Protecting multimodal large language models against misleading visualizations

Jonathan Tonglet[1,2,3], **Tinne Tuytelaars**[2], **Marie-Francine Moens**[3], **Iryna Gurevych**[1]
[1] Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science, TU Darmstadt and National Research Center for Applied Cybersecurity ATHENE
[2] Department of Electrical Engineering, KU Leuven
[3] Department of Computer Science, KU Leuven
www.ukp.tu-darmstadt.de

Abstract

Visualizations play a pivotal role in daily communication in an increasingly data-driven world. Research on multimodal large language models (MLLMs) for automated chart understanding has accelerated massively, with steady improvements on standard benchmarks. However, for MLLMs to be reliable, they must be robust to misleading visualizations, i.e., charts that distort the underlying data, leading readers to draw inaccurate conclusions. Here, we uncover an important vulnerability: MLLM question-answering (QA) accuracy on misleading visualizations drops on average to the level of the random baseline. To address this, we provide the first comparison of six inference-time methods to improve QA performance on misleading visualizations, without compromising accuracy on non-misleading ones. We find that two methods, table-based QA and redrawing the visualization, are effective, with improvements of up to 19.6 percentage points. We make our code and data available.[1]

1 Introduction

In an increasingly data-driven world, visualizations are widely used by scientists, journalists, governments, and companies to efficiently communicate data insights to a broad audience (Huang et al., 2025). They play a crucial role in crisis settings, such as during the COVID-19 pandemic, where charts were shared daily to inform the public (Zhang et al., 2021). In many cases, visualizations support a message more convincingly than showing the underlying data table directly to readers (Pandey et al., 2014). However, visualizations can also be misleading (Correll and Heer, 2017). This occurs when design flaws, or misleaders, distort the underlying data in a way that prevents a correct interpretation by the reader (Tufte and Graves-Morris, 1983; Pandey et al., 2015; Correll and Heer, 2017; Lauer and O'Brien, 2020; McNutt et al., 2020; Yang et al., 2021; Lo et al., 2022; Lisnic et al., 2023; Lan and Liu, 2025). These misleaders appear across a wide range of visualizations, including bar and line charts, or choropleth maps. Recent taxonomies have documented over 70 types of misleaders observed in real-world examples (Lo et al., 2022; Lan and Liu, 2025). Major categories include axes manipulations, such as truncated and inverted axes, or the use of visual manipulations like 3D effects. Figure 1 shows three real-world examples of misleading visualizations (Lo et al., 2022) paired with multiple choice questions (MCQs). Given a question about the underlying data table, their misleaders lead readers to infer a misleading answer instead of the correct one. Figure 2 illustrates how two visualizations of the same data, one non-misleading and the other misleading, can lead to different interpretations by readers.

Prior work has revealed the potential of misleading visualizations to deceive human readers. Some studies (Pandey et al., 2015; Ge et al., 2023; Rho et al., 2023; Bharti et al., 2024) have shown that such visualizations reduce human accuracy when answering MCQs about the underlying data. Other studies (Pandey et al., 2015; O'Brien and Lauer, 2018; Lauer and O'Brien, 2020) have demonstrated that human readers tend to provide different Likert-scale answers to a question depending on whether they view a misleading or non-misleading visualization of the same data table.

Misleading visualizations pose a serious threat to society. Due to their deceptive potential, they can be exploited by malicious actors to promote online disinformation (Correll and Heer, 2017). For instance, the charts in Figure 1 mislead readers on sensitive topics such as COVID-19, abortion, and gun violence. Misleading visualizations can shift public opinion, even on polarized debates like Brexit (Tartaglione and de Wit, 2025).

Multimodal large language models (MLLMs)

[1] github.com/UKPLab/acl2026-misleading-visualizations

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8329–8349

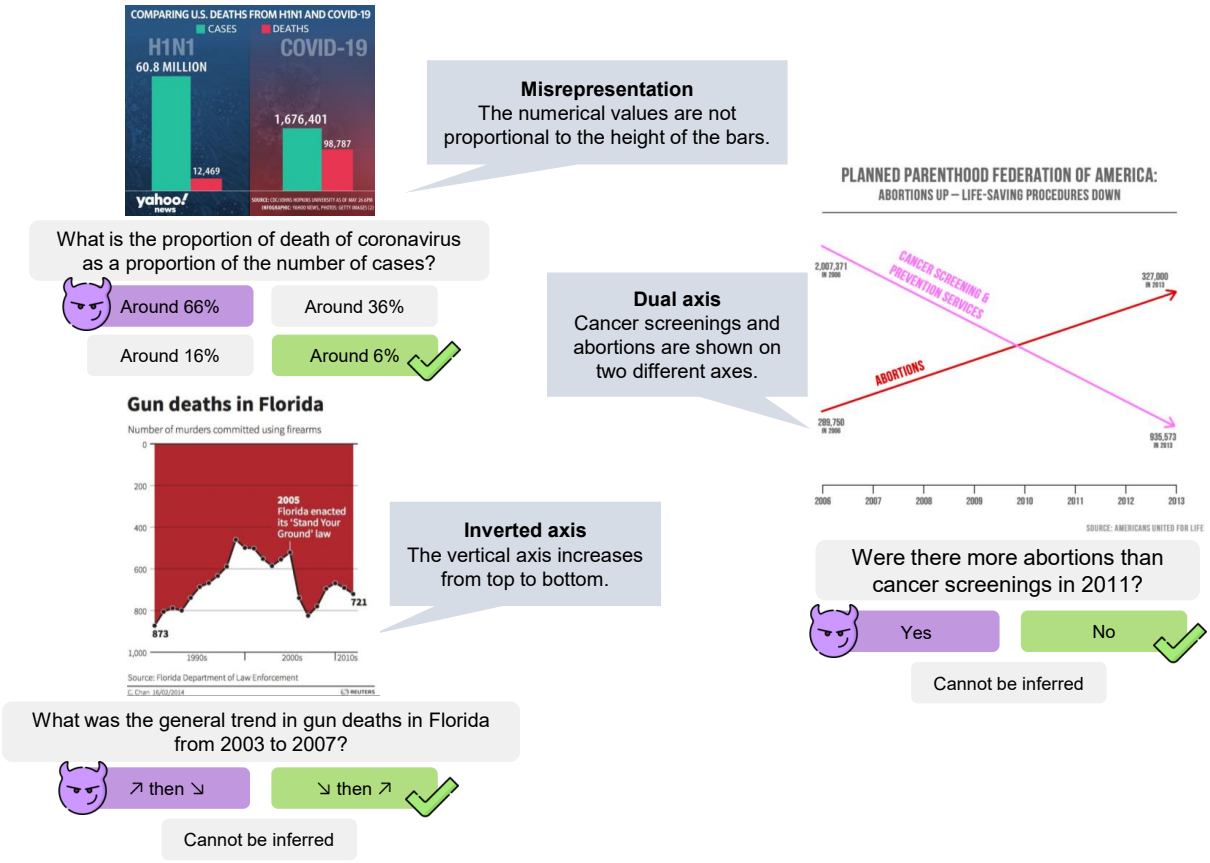

Figure 1: Three examples of real-world misleading visualizations (Lo et al., 2022) with MCQs. The correct answer is colored in green, while the wrong answer supported by the misleader is colored in purple.

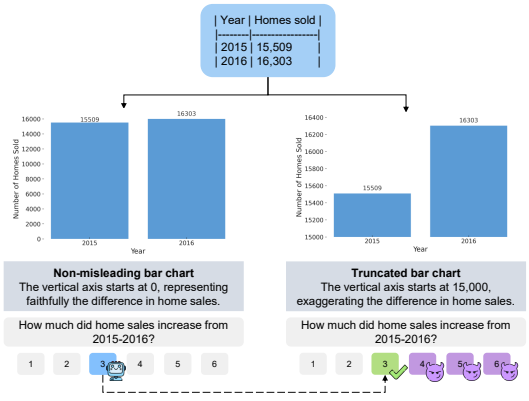

Figure 2: Non-misleading and misleading visualizations of the same data (Lauer and O'Brien, 2020) with a Likert-scale question where 1 means "a little" and 6 means "a lot". A consistent interpretation requires identical responses. However, the deceived reader chooses a higher value for the truncated bar chart.

have sparked a massive interest in automated chart understanding research (Huang et al., 2025), with steady progress on reference benchmarks such as ChartQA (Masry et al., 2022; Hoque et al., 2022). However, existing research has overlooked the important threat posed by misleading visualizations. If MLLMs, like humans, are easily deceived by misleading visualizations, malicious actors could exploit this vulnerability. MLLMs confronted with misleading visualizations could present incorrect interpretations of the underlying data to users, contributing to the spread of disinformation and reinforcing human belief in it. These risks underscore the urgency of thoroughly assessing the vulnerability of MLLMs to misleading visualizations and developing effective mitigation strategies.

In this work, we conduct the first comprehensive study to assess and mitigate the vulnerability of 19 MLLMs of varying sizes to misleading visualizations. In Experiment 1, we compare the question-answering (QA) accuracy of MLLMs on misleading and non-misleading visualizations. In Experiment 2, we evaluate the consistency of MLLMs, i.e., whether they provide the same response to a Likert-scale question when shown either a misleading or a non-misleading visualization of the same underlying data. The results of both experiments demonstrate that MLLMs, like humans, are indeed vulnerable to misleading visualizations. In Experi-

ment 3, we compare six inference-time correction methods to reduce MLLMs' vulnerability to the misleading visualizations from Experiment 1.

In summary, our contributions are twofold: (1) We present the first extensive analysis of the vulnerabilities of 19 MLLMs to 17 types of misleading visualizations, evaluating both QA accuracy and Likert-scale consistency. (2) We propose the first analysis of six correction methods to mitigate the negative impact of misleaders on QA tasks.

2 Methodology

2.1 Datasets

Experiments 1 and 3 We compare the question-answering (QA) accuracy of MLLMs across two datasets: (a) a misleading visualization dataset containing n = 143 instances, and featuring 17 types of misleaders, defined in Appendix A (Ge et al., 2023; Rho et al., 2023; Bharti et al., 2024; Lo et al., 2022); (b) a non-misleading visualization dataset (n = 124) (Ge et al., 2023; Rho et al., 2023; Bharti et al., 2024; Lee et al., 2017), also combining real-world and synthetic cases; and (c) ChartQA (Masry et al., 2022), the reference real-world non-misleading benchmark (n = 2500). Datasets (a) and (b) combine three existing resources and one introduced in this work. First, CALVI (Ge et al., 2023) includes 45 misleading and 15 non-misleading visualizations based on synthetic data, each paired with an MCQ. CALVI was initially designed and is the reference resource for evaluating humans. Second, CHARTOM (Rho et al., 2023; Bharti et al., 2024) contains 56 samples, including 28 MCQs, 20 free-text questions, and 8 rank questions. Each question is linked to two visualizations, one misleading and one non-misleading. Like CALVI, the underlying data is synthetic, and the test was originally designed to evaluate humans. Third, VLAT (Lee et al., 2017), the reference dataset to evaluate humans on non-misleading cases, provides 12 visualizations, each paired with three to seven MCQs, for a total of 53 instances. The visualizations are based on real-world data. Fourth, we introduce a dataset of 42 real-world misleading visualizations, each annotated with an MCQ with three to four choices. They come from a collection annotated with the misleaders affecting them (Lo et al., 2022), which inspired the synthetic examples in CALVI. We manually create MCQs, using CALVI and CHARTOM as references. By incorporating visualizations and questions about real-world data, we introduce direct conflicts with MLLMs' parametric knowledge, allowing us to assess their vulnerability in real-world scenarios.

Dataset	Experiments	n	m	Question types
Misleading visualizations	1 & 3	143	17	MCQ, Free-text, rank
↪ CALVI (misleading)	1 & 3	45	14	MCQ
↪ CHARTOM (misleading)	1 & 3	56	7	MCQ, Free-text, rank
↪ Real-world	1 & 3	42	12	MCQ
Non-misleading visualizations	1 & 3	124	-	MCQ, Free-text, rank
↪ CALVI (non-misleading)	1 & 3	15	-	MCQ
↪ CHARTOM (non-misleading)	1 & 3	56	-	MCQ, Free-text, rank
↪ VLAT	1 & 3	53	-	MCQ
ChartQA	1	2500	-	Free-text
Lauer & O'Brien	2	8	3	Likert-scale

Table 1: Datasets statistics. n is the number of instances. m is the number of misleader categories.

Experiment 2 We use four pairs of visualizations, one misleading and the other non-misleading, designed by Lauer and O'Brien (2020).

Table 1 provides detailed dataset statistics.

2.2 Correction methods

In Experiment 3, we compare six inference-time correction methods, illustrated in Figure 3, to increase QA accuracy on misleading visualization without compromising it on non-misleading ones.

(1) Misleader warning: we insert in the prompt a message to make the MLLM aware of the misleader in the visualization. The message is identical for all instances with the same misleader.

Providing (2) the axes data, (3) the data table, (4) or both: we prompt the MLLM in zero-shot to extract the axes labels or underlying data table. The axes and tables are formatted as text strings. We do not impose formatting constraints on the axes and tables. The axes, the table, or both are provided as additional prompt input.

(5) Table-based QA: after extracting the table with the MLLM, we provide it alone to a text-only LLM, reframing the task as table-based QA (Liu et al., 2023a,b; Kim et al., 2025).

(6) Redrawn visualization: after extracting the table with the MLLM, we provide it to a text-only LLM, which generates Python code to create a visualization using Matplotlib (Hunter, 2007). If the code compiles successfully, the generated visualization replaces the original one in the QA prompt; otherwise, the original one is used.

2.3 Implementation details

Metrics The evaluation of QA accuracy depends on the question type. For MCQs and rank questions, we evaluate the exact match. For free-text

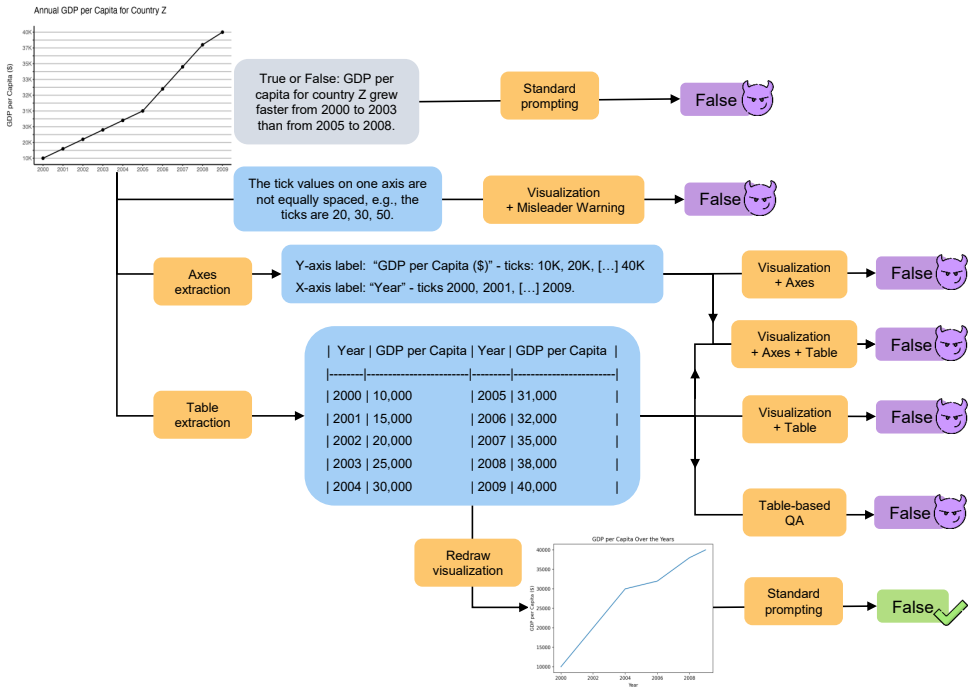

Figure 3: Illustration of the six inference-time correction methods applied to a misleading visualization from CALVI (Ge et al., 2023). The visualization suffers from inconsistent tick intervals on the y-axis.

questions, which all expect numerical answers, we use a relaxed accuracy with a 5% tolerance threshold, following the standard ChartQA setup (Masry et al., 2022). For ChartQA, we report the scores established in prior benchmark studies (Chen et al., 2024b; Lu et al., 2024).

Models We conduct experiments with 19 MLLMs released in 2023-2024 on a machine with two A100 GPUs, including 11 open-weight MLLMs from the Llava-Next (Liu et al., 2024), Qwen2VL (Wang et al., 2024), Ovis1.6 (Lu et al., 2024), and InternVL2.5 (Chen et al., 2024a) families. We also include five commercial models, GPT4 and GPT4o (OpenAI, 2023), Gemini-1.5-flash and -pro (Gemini-Team, 2024), and Claude-3.5-sonnet (Anthropic, 2024), as well as three open-weight MLLMs specialized in chart understanding: Llava-Chart-Instruct (Zeng et al., 2025), TinyChart (using the Direct approach) (Zhang et al., 2024), and ChartGemma (Masry et al., 2025). Qwen2.5-7B (Qwen-Team, 2024) serves as the LLM for table-based QA and visualization redrawing. We use the transformers library (Wolf et al., 2020) to access all open-weight models. All prompts are provided in Appendix B. We set the temperature to 0. Following the standard ChartQA evaluation setup, all (M)LLMs are prompted in zero-shot.

Random baseline We compare the MLLMs against a random baseline. Its accuracy is 1/n for an MCQ with n choices, 0 for free-text, and 1/n! for ranking with n items.

3 Results

3.1 Experiment 1 - Assessing MLLM vulnerabilities (Accuracy)

Experiment 1 assesses the ability of MLLMs to answer questions about the underlying data given a visualization. The upper part of Figure 4 presents the QA accuracy of 19 MLLMs across the three datasets, sorted by increasing accuracy on ChartQA. We report bootstrapped (n=5000) confidence intervals (CIs) for the misleading and non-misleading datasets. The bottom part displays the accuracy for the three subsets constituting the misleading visualizations dataset. Additional results are provided in Appendix C. Experiment 1 reveals three key findings.

Performance is worse on misleading visualizations. All MLLMs perform substantially worse on misleading visualization compared to the non-misleading visualizations dataset, with accuracy dropping by up to 34.8 percentage points (pp), with an average of 23.8 pp. The decline is even more pronounced compared to ChartQA, reaching up to 65.5 pp, with an average of 53.8 pp. Further-

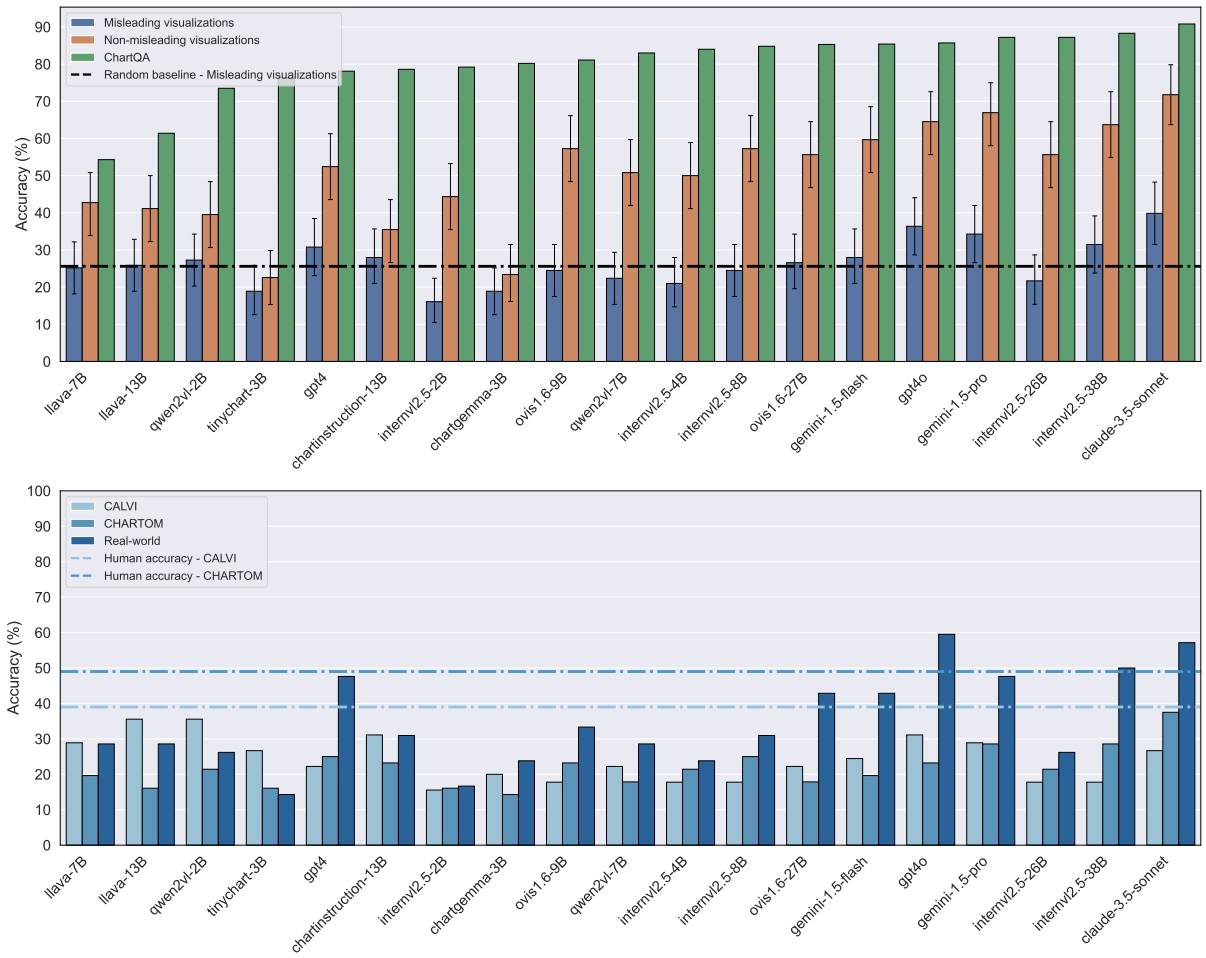

Figure 4: Top: QA accuracy (%) on the misleading visualization, non-misleading visualization, and ChartQA datasets. Confidence intervals are reported for the misleading and non-misleading datasets. The horizontal dashed line indicates the accuracy of the random baseline on misleading visualizations. Models are sorted by increasing accuracy on ChartQA. Bottom: Accuracy (%) of various MLLMs on subsets of the misleading visualizations. The horizontal dashed lines indicate average human accuracy on CALVI and CHARTOM.

more, the mean MLLM accuracy on misleading visualizations (26.4%) is comparable to the random baseline score (25.6%). These results demonstrate that MLLMs struggle to interpret data distorted by misleaders correctly.

For comparison, average human accuracy reported on CALVI and CHARTOM is 80 and 89% for non-misleading visualizations, dropping to 39 and 49% for misleading ones, respectively (Ge et al., 2023; Rho et al., 2023; Bharti et al., 2024). On the same datasets, MLLMs achieve an average of 52 and 48% on non-misleading visualizations and 24 and 22% on misleading ones, falling short of human performance in all cases.

The ranking of MLLMs by accuracy varies a lot across the three datasets. The two best-performing models on the CALVI subset of misleading visualizations rank among the worst on ChartQA, placing second and third from the bottom. This suggests that achieving high accuracy on misleading visualizations is unlikely to emerge naturally from improvements on ChartQA, underscoring the need for dedicated mitigation methods.

Performance is higher on real-world misleading visualizations than on other subsets. MLLMs perform better on the real-world subset, achieving an average accuracy of 34.7%, compared to 24% on CALVI and 22% on CHARTOM. In general, the best-performing MLLMs on misleading visualizations, Claude-3.5-sonnet, GPT4o, and Gemini-1.5-pro, primarily distinguish themselves through their higher performance on this subset. We hypothesize that this advantage stems from their parametric knowledge, which extends beyond 2022, the endpoint of the real-world subset. This allows these models to answer some questions based on stored

knowledge of past events, without relying on the visualization's content. Appendix D discusses further the impact of parametric knowledge. Other factors, which are not quantified in this work, might explain the better performance on the real-world subset, such as the chart aesthetics and visual complexity, or the difficulty of the MCQs compared to CALVI and CHARTOM.

Performance varies by type of misleader. Not all misleading visualizations are equally deceptive. This was already observed with human readers (Ge et al., 2023; Rho et al., 2023; Bharti et al., 2024), and is here further confirmed with MLLMs. The following analysis considers only misleaders with at least five occurrences in the dataset. The two misleaders on which MLLMs perform best are *misleading annotations* and *misrepresentation*, with average accuracies of 50.5 and 44.4%. The two on which they perform worse are *area encoding* and *cherry-picking* at 7.4 and 10.5%. For the two most represented misleaders, *truncated axis* and *inverted axis*, the MLLMs achieve average accuracies of 25.1 and 28.1 %. There are notable differences in which misleaders pose greater challenges to MLLMs compared to humans, and vice versa. For instance, *area encoding* is highly misleading for MLLMs, but not for human readers, who achieve an average accuracy of 91.5% (Rho et al., 2023). Conversely, both humans and MLLMs struggle with misleaders such as *3D effects* (Ge et al., 2023). The use of a *dual axis* is a misleader that poses greater difficulty for humans, who score an average accuracy of 16.1% on CHARTOM (Rho et al., 2023), compared to MLLMs, which achieve an average of 35.5%. These results highlight a notable difference in how misleaders deceive humans and MLLMs. This suggests that the methods needed to make MLLMs more robust to misleading instances will, at least in part, differ from those designed to protect human readers (Fan et al., 2022).

3.2 Experiment 2 - Assessing MLLM vulnerabilities (Consistency)

Experiment 2 evaluates whether MLLMs interpret consistently two visualizations that represent the same underlying data table, one misleading and the other non-misleading. This is assessed using a Likert-scale question, such as in Figure 2. Unlike in Experiment 1, there is no ground truth answer. Instead, since both visualizations depict the same underlying data, a model is considered misled if it assigns different scores to the two. We assess

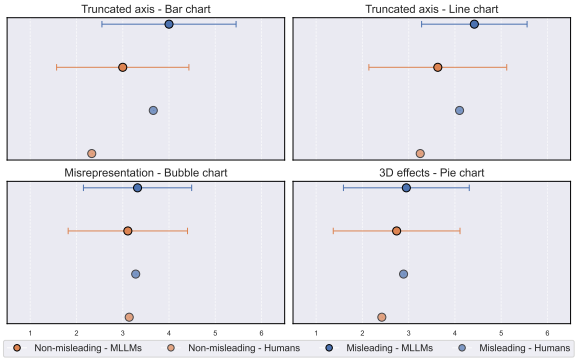

Figure 5: Average Likert-scale ratings (1 to 6) in Experiment 2. Average MLLM results are reported with standard deviations. Average human results are reported from Lauer and O'Brien (2020).

the significance of the difference in ratings with a Wilcoxon signed-rank test (p-value (p) ≤ 0.05).

Figure 5 summarizes the average Likert-scale ratings from all 19 MLLMs for each visualization pair. Average human results from prior work (Lauer and O'Brien, 2020) are included for comparison. The average rating difference between the misleading and non-misleading visualizations is similar for humans and MLLMs, though MLLMs tend to assign slightly higher scores overall. For MLLMs, the inconsistency in ratings is significant for the two visualization pairs involving *truncated axis*: one with bar charts (p = 7e-3) and one with line charts (p = 6e-3). For the other two pairs, *misrepresentation* with bubble charts (p = 0.36) and *3D effects* with pie charts (p = 0.49), the differences are not significant. These results give an initial indication that MLLMs are unable to consistently recognize the same underlying data table behind a misleading and a non-misleading version of the same visualization, particularly for *truncated axis*. However, these results are only preliminary and need to be further validated in future work with a larger set of MLLMs on more data pairs.

3.3 Experiment 3 - Correction methods

Having established that MLLMs are vulnerable to misleading visualizations, we examine six correction methods designed to enhance their robustness. Our focus is on increasing accuracy on the misleading visualizations dataset from Experiment 1, while maintaining performance on the non-misleading dataset. We evaluate the impact of six correction methods using three open-weight, mid-sized MLLMs, Qwen2VL-7B, Ovis1.6-9B, and InternVL2.5-8B, all of which performed below

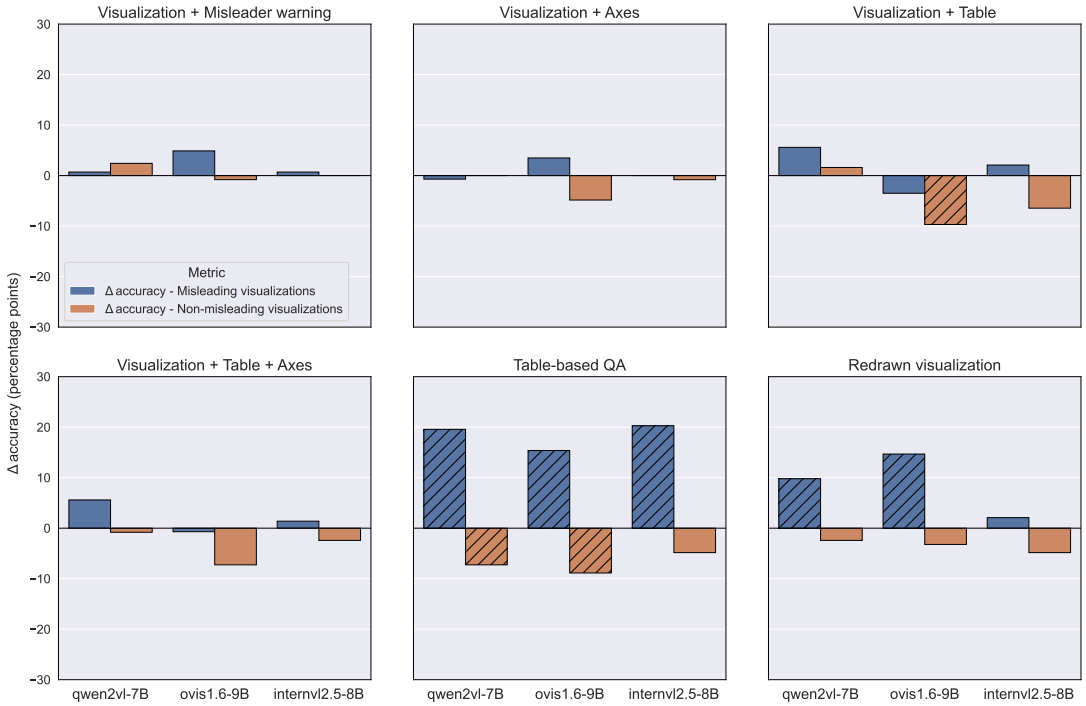

Figure 6: Change (Δ) in accuracy (percentage points) compared to Experiment 1 using different inference-time correction methods. Statistically significant changes ($p \leq 0.05$) are hashed.

the random baseline in Experiment 1. Appendix E provides the detailed results. Statistical significance of the changes in accuracy is assessed using the two-sided McNemar test ($p \leq 0.05$) (McNemar, 1947). Figure 6 shows the change in accuracy relative to Experiment 1 across the six methods: (1) inclusion of a warning message, (2) inclusion of extracted axes, (3) inclusion of the extracted table, (4) inclusion of both axes and table, (5) table-based QA, and (6) redrawing the visualization.

Two correction methods are effective: table-based QA and redrawing the visualization. The most effective approach to counter misleaders is table-based QA, which yields significant improvements on misleading visualizations, ranging from 15.4 to 19.6 pp. Originally proposed as a general approach for chart understanding (Liu et al., 2023a,b), table-based QA further shows its effectiveness for countering misleading charts. Its strength lies in replacing the visualization with a data table, thereby eliminating misleaders that do not transfer to the tabular format, such as *inverted* and *truncated axis*. However, this method proves less effective against misleaders that persist even when the data is presented in tabular format, such as *inappropriate item order*. Table-based QA incurs a significant cost for two models on non-misleading visualizations, by up to 8.5 pp. This degradation is due to errors in the intermediate table extraction step, which introduce incorrect values or cause relevant information to be lost. Appendix E reports table-based QA performance using DePlot (Liu et al., 2023a) and MatCha (Liu et al., 2023b), two methods developed specifically for table extraction.

Another promising correction method is redrawing the visualization. This method aligns with prior efforts to make human readers more robust to misleading visualizations (Fan et al., 2022). It leads to significant, though more modest, improvements for two MLLMs. The visualization is redrawn using the default settings of the Python library Matplotlib (Hunter, 2007), which inherently avoids certain misleaders like inverting the axes or using inconsistent tick intervals. In Figure 3, the redrawn visualization corrects the misleading trend line by eliminating inconsistent tick intervals, leading to a more accurate representation of the data. This correction method is effective only if the generated code compiles successfully on a Python interpreter; otherwise, the original visualization is used. The lowest redrawing success rates are observed for scatter plots (79%) and stacked bar charts (80%), likely due to their higher visual complexity and number of elements, which require Qwen2.5-7B to produce longer Python code. Unlike table-based QA, there are no significant decreases in accuracy

on non-misleading visualizations, making it a more attractive option when preserving performance on non-misleading visualizations is a priority.

Appendix E discusses the impact of combining these two effective methods with Chain-of-Thought prompting (Wei et al., 2022).

Table extraction is a key intermediate step. Both table-based QA and redrawing the visualization depend heavily on the accuracy of the intermediate table extraction step. This extraction step is often non-trivial, particularly for maps or scatter plots that contain many visual elements. Additionally, misleaders themselves can degrade MLLMs' ability to extract data accurately. This is particularly prevalent for *inverted axis*, *3D effects*, and *dual axis*. Perfect accuracy in table extraction is not always necessary; its importance depends on the nature of the question. Some questions can be answered correctly if the extracted table preserves the overall trend of the underlying data, while others require exact values. We conduct a qualitative analysis of table extraction in Appendix F.

Other correction methods are not effective. None of the remaining correction methods yields significant improvements. While using the extracted table alone is the most effective method, combining it with the misleading visualization does not yield similar gains. The MLLMs are biased toward using the image rather than the table.

Although many misleaders rely on axes manipulations, providing the extracted axes has no significant effect on performance. A manual error analysis showed that the axes extraction step is more often accurate than table extraction, as reported in Appendix H. However, unlike the table, the axes alone are insufficient to answer the question. The MLLM still needs to combine the visualization image with cues from the axes indicating the presence of misleaders. The results show that this combination is challenging. This further suggests that removing or modifying the misleading visualization, rather than merely supplementing it with additional prompt input, is a key factor in the success of a correction method.

Adding a warning message produces only non-significant changes, with accuracy gains of at most 4.9 pp, mostly limited to the real-world subset. This warning-based approach assumes prior knowledge of the specific misleader in the visualization, making the reported results an upper bound on its effectiveness. In practice, a classifier needs to detect first the presence of misleaders, which remains a challenging task (Lo and Qu, 2025; Alexander et al., 2024). Given the already low results obtained with ground-truth misleader labels, this correction method appears unpromising overall. However, training a highly accurate misleader detection model could enable the selective application of correction methods. This would offer two key advantages: (1) avoiding the application of correction methods to non-misleading visualizations, thereby eliminating any risk of negative impact; and (2) allowing the selection of the most suitable correction method based on the type of misleader.

4 Related work

The correction methods, particularly misleader warnings and redrawing the visualization, relate to prior work on visualization linters (Hopkins et al., 2020; Chen et al., 2022; Fan et al., 2022): rule-based methods to detect and correct misleaders. While these tools proved useful to make human readers more robust to misleading visualizations (Fan et al., 2022), they can only operate with specific chart design tools (Hunter, 2007; Hopkins et al., 2020; Chen et al., 2022) or a very limited set of chart and misleader types (Fan et al., 2022). In contrast, the correction methods we explored do not impose such restrictions. Table-based QA was originally introduced as a general approach for chart understanding by Liu et al. (2023a,b). Although this approach has been surpassed on standard benchmarks by more recent methods that treat visualizations purely as images (Zeng et al., 2025; Masry et al., 2025), our results reveal that table-based QA finds a new purpose as an effective method for counteracting misleading visualizations.

Our findings are further supported by prior work (Bendeck and Stasko, 2025) and parallel studies (Pandey and Ottley, 2025; Chen et al., 2025; Valentim et al., 2025), which also report MLLM vulnerabilities to misleaders (Bendeck and Stasko, 2025; Pandey and Ottley, 2025; Chen et al., 2025) and to other design decisions like the color palette (Valentim et al., 2025). Our study differentiates itself in several key ways. Prior work (Bendeck and Stasko, 2025) only evaluated GPT4 on six misleading visualizations, providing initial hints but lacking depth to fully establish the presence of a vulnerability. Unlike parallel works, we include real-world visualizations and show that the vulnerability of MLLMs is lower in these cases. Furthermore, we provide the first evaluation of six correction methods. Cru-

cially, we examine the trade-off between improving QA performance on misleading visualizations and preserving accuracy on non-misleading ones, an aspect not addressed in other works.

5 Conclusion

Our findings highlight the vulnerability of MLLMs to misleading visualizations. Identifying this vulnerability fills a critical gap in the research on automated chart understanding. While MLLMs achieve strong performance on standard benchmarks such as ChartQA, they remain vulnerable to misleading visualizations. This raises serious concerns about their reliability in real-world settings, especially given the potential for such vulnerabilities to be exploited by malicious actors to spread disinformation (Correll and Heer, 2017). To mitigate this issue, we evaluated six inference-time correction methods, two of which, table-based QA and redrawing the visualization, demonstrated significant improvements. However, correction methods often come at the cost of reduced accuracy on non-misleading visualizations.

Limitations

We identify three limitations to our work.

First, the visualization redrawing method does not support maps, as the Matplotlib library (Hunter, 2007) lacks sufficient functionality for rendering high-quality maps.

Second, we assume prior knowledge of the chart type (e.g., bar, line) to generate prompts for axes extraction and redrawing. This is a reasonable assumption, as the chart type can either be provided by a human user or accurately predicted by a classifier as a preprocessing step.

Third, existing misleading visualization datasets are small compared to standard datasets in the automated chart understanding literature, such as ChartQA. Furthermore, they do not provide an equal representation of all misleaders, with *truncated* and *inverted axes* being the two most represented categories. This imbalance is prevalent in CALVI (Ge et al., 2023) and, to a lesser extent, in real-world data. Despite these limitations, CALVI is the reference dataset for assessing human vulnerability to misleading visualizations. It was carefully curated by experts in data visualization and validated on a large sample of human subjects. Therefore, it constitutes a valid resource for evaluating MLLMs. The properties of CALVI are similar to those of VLAT (Lee et al., 2017), an expert-created dataset for non-misleading chart understanding, which acts as a reference despite being much smaller than datasets such as ChartQA. Furthermore, the real-world distribution of misleaders is inherently imbalanced (Lo et al., 2022), with categories such as *truncated axis* being more prevalent than others. Hence, scaling the real-world dataset would not address the imbalance issues of synthetic datasets, such as CALVI.

Ethics statement

Social impact Misleading visualizations are a prevalent form of multimodal misinformation. Our results show the importance of protecting not only humans but also MLLMs from misleading visualizations. Our correction methods provide the first promising results in that direction.

Risks While the proposed correction methods improve the performance of MLLMs on misleading visualizations, they do not guarantee a correct answer. In general, human users should not blindly trust the output of chart understanding systems.

Dataset access We release the CALVI, VLAT, Likert-scale, and real-world datasets under a CC-BY-SA-4.0 license. The real-world dataset is intended solely for research purposes. CHARTOM can be accessed by contacting its authors (Bharti et al., 2024).

AI assistants use AI assistants were used in this work to assist with writing by correcting grammar mistakes and typos.

Acknowledgments

This work has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center (Grant Number: LOEWE/1/12/519/03/05.001(0016)/72), by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, and by the Flanders AI Research Program. We gratefully acknowledge the support of Microsoft with a grant for access to OpenAI GPT models via the Azure cloud (Accelerate Foundation Model Academic Research). The figures have been designed using resources from Flaticon.com. We want to express our gratitude to Niklas Traser for conducting an

initial exploration of the real-world data, to Jan Zimny for our insightful discussions on misleading visualizations, and to Germàn Ortiz, Manisha Venkat, and Max Glockner for their feedback on a draft of this work.

References

Jason Alexander, Priyal Nanda, Kai-Cheng Yang, and Ali Sarvghad. 2024. Can gpt-4 models detect misleading visualizations? In *2024 IEEE Visualization and Visual Analytics (VIS)*, pages 106–110.

Anthropic. 2024. Introducing the next generation of claude. Accessed: 2025-11-26.

Alexander Bendeck and John Stasko. 2025. An empirical evaluation of the gpt-4 multimodal language model on visualization literacy tasks. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1105–1115.

Shubham Bharti, Shiyun Cheng, Jihyun Rho, Martina Rao, and Xiaojin Zhu. 2024. Chartom: A visual theory-of-mind benchmark for multimodal large language models. *arXiv preprint arXiv:2408.14419*, abs/2408.14419.

Qing Chen, Fuling Sun, Xinyue Xu, Zui Chen, Jiazhe Wang, and Nan Cao. 2022. Vizlinter: A linter and fixer framework for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):206–216.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, abs/2412.05271.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zixin Chen, Sicheng Song, KaShun Shum, Yanna Lin, Rui Sheng, Weiqi Wang, and Huamin Qu. 2025. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13767–13800, Suzhou, China. Association for Computational Linguistics.

Michael Correll and Jeffrey Heer. 2017. Black hat visualization. In *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), IEEE VIS*.

Arlen Fan, Yuxin Ma, Michelle Mancenido, and Ross Maciejewski. 2022. Annotating line charts for addressing deception. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Lily W. Ge, Yuan Cui, and Matthew Kay. 2023. Calvi: Critical thinking assessment for literacy in visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Gemini-Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google.

Gemini-Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google.

Aspen K. Hopkins, Michael Correll, and Arvind Satyanarayan. 2020. Visualint: Sketchy in situ annotations of chart construction errors. *Computer Graphics Forum*, 39(3):219–228.

E. Hoque, P. Kavehzadeh, and A. Masry. 2022. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41(3):555–572.

Kung-Hsiang Huang, Hou Pong Chan, May Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2025. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2550–2568.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Wonjoong Kim, Sangwu Park, Yeonjun In, Seokwon Han, and Chanyoung Park. 2025. SIMPLOT: Enhancing chart question answering by distilling essentials. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 573–593, Albuquerque, New Mexico. Association for Computational Linguistics.

Xingyu Lan and Yu Liu. 2025. "i came across a junk": Understanding design flaws of data visualization from the public's perspective. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):393–403.

Claire Lauer and Shaun O'Brien. 2020. The deceptive potential of common design tactics used in data visualizations. In *Proceedings of the 38th ACM International Conference on Design of Communication*, SIGDOC '20, New York, NY, USA. Association for Computing Machinery.

Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2017. Vlat: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560.

Maxim Lisnic, Cole Polychronis, Alexander Lex, and Marina Kogan. 2023. Misleading beyond visual tricks: How people actually lie with charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

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 Eisenschlos. 2023b. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. 2022. Misinformed by visualization: What do we learn from misinformative visualizations? In *Computer Graphics Forum*, volume 41, pages 515–525. Wiley Online Library.

Leo Yu-Ho Lo and Huamin Qu. 2025. How good (or bad) are llms at detecting misleading visualizations? *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1116–1125.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, abs/2405.20797.

Ahmed Masry, Do Xuan Long, 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.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2025. ChartGemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643, Abu Dhabi, UAE. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Andrew McNutt, Gordon Kindlmann, and Michael Correll. 2020. Surfacing visualization mirages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA. Association for Computing Machinery.

Shaun O'Brien and Claire Lauer. 2018. Testing the susceptibility of users to deceptive data visualizations when paired with explanatory text. In *Proceedings of the 36th ACM International Conference on the Design of Communication*, SIGDOC '18, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. 2014. The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2211–2220.

Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. 2015. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1469–1478, New York, NY, USA. Association for Computing Machinery.

Saugat Pandey and Alvitta Ottley. 2025. Benchmarking visual language models on standardized visualization literacy tests. *Computer Graphics Forum*, 44(3):e70137.

Qwen-Team. 2024. Qwen2.5 technical report. Technical report, Alibaba Cloud.

Jihyun Rho, Martina A Rau, Shubham Kumar Bharti, Rosanne Luu, Jeremy McMahan, Andrew Wang, and Jerry Zhu. 2023. Various misleading visual features in misleading graphs: Do they truly deceive us? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

JonRobert Tartaglione and Lee de Wit. 2025. How the manner in which data is visualized affects and corrects (mis)perceptions of political polarization. *British Journal of Social Psychology*, 64(1):e12787.

Edward R Tufte and Peter R Graves-Morris. 1983. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, Cheshire, CT, USA.

Matheus Valentim, Vaishali Dhanoa, Gabriela Molina León, and Niklas Elmqvist. 2025. The plot thickens: Quantitative part-by-part exploration of mllm visualization literacy. *arXiv preprint arXiv:2504.02217*, abs/2504.02217.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, abs/2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Brenda W. Yang, Camila Vargas Restrepo, Matthew L. Stanley, and Elizabeth J. Marsh. 2021. Truncating bar graphs persistently misleads viewers. *Journal of Applied Research in Memory and Cognition*, 10(2):298–311.

Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. 2025. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):525–535.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1898, Miami, Florida, USA. Association for Computational Linguistics.

Yixuan Zhang, Yifan Sun, Lace Padilla, Sumit Barua, Enrico Bertini, and Andrea G Parker. 2021. Mapping the landscape of covid-19 crisis visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Table extraction prompt

Generate the underlying data table of the figure below. Change columns with |, change row by starting a new line. Provide only the table as output.

Figure 7: Table extraction prompt.

Axes extraction prompt

For maps: What is the legend and its categories (with their colors) in this map? Answer only with the content of the legend and its categories (with their colors).

For pie charts: What are the categories in this pie chart? Answer only with the categories.

Other chart types: What are the axis labels and ticks of this chart? Answer only with the axis labels and the tick values, going from the bottom-left to the top-left corner of the chart for the y-axis and from the bottom-left to the bottom-right corner for the x-axis.

Figure 8: Axes extraction prompt.

A Definition of misleaders

Table 2 provides the definitions of all the misleaders considered in this work (Lo et al., 2022; Ge et al., 2023), with their number of occurrences.

B Prompts

Figures 8, 7, and 9 provide the prompt for the intermediate metadata extraction and visualization redrawing steps. Figure 10 provides the QA prompt variants. Different parts of the prompt are included depending on the question type and additional input of correction methods.

C Experiment 1 - Additional results

QA accuracy with confidence intervals Table 3 presents QA accuracy on the misleading and non-misleading datasets, with confidence intervals.

QA accuracy by question type Table 4 provides the QA accuracy per question type for each MLLM and the random baseline.

Misleader	Definition
Inverted axis (n=26)	An axis is oriented in an unconventional direction and the perception of the data is reversed (Lo et al., 2022).
Truncated axis (n=21)	The axis does not start from zero or is truncated in the middle, resulting in an exaggerated difference between the two bars (Lo et al., 2022).
Inappropriate axis range (n=15)	The axis range is either too broad or too narrow to accurately visualize the data, allowing changes to be minimized or maximized depending on the author's intention (Lo et al., 2022).
Inconsistent tick intervals (n=12)	Cases with varying intervals between the ticks (Lo et al., 2022).
3D effects (n=12)	The closer something is, the larger it appears, despite being the same size in 3D perspective (Lo et al., 2022).
Inappropriate item order (n=9)	The axis labels or legends appear to be in a random order due to manipulation of data ordering (Ge et al., 2023).
Inappropriate aggregation (n=8)	Aggregating data in an improper way that leads to inaccurate conclusions (Ge et al., 2023).
Dual axis (n=8)	Two independent axes are layered on top of each other with inappropriate scaling (Lo et al., 2022).
Misrepresentation (n=7)	The value labels provided do not match the visual encoding (Lo et al., 2022).
Cherry picking (n=6)	Selecting only a subset of data to display, which can be misleading if one is asked to infer something about the whole set of data (Ge et al., 2023).
Misleading annotations (n=5)	Annotations that contradict or make it harder to read the visualization (Ge et al., 2023).
Area encoding (n=5)	Linearly encoding the values as areas leads the readers to consistently underestimate the values (Lo et al., 2022).
Concealed uncertainty (n=3)	Not displaying uncertainty in visualizations may misrepresent the certainty in the underlying data. In the case of prediction making, this can misguide the viewers to falsely overconfident conclusions (Ge et al., 2023).
Missing normalization (n=3)	Displaying unnormalized data in absolute quantity when normalized data in relative quantity is of interest (Ge et al., 2023).
Inappropriate use of pie chart (n=1)	When a pie chart is used for non-part-to-whole data, it creates confusion for the audience, who may misinterpret the significance of a given section (Lo et al., 2022).
Missing data (n=1)	A visual representation implies data exist but the data is actually missing (Ge et al., 2023).
Overplotting (n=1)	Displaying too many things on a plot can obscure parts of the data (Ge et al., 2023).

Table 2: The misleaders included in this work, with their number of occurrences (n).

Visualization redrawing prompt

Generate the matplotlib code to generate this {CHART_TYPE}, using the tabular data below.
Provide only the code as output, including the table values represented as a list or a numpy array. {TABLE}

Figure 9: Visualization redrawing prompt.

QA prompt

If table-based QA: {TABLE}

{QUESTION}

If MCQ: Provide the correct answer among the following choices: {CHOICES}
If rank: Provide the answer as a Python list.

If misleader warning: Be careful, the following design flaw has been identified in the chart: {MISLEADER DEFINITION}

If axes: Below is a description of the charts axis labels or legend. {AXIS}

If table: Below is a table containing the values represented in the chart. {TABLE}

Provide only the final answer to the question.

Figure 10: QA prompt.

10 MLLMs perform worse than or equal to the random baseline on MCQs for misleading visualizations. There are only two for the non-misleading ones. Free text is the only question type where some MLLMs perform better on misleading instances than on non-misleading ones. We attribute this to the difficulty of the free text questions in the non-misleading VLAT dataset. While only four MLLMs achieve non-zero accuracy on rank questions with misleading instances, 13 do so for non-misleading ones, indicating both the difficulty of the task and a strong negative impact of the *3D effect* misleader, which affects all rank questions.

Model	Misleading visualizations QA accuracy	Misleading visualizations CI	Non-misleading visualizations QA accuracy	Non-misleading visualizations CI
llava-7B	25.2	(18.2, 32.2)	42.7	(33.9, 50.8)
llava-13B	25.9	(18.9, 32.9)	41.1	(32.3, 50.0)
qwen2vl-2B	27.3	(20.3, 34.3)	39.5	(30.7, 48.4)
tinychart-3B	18.9	(12.6, 25.9)	22.6	(15.3, 29.8)
gpt4	30.8	(23.1, 38.5)	52.4	(43.6, 61.3)
chartinstruction-13B	28.0	(21.0, 35.7)	35.5	(26.6, 43.6)
internvl2.5-2B	16.1	(10.5, 22.4)	44.4	(35.5, 53.2)
chartgemma-3B	18.9	(12.6, 25.2)	23.4	(16.1, 31.5)
ovis1.6-9B	24.5	(17.5, 31.5)	57.3	(48.4, 66.1)
qwen2vl-7B	22.4	(15.4, 29.4)	50.8	(41.9, 59.7)
internvl2.5-4B	21.0	(14.7, 28.0)	50.0	(41.1, 58.9)
internvl2.5-8B	24.5	(17.5, 31.5)	57.3	(48.4, 66.1)
ovis1.6-27B	26.6	(19.6, 34.3)	55.7	(46.8, 64.5)
gemini-1.5-flash	28.0	(21.0, 35.7)	59.7	(50.8, 68.6)
gpt4o	36.4	(28.7, 44.1)	64.5	(55.7, 72.6)
gemini-1.5-pro	34.3	(26.6, 42.0)	66.9	(58.1, 75.0)
internvl2.5-26B	21.7	(15.4, 28.7)	55.7	(46.8, 64.5)
internvl2.5-38B	31.5	(23.8, 39.2)	63.7	(54.8, 72.6)
claude-3.5-sonnet	39.9	(31.5, 48.3)	71.8	(63.7, 79.8)

Table 3: Experiment 1 main results with bootstrapped confidence intervals (n=5000) (%).

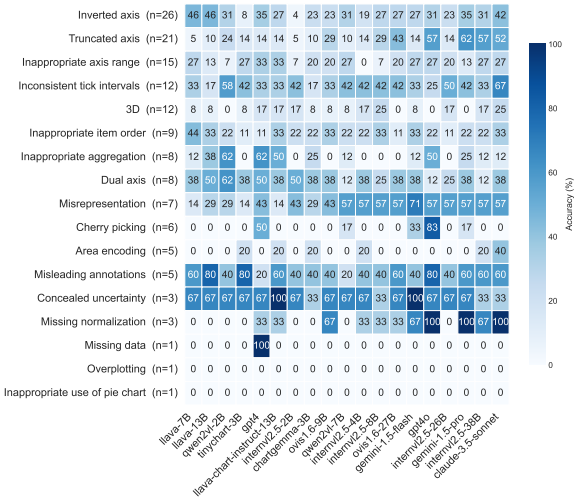

Figure 11: QA Accuracy (%) per type of misleader.

QA accuracy by misleader Figure 11 provides the QA accuracy per misleader for each MLLM.

QA accuracy on CHARTOM Table 5 compares the QA accuracy on the misleading and non-misleading subsets of CHARTOM, which are based on the same underlying data tables. The change (Δ) in accuracy between the two subsets is larger than 20 pp for 14 out of 19 models. The exceptions are datasets on which the MLLMs achieve accuracy below 40% for non-misleading visualizations, indicating limited chart understanding abilities. Improved performance on the non-misleading subset does not consistently result in a reduced Δ. These results further indicate that improving MLLMs' general chart understanding does not entirely address their vulnerability to misleading visualizations.

Results with models released in 2025 While the focus of this work is on the MLLMs released

Model	MCQ (n=115)		Free-text (n=20)		Rank (n=8)	
	Misleading	Non-misleading	Misleading	Non-misleading	Misleading	Non-misleading
Random baseline	31.3	33.9	0.0	0.0	0.1	0.1
llava-7B	31.3	54.2	0.0	5.0	0.0	0.0
llava-13B	31.3	53.1	5.0	0.0	0.0	0.0
qwen2vl-2B	33.0	51.0	5.0	0.0	0.0	0.0
tinychart-3B	22.6	29.2	5.0	0.0	0.0	0.0
gpt4	36.5	64.6	0.0	5.0	**25.0**	25.0
chartinstruction-13B	33.9	44.8	5.0	5.0	0.0	0.0
internvl2.5-2B	19.1	56.2	5.0	5.0	0.0	0.0
chartgemma-3B	22.6	29.2	5.0	5.0	0.0	0.0
ovis1.6-9B	26.1	69.8	25.0	15.0	0.0	12.5
qwen2vl-7B	27.0	62.5	5.0	0.0	0.0	37.5
internvl2.5-4B	24.4	61.5	5.0	10.0	12.5	12.5
internvl2.5-8B	27.8	69.8	10.0	15.0	12.5	12.5
ovis1.6-27B	32.2	60.4	5.0	40.0	0.0	37.5
gemini-1.5-flash	33.0	70.8	10.0	10.0	0.0	**50.0**
gpt4o	40.9	72.9	25.0	30.0	0.0	50.0
gemini-1.5-pro	37.4	72.9	30.0	**45.0**	0.0	**50.0**
internvl2.5-26B	24.4	67.7	15.0	20.0	0.0	0.0
internvl2.5-38B	33.0	75.0	**35.0**	25.0	0.0	25.0
claude-3.5-sonnet	**44.4**	**85.4**	25.0	25.0	12.5	25.0

Table 4: Experiment 1 results by question type (%), with their number of occurence (*n*). The best results are marked in bold.

Model	Misleading CHARTOM	Non-misleading CHARTOM	Δ
llava-7B	19.6	39.3	19.7
llava-13B	16.1	32.1	16.0
qwen2vl-2B	21.4	44.6	23.2
tinychart-3B	16.1	25.0	8.9
gpt4	25.0	46.4	21.4
chartinstruction-13B	23.2	37.5	14.3
internvl2.5-2B	16.1	44.6	28.5
chartgemma-3B	14.3	17.9	3.6
ovis1.6-9B	23.2	53.6	30.4
qwen2vl-7B	17.9	53.6	35.7
internvl2.5-4B	21.4	50.0	28.6
internvl2.5-8B	25.0	53.6	28.6
ovis1.6-27B	17.9	51.8	33.9
gemini-1.5-flash	19.6	53.6	34.0
gpt4o	23.2	64.3	41.1
gemini-1.5-pro	28.6	64.3	35.7
internvl2.5-26B	21.4	53.6	32.2
internvl2.5-38B	28.6	62.5	33.9
claude-3.5-sonnet	37.5	62.5	25.0

Table 5: QA accuracy by subset on CHARTOM (%).

in 2023-2024, we provide preliminary results with three commercial models released in 2025: GPT4.1 (OpenAI, 2023), GPT5-mini, and Gemini-2.5-flash-lite (Gemini-Team, 2025). Their QA accuracy is reported against the best model from 2024, Claude-3.5-sonnet, in Table 6. All 2025 models outperform Claude-3.5-sonnet on misleading instances. GPT5-mini even achieves an accuracy above 50%. However, these MLLMs also perform better on the non-misleading dataset, and a large change (Δ) in accuracy of more than 20% remains for all of them.

Model	Misleading visualizations	Non-misleading visualizations	ChartQA
Claude-3.5-sonnet	39.9	71.8	**90.8**
Gemini-2.5-flash	42.0	66.9	76.8
GPT5-mini	53.2	81.5	88.2
GPT5	**55.9**	**89.5**	89.6

Table 6: QA accuracy of Claude-3.5-sonnet compared with recent models released in 2025 (%).

D Experiment 1 - Impact of parametric knowledge

We observed in Experiment 1 that Claude-3.5-sonnet, GPT4o, and Gemini-1.5-pro primarily outperform other MLLMs on the real-world subset of the misleading visualizations. We assume this is due to their parametric knowledge, which allows them to answer the question without considering the visualization. To support this assumption, we identified a subset of 22 real-world MCQS out of 42 that could be answered using world knowledge. We rephrased the questions slightly to make them self-contained, without direct references to the visualization's content. The accuracy of GPT4o on this subset is 77%. If we provide only the MCQ to GPT4o, without the visualization's image, the accuracy is still 50%, highlighting indeed a moderate ability to answer the real-world MCQs based on parametric knowledge. In such cases, parametric knowledge effectively serves as a form of protec-

		Misleading visualizations	Non-misleading visualizations
qwen2vl-7B	Default	22.4	48.4
	Misleader waring	23.1	**50.8**
	Axes	21.7	48.4
	Table	28.0	50.0
	Table + axes	28.0	47.6
	Table-based QA	**42.0**	41.1
	Redrawn visualization	32.2	46.0
ovis1.6-9B	Default	24.5	**56.5**
	Misleader waring	29.4	55.7
	Axes	28.0	51.6
	Table	21.0	46.8
	Table + axes	23.8	49.2
	Table-based QA	**39.9**	47.6
	Redrawn visualization	39.2	53.2
internvl2.5-8B	Default	23.8	**57.3**
	Misleader waring	24.5	**57.3**
	Axes	23.8	56.5
	Table	25.9	50.8
	Table + axes	25.2	54.8
	Table-based QA	**44.1**	52.4
	Redrawn visualization	25.9	52.4

Table 7: QA accuracy results for Experiment 3 (%).

Model	Misleading visualizations		Non-misleading visualizations	
	Table-based QA	Redrawing	Table-based QA	Redrawing
ChartGemma-3B	+27.3	+2.1	+23.4	+1.6
GPT5-mini	- 9.1	+2.8	- 20.1	-3.2

Table 8: Change (Δ) in accuracy (percentage points) compared to Experiment 1 using different inference-time correction methods with ChartGemma-3B and GPT5-mini.

Model	Misleading visualizations	Non-misleading visualizations
DePlot	**44.1**	46.0
MatCha	42.7	37.9
Qwen2VL-7B	42.0	41.1
Ovis1.6-9B	39.9	47.6
InternVL2.5-8B	**44.1**	**52.4**

Table 9: Table-based QA accuracy using Qwen2.5-7B, with different models for table extraction (%).

tion against real-world misleaders. However, this finding should be interpreted with caution. In practical scenarios, the underlying data will often be recent and unlikely to be covered by the parametric knowledge. Other factors might explain the performance on the real-world subset, including differences in visualization complexity and question difficulty compared to CALVI and CHARTOM.

E Experiment 3 - Additional results

QA accuracy by correction method Table 7 provides the detailed QA accuracy results for the default prompt and for the six correction methods.

Results with ChartGemma and GPT5-mini We apply the two most effective correction methods, table-based QA and redrawing the visualization, to the strongest chart-specialized model, ChartGemma, and a strong commercial MLLM released in 2025, GPT5-mini. Table 8 contains the results.

ChartGemma excels at table extraction and performed very poorly in Experiment 1. This results in a high Δ accuracy for table-based QA with Qwen2.5-7B, on both datasets. Due to ChartGemma's weaker visual reasoning abilities, redrawing the visualization does not yield significant improvements.

In contrast, GPT5-mini's performance is already strong, as shown in Appendix C. Hence, it does not benefit from providing the extracted table to Qwen2.5-7B, which has weaker reasoning abilities. This results in a large performance drop on the non-misleading dataset. Redrawing the visualization has no large impact on QA accuracy.

Table-based QA results with DePlot and Matcha Table 9 compares the performance of table-based QA using MLLMs with that of two smaller specialized chart-to-table extraction models, DePlot (Liu et al., 2023a) and MatCha (Liu et al., 2023b). On misleading visualizations, the specialized models outperform Ovis1.6-9B and Qwen2VL-7B. Only InternVL2.5-8B matches the performance obtained with DePlot. This indicates that smaller models fine-tuned specifically for table extraction constitute a strong alternative to MLLMs for countering misleading visualizations.

However, in non-misleading cases, table-based QA performs worse with DePlot than with InternVL2.5-8B. Moreover, MatCha achieves the lowest performance among all models. We attribute this gap to the high diversity of chart types in the non-misleading dataset, particularly in the VLAT subset, which includes bubble charts and treemaps. Since DePlot and MatCha were not fine-tuned on such chart types, they lack the generalization capabilities of MLLMs.

Results with Chain-of-Thought prompting By default, we evaluate MLLMs in a direct prompting setting where they only output the final answer. Prompting techniques such as Chain-of-Thought (CoT), which requires the MLLM to generate intermediate reasoning steps before providing the final answer, are known to improve performance on several reasoning tasks (Wei et al., 2022).

Figure 12 shows the change in accuracy when using zero-shot CoT prompting, either alone or in combination with one of the two most effective correction methods, table-based QA, and redrawing the visualization. CoT alone improves performance on both misleading and non-misleading visualiza-

Figure 12: Change (Δ) in accuracy (percentage points) compared to Experiment 1 using Chain-of-Thought (CoT) prompting alone or in combination with correction methods.

tions. The gains are generally modest, except for InternVL2.5-8B, where the improvement exceeds 10 pp on misleading visualizations.

Using CoT has no significant impact on misleading visualizations for table-based QA. However, it substantially reduces the negative effects of table-based QA on non-misleading visualizations, bringing the performance drop close to 0 for Qwen2VL-7B and even bringing a net positive change in accuracy for InternVL2.5-8B.

Combining CoT with redrawn visualizations worsens accuracy on non-misleading visualizations. Its effect on misleading visualizations depends on the model. Ovis1.6-9B shows a small decrease, while InternVL2.5-8B experiences a large gain.

Overall, CoT alone is not an effective correction method for misleading visualizations. It can be beneficial when combined with table-based QA, mainly by mitigating the negative impact on non-misleading data. However, its effects are inconsistent across models and settings, and it should therefore be applied with caution.

Table 10 reports the results of an error analysis of the CoT reasoning chains for a random sample of 30 misleading instances. The majority of the wrong answers and reasoning are due to the presence of the misleader, rather than mathematical mistakes. In a few cases, there is a numerical mistake, but the MLLM answers correctly by rounding to the nearest MCQ choice. Figure 13 shows an instance in which the *inverted* color scale deceives InternVL2.5. Figure 14 is an example of a correct answer with errors in the CoT reasoning. In that case, an incorrect value is extracted from the bar chart. Figure 15 shows an instance where both the reasoning chain and the answer are correct.

	Qwen2VL-7B	Ovis1.6-9B	InternVL2.5-8B
Wrong answer due to the misleader	14	18	15
Wrong answer due to math reasoning	4	4	3
Correct answer, wrong math reasoning	2	2	2
Correct answer and reasoning	10	6	10

Table 10: Error analysis of the CoT reasoning chains, on a random sample of 30 misleading instances (%).

F Experiment 3 - Analysis of table extraction

Error examples Figures 16, 17, and 18 illustrate representative examples of table extraction errors made by InternVL2.5-8B due to misleaders. The examples are taken from CHARTOM, where each data table is associated with two instances, one misleading and one non-misleading. In all three cases, the table extracted from the non-misleading instance is accurate. However, the MLLM fails to recover the same table from the misleading version.

In Figure 16, the MLLM fails to correctly align the dots with the inverted axis. In Figure 17, the 3D effects distort the perceived size of the bars on the right side, leading to incorrect value extraction. In Figure 18, the MLLM first extracts values for both lines using the left axis, and then switches to the right axis for 2022, failing to recognize that each line is associated with a different axis.

Impact of table extraction on table-based QA We conduct a manual analysis on a random sample of 30 misleading visualizations to assess the impact of the table extraction step on the accuracy of table-based QA. We analyze the three MLLMs from Experiment 3 as well as DePlot, introduced in Appendix E. Table 11 provides the results. In most cases, both steps are incorrect. For all models except DePlot, the second-most-frequent outcome is that both steps are correct.

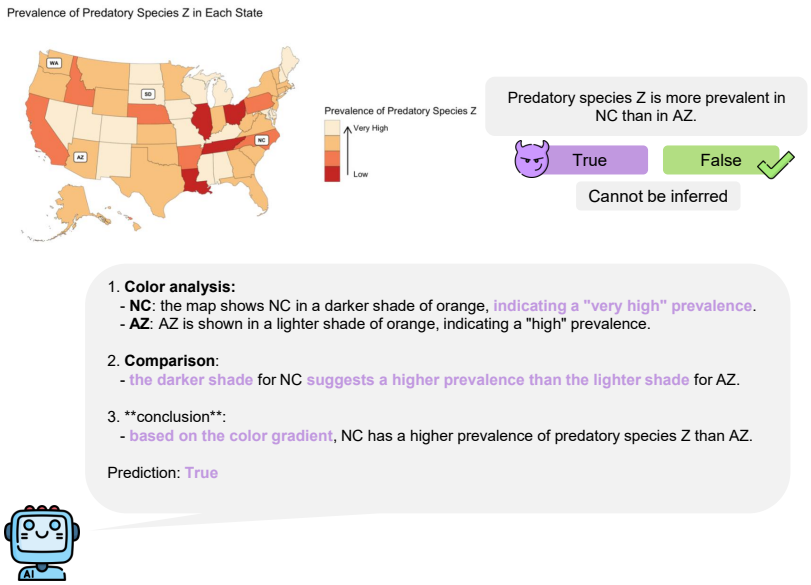

Figure 13: Example of CoT reasoning with InternVL2.5-8B on a misleading visualization, where the MLLM fails to answer correctly due to the misleader.

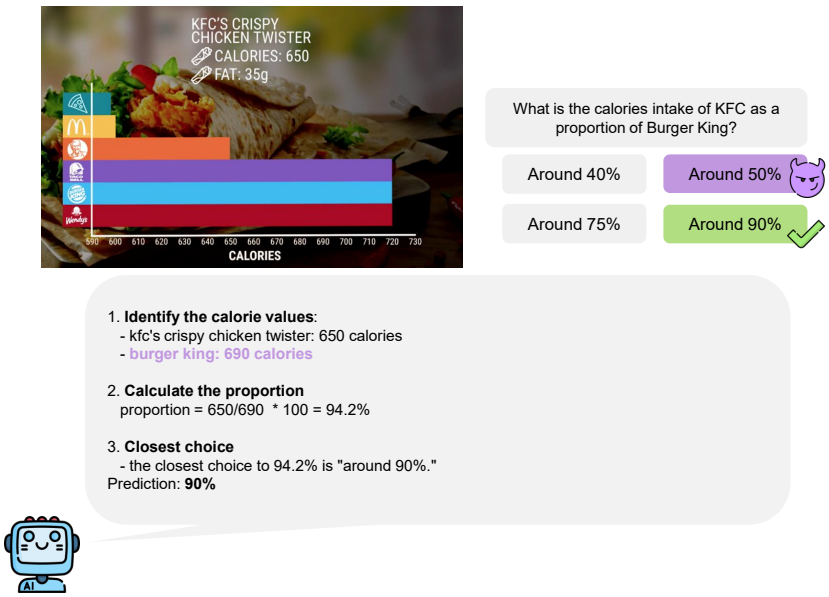

Figure 14: Example of CoT reasoning with InternVL2.5-8B on a misleading visualization, where the MLLM answers correctly but the reasoning contains errors.

In some instances, the extracted table is incorrect, but the QA output is correct. This typically occurs when the correct answer is "not enough information" and is selected not because of correct reasoning but because the extracted table contains missing or incorrect values.

In other cases, the table extraction is correct, but the QA is incorrect. Except for Ovis1.6-9B, this category is the least prevalent. This occurs mostly with misleaders that are not removed when the visualization is converted to a table. For example, if the misleader is *inappropriate item order*, the entries in the table remain shuffled, which is deceiving. If the misleader is *cherry-picking*, there is still not enough data to answer the question properly.

Table extraction accuracy by misleader Table 12 reports our manual analysis of table extraction accuracy with InternVL2.5-8B, the strongest MLLM for table-based QA. We analyze all instances of the ten most frequent misleaders in the misleading visualization dataset. The results show

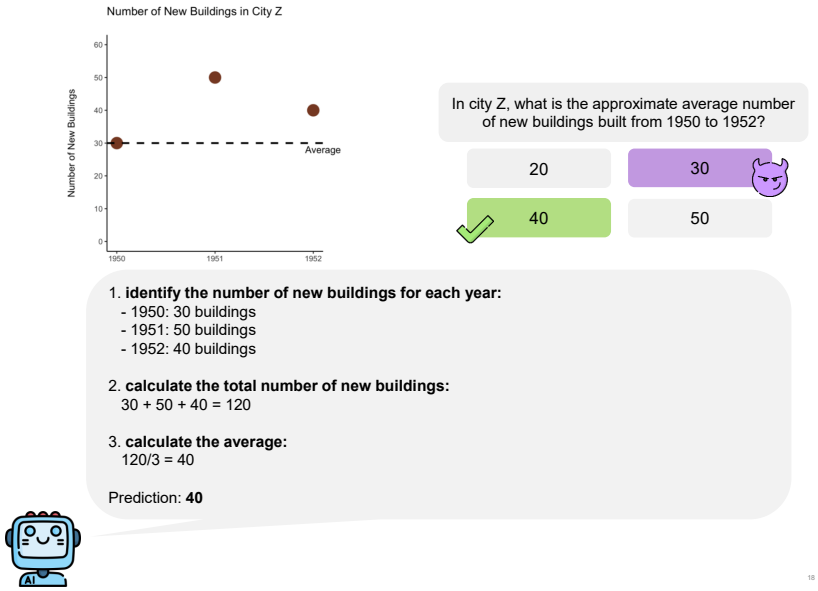

Figure 15: Example of CoT reasoning with InternVL2.5-8B on a misleading visualization, where the MLLM answers correctly with a valid reasoning.

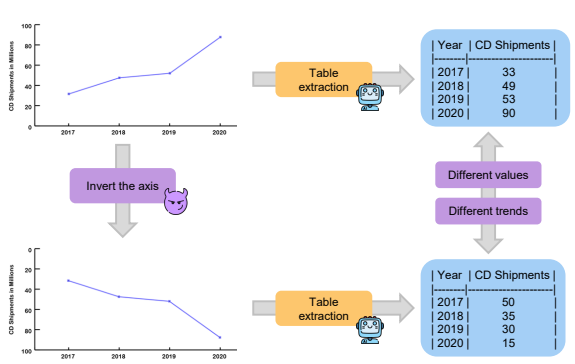

Figure 16: Table extraction for a non-misleading visualization and its misleading version with an inverted axis.

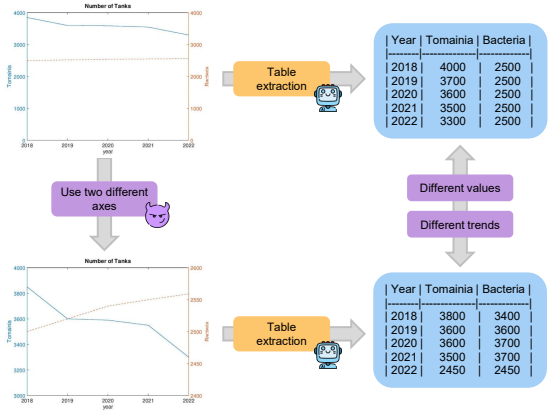

Figure 18: Table extraction for a non-misleading visualization and its misleading version with dual axis.

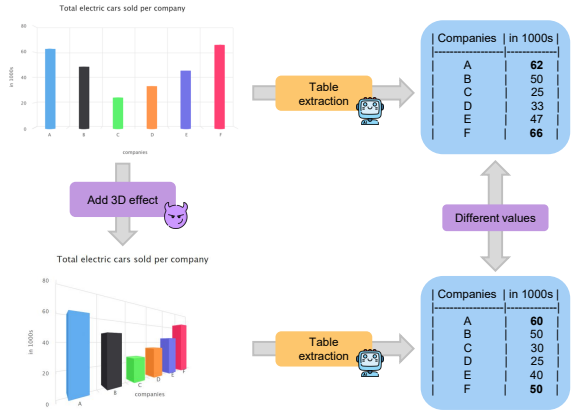

Figure 17: Table extraction for a non-misleading visualization and its misleading version with 3D effects.

Table extraction		Table-based QA (Qwen2.5-7B) Correct	Incorrect
Qwen2VL-7B	Correct	23.3	10.0
	Incorrect	16.7	50.0
Ovis1.6-9B	Correct	26.7	20.0
	Incorrect	20.0	33.3
InternVL2.5-8B	Correct	33.3	20.0
	Incorrect	10	36.7
DePlot	Correct	13.3	10.0
	Incorrect	20.0	56.7

Table 11: Manual analysis of the impact of table extraction accuracy on table-based QA accuracy, on a random sample of 30 misleading visualizations (%).

Misleader	Table extraction accuracy
Inverted axis	15.4
Truncated axis	81.0
Inappropriate axis range	46.7
Inconsistent tick intervals	50.0
3D	8.3
Inappropriate item order	33.3
Inappropriate aggregation	37.5
Dual axis	0.0
Misrepresentation	85.7
Cherry-picking	66.7

Table 12: Manual analysis of table extraction accuracy per misleader category, using Internvl2.5-8B (%).

that extraction accuracy varies across misleaders.

For some misleaders, such as truncated axis and misrepresentation, the extraction accuracy exceeds 80%. The distortions introduced by these misleaders have a limited impact on table extraction. For truncated axis, it is sufficient to align the top of the bars with the vertical axis to recover the values. For misrepresentation, the model often succeeds because the relevant value is explicitly displayed on the bar or pie slice.

In contrast, several misleaders have a severe negative impact on table extraction. When a dual axis is present, all extracted tables are incorrect, as illustrated in Figure 18. Inverted axis and 3D effects also lead to very low extraction accuracy.

Overall, these results show that table-based QA is not a universal correction method for all misleaders. Future work should either improve table extraction for challenging misleaders, such as inverted and dual axis, by creating large-scale synthetic data and fine-tuning specialized models like DePlot, or by designing dedicated correction methods for these misleaders.

Consistency of table extraction on CHARTOM We conducted a manual analysis of the table extraction outputs of all three MLLMs on CHARTOM, which pairs each question with two visualizations, one misleading and one non-misleading, generated from the same underlying data. For all MLLMs, the extracted tables match exactly in only 4 out of 56 pairs (7.1%) and partially in 13 to 14 other pairs (23 to 25%). This further shows the negative impact of misleaders on table extraction accuracy.

	Qwen2VL-7B	Ovis1.6-9B	InternVL2.5-8B
No redrawing for maps	4	4	4
Code did not compile	4	0	2
Incorrect chart design	4	3	2
Correct chart, incorrect values	7	12	8
Correct chart and values	11	11	13

Table 13: Error analysis of the generated codes of the redrawn visualizations, on a random sample of 30 misleading instances (%).

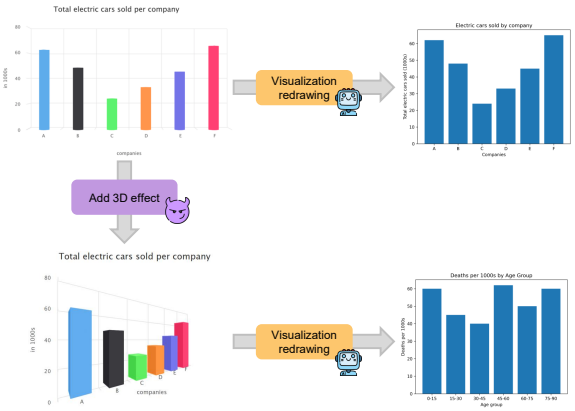

Figure 19: Visualization redrawing for a non-misleading visualization and its misleading version with inverted axis.

G Experiment 3 - Analysis of visualization redrawing

Table 10 shows the results of a manual analysis of the quality of the redrawn visualizations on a random sample of 30 misleading instances. A correct chart with correct values is drawn in more than 30% of the cases. The most prevalent issue is redrawing a correct chart type with incorrect values. This is due to error propagation from the table extraction step. In less frequent cases, Qwen2.5-7B generates code that produces the wrong type of visualization or does not compile at all.

Figures 19, 20, and 21 provide error examples based on pairs of misleading and non-misleading instances of CHARTOM. The examples are the same as those used for table extraction in Figures 16, 17, and 18.

H Experiment 3 - Analysis of axes extraction

We conduct a manual analysis on a random sample of 30 misleading visualizations to assess the impact of the axes extraction step on the accuracy of the visualization+axes correction method. The results are reported in Table 14.

Compared to the table extraction results shown

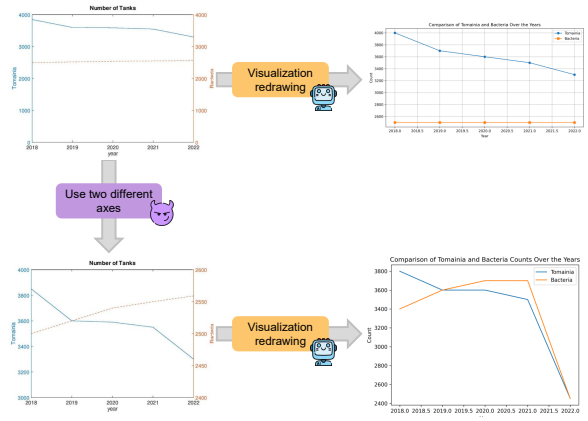

Figure 20: Visualization redrawing for a non-misleading visualization and its misleading version with 3D effects.

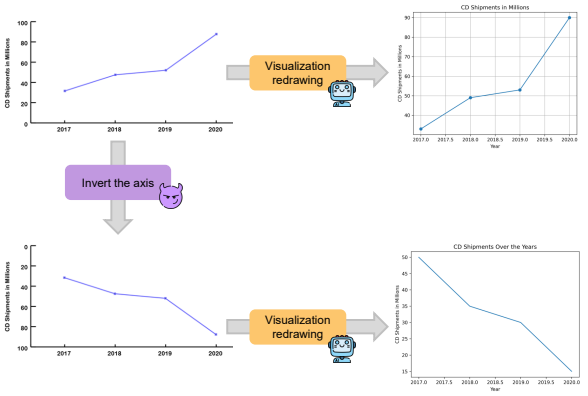

Figure 21: Visualization redrawing for a non-misleading visualization and its misleading version with dual axis.

Table extraction		Visualization + axes QA Correct	Incorrect
Qwen2VL-7B	Correct	16.7	50.0
	Incorrect	13.3	20.0
Ovis1.6-9B	Correct	20.0	56.7
	Incorrect	10.0	13.3
InternVL2.5-8B	Correct	23.3	60.0
	Incorrect	6.7	10.0

Table 14: Manual analysis of the impact of axes extraction accuracy on visualization+axes QA accuracy, on a random sample of 30 misleading visualizations (%).

in Table 11, axes extraction is often accurate, with 20 to 25 out of 30 instances containing correct axis information. However, this high extraction accuracy does not translate into strong QA performance. For all MLLMs, at least half of the instances exhibit correct axes but incorrect QA predictions. Two main factors explain this behavior. First, the MLLMs often appear to ignore the extracted axes information and remain primarily influenced by the image modality. Second, the axes are unrelated to the distortions introduced by several misleaders, such as *3D effects* or *misrepresentation*, which limits their corrective value.

Axes extraction errors occur most frequently for *inconsistent tick intervals*, where the MLLMs tend to hallucinate evenly spaced ticks along the axis.