Understanding Graphical Perception in Data Visualization through Zero-shot Prompting of Vision-Language Models

Grace Guo *, **Jenna Jiayi Kang** *, **Raj Sanjay Shah** *,
Hanspeter Pfister, Sashank Varma
Harvard University, New York University, Georgia Institute of Technology

Abstract

Vision Language Models (VLMs) have been successful at many chart comprehension tasks that require attending to both the images of charts and their accompanying textual descriptions. However, it is not well established how VLM performance profiles map to human-like behaviors. If VLMs can be shown to have human-like chart comprehension abilities, they can then be applied to a broader range of tasks, such as designing and evaluating visualizations for human readers. This paper lays the foundations for such applications by evaluating the accuracy of zero-shot prompting of VLMs on graphical perception tasks with established human performance profiles. Our findings reveal that VLMs perform similarly to humans under specific task and style combinations, suggesting that they have the potential to be used for modeling human performance. Additionally, variations to the input stimuli show that VLM accuracy is sensitive to stylistic changes such as fill color and chart contiguity, even when the underlying data and data mappings are the same.

1 Introduction and Related Work

Vision Language Models (VLMs) are capable of synthesizing information in both the vision and language input modalities, leading to their application in healthcare diagnostics (19), autonomous vehicles (16), interactive robotic applications (24), and other domains. In our domain of interest, *data visualization*, VLMs have also been used for a range of tasks that require attending to both the images of charts and graphs and their accompanying textual descriptions (5, 12, 13, 14, 23), from simple tasks such as data extraction (14) and question answering (2, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18, 20, 22) to more complex tasks such as chart generation and refinement (5).

Recent research has evaluated whether VLMs show human-like visualization comprehension abilities using visualization literacy tests (1). Such tests consist of questions that measure the ability of humans to comprehend and extract information from visualizations. Studies with GPT-4 show that it can reason about visualizations, identify trends, and suggest best design practices. Yet, the model struggles with simple tasks like value retrieval and color distinctions in charts. If VLMs show human-like visualization comprehension abilities, they can be used to design and evaluate visualizations, e.g., identifying potential sources of cognitive processing (over)load. However, doing so requires establishing that VLM performance profiles map to human-like behaviors. *Here, we lay the foundations for such applications by evaluating the accuracy of VLMs when performing graphical perception tasks.*

* Equal contribution.
Email: gguo31@g.harvard.edu, jennakang@nyu.edu, rajsanjayshah@gatech.edu, pfister@seas.harvard.edu, varma@gatech.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

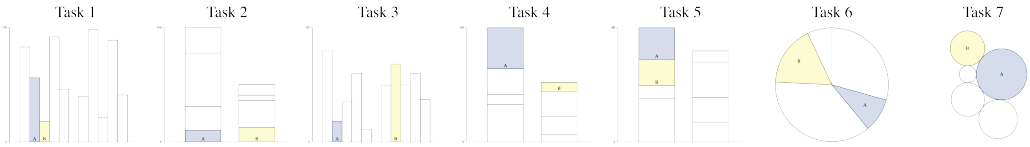


Figure 1: Examples of the seven tasks in our study, adapted from (6). For each visualization, the VLM was prompted to compare the two segments in blue and yellow (also labeled A and B, respectively).

Graphical perception tasks require elementary perceptual operations such as retrieving numerical values from positional encodings, lengths, and angles. They were introduced by Cleveland and McGill in 1984 (3), in a series of experiments where participants were asked to extract two numerical quantities from a chart and to judge the proportion of the smaller quantity against the larger (see Tasks 1-6 of Fig. 1). Heer and Bostock (6) later replicated this study with a larger pool of participants recruited from the crowd-sourcing platform MTurk. They also extended the stimuli to include other types of judgment tasks, such as area judgments (Fig. 1, Task 7). *These studies revealed potential sources of cognitive processing load in the complexity of common visualizations.* Inspired by these studies, other prior work have used the same stimuli to investigate relational reasoning in CNNs compared to human performance (4). We extend these experiments to evaluate the human-like performance of popular VLMs in a zero-shot, out-of-the-box manner.

In this work, we evaluate whether VLMs can simulate human graphical perception performance when performing the same seven tasks from these seminal studies (3, 6). To do so, we first recreated the original stimuli, implementing 45 trials for each of the seven chart types shown in Figure 1. Each trial included a visualization with two segments highlighted. We then *zero-shot prompted* the GPT-4o-mini model (17) to 1) indicate which segment is smaller, and 2) estimated percentage of the smaller segment is the larger, in a procedure similar to the one used by Heer and Bostock (6). Overall, our contributions are:

- **Behavioral evaluation of VLMs on graphical perception tasks:** We assess whether GPT-4o-mini can simulate human-like behaviors by comparing the accuracy and confidence of the VLM in interpreting visualizations against human performance profiles (3, 6).
- **Model performance across prompts:** We use four prompt variations to test the suitability of VLMs for modeling human graphical perception – with and without references to the target segment colors, and with and without generation of explanations/reasonings in the output template (21).
- **Model performance across stimuli:** We introduce variations in the stimuli as shown in Figure 3a to test how incidental factors influence the model's performance in interpreting visual data.
- **Model performance on new tasks:** We implement novel task variants, shown in Figure 3b, and evaluate whether VLMs show a performance decrement when the critical elements are contiguous.

2 Method

Our work adapts the stimuli and tasks from two prior human studies to evaluate the behavioral alignment of the graphical perception abilities of VLMs. To ensure the comparability of results across studies, we recreate the stimuli and prompt the VLM with the same probes in a zero-shot manner. Information for stimulus generation was taken from both (3, 6), whereas the text of the prompts was referenced from the experimental materials of (6). Our study included seven tasks from these studies plus two new variants. Each task consisted of 45 distinct trials.

Stimuli and Tasks. To create the stimuli (i.e., visualizations), we first generated ten numerical values using the formula from Cleveland and McGill (3):

$$s_i = 10 \times 10^{(i-1)/12}, i = 1, 2, ..., 10 \tag{1}$$

We then constructed all 45 possible unique pairs of these values. The ratios of these pairs ranged from 0.18 to 0.83. For each of the seven tasks from the original studies (3, 6), we generated 45 visualizations corresponding to these pairs. In each visualization, the segments encoding the values being compared were colored blue and yellow and also labeled "A" and "B" (Figure 1). All other values in the visualization (i.e., values not being compared) were generated randomly, with a few

constraints. For instance, the bottom of the bar segments being compared in Task 4 had to be unaligned; otherwise, the perceptual task would essentially become identical to Task 2.

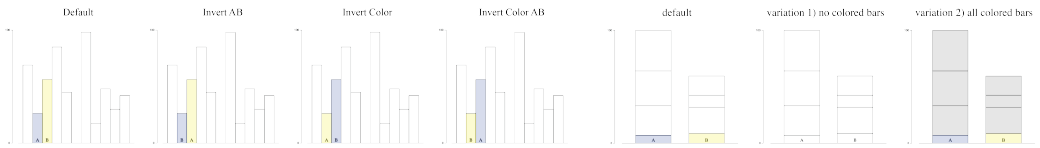
3 Experiments 1 and 2

Experiment 1. For each trial, the VLM was given a visualization and asked to respond to the probes:

1. Which of the two, blue (A) or yellow (B), shapes is smaller?
2. What percentage is the SMALLER marked shape of the LARGER? Enter a % between 0 and 100.

We vary the framing of the prompts in two ways. The first is the explicit mention of color in the probe (no color/has color). The "has color" prompt contains references to the colors of the two labeled segments. The second is requesting explanations in the model response (no explanation/has explanation). The "has explanation" prompt asks the model not just to provide an answer (such as identifying which visual element is smaller or the proportion between two segments), but also to generate the reasoning behind its decision. See Appendix 8 for the exact inputs to the VLM.

Finally, since VLMs may exhibit bias towards left/right layouts and A/B labels, we added three stimuli variations that inverted the order of colors and A/B labels (Figure 2a).



(a) (b)

Figure 2: (a) Experiment 1 inverted colors and AB labels. (b) Stimulus variations in Experiment 2.

Experiment 2. In addition to the stimuli used in prior studies (i.e., Figure 1), we created two new variations of all ($7 \times 45 =$) 315 visualizations and evaluated VLM comprehension of these stimuli using the same probes and prompt framings. These variations are, first, no colored bars, and second, all colored bars; Figure 2b). Other than fill color changes, the stimuli in the variations were identical.

3.1 Results

Prompts		Stimuli	Invert		Tasks								Data fit	
Clr.	Exp.		AB	Clr.	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Overall	ρ	p
		Def.			1.00	0.89	1.00	0.74	0.69	0.75	0.63	0.81	0.90	0.006
✗	✗	All clr.	✗	✗	0.88	0.44	0.92	0.58	0.53	0.62	0.62	0.66	0.18	0.699
		No clr.			0.53	0.44	0.44	0.53	0.53	0.38	0.44	0.47	0.19	0.679
		Def.			1.00	0.96	1.00	0.82	0.69	0.82	0.73	0.86	0.89	0.007
✗	✓	All clr.	✗	✗	0.96	0.56	0.91	0.76	0.58	0.80	0.69	0.75	0.39	0.383
		No clr.			0.44	0.49	0.51	0.47	0.58	0.60	0.49	0.51	0.52	0.229
		Def.			1.00	0.82	1.00	0.76	0.82	0.71	0.71	0.83	0.73	0.060
✓	✗	All clr.	✗	✗	0.78	0.47	0.80	0.58	0.62	0.67	0.69	0.66	0.07	0.879
		No clr.			0.62	0.44	0.47	0.44	0.56	0.38	0.44	0.48	0.37	0.413
		Def.			1.00	0.91	1.00	0.75	0.73	0.91	0.73	0.87	0.84	0.017
✓	✓	All clr.	✗	✗	0.87	0.56	0.91	0.64	0.56	0.87	0.80	0.75	0.22	0.638
		No clr.			0.56	0.51	0.44	0.49	0.56	0.51	0.40	0.50	0.36	0.423
			✓	✗	0.98	0.78	0.98	0.78	0.58	0.87	0.98	0.85	0.19	0.688
✓	✓	Def.	✗	✓	1.00	0.93	1.00	0.82	0.73	0.58	0.69	0.82	0.83	0.021
			✓	✓	1.00	0.83	1.00	0.74	0.71	0.86	1.00	0.88	0.22	0.632

Table 1: Accuracy of GPT-4o-mini on probe 1 ("smaller than") judgments. For each combination of prompt and stimuli variation, we calculate Spearman's rank correlation relative to human judgments reported in Figure 4 of Heer and Bostock (6), taking negative of ρ due to the opposite rankings that log error and accuracy yield. **Prompt variations - Clr:** Has color; **Exp:** Has explanation; **Stimuli variations - Def:** Default stimuli; **All clr:** All color stimuli; **No clr:** No color stimuli. See Figure 3a for examples of **Def, All clr, and No clr**. **Inversion variations - AB:** Invert labels for A and B ; **Clr** Invert colors associated with A and B.

We test both probes by Direct Probing to elicit VLM judgments, where we ask the VLM about its current state.

Table 1 shows GPT-4o-mini accuracies on probe 1 across the seven tasks, four prompt framings, three stimuli variations, and three combinations of inverted color and A/B labels explored in experiments 1 and 2. Note that for the inverted color and A/B labels conditions, we only looked at prompts with both color and requested explanations (**Has Color, Has explanation**), and the default stimuli condition (Figure 2b, left).

The key takeaways are as follows:

Experiment 1 - Prompt Sensitivity: Overall, the model performed best in the **Has Color, Has Explanation** prompt condition. Removing either cue from the prompt (either no color or no explanation) led to a small drop in model accuracy. Removing both color and explanations led to a substantial decline, as seen in the **No Color, No Explanation** condition. This demonstrates that explicit mentions of color and requesting explanations play large roles in enhancing the model's graph comprehension.

Experiment 1 - Color and Label Inversion: Inverting color and A/B labels do not affect model performance, with overall model accuracies remaining high. However, there is a decline in data fit when A/B labels are inverted. We discuss this further in Section 5

Experiment 2 - Input Stimuli: The model generally performed better on "Default" stimuli than "All Color" stimuli and better on these than "No Color" stimuli. This suggests that model performance can be impacted through stylistic changes, even when the data and data mappings used are the same.

4 Experiment 3

Experiment 3. To disentangle the effect of contiguous segments on model performance, we created variations of Tasks 5 and 6 (henceforth 5B and 6B) that change whether the segments used for comparison are contiguous with one another (Figure 3b). In Task 5, the segments being compared are always contiguous, whereas in Task 5B, they are always separated by another segment. In Task 6, the segments are always separated by other segments, whereas in Task 6B, they are always contiguous.

Since Experiment 1 demonstrates that model accuracy is highest when the prompt framing includes color and explanation, we use this framing here for Experiment 3 as well. Similarly, based on Experiment 2 results, we apply the best-performing default variant in this experiment (Figure 3b).

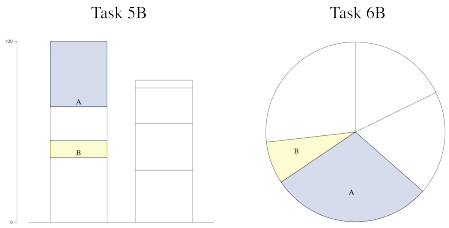


Figure 3: Task variations in Experiment 3.

Prompts		Stimuli	Invert		Tasks			
Clr.	Exp.		AB	Clr.	Task 5	Task 5B	Task 6	Task 6B
✓	✓	Def.	✗	✗	0.73	0.84	0.92	0.76
✓	✓	Def.	✓	✗	0.58	0.64	0.87	0.83
			✗	✓	0.73	0.71	0.58	0.82
			✓	✓	0.71	0.71	0.86	0.93

Table 2: Accuracy of GPT-4o-mini on probe 1 ("smaller than") judgments on Experiment 3 task variants. Task 5 and 6 accuracies were copied from Table 1 for comparison.

4.1 Results

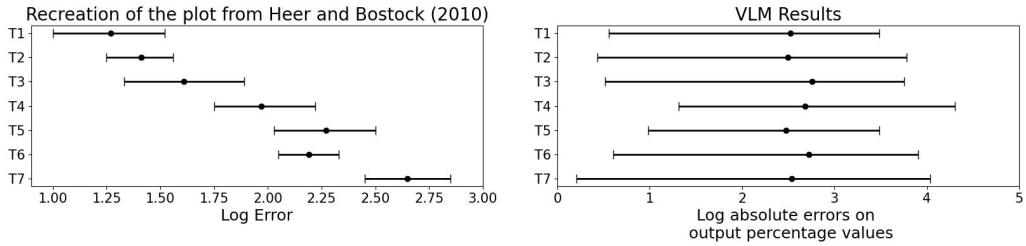
Experiment 3 - Segment Contiguity: There was an effect of segment contiguity on model performance. For the default condition, the model was less accurate when the segments being compared were contiguous than when they were well-separated (VLM performance Task 5B $>$ Task 5; Task 6B $<$ Task 6). However, inverting segment colors also inverts this relationship, causing contiguous segments to perform better than separate segments.

5 Discussion

Comparison to human performance: To evaluate the relationship between VLM performance and human performance, we conducted a rank-order correlation analysis (Table 1, Data fit). We ordered

the difficulty of the seven tasks for the VLM by their accuracy on probe 1 and for humans by their log error values from Heer and Bostock (6). (Note that these two approaches rank the results in descending and ascending order, respectively, so we take the negative value of calculated ρ.) There is greatest correspondence between VLM and humans ($\rho = 0.90$) on the relative difficulty of the seven tasks for the default prompt framing (**No Color, No Explanation**) and the default stimulus presentation. More broadly, there is a strong correlation across prompt variations for default stimuli. Conversely, there is a consistently low correlation for "All Color" stimuli.

Interestingly, Experiment 1 suggests that there is an effect of label order on model correlation to human performance even when average model accuracies remain high. Inverting the layout of A/B labels leads to a decline in data fit. However, we do not see similar effects when color is inverted.



(a) (b)

Figure 4: Accuracy of VLMs on proportion judgments (probe 2).

In addition to probe 1 accuracies reported above, we also evaluate the accuracy of the proportion judgments (probe 2) by replicating Heer and Bostock (6) and calculating the log absolute error ($\log_2(|\text{Judged}\% - \text{True}\%| + \frac{1}{8})$ and 95% confidence intervals for the seven tasks. Note that unlike for probe 1, the VLM does not align with human performance. Whereas humans showed systematic differences in the accuracy of proportion judgments across the tasks (Figure 4a), there was no statistical difference in the accuracy of the VLM (Figure 4b).

6 Conclusion

This paper reports the initial findings of evaluating GPT-4o-mini in a zero-shot manner on graphical perception tasks with established human performance profiles (3, 6). The study assesses the model's ability to extract and compare data from segments in a visualization. Our results show that VLMs perform similarly to humans when 1) both color and explanations are present in the prompt template, 2) segments are colored in the visualization, and 3) segments are non-contiguous. This suggests that, for certain combinations of task and visualization type, VLMs have the potential to design and evaluate visualizations by modeling human performance.

Looking ahead, the findings here may be useful for predicting and explaining VLM performance on more complex chart types, as seen in real-world applications. For instance, the effect of segment contiguity, documented here in the novel comparisons between Task 5 and 5B and Task 6 and 6B, may result in lower accuracies on ChartQA tasks (10) for stacked bar charts and pie charts overall. Future work can also evaluate human performance on the Task 5B and 6B variations introduced here to establish whether VLMs can generate new predictions about human performance on novel chart comprehension tasks.

7 Acknowledgements

The authors would like to thank the anonymous reviewers for their thoughtful and detailed feedback on the paper. This work is supported in part by NIH grant 1U01CA284207.

References

[1] Alexander Bendeck and John Stasko. An empirical evaluation of the gpt-4 multimodal language model on visualization literacy tasks. *IEEE Transactions on Visualization and Computer*

Graphics, 2024.

[2] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521, 2020.

[3] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[4] Zhenxing Cui, Lu Chen, Yunhai Wang, Daniel Haehn, Yong Wang, and Hanspeter Pfister. Generalization of cnns on relational reasoning with bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

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

[6] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 203–212, 2010.

[7] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

[8] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

[9] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. Opencqa: Open-ended question answering with charts. *arXiv preprint arXiv:2210.06628*, 2022.

[10] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[11] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023.

[12] Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartinstruct: Instruction tuning for chart comprehension and reasoning. *arXiv preprint arXiv:2403.09028*, 2024.

[13] Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *arXiv preprint arXiv:2407.04172*, 2024.

[14] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*, 2024.

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

[16] Jian Nie, Jun Yan, Huilin Yin, Lei Ren, and Qian Meng. A multimodality fusion deep neural network and safety test strategy for intelligent vehicles. *IEEE transactions on intelligent vehicles*, 6(2):310–322, 2020.

[17] OpenAI. Gpt-4 technical report, 2023.

[18] Hrituraj Singh and Sumit Shekhar. 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, 2020.

[19] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[20] Zhuoyue Wan, Yuanfeng Song, Shuaimin Li, Chen Jason Zhang, and Raymond Chi-Wing Wong. Datavist5: A pre-trained language model for jointly understanding text and data visualization. *arXiv preprint arXiv:2408.07401*, 2024.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

[23] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.

[24] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

8 Appendix / supplemental material

Input prompts to VLM

Here are the different kinds of prompts used in the study.

Prompt - No Color, No Explanation

```
Which of the two, (A) or (B), shapes is smaller?

 - When the inputs are bar charts or wedges compare the
    lengths/ height.
 - When the inputs are pie charts or circles compare the area.

Select one of the following:

A. The marked shape (A) is smaller.
B. The marked shape (B) is smaller.

What percentage is the SMALLER marked shape of the LARGER?
   Enter a percentage between 0 and 100.

Output in JSON format:
{
  "Is A smaller than B": true/false,
  "percentage": "XX%"
}
```

Prompt - Has Color, No Explanation

```
Which of the two, blue (A) or yellow (B), shapes is smaller?

 - When the inputs are bar charts or wedges compare the
    lengths/ height.
 - When the inputs are pie charts or circles compare the area.

Select one of the following:

A. The marked blue shape (A) is smaller.
B. The marked yellow shape (B) is smaller.

What percentage is the SMALLER marked shape of the LARGER?
   Enter a percentage between 0 and 100.

Output in JSON format:
{
  "Is A smaller than B": true/false,
  "percentage": "XX%"
}
```

Prompt - No Color, Has Explanation

```
Which of the two, (A) or (B), shapes is smaller?

 - When the inputs are bar charts or wedges compare the
    lengths/ height.
 - When the inputs are pie charts or circles compare the area.

Select one of the following:

A. The marked shape (A) is smaller.
B. The marked shape (B) is smaller.

What percentage is the SMALLER marked shape of the LARGER?
   Enter a percentage between 0 and 100.

Output in JSON format:
{
  "explanation for smaller or bigger": "...",
  "Is A smaller than B": true/false,
  "explanation for percentage": "...",
  "percentage": "XX%"
}
```

Prompt - Has Color, Has Explanation

```
Which of the two, blue (A) or yellow (B), shapes is smaller?

 - When the inputs are bar charts or wedges compare the
    lengths/ height.
 - When the inputs are pie charts or circles compare the area.

Select one of the following:

A. The marked blue shape (A) is smaller.
B. The marked yellow shape (B) is smaller.
```

```
What percentage is the SMALLER marked shape of the LARGER?
   Enter a percentage between 0 and 100.

Output in JSON format:
{
  "explanation for smaller or bigger": "...",
  "Is A smaller than B": true/false,
  "explanation for percentage": "...",
  "percentage": "XX%"
}
```

Limitations and Future Work

This study has a few limitations to be acknowledged. Our analysis was conducted using only one VLM, limiting the findings' generalizability, as different VLMs may exhibit varying performance characteristics. In particular, we expect that VLMs fine-tuned for chart comprehension or chart question-answering tasks will outperform more general models like GPT-4o-mini. Future work should thus consider testing multiple VLMs to create a more comprehensive evaluation. We also observed significant uncertainty in the model's performance on tasks involving percentage judgments, which points to lower model performance on these proportion-type judgments. Further testing would be useful to better understand and potentially mitigate this uncertainty.

Another limitation is that the input stimuli used for these experiments may not resemble the types of visualizations found in the training data of GPT-4o-mini. This mismatch could have contributed to suboptimal performance in certain tasks. Future studies can modify the input stimuli to match visualization styles in the training data better to more precisely evaluate model accuracy and reliability.