Plot Twist: Multimodal Models Don't Comprehend Simple Chart Details

Yasaman Razeghi ◇ **Ishita Dasgupta** ♠ **Fangyu Liu**♠
Vinay Ramasesh♠ **Sameer Singh**◇
◇University of California, Irvine ♠Google Deepmind
{yrazeghi, sameer}@uci.edu

Abstract

Recent advances in multimodal models show remarkable performance in real-world benchmarks for chart and figure understanding like ChartQA that involve interpreting trends, comparing data points, and extracting insights from visuals. In this paper, we investigate the extent to which these models truly comprehend the underlying information in charts by posing direct, elementary questions about simple features such as axes ranges and values to examine their fundamental visual understanding abilities in the context of charts. Our questions are applied to two sets of figures: synthetic and real-world. The empirical evaluation of 5 popular multimodal models on our dataset reveals shortfalls in understanding charts and figures, contrary to what their performance on complex benchmarks might suggest. For instance, Gemini Pro Vision only achieves 57.9% accuracy on our elementary set of questions on real-world plots, while other popular multimodal models showed similar or less performance. This work highlights an important limitation of current multimodal models, and cautions against overly optimistic interpretations of their abilities based on results of canonical evaluations.

1 Introduction

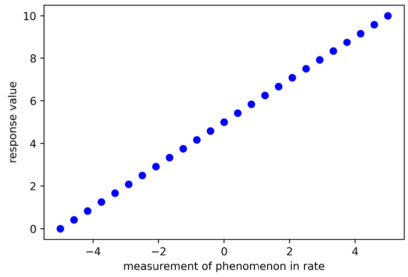
Assessing chart understanding capabilities offers a crucial benchmark for evaluating foundational models' reasoning skills beyond text. Significant efforts have been made to develop benchmarks for chart understanding, such as ChartQA, that features complex, human-written questions reflecting real-world applications (Methani et al., 2019; Masry et al., 2022). Multimodal models have recently made significant progress on these evaluation benchmarks (Gemini-Team et al., 2023; Chen et al., 2023; OpenAI et al., 2023). While these models perform well on complex tasks, how do they fare with more elementary aspects of chart understanding? Can they reliably answer basic questions about the chart?

Previous work has showed that real world datasets – while very useful for ensuring the practical application – can often contain statistical patterns such that model can do well without fully understanding the relevant capability (Goyal et al., 2017; McCoy et al., 2019). Moreover, testing basic capabilities can highlight important limitations of models that do not appear in complex benchmarks (Ribeiro et al., 2020). Complex capabilities examined in these real-world benchmarks, such as obtaining insights from visualizations, are also made up of many steps: understanding the image, domain knowledge, and reasoning. This makes it harder to diagnose the cause for failures.

In this work, we probe multimodal models to understand whether they can answer elementary questions about the specific visual content in charts. This is a core capability that is essential for any model claiming proficiency in chart comprehension. We evaluate this understanding by constructing elementary probing questions. These elementary questions include straightforward questions that measure fundamental skills like identifying axis extremes and extracting plot values on synthetic plots. We first pose these elementary questions on basic, synthetic plots. Then, we select a subset of real-world ChartQA test plots and pose our simple questions to them. This allows us to directly compare model performances on complex ChartQA queries versus performance on our elementary ones (examples in Figure 1).

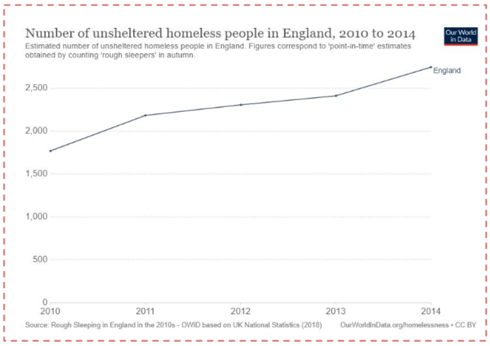
Our findings uncover shortcomings in these models regarding fundamental aspects of chart understanding. For example, PaLI-3 only gets 37.7% accuracy on our straightforward questions on the real-world plots. Moreover, other models such as Gemini Pro Vision and GPT-4V also get less than 60% performance. We further evaluate the robustness of these models and find that more powerful models, such as Gemini Pro Vision and GPT-4V, are often susceptible to the presence of text anno-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 5922–5937
November 12-16, 2024



What is the minimum x value among the set of points in the figure?	4 ✗
What is the minimum value for the range on the x axis?	0 ✗

(a) Synthetic plot



(chartQA question) When does the line reach the peak?	2014 ✓
(Our question) What is the minimum value for the range on the x axis?	0 ✗

(b) Real-world plot (from ChartQA)

Figure 1: An example of our evaluation method on PaLI-3. It shows a question in our Synthetic set (top) and a question in the ChartQA dataset with its corresponding question in our real-world subset below

tations over the actual data presented in the plots, which negatively affects their accuracy. This study highlights critical limitations of current multimodal models and underscores the importance of rigorous and thorough testing especially given limited public knowledge of the data used to train them. [1]

2 Setup

In the following section, we introduce our evaluation method, designed specifically to assess the understanding of elementary features in both synthetic and real-world chart and figure understanding by multimodal models.

Evaluation Method We propose a two-pronged evaluation approach. Using synthetic data for elementary questions allows us to identify failure modes in a cost-effective manner. Subsequently, we need to determine whether these failure modes propagate to real-world scenarios and applications. To facilitate this, we create a probing dataset containing two distinct subsets of plot-question pairs. The first subset, the basic synthetic plots, and elementary questions offers a controlled environment to scrutinize specific aspects of model performance. The second subset, the real-world plots, and the same elementary questions consist of real-world figures, for which we have randomly selected a set of plots from real-world images in the ChartQA test set, supplementing them with our straightforward elementary questions. While the synthetic subset is ideal for in-depth analysis and straightforward to expand, the real-world subset allows us to test whether our findings generalize to real-world charts, even though creating this subset requires more effort and resources due to the manual process involved. Note that even though our new dataset is human-generated, it was generated to explicitly contain simple questions that only require visual understanding. This controls for the bias that creeps into very open-ended human-generated datasets. Details on these probing plots and questions are provided in the Appendix A.

Models We evaluate multimodal models that demonstrated reasonable performance on already established chart understanding benchmarks such as ChartQA. We include Gemini Pro Vision (Gemini-Team et al., 2023), GPT-4V (OpenAI et al., 2023), PaLI-3 (Chen et al., 2023), ChartLlama (Han et al., 2023) and CogVLM (Wang et al., 2024) models in our empirical analysis.

Metrics In our evaluation framework, we employ a relaxed accuracy measure for numeric answers to accommodate minor inaccuracies following previous work (Methani et al., 2020; Masry et al., 2022; Liu et al., 2022, 2023a). Specifically, we deem a numerical answer correct if it falls within 5% relative range of the "gold standard" answer and for non-numerical answer we use exact matching. However, this accuracy metric does not have a symmetric error range for small vs large values – for example, it is much more restrictive for question querying the minimum values in comparison to those querying the maximum values. Recognizing that many of our simple questions often pertain to ranges, we adopt a range-based metric to evaluate models' answers. This metric, which we term "range-based accuracy," allows for a margin of error up to 5%

[1]The dataset used in this paper is available at Chart101.

Models↓	Standard		Range-Based	
	Acc	Collective Acc	Acc	Collective Acc
Gemini Pro Vision	$52.6 \pm 0.8\%$	$25.9 \pm 1.7\%$	$73.7 \pm 0.7\%$	36.5 ± 1.8 %
GPT-4V	$50.0 \pm 0.8\%$	$23.8 \pm 1.6\%$	$68.4 \pm 0.8\%$	33.4 ± 1.8 %
PaLI-3	$31.0 \pm 0.8\%$	8.0 ± 1.0 %	$43.1 \pm 0.8\%$	21.6 ± 1.7 %
ChartLlama	$10.6 \pm 0.5\%$	$0.1 \pm 0.1\%$	$21.3 \pm 0.7\%$	6.8 ± 0.9 %
CogVLM	$30.3 \pm 0.7\%$	5.5 ± 0.9 %	$47.7 \pm 0.8\%$	19.0 ± 1.5 %

Table 1: **Elementary Questions on Synthetic Plots:** The table displays accuracy rates using standard metrics in the left columns and Range-Based Accuracy in the right columns. These results highlight the overall low performance of models with simple chart understanding questions.

Model	Plot Type		
	bar	pie	scatter
Gemini Pro Vision	53.2	88.5	37.1
GPT-4V	42.2	87.2	40.3
CogVLM	27.3	51.9	23.4
PaLI-3	26.5	65.8	38.1
ChartLlama	9.8	26.9	4.2

Table 2: **Breakdown by Plot Types:** Range-Based accuracy on the different for synthetic plots.

Models↓	ChartQA Qs	Elementary Qs
Gemini Pro V	$67.4 \pm 2.5\%$	$57.9 \pm 1.5\%$
GPT-4V	$64.0 \pm 2.6\%$	$58.0 \pm 1.5\%$
PaLI-3	$69.7 \pm 2.5\%$	$37.7 \pm 1.5\%$
ChartLlama	$30.3 \pm 2.4\%$	$25.8 \pm 1.3\%$
CogVLM	$64.0 \pm 2.6\%$	$49.8 \pm 1.5\%$

Table 3: **Questions on Real Plots:** Overall standard accuracy on ChartQA Plots comparing original vs. our simple questions. The low performance of models on our elementary questions vs. complex questions on the same plots reveals that they struggle to answer simple questions on the same visual data. This discrepancy highlights a critical gap in their ability to consistently interpret visual information.

of the entire range under consideration. We also define collective accuracy; this metric assesses the correctness of responses to a full set of questions associated with a single figure as a single number. This metric underscores the model's capacity for a comprehensive and accurate interpretation of all the basic visual features we measure for that figure.

3 Results

Models struggle to answer basic synthetic chart questions reliably. We initially assess model performance on our synthetic subset. As demonstrated in Table 1, all of our models show poor performance on both versions of our accuracy metric; with the best model Gemini Pro Vision getting 52.6% accuracy. However, even for this model, the collective accuracy of 25.9% indicates a limited comprehensive understanding of these basic chart questions on the whole chart. Public models perform considerably worse, even ChartLlama, which is specialized explicitly for charts. These findings highlight the significance of our straightforward benchmark in pinpointing the limitations of current models. We analyze model accuracy by plot type to investigate the challenges with different plots. As Table 2 shows, models find scatter plot questions more challenging than pie chart questions. This is likely because answers for pie charts are often explicit in the figures (see Figure 6), whereas scatter plots require models to interpret min/max values or ranges using less explicit cues like x-ticks. Further analysis of challenging question types for each model is in Appendix C

Accuracy gap between elementary and complex questions on the same plots. We explore whether the difficulties models face with basic chart understanding questions in synthetic settings are also evident in real-world scenarios. We ask our elementary questions on a subset of the ChartQA test sets. The ChartQA images are chosen independently of the questions paired with them in the original dataset. In the right two columns of Table 3, we compare model performance on these simple questions to their performance on the original ChartQA questions, which involve more complex reasoning. As shown, there is often a high drop in performance, such as around 10% for Gemini Pro Vision. The low performance on elementary ques-

Type of Title	Gemini Pro Vision	GPT-4V
Correct Title	77.7 %	79.4 %
No Title	37.1 %	40.3 %
Misleading	32.4 %	32.5 %

Table 4: **Changing Titles in Plots:** Range-Based accuracy across different titles. The results highlight the impact of textual information on model performance.

tions particularly highlights concerns regarding the ability of models to answer simple questions on non-synthetic plots. This is problematic because it suggests that these models may struggle to handle basic tasks even in real-world scenarios, where accuracy and reliability are crucial.

4 Robustness Tests

One key aspect of chart understanding is the models' resilience to visual changes that do not affect the informational content but only the visual presentation, such as the choice of the plotting library or variations in phrasing. Our synthetic subset allows us to comprehensively evaluate model robustness against these changes. We make targeted visual modifications to charts to assess the models' ability to maintain accurate interpretation despite superficial alterations. This evaluation is crucial for determining the real-world utility and robustness of multimodal models in chart comprehension.

Model dependence on textual cues in plots. We compare three scenarios: 1) plots without any title (baseline), 2) plots with a title containing the correct answer, and 3) plots with a title providing misleading, incorrect answers. In all cases, models are instructed to base their answers on the figure itself. The results are presented in Table 4. Our findings reveal that including the correct answer in the plot title largely enhances model performance on our dataset, with an improvement of over 15% observed for both Gemini Pro Vision and GPT-4V models. When comparing the *misleading title* scenario to the *no title* scenario, we observe that Gemini Pro Vision, in particular, is swayed by the presence of misleading textual information, suggesting a bias towards text in the figure over an accurate understanding of the plot itself.

Model dependence on visual modifications. We make minor visual modifications to the synthetic figures while ensuring they convey the same information and then pose our questions. Modifications include altering scatter plot markers, introducing grids, or switching the data visualization library from Matplotlib to JavaScript. Results are displayed in Table 5. The top row shows the range-based accuracy for PaLI-3 on the scatter plot subset at 29.5%. The table reveals that even slight visual changes highly impact the model's performance in answering the same question. For instance, adding grids to the figures reduces accuracy to 21.3%, while changing the plot style from Matplotlib's default to JavaScript-amchart improves accuracy to 43.7%. These findings highlight model's lack of robustness, as its performance is greatly affected by such small visualization changes.

Category		Range Based Acc.
Original		29.5 %
Marker	x	25.5 %
	D	28.8 %
	O	31.5 %
	+	26.9 %
Grid		21.3 %
Plot	JS-plotly	36.9 %
	JS-highchart	39.7 %
	JS-amchart	43.7 %

Table 5: **Visual Changes in Plots:** PaLI-3 Performance on figures with same informational content but small visual changes. The variability in performance highlights a lack of robustness in chart understanding.

5 Related Work

Benchmarks for multimodal reasoning. With recent advancements in foundation multimodal models, extensive efforts have been made to create valuable evaluation benchmarks for assessing multimodal models in various domains such as math reasoning (Lu et al., 2024; Cherian et al., 2022), geometric reasoning (Kazemi et al., 2023; Lu et al., 2021), geometric reasoning for coding (Rismanchian et al., 2024), visual question answering on natural images (Liu et al., 2023b; Agrawal et al., 2016; Gurari et al., 2018), medical question answering (Zhang et al., 2023), hallucination detection (Guan et al., 2024; Li et al., 2023b) and comprehensive multimodal capabilities on real-world images (Yu et al., 2023; Aho and Ullman, 1972; Fu et al., 2024; Liu et al., 2023c; Li et al., 2023a; Xu et al., 2023). In this work, we delve into the chart understanding capabilities of foundational multimodal models, exploring a critical and valuable skill set.

Chart/Figure reasoning benchmarks. Specifically, there have been valuable efforts in developing benchmarks for chart understanding, ranging from synthetic benchmarks (Kafle et al., 2018; Singh and Shekhar, 2020) featuring yes/no questions (Kahou et al., 2018), to those focusing on understanding real-world charts (Masry et al., 2022; Methani et al., 2020; Xia et al., 2024). In this work, we show a method that bridges the gap between the synthetic creation of benchmarks for figure understanding and real-world applications within the benchmark. Our evaluation method demonstrates that assessing models on basic, fundamental questions about chart understanding can uncover crucial insights into model vulnerabilities. These vulnerabilities might be overlooked if evaluations are conducted solely on synthetic or real-world images.

Multimodal foundation models. Recently, there has been a significant emergence of generalist foundational multimodal models capable of answering questions about images and reasoning upon them. Closed-source models such as Gemini Pro Vision (Gemini-Team et al., 2023), GPT-4V (OpenAI et al., 2023) (OpenAI et al., 2023), PaLI-3 (Chen et al., 2023) stand alongside open-source counterparts like LLaVA-1.5 (Liu et al., 2023b), Mini-GPT4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023) and CogVLM (Wang et al., 2024). Additionally, there are models with a specific focus on chart understanding, such as MatCha (Liu et al., 2022), ChartLlAMA(Han et al., 2023) and ChartVLM (Xia et al., 2024). In this work, we evaluate all the aforementioned models that are accessible to us and have demonstrated even a slight capability for chart understanding.

6 Conclusion

In this paper, we present a diagnostic method for evaluating multimodal foundation models, with a focus on chart understanding capabilities. Our approach combines the precision of controlled synthetic evaluations with the real-world relevance of natural data scenarios. This dual strategy is essential in the current landscape, where models often lack transparency regarding training data and operate behind APIs. Our evaluation method complements real-world datasets like ChartQA, emphasizing the need to look beyond unified metrics and thoroughly assess models to identify their failure modes. By exposing subtle limitations, our method lays the groundwork for more effective benchmarks that accurately capture previously hidden model weaknesses. We believe this work will inspire further innovation in the field, promoting a holistic and nuanced approach to model evaluation.

7 Limitations

This study emphasizes the value of using direct, unit testing with real-world application evaluations for multimodal models in the context of chart understanding. While our approach effectively identifies clear limitations and challenges within these models, there are limitations to our study:
Lack of Proposed Solutions: While we identify various model limitations, our study does not offer specific solutions to these issues. Our insights are pivotal for pinpointing effective remedies.
Causes of the Shortcomings: One limitation of this study is the ambiguity regarding the precise causes behind the observed model shortcomings. Although we hypothesize that a distributional shift between the training data and our evaluation set might play a role, further investigation is needed to confirm this and understand this. We encourage continued research and improvement in the field, enhancing the robustness and applicability of multimodal models across various real-world tasks.

8 Acknowledgment

We would like to thank the members of UCI-NLP, for valuable discussions and feedback on this work. This material is sponsored in part by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research, the NSF CAREER award number IIS-2046873 and the NSF award IIS-2008956. Yasaman Razeghi has been doing a part-time internship at Google DeepMind while working on this paper.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil

Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023. Pali-3 vision language models: Smaller, faster, stronger. *Preprint*, arXiv:2310.09199.

Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, K Smith, and Joshua B Tenenbaum. 2022. Are deep neural networks smarter than second graders?. arxiv.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Gemini-Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Kli-

menko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo

Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *Preprint*, arXiv:1802.08218.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *Preprint*, arXiv:2311.16483.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. *Preprint*, arXiv:1801.08163.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos.

2022. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *Preprint*, arXiv:2105.04165.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2019. Data interpretation over plots. *CoRR*, abs/1909.00997.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sina Rismanchian, Yasaman Razeghi, Sameer Singh, and Shayan Doroudi. 2024. Turtlebench: A visual programming benchmark in turtle geometry. ArXiv preprint.

Hrituraj Singh and Sumit Shekhar. 2020. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284, Online. Association for Computational Linguistics.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *Preprint*, arXiv:2402.12185.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *Preprint*, arXiv:2306.09265.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *Preprint*, arXiv:2305.10415.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

A Data Creation

Synthetic Plot-question This subset comprises a series of scatter plots, bar plots and pie charts each created using the Matplotlib or javascript libraries. These plot types are identical to those found in the ChartQA dataset, ensuring consistency and relevance in our analysis. These include scatter plots, bar charts, and pie charts. We have then carefully formulated a set of fundamental questions for each plot type aimed at basic visual understanding. We create all primary subsets of the synthetic data using the Matplotlib library. We begin by automatically generating 50 plots for each subset, followed by a manual review of each plot to ensure they meet our quality standards and are free from ambiguity. The number of questions for each subset is in Table 7.

Scatter Plots We depict straightforward mathematical functions, such as $x = y$ $x = 2x$, etc., each represented using 25 default blue marker points in scatter plots for the main subset. Each plot is accompanied by eight corresponding direct simple questions focusing on the minimum and maximum values and ranges for the x and y axes. These questions are detailed in Table 6.

Bar Charts We automatically create bar charts with the default blue Matplotlib library. We randomly sample the number of bars for each chart, ranging from 1 to 5, and assign the values of each bar randomly within the range of [-200, 200]. Each plot is accompanied by five corresponding direct simple questions focusing on the minimum and maximum values of the bars and ranges for the y axes. These questions are detailed in Table 6.

Pie Charts We automatically generate pie charts with the number of categories randomly sampled between 1 and 10. We ensure that each pie chart represents a total sum value of 100%, which is the most common use case for pie charts. The actual values are explicitly written within the categories. These questions are detailed in Table 6.

Real World Plots In this subset, we randomly sampled plots from the ChartQA test set and adapted our questions to these plots. The questions were minimally edited to ensure each question's relevance to the specific chart type and context. The ground truth answers were then included. During the annotation process, we randomly selected plots from the ChartQA test sets and ensured that our added questions meet two criteria: 1. They involve minimal modifications from our set of questions in the synthetic set, and 2. They are devoid of ambiguity. This subset is created manually, resulting in 218 questions on 70 different plots from the ChartQA human-annotated test set. For comparison, examples of comparison between our additional Questions and ChartQA test questions are presented in Figure 4.

Robustness Tests Subsets Our primary dataset, as previously outlined, consists of two distinct subsets. These subsets bridge synthetic chart understanding questions with real-world scenarios, aiming to evaluate models' fundamental abilities to comprehend charts. Another fundamental aspect of image comprehension, especially with charts, is the resilience of models to invariant visual changes in the plots. These alterations do not modify the charts' informational content but solely affect their visual presentation, such as the libraries used for plot creation or color variations. Our synthetic subset specifically facilitates a comprehensive evaluation of model robustness against these changes. By introducing targeted modifications to visual aspects of charts, we can assess model performance in maintaining accurate interpretation despite superficial alterations. This evaluation is crucial for determining the real-world utility and robustness of multimodal models in chart comprehension. To that end, we create multiple additions to the subset, changing one specific visual parts of the charts to study the models robustness to such changes. These edits include changing the choice of plot library, changing the markers of the plots, adding or removing grids, adding misleading text in the charts etc.

B Experimental Details

All the experiments in this paper are performed in April 2024.

B.1 Sampling Method

Our sampling method for assessing model performance involves querying each model five times with the same set of questions and averaging the obtained metrics to mitigate the effects of nondeterminism inherent in model responses. For the models PaLI-3, Gemini Pro Vision, and GPT-4V, we utilize the default temperature setting to closely mirror their typical usage in real-world applications. Conversely, for the models CogVLM and ChartLlama, we adjust the temperature to 0.7, based on preliminary tests indicating optimized performance at this setting. This method ensures that our evaluation reflects both the robustness and the real-world applicability of these multimodal models.

B.2 Model Sizes

We include Gemini Pro Vision (Gemini-Team et al., 2023), GPT-4V (OpenAI et al., 2023), PaLI-3 (Chen et al., 2023), ChartLlama (Han et al., 2023) and CogVLM (Wang et al., 2024) models in our empirical analysis. We use Gemini 1.0 Pro Vision, and GPT-4V through their APIs. All experiments for these two models are performed in the first week of April 2024 (mentioning the data as the models behind APIs can change over time). We use ChartLlama-13B and CogVLM-17B. The PaLI-3 model is of size 5B.

B.3 Prompts

For all our question-answering tasks, we use the prompt "Answer the question based on the Figure + [Question]." for PaLI-3, CogVLM and ChartlLamma. For the robustness test of exploring models' dependence on textual cues in the plot, we further emphasize the figure by changing the prompt to "Answer the question only based on the figure + [Question]." For automated extraction of the answers, we instruct Gemini Pro Vision and GPT with another prompt as presented in Figure 2.

> Please answer the following question based on the plot/figure with the response in the same format: `"The answer is ANS. I hope the answer is correct."` In which the ANS represents the correct answer. Be concise and accurate in your reply.

Figure 2: Prompt for Gemini and GPT Models

B.4 Automated Evaluation

To facilitate automated evaluation, we instruct all models to format their respon'ses in a specific structure: "The answer is ANSWER." While Gemini Pro Vision and GPT-4V consistently adhere to this format, the other models frequently deviate from it. To address this inconsistency, we employ the GPT-3 turbo model to reformat the responses into the required structure before extracting the answers. This additional step ensures uniformity in response formatting across all tested models, enabling more accurate automated analysis. Prompt is in Figure 3

Synthetic subset	Questions
scatter plots	What is the maximum/minimum value among the set of points plotted in the figure? What is the approximate maximum/minimum value for the range on the x/y-axis?
bar chart	What is the maximum/minimum value among the set of bars plotted in the figure? How many bars are in the figure? What is the maximum/minimum value for the range on the y-axis?
pie chart	What is the maximum/minimum value among the set of categories in the figure? How many categories are presented?

Table 6: Question templates for synthetic dataset

Synthetic subset	Number of Questions
scatter plots	336
bar chart	250
pie chart	141

Table 7: Question templates for synthetic dataset

C More Analysis

What type of questions are the hardest? Following the approach of segmenting model performance by plot types, our use of synthetic data facilitates a similar analysis based on question types. This approach enables us to evaluate and pinpoint the particular performance characteristics of each model, allowing for a detailed investigation into the distinct behavioral patterns and challenges models exhibit when responding to different kinds of questions. The outcomes of this question-type-specific performance evaluation are presented in Figures 9 and 8, which detail the distinct range-based accuracy and response patterns of each model across the range of question types examined. For example, we first observe distinct behaviors among the models: while GPT-4V demonstrates its strongest performance on questions concerning the minimum range on the x-axis, Gemini Pro Vision and PaLI-3 struggle the most with this type of question when dealing with scatter plots. As discussed earlier, these detailed insights into models' specific limitations are vital for understanding the reliability of model responses and guiding the development of effective solutions.

Extract the concise answer from the model's response as shown in the examples below, making sure the answer is in this format: "The answer is ANS. I hope the answer is correct."

Example 1:

Question: "How many food items are shown in the bar graph?"
Model Answer: "<extra_id_0> 0"
Extracted Answer: The answer is 0. I hope the answer is correct.

Example 2:

Question: "How many bars are in the figure?"
Model Answer: "<extra_id_0> There are three bars in the figure."
Extracted Answer: The answer is three. I hope the answer is correct.

Example 3:

Question: "Find missing data of the sequence 24, _, 32, 33, 42?"
Model Answer: "<extra_id_0> 33"
Extracted Answer: The answer is 33. I hope the answer is correct.

Example 4:

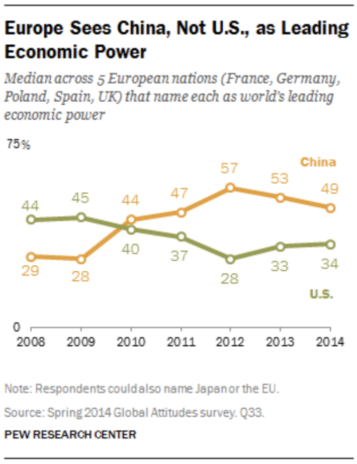
Question: "Which country has the highest secondary graduation rate in 2018?"
Model Answer: "<extra_id_0> Italy"
Extracted Answer: The answer is Italy. I hope the answer is correct.

Your Task:

Given the question and model answer below, extract the concise answer.

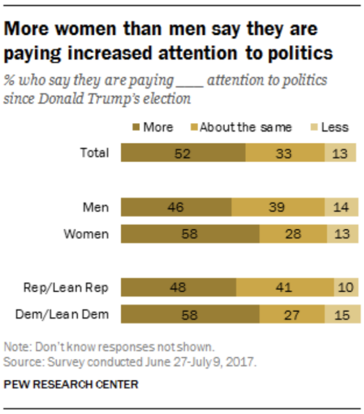
Question: "{question}"
Model Answer: "{model_raw_output}"
Extracted Answer:

Figure 3: Prompt for Extraction of Answers for Automated Evaluation



ChartQA: What is the greatest gap value between the orange and the green lines?

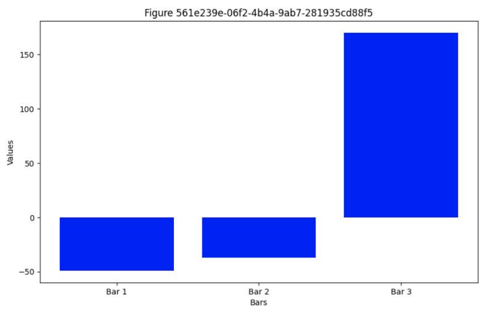
Our added Q: What is the maximum approximate value for the range on the x-axis?



ChartQA: What's the total value of the More bar?

Our added Q: How many bars are plotted in the figure?

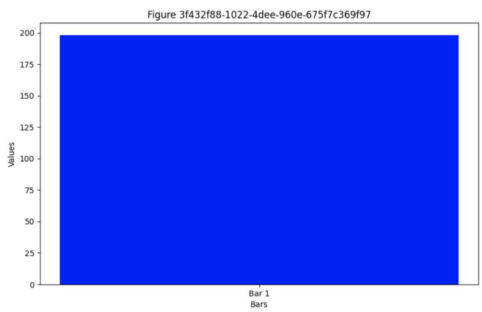
Figure 4: More examples of questions from our simplified dataset alongside those from the ChartQA dataset, with corresponding questions from our set. ChartQA human-written questions vary in complexity, from the more straightforward at the bottom of the example to those requiring complex reasoning at the top. In contrast, our questions are consistently structured to be simple.



Q1: What is the maximum value among the set of bars plotted in the figure?
Accepted Range of Answers for Accuracy: [161.5, 178.5]
Accepted Range of Answers for Range-Based Accuracy: [158, 182]

Q2: What is the minimum value among the set of bars plotted in the figure?
Accepted Range of Answers for Accuracy: [-51.45, -46.55]
Accepted Range of Answers for Range-Based Accuracy: [-61, -37]

Q3: How many bars are in the figure?
Accepted Answer for Accuracy: 3
Accepted Answer for Range-Based Accuracy: 3

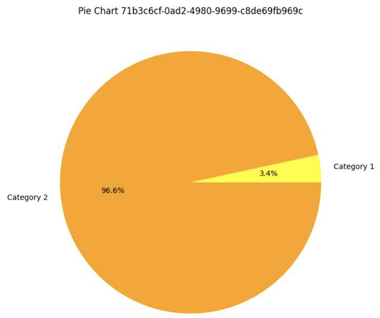


Q1: What is the maximum value among the set of bars plotted in the figure?
Accepted Range of Answers for Accuracy: [188.1,207.9]
Accepted Range of Answers for Range-Based Accuracy: [178.6,208.4]

Q2: What is the minimum value among the set of bars plotted in the figure?
Accepted Range of Answers for Accuracy: [188.1,207.9]
Accepted Range of Answers for Range-Based Accuracy: [178.6,208.4]

Q3: How many bars are in the figure?
Accepted Answer for Accuracy: 1
Accepted Answer for Range-Based Accuracy: 1

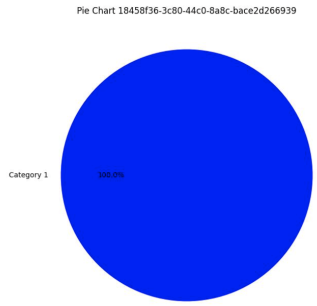
Figure 5: Examples of questions from our synthetic dataset for the barcharts



Q1: What is the maximum value among the set of categories in the figure?
Accepted Range of Answers for Accuracy: [91.7, 101.4]
Accepted Range of Answers for Range-Based Accuracy: [91.6, 101.6]

Q2: What is the minimum value among the set of categories in the figure?
Accepted Range of Answers for Accuracy: [3.23, 3.57]
Accepted Range of Answers for Range-Based Accuracy: [0, 8.4]

Q3: How many categories are presented?
Accepted Answer for Accuracy: 2
Accepted Answer for Range-Based Accuracy: 2

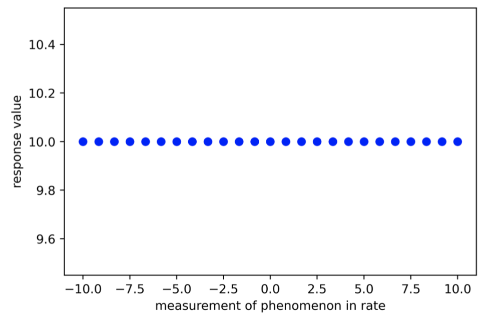


Q1: What is the maximum value among the set of categories in the figure?
Accepted Range of Answers for Accuracy: [95, 100]
Accepted Range of Answers for Range-Based Accuracy: [95, 100]

Q2: What is the minimum value among the set of categories in the figure?
Accepted Range of Answers for Accuracy: [95, 100]
Accepted Range of Answers for Range-Based Accuracy: [95, 100]

Q3: How many categories are presented?
Accepted Answer for Accuracy: 1
Accepted Answer for Range-Based Accuracy: 1

Figure 6: Examples of questions from our synthetic dataset for the piecharts



Q1: What is the minimum x value among the set of points plotted in the figure?
Accepted Range of Answers for Accuracy: [-10.5, -9.5]
Accepted Range of Answers for Range-Based Accuracy: [-11.1, -8.9]

Q2: What is the maximum x value among the set of points plotted in the figure?
Accepted Range of Answers for Accuracy: [9.5, 10.5]
Accepted Range of Answers for Range-Based Accuracy: [8.9, 11.1]

Q3: What is the minimum y value among the set of points plotted in the figure?
Accepted Range of Answers for Accuracy: [9.5, 10.5]
Accepted Range of Answers for Range-Based Accuracy: [9.95, 10.05]

Q4: What is the maximum y value among the set of points plotted in the figure?
Accepted Range of Answers for Accuracy: [9.5, 10.5]
Accepted Range of Answers for Range-Based Accuracy: [9.95, 10.05]

Q5: What is the approximate minimum value for the range on the y-axis?
Accepted Range of Answers for Accuracy: [9.01, 9.96]
Accepted Range of Answers for Range-Based Accuracy: [9.43, 9.54]

Q6: What is the approximate maximum value for the range on the y-axis?
Accepted Range of Answers for Accuracy: [10.02, 11.07]
Accepted Range of Answers for Range-Based Accuracy: [10.49, 10.65]

Q7: What is the approximate minimum value for the range on the x-axis?
Accepted Range of Answers for Accuracy: [-11.55, -10.45]
Accepted Range of Answers for Range-Based Accuracy: [-12.1,-9.9]

Q8: What is the approximate maximum value for the range on the x-axis?
Accepted Range of Answers for Accuracy: [10.45, 11.55]
Accepted Range of Answers for Range-Based Accuracy: [9.9, 12.1]

Figure 7: Examples of questions from our synthetic dataset for the scatterplot

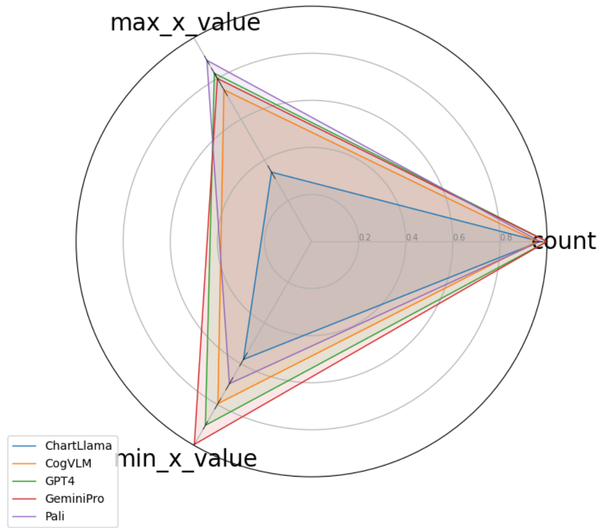
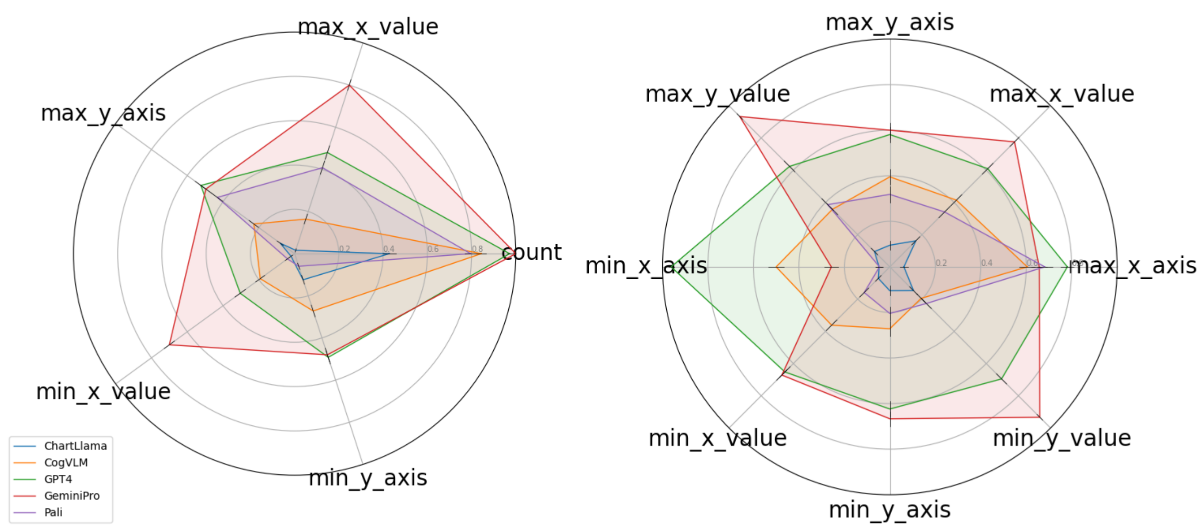


Figure 8: Radar chart depicting the range-based accuracy of different models in response to various question types in our pie charts.



(a) Synthetic Bar Chart Questions

(b) Synthetic Scatter Plot Questions

Figure 9: Radar chart depicting the range-based accuracy of different models in response to various question types, highlighting the distinct limitations each model exhibits with respect to specific types of questions. ,