An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks

Alexander Bendeck and John Stasko

Abstract—Large Language Models (LLMs) like GPT-4 which support multimodal input (i.e., prompts containing images in addition to text) have immense potential to advance visualization research. However, many questions exist about the visual capabilities of such models, including how well they can read and interpret visually represented data. In our work, we address this question by evaluating the GPT-4 multimodal LLM using a suite of task sets meant to assess the model's visualization literacy. The task sets are based on existing work in the visualization community addressing both automated chart question answering and human visualization literacy across multiple settings. Our assessment finds that GPT-4 can perform tasks such as recognizing trends and extreme values, and also demonstrates some understanding of visualization design best-practices. By contrast, GPT-4 struggles with simple value retrieval when not provided with the original dataset, lacks the ability to reliably distinguish between colors in charts, and occasionally suffers from hallucination and inconsistency. We conclude by reflecting on the model's strengths and weaknesses as well as the potential utility of models like GPT-4 for future visualization research. We also release all code, stimuli, and results for the task sets at the following link: https://doi.org/10.17605/OSF.IO/F39J6

Index Terms—Visualization Literacy, Large Language Models, Natural Language.

1 Introduction

Over the past few years, advances in the field of Natural Language Processing (NLP) have led to the wide commercial availability of Large Language Models (LLMs). LLMs have been shown to be both powerful and practical aids for completing a wide variety of tasks related to text generation [19, 48]. Among the most well-known and widely-used models are those in the Generative Pre-trained Transfomer (GPT) series, such as GPT-3.5 and GPT-4. Such models have demonstrated great utility in fields including education, healthcare, and more [48], with visualization research being no exception. Existing work has already assessed the ability of LLMs to generate code for creating visualizations [21, 50], help users author data-driven articles [68], and complete various other visualization and data analysis tasks [14].

The most recent versions of some LLMs, including GPT-4 [1], have begun to support multimodal input, enabling users to prompt such models with images in addition to textual questions or instructions. For visualization researchers, the potential applications of multimodal LLMs seem endless. For instance, such models could accelerate existing lines of research by being incorporated into systems for chart question answering [28] or into browser extensions for helping users read charts or detect deceptive and misinformation-laden visualizations online [31]. With sufficient knowledge about visualization design principles, multimodal LLMs could even critique visualization designs or serve as education aids [14] for students learning about visualization.

However, given the novel and fast-changing nature of these models, it is difficult to proceed in an informed and responsible manner without understanding the ability of LLMs to read and interpret visualizations – that is, such models' *visualization literacy*. While standard visualization literacy assessments have been developed and deployed for humans (e.g., the VLAT [40]), no such evaluation has yet been published for multimodal LLMs. Thus, the goal of this paper is to evaluate the state-of-the-art GPT-4 model with vision (also known as GPT-4V) on a set of visualization literacy tasks to understand the model's capabilities and establish a baseline level of performance for multimodal LLMs.

As we curated task sets for our evaluation, we sought to leverage existing materials and experiments in the visualization literature to emphasize the relevance and applicability of vision-capable LLMs for work familiar to our research community. In Section 4, we utilize the VLAT [40], a visualization literacy assessment test originally designed for non-expert consumers of data visualizations, to compare GPT-4's visualization literacy to that of humans. In Section 5, we assess GPT-4 on the task of chart question answering (CQA) using an existing system [32] and its released dataset as a point of comparison. In Section 6, we replicate and extend an analysis of common deceptive visualization design techniques [58] to investigate GPT-4's susceptibility to being fooled by such designs. In Section 7, we consolidate experimental conditions across two past studies investigating human responses to visualizations with misaligned titles [34, 35] to see if GPT-4 can discern title-visualization discrepancies, such as titles which are selective or contradictory based on the presented data.

Overall, we find that GPT-4 can perform several tasks when reading and interpreting charts, including recognizing high-level trends, identifying extreme values, and making comparisons between data points. When provided a visualization along with the underlying data, the model can complete and explain complex, multi-step computations with relative ease. The model can also critique visualizations while referencing basic best-practices and potential pitfalls in visualization design, as well as assess the interplay between charts and their titles. On the other hand, GPT-4 struggles with some tasks that are quite easy for humans, including simple value retrieval when not provided the dataset used to produce a visualization. GPT-4 also lacks the ability to reliably distinguish between colors in charts, especially for visualizations like stacked bar charts when many colors may be present at once. The model is susceptible to visual deception using some techniques (e.g., inverting axes) that are effective at fooling humans. Finally, GPT-4's well-documented hallucination and inconsistency issues [2, 12, 39, 80] extend to its vision capabilities [27, 79], though the model seems more prone to such problems on certain tasks compared to others.

In summary, the main contributions of this work are as follows:

- We curate a suite of task sets for evaluating the visualization literacy of multimodal LLMs like GPT-4, drawing from existing work in the visualization community.
- We evaluate GPT-4's visualization literacy using our task sets to understand the state of the art in vision-capable LLM performance on these tasks.
- We release the code, stimuli, and results for our task sets as supplemental material to encourage replication, adaptation, and augmentation of our work on both current and future LLMs.

- *Alexander Bendeck and John Stasko are with Georgia Institute of Technology. E-mails: abendeck3@gatech.edu, john.stasko@cc.gatech.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

- We reflect on this work and outline promising directions for future research, including additional evaluative tasks of potential value and novel applications of LLMs in visualization.

2 Related Work

2.1 Visualization Literacy

As data and data visualizations proliferate, especially online, citizens' ability to read and interpret visualizations – that is, *visualization literacy* – is being noted as an important skill [7,26,51]. Over the past decade, visualization researchers have proposed approaches to both assess [7,8,40] and improve [36,61] human visualization literacy. Much of this work builds upon older fundamental studies on graph comprehension [10], including in education research [18,25,70]. Among the most well-known and widely-used tools for measuring visualization literacy is the VLAT [40], a Visualization Literacy Assessment Test consisting of 53 multiple-choice questions across 12 common visualization types. The VLAT's questions cover a set of analysis tasks previously identified as common in data visualization [3,9].

The authors of the VLAT also provide a concise definition of visualization literacy which is helpful for our use case: "Visualization literacy is the ability and skill to read and interpret visually represented data in and to extract information from data visualizations." [40] Given this definition, we take a broad view of visualization literacy to also include other lines of work which have studied the impact of various visualization design choices on human interpretation. These design choices include deceptive encodings or axes [37,56,58], misaligned or misleading titles [34,35], and visual embellishments or "chart junk" [4,42].

While research has demonstrated that LLMs can compose and comprehend text, we have little understanding of multimodal LLMs' visualization literacy. In this work, we build upon existing research – including the VLAT and other studies mentioned above – to create a suite of LLM visualization literacy task sets and utilize it to assess the vision-capable GPT-4 model.

2.2 Chart Question Answering

Chart question answering (CQA) is a line of research which provides perhaps the closest analog to visualization literacy for systems. A recent EuroVis STAR report on the topic [28] reviews the capabilities of many recent works and defines the overarching problem as follows: "The goal of a chart question answering system is to automatically answer a natural language question about a chart to facilitate visual data analysis." Comparing this definition to that of visualization literacy above, it is not difficult to argue that a capable CQA system could feasibly be considered to have good visualization literacy.

While a handful of CQA systems have been developed over the past half-decade [11,30,32,52,65], in this work we particularly leverage the dataset and pipeline from Kim et al. [32] as a point of reference for assessing GPT-4 on the task of CQA, for several reasons. First, their work is one of only two which utilizes and publicly releases a high-quality set of human-generated questions [28]. Second, their system is again one of only two which supports "open-vocabulary" responses – that is, it can generate responses which do not rigidly follow a predefined vocabulary or template set, and which can provide responses consisting of numbers, words, or sentences [28]. Finally, the researchers who published this work are from the visualization research community, rather than other communities like computer vision, making their work likely more familiar and digestible for this audience.

2.3 LLMs in Visualization Research

Research at the intersection of large language models and visualization can be roughly divided into two types: visualization for LLMs, and LLMs for visualization [75]. The former refers to research where visualization is employed to help users understand and better utilize LLMs. Such work includes using visualization to explain the inner workings of models [20, 43, 44, 76], aid in effective prompt engineering [23, 67, 73, 74], and understand and evaluate model performance [16, 17, 45, 71, 78]. Meanwhile, the second category refers to visualization researchers using LLMs as powerful tools to advance the state of the art in our community. Such work has already employed LLMs to help users create visualizations by generating code [21,50], charts [41,77], or titles [46]; implement more flexible and powerful visual analytics systems [53,60,62,64]; and edit and author data-driven stories or videos [13,63,68,69]. We consider our work to be in support of this second umbrella of "LLMs for visualization". By demonstrating the current abilities and limitations of multimodal LLMs and providing a way to assess these models as they improve, we hope to elucidate their potential utility in future visualization research projects.

2.4 LLM Assessments on Domain-Specific Tasks

LLMs have been assessed on a variety of benchmark tasks in different domains. Studies have demonstrated that models like GPT-3 and GPT-4 perform reasonably well on standardized tests in higher education and processional licensing [54], including the Law School Admission Test (LSAT) [15], Medical College Admission Test (MCAT) [6], and U.S. Medical Licensing Examination [55]. Along these lines, visualization researchers recently found that GPT-4 could score 80% on quizzes and homework from Harvard's CS171 data visualization course [14].

Given their relative novelty, multimodal LLMs have not been assessed as thoroughly. While some prior work has assessed such models' caption generation capabilities [29], the existing work closest to ours is likely HalluisonBench [27], which introduces a diagnostic suite of tasks and stimuli for vision-capable language models. Although some HallusionBench stimuli are charts, both the chart types (dominated by bar charts) and question types (all yes/no) are quite limited. HallusionBench also has no concept of visualization literacy; the focus is instead on inducing LLM hallucinations by modifying factual charts. We craft our suite of task sets to specifically assess visualization literacy on a broader set of tasks and stimuli relevant for our research community.

3 Overall Evaluation Approach

Results for all experiments were collected by running the task sets using the OpenAI API[1] accessed using the official Python library. In particular, we used the `gpt-4-1106-vision-preview` model, which was OpenAI's most advanced vision-capable model that was commercially available during the time period between January 15, 2024 and February 15, 2024 when we ran the task sets. To configure our API calls to achieve as consistent and deterministic behavior as possible, we always set the `temperature` parameter to be 0. However, given the propensity of models like GPT-4 to provide non-deterministic responses in spite of a 0 temperature, each task set was run three times (as in similar prior work [27]). Finally, to ensure task independence and avoid per-minute API rate limits, each task was provided to the LLM individually (as opposed to one large request), and each task's prompt was prepended with any instructions that applied to its task set.

We next discuss the task sets and results for the VLAT (Section 4), Chart Question Answering task set (Section 5), Deceptive Visualizations task set (Section 6), and Visualizations with Misaligned Titles task set (Section 7). In each section, details about the specific prompts and scoring for the corresponding task set are provided as appropriate. Throughout our evaluation, we conducted as little specialized prompt engineering as possible; see Section 8 for more prompting details.

4 Visualization Literacy Assessment Test (VLAT)

4.1 Setup

Our motivation for evaluating GPT-4 using the VLAT was to compare the visualization literacy of GPT-4 to that of humans, while also getting an understanding of the model's performance across a variety of common tasks. To set up the test, we constructed our prompts to the LLM as follows. Each question was provided to the LLM individually, with the answer choices appended at the end and separated by semicolons. Each question was also prepended with adapted preliminary instructions from the real VLAT, specifying basic information like that we are asking a multiple-choice question based on the accompanying image. As in the original test, we stated that the model should skip questions (i.e., answer "Omit") rather than guess. (All prompts, along with stimuli and results, are released as supplemental material.)

[1] `https://platform.openai.com/docs/introduction`

For scoring the results, we sought to penalize inconsistent responses. We therefore only consider GPT-4 to answer a question correctly if its response is correct on all three runs. We also only consider GPT-4 to omit a question if it responds "Omit" on all three runs. In all other cases, we score the overall response for a question as incorrect. This is an important detail because the VLAT scoring scheme penalizes incorrect answers more harshly than omissions.

Tasks. We utilized the final version of the VLAT, consisting of 53 multiple-choice test items which cover 8 task types (see Table 2) and 12 visualization types (see Table 3).

4.2 Results

Overall performance. Using the VLAT scoring scheme from the original paper [40] (which includes a guessing penalty), GPT-4 achieved a score of 19.67. Considering that the human score distribution on the VLAT had a mean of 28.82 and a standard deviation of 8.16, this means that GPT-4 scored quite poorly – just outside of one standard deviation below the human mean. This puts GPT-4's visualization literacy around the 16th percentile of humans. GPT-4 also tended to struggle more with questions that were empirically found to be difficult for humans by the original VLAT authors, as shown in Table 1.

Table 1: VLAT results broken down by question difficulty (as classified in the original VLAT paper), ordered by increasing difficulty.

Difficulty	correct	omit	incorrect	% correct
Easy	12	1	4	70.6
Moderate	10	3	6	52.6
Hard	6	5	6	35.3

Breakdown by task type. As mentioned above, the questions in the VLAT were each associated with an analytic task type taken from existing task taxonomies in information visualization [3, 9]. A breakdown of GPT-4's performance based on the task type associated with each question, provided in Table 2, reveals divergent performance between different task types. While GPT-4 can consistently find trends and can more often than not correctly make comparisons and find extrema, it struggles to retrieve precise values.

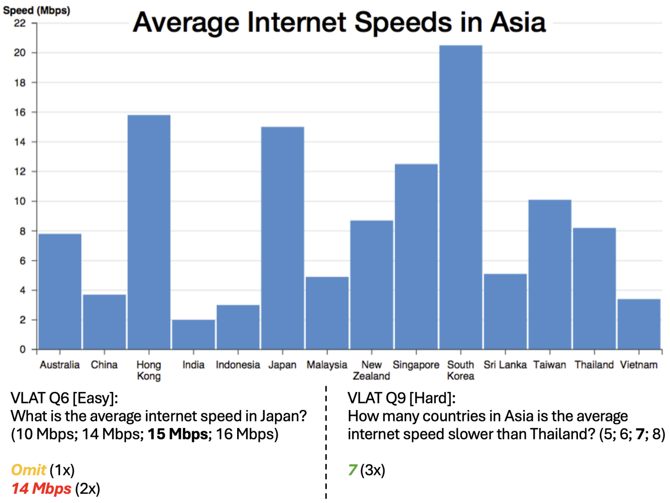
Table 2: VLAT results broken down by task type, ordered in decreasing order by percent correct.

Task type	correct	omit	incorrect	% correct
Identify hierarchy	1	0	0	100.0
Find trends	4	0	1	80.0
Make comparisons	9	1	3	69.2
Find extremum	8	2	2	66.7
Find anomalies	1	0	1	50.0
Determine range	2	0	3	40.0
Find clusters	1	0	2	33.3
Retrieve value	3	6	4	23.1

Breakdown by chart type. The VLAT contains 12 different chart types. While task types are not equally distributed among all charts types, charts generally have some baseline overlap in the types of questions asked (e.g., almost all chart types contained at least one "Retrieve value" question and "Find extremum" question each). A breakdown of GPT-4's performance based on the chart associated with each question is provided in Table 3. Of note is the fact that out of the 9 questions with omitted responses, 8 of these occurred on visualizations which utilized multiple colors to encode the data (stacked bar, 100% stacked bar, stacked area, and pie).

4.3 Discussion

Based on the limitations of GPT-4's vision capabilities which are outlined by OpenAI (as of March 2024) [57], the identified failure cases are not altogether surprising. OpenAI states that the model may struggle with both "graphs or text where colors or styles like solid, dashed, or dotted lines vary" and "tasks requiring precise spatial localization". The combination of these limitations reasonably makes it quite difficult for GPT-4 to retrieve exact values (note from Table 2 that 6 out of 9 omitted responses are for "Retrieve value" questions) and successfully answer questions on visualizations where color is a key encoding (see Table 3). We will see more examples of where misinterpretations of color and poor spatial localization impede GPT-4's ability to correctly answer questions in the next section on chart question answering.



VLAT Q6 [Easy]:
What is the average internet speed in Japan? (10 Mbps; 14 Mbps; **15 Mbps**; 16 Mbps)

Omit (1x)
14 Mbps (2x)

VLAT Q9 [Hard]:
How many countries in Asia is the average internet speed slower than Thailand? (5; 6; **7**; 8)

7 (3x)

Fig. 1: Questions and GPT-4's responses for two items on the VLAT. **Bottom-left:** Incorrect answer to a value retrieval question which humans found easy. **Bottom-right:** Correct answer to an empirically more difficult question requiring multiple comparisons.

However, it may be somewhat unintuitive that in spite of such limitations, GPT-4 can consistently answer some questions correctly which require multiple steps and are arguably (and, based on the original VLAT paper, empirically) more difficult than simple value retrieval. For instance, consider the chart and two questions in Figure 1. Although GPT-4 is unable to do a simple value retrieval based on the height of Japan's bar (Figure 1, bottom-left), it is able to correctly count the number of bars which are shorter than Thailand's (Figure 1, bottom-right). This indicates that in spite of GPT-4's struggles at retrieving exact values, it can perform decently at tasks requiring only rough value retrieval or visual comparison; note again from Table 2 that GPT-4 does much better at making comparisons and finding extrema than retrieving values. In the next section, we present more data on GPT-4's performance across different analytic tasks, including when the dataset underlying the visualization at hand is provided to the model.

Table 3: VLAT results broken down by chart type, ordered in decreasing order by percent correct.

Chart type	correct	omit	incorrect	% correct
Treemap	3	0	0	100.0
Line	4	0	1	80.0
Stacked area	4	2	0	66.7
100% stacked bar	2	1	0	66.7
Pie	2	1	0	66.7
Bubble	4	0	3	57.1
Scatterplot	4	0	3	57.1
Bar	2	0	2	50.0
Histogram	1	1	1	33.3
Choropleth map	1	0	2	33.3
Area	1	0	3	25.0
Stacked bar	0	4	1	0.0

5 Chart Question Answering

5.1 Setup

In this section, we compare GPT-4's ability to answer questions about visual charts with that of an existing chart question answering pipeline – namely, that of Kim et al. [32]. In addition to the reasons for choosing their work outlined in Section 2.2, the setup of their question answering pipeline also affords utility for testing GPT-4's capabilities. Specifically, their pipeline has 3 stages: (1) data extraction, (2) question processing and answering, and (3) explanation generation. We thus evaluate GPT-4 in two ways: as a replacement to the entire pipeline, and as a stand-in for Stages 2 and 3 after data extraction. While we briefly discuss GPT-4's explanations in Section 5.3, this is not a main focus of our work. We then compare the performance of GPT-4 when it is directly provided the data underlying the charts to when it is not.

Before assessing GPT-4 on the released question set from Kim et al., we sought to recreate the question categorization outlined in their paper. The paper describes each question as either "lookup" (single value retrieval) or "compositional" (needing several operations), and as either "visual" (referencing specific visual features of the chart such as color, length, etc.) or "non-visual". However, the released question set does not contain these categorizations. We thus contacted the corresponding author via email; while he was unable to find the precise labels for each question, he did provide us with a more granular breakdown of question category counts for each chart. We utilized this breakdown as a guide to manually re-categorize all questions. While we were unable to precisely match the categorization in the original paper, our count differed only in that we counted three more "non-visual lookup" questions and correspondingly three fewer "non-visual compositional" questions. The deviation from the initial category counts is thus only 3, a very small percentage of the 629 total questions. The count of questions according to our categorization is provided in Table 4.

Table 4: Breakdown of questions in our set based on categorization of "lookup" vs. "compositional" and "visual" vs. "non-visual".

	# Questions		
	Lookup	Compositional	**Total**
Visual	52 (8%)	24 (4%)	**76 (12%)**
Non-Visual	141 (22%)	412 (66%)	**553 (88%)**
Total	**193 (30%)**	**436 (70%)**	**629**

To run the task set, we then constructed our prompts to the LLM as follows. Each question was provided to the LLM individually. Each question was also prepended with instructions specifying that we are asking a question based on the accompanying image (and, in the data-provided condition, the accompanying dataset) and to answer based only on this information and not external knowledge. For the results, we again only consider GPT-4 to have answered a question correctly if its response was correct on all three runs. For questions with non-numeric answers, responses needed to be exactly correct. For questions with numeric answers, we considered responses to be correct if within 5% of the correct answer, as in prior work [52].

Tasks. The question set from Kim et al. consists of 629 questions across 47 bar charts (32 simple, 8 grouped, 7 stacked) and 5 line charts.

5.2 Results

Overall performance and breakdown by question type. When provided the underlying data along with the visualization, GPT-4 is able to achieve an accuracy of 87% on the CQA dataset from Kim et al., which outperforms their system quite significantly. However, when not provided the underlying data, GPT-4's performance degrades notably to 31% overall, well below Kim et al.'s system. Figure 2 reveals stark performance differences between the different question categories. Among both categories of "compositional" questions, GPT-4 with data is able to double the accuracy of prior work, and even GPT-4 without data performs comparably to prior work. Recall from Table 4 that the majority of questions in the set are "non-visual compositional", and so a large part of GPT-4's better overall performance (when provided the data) can be attributed to this category. On "visual lookup" questions, GPT-4 with data still underperforms Kim et al.'s system, though its performance is competitive. On "non-visual lookup" questions, GPT-4 achieves nearly perfect accuracy when provided the data, but posts a poor 13% accuracy when it is not – echoing the model's struggles with value retrieval from the previous section.

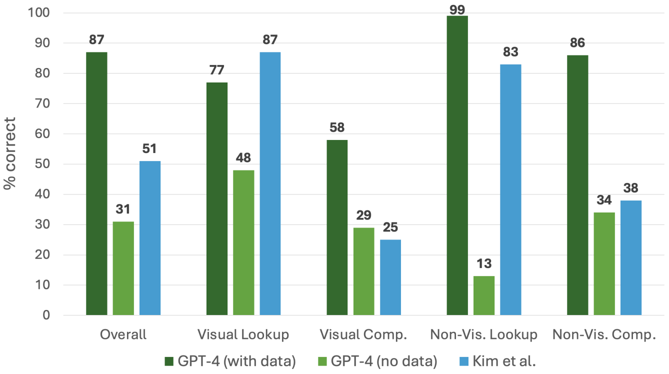


Fig. 2: CQA accuracy of GPT-4 with and without data provided, compared to the 3-stage pipeline from Kim et al.

Breakdown by task type. The questions from Kim et al. are not all associated with a specific analytic task like the questions in the VLAT. However, the more granular question breakdown provided to us by the corresponding author did specify a few sub-categories of tasks for "compositional" questions which allow for some high-level comparison to the results from the VLAT. Table 5 shows the accuracy breakdown for "lookup" tasks and for four "compositional" task types: compute derived value, find extrema, make comparisons, and "multiple" (meaning some combination of the former – e.g., computing the difference between maximum and minimum values).

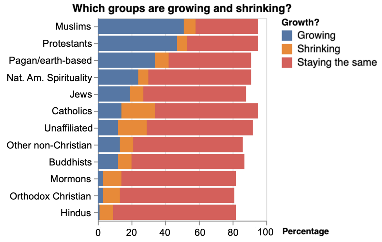
Table 5: CQA accuracy of GPT-4 with and without data on various tasks.

Task	# Questions	Accuracy w/ data	Accuracy w/o data
Compute derived value	125	96%	7%
Lookup	193	93%	23%
Find extrema	267	87%	52%
Make comparisons	25	84%	44%
Multiple	70	69%	37%

For the tasks with the highest accuracy given data – i.e., lookup and computing derived values – the dropoff in accuracy when data is no longer provided is quite significant. For the other tasks, however, performance does not degrade so severely without data. Even questions requiring "multiple" operations had a higher accuracy than lookup when not provided data. For instance, further inspection revealed that GPT-4 was sometimes still able to correctly make comparisons between extreme values even without data. Note that the accuracy without data for tasks which overlap with those from the VLAT looks relatively similar to the previous section, with accuracy finding extrema and making comparisons (which only require rough value retrieval) outpacing performance on precise value lookup.

5.3 Discussion

Visual lookup and color. From Figure 2, "visual lookup" is the only question category where Kim et al.'s system outperforms GPT-4 with data. Inspection of failure cases revealed that a large proportion of these occurred due to GPT-4's inability to consistently interpret multiple colors in charts. While on some occasions the model was confused by very close color hues or simply hallucinated issues with chart legends, GPT-4 also showed a tendency to misread relatively unambiguous color legends (see Figure 3). Such issues were consistently replicated across all three runs, regardless of whether data was provided or not (since the data does not contain this visual encoding information).



CQA Q85:
What is the percentage of red Protestants?

The red portion [...] indicates the percentage that is ***shrinking****. According to the provided data, 6% of Protestants are in the shrinking category.*

Fig. 3: GPT-4 consistently fails to correctly interpret the color legend on this stacked bar chart. Relevant text bolded for emphasis.

We note that while Stage 1 of Kim et al.'s pipeline did include a component to extract data encodings, we did not utilize this in any runs on GPT-4. Based on our VLAT experiment, we already knew GPT-4's limitations with value retrieval, and so we utilized the data extraction component of Kim et al.'s pipeline to compensate. However, we wanted to observe the limitations of GPT-4 in other areas such as interpreting visual encodings so we could judge the model's visualization literacy capabilities, as opposed to simply measuring its power as a question-answering engine with all data and encodings provided cleanly as input.

Axes and titles. One particular strength of GPT-4 is in reading numeric axes and titles of visualizations. The model correctly answered all 18 questions about axes and titles consistently across all runs. Figure 4 shows a failure case from Kim et al. which GPT-4 can correctly answer with ease: interpreting the increments along the y-axis of a chart. As opposed to color, meta-details about chart axes and titles which are not included directly in the dataset appear to be relatively easy for the model to understand.

CQA Q268:
What number of years is used as increments along the y-axis?

Kim et al.'s system:
19. I counted the number of bars.

GPT-4:
The increments along the y-axis are in ***5-year intervals****.*

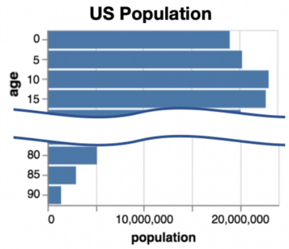


Fig. 4: GPT-4 can easily succeed on some failure cases from Kim et al., especially those related to reading axes.

Generating (visual) explanations. Beyond just question answering, a large part of Kim et al.'s paper focused on explanation generation – how their system can reference specific visual features in the chart to explain its answer. We explore this angle only briefly, since properly evaluating GPT-4's explanations would likely entail a more manual and subjective analysis than we aim to undertake in this work. GPT-4's explanations for its answers occasionally made little to no reference to the chart at all, simply answering in the form "Based on the data provided, the answer is X". Conducting a meaningful evaluation of GPT-4's ability to provide visual explanations when answering questions would likely require somewhat extensive prompt engineering to nudge the model towards producing the types of explanations that are desired.

However, it is worthwhile to note that GPT-4 did occasionally demonstrate the potential to generate intuitive and meaningful explanations for its answers during our experiments. Consider the chart and question in Figure 5 along with the model's visual explanation. In this case, GPT-4 is able to provide quite a good and easily verifiable explanation for its answer to a simple extremum identification question. Future work can consider testing the limits and common failure cases for LLMs generating visual explanations in a CQA setting and should take a cautious approach given the potential for hallucination.

CQA Q349:
What month has the most sun?

Based on the provided chart, ***August*** *has the most sun, with the* ***yellow segment*** *(representing sunny days) being the* ***tallest in that month's column****.*

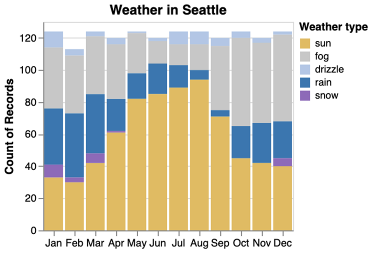


Fig. 5: GPT-4 shows promise in providing visual explanations when answering questions about charts. Relevant text bolded for emphasis.

Additionally, recall from Figure 2 that GPT-4's biggest performance win over Kim et al. is on "compositional" questions requiring multiple operations. For such questions, GPT-4 was generally able to provide quite detailed and easy-to-follow steps to compute the correct answer, even if these explanations rarely referenced visual chart features. This ability of the model to walk a user through multi-step computations, while already well-known [24], further increases the value of such models as force multipliers in chart question answering research.

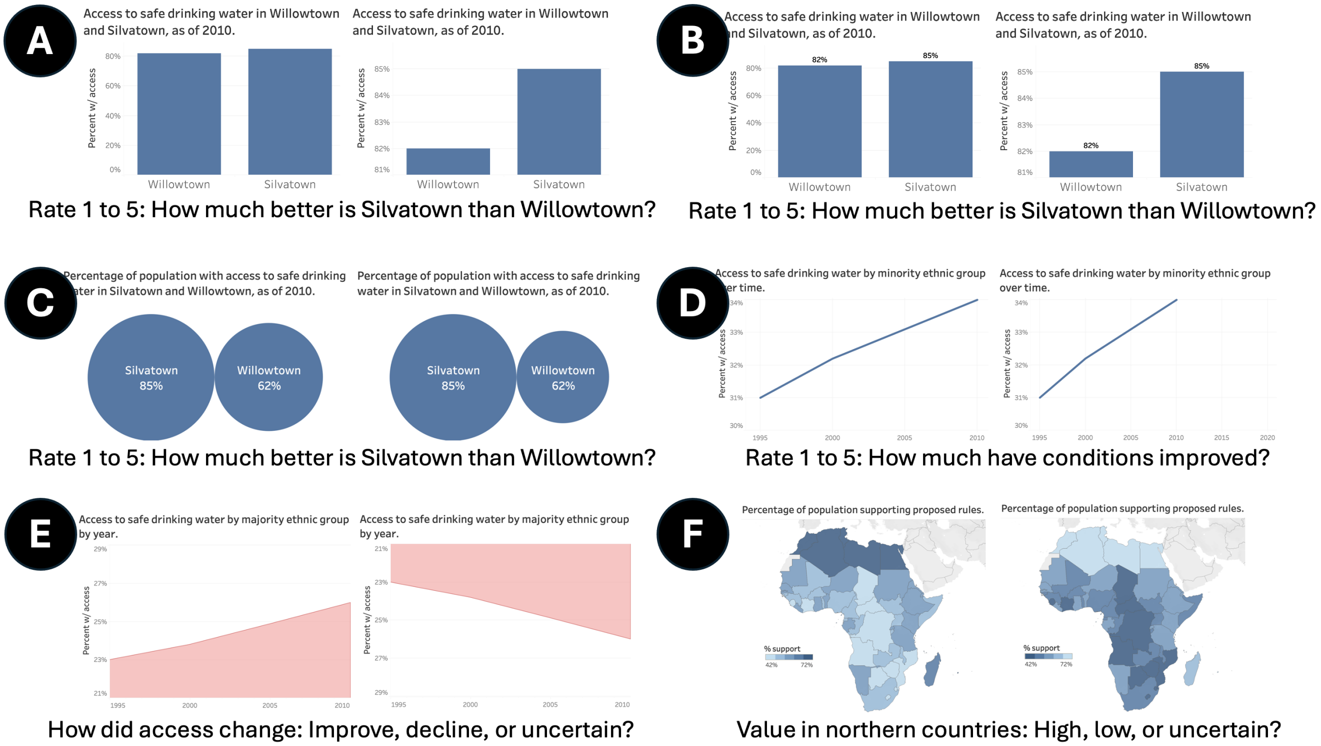
6 Deceptive Visualizations

6.1 Setup

In Section 4, we compared the ability of GPT-4 to answer questions about visualizations to that of humans. However, the VLAT is not designed to evaluate one's ability to answer questions about *deceptive visualizations* – that is, visualizations whose authors have made specific design choices to deceive viewers. As mentioned above, prior work has studied the effects of common distortions in visualizations and found a large effect on readers' takeaways [58]. In this section, we investigate whether GPT-4 can be misled by similar design tactics which have been shown successful at deceiving humans.

Our first step in curating this task set was to replicate the visual stimuli used by Pandey et al. [58], as their study materials were not released. We designed our stimuli to be as close as possible to those depicted in their paper. We also augmented those stimuli with two additional control-deceptive chart pairs, resulting in the following list:

- **Truncated axis** (Figure 6-A): The visualizations are bar charts with two bars each. In the control condition, the y-axis starts at 0. In the deceptive condition, the y-axis starts at a much higher values to exaggerate the differences between the bar heights.
- **Truncated axis, with data labels** (Figure 6-B): This chart pair is equivalent to the first, except exact bar height values are written on top of each bar for both conditions. Such a stimulus was not evaluated in the original paper, but we decided to include it given our observations about GPT-4's poor value retrieval capabilities.
- **Area as quantity** (Figure 6-C): The visualizations are very simple "bubble charts" where two circles are shown with different sizes and the numeric values shown inside each circle. In the control condition, circle area encodes the value. In the deceptive condition, circle radius encodes the value, exaggerating the size difference.
- **Aspect ratio** (Figure 6-D): The visualizations are line charts. Compared to the control condition, the deceptive condition shows the same data, but extends the x-axis to adjust the chart aspect ratio so that the slope of the line looks steeper.
- **Inverted axis** (Figure 6-E): The visualizations are area charts. In the control condition, the y-axis increases as it goes up. In the deceptive condition, the y-axis is inverted and increases as it goes down, to make it look like the quantity being depicted is actually decreasing.



Rate 1 to 5: How much better is Silvatown than Willowtown?

Rate 1 to 5: How much better is Silvatown than Willowtown?

Rate 1 to 5: How much better is Silvatown than Willowtown?

Rate 1 to 5: How much have conditions improved?

How did access change: Improve, decline, or uncertain?

Value in northern countries: High, low, or uncertain?

Fig. 6: Visualization stimuli for deceptive design task set. For each pair, the control condition is shown on the left and the deceptive condition on the right. Underneath each pair is a simplified version of the main message question which was asked to see if GPT-4 could be “fooled”.

- **Inverted color ramp** (Figure 6-F): The visualizations are choropleth maps. In the control condition, the map has a single-hue color ramp where darker colors indicate higher values. In the deceptive condition, the color ramp is inverted so lighter colors indicate higher values. This type of deception was not evaluated in the original paper, but we decided to include it given our observations about GPT-4’s poor performance discerning colors.

Tasks. Like Pandey et al., we posed questions about the main message for each of the 12 individual charts to gauge GPT-4’s judgement on either the magnitude or the direction of the effect shown in the chart, depending on the distortion type. Figure 6 shows a shortened (for the figure) version of the question asked for each visualization type. The actual questions were worded as similarly to those in the original paper as possible. We were additionally interested in whether GPT-4 could detect deceptive design choices when prompted to do so. To achieve this, we added two additional question types. First, for each of the 12 charts, we asked “Are any misleading design tactics being used in this chart?” Second, we presented each of the 6 chart *pairs* and asked, “Do these two charts, displayed side by side, show the same data?”

6.2 Results

Truncated axis. When asked to rate the difference between the two bars, GPT-4 was fooled by the truncated axis. The model consistently rated the difference between the bars as a 1/5 for the control condition and a 5/5 for the deceptive condition. In spite of this, the model was able to detect the misleading design tactic in the deceptive condition, describing it as follows: “Truncated Y-Axis: The y-axis starts at 81% instead of 0%, which exaggerates the difference [...] between Willowtown and Silvatown.” Likewise, when shown the two visualizations side-by-side, GPT-4 identified the charts as equivalent and mentioned the axis distortion on one of the charts.

Truncated axis with data labels. When asked to rate the difference between the two bars with data labels, GPT-4 still appeared to be influenced by the truncated axis, though not as strongly. The model consistently rated the difference between the bars as a 1/5 for the control condition as before, but now only rated a 2/5 for the deceptive condition. The model was again able to detect the misleading design tactic in the deceptive condition and identify the charts as equivalent when side-by-side.

Area as quantity. GPT-4 was not influenced by the size distortion in the bubble chart. The model rated the difference between them as a 4/5 or 5/5 interchangeably for both conditions across the repeated runs. The model was wary of the pitfalls of this technique, noting that perception of exaggerated differences “is a common issue when using circle sizes to represent quantities”. However, it erroneously detected this distortion in both the deceptive *and* control conditions. When shown the two visualizations side-by-side, GPT-4 identified the charts as equivalent and mentioned the size distortion. However, its assessment of the difference between the charts was inconsistent, sometimes identifying the the deceptive chart as being more accurate and once hallucinating that the control condition showed two circles of the same size.

Aspect ratio. GPT-4 was not at all influenced by the aspect ratio distortion. The model consistently rated the change in conditions over time as a 2/5. The model did not detect the aspect ratio distortion when asked to assess the deceptive condition for misleading tactics (though it did mention the truncated axis). It correctly assessed the juxtaposed charts as showing equivalent data and did not mention the aspect ratio difference, but hallucinated that both charts had “the same time frame on the x-axis (from 1995 to 2020)” even though the control condition chart ends at 2010.

Inverted axis. Out of all the deceptive conditions, GPT-4 was most thoroughly fooled by the inverted axis distortion. The model consistently stated that the data was increasing in the control condition and decreasing in the deceptive condition. When prompted to assess the charts for misleading tactics, the model mentioned the truncated y-axis for both conditions, but ignored the axis inversion in the deceptive condition. GPT-4 also assessed the charts as not equivalent when side-by-side, stating the following: “The chart on the left shows an increasing trend over time, while the chart on the right shows a decreasing trend over the same time period.”

Inverted color ramp. GPT-4 seemed generally not fooled by this distortion, consistently identifying the values in northern Africa as high in the control condition and doing so 2 out of 3 times on the deceptive condition (responding "uncertain" once). However, the model's response to the other two questions suggests quite severe hallucination when viewing these charts. The model made several critiques of the visualizations when prompted, but did not identify the deceptive condition to have a misleading color ramp, in fact stating that the deceptive visualization had a color ramp with "darker shades indicating higher support". Even when shown the maps side-by-side, the model concluded that the "color coding for each country is identical in both maps".

6.3 Discussion

Overall, the results presented in this section indicate that GPT-4 can be fooled, or at least confused, by some common deceptive visualization techniques that have also been found misleading for humans (see Figure 7). GPT-4's robustness to visual deception depends heavily on the task at hand; for instance, although GPT-4 can detect a truncated axis in a bar chart as misleading when prompted, it is also susceptible to being misled when asked to draw conclusions the data being presented in such a chart. The model also shows knowledge of visualization design pitfalls such as encoding values using circle size, even if it is not always able to reliably assess the problem when shown a visualization. Finally, we see again that hallucination is a limitation of GPT-4 in its ability to interpret visualizations, especially when color is utilized as a key encoding like in the choropleth maps.

Deceptive tactic	Fooled?	Detects (main) tactic?	Identifies chart equivalence?
Truncated axis	Yes (large effect)	Yes	Yes
Truncated axis w/ labels	Yes (small effect)	Yes	Yes
Area as quantity	No	Yes, but also detects for control	Yes*
Aspect ratio	No	No	Yes*
Inverted axis	Yes	No	No
Inverted color ramp	No	No*	Yes*

Fig. 7: Summary of results across deceptive design tactics. An asterisk (*) indicates that the response(s) contained evidence of hallucination.

7 Visualizations with Misaligned Titles

7.1 Setup

Visualizations do not need to use deceptive visual encodings or axes to mislead the viewer. Text that accompanies a visualization, whether as a title, subtitle, or in an accompanying article or social media post, can have a large impact on the reader's takeaways [33, 66] and can potentially be used to mislead. LLMs like GPT-4 could be powerful tools to help detect misalignment between visualizations and accompanying text, given their demonstrated ability to evaluate and improve text written by humans [19, 48]. In this section, we focus specifically on deceptive visualization titles, assessing the ability of GPT-4 to detect title-chart misalignment across several conditions previously studied by Kong et al. [34, 35] and making use of the released experimental materials from their work.

In curating the stimuli for these tasks, we consolidated conditions from across two papers by Kong et al. which studied human assessments of misaligned titles. Their first paper [34] focused on **selective** titles, such as in the visualization shown in Figure 8, where multiple phenomena are shown simultaneously in a chart but the title only makes reference to one while ignoring the other. Their second paper [35] extended the first to additionally study charts with **miscued** titles, which make reference to a visually de-emphasized phenomenon in the chart, and **contradictory** titles, which do not correspond to the sole phenomenon shown in the chart. We utilized these conditions, along with a **control** condition where the title properly corresponds to the sole phenomenon in the chart, in this experiment. Using the stimuli released as supplemental material by Kong et al. and making any modifications as necessary, we ended up with 4 visualizations for each condition (i.e., 4 visualizations with control titles, 4 with selective titles, etc.).

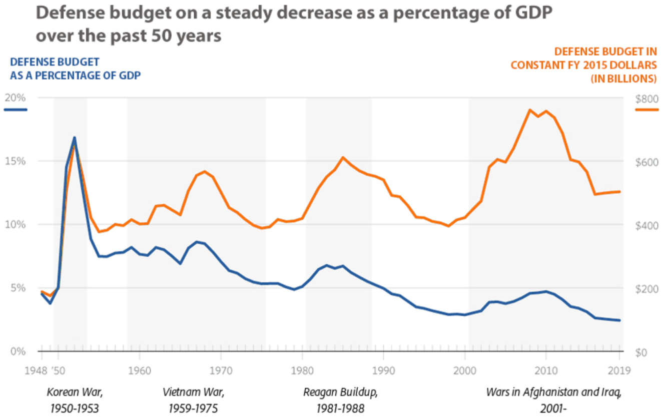


Fig. 8: One example of a visualization where the title is **selective**, referencing only the blue line. Keeping the title the same but slightly modifying the visualization would create the other conditions. Without the orange line, this would be a **control** condition; with the blue line grayed out and dotted but still present, this would be a **miscued** condition; and without the blue line, this would be a **contradictory** condition.

Tasks. For each of our 16 stimuli, we asked two questions taken from the surveys completed by participants in Kong et al.'s second paper: "Do you find the title appropriate for this visualization?" and "Does the title tell the whole story?" We initially planned to only ask the first question to compare GPT-4's assessments with those of humans, most of whom found titles appropriate even in the contradictory condition. However, during the development of this experiment, we found meaningful differences in the results based on the question wording which highlight the importance of prompt engineering in utilizing LLMs to interpret visualizations. Note that the results below outline our observations for each condition across all visualizations, but we will repeatedly refer to Figure 8 as a running example.

7.2 Results

Control. We were somewhat surprised by the results of the control condition. While our expectation was that the control visualizations (Figure 8 without the orange line) would generally get a stamp of approval from GPT-4, this was not the case. Both when we asked whether the titles were "appropriate" and "told the whole story", GPT-4 consistently took issue with some aspect of the control titles. For instance, in the control condition corresponding to Figure 8, the model objected to the characterizing the decrease as "steady" given that "there are periods of increase and fluctuation". To test this sensitivity, we then changed the word "steady" in the title to "general", but this still did not satisfy GPT-4. Finally, we prompted GPT-4 to write its own title for the visualization and put this title ("Trends in Defense Spending as a Percentage of GDP Over the Last 70 Years") on top of the chart, which earned approval from the model. Note that we did not keep these modified titles, only using them briefly as a sensitivity test.

Selective. In the selective condition, we saw a notable difference in the results depending on whether we asked if the title was "appropriate" or told the "whole story". When asked about appropriateness, GPT-4 would generally focus on similar minutiae as in the control condition and miss the point of the title being selective. However, when we asked whether selective titles told the whole story, the model almost always identified the main issue with the title and provided a relevant explanation. Consider the following part of a response for the visualization in Figure 8: "[...] the chart shows two lines: one representing the defense budget as a percentage of GDP (blue line), and the other showing the defense budget in constant FY 2015 dollars (orange line). [...] the title does not reflect the complexity shown in the chart, where the actual defense budget in constant dollars does not always decrease."

Miscued. The miscued condition produced extremely similar results to the selective condition. GPT-4 still often missed the main point when we asked about appropriateness, but identified it when we asked about telling the whole story. Indeed, regardless of the prompt, the model showed little understanding of how the visual misdirection itself could contribute to the title misalignment. Responses did not seem to indicate that the miscued component of the visualization was being visually called out or emphasized in any way – only that it was, for instance, a "solid" line compared to a "dashed" line.

Contradictory. In the contradictory condition, the difference in responses between asking about appropriateness or telling the whole story mostly disappeared. Although the model could still get distracted by smaller issues, the absence of a second phenomenon in the data to fixate on resulted in the assessments of appropriateness focusing more on the contradiction between the title and the displayed data. For instance, for the contradictory version of Figure 8 (without the blue line), the model concludes that the title is not appropriate since the graph "shows fluctuations in the defense budget in constant FY 2015 dollars".

7.3 Discussion

Overall, GPT-4 appears much more sensitive to potentially inappropriate titles than humans. Consider that in Kong et al.'s second paper, 79% of survey responses for miscued conditions and even 60% of responses for contradictory conditions considered the titles to be appropriate. By contrast, GPT-4 even found control titles to be inappropriate due to subtleties in wording. (We note that the model is much less sensitive to the type of subtle *visual* manipulation employed in the miscued conditions, never mentioning the clear visual emphasis placed on the data which did not support the title.) In some sense, we can consider the model's flagging of inappropriate titles to have extremely high recall, but relatively low precision. If there is even a small issue with a title, it will likely be found, but identified issues may not be of extreme interest or relevance. In particular, we saw that even with selective and miscued titles, GPT-4 often explained that titles inappropriate without mentioning the part of the visualization being ignored by the title.

The results for the selective and miscued conditions also showcase the importance of prompt engineering in utilizing LLMs like GPT-4. For both conditions, we consistently observed meaningful differences in responses based on whether we asked if the title was "appropriate" or if it told "the whole story". When asked the former, GPT-4 would often fixate on relatively small issues with the title and ignore the fact that the title was ignoring half of the visualization. When asked the latter, the model showed the ability to thoroughly and clearly explain the discrepancy between the title and the visualization. The impact of prompt engineering when utilizing LLMs is well-documented [23, 67, 73, 74], and the results of this section further demonstrate the importance of composing specific prompts that are highly tailored to the task at hand.

8 Sensitivity Analysis

To ensure that our results are reasonably general and are not simply artifacts of particular visual stimuli, questions, or prompts, here we include a sensitivity analysis. It consists of two main parts: (1) a replication study where we created similar stimuli to those used in the original task sets to test the potential effects of GPT-4's knowledge base; and (2) a brief discussion of how our prompts were constructed, as well as a few additional prompt variation tests to investigate the effects of different prompts for each task (using the original stimuli).

Replication study. For the VLAT, we created 12 charts and 53 questions of the same types as the published test, but using original and artificially generated data. Results indicated some sensitivity of GPT-4 to the particular multiple-choice responses given, especially for value retrieval tasks. However, in general the model performed similarly as on the real VLAT and struggled on the same types of questions as reported earlier. For the CQA task, we created two charts with synthetic data (one regular and one stacked bar) and wrote 40 questions across the categories of lookup, finding extrema, making comparisons, and computing derived values. The accuracy on these questions closely mirrored that reported in Table 5. For the deceptive design task, we re-created stimuli A through E from Figure 6 (as F was not in the original work and thus could not be in GPT-4's knowledge base) with new, synthetic data. The results were almost identical to those reported in Figure 7 except that the model was able to correctly detect the inverted axis deceptive tactic on our chart. For the title misalignment task, we created a new set of bar chart visualizations with control, selective, miscued, and contradictory conditions. While there seemed to be a bit more hallucination by the model in the test on our title misalignment stimuli, the results were comparable to the main task. GPT-4 still sometimes objected to "correct" titles and was relatively sensitive to question wording of whether the title was "appropriate" or "told the whole story". Overall, we found little evidence that GPT-4 had benefited from its knowledge base during the main task sets as the model's performance was similar on our original stimuli with original, synthetic data. All stimuli, questions, and results for the replication study are released as supplemental material.

Prompt variations between task sets. For each task set, we had to briefly iterate on the prompt format before we saw decent results. The prompts for all task sets followed a similar pattern wherein a preface was prepended to each question and associated image and then fine-tuned for each task set. (Note that the questions themselves were not engineered as they were taken verbatim from prior work.) The basic preface template was: "I am about to show you an image and ask you a question about that image. Answer as best you can based only on the chart and not external knowledge." While we did not find a major difference between responses when we added the directive to answer based only on the chart, we kept it as a precaution. We found that the vanilla prompt preface worked well for the deceptive design and title misalignment tasks. To achieve reasonable performance on the VLAT, we had to add that the question would be multiple-choice and append the instructions from the actual VLAT so the model would omit answers when appropriate rather than guess. For the CQA task set, we had to mention that a CSV or JSON dataset associated with the chart would be included (when appropriate). We also asked the model to "provide a visual explanation" for CQA answers as a test of its abilities. While GPT-4 did not always provide visual explanations (and just referred to the data), this wording generally produced good results.

Prompt variation tests. We conducted a brief series of follow-up prompt variation tests across the task sets to determine the effect of various changes on the responses. Several tests were not intended to induce any changes in model behavior; these included switching the order of the question and image in the prompt and changing delimiters between answer choices (VLAT only). These changes seemed to have little impact on the correctness of responses for each question, but sometimes led to noticeable variations in the incorrect or hallucinatory responses. Other tests involved giving the model an unreasonable or ill-formed task, such as by providing a contradictory dataset and chart (CQA only) or a question which referred to entities absent from the chart or data. In these types of tests, GPT-4 usually noticed the inconsistency and refused to directly answer the question, though on rare occasions it instead produced an erroneous or hallucinatory answer.

9 General Discussion

9.1 Results Summary

Strengths of GPT-4. Even when not provided the data underlying a visualization, GPT-4 was able to complete several common analytic tasks, namely recognizing high-level trends, finding extrema, and making comparisons between values. This was observed on both the VLAT and the CQA question set, although the accuracy for such questions was a bit lower on the CQA questions (without data), which may be more indicative of the model's abilities given the larger sample size. When provided data on the CQA questions, GPT-4 showed remarkable reliability at retrieving values as well as performing and explaining computations, even ones which required multiple steps. Faced with visualizations that utilized deceptive tactics, the model showed some basic knowledge of visualization design best practices, including the potential of truncated axes on charts to mislead viewers and the pitfalls of using circle size as a quantitative variable encoding. When asked to assess chart titles at varying levels of misalignment, GPT-4 also showed the ability to make nuanced and subtle assessments of these titles and their relationships with the charts (depending on the prompt).

Weaknesses of GPT-4. A primary limitation of GPT-4's visualization literacy is its inability, or even occasional unwillingness to attempt, to retrieve values from a chart when not given the underlying data. Precise value retrieval is an understandably difficult task even for humans when axis intervals are large. However, on several occasions GPT-4 could not even provide ballpark values, or failed on exact value retrieval tasks that would be reasonable for a human (recall Figure 1). This is exacerbated by the model's other main visual limitation: its inability to reliably distinguish between colors (recall Figure 3). In addition to the VLAT and CQA sections, issues with color also surfaced during the deceptive design tasks, when the model hallucinated that two choropleth maps with inverse color ramps were identical. However, we note that it is difficult to completely isolate issues of value retrieval and even color interpretation (e.g., reading a color legend) from the model's observed struggles with spatial reasoning more generally [72]. The deceptive design evaluation further revealed that GPT-4 can be fooled by common distortions such as truncated and inverted axes – even though it recognizes the former as a deceptive technique when prompted. In both the deceptive design and title misalignment sections, although GPT-4 showed an ability to evaluate visualizations with a critical eye, it would sometimes make nitpicky judgements or fixate on details that paled in comparison to a more pressing design issue. Finally, across all task sets, the model demonstrated hallucination and inconsistency issues. These were most numerous during the CQA task set, and included but were not limited to: providing different answers to almost identical questions; answering questions with words or items that did not appear in the visualization; erroneously noting discrepancies between the visualization and the provided data; answering questions as though bars in a chart were sorted in ascending or descending order, even though they were not; and incorrectly reading color legends.

9.2 Limitations of Our Approach

Missing visualization types and task sets. While we aimed to cover a reasonable set of visualizations, our task sets are by no means exhaustive. In particular, we do not include any geographic visualizations besides a few choropleth maps, and we do not include any graph or network visualizations (which have their own associated analytic tasks [38] beyond those assessed here). In order to make reasonable comparisons between our results and those of the prior studies on which we built our evaluation, we also leaned towards choosing task sets from work which either released much of their materials or utilized stimuli which were easy to replicate. Regardless, we believe that our selected task sets provide sufficient breadth and depth to constitute a useful contribution – comparing GPT-4 to both human and machine question answering on charts, and then measuring the model's robustness in the face of both visual and textual deception techniques.

GPT-4's extensive knowledge base. The existing works upon which we based this evaluation (e.g., the VLAT) and some real-world datasets or phenomena which were visualized in the chart stimuli are several years old. GPT-4 could thus feasibly know of these research works and/or datasets and use such knowledge in completing our task sets. We attempted to mitigate and detect any such issues along this vein through our sensitivity analysis (Section 8), which showed little evidence of GPT-4 benefiting from prior knowledge of the tasks or stimuli. Based on these considerations, we ultimately judged that the benefits of being able to compare GPT-4's performance to humans and existing systems by utilizing prior work outweighed the potential drawbacks.

LLMs as a moving target. LLMs like GPT-4 can be quite fickle and inconsistent in their behavior [19, 48]. We aimed to avoid extensive prompt engineering to keep the tasks as similarly worded as possible to the original works which inspired them. However, to achieve reasonable results, it was sometimes necessary to manipulate question choices (e.g., to detect misaligned titles in Section 7) or prompts (see Section 8). Future work can further test the effects of prompt engineering and follow-up prompting on LLM visualization literacy task performance. Furthermore, despite setting the model `temperature` parameter to 0 in our API calls and running each task three times to maximize replicability and reliability of our results, there are still inherent difficulties in attempting to generally characterize the behavior of models like GPT-4. Finally, GPT-4 is only one multimodal LLM, and we fully anticipate that newer and more powerful models will soon succeed it. Although this work only represents a snapshot of one model at this point in time, we believe there is significant value in (1) assessing GPT-4's performance to better understand the current state of the art, and (2) releasing the materials of our suite of task sets to facilitate the tracking of vision-capable LLMs as they improve over the coming years.

9.3 Future Work

Evaluating visualization literacy of other LLMs. A clear direction for future work is leveraging our released task sets to evaluate the visualization literacy of multimodal LLMs besides GPT-4, including newer GPT models. Other models including Google's Gemini have already been evaluated on some CQA tasks [59], but not on visualization literacy tasks more broadly. We note that as proprietary models, the exact training data input and internal parameters for GPT-4 and Gemini have not been released. This makes it difficult to investigate the internal workings of the models and determine why they behave the way they do. For this reason, we especially advocate for future work to evaluate models with architectures that are open-source, such as LLaVA [47] and Fuyu-8B [5]. Using open-source models would allow researchers to tinker with model parameters and training to assess the impact of such changes on visualization literacy task performance. While these models are often introduced and evaluated on more general visual question answering tasks, their performance on visualization literacy tasks remains unexplored. We encourage the creation of a visualization literacy leaderboard for LLMs, akin to the MathVista mathematical reasoning benchmark [49], to continue the systematic study of models' capabilities. This would also help establish an understanding of the relative strengths and weaknesses between different models regarding visualization literacy. Finally, although our evaluation and its released materials may be incorporated into the knowledge bases of future LLMs, we believe our methodology can provide a lasting roadmap to evaluate such models for their use in visualization research.

Leveraging LLMs in visualization research. Based on the capabilities of GPT-4 as observed throughout our evaluation, we have identified a few potentially fruitful applications of LLMs in visualization research. As we saw in Section 5, GPT-4 is powerful at answering many types of questions about charts, especially compositional questions requiring multiple steps. Given that Kim et al. [32] cite limitations with the Sempre question answering engine they employed as a bottleneck on the performance of their system, GPT-4 at a minimum provides a powerful new backbone for CQA systems. We also found in Section 6 that GPT-4 has basic knowledge of visualization design best practices and can detect when visualizations fail to adhere to these guidelines. We believe that once LLMs' tendency to hallucinate becomes less frequent, they could be employed in web browser extensions [22], design tools, or education aids [14] to help different users create and consume visualizations. Finally, given GPT-4's ability to generate and assess visualization titles (Section 7), LLMs can be used to develop automatic chart captioning and titling systems [46].

10 Conclusion

In this paper, we empirically evaluate the visualization literacy of the GPT-4 multimodal large language model (LLM) on a suite of task sets based on prior work in the visualization community. The suite includes a basic visualization literacy assessment test, a chart question answering task set, a section on assessing deceptively designed visualizations, and a section on detecting title-chart misalignment. We find that GPT-4 performs well at identifying trends and extreme values, and also has some knowledge of visualization design best-practices, but struggles with retrieving values and distinguishing colors, and occasionally suffers from inconsistency and hallucination. We believe that by showcasing the current abilities and limitations GPT-4 and releasing our task sets to provide an assessment mechanism for multimodal LLMs moving forward, we have paved the way for visualization researchers to utilize these models in a more confident, informed, and responsible manner.

Supplemental Materials

Supplemental materials are available online at https://doi.org/10.17605/OSF.IO/F39J6 and additionally submitted along with this paper. In particular, they include all code, stimuli, and results for the main task sets and replication study.

Acknowledgments

The authors wish to thank Gaurav Verma from Georgia Institute of Technology for reviewing an early draft of this paper and providing valuable input from the perspective of an LLM researcher.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv*, 2023. doi: 10.48550/arXiv.2303.08774 1

[2] H. Alkaissi and S. I. McFarlane. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2), 2023. doi: 10.7759/cureus.35179 1

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on InfoVis*, pp. 111–117, 2005. doi: 10.1109/INFVIS.2005.1532136 2, 3

[4] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proc. CHI*, pp. 2573–2582. ACM, 2010. doi: 10.1145/1753326.1753716 2

[5] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar. Introducing our multimodal models, 2023. https://www.adept.ai/blog/fuyu-8b/. 9

[6] V. L. Bommineni, S. Bhagwagar, D. Balcarcel, C. Davatzikos, and D. Boyer. Performance of chatgpt on the mcat: the road to personalized and equitable premedical learning. *MedRxiv*, 2023. doi: 10.1101/2023.03.05.23286533 2

[7] K. Börner, A. Bueckle, and M. Ginda. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *PNAS*, 116(6):1857–1864, 2019. doi: 10.1073/pnas.1807180116 2

[8] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A principled way of assessing visualization literacy. *IEEE Trans. Visual Comput. Graphics*, 20(12):1963–1972, 2014. doi: 10.1109/TVCG.2014.2346984 2

[9] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Trans. Visual Comput. Graphics*, 19(12):2376–2385, 2013. doi: 10.1109/TVCG.2013.124 2, 3

[10] C. M. Carswell. Choosing specifiers: An evaluation of the basic tasks model of graphical perception. *Hum. Factors*, 34(5):535–554, 1992. doi: 10.1177/001872089203400503 2

[11] R. Chaudhry, S. Shekhar, U. Gupta, P. Maneriker, P. Bansal, and A. Joshi. Leaf-qa: Locate, encode & attend for figure question answering. In *Proc. WACV*, pp. 3512–3521. IEEE, 2020. doi: 10.1109/WACV45572.2020.9093269 2

[12] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, and Y. Xiao. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proc. CIKM*, pp. 245–255. ACM, 2023. doi: 10.1145/3583780.3614905 1

[13] Z. Chen, Q. Yang, X. Xie, J. Beyer, H. Xia, Y. Wu, and H. Pfister. Sporthesia: Augmenting sports videos using natural language. *IEEE Trans. Visual Comput. Graphics*, 29(1):918–928, 2022. doi: 10.1109/TVCG.2022.3209497 2

[14] Z. Chen, C. Zhang, Q. Wang, J. Troidl, S. Warchol, J. Beyer, N. Gehlenborg, and H. Pfister. Beyond generating code: Evaluating gpt on a data visualization course. In *IEEE VIS EduVis Workshop*, pp. 16–21, 2023. doi: 10.1109/EduVis60792.2023.00009 1, 2, 9

[15] J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz. Chatgpt goes to law school. *J. Legal Educ.*, 71(3):387–400, 2022. 2

[16] A. Coscia and A. Endert. Knowledgevis: Interpreting language models by comparing fill-in-the-blank prompts. *IEEE Trans. Visual Comput. Graphics*, 2023. To appear. doi: 10.1109/TVCG.2023.3346713 2

[17] A. Coscia, L. Holmes, W. Morris, J. S. Choi, S. Crossley, and A. Endert. iscore: Visual analytics for interpreting how language models automatically score summaries. In *Proc. IUI*. ACM, 2024. doi: 10.1145/3640543.3645142 2

[18] F. R. Curcio. Comprehension of mathematical relationships expressed in graphs. *J. Res. Math. Educ.*, 18(5):382–393, 1987. doi: 10.5951/jresematheduc.18.5.0382 2

[19] R. Dale. Gpt-3: What's it good for? *Nat. Lang. Eng.*, 27(1):113–118, 2021. doi: 10.1017/S1351324920000601 1, 7, 9

[20] J. F. DeRose, J. Wang, and M. Berger. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Trans. Visual Comput. Graphics*, 27(2):1160–1170, 2021. doi: 10.1109/TVCG.2020.3028976 2

[21] V. Dibia. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proc. ACL*, pp. 113–126. ACL, 2023. doi: 10.18653/v1/2023.acl-demo.11 1, 2

[22] A. Fan, Y. Ma, M. Mancenido, and R. Maciejewski. Annotating line charts for addressing deception. In *Proc. CHI*, pp. 80:1–80:12. ACM, 2022. doi: 10.1145/3491102.3502138 9

[23] Y. Feng, X. Wang, K. K. Wong, S. Wang, Y. Lu, M. Zhu, B. Wang, and W. Chen. Promptmagician: Interactive prompt engineering for text-to-image creation. *IEEE Trans. Visual Comput. Graphics*, 30(1):295–305, 2024. doi: 10.1109/TVCG.2023.3327168 2, 8

[24] S. Frieder, L. Pinchetti, , R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, and J. Berner. Mathematical capabilities of chatgpt. In *NeurIPS*, vol. 36, pp. 27699–27744, 2023. 5

[25] S. N. Friel, F. R. Curcio, and G. W. Bright. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *J. Res. Math. Educ.*, 32(2):124–158, 2001. doi: 10.2307/749671 2

[26] M. Galesic and R. Garcia-Retamero. Graph literacy: A cross-cultural comparison. *Med. Decis. Mak.*, 31(3):444–457, 2011. doi: 10.1177/0272989X10373805 2

[27] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proc. CVPR*, pp. 14375–14385. IEEE, 2024. 1, 2

[28] E. Hoque, P. Kavehzadeh, and A. Masry. Chart question answering: State of the art and future directions. *Comput. Graphics Forum*, 41(3):555–572, 2022. doi: 10.1111/cgf.14573 1, 2

[29] K.-H. Huang, M. Zhou, H. P. Chan, Y. R. Fung, Z. Wang, L. Zhang, S.-F. Chang, and H. Ji. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the ACL*, pp. 1314–1326. ACL, 2024. doi: 10.18653/v1/2023.findings-acl.85 2

[30] K. Kafle, R. Shrestha, S. Cohen, B. Price, and C. Kanan. Answering questions about data visualizations using efficient bimodal fusion. In *Proc. WACV*, pp. 1498–1507. IEEE, 2020. doi: 10.1109/WACV45572.2020.9093494 2

[31] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D. L. Arendt, S. Shaikh, and W. Dou. Vulnerable to misinformation? verifi! In *Proc. IUI*, pp. 312–323. ACM, 2019. doi: 10.1145/3301275.3302320 1

[32] D. H. Kim, E. Hoque, and M. Agrawala. Answering questions about charts and generating visual explanations. In *Proc. CHI*, pp. 1–13. ACM, 2020. doi: 10.1145/3313831.3376467 1, 2, 4, 9

[33] D. H. Kim, V. Setlur, and M. Agrawala. Towards understanding how readers integrate charts and captions: A case study with line charts. In *Proc. CHI*, pp. 610:1–610:11. ACM, 2021. doi: 10.1145/3411764.3445443 7

[34] H.-K. Kong, Z. Liu, and K. Karahalios. Frames and slants in titles of visualizations on controversial topics. In *Proc. CHI*, pp. 438:1–438:12. ACM, 2018. doi: 10.1145/3173574.3174012 1, 2, 7

[35] H.-K. Kong, Z. Liu, and K. Karahalios. Trust and recall of information across varying degrees of title-visualization misalignment. In *Proc. CHI*, pp. 346:1–346:13. ACM, 2019. doi: 10.1145/3290605.3300576 1, 2, 7

[36] B. C. Kwon and B. Lee. A comparative evaluation on online learning approaches using parallel coordinate visualization. In *Proc. CHI*, pp. 993–997. ACM, 2016. doi: 10.1145/2858036.2858101 2

[37] C. Lauer and S. O'Brien. How people are influenced by deceptive tactics in everyday charts and graphs. *IEEE Trans. Prof. Commun.*, 63(4):327–340, 2020. doi: 10.1109/TPC.2020.3032053 2

[38] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry. Task taxonomy for graph visualization. In *Proc. AVI BELIV Workshop*, pp. 1–5. ACM, 2006. doi: 10.1145/1168149.1168168 9

[39] M. Lee. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10):2320, 2023. doi: 10.3390/math11102320 1

[40] S. Lee, S.-H. Kim, and B. C. Kwon. Vlat: Development of a visualization literacy assessment test. *IEEE Trans. Visual Comput. Graphics*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920 1, 2, 3

[41] G. Li, X. Wang, G. Aodeng, S. Zheng, Y. Zhang, C. Ou, S. Wang, and C. H. Liu. Visualization generation with large language models: An evaluation. *arXiv*, 2024. doi: 10.48550/arXiv.2401.11255 2

[42] H. Li and N. Moacdieh. Is "chart junk" useful? an extended examination of visual embellishment. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 58(1):1516–1520, 2014. doi: 10.1177/1541931214581316 2

[43] R. Li, W. Xiao, L. Wang, H. Jang, and G. Carenini. T3-vis: visual analytic for training and fine-tuning transformers in nlp. In *Proc. EMNLP*, pp. 220–230. ACL, 2021. doi: 10.18653/v1/2021.emnlp-demo.26 2

[44] Y. Li, J. Wang, X. Dai, L. Wang, C.-C. M. Yeh, Y. Zheng, W. Zhang, and K.-L. Ma. How does attention work in vision transformers? a visual analytics attempt. *IEEE Trans. Visual Comput. Graphics*, 29(6):2888–2900, 2023. doi: 10.1109/TVCG.2023.3261935 2

[45] Z. Li, X. Wang, W. Yang, J. Wu, Z. Zhang, Z. Liu, M. Sun, H. Zhang, and S. Liu. A unified understanding of deep nlp models for text classification. *IEEE Trans. Visual Comput. Graphics*, 28(12):4980–4994, 2022. doi: 10.1109/TVCG.2022.3184186 2

[46] C. Liu, Y. Guo, and X. Yuan. Autotitle: An interactive title generator for visualizations. *IEEE Trans. Visual Comput. Graphics*, 30(8):5276–5288, 2024. doi: 10.1109/TVCG.2023.3290241 2, 9

[47] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, vol. 36, pp. 34892–34916, 2023. 9

[48] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017, 2023. doi: 10.1016/j.metrad.2023.100017 1, 7, 9

[49] 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. In *ICLR*, 2024. 9

[50] P. Maddigan and T. Susnjak. Chat2vis: Generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *IEEE Access*, 11:45181–45193, 2023. doi: 10.1109/ACCESS.2023.3274199 1, 2

[51] A. V. Maltese, J. A. Harsh, and D. Svetina. Data visualization literacy: Investigating data interpretation along the novice—expert continuum. *Journal of College Science Teaching*, 45(1):84–90, 2015. 2

[52] 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. In *Findings of the ACL*, pp. 2263–2279. ACL, 2022. doi: 10.18653/v1/2022.findings-acl.177 2, 4

[53] A. Narechania, A. Karduni, R. Wesslen, and E. Wall. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Trans. Visual Comput. Graphics*, 28(1):486–496, 2022. doi: 10.1109/TVCG.2021.3114820 2

[54] P. Newton and M. Xiromeriti. Chatgpt performance on multiple choice question examinations in higher education. a pragmatic scoping review. *Assessment & Evaluation in Higher Education*, pp. 1–18, 2023. doi: 10.1080/02602938.2023.2299059 2

[55] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv*, 2023. doi: 10.48550/arXiv.2303.13375 2

[56] S. O'Brien and C. Lauer. Testing the susceptibility of users to deceptive data visualizations when paired with explanatory text. In *Proc. SIGDOC*, pp. 7:1–7:8. ACM, 2018. doi: 10.1145/3233756.3233961 2

[57] OpenAI. "Vision - OpenAI API". `platform.openai.com`. Accessed: Mar. 28, 2024. [Online]. Available: `https://platform.openai.com/docs/guides/vision/limitations`. 3

[58] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proc. CHI*, pp. 1469–1478. ACM, 2015. doi: 10.1145/2702123.2702608 1, 2, 5

[59] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*, 2024. doi: 10.48550/arXiv.2403.05530 9

[60] L. E. Resck, J. R. Ponciano, L. G. Nonato, and J. Poco. Legalvis: Exploring and inferring precedent citations in legal documents. *IEEE Trans. Visual Comput. Graphics*, 29(6):3105–3120, 2023. doi: 10.1109/TVCG.2022.3152450 2

[61] P. Ruchikachorn and K. Mueller. Learning visualizations by analogy: Promoting visual literacy through visualization morphing. *IEEE Trans. Visual Comput. Graphics*, 21(9):1028–1044, 2015. doi: 10.1109/TVCG.2015.2413786 2

[62] L. Shen, E. Shen, Y. Luo, X. Yang, X. Hu, X. Zhang, Z. Tai, and J. Wang. Towards natural language interfaces for data visualization: A survey. *IEEE Trans. Visual Comput. Graphics*, 29(6):3121–3144, 2023. doi: 10.1109/TVCG.2022.3148007 2

[63] L. Shen, Y. Zhang, H. Zhang, and Y. Wang. Data player: Automatic generation of data videos with narration-animation interplay. *IEEE Trans. Visual Comput. Graphics*, 30(1):109–119, 2024. doi: 10.1109/TVCG.2023.3327197 2

[64] C. Shi, F. Nie, Y. Hu, Y. Xu, L. Chen, X. Ma, and Q. Luo. Medchemlens: An interactive visual tool to support direction selection in interdisciplinary experimental research of medicinal chemistry. *IEEE Trans. Visual Comput. Graphics*, 29(1):63–73, 2023. doi: 10.1109/TVCG.2022.3209434 2

[65] H. Singh and S. Shekhar. Stl-cqa: Structure-based transformers with localization and encoding for chart question answering. In *Proc. EMNLP*, pp. 3275–3284. ACL, 2020. doi: 10.18653/v1/2020.emnlp-main.264 2

[66] C. Stokes, V. Setlur, B. Cogley, A. Satyanarayan, and M. A. Hearst. Striking a balance: reader takeaways and preferences when integrating text and charts. *IEEE Trans. Visual Comput. Graphics*, 29(1):1233–1243, 2023. doi: 10.1109/TVCG.2022.3209383 7

[67] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Trans. Visual Comput. Graphics*, 29(1):1146–1156, 2023. doi: 10.1109/TVCG.2022.3209479 2, 8

[68] N. Sultanum and A. Srinivasan. Datatales: Investigating the use of large language models for authoring data-driven articles. In *Proc. VIS*, pp. 231–235. IEEE, 2023. doi: 10.1109/VIS54172.2023.00055 1, 2

[69] M. Sun, L. Cai, W. Cui, Y. Wu, Y. Shi, and N. Cao. Erato: Cooperative data story editing via fact interpolation. *IEEE Trans. Visual Comput. Graphics*, 29(1):983–993, 2023. doi: 10.1109/TVCG.2022.3209428 2

[70] H. Wainer. Understanding graphs and tables. *Educ. Res.*, 21(1):14–23, 1992. doi: 10.3102/0013189X021001014 2

[71] X. Wang, R. Huang, Z. Jin, T. Fang, and H. Qu. Commonsensevis: Visualizing and understanding commonsense reasoning capabilities of natural language models. *IEEE Trans. Visual Comput. Graphics*, 30(1):273–283, 2024. doi: 10.1109/TVCG.2023.3327153 2

[72] A. Wu, K. Brantley, and Y. Artzi. A surprising failure? multimodal llms and the nlvr challenge. *arXiv*, 2024. doi: 10.48550/arXiv.2402.17793 9

[73] S. Wu, H. Shen, D. S. Weld, J. Heer, and M. T. Ribeiro. Scattershot: Interactive in-context example curation for text transformation. In *Proc. IUI*, pp. 353–367. ACM, 2023. doi: 10.1145/3581641.3584059 2, 8

[74] T. Wu, M. Terry, and C. J. Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proc. CHI*, pp. 385:1–385:22. ACM, 2022. doi: 10.1145/3491102.3517582 2, 8

[75] W. Yang, M. Liu, Z. Wang, and S. Liu. Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media*, 10:399–424, 2024. doi: 10.1007/s41095-023-0393-x 2

[76] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viegas, and M. Wattenberg. Attentionviz: A global view of transformer attention. *IEEE Trans. Visual Comput. Graphics*, 30(01):262–272, 2024. doi: 10.1109/TVCG.2023.3327163 2

[77] L. Ying, X. Shu, D. Deng, Y. Yang, T. Tang, L. Yu, and Y. Wu. Metaglyph: Automatic generation of metaphoric glyph-based visualization. *IEEE Trans. Visual Comput. Graphics*, 29(1):331–341, 2023. doi: 10.1109/TVCG.2022.3209447 2

[78] X. Zhang, J. P. Ono, H. Song, L. Gou, K.-L. Ma, and L. Ren. Sliceteller: A data slice-driven approach for machine learning model validation. *IEEE Trans. Visual Comput. Graphics*, 29(1):842–852, 2023. doi: 10.1109/TVCG.2022.3209465 2

[79] Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, and H. Yao. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*, 2024. 1

[80] G. Zuccon, B. Koopman, and R. Shaik. Chatgpt hallucinates when attributing answers. In *Proc. SIGIR-AP*, pp. 46–51. ACM, 2023. doi: 10.1145/3624918.3625329 1