skills that are considered to be important when assessing human data visualization literacy.

total. While in VLAT and GGR each plot is paired with an uneven number and variety of types of questions, in HOLF each plot was paired with all six question types, making it possible to disentangle the impact of various plot attributes from properties of the underlying dataset. This test was originally used in controlled laboratory settings to characterize human judgments concerning which of several plots would be most informative to other people for answering a particular question [41].

HOLF-Multi HOLF-Multi is a 216-item extension of HOLF containing 72 bar, line, and scatter plots [40]. What distinguishes HOLF-Multi from HOLF is a larger variety of graph types. These plots were generated from the same eight datasets as in HOLF, and each plot was paired with 3 questions, yielding a total of 24 unique questions.

ChartQA-Human ChartQA [56] is a data visualization understanding benchmark containing plots obtained from various web sources such as Statista and Pew Research. An initial set of questions about them was generated by a combination of human participants and language models, which was then refined by the benchmark developers. Vision-language models are routinely evaluated on the test split of this benchmark, which consists of 2,490 questions pertaining to 1,509 plots. Here we consider only the set of items in ChartQA that require numerical responses. We constructed ChartQA-Human by sampling a random subset of 125 items from the ChartQA test set such that different types of graphs, data sources, and question styles i.e., human-written vs. template-based) appeared in roughly equal proportion to their relative frequency in the full set of ChartQA items.

2.2 Task categories

Because these six tests were developed independently of one another, they used ways of organizing items into task categories that were not commensurate with one another (e.g., “find trends & correlations” and “read beyond the data”). To conduct analyses that spanned these different tests, we defined a common set of task categories that could be applied to all tests: *value identification*, where participants retrieve an individual value appearing in a plot (e.g., finding the maximum value); *arithmetic computation*, where participants are expected to perform simple arithmetic operations over multiple values displayed in the plot (e.g., finding the average of two values); and *statistical inference*, where participants must estimate latent parameters in a statistical model based on the values shown (e.g., judge the strength of trends or presence of clusters). The only exception was ChartQA-Human which did not initially specify any task categories to organize the test items it contains.

2.3 Measuring data visualization understanding in humans

Where available, we leveraged existing human behavioral data, and where necessary, collected new data by conducting studies with human participants.

GGR and VLAT Data were collected in a previous study with 1,135 human participants recruited via Prolific [52]. Each participant was asked to complete both of these tests in a single session with test order randomized across participants.

CALVI Data were collected in a previous study with 497 participants ¹ [31]. Participants were recruited via Prolific and given a 30-item test: 15 were randomly sampled from the set of 45 misleading items, while the other 15 were always the same set of non-misleading items.

HOLF and HOLF-Multi Data were collected in a previous study with 531 participants on HOLF ² and 1,743 participants on HOLF-Multi [40, 41]. In both studies, each participant was presented with eight items drawn from the full set of test items, such that they only saw one plot and question pertaining to each of the eight datasets.

ChartQA-Human We recruited 50 participants via Prolific in the present study to complete ChartQA-Human, a 125-item representative subset of the ChartQA benchmark. Each participant completed a set of 25 items sampled at random from the full set of 125 items. Participants provided informed consent and were compensated for their time (\$15.50 per hour). All study procedures were carried out in accordance with the cognizant university IRB.

2.4 Measuring data visualization understanding in models

Model suite To determine which vision-language models to include in our evaluation, we prioritized those that achieved strong performance on other benchmarks that involve reasoning over visual and linguistic inputs [47, 85]. In addition, to improve the robustness of our findings, we sought to include a suite of models that was reasonably diverse and representative of current modeling approaches with respect to architecture, size, and pre-training protocol. We selected eight models in total, which included three pairs of models that shared similar architectures and training regimes. **Blip2-FlanT5-4B** and **Blip2-FlanT5-11B** used the 4B-parameter (FlanT5-XL) and 11B-parameter (FlanT5-XXL) versions of the FlanT5 language model respectively [14]. Both models used the same BLIP-2 pre-training regimen [48]. Similarly, **LLaVA1.5-Vicuna-7B** and **LLaVA1.5-Vicuna-13B** used the same CLIP-ViT-L-336px vision encoder and the 7B-parameter and 13B-parameter version of the Vicuna language model respectively. Both models were trained using the LLaVA-1.5 framework [51, 67]. **MatChar-0.3B** [50] augments the pre-training of **Pix2Struct-0.3B** [45] with additional tasks intended to enhance its general visual and quantitative reasoning performance. We also included **LLaVA1.6-Yi-34B** [51], which uses the 34B-parameter version of the Yi language model [84]

¹ Downloaded on May 2024 at: <https://osf.io/pv67z>.

² Downloaded on January 2024 at:

https://github.com/cogtoolslab/davinci_public2023.

Test	Value Identification	Arithmetic Computation	Statistical Inference
GGR	read the data	read between the data	read beyond the data
VLAT	retrieve value, find extremum, determine range	make comparisons	find correlations/trends, characterize distribution, find anomalies, find clusters
CALVI	retrieve value	make comparisons	find correlations/trends, find extremum, make predictions, aggregate values
HOLF	retrieve value	make comparisons, determine range	—
HOLF-Multi	retrieve value	make comparisons	predict trend
ChartQA-Human	—	—	—

Table 1: The relationship between the original task categories and the common set of task categories that can be applied across all six tests. Each row contains the names of the task categories originally defined in each test.

trained using the LLaVA-1.5 framework. Finally, we evaluated **GPT-4V**, a highly performant proprietary model³ [62].

Extracting model output Each model was evaluated on all 851 test items from GGR, VLAT, CALVI, HOLF, HOLF-Multi, and ChartQA-Human. The input to models consisted of two components: (1) an image containing a data visualization and (2) a text prompt containing a question about the visualization written in English. General instructions describing the nature of the task were prepended to each question. In addition, all prompts were formatted to match the model-specific prompts used during training (e.g., prepending the word *Question:* before each question; see Appendix for details).

To assess the test-retest reliability of responses generated by models, we presented each test item 10 times to every model, yielding a total of 8,510 responses per model. We explored commonly used strategies for improving the diversity and fluency of model outputs, including nucleus sampling [38, 70], a decoding procedure wherein sampling is performed over the smallest possible set of words whose cumulative probability exceeds a probability threshold, *top-p*. Specifically, we identified the combination of *top-p* and *temperature* values for each model that produced the best performance on one test (in our experiments, VLAT), and then used these same model-specific *top-p* and *temperature* values for the remaining tests.

Processing model output Determining whether a model had correctly answered a question usually required further processing and validation (Figure 3; see Appendix for details). For instance, 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, **LLaVA1.5-Vicuna-7B** and **LLaVA1.5-Vicuna-13B** often returned the full input prompt as part of its response, so we applied further processing to excise the prompt from any responses that included them. Following prior work [85], we also used GPT-4⁴ to extract only the relevant information in the correct response format from the raw model output. For items that required a floating-point answer, any strings prefixed or suffixed to the floating-point value (e.g. “\$” in “\$3.27” or “cm” in “4cm”) were removed.

Setting model hyperparameters The *max_new_tokens* parameter for all models was set to 270, a relatively high value in order to reduce the likelihood of obtaining a prematurely truncated response. We conducted a grid search over possible combinations of *temperature* and *top-p* parameters that maximized each model’s performance on VLAT, then used these values when evaluating that model on the remaining assessments. We considered *temperature* values of 0.2, 0.4, 0.6, 0.8, and 1.0 when *top-p* was set to 1.0; and *top-p* values ranging between

³Evaluation done through Azure OpenAI using model GPT-4V version vision-preview from April-May 2024.

⁴Evaluation done through Azure OpenAI using model GPT-4 version 1106-preview from April-May 2024.

Model	<i>top-p</i>	<i>temperature</i>
Blip2-FlanT5-4B	0.6	1.0
Blip2-FlanT5-11B	1.0	1.0
LLaVA1.5-Vicuna-7B	0.4	1.0
LLaVA1.5-Vicuna-13B	1.0	0.4
LLaVA1.6-Yi-34B	1.0	0.4
Pix2Struct-0.3B	0.8	1.0
MatCha-0.3B	0.4	1.0
GPT-4V	1.0	0.2

Table 2: *Top-p* and *temperature* hyperparameters used in the current model evaluation study.

0.2, 0.4, 0.6, 0.8, and 1.0 where *temperature* was set to 1.0 (Table 2).

2.5 Statistical analyses

Overall, our statistical analyses aimed to account for reliable variation in model behavior across vision-language models and human participants. We additionally explored the contribution of other factors, including test type, question type, and graph type. Towards this end, we fit generalized mixed-effects linear regression models to assess the relative contribution of each of these factors in predicting model and human responses. We used non-parametric resampling methods to provide quantitative estimates of effect sizes for each factor.

Linear Models We constructed linear regression models to assess the effect of different predictors (i.e., graph type, task type, model) on visualization understanding performance. We used nested model comparisons as our general approach to hypothesis testing as it provides a unified framework that goes beyond the narrower set of cases considered by traditional hypothesis tests (e.g., *t*-tests, ANCOVA).

For example, to assess whether different vision-language models reliably varied in performance, we fit a mixed-effects logistic regression model predicting whether a response was correct or incorrect from model type, fitting random intercepts for each test item. We then compared the fit of this model to a null model that included only the random-effects term for item. In more targeted analyses comparing **GPT-4V** and **Humans** across items involving different types of graphs, we fit a mixed-effects linear regression model predicting proportion correct from “agent type” (i.e., all models and **Humans**), graph type, and their interaction, with variation across individual items modeled using random intercepts. To assess the degree to which any performance gap between **GPT-4V** and **Humans** differed across items involving different graph types, we compared the above model to one without the interaction term. We conducted an analysis following the same structure to compare performance by **GPT-4V** and **Humans** across different task categories.

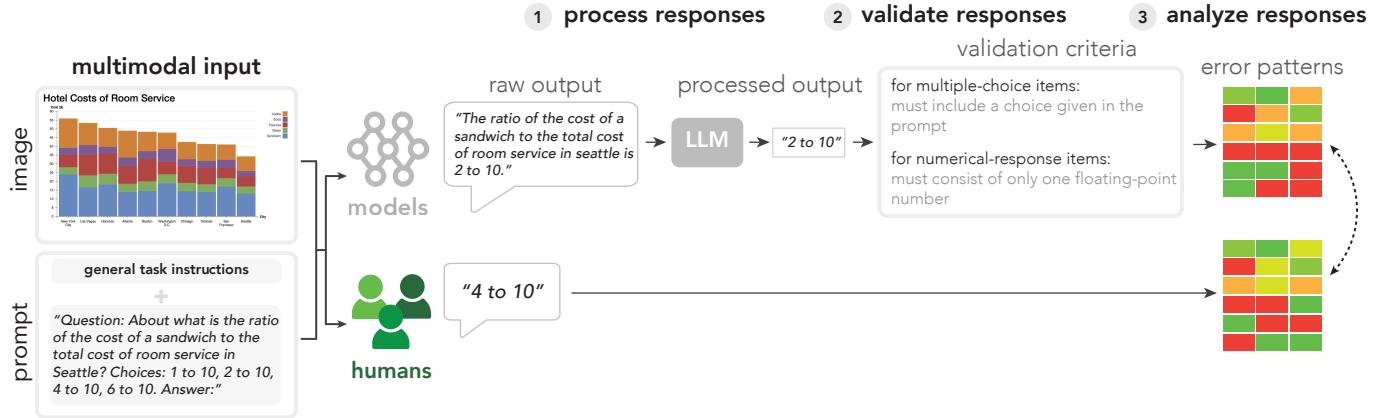


Fig. 3: Procedure for processing and validating model responses for comparison to human responses. All vision-language models were presented with every test item 10 times. Each test item consisted of an image containing a data visualization and a question accompanying it, preceded by general task instructions. The raw output generated by each model was then processed independently by a different large-language model to extract the response in the correct format. These processed outputs were then scored and the pattern of errors compared to human error patterns.

Confidence Intervals To estimate uncertainty in our point estimates of performance, we constructed 95% confidence intervals using bootstrap resampling. For each model, we resampled items with replacement 1,000 times, each time re-computing performance and retaining only items for which valid responses were ever generated. To estimate differences between any two groups, we constructed 95% confidence intervals based on the sampling distribution for the difference between the point estimates for each group derived on each bootstrap resampling iteration.

3 RESULTS

Our core finding is that current state-of-the-art vision-language models consistently underperform humans on reasoning tasks involving data visualizations, and that this gap is especially pronounced for assessments that were developed to measure these skills in humans.

3.1 How often do models produce valid responses?

First, we determined which model responses were correctly formatted, and thus amenable to further analysis. For multiple-choice questions in GGR, VLAT, and CALVI, 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, HOLF, HOLF-Multi, and ChartQA-Human, a response was considered to be valid if it contained a single floating-point value.

Using these criteria, we computed the proportion of valid responses generated for each test item by every model (Figure 4). We found that no model always provided valid responses. When pooling all items across tests, we found that **LLava1.6-Yi-34B** produced the lowest proportion of valid responses (average: 0.32; 95% CI = [0.30, 0.34]; 2735/8510 responses were valid) and **MatCha-0.3B** produced the highest proportion of valid responses (average: 0.83; 95% CI = [0.82, 0.84]; 7082/8510 responses were valid). However, the BLIP-2 variants stood out for most consistently producing a high proportion of valid responses across all tests (**Blip2-FlanT5-4B**: 0.82; 95% CI = [0.75, 0.88]); **Blip2-FlanT5-11B**: 0.80; 95% CI = [0.72, 0.86]).

These results suggest that reliably extracting task-relevant output from these models remains challenging. This limitation has implications for the way that sound comparisons between model and human performance can be made, depending on whether invalid responses are considered to be incorrect responses generated under fair evaluation settings, and thus reflect limitations of the model, or are considered to be the product of limitations in our evaluation protocol. To ensure that our conclusions are not dependent on this choice, we conducted subsequent analyses under both ways of interpreting invalid responses from models.

3.2 How often do models produce accurate answers?

Next, we compared the accuracy of the responses achieved by models to that by human participants (Figure 5). We established an upper and lower bound on estimates of model performance by computing accuracy when considering only valid responses (upper bound) and when considering all responses, including invalid ones, where invalid responses were marked as incorrect (lower bound). For GGR, VLAT, and CALVI (the ‘multiple-choice’ tests), we computed the proportion of correct responses produced by humans and models. For the 9 items requiring numerical responses in GGR, responses were only deemed correct if they *exactly* matched the ground-truth answer provided by the original test designers, to ensure fair comparisons between vision-language models and human responses to items on this test. For HOLF, HOLF-Multi, and ChartQA (the ‘numerical-response’ tests), following prior work [56, 57], we computed a *relaxed accuracy* metric, which considers a response to be correct if it falls within 5% of the correct answer. The same standard was applied to both human and vision-language model responses.

We found that models reliably differed in performance from one another ($\chi^2(7) = 3,591, p < .001$). We further found that considering only valid responses inflated estimates of model performance on the numerical-response tests to some degree (Δ proportion correct: 0.041; 95% CI = [0.024, 0.057]), with a more modest impact on estimates of model performance on multiple-choice tests (Δ proportion correct: 0.12; 95% CI = [0.066, 0.18]). These results suggest the value of jointly considering both stricter and more lenient ways of assessing model performance to more clearly establish the range of expected performance levels for any given model.

When examining only the valid responses generated by models (Figure 5), we found that **GPT-4V** was the best performing model on five out of the six tests. However, it performed reliably worse than human participants on GGR (Δ mean proportion correct (model – human): -0.56; 95% CI = [-0.78, -0.30]), HOLF (Δ mean relaxed accuracy: -0.14; 95% CI = [-0.25, -0.15]), and HOLF-Multi (Δ mean relaxed accuracy: -0.07; 95% CI = [-0.16, -0.02]). It did approach human performance on VLAT (Δ mean proportion correct: -0.12; 95% CI = [-0.26, 0.01]) and ChartQA-Human (Δ mean relaxed accuracy: -0.060; 95% CI = [-0.14, 0.020]). **Pix2Struct-0.3B** performed best among models on CALVI, and also at a level approaching human performance (Δ mean proportion correct: -0.23; 95% CI = [-0.42, -0.018]). Among those items for which **Pix2Struct-0.3B** could generate a valid response at all, the gap between **Pix2Struct-0.3B** and **Humans** was all but closed for the misleading items (Δ mean proportion correct: -0.01; 95% CI = [-0.28, 0.22]), but not for the non-misleading items (Δ mean proportion correct: -0.42; 95% CI = [-

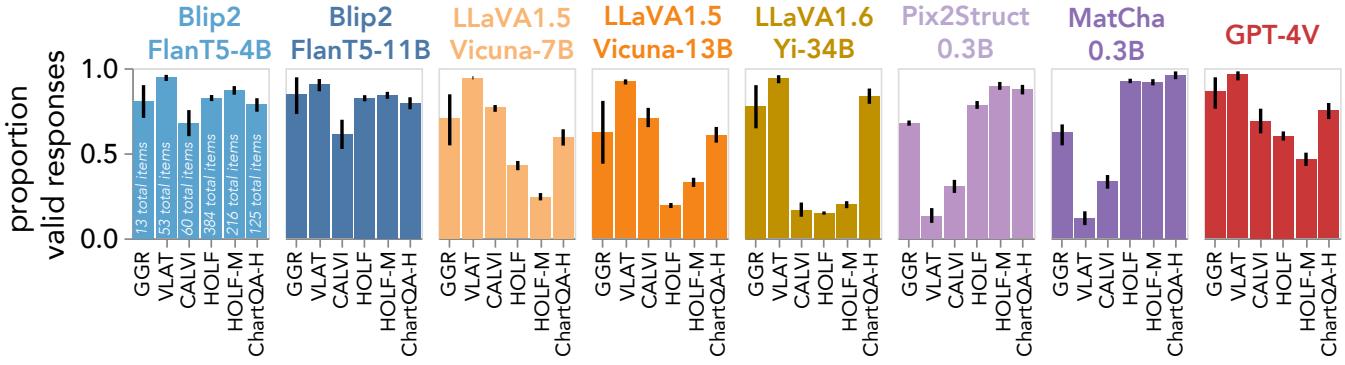


Fig. 4: Proportion of valid responses produced by each model on each assessment.

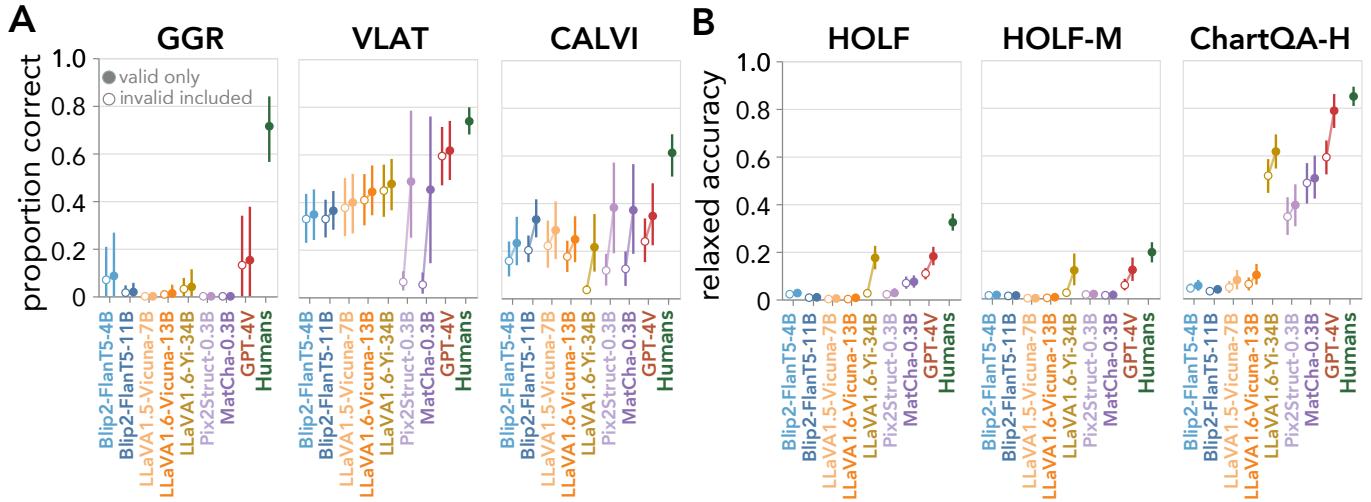


Fig. 5: Human and model performance on (A) the mean proportion correct in multiple-choice assessments (GGR, VLAT, and CALVI) and (B) the mean relaxed accuracy in numerical-response assessments (HOLF, HOLF-multi, and ChartQA). Relaxed accuracy is calculated by the proportion of responses that fall within 5% of the correct answer. Empty circles represent estimates of model performance based on all responses, with any invalid responses marked as incorrect. Filled circles represent estimates of model performance based only on valid responses, and therefore reflect an upper bound on model performance. All error bars represent bootstrapped 95% confidence intervals.

0.90, -0.22].

We also compared model performance to that of humans using all model responses, with any invalid model responses marked as incorrect. Under these conditions, we found that **GPT-4V** performed best among all models on all six assessments, including CALVI. Again we found that **GPT-4V** performed worse than human participants on several of the tests: GGR (Δ mean proportion correct (model – human): -0.58; 95% CI = [-0.76, -0.35]), CALVI (Δ mean proportion correct: -0.37; 95% CI = [-0.50, -0.23]), HOLF (Δ mean relaxed accuracy: -0.37; 95% CI = [-0.50, -0.23]), and HOLF-Multi (mean relaxed accuracy: -0.14; 95% CI = [-0.19, -0.09]). It did approach human-level performance on VLAT (Δ mean proportion correct: -0.12; 95% CI = [-0.25, 0.01]). However, by contrast with what we found when examining only valid responses, **GPT-4V** did not achieve human-level on ChartQA-Human (Δ mean relaxed accuracy: -0.26; 95% CI = [-0.34, -0.17]).

These results suggest meaningful variation across assessments and evaluation strategies with respect to the apparent size of the gap in performance between current vision-language models and humans on data visualization understanding tasks. In particular, we found that when we considered only items for which models could generate valid responses, the model-human performance gap narrowed considerably, especially for the subset of items from ChartQA, which is widely used to benchmark multimodal reasoning capabilities in the machine learning literature. However, this gap widened substantially when we considered

all responses generated by models, including those on items for which it never produced a properly formatted response. Taken together, our analyses show a reliable gap in performance between these models and human participants on several of the tests in our suite, including GGR, CALVI and HOLF, suggesting the value of using a diversity of independently designed measures for identifying opportunities for further model development.

3.3 How does model and human performance vary across different types of graphs and tasks?

We next examined the degree to which the model-human performance gap varied across different categories of test items, regardless of which test they had come from. Specifically, we examined variation in performance that could be attributable to the type of graph shown (e.g., bar plot vs. scatter plot) or the type of task being performed (e.g., value identification vs. arithmetic computation). Here we focused on the comparison between **GPT-4V** and **Humans**, because **GPT-4V** was the most consistently high-performing model on our suite of tests.

We found that **Humans** consistently outperformed **GPT-4V** on most types of graphs, regardless of whether an item came from a multiple-choice or numerical-response test (Figure 6 left). The exceptions were stacked area and bubble charts that required multiple-choice responses and pie charts that required numerical responses. For these item categories, **GPT-4V** scored higher than humans. However, we did find that the kinds of graphs on which **Humans** did well also tended to

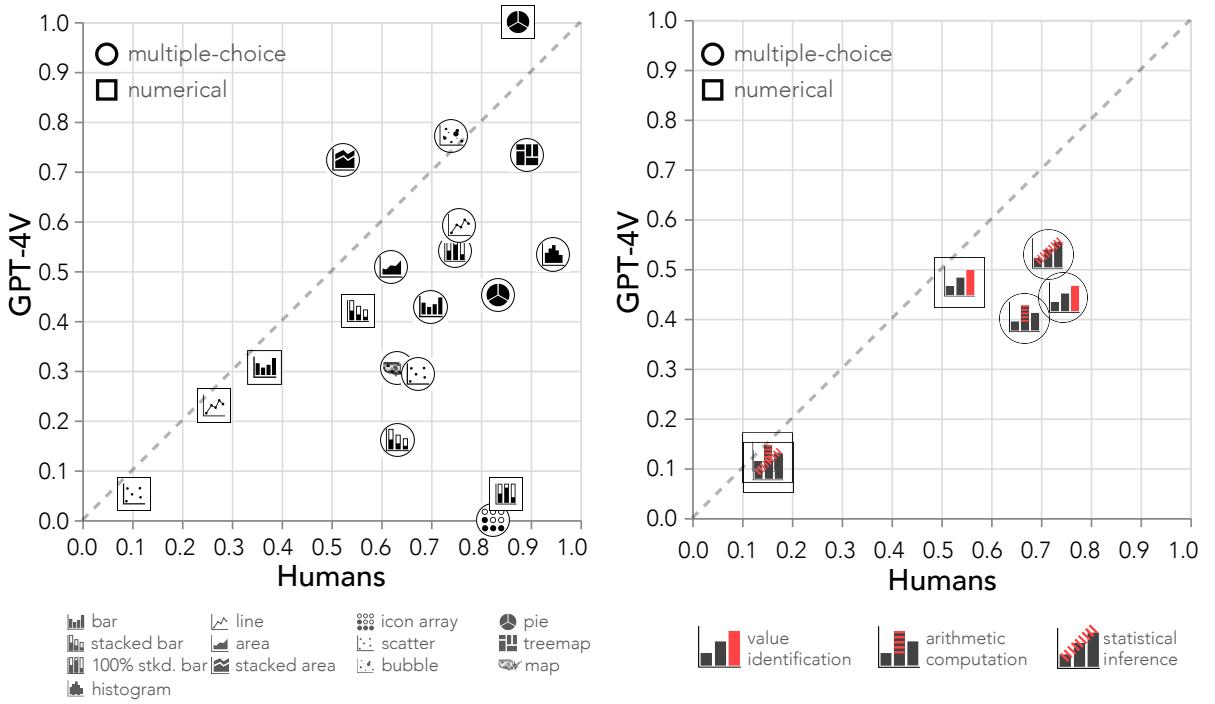


Fig. 6: Mean proportion of correct responses between **GPT-4V** and **Humans** across different categories of *graphs* (left) and *tasks* (right).

be the ones on which **GPT-4V** also performed well (Pearson’s $r = 0.40$, $p = .091$). Nevertheless, we found that the size of the gap between **GPT-4V** and **Humans** did reliably vary across different types of graphs ($\chi^2(12) = 23.993$, $p = 0.020$).

We found that **Humans** outperformed **GPT-4V** across all three task types (i.e., value identification, arithmetic computation, statistical inference; Figure 6 right). We further found the kinds of tasks on which **Humans** performed well were also often the ones on which **GPT-4V** also performed well (Pearson’s $r = 0.94$, $p = 0.005$) though the size of the gap did vary across tasks ($\chi^2(2) = 21.288$, $p < .001$).

Taken together, this more detailed comparison between **GPT-4V** and **Humans** suggests that some of the categories of items that were more difficult for **Humans** were also relatively more difficult for **GPT-4V**, even though **GPT-4V** achieved lower overall performance relative to **Humans**.

3.4 How similar were the error patterns generated by models and humans?

To more thoroughly investigate any covariation between human and model performance, we analyzed the full set of error patterns produced by humans and all models across the six assessments (Figure 7). Comparing error patterns across items is valuable because they could reveal aspects of how model and human behavior relate to each other that might not be apparent based on analyses of average performance on entire tests or pre-defined categories of items. For each model, we constructed two versions of an error-pattern vector with length equal to the total number of items, where each element represented the proportion of correct responses it generated for a single item. One version of this error-pattern vector was derived from those items for which that model had generated at least one valid response; the second version was defined for all items, including ones where the model generated only invalid responses, which were marked as incorrect. We then constructed an analogous error-pattern vector for humans, where each element represented the proportion of correct responses across all participants who had been given that item. Next, we computed how correlated the error-pattern vectors were between humans and all models. We computed a human “noise ceiling” reflecting how well any model could be expected to approximate human error patterns,

given the variability in our estimates of human performance. We estimated this noise ceiling by constructing the sampling distribution of the Spearman-Brown-corrected correlation between error patterns computed on randomly partitioned halves of the human data.

When restricting estimates of human-model consistency to the items for which we obtained valid responses, we found that all models across all six tests consistently fell far short of the human noise ceiling, with different models being closer to that ceiling for different tests. **GPT-4V** was the closest for GGR (0.21; 95% CI = [0.020, 0.45]; noise ceiling: 1.00; 95% CI = [1.00, 1.00]), HOLF-Multi (0.63; 95% CI = [0.50, 0.74]; noise ceiling: 0.99; 95% CI = [0.98, 0.99]), and ChartQA-Human (0.43; 95% CI = [0.29, 0.56]; noise ceiling: 0.89; 95% CI = [0.80, 0.94]). **LLaVA1.5-Vicuna-13B** was closest for VLAT (0.32; 95% CI = [0.17, 0.47]; noise ceiling: 1.00; 95% CI = [1.00, 1.00]). **Pix2Struct-0.3B** was closest for CALVI (0.51; 95% CI = [0.25, 0.78]; noise ceiling: 0.99; 95% CI = [0.99, 1.00]). Finally, **LLaVA1.6-Yi-34B** was closest for HOLF (0.42; 95% CI = [0.34, 0.50]; noise ceiling: 0.89; 95% CI = [0.88, 0.91]).

When estimating human-model consistency when considering all items, including those for which a given model generated only invalid responses, **GPT-4V** was the closest to **Humans** across all assessments, but it still fell short of the human noise ceiling in every test, including GGR (0.21; 95% CI = [0.04, 0.42]), VLAT (0.38; 95% CI = [0.19, 0.57]), CALVI (0.31; 95% CI = [0.12, 0.50]), HOLF (0.44; 95% CI = [0.37, 0.50]), HOLF-Multi (0.55; 95% CI = [0.43, 0.65]), and ChartQA-Human (0.34; 95% CI = [0.22, 0.45]).

Taken together, these findings suggest that when scrutinizing the patterns of performance from models and humans more comprehensively, current vision-language models generate error patterns that are reliably distinguished from those produced by humans. While proprietary systems like **GPT-4V** was most aligned with human behavior among the models in our suite, open-source models such as **LLaVA1.6-Yi-34B** and **Pix2Struct-0.3B** did not necessarily lag far behind. These findings suggest promising opportunities for developing open and human-aligned models of visualization understanding.

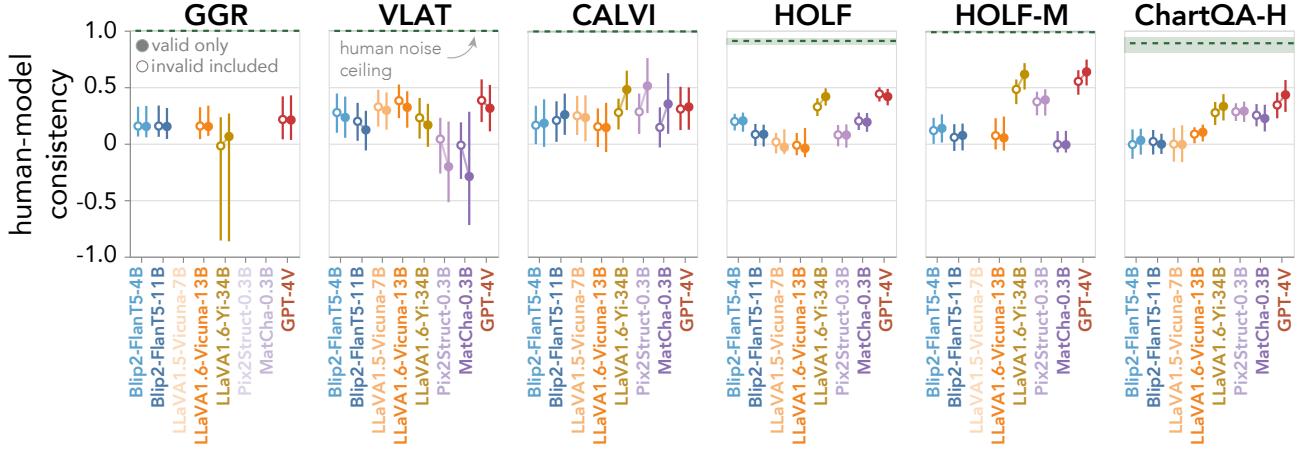


Fig. 7: Pearson correlation between error patterns produced by each model and **Humans** on the same assessments. Empty circles represent estimates of model-human correlation based on all responses, with any invalid responses marked as incorrect. Filled circles represent estimates of model-human correlation based only on valid responses. All error bars represent bootstrapped 95% confidence intervals. Human noise ceiling constructed by estimating the sampling distribution of the Spearman-Brown-corrected correlations between split-halves of the human data.

4 DISCUSSION

A key open challenge in cognitive science is to develop mechanistic accounts of the mental processes that enable people to read and understand a wide variety of symbolic displays of information, including data visualizations. Here we asked to what degree vision-language models, an emerging class of AI systems that can operate over both text and images [39, 49, 53, 85], might provide the basis for future development of computational cognitive models of human visualization understanding. We constructed a suite of visualization literacy benchmarks, CHART-6, which combines six tests that were developed independently by researchers across different disciplinary communities. This suite included five assessments intended to measure data visualization understanding in humans, GGR, VLAT, CALVI, HOLF, & HOLF-Multi, as well as a representative subset of items from ChartQA, a commonly used benchmark that was developed to measure these skills in AI systems. We evaluated a set of eight state-of-the-art vision-language models and compared these models’ performance to that of human participants. Even when considering only valid responses from models (and thereby, if anything, *overestimating* their performance), we found that models still performed worse than human participants, on average. At the same time, the categories of items on which **GPT-4V**, the most performant model, performed relatively well were also often those that human participants did well on, suggesting some degree of alignment in relative performance levels achieved by this model and humans. Nevertheless, further inspection of all models’ patterns of responses across the full set of test items revealed that no model generated responses that approached the human noise ceiling. Our results contribute to a growing body of cognitive-AI benchmarking efforts that employ large-scale controlled experimentation to reveal gaps between humans and AI systems on a common set of real-world tasks involving complex, naturalistic inputs [4, 23, 55, 58, 60, 74]. Taken together, our findings suggest that while vision-language models show promise as a class of models that can reason over a broad class of visualization and question types, there remain opportunities to improve their alignment with human behavior, which would enhance their value as potential scientific models of the computations involved in visualization understanding.

An outstanding question concerns where the identified gaps between models and humans come from and how to close them. Data visualization literacy is often acquired by humans through formal education and training. While modern vision-language models are trained on very large datasets that likely include data visualizations [44], they generally do not engage with these inputs or receive social feedback in the ways that human learners do [2, 35, 64]. An important future direction will 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 beyond current pre-training strategies [34]. This stands to not only help close the quantitative alignment gap but potentially mitigate qualitative differences between vision-language models and humans. For example, text-based annotations embedded in data visualizations seem to influence model performance [82, 83] to a greater degree than they do human performance [68, 76], although more direct comparisons between models and humans, similar to the present work, is needed. Moreover, other work suggests that vision-language models often fail to detect visual properties that are generally salient to humans, such as intersections between lines, overlap between shapes, and the number of simple visual elements in a scene [69] — all foundational abilities needed to succeed on the visualization understanding tasks investigated in the present work. One possibility is that the gaps between human and model performance on the data visualization understanding tasks in CHART-6 can be explained, at least in part, by general limitations in models’ visual processing capabilities. Future work should seek to elucidate the relationship between model performance on a broad suite of both perceptual tasks [29, 37, 59] and data visualization understanding tasks to more directly evaluate this claim.

Another important future direction will be to develop more unified measures of data visualization literacy. Currently, the landscape of assessments and benchmarks for measuring these skills is fragmented, and there is a lack of consensus regarding the key components of data visualization literacy and exactly how to measure them [9, 10, 12, 31, 41, 46, 56, 80, 86]. Furthermore, there might also be important aspects of data visualization literacy that are not well captured by existing measures. Many benchmarks used in the machine learning literature [17, 37, 53, 80, 82] contain a large number of graphs that are similar to those that can be found in real-world settings, yet the questions accompanying them are often simpler than would be expected for a comprehensive measure of data visualization understanding. Meanwhile, assessments of data visualization literacy designed for humans often contain fewer items, but tap into a broader array of skills [30, 31, 46]. Future work could analyze the properties of existing measures [12] and leverage the resulting insights to develop scalable procedures for developing more comprehensive measures [18]. Adaptive testing methods might also be used to more efficiently administer these comprehensive tests to humans [19].

Data visualizations are a versatile tool for supporting discovery, communication, and learning. We envision CHART-6 being used to track the progress of artificial systems towards achieving human-like behavior on tasks involving data visualizations, and thus a procedure

for identifying promising systems for further investigation as candidate computational models of the cognitive operations involved in these tasks. Here we found that many current vision-language models show promise, but still fall short of providing a strong foundation for developing cognitive models. As progress is made on this front, we believe it to be likely that AI systems displaying more human-like understanding of visual, linguistic, and mathematical concepts could be used to design more effective STEM learning environments and tools to support scientific communication.

REFERENCES

- [1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1
- [2] B. Alper, N. H. Riche, F. Chevalier, J. Boy, and M. Sezgin. Visualization literacy at elementary school. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 5485–5497, 2017. 8
- [3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117. IEEE, 2005. 1
- [4] D. Bear, E. Wang, D. Mrowca, F. Binder, H.-Y. Tung, P. RT, C. Holdaway, S. Tao, K. Smith, F.-Y. Sun, F.-F. Li, N. Kanwisher, J. Tenenbaum, D. Yamins, and J. Fan. Physion: Evaluating physical prediction from vision in humans and machines. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021. 2, 8
- [5] J. Bertin. *Graphics and graphic information processing*. Walter de Gruyter, 1981. 1
- [6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models, 2022. 1
- [7] T. Bonnen, S. Fu, Y. Bai, T. O’Connell, Y. Friedman, N. Kanwisher, J. Tenenbaum, and A. Efros. Evaluating multiview object consistency in humans and image models. *Advances in Neural Information Processing Systems*, 37:43533–43548, 2024. 2
- [8] K. Börner, A. Bueckle, and M. Ginda. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, 116(6):1857–1864, 2019. doi: [10.1073/pnas.1807180116](https://doi.org/10.1073/pnas.1807180116) 1
- [9] K. Börner, A. Maltese, R. N. Balliet, and J. Heimlich. Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, 15(3):198–213, 2016. 8
- [10] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1963–1972, 2014. doi: [10.1109/TVCG.2014.2346984](https://doi.org/10.1109/TVCG.2014.2346984) 1, 8
- [11] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013. doi: [10.1109/TVCG.2013.124](https://doi.org/10.1109/TVCG.2013.124) 1
- [12] E. Brockbank, A. Verma, H. Lloyd, H. Huey, L. Padilla, and J. E. Fan. Evaluating convergence between two data visualization literacy assessments. *Cognitive Research: Principles and Implications*, 2025. 8
- [13] P. A. Carpenter and P. Shah. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4(2):75–100, 1998. doi: [10.1037/1076-898X.4.2.75](https://doi.org/10.1037/1076-898X.4.2.75) 1
- [14] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 1, 3
- [15] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: [10.1080/01621459.1984.10478080](https://doi.org/10.1080/01621459.1984.10478080) 1
- [16] N. R. Council. *Developing assessments for the next generation science standards*. National Academies Press, 2014. 1
- [17] Y. Cui, L. W. Ge, Y. Ding, L. Harrison, F. Yang, and M. Kay. Promises and Pitfalls: Using Large Language Models to Generate Visualization Items. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1094–1104, Jan. 2025. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: [10.1109/TVCG.2024.3456309](https://doi.org/10.1109/TVCG.2024.3456309) 8
- [18] Y. Cui, W. G. Lily, Y. Ding, L. Harrison, F. Yang, and M. Kay. Promises and pitfalls: Using large language models to generate visualization items. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 8
- [19] Y. Cui, W. G. Lily, Y. Ding, F. Yang, L. Harrison, and M. Kay. Adaptive assessment of visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):628–637, 2023. 8
- [20] G. Cumming and S. Finch. Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2):170–180, 2005. doi: [10.1037/0003-066X.60.2.170](https://doi.org/10.1037/0003-066X.60.2.170) 1
- [21] F. R. Curcio. Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, 18(5):382–393, 1987. doi: [10.2307/749086](https://doi.org/10.2307/749086) 1
- [22] J. E. Fan, W. A. Bainbridge, R. Chamberlain, and J. D. Wammes. Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, 2(9):556–568, 2023. 1
- [23] T. Fel, I. Felipe, D. Linsley, and T. Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432, 2022. 2, 8
- [24] E. E. Firat, A. Joshi, and R. S. Laramee. Interactive visualization literacy: The state-of-the-art. *Information Visualization*, 21(3):285–310, 2022. doi: [10.1177/14738716221081831](https://doi.org/10.1177/14738716221081831) 1
- [25] A. R. Fox. Theories and models in graph comprehension. *Visualization Psychology*, pp. 39–64, 2023. 1
- [26] S. L. Franconeri, L. Padilla, P. Shah, J. M. Zacks, and J. Hullman. The science of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3):110–161, 2021. 1
- [27] S. N. Friel, F. R. Curcio, and G. W. Bright. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32(2):124–158, 2001. doi: [10.2307/749671](https://doi.org/10.2307/749671) 1, 2
- [28] E. Fry. Graphical literacy. *Journal of Reading*, 24(5):383–389, 1981. 1
- [29] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024. 8
- [30] M. Galesic and R. Garcia-Retamero. Graph literacy: A cross-cultural comparison. *Medical decision making*, 31(3):444–457, 2011. doi: [10.1177/0272989X10373805](https://doi.org/10.1177/0272989X10373805) 2, 8
- [31] L. W. Ge, Y. Cui, and M. Kay. Calvi: Critical thinking assessment for literacy in visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, article no. 815, 18 pages. Association for Computing Machinery, New York, NY, USA, 2023. doi: [10.1145/3544548.3581406](https://doi.org/10.1145/3544548.3581406) 2, 3, 8
- [32] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018. 2
- [33] D. J. Gillan and R. Lewis. A componential model of human interaction with graphs: 1. Linear regression modeling. *Human Factors*, 36(3):419–440, 1994. doi: [10.1177/001872089403600303](https://doi.org/10.1177/001872089403600303) 1
- [34] A. Gupta, V. Gupta, S. Zhang, Y. He, N. Zhang, and S. Shah. Enhancing question answering on charts through effective pre-training tasks. *arXiv preprint arXiv:2406.10085*, 2024. 8
- [35] H. Gweon, J. Fan, and B. Kim. Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048, 2023. doi: [10.1098/rsta.2022.0048](https://doi.org/10.1098/rsta.2022.0048) 8
- [36] M. Hegarty. Multimedia learning about physical systems. In R. E. Mayer, ed., *The Cambridge Handbook of Multimedia Learning*, pp. 447–466. Cambridge University Press, Cambridge, UK, 2005. 1
- [37] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 8
- [38] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *International Conference on Learning Representations*, 2020. 4
- [39] K.-H. Huang, H. P. Chan, Y. R. Fung, H. Qiu, M. Zhou, S. Joty, S.-F. Chang, and H. Ji. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models, 2024. 8
- [40] H. Huey. *Adaptive Visualization Strategies Across Drawings, Diagrams, and Data Visualizations*. University of California, San Diego, 2024. 2, 3
- [41] H. Huey, L. A. Oey, H. Lloyd, and J. E. Fan. 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, 2023. 2, 3, 8
- [42] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024. doi: [10.1098/rsta.2023.0254](https://doi.org/10.1098/rsta.2023.0254) 1
- [43] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum*, vol. 37(3), pp. 157–167. Wiley Online Library, 2018. doi: [10.1111/cgf.13409](https://doi.org/10.1111/cgf.13409) 1
- [44] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024. 1, 8
- [45] K. Lee, M. Joshi, I. R. Tunc, H. Hu, F. Liu, J. M. Eisenschlos, U. Khandawal, P. Shaw, M.-W. Chang, and K. Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. *Proceedings of the 40th International Conference on Machine Learning*, 202:18893–18912, 23–29 Jul 2023. 3
- [46] S. Lee, S.-H. Kim, and B. C. Kwon. Vlat: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560, 2017. doi: [10.1109/TVCG.2016.2598920](https://doi.org/10.1109/TVCG.2016.2598920) 1, 2, 8
- [47] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan. Seed-bench: Benchmarking multimodal large language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, June 2024. 3
- [48] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning*, 202:19730–19742, 23–29 Jul 2023. 3
- [49] Z. Li, H. Miao, V. Pascucci, and S. Liu. Visualization literacy of multimodal large language models: A comparative study, 2024. 8
- [50] F. Liu, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, and J. Eisenschlos. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12756–12770, July 2023. doi: [10.18653/v1/2023.acl-long.714](https://doi.org/10.18653/v1/2023.acl-long.714) 3
- [51] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024. 1, 3
- [52] H. Lloyd, H. Huey, E. Brockbank, L. Padilla, and J. E. Fan. What is graph comprehension and how do you measure it? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45, 2023. 3
- [53] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *The Twelfth International Conference on Learning Representations*, 2024. 1, 8
- [54] A. Lundgard and A. Satyanarayan. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1073–1083, 2022. doi: [10.1109/TVCG.2021.3114770](https://doi.org/10.1109/TVCG.2021.3114770) 1
- [55] R. Marjeh, I. Sucholutsky, P. van Rijn, N. Jacoby, and T. L. Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, 2024. 2, 8
- [56] A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, May 2022. doi: [10.18653/v1/2022.findings-acl.177](https://doi.org/10.18653/v1/2022.findings-acl.177) 1, 3, 5, 8
- [57] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. Plotqa: Reasoning over scientific plots. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020. 1, 5
- [58] K. Mukherjee, H. Huey, X. Lu, Y. Vinker, R. Aguina-Kang, A. Shamir, and J. Fan. Seva: Leveraging sketches to evaluate alignment between human and machine visual abstraction. *Advances in Neural Information Processing Systems*, 36:67138–67155, 2023. 2, 8
- [59] K. Mukherjee, D. Ren, D. Moritz, and Y. Assogba. Encqa: Benchmarking vision-language models on visual encodings for charts. *arXiv preprint*, 2025. 8
- [60] K. Mukherjee, T. T. Rogers, and K. B. Schloss. Large language models estimate fine-grained human color-concept associations. *arXiv preprint arXiv:2406.17781*, 2024. 2, 8
- [61] T. Munzner. *Visualization analysis and design*. CRC press, 2014. 1
- [62] OpenAI. Gpt-4 technical report, 2023. 1, 4
- [63] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, 3(1):29, Jul 2018. doi: [10.1186/s41235-018-0120-9](https://doi.org/10.1186/s41235-018-0120-9) 1
- [64] K. Peppler, A. Keune, and A. Han. Cultivating data visualization literacy in museums. *Information and Learning Sciences*, 122(1/2):1–16, 2021. 8
- [65] S. Pinker. *A theory of graph comprehension.*, pp. 73–126. Artificial intelligence and the future of testing. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1990. 1
- [66] G. J. Quadri and P. Rosen. A Survey of Perception-Based Visualization Studies by Task. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5026–5048, Dec. 2022. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: [10.1109/TVCG.2021.3098240](https://doi.org/10.1109/TVCG.2021.3098240) 1
- [67] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139:8748–8763, 18–24 Jul 2021. 3
- [68] M. D. Rahman, B. Doppalapudi, G. J. Quadri, and P. Rosen. A survey on annotations in information visualization: Empirical insights, applications, and challenges. *arXiv preprint arXiv:2410.05579*, 2024. 8
- [69] P. Rahmazadehgeravi, L. Bolton, M. R. Taesiri, and A. T. Nguyen. Vision language models are blind. *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024. 8
- [70] M. Renze and E. Guven. The effect of sampling temperature on problem solving in large language models, 2024. 4
- [71] P. Shah and E. G. Freedman. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, 3(3):560–578, 2011. doi: [10.1111/j.1756-8765.2009.01066.x](https://doi.org/10.1111/j.1756-8765.2009.01066.x) 1
- [72] P. Shah, E. G. Freedman, and I. Vekiri. The comprehension of quantitative information in graphical displays. *The Cambridge handbook of visuospatial thinking*, pp. 426–476, 2005. 1
- [73] P. Shah and J. Hoeffner. Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1):47–69, Mar 2002. doi: [10.1023/A:1013180410169](https://doi.org/10.1023/A:1013180410169) 1
- [74] T. Shu, A. Bhandwaldar, C. Gan, K. Smith, S. Liu, D. Gutfreund, E. Spek, J. Tenenbaum, and T. Ullman. Agent: A benchmark for core psychological reasoning. In *International conference on machine learning*, pp. 9614–9625. PMLR, 2021. 8
- [75] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82(398):454–465, 1987. 1
- [76] C. Stokes, C. X. Bearfield, and M. A. Hearst. The role of text in visualizations: How annotations shape perceptions of bias and influence predictions. *IEEE Transactions on Visualization and Computer Graphics*, 30(10):6787–6800, 2023. 8
- [77] E. R. Tufte. *The visual display of quantitative information*, vol. 2. Graphics press Cheshire, CT, 1983. 1
- [78] J. W. Tukey et al. *Exploratory data analysis*, vol. 2. Springer, 1977. 1
- [79] B. Tversky. Visualizing thought. *Topics in Cognitive Science*, 3(3):499–535, 2011. 1
- [80] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024. 1, 8
- [81] L. Wilkinson. *The grammar of graphics*. Springer, 2012. 1
- [82] Y. Wu, L. Yan, L. Shen, Y. Wang, N. Tang, and Y. Luo. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. *arXiv preprint arXiv:2405.07001*, 2024. 1, 8
- [83] Z. Xu, S. Du, Y. Qi, C. Xu, C. Yuan, and J. Guo. ChartBench: A Benchmark for Complex Visual Reasoning in Charts, June 2024. arXiv:2312.15915 [cs]. 8
- [84] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, K. Yu, P. Liu, Q. Liu, S. Yue, S. Yang, S. Yang, T. Yu, W. Xie, W. Huang, X. Hu, X. Ren, X. Niu, P. Nie, Y. Xu, Y. Liu, Y. Wang, Y. Cai, Z. Gu, Z. Liu, and Z. Dai. Yi: Open foundation models by 01.ai, 2024. 3
- [85] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. 1, 3, 4, 8
- [86] X. Zeng, H. Lin, Y. Ye, and W. Zeng. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning, 2024. 8
- [87] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. 1

A MODEL EVALUATION DETAILS

A.1 Prompt for administering test items

To construct the prompts used to administer the test items to each model, a model-specific prefix and suffix were combined with the original question text (Table 3; Table 4). None of the text in the original question was otherwise modified.

A.2 Prompt for response processing

Table 5 shows prompt templates that were used to extract an answer from a model’s response. All prompts were processed using OpenAI GPT-4 using the hyperparameter values: *max_tokens* set to 2000, *top-p* set to 1.0, and *temperature* set to 1.0. For items requiring a numerical response in GGR, the prompt template for continuous-response assessments was used.

Assessment type	Prompt template
continuous-response	Question: You will be presented with a series of data visualizations, each accompanied by a question. Your goal is to answer each question as accurately and as quickly as you are able. It is common for people to not be fully sure when answering these questions, but please do your best on each question, even if you have to make a guess. {Question} <u>Answer:</u>
multiple-choice	Question: The first part of this study consists of 53 multiple choice questions associated with visualizations, and you will be asked to choose the best answer for each question. You are required to provide an answer to the current question. Your answer must be one of the choices provided. Your answer must be one of the choices provided. {Question} Choices: {Choice 1}, {Choice 2}, {Choice 3}, {Choice 4}. <u>Answer:</u>

Table 3: Each prompt used to administer a test item embeds the original question text (Question) and possible choices (Choice 1-4), where applicable, within a prompt template with a model-specific prefix and suffix (terms underlined in the example above).

Model	Prompt prefix	Prompt suffix
Blip2-FlanT5-4B	Question:	\n <u>Answer:</u>
Blip2-FlanT5-11B	Question:	\n <u>Answer:</u>
LLaVA1.5-Vicuna-7B	USER: <image> \n	\nASSISTANT:
LLaVA1.5-Vicuna-13B	USER: <image> \n	\nASSISTANT:
LLaVA1.6-Yi-34B	USER:	\nASSISTANT:
Pix2Struct-0.3B	Question:	<u>Answer:</u>
MatCha-0.3B	Question:	<u>Answer:</u>
GPT-4V	Question:	\n <u>Answer:</u>

Table 4: Model-specific prefixes and suffixes, optionally containing an <image> token to indicate where an image should be inserted and a \n character to indicate where a new line should be inserted.

Assessment type	Prompt template
continuous-response	Please read the following example. Then extract the answer from the model response and type it at the end of the prompt. Hint: Please answer the question requiring a floating-point number with two decimal places and provide the final value, e.g., 1.23, 1.34, 1.45, at the end. Question: {Question} Model response: {Model Response} Extracted answer:
multiple-choice	Please read the following example. Then extract the answer from the model response and type it at the end of the prompt. Hint: Please answer the question and provide the correct option. Question: {Question} Choices: {Choice 1}, {Choice 2}, {Choice 3}, {Choice 4}. Model response: {Model Response} Extracted answer:

Table 5: Prompt templates which contain the corresponding question (Question) and choices (Choice 1-4) for a given model response (Model Response) to an item.