Benchmark It Yourself (BIY): Preparing a Dataset and Benchmarking AI Models for Scatterplot-Related Tasks

João Palmeiro [*] Diogo Duarte [†] Rita Costa [‡] Pedro Bizarro [§]
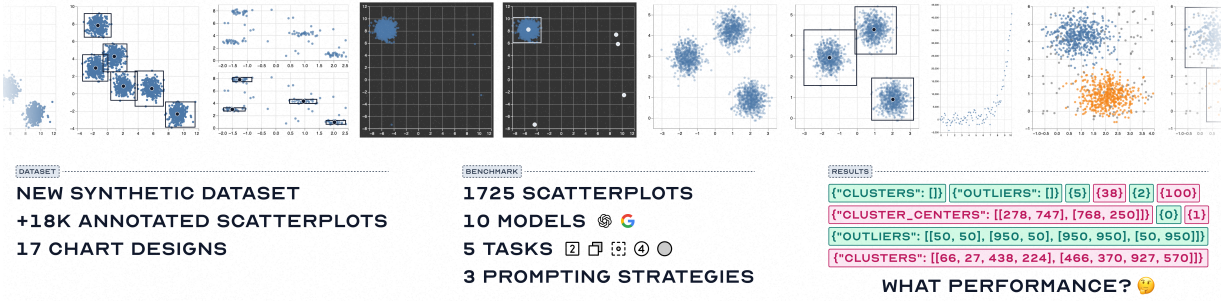
Feedzai

Figure 1: A high-level overview of the new dataset and benchmark for scatterplot-related tasks.

ABSTRACT

AI models are increasingly used for data analysis and visualization, yet benchmarks rarely address scatterplot-specific tasks, limiting insight into performance. To address this gap for one of the most common chart types, we introduce a synthetic, annotated dataset of over 18,000 scatterplots from six data generators and 17 chart designs, and a benchmark based on it. We evaluate proprietary models from OpenAI and Google using N-shot prompting on five distinct tasks derived from annotations of cluster bounding boxes, their center coordinates, and outlier coordinates. OpenAI models and Gemini 2.5 Flash, especially when prompted with examples, are viable options for counting clusters and, in Flash's case, outliers (90%+ Accuracy). However, the results for localization-related tasks are unsatisfactory: Precision and Recall are near or below 50%, except for Flash in outlier identification (65.01%). Furthermore, the impact of chart design on performance appears to be a secondary factor, but it is advisable to avoid scatterplots with wide aspect ratios (16:9 and 21:9) or those colored randomly. Supplementary materials are available at https://github.com/feedzai/biy-paper.

Index Terms: Computing methodologies—Artificial intelligence; Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

Large Language Models (LLMs), particularly multimodal models, are among today's key digital technologies. From promising academic results [31, 32] to new products and services [48, 67, 102], these AI models have demonstrated wide-ranging applicability, including for data analysis and visualization.

Nevertheless, when implementing a feature or developing new ideas, the ever-increasing number of options (such as models,

---

[*]e-mail: joao.palmeiro@feedzai.com

[†]e-mail: diogo.duarte@feedzai.com

[‡]e-mail: rita.costa@feedzai.com

[§]e-mail: pedro.bizarro@feedzai.com

prompting strategies, and multi-agent architectures [25]) can be a problem. Experimenting with different approaches can be costly and time-consuming, and it ultimately takes expertise to make a sound decision. One could simply resort to a frontier model or adapt an existing multi-agent system; however, performance remains uncertain.

For data analysis and visualization, several benchmarks offer guidance [15, 21, 42, 124]. Yet, when it comes to scatterplots, they lack representativeness and are not targeted at scatterplot-related tasks such as clustering and outlier detection [42, 46, 68, 69, 130, 136, 137] — a significant gap, as scatterplots are one of the most commonly used chart types [8, 12, 24] and a standard way to visualize two numerical variables to extract insights from their patterns. Moreover, existing benchmarks can be manipulated [95, 103] or require nuanced interpretation [5], rendering existing numbers more uncertain. Furthermore, models with vision capabilities have been shown to be brittle, highlighting their weaknesses [47, 65, 77, 93, 104, 117, 119].

In this work, we explore the application of AI models to a set of typical visual scatterplot tasks [98] using a new synthetic, annotated dataset, contributing to the common understanding of the actual capabilities of current models. When there is no access to the raw data or when datasets are large and extensively plotted — especially on dynamic platforms without relevant a priori context for the data — relying on chart images may be the only viable approach for extracting insights. These insights can then be used to generate descriptions for the charts or to complement multi-agent systems for data analysis and reporting. Our **main contributions** are: 1) a synthetic, annotated dataset (and its generation pipeline) for scatterplot-related tasks (3.1); 2) a comprehensive evaluation of the performance of ten proprietary models on said tasks (3.2); and 3) a list of considerations when designing charts and providing them as input to AI models (3.4).

2 RELATED WORK

2.1 Datasets

Multiple datasets are available for training and testing new models or for evaluating existing solutions. Chart understanding datasets are multimodal (each instance contains a chart image) and can be divided into categories according to task type. Three of the main ones are chart question answering (QA) [1, 13, 14, 19, 27, 37, 49,

58, 101, 132, 133, 140, 145], chart captioning [16, 40, 66, 76, 82, 85, 91, 142], and chart extraction [18, 144].

PlotQA [71], alongside FigureQA [46] and DVQA [45], is one of the first examples of chart QA datasets, consisting of a large set of annotations and question-answer pairs. VisText [106], Alt4Blind [75], and ChartSumm [92] are also rich in annotations but focus on chart descriptions or summaries. For chart extraction, ChartX [131] and SynthChartNet [78] map chart images to their corresponding tabular data (CSV) and a structured data format (OTSL [63]), respectively. There are also compilations, such as ChartSFT [70], DataVizQA [17], CHOCOLATE [43], and those assembled to pre- or post-train [138] models like Granite Vision [110], Llama Nemotron Nano VL [80], and Seed1.5-VL [33]. Lastly, datasets can be hybrid, combining chart understanding and chart code generation or chart extraction, for example [41, 55, 118, 120, 129, 135], or tasks across different domains like EMMA [34], MMMU [136, 137], and KITAB-Bench [36].

Our dataset is a hybrid of chart QA and extraction. It stands out from existing datasets by including complementary, scatterplot-oriented annotations to evaluate the clustering and outlier detection performance of AI models (or to train them), as well as by the ease, accuracy, and negligible cost of extending it.

2.2 Benchmarks and Experiments

Datasets like ChartQA [69], ChartQAPro [68], CharXiv [121], ChartMuseum [107], MMMU [136], and MMMU-Pro [137] are, in practice, benchmarks that are reported on when new models are released [6, 28, 61, 84, 109, 111]. Each has an expected prompt structure (though the final prompt can be adjusted depending on the model [59, 73, 139]) and at least one performance metric.

Beyond that, AI models have been evaluated from various perspectives, adding to the benchmark results. These experiments are as comprehensive as one can imagine, including visualization literacy testing [9, 39, 54, 56, 86], predicting the experiential impact of charts [51], evaluating the (positive) contribution of including chart images for data analysis [54], detecting misleading visualizations [60], and even evaluating the performance of LLMs on various tasks, taking SVG charts as input instead of raster images [134].

In any case, none of these benchmarks or experiments cover tasks related to multimodal (chart image + text) evaluation of scatterplots. The exceptions are PUB [87] and ChartInsights [130], the works closest to ours, which cover tasks for various chart types, including scatterplots. We chose to focus on scatterplots, tackling undocumented tasks such as outlier detection. We employ three different prompting strategies, a new set of models and metrics, and a distinct methodology for assembling the charts under evaluation. The scatterplots feature more complex patterns than the other works, resulting in a more diverse collection. Two other works [26, 97] are also worth mentioning, given the similar experimental setup, although they address different domains.

3 BENCHMARKING SCATTERPLOT-RELATED TASKS

In this section, we first introduce our dataset and the benchmark. Following that, we present the results and main considerations for future applications.

3.1 Dataset

Our synthetic dataset consists of 371 data samples, totaling **18,921 distinct scatterplots**. The data samples are obtained from six generators implemented in Python using NumPy [35], pandas [123], PyOD [143], scikit-learn [88], SciPy [115], and Shapely [29]. These data generators are designed to output Gaussian blobs with and without background noise (round clusters; 60 data samples each), Gaussian blobs with outliers (54), random patterns (scattered points with no clear pattern; 8), relationships (linear, exponential, and quadratic; 9), and, finally, geometric shape blobs (6 shapes; 180

in total). Each one is parameterized differently, with data samples containing between zero and six clusters. When outliers are injected, the contamination level varies between 0.001 and 0.01. We chose to inject a relatively small number of outliers, keeping them well-distanced from the clusters, in order to evaluate the detection of points that are clearly anomalous and relevant to report.

The scatterplot images are generated via Vega-Lite [99] and 17 different specifications [2, 7, 74, 77, 130]. Each corresponds to a chart design created from the default styling, including itself (*default*) ⣿ and its dark-themed counterpart ⣿ (2):

- Aspect ratios (5): 3:4, 4:3, 9:16, 16:9, and 21:9.
- Colors (2): colored clusters and randomly colored points.
- Opacity (2): full opacity and half the default opacity (0.35 instead of 0.7).
- Chart elements (2): points only and Y-axis only.
- Point shapes (2): random shapes and squared points.
- Point size (2): half the default size (15 instead of 30) and twice the default size (60).

The pipeline to combine the data samples and the Vega-Lite specifications is built on top of FastHTML [3], Playwright [72], and the Vega View API, generating an image and a set of annotations for each scatterplot. By creating a web application using FastHTML to render each scatterplot, launched in a separate thread, Playwright automatically interacts with this application to extract the scatterplot images and their corresponding annotations—all orchestrated from a Python script. The Vega View API is leveraged to transform Cartesian coordinates into screen coordinates (pixels) for each of the annotation types: cluster horizontal bounding boxes [141], cluster center coordinates, and outlier coordinates. Each image is saved in three sizes derived from Vega-Lite's standard dimensions: for a square image, approximately 150px, 300px (default), and 600px (the reference size used in the benchmark [10, 11, 116]).

Finally, the scatterplot images are optimized with Oxipng [38] to reduce the total file size and remove unnecessary metadata.

3.2 Benchmark

For the benchmark, the dataset was randomly sampled to focus on the stipulated analyses and balance costs. We tested **1,725 scatterplots**, featuring 15 chart designs (all except the point shape ones) for 115 data samples: 25 Gaussian blobs, 25 Gaussian blobs with background noise, 50 Gaussian blobs with outliers, 7 random patterns, and 8 relationships.

There were **five scatterplot-related tasks** under evaluation: cluster counting, cluster detection (via bounding boxes), cluster identification (from their respective centers), outlier counting, and outlier identification (from the points' coordinates). Each one was defined by a prompt structured with an instruction and a response format [93, 117]:

- ② **Cluster counting**: *How many clusters are there in the scatterplot? Answer with a number in curly brackets, e.g., {4}.*

- ⧉ **Cluster detection**: *Detect all clusters in the scatterplot. For each detected cluster, provide its bounding box in normalized coordinates [22, 94] (x1, y1, x2, y2), where (0, 0) is the top-left corner and (1000, 1000) is the bottom-right corner of the image. Answer with a JSON object, e.g., {"clusters": [[200, 30, 300, 50]]}.*

- ⸬ **Cluster identification**: *Identify each cluster in the scatterplot. For each, provide its center point in normalized coordinates (x, y), where (0, 0) is the top-left corner and (1000, 1000) is the bottom-right corner of the image. Answer with a JSON object, e.g., {"cluster_centers": [[250, 40], [700, 500]]}.*

- ④ **Outlier counting**: *How many outliers are there in the scatterplot? Answer with a number in curly brackets, e.g., {3}.*

- ◯ **Outlier identification**: *Identify each outlier in the scatterplot. For each, provide its location in normalized coordinates (x, y), where (0, 0) is the top-left corner and (1000, 1000) is the bottom-right corner of the image. Answer with a JSON object, e.g., {"outliers": [[150, 900], [820, 150]]}.*

In addition to the **zero-shot prompts** listed above, two other prompting strategies were used: **one-shot** and **few-shot** in turn-based conversation format [44, 52]. To this end, six examples were randomly selected from the complete dataset, before it was sampled, to represent each of the benchmarked data generators (plus an extra one so that the few-shot prompts would include two examples with outliers). We chose to include these two prompting strategies because they are easy to prepare (assuming there is labeled data), they do not require a higher number of inferences as prompt chaining strategies do, and they are expected to yield positive results [128].

The **models** tested are largely from OpenAI ⑤, as listed below. Their extensive adoption in various contexts [20, 96, 108, 114] motivated us to focus on this provider. Nevertheless, we also tested two of Google's Ⓖ models to begin comparing results across providers, especially for low-cost models.

- ⑤ *gpt-4.1-2025-04-14* (GPT-4.1), *gpt-4.1-mini-2025-04-14* (GPT-4.1 mini), *gpt-4.1-nano-2025-04-14* (GPT-4.1 nano), *gpt-4o-2024-08-06* (GPT-4o), and *gpt-4o-mini-2024-07-18* (GPT-4o mini) with the *temperature* set to 0.

- ⑤ *o3-2025-04-16* (o3) and *o4-mini-2025-04-16* (o4-mini) with the *temperature* set to 1 and *reasoning_effort* to *medium*.

- Ⓖ *gemini-2.5-flash* (Flash) and *gemini-2.5-flash-lite* (Flash-Lite) with the *temperature* set to 0 and no thinking[1] budget.

- Ⓖ *gemini-2.5-flash-lite* (Flash-Lite (Thinking)) with the *temperature* set to 0 and thinking budget to 8192 tokens [30, 79].

All OpenAI models were used with the image input detail level set to *high*, whereas Google's equivalent was set to *MEDIA_RESOLUTION_MEDIUM*[2].

3.3 Results

The benchmark was run between July 29th and August 5th, 2025, using the Batch APIs from OpenAI and Vertex AI (Google) [89]. The total cost was approximately \$666 [94] for 258,750 requests (115 *data samples* × 15 *chart designs* × 10 *models* × 5 *tasks* × 3 *prompting strategies*).

The raw responses were processed to extract the counts and denormalize the bounding boxes and points, according to the prompts. Between 70.2% and 99.95% of OpenAI's responses were considered valid per model, whereas for Google, the range was between 76.52%[3] and 91.91%. When breaking down the responses by chart design, the values are between 89.17% and 92.66%.

3.3.1 Counting

For the cluster and outlier counting tasks, we leverage two metrics: **Accuracy** (the proportion of correct counts) and **Mean Absolute Error (MAE)** [26, 126]. The former allows us to estimate how good the models and prompting strategies are, while the latter to quantify the deviation in the predicted counts.

Figure 2 shows the results for **Accuracy**. Regarding **MAE**, the best value for cluster counting is 0.03 (o3 when few-shot prompted), while the worst is 1.46 (Flash-Lite (Thinking) when one-shot prompted). However, although GPT-4o achieved an MAE
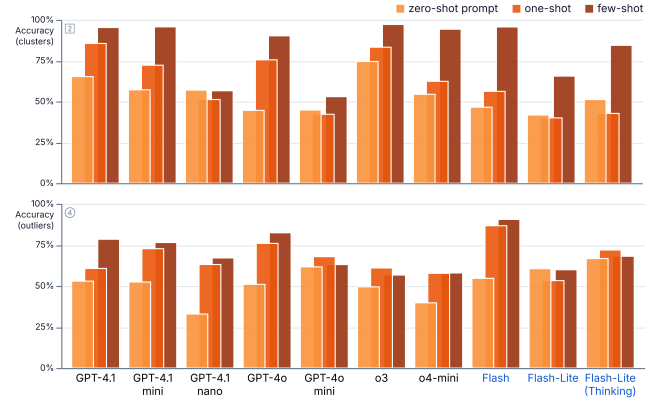
Figure 2: Accuracy for each model and prompting strategy. The results for cluster counting ② (top) using few-shot prompting are particularly promising for several models. On the other hand, Flash excels at the outlier counting ④ task (bottom) when one-shot (86.67%) and few-shot (90.49%) prompted.

of 0.29 for outlier counting when few-shot prompted, o4-mini achieves 18.54, which is much higher than expected given the relatively low number of outliers in the dataset.

On the other hand, the **consistency** in the predicted number of clusters ②□:: or outliers ④◯ varied considerably across models and prompting strategies. Consistency implies that a given scatterplot is predicted with the same number of clusters/outliers, whether by counting them or by checking the number of predicted bounding boxes or coordinates. The results show that models like o3 and GPT-4.1, when few-shot prompted, predicted the same amount of clusters for over 90% of scatterplots. Others, such as GPT-4.1 nano and mini (both zero-shot prompted), responded with the same number of clusters and outliers only 26.71% and 24.39% of the time, respectively. On average, only 61.4% of all scatterplots were predicted with the same number of clusters (57.7% for outliers).

The results exclusively for scatterplots with **no clustering patterns or outliers** also vary considerably. When the correct answer is simply *0*, GPT-4.1, o3, and Flash, when few-shot prompted, achieve 100% Accuracy in cluster counting. Flash, when one-shot and few-shot prompted, reaches 100% for outlier counting as well. However, in cluster counting, Accuracy is below 1% for Flash and Flash-Lite (one-shot), as well as GPT-4o (zero-shot). The lowest Accuracy in outlier counting is 50.89% for the zero-shot-prompted o4-mini.

3.3.2 Detection and Identification

For detection and identification tasks involving bounding boxes and point coordinates, we rely on **Precision** and **Recall** [53, 127]. More specifically, after obtaining the Precision and Recall values for each individual scatterplot, we average and report them considering the relevant groupings [122]. These two metrics allow us to verify if only relevant objects are being predicted, as well as if all relevant objects are indeed predicted. Plus, these metrics do not rely on models reporting confidence scores [122, 141].

For cluster detection, we start by matching predictions and targets based on their Intersection over Union (IoU) values [50, 125]. Then, we compute the confusion matrix. A True Positive corresponds to a predicted bounding box with an IoU of 0.75 or higher (**Precision and Recall @ IoU75**). Extra bounding boxes are considered False Positives, and if a target is not matched, we consider it a False Negative. This approach is also applied to identification tasks, where the threshold is a Euclidean distance of 10px [26] (**Precision and Recall @ 10px**). We opted for strict thresholds to determine whether a given model and prompting strategy can be used when numerical precision is required.

Figure 3: Recall for each model and prompting strategy. None surpass 25% Recall in the cluster detection ⬓ (top) and identification ⬚ (middle) tasks. Recall is also low for the outlier identification ◯ task (bottom), although Flash, when few-shot prompted, seems promising (65.01%). Lowering the IoU to 0.5 considerably changes the results for some models, but Recall remains low (o3, when few-shot prompted, reaches a Recall of 50.26%). Doubling the distance threshold to 20px does not change the conclusions.

Figure 3 shows the results for **Recall** (the distribution of Precision values is similar across models and prompting strategies). The results for scatterplots with **no clusters or outliers** vary significantly again depending on the model and prompting strategy. When the correct answer is an empty list, GPT-4.1, GPT-4o, and Flash, when few-shot prompted, respond correctly between 98.22% and 100% of the time on the various detection and identification tasks. However, particularly in the cluster identification task, five of the seven OpenAI models, when zero-shot or one-shot prompted, score as low as 0%.

3.3.3 Chart Design Performance

The motivation for benchmarking 15 chart designs stems from the need to evaluate the robustness of these models (three clusters should always be three clusters, regardless of the chart's aspect ratio or colors, for example) and to identify potential aesthetic choices that appear to negatively impact performance, which is often mitigable.

In the cluster counting task, the one with the best overall results, the best model when few-shot prompted (o3) achieves an Accuracy of 99.13% on the *default* chart design examples and on four others. However, for the *points_only* chart design, the Accuracy is only 90.43% — a gap of 8.7 percentage points (pp). These differences vary across models, prompting strategies, and chart designs.

Therefore, we evaluated the **effect of chart design on cluster counting accuracy** using logistic regression with the *default* chart design as the reference category. While the model's overall explanatory power [64] was low (*pseudo* $R^2 = 0.0013$), a likelihood ratio test indicated that chart design significantly improved model fit compared to the null model ($p < 0.001$). Relative to the *default* chart design (66.94% accuracy), the *half_opacity* design significantly improved accuracy by 2.42 pp (69.36%, *odds ratio* (OR) = 1.12, $p = 0.032$). Conversely, three chart designs significantly impaired performance: *16_9* (64.45%, $OR = 0.9$, $p = 0.03$), *21_9* (62.34%, $OR = 0.82$, $p < 0.001$), and *random_colors* (62.01%,

$OR = 0.81$, $p < 0.001$). These findings suggest that although chart design may explain a small proportion of the overall variability in accuracy, individual design choices can have nontrivial effects on performance.

3.4 Considerations

Based on the results, the main considerations when combining scatterplot images and AI models are as follows:

1. **Prioritize few-shot prompting.** Few-shot prompting consistently outperformed zero-shot prompting across all models and tasks, with top models (Figure 2) achieving over 90% Accuracy in counting tasks. Including examples is also useful for handling zero-answer scatterplots. Regardless, it can be worthwhile to determine the ideal number of examples for a given task to balance costs.

2. **Avoid using OpenAI and low-cost Google models with similar prompting strategies for localization tasks involving scatterplots.** Results are significantly better for counting tasks than for precise detection/identification tasks (Figure 3), where their performance is unreliable, compromising their usefulness.

3. **Invest in other components first, not chart design.** Chart design is fundamental for humans, but it's a secondary factor when fed into AI models. Even so, it can be beneficial to avoid chart designs with wide aspect ratios (16:9 and 21:9) or seemingly random colors. Adding opacity to points (for example, 0.35 in Vega-Lite and Vega-Altair [113]) can also be helpful.

4 CONCLUSION AND FUTURE WORK

In this work, we address the lack of benchmarks for scatterplot-related tasks by introducing a novel synthetic dataset and a comprehensive benchmark. This is our first step toward better gauging the real capabilities of AI models for chart understanding and their applicability to real-world contexts. Consequently, we intend to further study the impact of chart design on performance, analyze the trade-off between cost and performance, and examine invalid responses.

Subsequently, we plan to include more data generators, chart designs, and internal data, as well as enhance the benchmark's representativeness. To this end, we intend to incorporate new prompting strategies, tasks (such as identifying relationships between variables and cluster shapes), and proprietary and open [57] models. In parallel, we also plan to create a lightweight version of the dataset and benchmark [34, 90] to monitor the performance of new models and prompting strategies at low cost.

Following this phase, we will proceed to fine-tune small, open models [112] and further assess the trade-offs of using significantly less demanding models that can be run on our own infrastructure.

Finally, we will focus on chart captioning, leveraging the dataset to generate template-based descriptions for accessible scatterplots. Our goal is to evaluate the use of AI models to generate chart descriptions that can be used as `alt` text for chart images (for instance, when the data is plotted using the Canvas API [105] due to the dataset's size) [4, 23, 62, 81, 83, 100, 104]. Thus, extending our work to other chart types will only come later.

REFERENCES

[1] M. Akhtar, N. Subedi, V. Gupta, S. Tahmasebi, O. Cocarascu, and E. Simperl. ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In L. Ku, A. Martins, and V. Srikumar, eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 13921–13937. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.828 2

[2] A. M. Anis, H. Ali, and S. Sarfraz. On the Limitations of Vision-Language Models in Understanding Image Transforms, 2025. 2

[3] Answer.AI. FastHTML, 2024. 2

[4] K. Anukarnsakulchularp and D. Cook. Some Guidance for Writing alt text for Data Plots, 2024. Last accessed 8 August 2025. 4

[5] Artificial Analysis. gpt-oss-120B (high): API Provider Benchmarking & Analysis (Endpoint Evaluations), 2025. Last accessed 16 August 2025. 1

[6] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-VL Technical Report, 2025. 2

[7] A. Balaji, T. Ramanathan, and V. Sonathi. Chart-Text: A Fully Automated Chart Image Descriptor, 2018. 2

[8] L. Battle, P. Duan, Z. Miranda, D. Mukusheva, R. Chang, and M. Stonebraker. Beagle: Automated Extraction and Interpretation of Visualizations from the Web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–8. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174168 1

[9] A. Bendeck and J. Stasko. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1105–1115, 2025. doi: 10.1109/TVCG.2024.3456155 2

[10] L. Beyer. On the speed of ViTs and CNNs, 2024. Last accessed 17 August 2025. 2

[11] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai. PaliGemma: A versatile 3B VLM for transfer, 2024. 2

[12] J. A. Brown, L. Valade-DeMelo, E. Borawski, S. B. Robinson, and J. Dru. State of the Data Viz Industry survey. Technical Report 2024, Data Visualization Society, 2024. 1

[13] S. Chang, D. Palzer, J. Li, E. Fosler-Lussier, and N. Xiao. MapQA: A Dataset for Question Answering on Choropleth Maps. In *NeurIPS 2022 First Table Representation Workshop*. 2

[14] R. Chaudhry, S. Shekhar, U. Gupta, P. Maneriker, P. Bansal, and A. Joshi. LEAF-QA: Locate, Encode & Attend for Figure Question Answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3501–3510, 2020. doi: 10.1109/WACV45572.2020.9093269 2

[15] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Forty-first International Conference on Machine Learning*, 2024. 1

[16] S. S. Chintalapati, J. Bragg, and L. L. Wang. A Dataset of Alt Texts from HCI Publications: Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations in Scientific Papers. In J. Froehlich, K. Shinohara, and S. Ludi, eds., *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2022, Athens, Greece, October 23-26, 2022*, pp. 30:1–30:12. ACM, 2022. doi: 10.1145/3517428.3544796 2

[17] J. Cummings. DataVizQA, July 2024. 2

[18] K. Davila, R. Lazarus, F. Xu, N. Rodríguez Alcántara, S. Setlur, V. Govindaraju, A. Mondal, and C. V. Jawahar. CHART-Info 2024: A Dataset for Chart Analysis and Recognition. In A. Antonacopou-los, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal, eds., *Pattern Recognition*, pp. 297–315. Springer Nature Switzerland, Cham, 2025. 2

[19] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, J. Lu, T. Anderson, E. Bransom, K. Ehsani, H. Ngo, Y. Chen, A. Patel, M. Yatskar, C. Callison-Burch, A. Head, R. Hendrix, F. Bastani, E. Vander-Bilt, N. Lambert, Y. Chou, A. Chheda, J. Sparks, S. Skjonsberg, M. Schmitz, A. Sarnat, B. Bischoff, P. Walsh, C. Newell, P. Wolters, T. Gupta, K.-H. Zeng, J. Borchardt, D. Groeneveld, C. Nam, S. Lebrecht, C. Wittlif, C. Schoenick, O. Michel, R. Krishna, L. Weihs, N. A. Smith, H. Hajishirzi, R. Girshick, A. Farhadi, and A. Kembhavi. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, 2024. 2

[20] Devographics. State of Web Dev AI. Technical report, Devographics, 2025. 3

[21] H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. Wang, et al. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024. 1

[22] S. Edwardsson. Is Gemini 2.5 good at bounding boxes? Sort of..., 2025. Last accessed 17 August 2025. 2

[23] F. Elavsky and C. X. Bearfield. Playing telephone with generative models: "verification disability," "compelled reliance," and accessibility in data visualization, 2025. 4

[24] M. Friendly and D. Denis. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2):103–130, 2005. doi: 10.1002/jhbs.20078 1

[25] W. Fu-Hinthorn. Benchmarking Multi-Agent Architectures, 2025. Last accessed 19 August 2025. 1

[26] S. Gautam, M. A. Riegler, and P. Halvorsen. Point, Detect, Count: Multi-Task Medical Image Understanding with Instruction-Tuned Vision-Language Models. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 415–422. IEEE Computer Society, Los Alamitos, CA, USA, Jun 2025. doi: 10.1109/CBMS65348.2025.00090 2, 3

[27] Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 2

[28] Gemini Team, Google. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, 2025. 2

[29] S. Gillies, C. van der Wel, J. Van den Bossche, M. W. Taves, J. Arnott, B. C. Ward, and others. Shapely, May 2025. doi: 10.5281/zenodo.5597138 2

[30] Google. OpenAI compatibility, 2025. Last accessed 17 August 2025. 3

[31] K. Goswami, P. Mathur, R. Rossi, and F. Dernoncourt. ChartCitor: Answer Citations for ChartQA via Multi-Agent LLM Retrieval. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, p. 1668–1671. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3701716.3716886 1

[32] K. Goswami, P. Mathur, R. Rossi, and F. Dernoncourt. PlotGen: Multi-Agent LLM-based Scientific Data Visualization via Multimodal Retrieval Feedback. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, p. 1672–1676. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3701716.3716888 1

[33] D. Guo, F. Wu, F. Zhu, F. Leng, G. Shi, H. Chen, H. Fan, J. Wang, J. Jiang, J. Wang, J. Chen, J. Huang, K. Lei, L. Yuan, L. Luo, P. Liu, Q. Ye, R. Qian, S. Yan, S. Zhao, S. Peng, S. Li, S. Yuan, S. Wu, T. Cheng, W. Liu, W. Wang, X. Zeng, X. Liu, X. Qin, X. Ding, X. Xiao, X. Zhang, X. Zhang, X. Xiong, Y. Peng, Y. Chen, Y. Li, Y. Hu, Y. Lin, Y. Hu, Y. Zhang, Y. Wu, Y. Li, Y. Liu, Y. Ling, Y. Qin, Z. Wang, Z. He, A. Zhang, B. Yi, B. Liao, C. Huang, C. Zhang, C. Deng, C. Deng, C. Lin, C. Yuan, C. Li, C. Gou, C. Lou, C. Wei, C. Liu, C. Li, D. Zhu, D. Zhong, F. Li, F. Zhang, G. Wu, G. Li, G. Xiao, H. Lin, H. Yang, H. Wang, H. Ji, H. Hao, H. Shen, H. Li, J. Li, J. Wu, J. Zhu, J. Jiao, J. Feng, J. Chen, J. Duan, J. Liu, J. Zeng, J. Tang, J. Sun, J. Chen, J. Long, J. Feng, J. Zhan, J. Fang, J. Lu,

K. Hua, K. Liu, K. Shen, K. Zhang, K. Shen, K. Wang, K. Pan, K. Zhang, K. Li, L. Li, L. Li, L. Shi, L. Han, L. Xiang, L. Chen, L. Chen, L. Li, L. Yan, L. Chi, L. Liu, M. Du, M. Wang, N. Pan, P. Chen, P. Chen, P. Wu, Q. Yuan, Q. Shuai, Q. Tao, R. Zheng, R. Zhang, R. Zhang, R. Wang, R. Yang, R. Zhao, S. Xu, S. Liang, S. Yan, S. Zhong, S. Cao, S. Wu, S. Liu, S. Chang, S. Cai, T. Ao, T. Yang, T. Zhang, W. Zhong, W. Jia, W. Weng, W. Yu, W. Huang, W. Zhu, W. Yang, W. Wang, X. Long, X. Yin, X. Li, X. Zhu, X. Jia, X. Zhang, X. Liu, X. Zhang, X. Yang, X. Luo, X. Chen, X. Zhong, X. Xiao, X. Li, Y. Wu, Y. Wen, Y. Du, Y. Zhang, Y. Ye, Y. Wu, Y. Liu, Y. Yue, Y. Zhou, Y. Yuan, Y. Xu, Y. Yang, Y. Zhang, Y. Fang, Y. Li, Y. Ren, Y. Xiong, Z. Hong, Z. Wang, Z. Sun, Z. Wang, Z. Cai, Z. Zha, Z. An, Z. Zhao, Z. Xu, Z. Chen, Z. Wu, Z. Zheng, Z. Wang, Z. Huang, Z. Zhu, and Z. Song. Seed1.5-VL Technical Report, 2025. 2

[34] Y. Hao, J. Gu, H. W. Wang, L. Li, Z. Yang, L. Wang, and Y. Cheng. Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark, 2025. 2, 4

[35] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2 2

[36] A. Heakl, A. Sohail, M. Ranjan, R. Hossam, G. S. Ahmad, M. El-Geish, O. Maher, Z. Shen, F. Khan, and S. Khan. KITAB-Bench: A Comprehensive Multi-Domain Benchmark for Arabic OCR and Document Understanding, 2025. 2

[37] S. Hegde, P. Fazli, and H. Seifi. ChartQA-X: Generating Explanations for Visual Chart Reasoning, 2025. 2

[38] J. Holmer. Oxipng, 2016. 2

[39] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do LLMs Have Visualization Literacy? An Evaluation on Modified Visualizations to Test Generalization in Data Interpretation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2025. doi: 10.1109/TVCG.2025.3536358 2

[40] T.-Y. Hsu, C. L. Giles, and T.-H. Huang. SciCap: Generating Captions for Scientific Figures. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3258–3264. Association for Computational Linguistics, Punta Cana, Dominican Republic, Nov. 2021. doi: 10.18653/v1/2021.findings-emnlp.277 2

[41] L. Hu, D. Wang, Y. Pan, J. Yu, Y. Shao, C. Feng, and L. Nie. NovaChart: A Large-scale Dataset towards Chart Understanding and Generation of Multimodal Large Language Models. In *ACM Multimedia 2024*, 2024. 2

[42] K.-H. Huang, H. P. Chan, M. Fung, H. Qiu, M. Zhou, S. Joty, S.-F. Chang, and H. Ji. From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models. *IEEE Trans. on Knowl. and Data Eng.*, 37(5):2550–2568, Dec. 2024. doi: 10.1109/TKDE.2024.3513320 1

[43] K.-H. Huang, M. Zhou, H. P. Chan, Y. Fung, Z. Wang, L. Zhang, S.-F. Chang, and H. Ji. Do LVLMs Understand Charts? Analyzing and Correcting Factual Errors in Chart Captioning. In L.-W. Ku, A. Martins, and V. Srikumar, eds., *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 730–749. Association for Computational Linguistics, Bangkok, Thailand, Aug. 2024. doi: 10.18653/v1/2024.findings-acl.41 2

[44] Hugging Face. Prompt engineering, 2025. Last accessed 17 August 2025. 3

[45] K. Kafle, S. Cohen, B. Price, and C. Kanan. DVQA: Understanding Data Visualizations via Question Answering. In *CVPR*, 2018. 2

[46] S. E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, and Y. Bengio. FigureQA: An Annotated Figure Dataset for Visual Reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018. 1, 2

[47] A. Kanade and T. Ganu. Do You See Me : A Multidimensional Benchmark for Evaluating Visual Perception in Multimodal LLMs,

2025. 1

[48] Kanaries. VizGPT, 2025. Last accessed 17 August 2025. 1

[49] S. Kantharaj, X. L. Do, R. T. Leong, J. Q. Tan, E. Hoque, and S. Joty. OpenCQA: Open-ended Question Answering with Charts. In Y. Goldberg, Z. Kozareva, and Y. Zhang, eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11817–11837. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, Dec. 2022. doi: 10.18653/v1/2022.emnlp-main.811 2

[50] T. Kazmar. lap, 2016. 3

[51] S. G. Kim, J. Y. Choi, R. A. Rossi, E. Koh, and T. Y. Lee. Chart-to-Experience: Benchmarking Multimodal LLMs for Predicting Experiential Impact of Charts. In *18th IEEE Pacific Visualization Conference, PacificVis 2025, Taipei City, Taiwan, April 22-25, 2025*, pp. 340–345. IEEE, 2025. doi: 10.1109/PACIFICVIS64226.2025.00040 2

[52] Learn Prompting. N-Shot Prompting (Zero-Shot, One-Shot, Few-Shot, etc)), 2024. Last accessed 16 August 2025. 3

[53] M. Li, R. Zhang, J. Chen, J. Gu, Y. Zhou, F. Dernoncourt, W. Zhu, T. Zhou, and T. Sun. Towards Visual Text Grounding of Multimodal Large Language Model, 2025. 3

[54] V. R. Li, J. Sun, and M. Wattenberg. Does visualization help AI understand data?, 2025. 2

[55] Z. Li, D. Li, Y. Guo, X. Guo, B. Li, L. Xiao, S. Qiao, J. Chen, Z. Wu, H. Zhang, X. Shu, and S. Liu. ChartGalaxy: A Dataset for Infographic Chart Understanding and Generation, 2025. 2

[56] Z. Li, H. Miao, V. Pascucci, and S. Liu. Visualization Literacy of Multimodal Large Language Models: A Comparative Study. *CoRR*, abs/2407.10996, 2024. doi: 10.48550/ARXIV.2407.10996 2

[57] A. Liesenfeld and M. Dingemanse. Rethinking open source generative AI: open-washing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, p. 1774–1787. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3630106.3659005 4

[58] M. Lin, T. Xie, M. Liu, Y. Ye, C. Chen, and S. Liu. InfoChartQA: A Benchmark for Multimodal Question Answering on Infographic Charts, 2025. 2

[59] LLM-Core Xiaomi. The Evaluation Suite of Xiaomi MiMo-VL, 2025. 2

[60] L. Y. Lo and H. Qu. How Good (Or Bad) Are LLMs at Detecting Misleading Visualizations? *IEEE Trans. Vis. Comput. Graph.*, 31(1):1116–1125, 2025. doi: 10.1109/TVCG.2024.3456333 2

[61] S. Lu, Y. Li, Y. Xia, Y. Hu, S. Zhao, Y. Ma, Z. Wei, Y. Li, L. Duan, J. Zhao, Y. Han, H. Li, W. Chen, J. Tang, C. Hou, Z. Du, T. Zhou, W. Zhang, H. Ding, J. Li, W. Li, G. Hu, Y. Gu, S. Yang, J. Wang, H. Sun, Y. Wang, H. Sun, J. Huang, Y. He, S. Shi, W. Zhang, G. Zheng, J. Jiang, S. Gao, Y.-F. Wu, S. Chen, Y. Chen, Q.-G. Chen, Z. Xu, W. Luo, and K. Zhang. Ovis2.5 Technical Report, 2025. 2

[62] A. Lundgard and A. Satyanarayan. Accessible Visualization via Natural Language Descriptions. *IEEE Transactions on Visualization & Computer Graphics (Proc. IEEE VIS)*, 2022. doi: 10.1109/TVCG.2021.3114770 4

[63] M. Lysak, A. Nassar, N. Livathinos, C. Auer, and P. Staar. Optimized table tokenization for table structure recognition. In *International Conference on Document Analysis and Recognition*, pp. 37–50. Springer, 2023. 2

[64] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, 6(60):3139, 2021. doi: 10.21105/joss.03139 4

[65] R. Mahbub, M. S. Islam, M. T. R. Laskar, M. Rahman, M. T. Nayeem, and E. Hoque. The Perils of Chart Deception: How Misleading Visualizations Affect Vision-Language Models, 2025. 1

[66] A. Mahinpei, Z. Kostic, and C. Tanner. LineCap: Line Charts for Data Visualization Captioning Models, year=2022. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 35–39. doi: 10.1109/VIS54862.2022.00016 2

[67] Manus. Use cases from our official collection, 2025. Last accessed 17 August 2025. 1

[68] A. Masry, M. S. Islam, M. Ahmed, A. Bajaj, F. Kabir, A. Kartha, M. T. R. Laskar, M. Rahman, S. Rahman, M. Shahmohammadi, M. Thakkar, M. R. Parvez, E. Hoque, and S. Joty. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, eds., *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151. Association for Computational Linguistics, Vienna, Austria, July 2025. doi: 10.18653/v1/2025.findings-acl.978 1, 2

[69] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In S. Muresan, P. Nakov, and A. Villavicencio, eds., *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, Dublin, Ireland, May 2022. doi: 10.18653/v1/2022.findings-acl .177 1, 2

[70] F. Meng, W. Shao, Q. Lu, P. Gao, K. Zhang, Y. Qiao, and P. Luo. ChartAssisstant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training and Multitask Instruction Tuning, 2024. 2

[71] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. PlotQA: Reasoning over Scientific Plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[72] Microsoft. Playwright, 2019. 2

[73] Mistral AI. Mistral Evals, 2024. 2

[74] O. Moured, Y. Chen, R. Liu, S. Reiß, P. Torr, J. Zhang, and R. Stiefelhagen. CHAOS: Chart Analysis with Outlier Samples, 2025. 2

[75] O. Moured, S. A. Farooqui, K. Müller, S. Fadaeijouybari, T. Schwarz, M. Javed, and R. Stiefelhagen. Alt4Blind: A User Interface to Simplify Charts Alt-Text Creation. In *Computers Helping People with Special Needs: 19th International Conference, ICCHP 2024, Linz, Austria, July 8–12, 2024, Proceedings, Part I*, 2024. doi: 10.1007/978-3-031-62846-7_35 2

[76] O. Moured, J. Zhang, M. S. Sarfraz, and R. Stiefelhagen. AltChart: Enhancing VLM-Based Chart Summarization Through Multi-pretext Tasks. In E. H. Barney Smith, M. Liwicki, and L. Peng, eds., *Document Analysis and Recognition - ICDAR 2024*, pp. 349–366. Springer Nature Switzerland, Cham, 2024. 2

[77] S. Mukhopadhyay, A. Qidwai, A. Garimella, P. Ramu, V. Gupta, and D. Roth. Unraveling the Truth: Do VLMs really Understand Charts? A Deep Dive into Consistency and Robustness. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16696–16717. Association for Computational Linguistics, Miami, Florida, USA, Nov. 2024. doi: 10.18653/v1/2024.findings-emnlp.973 1, 2

[78] A. Nassar, A. Marafioti, M. Omenetti, M. Lysak, N. Livathinos, C. Auer, L. Morin, R. T. de Lima, Y. Kim, A. S. Gurbuz, et al. SmolDocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*, 2025. 2

[79] NuMind. NuMarkdown-8B-Thinking, July 2025. 3

[80] NVIDIA. Llama Nemotron VLM Dataset, August 2025. 2

[81] K. Nylund, J. Mankoff, and V. Potluri. MatplotAlt: A Python Library for Adding Alt Text to Matplotlib Figures in Computational Notebooks. *Computer Graphics Forum*, 44(3):e70119, 2025. doi: 10 .1111/cgf.70119 4

[82] J. Obeid and E. Hoque. Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In B. Davis, Y. Graham, J. Kelleher, and Y. Sripada, eds., *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 138–147. Association for Computational Linguistics, Dublin, Ireland, Dec. 2020. doi: 10.18653/v1/2020.inlg-1.20 2

[83] C. L. T. d. Oliveira, A. T. d. A. Silva, J. M. d. Morais, and M. P. Mota. Accessible Bar Charts Through Textual Description Templates. *Journal of the Brazilian Computer Society*, 29(1):1–18, Feb. 2023. doi: 10.5753/jbcs.2023.2301 4

[84] OpenAI. Introducing GPT-5, 2025. Last accessed 17 August 2025. 2

[85] Our World in Data and Jina AI. Our World In Data (OWID) Evaluation Dataset, July 2025. 2

[86] S. Pandey and A. Ottley. Benchmarking Visual Language Models on Standardized Visualization Literacy Tests. *Computer Graphics Forum*, 44(3):e70137, 2025. doi: 10.1111/cgf.70137 2

[87] A. Pawelec, V. S. Wesołowska, Z. Baczek, and P. Sankowski. PUB: Plot Understanding Benchmark and Dataset for Evaluating Large Language Models on Synthetic Visual Data Interpretation, 2024. 2

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2

[89] J. Phillips. daily-bench, 2025. 3

[90] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. 4

[91] L. Qiu and E. Chersoni. GenChaR: A Dataset for Stock Chart Captioning, 2024. 2

[92] R. Rahman, R. Hasan, A. A. Farhad, M. T. R. Laskar, M. H. Ashmafee, and A. R. M. Kamal. ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization of Long and Short Summaries. *Proceedings of the Canadian Conference on Artificial Intelligence*, jun 5 2023. https://caiac.pubpub.org/pub/ujhjycsw. 2

[93] P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, December 2024. 1, 2

[94] R. Ramachandran, A. Garjani, R. Bachmann, A. Atanov, O. F. Kar, and A. Zamir. How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks, 2025. 2, 3

[95] K. Robison. Meta got caught gaming AI benchmarks. *The Verge*, April 2025. Last accessed 17 August 2025. 1

[96] S. Ruan, R. Sheng, X. Wen, J. Wang, T. Zhang, Y. Wang, T. Dwyer, and J. Li. Qualitative Study for LLM-assisted Design Study Process: Strategies, Challenges, and Roles. *CoRR*, abs/2507.10024, 2025. doi: 10.48550/ARXIV.2507.10024 3

[97] T. Rädsch, L. Mayer, S. Pavicic, A. E. Kavur, M. Knopp, B. Öztürk, K. Maier-Hein, P. F. Jaeger, F. Isensee, A. Reinke, and L. Maier-Hein. Bridging vision language model (VLM) evaluation gaps with a framework for scalable and cost-effective benchmark generation, 2025. 2

[98] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, Data, and Designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018. doi: 10.1109/TVCG.2017.2744184 1

[99] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization & Computer Graphics (Proc. InfoVis)*, 2017. doi: 10.1109/tvcg.2016.2599030 2

[100] J. Schwabish, S. J. Popkin, and A. Feng. Do No Harm Guide: Centering Accessibility in Data Visualization. Technical report, Urban Institute, 2022. 4

[101] L. Shen, Qigqi, K. Ding, G. Meng, and S. Xiang. Rethinking Comprehensive Benchmark for Chart Understanding: A Perspective from Scientific Literature. *CoRR*, abs/2412.12150, 2024. doi: 10.48550/ARXIV.2412.12150 2

[102] M. Shen, Y. Li, L. Chen, and Q. Yang. From Mind to Machine: The Rise of Manus AI as a Fully Autonomous Digital Agent, 2025. 1

[103] S. Singh, Y. Nan, A. Wang, D. D'Souza, S. Kapoor, A. Üstün, S. Koyejo, Y. Deng, S. Longpre, N. A. Smith, B. Ermis, M. Fadaee, and S. Hooker. The Leaderboard Illusion, 2025. 1

[104] T. C. Smits, S. L'Yi, A. P. Mar, and N. Gehlenborg. AltGosling: automatic generation of text descriptions for accessible genomics data visualization. *Bioinformatics*, 40(12):btae670, 11 2024. doi: 10.1093/bioinformatics/btae670 1, 4

[105] B. Smus. Performance of canvas versus SVG, 2009. Last accessed 17 August 2025. 4

[106] B. J. Tang, A. Boggust, and A. Satyanarayan. VisText: A Benchmark for Semantically Rich Chart Captioning. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 2

[107] L. Tang, G. Kim, X. Zhao, T. Lake, W. Ding, F. Yin, P. Singhal, M. Wadhwa, Z. L. Liu, Z. Sprague, R. Namuduri, B. Hu, J. D. Rodriguez, P. Peng, and G. Durrett. ChartMuseum: Testing Visual Reasoning Capabilities of Large Vision-Language Models, 2025. 2

[108] X. Tang, A. Abdolrahmani, D. Gergle, and A. M. Piper. Everyday Uncertainty: How Blind People Use GenAI Tools for Information Access. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3713433 3

[109] C. Team, Z. Yue, Z. Lin, Y. Song, W. Wang, S. Ren, S. Gu, S. Li, P. Li, L. Zhao, L. Li, K. Bao, H. Tian, H. Zhang, G. Wang, D. Zhu, Cici, C. He, B. Ye, B. Shen, Z. Zhang, Z. Jiang, Z. Zheng, Z. Song, Z. Luo, Y. Yu, Y. Wang, Y. Tian, Y. Tu, Y. Yan, Y. Huang, X. Wang, X. Xu, X. Song, X. Zhang, X. Yong, X. Zhang, X. Deng, W. Yang, W. Ma, W. Lv, W. Zhuang, W. Liu, S. Deng, S. Liu, S. Chen, S. Yu, S. Liu, S. Wang, R. Ma, Q. Wang, P. Wang, N. Chen, M. Zhu, K. Zhou, K. Zhou, K. Fang, J. Shi, J. Dong, J. Xiao, J. Xu, H. Liu, H. Xu, H. Qu, H. Zhao, H. Lv, G. Wang, D. Zhang, D. Zhang, D. Zhang, C. Ma, C. Liu, C. Cai, and B. Xia. MiMo-VL Technical Report, 2025. 2

[110] G. V. Team, L. Karlinsky, A. Arbelle, A. Daniels, A. Nassar, A. Alfassi, B. Wu, E. Schwartz, D. Joshi, J. Kondic, N. Shabtay, P. Li, R. Herzig, S. Abedin, S. Perek, S. Harary, U. Barzelay, A. R. Goldfarb, A. Oliva, B. Wieles, B. Bhattacharjee, B. Huang, C. Auer, D. Gutfreund, D. Beymer, D. Wood, H. Kuehne, J. Hansen, J. Shtok, K. Wong, L. A. Bathen, M. Mishra, M. Lysak, M. Dolfi, M. Yurochkin, N. Livathinos, N. Harel, O. Azulai, O. Naparstek, R. T. de Lima, R. Panda, S. Doveh, S. Gupta, S. Das, S. Zawad, Y. Kim, Z. He, A. Brooks, G. Goodhart, A. Govindjee, D. Leist, I. Ibrahim, A. Soffer, D. Cox, K. Soule, L. Lastras, N. Desai, S. Ofekkoifman, S. Raghavan, T. Syeda-Mahmood, P. Staar, T. Drory, and R. Feris. Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence, 2025. 2

[111] V. Team, W. Hong, W. Yu, X. Gu, G. Wang, G. Gan, H. Tang, J. Cheng, J. Qi, J. Ji, L. Pan, S. Duan, W. Wang, Y. Wang, Y. Cheng, Z. He, Z. Su, Z. Yang, Z. Pan, A. Zeng, B. Wang, B. Chen, B. Shi, C. Pang, C. Zhang, D. Yin, F. Yang, G. Chen, J. Xu, J. Zhu, J. Chen, J. Chen, J. Chen, J. Lin, J. Wang, J. Chen, L. Lei, L. Gong, L. Pan, M. Liu, M. Xu, M. Zhang, Q. Zheng, S. Yang, S. Zhong, S. Huang, S. Zhao, S. Xue, S. Tu, S. Meng, T. Zhang, T. Luo, T. Hao, T. Tong, W. Li, W. Jia, X. Liu, X. Zhang, X. Lyu, X. Fan, X. Huang, Y. Wang, Y. Xue, Y. Wang, Y. Wang, Y. An, Y. Du, Y. Shi, Y. Huang, Y. Niu, Y. Wang, Y. Yue, Y. Li, Y. Zhang, Y. Wang, Y. Wang, Y. Zhang, Z. Xue, Z. Hou, Z. Du, Z. Wang, P. Zhang, D. Liu, B. Xu, J. Li, M. Huang, Y. Dong, and J. Tang. GLM-4.1V-Thinking and GLM-4.5V: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning, 2025. 2

[112] Unsloth. Vision Fine-tuning, 2025. Last accessed 17 August 2025. 4

[113] J. VanderPlas, B. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, A. Satyanarayan, E. Lees, I. Timofeev, B. Welsh, and S. Sievert. Altair: Interactive Statistical Visualizations for Python. *Journal of Open Source Software*, 3(32):1057, 2018. doi: 10.21105/joss.01057 4

[114] Vercel. State of AI. Technical Report Q1 2025, Vercel, 2025. 3

[115] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10. 1038/s41592-019-0686-2 2

[116] A. Visheratin. Breaking resolution curse of vision-language models, 2024. Last accessed 17 August 2025. 2

[117] A. Vo, K.-N. Nguyen, M. R. Taesiri, V. T. Dang, A. T. Nguyen, and D. Kim. Vision Language Models are Biased, 2025. 1, 2

[118] A. Vogel, O. Moured, Y. Chen, J. Zhang, and R. Stiefelhagen. RefChartQA: Grounding Visual Answer on Chart Images through Instruction Tuning, 2025. 2

[119] H. W. Wang, J. Hoffswell, S. M. Thazin Thane, V. S. Bursztyn, and C. X. Bearfield. How Aligned are Human Chart Takeaways and LLM Predictions? A Case Study on Bar Charts with Varying Layouts. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):536–546, 2025. doi: 10.1109/TVCG.2024.3456378 1

[120] S. Wang, S. Yang, W. Lin, Z. Guo, S. Cai, H. Huang, Y. Wang, J. Chen, and T. Jin. Omni-Chart-600K: A Comprehensive Dataset of Chart Types for Chart Understanding. In L. Chiruzzo, A. Ritter, and L. Wang, eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4051–4069. Association for Computational Linguistics, Albuquerque, New Mexico, Apr. 2025. doi: 10. 18653/v1/2025.findings-naacl.226 2

[121] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs, 2024. 2

[122] B. G. Weinstein, S. Marconi, M. Aubry-Kientz, G. Vincent, H. Senyondo, and E. P. White. DeepForest: A Python package for RGB deep learning tree crown delineation. *Methods in Ecology and Evolution*, 11(12):1743–1751, 2020. doi: 10.1111/2041-210X. 13472 3

[123] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, eds., *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010. doi: 10.25080/ Majora-92bf1922-00a 2

[124] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Dey, Shubh-Agrawal, S. S. Sandha, S. V. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum. LiveBench: A Challenging, Contamination-Free LLM Benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[125] Wikipedia contributors. Jaccard index — Wikipedia, the free encyclopedia, 2025. Last accessed 18 August 2025. 3

[126] Wikipedia contributors. Mean absolute error — Wikipedia, the free encyclopedia, 2025. Last accessed 18 August 2025. 3

[127] Wikipedia contributors. Precision and recall — Wikipedia, the free encyclopedia, 2025. Last accessed 18 August 2025. 3

[128] S. Willison. OpenAI reasoning models: Advice on prompting, 2025. Last accessed 17 August 2025. 3

[129] Y. Wu, L. Yan, L. Shen, Y. Mei, J. Wang, and Y. Luo. ChartCards: A Chart-Metadata Generation Framework for Multi-Task Chart Understanding, 2025. 2

[130] Y. Wu, L. Yan, L. Shen, Y. Wang, N. Tang, and Y. Luo. ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering. In Y. Al-Onaizan, M. Bansal, and Y. Chen, eds., *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 12174–12200. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.710 1, 2

[131] R. Xia, B. Zhang, H. Ye, X. Yan, Q. Liu, H. Zhou, Z. Chen, P. Ye, M. Dou, B. Shi, J. Yan, and Y. Qiao. ChartX & ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning, 2025. 2

[132] Y. Xu, L. Chen, L. Zhang, W. Wang, and Q. Jin. POLYCHARTQA: Benchmarking Large Vision-Language Models with Multilingual Chart Question Answering, 2025. 2

[133] Z. Xu, S. Du, Y. Qi, C. Xu, C. Yuan, and J. Guo. ChartBench: A Benchmark for Complex Visual Reasoning in Charts, 2024. 2

[134] Z. Xu and E. Wall. Exploring the Capability of LLMs in Performing Low-Level Visual Analytic Tasks on SVG Data Visualizations. In *IEEE Visualization and Visual Analytics, VIS 2024, St. Pete Beach, FL, USA, October 13-18, 2024*, pp. 126–130. IEEE, 2024. doi: 10. 1109/VIS55277.2024.00033 2

[135] Y. Yang, A. Patel, M. Deitke, T. Gupta, L. Weihs, A. Head, M. Yatskar, C. Callison-Burch, R. Krishna, A. Kembhavi, and C. Clark. Scaling Text-Rich Image Understanding via Code-Guided Synthetic Multimodal Data Generation. In W. Che, J. Nabende,

E. Shutova, and M. T. Pilehvar, eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17486–17505. Association for Computational Linguistics, Vienna, Austria, July 2025. doi: 10.18653/v1/2025.acl-long.855 2

[136] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*, 2024. 1, 2

[137] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, Y. Su, W. Chen, and G. Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark, 2025. 1, 2

[138] X. Zeng, H. Lin, Y. Ye, and W. Zeng. Advancing Multimodal Large Language Models in Chart Question Answering with Visualization-Referenced Instruction Tuning. *IEEE Trans. Vis. Comput. Graph.*, 31(1):525–535, 2025. doi: 10.1109/TVCG.2024.3456159 2

[139] K. Zhang, B. Li, P. Zhang, F. Pu, J. A. Cahyono, K. Hu, S. Liu, Y. Zhang, J. Yang, C. Li, and Z. Liu. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models, 2024. 2

[140] L. Zhang, A. Hu, H. Xu, M. Yan, Y. Xu, Q. Jin, J. Zhang, and F. Huang. TinyChart: Efficient Chart Understanding with Program-of-Thoughts Learning and Visual Token Merging. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1898. Association for Computational Linguistics, Miami, Florida, USA, Nov. 2024. doi: 10.18653/v1/2024.emnlp-main.112 2

[141] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. doi: 10.1109/TGRS.2024.3409624 2, 3

[142] Z. Zhang, W. Ma, and S. Vosoughi. Is GPT-4V (ision) All You Need for Automating Academic Data Visualization? Exploring Vision-Language Models' Capability in Reproducing Academic Charts. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8271–8288. Association for Computational Linguistics, Miami, Florida, USA, Nov. 2024. doi: 10.18653/v1/2024.findings-emnlp.485 2

[143] Y. Zhao, Z. Nasrullah, and Z. Li. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019. 2

[144] J. Zhu, Y. Zhou, Z. Wang, J. Yao, Y. Gu, Y. Yuan, and S. Liu. OrionBench: A Benchmark for Chart and Human-Recognizable Object Detection in Infographics, 2025. 2

[145] Z. Zhu, M. Jia, Z. Zhang, L. Li, and M. Jiang. MultiChartQA: Benchmarking Vision-Language Models on Multi-Chart Problems. In L. Chiruzzo, A. Ritter, and L. Wang, eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11341–11359. Association for Computational Linguistics, Albuquerque, New Mexico, Apr. 2025. doi: 10.18653/v1/2025.naacl-long.566 2