Do you "see" what I "see"? A Multi-panel Visual Question Answer Dataset for Large Language Model Chart Analysis

Daniel J. Evans[1[0000−0002−6047−6540]], J. P. Naiman[1[0000−0002−9397−6189]], JooYoung Seo[1[0000−0002−4064−6012]], and J. Stephen Downie[1[0000−0001−9784−5090]]

School of Information Sciences, University of Illinois, Urbana-Champaign, 61820, USA
djevans4@illinois.edu

Abstract. Digital libraries increasingly host scientific publications whose figures encode primary research findings, yet making this visual content searchable, accessible, and reusable remains an open challenge. Large language and multimodal models (LLMs/LMMs) provide the possibility to automate many figure parsing tasks such as determining data trends or extracting data from plots. In this work, we consider if these models can reliably support task-oriented visual data analysis when presented with synthetic, but realistic, scientific figures that contain multiple charts. Our Visual Question Answering (VQA) process scaffolds our tasks along a four-level description model from basic conceptual answers to interpretive and analytical understanding. Additionally, we outline our process for synthetically generating our single and multi-panel (charts in which there are multiple panels present) plots and testing the data with various LMMs using quantitative analysis. In general, we find LMMs become less accurate with increasing figure complexity as the number of panels, number of plot elements, and difficulty of questions increase. Finally, because LMM accuracy is expected to change rapidly, our analysis also considers cost relative to performance as a metric to inform future user decisions. We provide the detailed question-answer pairs used in this analysis, along with the code used to generate random plots, question-answer pairs and detailed bounding-box metadata for chart elements. The code and data are available at `https://github.com/ReadingTimeMachine/LLM_VQA_MultiPanel` and `https://huggingface.co/datasets/ReadingTimeMachine/visual_qa_multipanel`

Keywords: Visual Question and Answering · AI and Machine Learning Applications in Digital Libraries

1 Introduction

Scientific charts, plots, and diagrams are central to the research housed in digital libraries, and the ability to parse, search, and describe their content is a long-standing challenge in the digital libraries community [33, 57]. They grant coherence to large data sets by allowing readers to identify patterns, compare values, and examine particularly noteworthy aspects of a dataset [70]. However, for many readers, such as those who are blind or low-vision (BLV), interpreting those figures can be challenging. A recent audit of over 11,000 scientific PDFs

found only 2.4% met basic accessibility criteria [72], with figure and chart content among the least accessible elements [36, 46].

Large language models (LLMs) and large multimodal models (LMMs) have opened new avenues for interpreting visual data using natural language. By integrating image recognition with language-based reasoning, these systems are able to analyze and respond to questions about figures. Despite their promise, only recently has testing begun to understand how well these models perform task-oriented visual question analyses in a few-shot or zero-shot setting when faced with multiple paneled images. A growing body of work examines multi-panel chart benchmarks [49], [41], while others specifically target reasoning across multiple charts [61]. Multi-panel plots are charts in which there are multiple panels present. They are often complex with panels meant to be read in conjunction with each other. Systematic evaluation of cross-panel reasoning, where answering a question requires extracting and comparing information from multiple subplots within a single figure remains limited. This is especially problematic as multi-panel charts comprise ∼30–60% of scientific charts [38, 39, 69].

To address this gap, we present a preliminary study that evaluates several LMMs on a synthetically generated visual question answering (VQA) dataset designed for chart interpretation on multi-paneled charts. Our dataset scaffolds questions from basic factual identification to interpretive and analytical reasoning. This allows us to measure not only whether models can recognize visual features, but also whether they can generalize to more complex analysis tasks.

Specifically, we investigate:

- Can zero-shot LLMs reliably parse the multi-panel charts that populate digital library collections, and which question categories (description, comparison, inference) succeed or fail?
- Can LLMs surface chart content for BLV readers of scholarly publications, such as distributions, comparisons, trends, and what is the cost-accuracy trade-off across model tiers?

1.1 Related Work

VQA is a research process through which open-ended questions are asked and answered about an image and presented as a dataset [1]. These questions range in complexity from simple yes/no or multiple choice questions [30] to more advanced descriptions [2, 20] to summarization [62] and reasoning/claim checking [4]. Through the codification of a series of questions and answers, VQA provides a framework to benchmark the accuracy of LMMs to interpret images [10]. More recent work has begun to address multi-panel chart compositions [41, 49, 61], with some research including partial support in this space [68].

This codification underpins efforts to generate accessible figure captions [8,64, 67] and frame BLV-relevant queries that are often non-data and contextual [36].

While contextual information is useful, our dataset scaffolds those queries along a four-level figure description model [46]. This incorporates contextual information alongside other tasks as a description model.

1.2 Four-Level Figure Description Model

In what follows, we make use of the four-level figure description model as described by [46, 67] in order to quantify the accuracy of several LMMs on figure description tasks. During this study, we focus our efforts on Levels 1–3 of [46]. We describe our analysis levels in the table below:

Table 1. Four-level figure description model adapted from [46, 67].

Level	Description	Example
L1	Low-level "grammar of graphics" visual encoding	*"The graph shows a single line plot with x-axis 'Bob' (0–100) and y-axis 'Linda' (−20 to +20)"*
L2	Data trends or "data facts" [46, 65, 73]	*"The median of the y-axis data is 50"*
L3	Complex trends not evident from statistics, in informal language	*"The blue line appears linear, especially for large X"*
L4	Requires external context/domain expertise	Not included; left to future work

As shown in Table 1, we focus on Levels 1–3. Sighted readers typically find Level 3 and 4 the most useful descriptions, whereas BLV readers find Level 2 and Level 3 the most useful [46]. In our analysis, we test various multi-paneled plot "types." Extending [23,24,29,45], we provide a summary of existing VQA datasets (see Table 2). Included in this table are the "type" of chart and corresponding linked dataset, if available. We found that most VQA datasets incorporate multiple chart types, however, we found that few focus on one specifically, link to their original data, or include bounding boxes for chart elements.

Level 3 questions in our dataset ask about visual trends typically not reducible to a single statistic, for example whether a line appears linear, whether a distribution appears symmetric or skewed, or whether two panels show similar overall trends. Ground-truth answers for L3 questions are derived from the known distribution type used to generate the underlying data.

Additionally, as summarized in Table 2, few existing datasets include multi-panel figures. Recent work has begun to address this gap [41, 49, 61] by including multi-panel compositions. We also found a series of VQA datasets with partial support for multi-panel plots [15, 22, 40, 45, 68]. Yet these datasets primarily collect questions about individual charts within a multi-panel layout rather than posing questions that require reasoning across panels. Furthermore, we found no existing chart VQA dataset that includes contour plots, despite their widespread use in scientific and engineering publications for visualizing continuous surfaces and multi-variable relationships.

2 Methods

2.1 Figure Dataset Generation

Synthetically generated data can often fail to mimic realistic images [49, 54, 74], but has the advantage of enabling the storage of precise bounding box information and complex datasets used to create the figures [29]. We employ several methods to best align our synthetically generated figures with their features with "real" charts as faithfully as possible while enabling the sharing of a dataset which

Table 2. Overview of chart VQA datasets. Plot types: l=line, b=bar, s=scatter, p=pie, h=histogram, o=other. BB: bounding boxes included (P=partial). Data: original data accessible. MP: multi-panel figures included.

Name	Date	Plot Type	BB	Data	MP
FigureQA [30]	2018	lbp	✓	–	–
ChartLlama [20]	2023	so	–	–	–
ChartAssistant/SFT [53]	2024	lbspo	–	P	–
MMCA [44]	2024	lbspho	–	–	–
Chart2Text [32, 58]	2020/22	lbspo	P	✓	–
Chocolate [25]	2024	lb	–	✓	–
Leaf-qa [12]	2020	lbspo	✓	–	–
ChartQA [51]	2022	lbp	✓	✓	–
ChartCheck [4]	2023	lbspo	–	✓	–
OpenCQA [31]	2022	lbspo	✓	P	–
MapQA [11]	2022	o	✓	✓	–
MathVista [45]	2023	lbspo	–	–	P
SciGraphQA [40]	2023	o	–	–	P
VisText [67]	2023	lbo	P	✓	–
Chart-Info [14–16, 27, 28]	2019–24	lbo	✓	P	P
DocVQA [52]	2021	lbp	–	✓	P
ChartFC [3]	2023	b	✓	✓	–
RealCQA [2]	2023	lbso	✓	✓	–
StructChart [76]	2023	lbpho	–	✓	–
ChartSumm [62]	2023	lbp	–	–	–
ChartOCR [47]	2019	lbp	P	–	–
LineEX [60]	2023	l	✓	✓	–
SciCap [22]	2021	bso	–	–	P
AutoChart [77]	2021	lbs	✓	–	–
Unichart [50]	2023	lbp	P	–	–
PlotQA [54]	2020	lbs	✓	✓	–
MatCha [43]	2023	lbo	–	✓	–
AI2D [34]	2016	o	–	–	–
TQA [35]	2017	o	✓	✓	–
ChartQAPro [49]	2025	lbspo	–	–	✓
EvoChart-QA [26]	2025	lbspo	–	–	–
ChartMuseum [68]	2025	lbspo	–	–	P
ChartGalaxy [41]	2025	lbspo	–	✓	✓
RefChartQA [71]	2025	lbp	✓	P	–
ChartDQA [17]	2025	lbspho	–	✓	–
ChartQA-X [21]	2025	lbp	–	–	–
MultiChartQA-R [61]	2025	lbspo	–	–	✓
mChartQABench [75]	2025	lbpo	–	–	–
What Lies Beneath [56]	2026	h	✓	✓	–
Chart-HQA [13]	2026	lbpo	–	–	–
ChartNet [37]	2026	lbspho	P	✓	–

has complete information (dataset, generation parameters) and precise bounding boxes for all chart marks. Our current pipeline generates four plot "types" including line plots, histograms, contour and scatter plots. We generate these synthetic figures using the matplotlib package in Python. Relationships between the x and y data values are chosen from three types of distributions: linear relationships, gaussian mixture models, or random distributions. Colors for lines and bars are chosen at random, however, colors for scatter plot points are chosen from one of the three distribution options. Noise is added to each relationship at a prescribed level. Titles and x and y axis labels are taken from the 1000 most frequent words from a science-based NLP corpus mined from the arχiv [55]. Inline math formulas are substituted for words randomly, using data from the arχiv bulk downloads as in [55]. Fonts, fontsizes, scales (log or linear), x/y error bar existence and size, matplotlib plotting style, color maps, DPI, aspect ratio. Plot numbers and layout (horizontal, vertical or rows-columns) are chosen randomly.

The number of panels in a plot is chosen from a normal distribution centered around 6 with a standard deviation of 15. Each generated plot is automatically

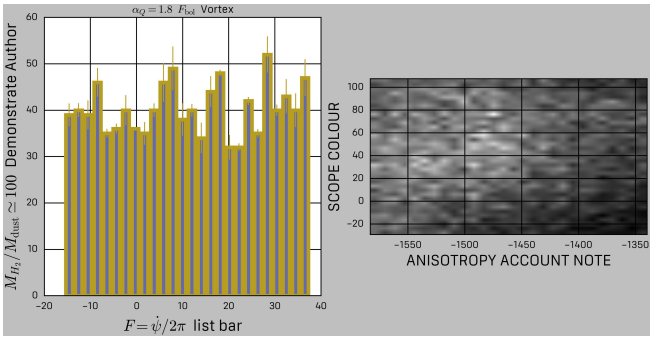

Fig. 1. Example of a multi-panel figure with complex data marks. On the left is a histogram, while on the right a contour plot is shown.

checked for overlapping figure boxes or labels, and if found, fontsize and layout is automatically updated until there are no overlaps. We re-run our generation code until there are >30% of figures with multi-panels (∼41% in this dataset).

The randomly selected plot parameters and data are saved in a JSON format, and used to generate VQA pairs using typical LLM/LMM prompting style [5,66]. While our synthetic figures do not fully capture the visual complexity of real scientific publications, particularly domain-specific color conventions and irregular layouts, they offer complete ground-truth metadata by including features are unavailable in real-figure datasets such as bounding boxes and generation parameters [42]. Furthermore, as seen in Table 2, histograms and contour plots are underrepresented in existing datasets but included in ours. All code used to generate these figures is shared on our Github repository.[1] Figure images and VQA dataset are stored on HuggingFace.[2] See Figure 1 for an exemplar multi-panel figure.

2.2 Question Data Generation

Along with randomly generated figures, our pipeline includes the generation of questions for specific panels in a one to several-panel figure. We restrict questions to those which can be addressed from a model's inspection of the figure alone, without requiring additional contextual information such as article text [25, 43], broader documents [50], or domain expertise [45]. We leave this level of analysis ("Level 4" of [46]) to future work, focusing instead on what can be extracted from the visual representation itself.

Questions are motivated by the work of [63], with prompts for LLMs distributed into three "Levels" of [46], either asked of the entire figure (Figure-level type) or of a specific panel (Plot-level type). When asked of a single panel, the panel location is specified with words (e.g., "second from the top, middle panel").

Individual question templates within each Level draw on question types established in prior chart VQA work. These include factual and counting questions comparable to those in ChartQA [51] and PlotQA [54], trend and comparison

[1] `https://github.com/ReadingTimeMachine/LLM_VQA_MultiPanel`
[2] `https://huggingface.co/datasets/ReadingTimeMachine/visual_qa_multipanel`

questions in the style of VisText [67] and OpenCQA [31], and interpretive reasoning questions adapted from RealCQA [2] and MathVista [45].

Question-answer pairs are generated in four parts and are validated by construction: Level 1 and Level 2 ground-truth answers are computed directly from stored generation parameters (e.g., panel count from figure metadata, means from underlying data arrays), while Level 3 answers derive from the known distribution type used to generate each panel. A manual review of sampled question-answer pairs confirmed correctness. We distinguish Figure-level questions, which require holistic parsing of the full multi-panel layout (e.g., panel counting), from Plot-level questions targeting individual panels. Neither category in the current dataset requires synthesizing values across multiple panels simultaneously. We identify this as the clearest direction for extending the benchmark. Representative question-answer pairs for each level are available in our dataset repository.

3 Results and Analysis

We first pass 200 figures and a single question to each of proprietary LMMs via their API: ChatGPT-5.4 [59], Claude-4.5 [7], and Gemini-3.1 [19]. If a token-overflow error is detected, figures are scaled down by a factor of two and the question is re-processed. We restrict our analysis to proprietary LMMs due to the fact that as of writing, open source LMMs generally do not perform as well on chart-specific VQA tasks (e.g., [20,68]). We consider two "classes" of LMMs. Lower level versions of these three models – GPT-nano with default reasoning (GPT in all figures), Gemini-flash-lite with variable thinking (thinking_budget = -1, Gemini), and Claude-Haiku (Claude). Mid-to-high level versions of models include GPT-nano with low verbosity but high reasoning (verbosity = 'low', reasoning_level = 'high', GPT-HR), GPT mini with low (high) verbosity (reasoning) (GPT-HRM), and Claude-Sonnet with thinking (thinking={'type':'adaptive'}, max_tokens = 10,000, Claude-T). These distinctions are not stringently defined, and simply mimic "buckets" of potential user parameter choice and model cost.

In order to compare to our pre-generated answers, we parsed LMM results and disregarded answers which did not fit our specifications for output (rejecting 130 questions). For example, we disregard the output if the model failed to return a numerical answer when one was expected. This results in a set of 33,560 (99.6%) answered questions for 200 image files. Each question was submitted as an independent API call with no shared context between questions. This single-query, zero-shot structure is distinct from a model-training setting, where correlated samples from the same figure could bias fitted parameters.

The accuracy of LMM answers from "Level 1" questions are shown in Figure 2. The left column shows estimates of the number of panels (a Figure-level question), comparing lower-level models (top) with higher-level reasoning models (bottom) as a function of the number of ground truth panels. Dashed lines indicate the 90% containment region of the distribution. Among the lower-level models, Gemini tracks the ground truth most closely, while GPT tends to slightly overestimate. The higher-level reasoning models (GPT-HR, GPT-HRM, and Claude-T) show equally comparable performance.

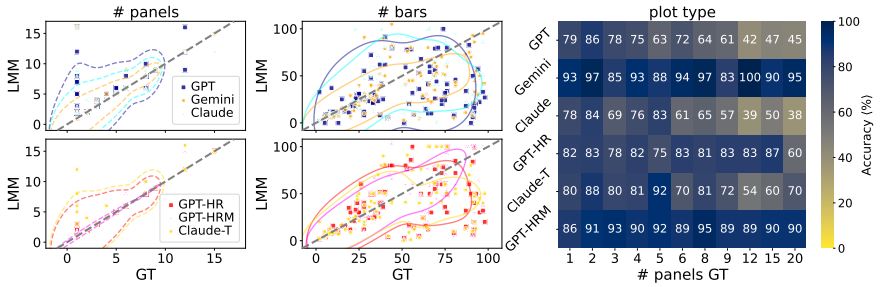

Fig. 2. Various "Level 1" VQA accuracies. Left-most plots show the LMM counts of number of figure panels as a function of ground truth (GT), and center-plots show the LMM counts of number of bars in histogram-type figures as a function of ground truth. Lower-level models (less complex, less reasoning/thinking) are top left/center plots, while mid-to-high level models are on the bottom left/center. Dashed gray line shows where LMM answer is equal to GT, dashed contours show 90% containment of distribution (left) and solid lines show 50% containment (center). The right-most plot shows mean accuracy per "# panels" bin of LMM selection of plot type (line, scatter, contour, histogram) as a function of GT "# panels".

The middle column of Figure 2 shows estimates of the number of bars for histograms (Plot-level questions), with contours indicating the 50% containment region. All models perform well at low bars but diverge as the number of bars increase, trending towards underestimation. The higher-level models (bottom-middle plot) show tighter contours, with GPT-HRM performing the best. The right panel of Figure 2 shows the accuracy of each model in identifying plot type (line, contour, scatter, histogram) as a function of the number of panels in the figure. Gemini achieves the highest overall accuracy across panel counts, while GPT and Claude showed a marked decrease in accuracy as the number of panels increased. Accuracy generally remains stable as the number of panels increases, for both Gemini and GPT-HRM.

Figure 3 shows the mean character error rate (CER) for each model's extraction of panel titles (left), x-axis labels (middle), and y-axis labels (right) as a function of the number of panels. CER was computed by first aligning the ground truth and LMM-returned lists of strings for each figure to minimize overall edit distance. Similar to the right-most panel of Figure 2, mean CER is calculated per ground truth "# panels". In alignment with OCR and LaTeX recognition work, CER is used as the evaluation metric as labels/titles may include mathematical formulas requiring sensitivity to character-level errors in symbolic notation [18].

Across all models, title extraction generally yields lower CER than axis label extraction. This is consistent with the higher relative probability of equations appearing in axis labels as axis labels allow up to 3 words with a 25% equation probability, compared to 5 words with a 10% equation probability for titles. Y-axis labels show notably higher CER than x-axis labels across most models, likely due to the difficulty of reading vertically oriented text and potential confusion with colorbar labels. Among the models, Gemini achieves the lowest CER overall, with GPT-HRM following closely behind.

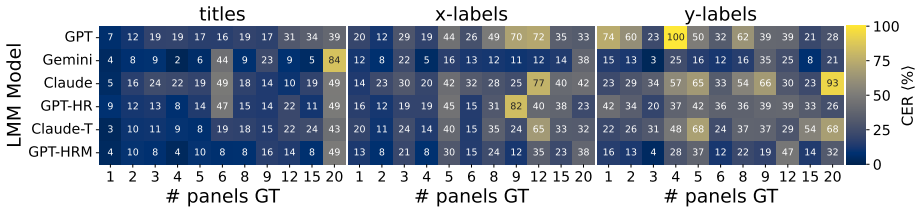

Fig. 3. Mean character error rate (CER) in each ground truth (GT) "# panels" bin as a function of GT "# panels" for figure panel titles and x/y axis labels. Titles are included with a probability of 50%, and x/y labels are always present.

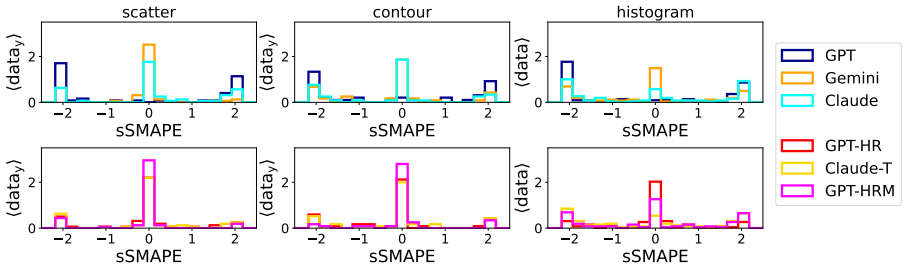

Fig. 4. LMM accuracy at estimating the mean of the data along the y-axis (scatter, contour) or in the distribution (histogram). Accuracy is measured as $sSMAPE = 2(\mu_{\text{LMM}} - \mu_{\text{GT}})/(|\mu_{\text{LMM}}| + |\mu_{\text{GT}}|)$ [48]. sSMAPE saturation occurs at ± 2, where $+2$ (-2) indicates instances where $\mu_{\text{LMM}} >> \mu_{\text{GT}}$ $(\mu_{\text{LMM}} << \mu_{\text{GT}})$.

Figure 4 shows the distribution of normalized error, for Level 2 questions asking models to estimate the mean value data value, broken down by a subset of plot types (contour, histogram, scatter). For scatter and contour plots, the mean data value along the y-axis is requested, while the mean data value (of the distribution) is requested for histograms. Here, the normalized error is calculated as a signed-SMAPE (sSMAPE) where $sSMAPE = 2(\mu_{\text{LMM}} - \mu_{\text{GT}})/(|\mu_{\text{LMM}}| + |\mu_{\text{GT}}|)$ [48]. We use this metric as unlike a simple weighting by μ_{GT}, sSMAPE has the advantage of not diverging when $\mu_{\text{GT/LMM}} \sim 0$. Zero-values of sSMAPE occur when $\mu_{\text{LMM}} \approx \mu_{\text{GT}}$, and saturation of the metric occurs at ± 2, where $+2$ (-2) indicates instances where $\mu_{\text{LMM}} >> \mu_{\text{GT}}$ $(\mu_{\text{LMM}} << \mu_{\text{GT}})$. Note that for histograms (right panels in Figure 4), the question asks about the mean of the underlying data rather than the mean along the axis, representing a subtly different, and more difficult, task, resulting in smaller peaks around sSMAPE≈0.

For lower level models (upper panels in Figure 4), Gemini shows the tightest clustering around zero across plot types while GPT struggles the most with the mean estimate along the y-axis. The mid/higher-level reasoning models (lower panels in Figure 4) generally improve upon their lower-level counterparts for mean estimation, suggesting that additional reasoning capability helps here.

3.1 Cost Analysis

In addition to answer accuracy measurements, there are considerations of cost. While ChatGPT-Mini with high reasoning (GPT-HRM) shows overall the high-

est accuracies, at ∼$55 it also results in the highest spend rate to answer all questions. Claude-Haiku with thinking (Claude-T) is the second most expensive (∼$36), however it does not meet the overall accuracy of Gemini (∼$25). Other, less expensive models like Claude-Haiku without thinking (Claude) and ChatGPT-nano with high reasoning (GPT-HR) are significantly less than these models (∼$13 and ∼$11, respectively), however the sacrifice in accuracy is significant. At $\lesssim$$1, ChatGPT-nano (GPT) is the least expensive model, however it is also overall the least accurate. These costs reflect API pricing at time of writing and should be interpreted as relative comparisons rather than absolute benchmarks, as model pricing and performance are subject to rapid change. They are included for methodological transparency.

Conclusions and Future Work

This paper presents methods to generate synthetic VQA datasets that can support evaluation of LLM/LMM tools for figure parsing in digital library settings. Our work has focused on multi-panel plots, an under represented figure format in the VQA community. Our error analysis shows that ChatGPT, Claude and Gemini succeed in describing several simple aspects of plots (e.g., number of panels), but that accuracy decreases with increasing figure and question complexity. In other words, LMMs become less accurate as the number of panels, number of plot elements, and difficulty of question increase. These results are broadly consistent with those of single-panel VQA (e.g., Table 6 in [24]), which also show increasing error rates for more complex questions. Our work considers accuracy in combination with cost to evaluate a model's "precision-per-cost". While we find Gemini currently provides the overall highest accuracy per cost, it is likely this will change rapidly with future versions of other LMMs.

Our work has used relatively simple metrics (binary comparisons, Levenshtein edit distance) to quantify accuracy. While this metric works for text like those in our figures, more free-form answers require more sophisticated metrics [20, 44]. Additionally, recently submitted work highlights the need to baseline any LMM accuracy to that of humans from a range of backgrounds [6].

Because BLV readers rely most heavily on Level 2 and Level 3 content, the accuracy declines we observe on these tasks (Figure 2-Figure 4) carry direct accessibility implications as the same complexity that challenges LMMs is the complexity most needed to make multi-panel figures usable for BLV readers. The present study focuses on establishing accuracy benchmarks, we leave direct evaluation of model outputs with BLV users for future work. While this work has focused on the estimations of the accuracy of "zero-shot" LLMs/LMMs, future work additionally will focus on the training of models using the synthetically generated plots discussed in this paper. As prior work has shown some promise in training and more complex prompting [9], we expect such methods to decrease the error rates reported here. However, validating this benchmark accuracy against a sample of real multi-panel figures from scientific publications also remains for follow-on work.

Acknowledgments. This work is supported by a NASA Astrophysics Data Analysis Program Grant (20-ADAP20-0225).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Batra, D., Parikh, D.: VQA: Visual Question Answering (Oct 2016). https://doi.org/10.48550/arXiv.1505.00468, http://arxiv.org/abs/1505.00468, arXiv:1505.00468 [cs]
2. Ahmed, S., Jawade, B., Pandey, S., Setlur, S., Govindaraju, V.: RealCQA: Scientific Chart Question Answering as a Test-Bed for First-Order Logic. In: Document Analysis and Recognition - ICDAR 2023. pp. 66–83. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-41682-8_5, https://link.springer.com/chapter/10.1007/978-3-031-41682-8_5, iSSN: 1611-3349
3. Akhtar, M., Cocarascu, O., Simperl, E.: Reading and Reasoning over Chart Images for Evidence-based Automated Fact-Checking. In: Vlachos, A., Augenstein, I. (eds.) Findings of the Association for Computational Linguistics: EACL 2023. pp. 399–414. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.findings-eacl.30, https://aclanthology.org/2023.findings-eacl.30/
4. Akhtar, M., Subedi, N., Gupta, V., Tahmasebi, S., Cocarascu, O., Simperl, E.: ChartCheck: Explainable Fact-Checking over Real-World Chart Images (Feb 2024). https://doi.org/10.48550/arXiv.2311.07453, http://arxiv.org/abs/2311.07453, arXiv:2311.07453 [cs]
5. Al-Hadhrami, S., Menai, M.E.B., Al-Ahmadi, S., Alnafessah, A.: A Critical Analysis of Benchmarks, Techniques, and Models in Medical Visual Question Answering. IEEE Access **11**, 136507–136540 (2023). https://doi.org/10.1109/ACCESS.2023.3335216, https://ieeexplore.ieee.org/document/10323452/
6. Anonymous: Anonymous title. ACM/IEEE Joint Conference on Digital Libraries (2025)
7. Anthropic: Claude 4 sonnet (2025), large language model
8. Belouadi, J., Ponzetto, S.P., Eger, S.: DeTikZify: Synthesizing Graphics Programs for Scientific Figures and Sketches with TikZ (May 2024). https://doi.org/10.48550/arXiv.2405.15306, http://arxiv.org/abs/2405.15306, arXiv:2405.15306 [cs]
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners **33**, 1877–1901 (2020), https://arxiv.org/abs/2005.14165
10. Cai, M., Liu, H., Park, D., Mustikovela, S.K., Meyer, G.P., Chai, Y., Lee, Y.J.: ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts (Apr 2024). https://doi.org/10.48550/arXiv.2312.00784, http://arxiv.org/abs/2312.00784, arXiv:2312.00784 [cs]
11. Chang, S., Palzer, D., Li, J., Fosler-Lussier, E., Xiao, N.: MapQA: A Dataset for Question Answering on Choropleth Maps (Nov 2022). https://doi.org/10.48550/arXiv.2211.08545, http://arxiv.org/abs/2211.08545, arXiv:2211.08545 [cs]

12. Chaudhry, R., Shekhar, S., Gupta, U., Maneriker, P., Bansal, P., Joshi, A.: LEAF-QA: Locate, Encode & Attend for Figure Question Answering. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 3501–3510. IEEE, Snowmass Village, CO, USA (Mar 2020). https://doi.org/10.1109/WACV45572.2020.9093269, https://ieeexplore.ieee.org/document/9093269/
13. Chen, X., Fang, Y., Li, J., Xiao, Q., Lin, J., Tang, S., Zhuang, Y.: Chart-HQA: A Benchmark for Hypothetical Question Answering in Charts. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13297–13303. ACM, Dublin Ireland (Oct 2025). https://doi.org/10.1145/3746027.3758288, https://dl.acm.org/doi/10.1145/3746027.3758288
14. Davila, K., Kota, B.U., Setlur, S., Govindaraju, V., Tensmeyer, C., Shekhar, S., Chaudhry, R.: ICDAR 2019 Competition on Harvesting Raw Tables from Infographics (CHART-Infographics). In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1594–1599 (Sep 2019). https://doi.org/10.1109/ICDAR.2019.00203, https://ieeexplore.ieee.org/document/8978105, iSSN: 2379-2140
15. Davila, K., Lazarus, R., Xu, F., Rodríguez Alcántara, N., Setlur, S., Govindaraju, V., Mondal, A., Jawahar, C.V.: CHART-Info 2024: A Dataset for Chart Analysis and Recognition. In: Lecture Notes in Computer Science, pp. 297–315. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-78495-8_19, https://link.springer.com/10.1007/978-3-031-78495-8_19, iSSN: 0302-9743, 1611-3349
16. Davila, K., Xu, F., Ahmed, S., Mendoza, D.A., Setlur, S., Govindaraju, V.: Icpr 2022: Challenge on harvesting raw tables from infographics (chart-infographics). In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 4995–5001. IEEE (2022)
17. Fan, W.C., Chen, Y.C., Liu, M., Jacobson, A., Yuan, L., Sigal, L.: In-Depth and In-Breadth: Pre-training Multimodal Language Models Customized for Comprehensive Chart Understanding (Jul 2025). https://doi.org/10.48550/arXiv.2507.14298, http://arxiv.org/abs/2507.14298, arXiv:2507.14298 [cs]
18. Gervais, P., Fadeeva, A., Maksai, A.: Mathwriting: A dataset for handwritten mathematical expression recognition (2025), https://arxiv.org/abs/2404.10690
19. Google DeepMind: Gemini 3.1 Pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/ (Feb 2026), large language model. Released February 19, 2026. API model identifier: gemini-3.1-pro
20. Han, Y., Zhang, C., Chen, X., Yang, X., Wang, Z., Yu, G., Fu, B., Zhang, H.: ChartLlama: A Multimodal LLM for Chart Understanding and Generation. CoRR (Jan 2023), https://openreview.net/forum?id=gMb0iC3eUv
21. Hegde, S., Fazli, P., Seifi, H.: ChartQA-X: Generating Explanations for Visual Chart Reasoning (Dec 2025). https://doi.org/10.48550/arXiv.2504.13275, http://arxiv.org/abs/2504.13275, arXiv:2504.13275 [cs]
22. Hsu, T.Y., Giles, C.L., Huang, T.H.K.: SciCap: Generating Captions for Scientific Figures (Oct 2021). https://doi.org/10.48550/arXiv.2110.11624, http://arxiv.org/abs/2110.11624, arXiv:2110.11624 [cs]
23. Huang, J., Chen, H., Yu, F., Lu, W.: From Detection to Application: Recent Advances in Understanding Scientific Tables and Figures. ACM Comput. Surv. **56**(10), 261:1–261:39 (Jun 2024). https://doi.org/10.1145/3657285, https://dl.acm.org/doi/10.1145/3657285
24. Huang, K.H., Chan, H.P., Fung, Y.R., Qiu, H., Zhou, M., Joty, S., Chang, S.F., Ji, H.: From Pixels to Insights: A Survey on Automatic Chart Understanding in the

Era of Large Foundation Models (Mar 2024). https://doi.org/10.48550/arXiv.2403.12027, http://arxiv.org/abs/2403.12027, arXiv:2403.12027 [cs] version: 2
25. Huang, K.H., Zhou, M., Chan, H.P., Fung, Y., Wang, Z., Zhang, L., Chang, S.F., Ji, H.: Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 730–749 (2024)
26. Huang, M., Lai, H., Zhang, X., Wu, W., Ma, J., Zhang, L., Liu, J.: EvoChart: A Benchmark and a Self-Training Approach Towards Real-World Chart Understanding (Sep 2024), https://arxiv.org/abs/2409.01577v2
27. ICDAR: Icdar 2023 competition on harvesting answers and raw tables from infographics (chart-infographics). https://chartinfo.github.io/ (Apr 2023), accessed September 5, 2025
28. ICPR: Competition on harvesting raw tables from infographics. https://chartinfo.github.io/index_2022.html (Nov 2020), accessed September 5, 2025
29. Jill P. Naiman, D.J.E., Seo, J.: What Lies Beneath: A Call for Distribution-based Visual Question & Answer Datasets. In: ACM/IEEE Joint Conference on Digital Libraries JCDL 2025. ACM/IEEE Joint Conference on Digital Libraries JCDL 2025 (Dec 2025)
30. Kahou, S.E., Atkinson, A., Michalski, V., Kádár, A., Trischler, A., Bengio, Y.: FigureQA: An Annotated Figure Dataset for Visual Reasoning (Feb 2018), https://openreview.net/forum?id=SyunbfbAb
31. Kantharaj, S., Do, X.L., Leong, R.T.K., Tan, J.Q., Hoque, E., Joty, S.: OpenCQA: Open-ended Question Answering with Charts (Oct 2022). https://doi.org/10.48550/arXiv.2210.06628, http://arxiv.org/abs/2210.06628, arXiv:2210.06628 [cs]
32. Kantharaj, S., Leong, R.T., Lin, X., Masry, A., Thakkar, M., Hoque, E., Joty, S.: Chart-to-Text: A Large-Scale Benchmark for Chart Summarization. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4005–4023. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.277, https://aclanthology.org/2022.acl-long.277/
33. Kashyap, A.R., Yang, Y., Kan, M.Y.: Scientific document processing: challenges for modern learning methods. International Journal on Digital Libraries **24**(4), 283–309 (2023). https://doi.org/10.1007/s00799-023-00352-7
34. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A Diagram Is Worth A Dozen Images (Mar 2016). https://doi.org/10.48550/arXiv.1603.07396, http://arxiv.org/abs/1603.07396, arXiv:1603.07396 [cs] version: 1
35. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5376–5384. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.571, http://ieeexplore.ieee.org/document/8100054/
36. Kim, J., Srinivasan, A., Kim, N.W., Kim, Y.S.: Exploring Chart Question Answering for Blind and Low Vision Users. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–15. CHI '23, Association for Computing Machinery, New York, NY, USA (Apr 2023). https://doi.org/10.1145/3544548.3581532, https://dl.acm.org/doi/10.1145/3544548.3581532

37. Kondic, J., Li, P., Joshi, D., Sanchez, I., Wiesel, B., Abedin, S., Alfassy, A., Schwartz, E., Caraballo, D., Cinar, Y.G., Scheidegger, F., Ross, S.I., Weidele, D.K.I., Hua, H., Arutyunova, E., Herzig, R., He, Z., Wang, Z., Yu, X., Zhao, Y., Jiang, S., Liu, M., Lin, Q., Staar, P., Lastras, L., Oliva, A., Feris, R.: ChartNet: A Million-Scale, High-Quality Multimodal Dataset for Robust Chart Understanding (Mar 2026), https://arxiv.org/abs/2603.27064v2
38. Lee, P.S., Howe, B.: Dismantling composite visualizations in the scientific literature. In: Proceedings of the International Conference on Pattern Recognition Applications and Methods - Volume 2. p. 79–91. ICPRAM 2015, SCITEPRESS - Science and Technology Publications, Lda, Setubal, PRT (2015). https://doi.org/10.5220/0005213100790091, https://doi.org/10.5220/0005213100790091
39. Lee, P.S., West, J.D., Howe, B.: Viziometrics: Analyzing Visual Information in the Scientific Literature. IEEE Transactions on Big Data **4**(1), 117–129 (Mar 2018). https://doi.org/10.1109/TBDATA.2017.2689038, https://ieeexplore.ieee.org/abstract/document/7888968?casa_token=DIiHFvOWs-wAAAAA:Ek6qnmKTdl7evW50NKVgKadyRp_YWQshgphKcx1O_UV701BZj-Bw7fy5QWIVWNNzCctgVKljwaOW, conference Name: IEEE Transactions on Big Data
40. Li, S., Tajbakhsh, N.: SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs (Aug 2023). https://doi.org/10.48550/arXiv.2308.03349, http://arxiv.org/abs/2308.03349, arXiv:2308.03349 [cs]
41. Li, Z., Li, D., Guo, Y., Guo, X., Li, B., Xiao, L., Qiao, S., Chen, J., Wu, Z., Zhang, H., Shu, X., Liu, S.: ChartGalaxy: A Dataset for Infographic Chart Understanding and Generation (Oct 2025). https://doi.org/10.48550/arXiv.2505.18668, http://arxiv.org/abs/2505.18668, arXiv:2505.18668 [cs]
42. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. arXiv preprint arXiv:2501.02189 (2025)
43. Liu, F., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Altun, Y., Collier, N., Eisenschlos, J.M.: MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering (May 2023). https://doi.org/10.48550/arXiv.2212.09662, http://arxiv.org/abs/2212.09662, arXiv:2212.09662 [cs]
44. Liu, F., Wang, X., Yao, W., Chen, J., Song, K., Cho, S., Yacoob, Y., Yu, D.: MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning pp. 1287–1310 (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.70, https://aclanthology.org/2024.naacl-long.70/
45. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts (Oct 2023), https://openreview.net/forum?id=KUNzEQMWU7
46. Lundgard, A., Satyanarayan, A.: Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. IEEE Transactions on Visualization & Computer Graphics (Proc. IEEE VIS) (Jan 2022). https://doi.org/10.1109/TVCG.2021.3114770, http://vis.csail.mit.edu/pubs/vis-text-model/
47. Luo, J., Li, Z., Wang, J., Lin, C.Y.: ChartOCR: Data Extraction From Charts Images via a Deep Hybrid Framework. pp. 1917–1925 (2021)
48. Makridakis, S.: Accuracy measures: theoretical and practical concerns. International Journal of Forecasting **9**(4), 527–529 (1993). https://doi.org/https://doi.org/10.1016/0169-2070(93)90079-3, https://www.sciencedirect.com/science/article/pii/0169207093900793

49. Masry, A., Islam, M.S., Ahmed, M., Bajaj, A., Kabir, F., Kartha, A., Laskar, M.T.R., Rahman, M., Rahman, S., Shahmohammadi, M., Thakkar, M., Parvez, M.R., Hoque, E., Joty, S.: ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering (Apr 2025). https://doi.org/10.48550/arXiv.2504.05506, http://arxiv.org/abs/2504.05506, arXiv:2504.05506 [cs] version: 2
50. Masry, A., Kavehzadeh, P., Do, X.L., Hoque, E., Joty, S.: UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning (Oct 2023), http://arxiv.org/abs/2305.14761, arXiv:2305.14761 [cs]
51. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 2263–2279. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.177, https://aclanthology.org/2022.findings-acl.177/
52. Mathew, M., Karatzas, D., Jawahar, C.V.: DocVQA: A Dataset for VQA on Document Images. pp. 2200–2209 (2021), https://arxiv.org/abs/2007.00398
53. Meng, F., Shao, W., Lu, Q., Gao, P., Zhang, K., Qiao, Y., Luo, P.: ChartAssistant: A Universal Chart Multimodal Language Model via Chart-to-Table Pretraining and Multitask Instruction Tuning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 7775–7803. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.463, https://aclanthology.org/2024.findings-acl.463/
54. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: PlotQA: Reasoning over Scientific Plots. pp. 1527–1536 (2020), https://arxiv.org/pdf/1909.00997
55. Naiman, J.P., Cosillo, M.G., Williams, P.K.G., Goodman, A.: Large synthetic data from the arχiv for ocr post correction of historic scientific articles. In: Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, September 26–29, 2023, Proceedings. p. 265–274. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-43849-3_23, https://doi.org/10.1007/978-3-031-43849-3_23
56. Naiman, J.P., Evans, D.J., Seo, J.: What Lies Beneath: A Call for Distribution-based Visual Question & Answer Datasets (Jan 2026). https://doi.org/10.48550/arXiv.2601.22218, http://arxiv.org/abs/2601.22218, arXiv:2601.22218 [cs]
57. Naiman, J.P., Williams, P.K.G., Goodman, A.: Figure and figure caption extraction for mixed raster and vector pdfs: Digitization of astronomical literature with ocr features. In: Linking Theory and Practice of Digital Libraries: 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022). Lecture Notes in Computer Science, vol. 13541. Springer (2022)
58. Obeid, J., Hoque, E.: Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In: Davis, B., Graham, Y., Kelleher, J., Sripada, Y. (eds.) Proceedings of the 13th International Conference on Natural Language Generation. pp. 138–147. Association for Computational Linguistics, Dublin, Ireland (Dec 2020). https://doi.org/10.18653/v1/2020.inlg-1.20, https://aclanthology.org/2020.inlg-1.20/

59. OpenAI: ChatGPT (GPT-5.4). https://openai.com/index/introducing-gpt-5-4/ (Mar 2026), large language model. Released March 5, 2026. API model identifier: gpt-5.4
60. P, S.V., Yusuf Hassan, M., Singh, M.: LineEX: Data Extraction from Scientific Line Charts. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6202–6210 (Jan 2023). https://doi.org/10.1109/WACV56688.2023.00615, https://ieeexplore-ieee-org.proxy2.library.illinois.edu/abstract/document/10030557, iSSN: 2642-9381
61. QiyuanFan, Zhang, J., WeitaoWang, Chen, S., Yu, Y., Zhang, Y.: MultiChartQA-R: A Benchmark for Multi-Chart Question Answering in Real-World Reasoning Scenarios (Oct 2025), https://openreview.net/forum?id=ibnb8sDM24
62. Rahman, R., Hasan, R., Farhad, A.A., Laskar, M.T.R., Ashmafee, M.H., Kamal, A.R.M.: ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization of Long and Short Summaries. CoRR (Jan 2023), https://arxiv.org/abs/2304.13620
63. Seo, J., Xia, Y., Lee, B., McCurry, S., Yam, Y.J.: MAIDR: Making Statistical Visualizations Accessible with Multimodal Data Representation (Mar 2024). https://doi.org/10.1145/3613904.3642730, http://arxiv.org/abs/2403.00717, arXiv:2403.00717 [cs]
64. Singh, N., Wang, L.L., Bragg, J.: FigurA11y: AI Assistance for Writing Scientific Alt Text. In: Proceedings of the 29th International Conference on Intelligent User Interfaces. pp. 886–906. ACM, Greenville SC USA (Mar 2024). https://doi.org/10.1145/3640543.3645212, https://dl.acm.org/doi/10.1145/3640543.3645212
65. Srinivasan, A., Drucker, S.M., Endert, A., Stasko, J.: Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. IEEE Transactions on Visualization and Computer Graphics **25**(1), 672–681 (Jan 2019). https://doi.org/10.1109/TVCG.2018.2865145, https://ieeexplore.ieee.org/document/8440860, conference Name: IEEE Transactions on Visualization and Computer Graphics
66. Sundar, A., Xu, J., Gay, W., Richardson, C., Heck, L.: cPAPERS: A Dataset of Situated and Multimodal Interactive Conversations in Scientific Papers (Jun 2024). https://doi.org/10.48550/arXiv.2406.08398, http://arxiv.org/abs/2406.08398, arXiv:2406.08398 [cs]
67. Tang, B., Boggust, A., Satyanarayan, A.: VisText: A Benchmark for Semantically Rich Chart Captioning. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7268–7298. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.401, https://aclanthology.org/2023.acl-long.401/
68. Tang, L., Kim, G., Zhao, X., Lake, T., Ding, W., Yin, F., Singhal, P., Wadhwa, M., Liu, Z.L., Sprague, Z., Namuduri, R., Hu, B., Rodriguez, J.D., Peng, P., Durrett, G.: ChartMuseum: Testing Visual Reasoning Capabilities of Large Vision-Language Models (2025). https://doi.org/10.48550/ARXIV.2505.13444, https://arxiv.org/abs/2505.13444, version Number: 3
69. Taschwer, M., Marques, O.: Automatic separation of compound figures in scientific articles. Multimedia Tools and Applications **77**(1), 519–548 (2018)
70. Tufte, E.R.: The visual display of quantitative information. Graphics Press, Cheshire, Connecticut (1983), publication Title: The visual display of quantitative information

71. Vogel, A., Moured, O., Chen, Y., Zhang, J., Stiefelhagen, R.: RefChartQA: Grounding Visual Answer on Chart Images through Instruction Tuning (Jun 2025). https://doi.org/10.48550/arXiv.2503.23131, http://arxiv.org/abs/2503.23131, arXiv:2503.23131 [cs]
72. Wang, L.L., Cachola, I., Bragg, J., Cheng, E.Y.Y., Haupt, C., Latzke, M., Kuehl, B., van Zuylen, M., Wagner, L., Weld, D.S.: Improving the accessibility of scientific documents: Current state, user needs, and a system solution to enhance scientific PDF accessibility for blind and low vision users. arXiv preprint arXiv:2105.00076 (2021)
73. Wang, Y., Sun, Z., Zhang, H., Cui, W., Xu, K., Ma, X., Zhang, D.: DataShot: Automatic Generation of Fact Sheets from Tabular Data. IEEE Transactions on Visualization and Computer Graphics **26**(1), 895–905 (Jan 2020). https://doi.org/10.1109/TVCG.2019.2934398, https://ieeexplore.ieee.org/document/8805442, conference Name: IEEE Transactions on Visualization and Computer Graphics
74. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., Chevalier, A., Arora, S., Chen, D.: CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. Advances in Neural Information Processing Systems **37**, 113569–113697 (Dec 2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/cdf6f8e9fd9aeaf79b6024caec24f15b-Abstract-Datasets_and_Benchmarks_Track.html
75. Wei, J., Xu, N., Chang, G., Luo, Y., Yu, B., Guo, R.: mChartQA and mChartQABench: A multimodal-only solution for complex chart question-answering. Pattern Recognition **172**, 112348 (Apr 2026). https://doi.org/10.1016/j.patcog.2025.112348, https://www.sciencedirect.com/science/article/pii/S003132032501009X
76. Xia, R., Peng, H., Ye, H., Li, M., Yan, X., Ye, P., Shi, B., Qiao, Y., Yan, J., Zhang, B.: StructChart: On the Schema, Metric, and Augmentation for Visual Chart Understanding (Dec 2024). https://doi.org/10.48550/arXiv.2309.11268, http://arxiv.org/abs/2309.11268, arXiv:2309.11268 [cs]
77. Zhu, J., Ran, J., Lee, R.K.W., Li, Z., Choo, K.: AutoChart: A Dataset for Chart-to-Text Generation Task. In: Mitkov, R., Angelova, G. (eds.) Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 1636–1644. INCOMA Ltd., Held Online (Sep 2021), https://aclanthology.org/2021.ranlp-1.183/