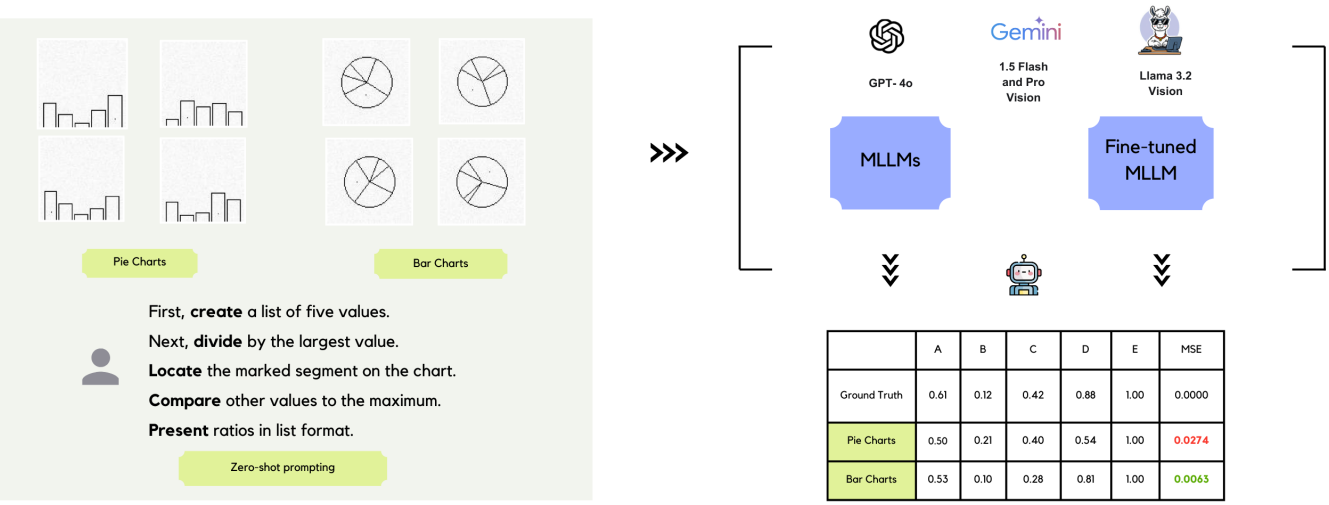
Evaluating 'Graphical Perception' with Multimodal LLMs

Rami Huu Nguyen*
University of Massachusetts Boston

Kenichi Maeda†
University of Massachusetts Boston

Mahsa Geshvadi‡
University of Massachusetts Boston

Daniel Haehn§
University of Massachusetts Boston



	A	B	C	D	E	MSE
Ground Truth	0.61	0.12	0.42	0.88	1.00	0.0000
Pie Charts	0.50	0.21	0.40	0.54	1.00	0.0274
Bar Charts	0.53	0.10	0.28	0.81	1.00	0.0063

Figure 1: **Computing Cleveland and McGill's Position-Angle Experiment using Multimodal Large Language Models**. We replicate the original experiment by asking MLLMs to interpret values in pie and bar charts using zero-shot prompting, where models follow instructions without prior examples. Results highlight that MLLMs predict values more accurately from bar charts (mean squared error (MSE) in green).

Abstract

Multimodal Large Language Models (MLLMs) have remarkably progressed in analyzing and understanding images. Despite these advancements, accurately regressing values in charts remains an underexplored area for MLLMs. For visualization, **how do MLLMs perform when applied to graphical perception tasks?** Our paper investigates this question by reproducing Cleveland and McGill's seminal 1984 experiment and comparing it against human task performance. Our study primarily evaluates fine-tuned and pretrained models and zero-shot prompting to determine if they closely match human graphical perception. Our findings highlight that MLLMs outperform human task performance in some cases but not in others. We highlight the results of all experiments to foster an understanding of where MLLMs succeed and fail when applied to data visualization.

Index Terms: Multimodal Large Language Models, Graphical Perception, Machine Perception, Deep Learning

1 Introduction

Nowadays, data visualization has become increasingly important in our lives [13, 20]. There has been a rising research focus on computational techniques for studying charts, and graphs. [7, 13, 14], which are applied in several applications, including data extraction, classification, visual Q&A (e.g., "computer, which section is greater?"), and design evaluation or synthesis. MLLMs have made significant progress in analyzing and understanding images [1, 6, 9, 18, 19, 23, 27]. Although MLLMs perform well in understanding charts, they struggle in generalization and face difficulties accurately answering chart-related questions [9, 20]. This requires the MLLMs to understand both language and information derived in charts and apply reasoning skills to provide correct answers [5, 16]. Most current MLLMs are pre-trained vision and knowledge, which means those models are trained before with general knowledge, and they might struggle with new application [8, 15, 17, 24], which potentially lead to incorrect visual understanding. Understanding images (computer vision) poses unique challenges as compared to understanding language [24] [26]. Language often relies on structured syntax and grammar, while chart data depends on spatial relationships, patterns, and context [22]. Hence, analyzing chart data might be more challenging for the MLLMs. What's more, the limitation of MLLMs also persists: MLLMs find it difficult to recognize small objects or tiny details in pictures [2, 6, 28]. Additionally, MLLMs currently have difficulty pinpointing the important details in the images that are unclear or absent in the images [25]. Also, humans use senses such as sight and language to understand the world and recognize new objects based on their knowledge.[17, 23]. Zero-shot prompting follows similar principles as human abilities, with its main purpose being to improve MLLMs using the zero-shot prompts to make them perform better without the need for additional training. Cleveland and McGill introduced the concept of graphical perception, explaining how humans visually interpret information from graphs [3, 4]. Cleveland and McGill defined elementary perceptual tasks as mental-visual processes and ranked how complex those tasks are

*e-mail: rami@mpsych.org
†e-mail: kenichi.maeda001@umb.edu
‡e-mail: mahsa.geshvadi001@umb.edu
§e-mail: daniel.haehn@umb.edu

for humans. Building on their work, our research is to compare fine-tuned MLLMs, trained for specific low-level graphical perception tasks, with pretrained MLLMs models using zero-shot prompting. Once those MLLMs models perform well on those tasks, they will establish a strong foundation for interpreting more complex visualizations.

1.1 Related Work

To study this graphical perception concept, Cleveland and McGill performed the position-angle experiment (comparing pie charts and bar charts) and the position-length experiment (where participants were asked to compare values in groups and divided charts [3, 4]. Then, the authors use this to redesign a statistical map via bars, framed rectangles, and Weber's law [10], using the proportional relation between an initial distribution density and perceivable change. Heer and Bostock later reproduced Cleveland and McGill's experiments by crowdsourcing participants on Amazon Mechanical Turk with similar findings [11]. Harrison et al. also repeated the studies while studying viewer emotions, again with similar results [10]. Talbot et al. explored how variation in bar charts affects human prediction [21]. Cleveland and McGill's idea of graphical perception does not rely on human-specific traits, and it targets the process of decoding information visually. Thus, it allows machines to perform similar tasks, such as MLLMs. Nonetheless, machines must match humans' graphical perception levels to function effectively in practice. Our research paper is inspired by Cleveland and McGill's, and Haehn et al. [7] also use CNNs for similar stimuli to investigate where machines perceive and reason visual relationships similar to human perception.

2 Experiment Setup

We compare MLLMs to human baselines across five experiments. **E1** estimates quantities from visual features based on Cleveland and McGill's elementary perceptual tasks. **E2** replicates their position-angle experiment, comparing pie and bar charts. **E3** reproduces their position-length experiment, analyzing grouped versus divided bar charts. **E4** evaluates bars and framed rectangles using their visual cue framework. **E5** conducts a Weber's law point cloud experiment.

2.1 Networks and Processes

As a starting point, we used three latest closed-source pretrained MLLMs, including **GPT-4o** with 1.8 trillion parameters, **Gemini 1.5 Flash** with 8 billion parameters, **Gemini 1.0 Vision Pro** (unavailable parameter data), and one open-source **Llama 3.2-Vision** with 11 billion parameters. For *fine-tuned MLLMs*, we used **Llama 3.2-Vision**, with 6.0 billion parameters, of which 94.4 million were trainable. We also produced 15 fine-tuned MLLMs (each experiment has three fine-tuned MLLMs) for this study. View our fine-tuning details in our supplementary material. Each experiment generated *5000 unique training images*, *1000 for unique validation*, and *55 for unique testing for each task*, add *added 5% noise* to each image, and ensure no data leakage. Detailed dataset generation and preprocessing are discussed in our supplementary material.

2.2 Measurements

Our research calculates models' performances using the midmean logistic absolute error metric (MLAE). Our research paper is inspired by Cleveland and McGill's methodology in their 1984 study [3] and defines our calculation as follow:

$$\text{MLAE} = \frac{1}{N}\sum_{i=1}^{N} \log_2\left(|\text{predicted}_i - \text{true}_i| + 0.125\right)$$

We also include standard error metrics such as mean squared error (MSE) and mean absolute error (MAE). Our fine-tuned models use Cross-Entropy loss instead of MLAE, applying the logarithm before averaging to ensure fair evaluation of small and large errors. d

2.3 Stimuli

We employed the stimuli generator designed by Haehn et al. [7] for each perceptual task, and the number of parameter values varied depending on each experiment.

2.4 Human Baselines

We gather human baseline measurements for the position-angle (E2) and position-length (E3) experiments from [3], with 51 participants. We also included human baseline measurements from Heer and Bostock's study [11], with 50 participants. In both experiments, each participant reviewed ten stimuli under each condition. We followed Haehn et al.'s paper [7] for E1, E4, and E5. The human baseline in Haehn et al.'s paper were gathered from 25 participants on Amazon Mechanical Turk for those three experiments. Each experiment has ten stimuli for participants (nine for E1, two for E4, and three for E5), including three practice stimuli for each condition.

2.5 Data Preprocessing

Initially, we collected **825 responses per task across all five experiments**, with each experiment has three runs. However, *invalid and missing responses* were found, primarily for pretrained models from E1 to E5 in each experiment. Therefore, we removed those invalid values from each initial dataset. To ensure fair comparisons between models, we excluded these invalid responses. However, since each model produced a different number of invalid responses, this resulted in an imbalance total number of valid responses per model. Since these invalid responses vary in each run, we balanced the dataset by randomly choosing the global minimum number of valid responses. Examples of our invalid responses and how we balanced the dataset are provided in our supplementary al material.

3 Experiment 1

Cleveland and McGill introduced ten elementary perceptual tasks that use graphical elements or visual marks to represent numerical values [3]. These tasks are the low-level building blocks for information visualizations. Examples are **estimating position on a common scale, position on non-aligned scales, length, direction (or slope), angle, area, volume, curvature, and shading (or ink density)**. To evaluate these tasks, we developed a visual representation as 100 x 100 raster images. Our main goal is to evaluate whether our MLLMs networks can regress the numerical value they encoded. We also added complexity to each task by generating multiple versions of elementary perceptual tasks. For instance, we created multiple angle degrees by changing the line's direction and the angle's size.

3.1 Hypothesis and Results

H1.1: Our fine-tuned MLLMs model can regress quantitative variables from graphical elements and outperform pretrained models and human graphical perception on these elementary perceptual tasks.

Figure 2 highlights our fine-tuned models have the highest MLAE errors compared to human perception in most tasks. In ***angle tasks***, our fine-tuned models achieve **MLAE = 5.01**, MAE = 51.65, SD = 1.66, while humans outperform them with **MLAE = 3.22**, SD = 0.54. In ***area tasks***, our fine-tuned models struggle with **MLAE = 10.09**, MAE = 1815.65, SD = 1.91, while humans perform better (**MLAE = 3.64**, SD = 0.38). In ***volume tasks***, our fine-tuned models show **MLAE = 8.63**, MAE = 2034.04, SD = 3.38, while humans perform better (**MLAE = 5.18**, SD = 0.40). In particular, the fine-tuned models only performed the best with ***curvature*** task with an **MLAE of -1.23, MAE = 1.19, SD = 1.98**; however,

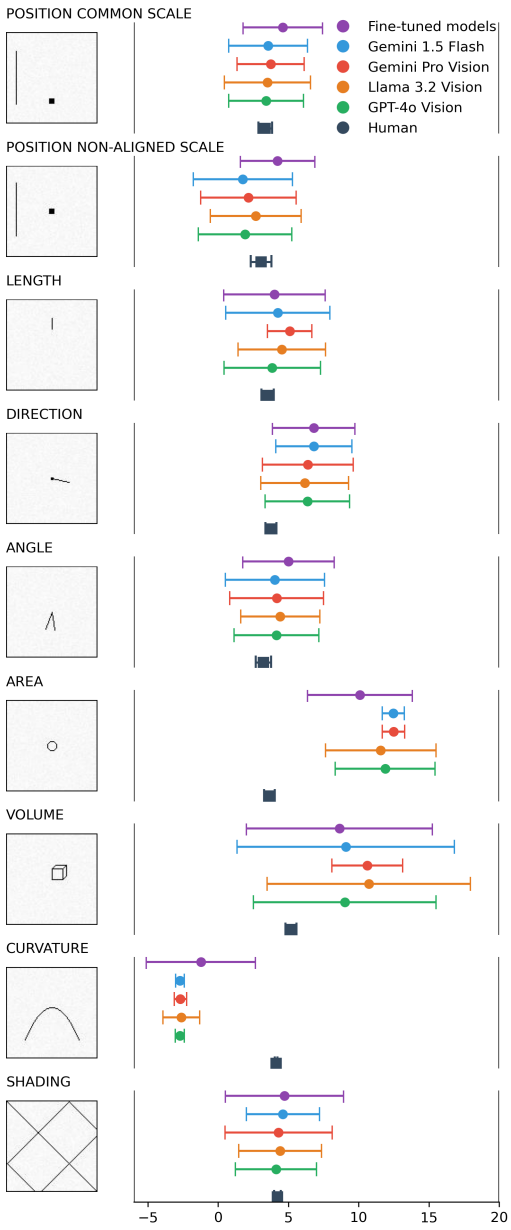


Figure 2: **Elementary perceptual tasks results for the most complex task parameterization**. In each column: Left: Example stimuli image. Right: MLAE and bootstrapped 95% confidence intervals for different networks. Lower MLAE scores are better.

it also underperformed all pretrained MLLMs with average MLAE ranging from -2.73 to -2.62. Our fine-tuned models only performed slightly better than two or three pretrained models, especially in area and volume. For instance, in the ***volume* tasks**, the difference is minimal as our fine-tuned MLLMs record an **MLAE of 8.63**, slightly higher than the pretrained Gemini 1.5 Flash, Gemini Pro Vision, Llama 3.2-Vision and GPT-4o which have **MLAE scores of 9.08, 10.62, , 10.72 and 9.01**, respectively. Across tasks, we compare the average regression performance of our networks and report statistically significant effects (**F = 10.303,** $p < 0.01$). Tukey's HSD post-hoc test highlights that **Gemini 1.5 Flash and GPT-4o significantly outperform our fine-tuned models**, while the difference between our fine-tuned models and Gemini 1.5 Flash and Llama 3.2-Vision is not statistically significant ($p > 0.01$). Therefore, **we do not accept H1.1**.

4 Experiment 2

Cleveland and McGill compare bar and pie charts by studying how humans perceive position ratios and angles [3]. Following Cleveland and McGill's proposed encoding, we generate rasterized images to evaluate how our networks perceive these two tasks. Each visualization consists of pie or bar charts representing numbers that sum to 100, with individual numbers ranging from 3 to 39. We modified our approach to minimize value differences. Cleveland and McGill created stimuli with a minimum scale difference of 0.1, but our models only process 100 x 100-pixel images as input; we can only minimally represent a difference of 1 pixel. In Cleveland and McGill's experiments [3], participants were asked to estimate the ratio of the four smaller segments to the known and marked most significant segments. Similarly, we marked the largest segment in each visualization with a single pixel dot. We tasked our networks with performing multiple regressions to estimate the ratio of the remaining four segments.

4.1 Hypothesis and Results

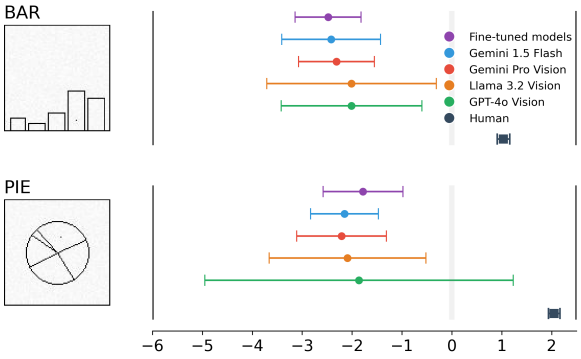


Figure 3: **Computational results of the position-angle experiment**. Left: Example stimuli. Right: MLAE and bootstrapped 95% confidence intervals (the lower, the better)

H2.1: Our fine-tuned models outperform human graphical perception for bar and pie charts tasks.

For the **bar chart** tasks, our *fine-tuned models* achieved **an MLAE of -2.48** (SD = 0.34, MAE: 0.06). For our pretrained models, *Gemini 1.5 Flash* achieved **an MLAE of -2.42** (SD = 0.51) and an MAE of 0.12, *Gemini 1.0 Pro Vision* showed **an MLAE of -2.31** (SD = 0.39) and MAE of 0.08, *Llama 3.2-Vision* had an **MLAE of -2.01**, (SD = 0.87) and MAE of 0.28, and *GPT-4o* had an MLAE of -2.01, (SD = 0.72) and MAE of 0.28. Comparing this to **the human baseline of MLAE: 1.035** (SD = 0.115) for bar charts, both pre-trained and fine-tuned models outperform human graphical perception, as a lower MLAE suggests better performance.

For the **pie chart** tasks, our *fine-tuned models* achieved **an average MLAE of -1.78** (SD = 0.41) and an MAE of 0.18. For our pretrained models, *Gemini 1.5 Flash* achieved **an MLAE of -2.15** (SD = 0.34) and MAE of 0.11, *Gemini 1.0 Pro Vision* showed **an MLAE of -2.21** (SD = 0.46) and MAE of 0.10, Llama 3.2-Vision had an **MLAE of -2.09**, (SD = 0.80 and MAE of 0.28, and *GPT-4o* had an MLAE of -1.86, (SD = 1.58) and MAE of 1.22. Comparing this to **the human baseline of MLAE: 2.05** (SD = 0.125) for pie charts, both pretrained and fine-tuned models outperform human graphical perception, as a lower MLAE suggests better performance. Based on the results from two tasks, **we accept H2.1**.

H2.2: Pie charts are more challenging for pre-trained and fine-tuned models than bar charts.

Figure 3 reveals that **our fine-tuned models outperformed pre-trained models for bar tasks**. Pie charts also have higher the average MLAE values than bar charts. CHANGED Across tasks, we compare the average regression performance of our networks and report statistically significant effects (F = 25.614, $p < 0.01$). Tukey's HSD post-hoc test highlight that our fine-tuned models performs significantly worse than Gemini 1.5 Flash and Gemini 1.0 Pro Vision ($p < 0.01$), but our fine-tuned models only outperforms GPT-4o ($p < 0.01$) with a mean difference of 0.15. Additionally, the difference between our fine-tuned models and Llama 3.2-Vision is not statistically significant ($p > 0.01$), indicating similar performance. Since the results do not fully support the hypothesis for across models, we **partially accept H2.2**.

5 Experiment 3

Cleveland and McGill evaluated the perception of position and length across five variations of the group and divided bar charts [3]. While both charts highlight identical information, the perceptual tasks are interpreted differently. A group of bar charts always requires the judgments of positions along a common scale, while a group of divided bar charts also involves length judgments. Types 1, 2, and 3 focus on judging positions along a common scale, whereas types 4 and 5 require length judgment. In their experiment, participants were asked to estimate the percentage of the smaller marked bar element of the larger one. Cleveland and McGill ranked the tasks from easiest (Type 1) to hardest (Type 5). We followed their method to generate data and created ten value pairs using the equation:

$$s_i = 10 \times 10^{\frac{i-1}{12}}, \quad i = 1, \ldots, 10$$

The experiment involves a dataset comprising a combination of unique labels across its subsets. For each chart type, we created visualization charts corresponding to these ground truth values. We ask our MLLMs to follow zero-shot prompts to estimate the percentage of the smaller value relative to the larger one, treating this as a single-value regression problem.

5.1 Hypothesis and Results

H3.1: Pre-trained and fine-tuned MLLMs work well and outperform human perception for all five types of tasks.

Our *fine-tuned models* achieved ranges from (**MLAE: -1.63**, SD = 0.74, MAE: 0.26) to (**MLAE: -1.48**, SD = 0.82, MAE: 0.29). Our *pretrained models* recorded the lowest error with ranges of (**MLAE: -1.85**, SD = 0.53, MAE: 0.17) to (**MLAE: -1.61**, SD = 0.82, MAE: 0.29). In comparison to human baseline data, both fine-tuned and pretrained models produced the lowest errors, with **Human Baseline 1** [3]: (MLAE: 1.4 to 2.72, SD from 0.14 to 0.175) and **Human Baseline 2** [11]: (MLAE: 1.25 to 2.24, SD from 0.175 to 0.25). Also, Figure 4 displays that the MLAE of all MLLMs performs below zero, suggesting that **its performance estimates are mostly precise for five tasks**. From the MLAE results above, we recognized that zero-shot prompting was effective for all pretrained models. The zero-shot prompts worked particularly well in these charts for all pretrained models. They specify the goal (e.g., compare bar heights), give detailed instructions on how to output answers (e.g., a scale from 0 to 1), and eliminate unnecessary complexity by using two words (no explanation). These prompts guide the models to concentrate solely on the visual comparison of bar and pie charts to produce single numeric outputs.

It also indicates that the general knowledge of pre-trained models enables them to recognize these tasks' structure. The lower average value of MLAE errors for this experiment might prove the pretrained models understand the concept of bar or pie charts and identify the relationship between different bars or pies based on visual presentation. Also, the pretrained model might use its pretrained mathematical and logical reasoning to compute height and compare ratios. From this result, we **accept H3.1.**

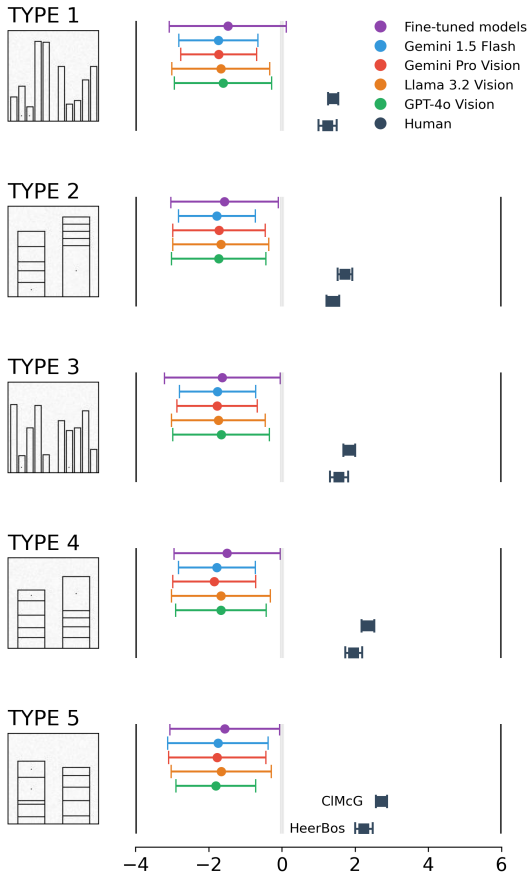


Figure 4: **Computational results of the position-length experiment**. Left: Type 1–5 stimuli for divided and grouped bar charts (as per Cleveland and McGill). Right: MLAE and bootstrapped 95% confidence intervals of our networks.

H3.2: Our fine-tuned models surpass pretrained models for all tasks.

It is apparent from Figure 4 that all pretrained models consistently recorded lower MLAE scores than our fine-tuned models. Across tasks, we evaluate the average regression performance of our networks and report statistically significant differences (**F = 36.66,** $p < 0.01$. Tukey's HSD post-hoc test reveals that our fine-tuned models perform significantly worse than Gemini 1.5 Flash, Gemini Pro Vision, Llama 3.2-Vision, and GPT-4o ($p < 0.01$). The mean differences range from **-0.13 to -0.22**, confirming that our fine-tuned models has the highest MLAE errors compared to all pretrained models. Since the differences are statistically significant across models, we conclude that **our fine-tuned models underperform consistently and thus we do not accept H3.2.**

6 Experiment 4

Visual cues are essential in graphical elements as they integrate into real-world variables. Cleveland and McGill designed an experiment using bars and framed rectangles to study how humans perceive the length and position of non-aligned scales [3]. Figure 5 highlights both variations on the left. It is difficult to estimate which bar is larger (bottom). Once the frame is added to the maximum length, the task transforms bar length into position judgment along aligned scales, making it easier to interpret. Cleveland and McGill theo-

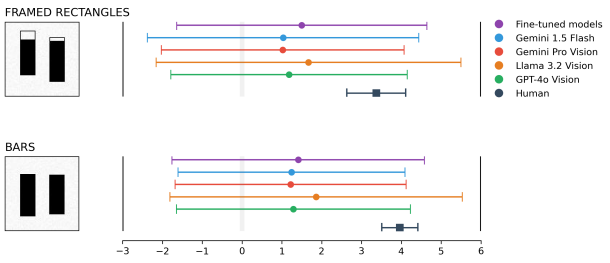


Figure 5: **Computational results of the bars-and-framed-rectangles experiment**. Left: Stimuli of two bars for length judgment (bottom) following Cleveland and McGill's setting. Perceiving which bar is longer is significantly easier for humans when a frame is added (top).

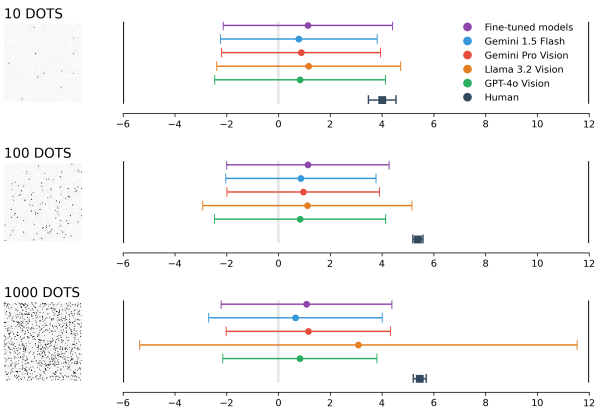


Figure 6: **Computational results of the point cloud experiment.** Left: We create 2D point clouds with 10, 100, and 1000 initial dots. Then, we add up to 10 new dots. For humans, it is possible to estimate how many dots are added if there are initially 10 points, but it is impossible to see how many dots are added when starting with 1000 dots.

rized that judging the white space in the frame could resemble a length judgment rather than a position judgment. Given this, they relate the tasks to Weber's Law: the perceivable difference within a distribution is proportional to its initial size [12]. In this experiment, Weber's Law implies that humans can find it easier to measure the differences in the whole space (framed scale) as its initial size is small. In contrast, estimating small changes in the length of black bars is harder. The Just Noticeable Difference (JND) is higher when the initial stimulus is smaller. We generated visualization charts aligned with these unique ground truth values for each framed and unframed task.

6.1 Hypothesis and Results

H4.1: Fine-tuned and pretrained models surpass human perception for framed and unframed tasks.

For *framed tasks*, our fine-tuned model recorded average ranges of (**MLAE: 1.50**, SD = 1.60, MAE: 4.02). Our pretrained models have errors with average ranges of (**MLAE: 1.03**, SD = 1.74, MAE: 3.04) for Gemini 1.5 Flash, (**MLAE: 1.02, SD** = 1.56, MAE: 2.80) for Gemini 1.0 Pro Vision, (**MLAE: 1.18**, SD = 1.51, MAE: 3.09) for GPT-4o, and (**MLAE: 1.67, SD** = 1.95, MAE: 6.98) for Llama 3.2-Vision. For *unframed tasks*, our fine-tuned models recorded average ranges of (**MLAE: 1.41**, SD = 1.62, MAE: 3.80). Our pretrained models have errors with average ranges of (**MLAE: 1.24**, SD = 1.46, MAE: 3.07) for Gemini 1.5 Flash, (**MLAE: 1.22**, SD = 1.48, MAE: 3.08) for Gemini 1.0 Pro Vision, (**MLAE: 1.29**, SD = 1.50, MAE: 3.30) for GPT-4o, and (**MLAE: 1.86**, SD = 1.87, MAE: 7.97) for Llama 3.2-Vision.

Our fine-tuned and pretrained models produced lower errors and outperformed human perceptions with **Human Baseline** (**MLAE: 3.371**, SD = 0.741) for framed tasks and **Human Baseline (MLAE: 3.961**, SD = 0.454) for unframed and framed tasks. In addition to this comparison, zero-shot prompting works effectively for framed and unframed tasks for pretrained models, especially Gemini 1.0 Pro Vision, as they have the lowest average MLAE errors, indicating those models generalize well without requiring any prior training.

H4.2: Our fine-tuned models work better than our pretrained models for two tasks.

Figure 5 highlights all pretrained models that produce lower average MLAE errors than our fine-tuned models for two tasks. However, an exception is shown here with the Llama 3.2-Vision model, **one of the pretrained models, which has underperformed our fine-tuned models**. Across tasks, we evaluate the average regression performance of our networks and report statistically significant differences (**F = 32.50,** $p < 0.01$). Tukey's HSD post-hoc test reveals that **our fine-tuned models performs significantly worse than Gemini 1.5 Flash, Gemini 1.0 Pro Vision, and GPT-4o** ($p <= 0.01$), with mean differences ranging from **-0.43 to -0.51**. However, its performance does not significantly differ from Llama 3.2-Vision ($p > 0.01$), indicating similar error levels between the two models. Since the differences are statistically significant across most models, we found that our fine-tuned models consistently underperforms except when compared to Llama 3.2-Vision. Therefore, we **partially accept H4.2**.

7 Experiment 5

7.1 Hypothesis and Results

We generated a 2D point cloud simulation inspired by Weber's law, where networks predict the number of dots (up to 10) added to an initial set of 10, 100, or 1000 dots. While humans can roughly estimate the number of added dots for 10 initial dots, it becomes harder for them to give answers and prone to random guessing with 100 or 1,000 initial dots.

H5.1: Our fine-tuned models perform relatively better than human perception and all pretrained models.

Our fine-tuned models performed across three tasks (10 dots, 100 dots, 1000 dots) with the following metrics:**10 dots: MLAE = 1.14**, SD = 1.67, MAE = 3.23; **100 dots: MLAE = 1.14**, SD = 1.60, MAE = 3.17; **1000 dots: MLAE = 1.09**, SD = 1.69, MAE = 3.12. These results significantly surpass the **Human Baseline:** (**MLAE = 4.0149**, SD = 0.5338), (**MLAE = 5.3891**, SD = 0.1945), (**MLAE = 5.4612**, SD = 0.2509) for the respective tasks. In *10 dots*, our pretrained models exceeded the range of an **MLAE from 0.79 to 0.88**, compared to our fine-tuned models at an **MLAE of 1.14**. Our fine-tuned models only outperformed Llama 3.2-Vision at an MLAE of 1.17. In *100 dots*, our fine-tuned models with an **MLAE = 1.14**, SD = 1.60, MAE = 3.17 underperformed all pretrained models, with the range of average MLAE from 0.84 to 1.12. In *1000 dots*, our fine-tuned models' performance with an **MLAE of 1.09**, SD = 1.69, MAE = 3.12 outperformed Llama 3.2-Vision, which recorded **MLAE = 3.09**, SD = 4.31, MAE = 503.27.

Across tasks, we evaluate the average regression performance of our networks and report statistically significant differences (**F = 37.44,** $p < 0.01$). Tukey's HSD post-hoc test highlights that our fine-tuned models perform significantly worse than Gemini 1.5 Flash and GPT-4o ($p < 0.01$), with mean differences of **-0.35 and -0.29**, respectively. However, its performance does not significantly differ from Gemini 1.0 Pro Vision ($p > 0.01$), indicating similar error levels between the two models. Additionally, our fine-tuned

models outperform Llama 3.2-Vision ($p < 0.01$) with a mean difference of **0.67**. Since the differences are statistically significant across most models, we found that our fine-tuned models generally underperform, except when compared to Gemini 1.0 Pro Vision and Llama 3.2-Vision. Therefore, we **partially accept H5.1**.

8 Conclusion

Our paper reports the findings of evaluating the performance of MLLMs on graphical perception tasks in zero-shot prompt settings. This study measures pretrained and fine-tuned models' abilities across five experiments, focusing on elementary perceptual tasks, position length, position angle, position non-aligned scale, and point cloud. Overall, our pretrained models outperform fine-tuned models in most cases; all MLLMs can evaluate visualizations more precisely using zero-shot prompting on curvature tasks and all tasks from E2 to E5.

Although our fine-tuned models mostly underperformed compared to pretrained models, the differences are minimal, and they perform well in volume and area tasks. Also, the post-hoc from our E2, highlights fine-tuned models that outperforms GPT-4o across bar and pie charts. This would be a strong opportunity for future research to fine-tune these models and enhance their performance. We trained our fine-tuned models with 94.4M parameters over five epochs using Parameter-Efficient Fine-Tuning (PEFT) and a LoRA configuration. Despite using fewer parameters, they performed closely to pretrained models, with only a small performance gap. These findings lay the groundwork for further enhancing fine-tuned models by scaling parameters and leveraging more diverse datasets. Since our MLLMs performed well in specific visual tasks, we wondered, "**Why do the MLLMs have wider MLAE error bars?**". This interesting point is explored in our supplement material.

Our findings also could be used for automating those zero-shot prompts to provide quicker insights for various real-world applications: business intelligence, scientific research, autonomous driving, robotics, etc. To achieve this purpose, our future studies will include more complex visualizations, diverse experiments, a wider range of MLLMs, color saturation and edge detection, and chain-of-thought reasoning to enhance MLLMs' ability to regress multiple values in visualizations.

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468*, 2016. 1

[2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

[3] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. 1, 2, 3, 4

[4] W. S. Cleveland and R. McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833, 1985. 1, 2

[5] Y. Du, H. Guo, K. Zhou, W. X. Zhao, J. Wang, C. Wang, M. Cai, R. Song, and J.-R. Wen. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv preprint arXiv:2311.01487*, 2023. 1

[6] G. Guo, J. J. Kang, R. S. Shah, H. Pfister, and S. Varma. Understanding graphical perception in data visualization through zero-shot prompting of vision-language models. *arXiv preprint arXiv:2411.00257*, 2024. 1

[7] D. Haehn, J. Tompkin, and H. Pfister. Evaluating 'graphical perception'with cnns. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):641–650, 2018. 1, 2

[8] S. C. Han, F. Cao, J. Poon, and R. Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond. *arXiv preprint arXiv:2410.05608*, 2024. 1

[9] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 1

[10] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014. 2

[11] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 203–212, 2010. 2, 4

[12] A. S. Householder and G. Young. Weber laws, the weber law, and psychophysical analysis. *Psychometrika*, 5(3):183–193, 1940. 5

[13] K. Kafle, B. Price, S. Cohen, and C. Kanan. Dvqa: Understanding data visualizations via question answering. *arXiv preprint arXiv:1801.08163*, 2018. 1

[14] S. E. Kahou, V. Michalski, A. Atkinson, A. Kadar, A. Trischler, and Y. Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2018. 1

[15] J. Lee, Y. Wang, J. Li, and M. Zhang. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*, 2024. 1

[16] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoob, and D. Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1287–1310, 2024. doi: 10.18653/v1/2024.naacl-long.70 1

[17] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1

[18] R. Luo, H. Zhang, L. Chen, T.-E. Lin, X. Liu, Y. Wu, M. Yang, M. Wang, P. Zeng, L. Gao, H. T. Shen, Y. Li, X. Xia, F. Huang, J. Song, and Y. Li. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*, 2024. 1

[19] T. Lv, Y. Huang, J. Chen, Y. Zhao, Y. Jia, L. Cui, S. Ma, Y. Chang, S. Huang, W. Wang, L. Dong, W. Luo, S. Wu, G. Wang, C. Zhang, and F. Wei. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2024. 1

[20] 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. 1

[21] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 20:2152–2160, 2014. 2

[22] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, Y. Li, and N. Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *arXiv preprint arXiv:2406.14852*, 2024. 1

[23] T. Wang, Y. Liu, J. C. Liang, J. Zhao, Y. Cui, Y. Mao, S. Nie, J. Liu, F. Feng, Z. Xu, C. Han, L. Huang, Q. Wang, and D. Liu. M^2pt: Multimodal prompt tuning for zero-shot instruction learning. *arXiv preprint arXiv:2409.15657*, 2024. 1

[24] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, and J. Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 1

[25] P. Wu and S. Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 1

[26] L. Xie, L. Wei, X. Zhang, K. Bi, X. Gu, J. Chang, and Q. Tian. Towards agi in computer vision: Lessons learned from gpt and large language models. *arXiv preprint arXiv:2306.08641*, 2023. 1

[27] X. Zeng, H. Lin, Y. Ye, and W. Zeng. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *arXiv preprint arXiv:2407.20174*, 2024. 1

[28] J. Zhang, J. Hu, M. Khayatkhoei, F. Ilievski, and M. Sun. Exploring perceptual limitation of multimodal large language models. *arXiv preprint arXiv:2402.07384*, 2024. 1