When Visuals Aren't the Problem: Evaluating Vision-Language Models on Misleading Data Visualizations

Harsh Nishant Lalai * **Raj Sanjay Shah** * **Hanspeter Pfister**
Sashank Varma **Grace Guo**
Birla Institute of Technology and Science, Pilani
Georgia Institute of Technology Harvard University

Abstract

Visualizations help communicate data insights, but deceptive data representations can distort their interpretation and propagate misinformation. While recent Vision Language Models (VLMs) perform well on many chart understanding tasks, their ability to detect misleading visualizations, especially when deception arises from subtle reasoning errors in captions, remains poorly understood. Here, we evaluate VLMs on misleading visualization-caption pairs grounded in a fine-grained taxonomy of reasoning errors (e.g., Cherry-picking, Causal inference) and visualization design errors (e.g., Truncated axis, Dual axis, inappropriate encodings). To this end, we develop a benchmark that combines real-world visualization with human-authored, curated misleading captions designed to elicit specific reasoning and visualization error types, enabling controlled analysis across error categories and modalities of misleadingness. Evaluating many commercial and open-source VLMs, we find that models detect visual design errors substantially more reliably than reasoning-based misinformation, and frequently misclassify non-misleading visualizations as deceptive. Overall, our work fills a gap between coarse detection of misleading content and the attribution of the specific reasoning or visualization errors that give rise to it.

1 Introduction

Visualizations are often used to communicate data-driven insights and to convey complex information effectively. When paired with well-crafted captions, visualizations can improve understanding and decision-making across domains ranging from journalism (Weber and Rall, 2012; Fu and Stasko, 2023) to scientific research (Mogull and Stanfield, 2015; Duarte et al., 2022). However, the same communicative power that makes visualizations impactful makes them susceptible to misrepresentations. Misleading captions and deceptive data representations can distort interpretation, propagate misinformation, and break public trust in data communication (Pandey et al., 2015; Parks and Yeh, 2021; Akhtar et al., 2024).

Defining misleading visualizations:

We adopt the definition provided by Richards (2013) and Pandey et al. (2015) as *"a graphical depiction of information, designed with or without an intent to deceive, that may create a belief about the message and/or its components, which varies from the actual message."* Importantly, under this definition, a mislead does not require malicious intent; even well-intentioned visualizations can mislead through ambiguous framing, selective emphasis, or incorrect interpretation.

Prior work on visualization misinformation has largely focused on flawed visual design, from early notions of graphical integrity (Tufte and Graves-Morris, 1983) to subsequent taxonomies and descriptions of design errors (Pandey et al., 2015; Correll and Heer, 2017; Lo et al., 2022). However, recent studies have shown that misleading real-world visualizations often arise not only from flawed visual encodings (e.g., truncated axes or dual axes), but also from subtle *reasoning errors* in how captions describe or infer meaning from the data (Lisnic et al., 2023; Lan and Liu, 2024). Such errors can appear even when the visualization itself is plausible or professionally produced. Figure 1 shows a real-world example of a chart with both visualization design and reasoning errors. The design includes a misleading dual-axis, while the caption cherry-picks a short-term spike in vaccinations to convey a distorted yet seemingly credible message.

Recent Vision-Language Models show strong performance on many chart understanding and multimodal reasoning tasks (Masry et al., 2022; Islam et al., 2024). This progress raises a natural question: *can VLMs detect misleading visualizations and ac-*

* Equal contribution. Emails: lalaiharsh26@gmail.com, rajsanjayshah@gatech.edu, gguo31@g.harvard.edu. Code and dataset available at GitHub and HuggingFace respectively.

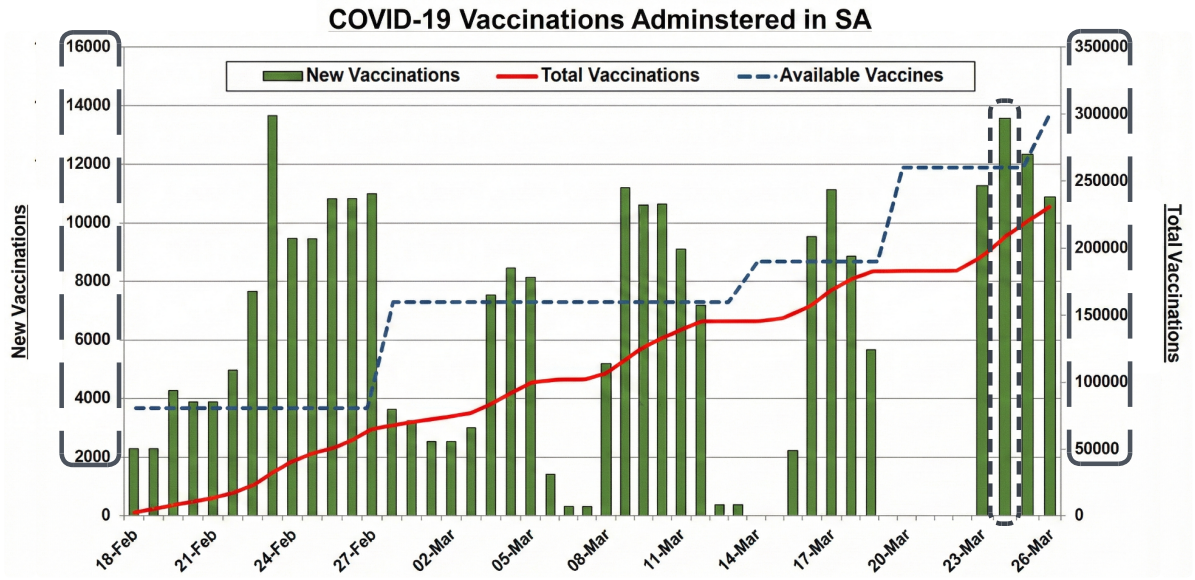


Figure 1: Example of a misleading chart-caption pair with both visual design and reasoning errors. The chart contains a **dual-axis** visualization design error, which may be confusing because viewers must mentally map each axis to its corresponding visual representation (bar or line). The caption also introduces a reasoning error by extrapolating a **cherry-picked** short-term increase to a broader **causal claim**. Together, these factors can distort interpretation without altering the underlying data.

curately attribute them to specific documented reasoning and visualization design error types? While existing benchmarks primarily focus on fact verification or chart-based Q&A, they provide limited insight into how models handle reasoning-based misinformation embedded in visualization-caption pairs. In contrast, we study whether models can attribute misleadingness to specific error types and disentangle whether it arises from the caption, the visualization, or both.

For this, we introduce a benchmark comprising real-world visualizations with human-authored and curated misleading captions designed to elicit specific error types, enabling controlled analysis across error types and modalities (caption, visualization, or both). We assess a range of frontier commercial and open models and provide a diagnostic analysis of where current systems succeed and fail across error types, including their tendency to over-flag non-misleading examples. Lastly, we discuss learnings towards the real-world deployment of such systems.

2 Related Works

2.1 Visualization Design and Misleading Communication

Visualization research has long documented how charts can mislead audiences even without manipulating the underlying data, with many early studies focused on categorizing and describing visual distortion techniques, such as truncated or inverted axes, misleading aspect ratios, inappropriate legends, etc. (Pandey et al., 2015; Correll and Heer, 2017; Lo et al., 2022). These detailed, descriptive error taxonomies enable researchers to better discuss the extent and impact of visualization misinformation (Correll and Heer, 2017). They also help researchers explain why misinterpretation occurs (e.g., the graphical literacy of designers and audiences (Lo et al., 2022; Lan and Liu, 2024)) and develop tools to automatically detect and mitigate the effect of different errors (Chen et al., 2021).

However, recent work has shown that misleading visualizations are not limited to graphical distortions. Lisnic et al. (2023) conducted a large-scale analysis of COVID-19 charts shared on X (formerly Twitter) and found that the majority of misleading cases stemmed not from visual design flaws, but from *reasoning errors in the accompanying captions*. Similar work by Lan and Liu (2024) also found reasoning errors in an online gallery of misleading visualizations curated by the public. Taken together, these findings suggest that *focusing solely on visual distortions significantly underestimates how visualizations are used to misinform in real-world settings.*

Our work builds on and extends these prior stud-

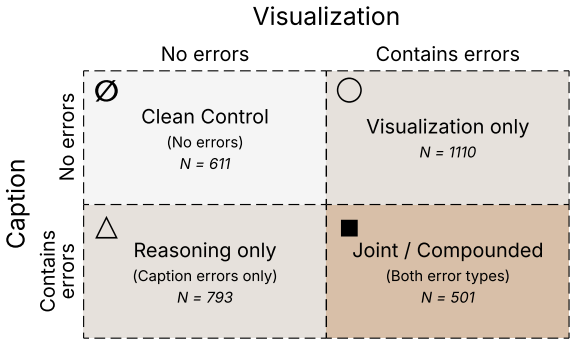


Figure 2: Structure of our dataset organized as a 2×2 grid based on the presence or absence of misleading content in captions and visualizations. Counts denote the number of chart-caption pairs in each cell. Symbols denote error composition: ∅ no errors, △ caption-only errors, ◯ visualization-only errors, ■ joint errors.

ies by examining a combined taxonomy of *visual design and reasoning errors*. Unlike prior findings that describe how humans produce and interpret misleading visualizations, we examine the potential of VLMs as scalable, automated detectors of visual and reasoning errors, grounded in taxonomies derived from human misinformation practices.

2.2 Misinformation Detection Capabilities of VLMs

Most prior work has focused on chart understanding: evaluating whether a VLM can interpret visual elements, extract values, and answer questions about data visualizations (Guo et al., 2024). Early benchmarks such as FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018), PlotQA (Methani et al., 2020), ChartQA (Masry et al., 2022), and LEAF-QA (Chaudhry et al., 2020) primarily assessed a model's ability to interpret and answer questions about charts under the assumption that the charts themselves are truthful. Other works extend this line of work to chart summarization, chart-to-table conversion, and chart-based fact verification (Masry et al., 2023; Islam et al., 2024; Lo, 2024).

More recently, research has focused on multimodal misinformation detection in visualizations, examining visual cues of deception, though typically without reasoning about accompanying text (Alexander et al., 2024; Chen et al., 2025; Wu et al., 2025). Some efforts focus on visual distortions in charts, including detecting design-principle violations (e.g., truncated axes or misleading scales; Tonglet et al. (2025b)) and evaluating VLMs' vulnerability to such distortions using inference-time correction strategies (e.g., table extraction and redrawing; Tonglet et al. (2025a)). Other studies address image-text inconsistencies, like identifying out-of-context image captions in which a real image is paired with an incorrect description (Kalla et al., 2024). However, such approaches are *restricted to visual distortions in charts and do not evaluate whether models can detect reasoning-based deception in which the chart's caption draws false conclusions from the data (see Appendix Table 24).* Motivated by this, we focus on detecting misinformation in visualization-caption pairs that explicitly differentiate visualization design errors from reasoning errors in the caption.

3 Problem Setup and Methodology

We structure our problem along two orthogonal dimensions: whether the visualization is misleading and whether the caption is misleading, producing a 2×2 decomposition that isolates different modes of misinformation (Figure 2).

3.1 Error Taxonomy

We adopt the taxonomy of visualization design errors and caption-level reasoning errors introduced by Lisnic et al. (2023). Table 1 provides the abbreviated descriptions of each error category. The reasoning taxonomy comprises seven caption-level errors: cherry-picking, setting an arbitrary threshold, causal inference, failure to account for statistical nuance, incorrect interpretation of the chart, issues with data validity, and misrepresentation of scientific studies. The visualization taxonomy similarly includes seven error types: truncated axis, dual axis, value encoded as area or volume, inverted axis, uneven binning, unclear encoding, and inappropriate encoding. Each chart-caption pair may contain zero, one, or multiple errors, reflecting the fact that misleading communication often combines several forms of distortion. Detailed definitions, distributions, and examples for all error categories are in Appendix A.2, A.3, and A.4.

3.2 Dataset Construction

To support a controlled analysis of these error sources, we construct a dataset organized to isolate different modes of misleadingness[1]. Samples are drawn from multiple sources of *real-world charts*, such as X or Reddit, and then used to populate the 2×2 grid.

[1]We provide the whole dataset at https://huggingface.co/datasets/MaybeMessi/MisVisBench as an artifact under the CC BY-NC-SA 4.0 license.

Reasoning Errors	
Cherry-picking	Selectively highlighting data subsets that support a claim while ignoring context.
Causal Inference	Claiming causation based solely on correlation or temporal association.
Setting an Arbitrary Threshold	Introducing an unjustified cutoff to frame comparisons as meaningful.
Statistical Nuance	Ignoring uncertainty, baselines, or statistical significance when interpreting data.
Incorrect Reading of the Chart	Misinterpreting trends or values shown in the visualization.
Issues with Data Validity	Questioning data reliability or integrity without substantiated evidence.
Misrepresentation of Studies	Exaggerating or selectively citing scientific findings to support a claim.
Visualization Errors	
Truncated Axis	Manipulating axis ranges to exaggerate visual differences or trends.
Dual Axis	Using multiple axes with unrelated scales to suggest misleading associations.
Values as Area / Volume	Encoding values via area or volume, leading to perceptual distortion.
Inverted Axis	Reversing axis direction in a way that obscures or flips trends.
Uneven Binning	Using non-uniform bins to distort distributions or comparisons.
Unclear Encoding	Using ambiguous or insufficiently labeled visual elements.
Inappropriate Encoding	Applying a chart type unsuitable for the data semantics.

Table 1: Abbreviated definition of the error types used in our dataset. Full descriptions and examples are provided in Appendix Tables 6, 7, 8, and 9.

Populating the 2×2 grid. Each chart-caption pair is assigned to one of four conditions depending on whether the caption and/or the visualization contains misleading content (Figure 2). △ Chart-caption pairs with misleading captions and non-misleading visualizations are drawn from Lisnic et al. (2023) (CC BY 4.0). ◯ For samples with misleading visualizations and non-misleading captions, we combine examples from Lisnic et al. (2023) with additional visualizations obtained from the *r/DataIsUgly subreddit*. ■ For cases where both the visualization and caption are misleading, we reuse charts exhibiting visualization design errors from Lisnic et al. (2023) and author new captions that introduce specific reasoning errors. ∅ Finally, non-misleading chart-caption pairs are collected from the *r/DataIsBeautiful* subreddit and manually verified to ensure that neither visualization design errors nor reasoning errors are present.

Note. For samples collected from *r/DataIsUgly* (◯ condition), the authors manually annotated (see annotation interface in Appendix Figure 7) the visualization error types in the charts. Initially, annotation guidelines were refined through pilot labeling and discussion among the authors to clarify category boundaries and resolve ambiguities. To assess annotation reliability, a subset of 50 samples was independently annotated by multiple authors, yielding a Krippendorff's α of 0.81 across visualization error categories. Following this validation step, the remaining samples were individually annotated using the finalized guidelines. Annotation and verification details for all samples in our dataset, including those manually annotated by the authors as well as those inherited or curated from external sources, are provided in Appendix A.6.

Dataset statistics. A sample in our dataset can contain one or more reasoning and visualization errors. While most samples contain a single error, a considerable subset includes multiple errors. We provide the exact statistics in Appendix Table 10.

3.3 Task Definition

We study whether VLMs can identify and attribute misleadingness in chart-caption pairs by formulating two related multi-label classification tasks: reasoning-error and visualization-error classification. For each sample, the model is provided with the visualization, the accompanying caption (if any), and natural language descriptions of the relevant error categories. The two tasks are evaluated independently to isolate model behavior on caption-level reasoning versus visual design errors. Models are asked to predict the set of applicable error categories and provide a brief justification for each prediction; if no error applies, the model outputs `["None"]`. Details on the prompts, their construction, and ablations are in Appendix A.1, A.10.

3.4 Models Studied

We explore a set of widely used proprietary and open-source vision-language models that report strong performance on existing multimodal benchmarks (Chiang et al., 2024). The models span multiple families and architectural choices, including general-purpose frontier models and a chart-specialized variant (Chhipa, 2025). Table 2 summarizes the models included in our study. All models

Family	Symbol	Model
Gemini (Comanici et al., 2025)	3-P	Gemini-3-Pro-Preview
	2.5-P	Gemini-2.5-Pro
	2.5-F	Gemini-2.5-Flash
GPT (OpenAI, 2025)	5	GPT-5 (25-08-07)
	5-Mini	GPT-5-mini (25-08-07)
Qwen (Bai et al., 2025a,b; Chhipa, 2025)	3	Qwen3-VL-30B-A3B
	2.5	Qwen2.5-VL-7B
	2.5-ChartQA	Qwen2.5-VL-7B-ChartQA

Table 2: Vision-language models studied in our analysis. Models are grouped by family, with symbol identifiers used throughout the paper.

are evaluated using a consistent inference configuration. To account for stochastic decoding, we run each model multiple times and observe consistent performance across runs, indicating that our conclusions are not driven by a single favorable generation. Additional details on inference configuration, retry policies, and run stability are provided in Appendix A.5 and A.9.

3.5 Evaluation Measures

We use multiple evaluation measures to characterize how models identify and attribute misleadingness in chart-caption pairs. As each sample may contain multiple reasoning and visualization errors, we adopt metrics that capture both partial detection and complete attribution. Following prior work (Tonglet et al., 2025b), we report weighted F1, Partial Match, and Exact Match scores, computed separately for reasoning errors, visualization errors, and their combination. The F1 score provides fine-grained per-error performance; PM captures the model's ability to identify at least some of the misinformation present (useful in multi-label settings); and EM sets a strict bar for complete and accurate error detection across modalities.

F1 Score We calculate per-error-type F1 scores for each reasoning and visualization error category. We also compute weighted F1 scores separately for (i) *reasoning error classification*, where the score is the weighted average over the 7 reasoning error categories, and (ii) *visualization error classification*, where the score is the weighted average over the 7 visualization error categories. We also calculate a *combined weighted* F1 score computed as a weighted average over all 14 categories (7 reasoning + 7 visualization). We additionally report macro-averaged F1 scores for reasoning error classification, visualization error classification, and the combined setting in Appendix Table 13.

Metric	3-P	2.5-P	2.5-F	5	5-Mini	3	2.5	2.5-ChartQA
F1	**0.57**	0.59	0.52	0.56	0.53	0.47	0.27	0.24
PM	**0.89**	0.84	0.75	0.86	0.82	0.78	0.81	0.77
EM	**0.23**	0.06	0.02	0.12	0.06	0.03	0.16	0.16

Table 3: The performance of various VLMs on our dataset. We report the combined weighted F1, Partial Match, and Exact Match scores.

Partial Match (PM) is computed at three levels: (i) *reasoning-only*, where a sample is counted as a match if the predicted reasoning error set overlaps with the ground-truth reasoning error set; (ii) *visualization-only*, defined analogously for visualization errors; and (iii) *combined*, where a sample is counted as a partial match with *any subset* of the reasoning or the visualization errors.

Exact Match (EM) is also computed at three levels: (i) *reasoning-only*, where the predicted reasoning error set must exactly equal the ground-truth reasoning error set; (ii) *visualization-only*, defined analogously for visualization errors; and (iii) *combined*, where a sample is counted as an exact match only if the model achieves an exact match on *both* the reasoning and the visualization errors.

4 Findings

Finding 1: VLMs struggle to reliably detect and classify misinformation errors in real-world visualization-caption pairs.

Across all evaluated models, misinformation error detection remains challenging, with no system achieving strong performance on the full benchmark. Even the highest-performing VLMs achieve only mid-range weighted F1 scores (Best model: 3-P, 0.57), indicating limited ability to identify and attribute misinformation in real-world visualization-caption pairs (Table 3). Notably, this limitation persists even for models explicitly fine-tuned for chart understanding: 2.5-ChartQA performs comparably to general-purpose versions of the Qwen family and substantially below other frontier systems. This suggests that relatively strong performance on existing chart-centric benchmarks primarily serves other goals, for example, value extraction and factual question answering, and that these capabilities may not *consistently transfer to reasoning about misleading framing, selective interpretation, or multimodal misinformation.* Overall, the uniformly low F1 scores indicate that detecting and attributing the full spectrum of reasoning and visualization errors remains a difficult task for current state-of-the-art VLMs.

Model	Reasoning Errors			Visualization Errors		
	F1	PM	EM	F1	PM	EM
3-P	0.52	0.66	0.45	0.63	0.69	0.51
2.5-P	0.56	0.58	0.25	0.62	0.58	0.24
2.5-F	0.46	0.40	0.08	0.58	0.56	0.27
5	0.53	0.63	0.31	0.59	0.65	0.39
5-Mini	0.47	0.48	0.15	0.59	0.67	0.38
3	0.46	0.59	0.31	0.47	0.42	0.08
2.5	0.26	0.61	0.47	0.29	0.50	0.38
2.5-ChartQA	0.22	0.56	0.42	0.26	0.47	0.38

Table 4: Performance of the VLMs on reasoning and visualization error classification on the whole dataset. We report each score separately for reasoning and visualization errors. Models consistently achieve higher scores on visualization error detection than reasoning errors, suggesting greater difficulty in identifying and reasoning about misinformation embedded in captions.

This difficulty is also echoed in the distinct gap between Partial Match and Exact Match scores across all models. While PM scores are relatively high, showing that models frequently identify some form of misleadingness, EM scores remain extremely low, showing that models *rarely recover the complete and correct set of errors present in a given example.* This disparity indicates that current VLMs may operate at the level of coarse detection rather than precise attribution: they may often sense that a visualization-caption pair is problematic, but fail to fully enumerate or correctly describe the underlying reasoning and visualization errors. As a result, partial detection can overstate models' actual ability to perform fine-grained multimodal error attribution, as partial overlap may mask missing, spurious, or mislocalized error predictions.

Finding 2: VLMs are systematically better at detecting visual deception than reasoning-based deception, even when the latter is isolated.

All the models achieve lower weighted F1 scores on reasoning-error classification than on visualization-error classification on the whole benchmark (Table 4), *highlighting the greater difficulty of reasoning-based misleads*. Importantly, no model reverses this trend, indicating a systematic asymmetry rather than model-specific variation. For instance, closed-source models (and series) exhibit a 6 to 12 point gap in F1 scores between the two tasks, and this gap becomes smaller for the open-sourced models (3, 2.5, 2.5-ChartQA); however, even these fail to detect caption-based reasoning errors better than visualization errors.

This asymmetry persists under controlled conditions that isolate a single source of misleadingness. When evaluated on samples containing misleading visualizations with non-misleading captions ◯, models achieve substantially higher performance than on samples containing misleading captions paired with otherwise standard visualizations △ (Figure 3). Because these subsets remove cross-modal confounds, the resulting performance gap provides direct evidence that reasoning-based deception, independent of visual distortions, poses a greater challenge for current VLMs. *Notably, this pattern contrasts with a broad body of prior work showing that models often perform better on text-based reasoning than on image-based understanding (Sim et al., 2025; Park et al., 2025). Our results suggest that this advantage does not straightforwardly extend to settings in which textual claims must be evaluated against visual evidence rather than in isolation.*

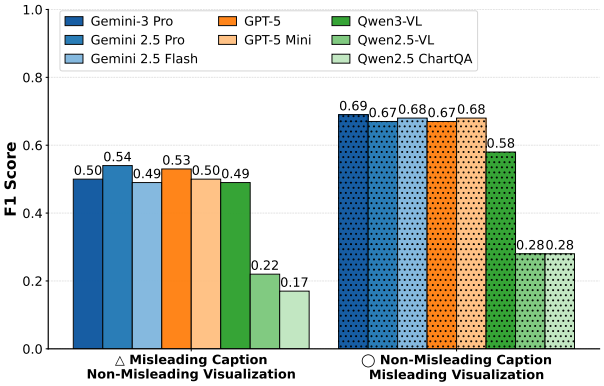


Figure 3: Combined weighted F1 scores for VLMs on benchmark subsets containing only one modality of misinformation.

Finding 3: VLMs succeed on surface-level error patterns but struggle with epistemic and context-dependent reasoning errors.

F1, PM, and EM scores vary substantially across error categories (see Table 5). Errors with salient visual structures or recurring linguistic templates are detected with relatively higher accuracy. In contrast, errors that require epistemic judgment, statistical reasoning, or careful alignment between captions and underlying data remain challenging.

Among reasoning errors, models perform best on categories such as causal inference (F1: 0.59 - 0.71), the use of arbitrary thresholds (F1: 0.50 - 0.56), and cherry-picking (F1: 0.43 - 0.53). These errors often involve *recognizable textual cues*, such as explicit cause-and-effect language, selectively framed time windows, or highlighted cutoffs, that can be easily identified without engagement with

Reasoning Errors							
Model	**Cherry Picking**	**Causal Inference**	**Setting an Arbitrary Threshold**	**Statistical Nuance**	**Incorrect Reading of the Chart**	**Issues with Data Validity**	**Misrepresentation of Studies**
3-P	0.48	0.71	0.51	0.11	0.04	0.26	0.24
2.5-P	0.57	0.67	0.62	0.12	0.10	0.04	0.15
2.5-F	0.43	0.59	0.51	0.10	0.03	0.08	0.28
5	0.53	0.68	0.56	0.13	0.04	0.13	0.22
5-Mini	0.49	0.59	0.50	0.10	0.03	0.08	0.26
3	0.45	0.60	0.49	0.18	0.02	0.05	0.15
2.5	0.46	0.20	0.19	0.12	0.02	0.07	0.04
2.5-ChartQA	0.41	0.17	0.16	0.03	0.03	0.00	0.04
Visualization Errors							
Model	**Truncated Axis**	**Dual Axis**	**Values as Area/ Volume**	**Inverted Axis**	**Uneven Binning**	**Unclear Encoding**	**Inappropriate Encoding**
3-P	0.66	0.89	0.71	0.49	0.12	0.40	0.24
2.5-P	0.66	0.92	0.66	0.44	0.13	0.33	0.16
2.5-F	0.58	0.86	0.68	0.44	0.08	0.34	0.16
5	0.54	0.89	0.73	0.15	0.11	0.36	0.20
5-Mini	0.60	0.89	0.66	0.36	0.16	0.37	0.18
3	0.17	0.81	0.59	0.11	0.12	0.28	0.12
2.5	0.16	0.75	0.14	0.09	0.00	0.14	0.07
2.5-ChartQA	0.12	0.68	0.16	0.04	0.00	0.12	0.07

Table 5: Per-error F1 scores for reasoning and visualization error classification on the full dataset. Models perform relatively well on some visually distinctive errors (e.g., Dual Axis and Values as Area) and some linguistic reasoning errors (e.g., Causal Inference), but struggle on errors requiring statistical interpretation or careful chart reading.

the underlying data-generating process. Similarly, several visualization errors with *visually distinctive patterns*, such as dual axes, truncated axes, and value-as-area encodings, achieve comparatively higher detection performance. These error types share consistent perceptual or structural signatures that appear amenable to pattern-based recognition.

In contrast, VLMs struggle markedly with errors that require contextual or epistemic reasoning. Reasoning categories such as incorrect reading of the chart (F1: 0.02 - 0.06), failure to account for statistical nuance (F1: 0.10 - 0.18), issues with data validity (F1: 0.05 - 0.26), and misrepresentation of scientific studies show *uniformly* low performance. Correct identification often requires aligning textual claims with visual trends and reasoning about omitted baselines, uncertainty, or the plausibility of scientific assertions. Notably, categories such as Failure to Account for Statistical Nuance and Unclear Encoding also exhibit high false positive rates, indicating that models frequently over-predict these context-dependent errors (See Appendix Table 22).

A similar pattern is observed within visualization error detection. While visually salient distortions are often recognized (e.g., dual-axis), more subtle design flaws, such as uneven binning, inappropriate encodings, or inverted axes, remain difficult for most models. Detecting these errors requires precise spatial comparison or knowledge of visualization design principles, which current VLMs do not consistently demonstrate. As a result, even classic visualization pitfalls evade detection when they lack strong visual regularities.

Finding 4: VLMs frequently over-flag non-misleading visualizations-caption pairs (∅) as misleading, indicating a false positive bias.

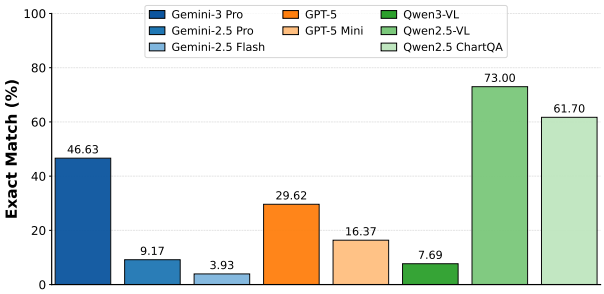


Figure 4: EM scores on the *Non-Misleading Caption, Non-Misleading Visualization (case ∅)* subset of the benchmark. Most VLMs incorrectly flag clean examples as containing at least one error.

In addition to struggling with accurate error attribution, many VLMs frequently misclassify non-misleading visualization-caption pairs as containing one or more errors (refer to Figure 4). On the subset containing no reasoning or visualization errors (case ∅), several models achieve low Exact Match rates, indicating frequent false positives

even when both modalities are clean. This suggests that models often label inputs as misleading even without explicit evidence.

This over-flagging pattern suggests a calibration issue: in the absence of clear error signals, models tend to default toward predicting the presence of an error. In many realistic deployment settings, such as social media platforms, most of the data visualizations are expected to be non-misleading. In such contexts, a tendency to over-flag benign content makes deployment problematic.

Lastly, it is important to note that this tendency is not uniform across models. Some open-source models (InternVL$_{2.5\text{-ChartQA}}$, InternVL$_{2.5}$) achieve higher exact-match accuracy on clean examples, but this improved calibration comes at the cost of lower detection rates on misleading cases. Detailed false positive rates across error categories are reported in Appendix A.12.

5 Discussion

Our findings highlight several key considerations for real-world deployment of VLMs in misinformation detection systems. A likely deployment scenario for such systems is the monitoring of visualization-caption pairs on social media platforms, where charts are frequently used to support claims in public discourse. In these environments, the majority of visualizations are expected to be benign, with misleading cases constituting a relatively small fraction of overall content. Under this base-rate assumption, effective deployment requires not only the ability to flag genuinely misleading visualizations but also the capacity to reliably recognize error-free cases and to provide accurate explanations when intervention occurs.

Effective deployment requires both accurate detection and justification. Prior work in visualization research has emphasized that misleadingness is rarely binary, noting that "the notion that a visualization is either deceptive or not elides the subtlety of many [misinformation] techniques" (Correll and Heer, 2017). Our results align with this perspective: although several VLMs achieve moderately high Partial Match scores, their consistently low Exact Match performance indicates that models often detect misleadingness without correctly attributing underlying reasoning or visualization errors. In deployment settings, such partial detection is insufficient. Flagging content as misleading without accurately identifying *why* it is misleading risks producing incorrect or uninformative justifications, limiting the system's utility for diagnosis, explanation, or downstream moderation.

This limitation is particularly consequential in domains such as health communication, political discourse, and financial reporting, where mischaracterizing the basis of a misleading claim can propagate incorrect conclusions or undermine trust. Accurately distinguishing between different misinformation techniques is not only important for detection, but also for building tools that raise awareness, support human judgment, and mitigate the effects of deceptive framing (Correll and Heer, 2017; Chen et al., 2021). As such, while VLMs show *some* potential for detecting misleading visualizations, their deployment as standalone diagnostic systems would require substantial improvements in attribution accuracy and explanation reliability.

Reasoning-based errors pose a unique challenge. Compared to visual distortions, reasoning errors require contextual understanding and alignment between the caption and the visual evidence. Our results show that models consistently underperform on these categories, despite the community view that models handle text more effectively than visual inputs. Furthermore, visualization researchers have posited that reasoning errors in deceptive visualizations are so persuasive and persistent because they "generally do not contain formal logical fallacies, as the conclusion always logically follows from the presented premises" (Lisnic et al., 2023). In such cases, simple fact-checking or surface-level verification is insufficient: effective detection requires identifying how claims are derived from the data, not merely whether the data itself is accurate.

Over-flagging further complicates deployment. Beyond missed or incomplete detections, our results show that many VLMs frequently misclassify non-misleading visualization-caption pairs as deceptive. This false-positive bias suggests a calibration issue in which models default toward flagging content under uncertainty, rather than reliably recognizing error-free cases. In deployment settings dominated by benign visualizations, this tendency can substantially reduce practical utility by massively flagging samples for human review.

In the current state of VLMs, we recommend against deploying current VLMs as standalone systems for detecting or moderating misleading data visualizations, and instead limit their use to carefully designed human-in-the-loop settings until sub-

stantial improvements are achieved in error attribution, reasoning robustness, and calibration.

6 Conclusion

In this work, we evaluated whether current vision-language models can detect misinformation in visualization-caption pairs by jointly modeling visual design errors and caption-level reasoning errors. Unlike prior benchmarks that largely assume truthful visualizations, our benchmark explicitly includes reasoning errors in captions, which is an important but often neglected dimension of real-world misinformation. Across a diverse set of models, we find that while VLMs are relatively effective at identifying certain perceptual distortions, they struggle substantially with subtle design guideline violations and reasoning errors that require contextual interpretation, statistical nuance, or the evaluation of claims against the visual evidence. These results suggest that strong performance on chart understanding benchmarks does not directly translate into robust multimodal misinformation detection. Our benchmark and findings highlight concrete limitations of current VLMs and provide a foundation for developing models and evaluations that better emulate how misinformation manifests in real-world visual communication.

Limitations

Limitations to our work are as follows: (1) No task-specific fine-tuning. We evaluate models using their default inference configurations and do not explore whether task-specific fine-tuning or instruction tuning on our benchmark could improve performance. While this choice reflects realistic out-of-the-box deployment scenarios, fine-tuning may meaningfully alter both detection accuracy and calibration behavior. (2) Model coverage is not exhaustive. Although we evaluate a diverse set of proprietary and open-source VLMs based on our resource constraints, our open-source analysis primarily focuses on the Qwen family. Future work could extend this evaluation to a broader range of community models to better assess generalizability across architectures and training paradigms. (3) Static visualizations only. Our benchmark focuses on static chart-caption pairs and excludes interactive or animated visualizations, which are increasingly common in online settings. Detecting misleadingness in such formats may pose additional challenges not captured here. (4) No external knowledge or verification tools. Unlike current agentic systems, models are evaluated without access to external sources, such as fact-checking databases or domain-specific knowledge bases. As a result, performance on reasoning errors involving data validity or scientific misrepresentation may underestimate what could be achieved with retrieval-augmented or tool-assisted systems. (5) No downstream user impact analysis. Our evaluation focuses on model performance and error attribution accuracy, and does not examine how model outputs influence human judgment, trust, or decision-making. Understanding how partial detections or incorrect explanations affect users is an important direction for future work. (6) Inherited dataset characteristics. A subset of our benchmark reuses samples from Lisnic et al. (2023). As with any such dataset, this portion inherits the characteristics and potential limitations of the original resource, while the remainder of our benchmark is constructed from additional sources.

Ethical considerations

This work relies on a benchmark constructed from publicly available visualization-caption pairs sourced from online platforms such as X (formerly Twitter) and Reddit.

Data sourcing and platform compliance. All visualizations in the benchmark are derived from publicly available posts on X and Reddit. Note: Lisnic et al. (2023) provided curated sets of visualizations from *X* in the form of tweet ID and their own annotations. To adhere to platform policies and content-sharing requirements, we do not redistribute raw social media content. Instead, we release only tweet IDs and Reddit post IDs, allowing data to be extracted (rehydrated) in accordance with the respective platforms' terms of service. We do not include private, deleted, or access-restricted content, nor do we collect or infer personally identifiable or sensitive user information.

Annotation and caption construction. To enable controlled analysis of misinformation mechanisms, some captions are human-authored (by the project team) or curated to introduce specific reasoning errors grounded in prior visualization research. These captions are intended to model common misleading practices rather than to endorse the claims they express. Dataset documentation distinguishes between original and researcher-authored content.

Risks of misuse and over-interpretation. Because the benchmark labels specific reasoning and visualization errors, models trained or evaluated on it could be misused for unsupervised moderation or for generating misleading content.

Overall, our dataset is intended to support diagnostic evaluation and responsible research on multimodal misinformation detection, and should be used with an awareness of its scope and limitations.

References

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. Chartcheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937.

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. IEEE.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3512–3521.

Qing Chen, Fuling Sun, Xinyue Xu, Zui Chen, Jiazhe Wang, and Nan Cao. 2021. Vizlinter: A linter and fixer framework for data visualization. *IEEE transactions on visualization and computer graphics*, 28(1):206–216.

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.

Prakash Chandra Chhipa. 2025. Askanythingincharts-qwen2.5-7b: Fine-tuned qwen2.5-vl for chart understanding.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Michael Correll and Jeffrey Heer. 2017. Black hat visualization. In *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), IEEE VIS*, volume 1, page 10.

Ana Duarte, Miguel Carvalhais, and Pedro Amado. 2022. The role of data visualization in science communication: Principles, encoding, and design patterns. In *International Conference on Design and Digital Communication*, pages 753–764. Springer.

Yu Fu and John Stasko. 2023. More than data stories: Broadening the role of visualization in contemporary journalism. *IEEE Transactions on Visualization and Computer Graphics*.

Grace Guo, Jenna Jiayi Kang, Raj Sanjay Shah, Hanspeter Pfister, and Sashank Varma. 2024. Understanding graphical perception in data visualization through zero-shot prompting of vision-language models. *arXiv preprint arXiv:2411.00257*.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749, Bangkok, Thailand. Association for Computational Linguistics.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning? an extensive investigation into the capabilities and limitations of lvlms. *arXiv preprint arXiv:2406.00257*.

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

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

Jayateja Kalla, Soma Biswas, and 1 others. 2024. Covlm: Leveraging consensus from vision-language models for semi-supervised multimodal fake news detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 1197–1214.

Xingyu Lan and Yu Liu. 2024. "i came across a junk": Understanding design flaws of data visualization from the public's perspective. *IEEE Transactions on Visualization and Computer Graphics*.

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*, pages 1–21.

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. 2024. How good (or bad) are llms at detecting misleading visualizations? *IEEE Transactions on Visualization and Computer Graphics*.

Yu Ho Lo. 2024. *On Understanding Misleading Visualizations, Automatic Detection, and Prevention*. Hong Kong University of Science and Technology (Hong Kong).

Ridwan Mahbub, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, and Enamul Hoque. 2025. The perils of chart deception: How misleading visualizations affect vision-language models. *arXiv preprint arXiv:2508.09716*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

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

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Scott A Mogull and Candice T Stanfield. 2015. Current use of visuals in scientific communication. In *2015 IEEE international professional communication conference (IPCC)*, pages 1–6. IEEE.

OpenAI. 2025. Gpt-5. Large language model.

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*, pages 1469–1478.

Simon Park, Abhishek Panigrahi, Yun Cheng, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2025. Generalizing from simple to hard visual reasoning: Can we mitigate modality imbalance in vlms? *arXiv preprint arXiv:2501.02669*.

Jonathan Parks and D Dante Yeh. 2021. How to lie with statistics and figures. *Surgical infections*, 22(6):611–619.

Jef Richards. 2013. *Deceptive advertising: Behavioral study of a legal concept*. Routledge.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470.

Minjun Son and Sungjin Lee. 2025. Advancing multimodal large language models: optimizing prompt engineering strategies for enhanced performance. *Applied Sciences*, 15(7):3992.

Jonathan Tonglet, Tinne Tuytelaars, Marie-Francine Moens, and Iryna Gurevych. 2025a. Protecting multimodal large language models against misleading visualizations. *arXiv preprint arXiv:2502.20503*.

Jonathan Tonglet, Jan Zimny, Tinne Tuytelaars, and Iryna Gurevych. 2025b. Is this chart lying to me? automating the detection of misleading visualizations. *arXiv preprint arXiv:2508.21675*.

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

Wibke Weber and Hannes Rall. 2012. Data visualization in online journalism and its implications for the production process. In *2012 16th International Conference on Information Visualisation*, pages 349–356. IEEE.

Jiaying Wu, Fanxiao Li, Zihang Fu, Min-Yen Kan, and Bryan Hooi. 2025. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models. *arXiv preprint arXiv:2505.15489*.

A Appendices

A.1 Prompts used in the paper

The prompts used for our tasks were iteratively refined in consultation with visualization experts on the author team. These experts reviewed early versions of the prompt and provided feedback on clarity and specificity. In each round, experts reviewed the task instructions, the definitions of reasoning and visualization error categories, and representative examples. Their review protocol focused on three aspects. First, they validated the error definitions to ensure that categories were theoretically grounded. Second, they improved instruction clarity by revising ambiguous phrasing. Third, they tested the prompts on representative sample cases across all conditions and reviewed model outputs to identify potential confusions. Based on this feedback, we arrived at the final versions listed below.

Reasoning Error Classification Prompt

You will be provided with a visualization, its accompanying caption, and descriptions of reasoning errors. These reasoning errors represent ways in which people use captions to spread misinformation.

Your task is to carefully examine the image and its accompanying caption. Then, based on the information and the descriptions of reasoning errors, determine which kinds of misinformation, if any, are being propagated.

If none of the reasoning errors apply, classify the reasoning error as "None."

Please classify which reasoning errors are present and explain your reasoning. If more than one classification applies, include all applicable classifications in a list. Even if only one classification applies, the "classification" field must still be a list.

Only provide output in the following JSON format:

{

"reason": "[Explanation]",

"classification": ["Cherry-picking/Causal inference/Setting an arbitrary threshold/Failure to account for statistical nuance/Incorrect reading of chart/Issues with data validity/Misrepresentation of scientific studies/None"]

}

Image: {image}

Accompanying Text: {caption}

Error Descriptions: {reasoning_error_descriptions}

Visualization Error Classification Prompt

You will be provided with a visualization, its accompanying caption, and descriptions of the visualization error. These visualization errors represent ways in which people use visualization to spread misinformation.

Your task is to carefully examine the image and its accompanying caption. Then, based on the information and the descriptions of visualization errors, determine which kinds of misinformation, if any, are being propagated here.

If none of the visualization errors apply, you may classify the visualization error as "None."

Please classify which visualization errors are present and explain your reasoning. If more than one classification applies, include all applicable classifications in a list. Even if there is only one classification, the "classification" field must still be a list.

Only provide output in the following JSON format:

{

"reason": "[Explanation]",

"classification": ["Truncated axis/Dual axis/Value as area or volume/Inverted axis/Uneven binning/Unclear encoding/Inappropriate encoding/None"]

}

Image: {image}

Accompanying Text: {caption}

Error Descriptions: {visualizaton_error_descriptions}

A.2 Error Descriptions

We present detailed descriptions of the reasoning (Table 6) and visualization errors (Table 7) used in our benchmark. The descriptions are adapted from Lisnic et al. (2023), and further refined through feedback from visualization graduate students.

Error Type	Description
Setting an Arbitrary Threshold *Shorthand: Arb. Thr.*	Setting a benchmark or threshold that lacks a solid factual basis or recognition by standard authorities. The arbitrary threshold is used to judge or compare data, leading to potentially misleading conclusions because it appears meaningful but isn't supported by official criteria. Key Characteristics: - Unjustified Benchmarks: The threshold is chosen without logical reasoning, often to support a specific argument. - Selective Highlighting: Data aligning with the threshold is emphasized, ignoring the broader context. - Visual Manipulation: Annotations or labels make the threshold seem more significant than it is. - Lack of Context: The threshold is presented without proper context or comparison to recognized standards.
Cherry-picking *Shorthand: Cher. Pick.*	Selectively presenting data points that support a specific argument while ignoring those that don't. This can create a biased and misleading representation of the data. Key Characteristics: - Selective Data Points: Only data that supports the argument is shown. - Ignoring Context: Broader data context that might contradict the argument is omitted. - Overemphasis: Overemphasizing certain data points to sway opinion.
Causal Inference *Shorthand: Caus. Infer.*	Assuming a cause-and-effect relationship between two variables based on their correlation, without sufficient evidence to support such a claim. Key Characteristics: - Correlation Assumed as Causation: Assuming that because two variables are correlated, one must cause the other. - Lack of Evidence: No rigorous evidence to support the causal link. - Ignoring Other Factors: Failing to consider other variables that might influence the outcome.
Issues with Data Validity *Shorthand: Data Val.*	Questioning the accuracy or reliability of the data without sufficient justification, often to cast doubt on the conclusions drawn from the data. Key Characteristics: - Questioning Data Accuracy: Raising doubts about the data without solid evidence. - Suggesting Manipulation: Implying that data has been manipulated to fit a narrative. - Ignoring Explanations: Overlooking valid reasons for data inconsistencies.
Failure to Account for Statistical Nuance *Shorthand: Stat. Nu.*	Oversimplifying complex statistical data, and ignoring important details that are crucial for accurate interpretation. Key Characteristics: - Oversimplification: Ignoring complex statistical relationships and nuances. - Lack of Comparison: Failing to compare with relevant control groups or baselines. - Misinterpretation: Drawing conclusions without considering statistical significance or variability.
Misrepresentation of Scientific Studies *Shorthand: Mis. Sci.*	Selectively citing studies or exaggerating their findings to support a specific argument, often ignoring the broader scientific consensus. Key Characteristics: - Selective Citation: Citing only studies that support the argument. - Exaggeration: Overstating the significance or certainty of study findings. - Ignoring Consensus: Overlooking the broader context or scientific consensus.
Incorrect Reading of the Chart *Shorthand: Chart Read*	Misinterpreting the data presented in a chart, often due to visual distortions or a lack of understanding of the chart's design. Key Characteristics: - Visual Distortion: Misinterpreting data due to design issues like truncated axes or misleading scales. - Misreading Data: Incorrectly interpreting the data points or trends shown in the chart. - Lack of Understanding: Failing to understand the chart's design or the data it represents.

Table 6: Descriptions of reasoning errors in captions that contribute to misinformation. The shorthand names are used in subsequent tables for compact reporting.

Error Type	Description
Truncated Axis *Shorthand: Trunc. Axis*	Shortening the axis scale in a chart to exaggerate the appearance of differences or trends in the data. Key Characteristics: - Exaggerated Trends: Small differences in data appear more significant due to axis truncation. - Misleading Scales: The axis starts at a value other than zero without clear justification. - Distorted Proportions: Viewers perceive larger changes than actually exist.
Dual Axis *Shorthand: Dual Axis*	Using two separate vertical axes to plot unrelated or loosely related variables, often creating misleading visual correlations. Key Characteristics: - Misaligned Scales: The scales of the two axes are unrelated, leading to false visual patterns. - Forced Correlation: Unrelated datasets appear correlated due to shared chart space. - Overloading Information: Multiple axes make the chart harder to interpret accurately.
Value as Area or Volume *Shorthand: Area/Vol.*	Using shapes or 3D volumes to represent data although it is known that humans are poor at visually distinguishing differences in area or volume. Key Characteristics: - Exaggerated Perception: Changes in size appear larger than the actual proportional difference. - Misleading Scaling: Areas or volumes do not accurately reflect the data values. - Ineffective Comparison: Viewers struggle to interpret exact values or differences.
Inverted Axis *Shorthand: Inv. Axis*	Reversing the direction of an axis, which can confuse the audience and lead to misinterpretation of trends or comparisons. Key Characteristics: - Reversed Direction: An axis increases in value downward or to the left instead of the standard upward or rightward directions. - Misleading Trends: Data trends appear opposite to their actual direction. - Lack of Clarity: The inversion is not clearly labeled or explained.
Uneven Binning *Shorthand: Uneven Bin.*	Grouping data into bins of unequal size or creating bins that do not span the data distribution, leading to biased or misleading visual distributions. Key Characteristics: - Inconsistent Intervals: Bin sizes vary without justification, skewing the data representation. - Disproportionate Emphasis: Certain bins appear more significant due to size differences. - Misleading Comparisons: Data is harder to compare accurately across bins.
Unclear Encoding *Shorthand: Unclr. Enc.*	Using visual elements that are difficult to interpret or lack sufficient labeling, leading to confusion about what the chart represents. Key Characteristics: - Ambiguous Visuals: Symbols, colors, or patterns are non-standard or not clearly explained. - Missing Labels: Key elements like axes, legends, or annotations are absent or unclear. - Overloaded Design: Too many visual elements representing multiple data variables in a single chart, making interpretation difficult.
Inappropriate Encoding *Shorthand: Inappr. Enc.*	Choosing a visual representation that is unsuitable for the type of data, making interpretation inaccurate or misleading. Key Characteristics: - Misaligned Visuals: The chosen chart type or visual encoding does not match the data variable. - Distorted Representation: Data relationships are inaccurately emphasized or diminished, resulting in ineffective comparisons.

Table 7: Descriptions of visualization errors that mislead viewers through visualization design. The shorthand names are used in subsequent tables for compact reporting.

A.3 Examples from the Benchmark

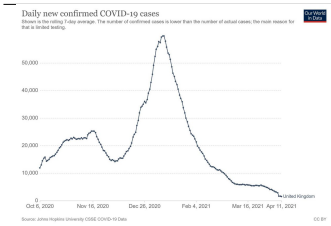
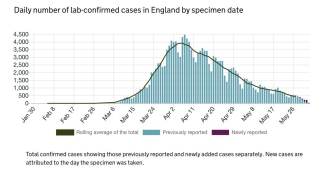
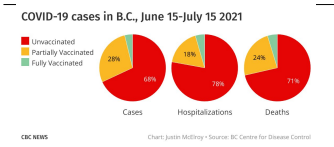
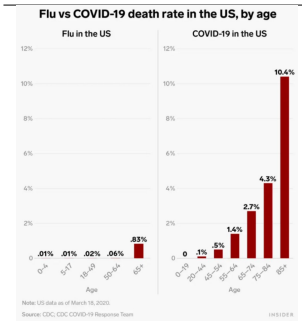
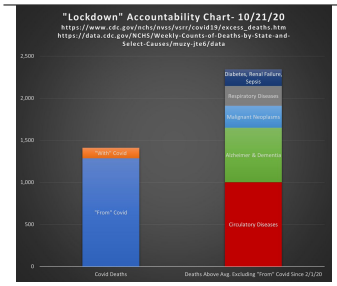
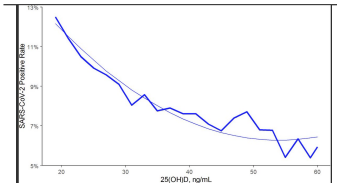
Visualization	Caption	Reasoning Error
File:1918 spanish flu waves.gif 	Reminder: Just because we've hit a peak does not mean we've hit THE peak.	**Cherry-picking**
	The positive impact of the UK's vaccination efforts in one graph	**Causal inference**
	This in a country of 56 million. Lift lockdown now, the virus is just gone.	**Setting an arbitrary threshold**
	The numbers absolutely speak for themselves. Get vaccinated!	**Failure to account for statistical nuance**
	The flu is 10 times less deadly - particularly for elderly - than Covid!	**Incorrect reading of chart**
	This is a test of our humanity	**Issues with data validity**
	SARS-CoV-2 positivity rates associated with circulating 25-hydroxyvitamin D levels (https://tinyurl.com/5n9xm536)	**Misrepresentation of scientific studies**

Table 8: Examples of misleading captions and their associated errors paired with visualizations.

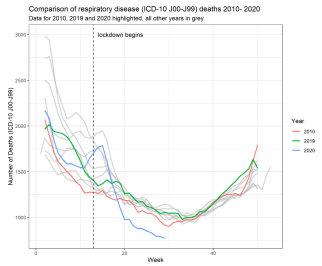
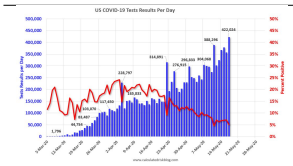
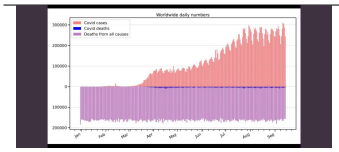
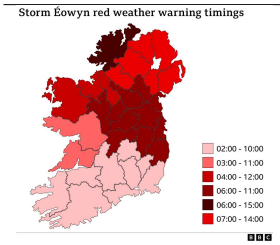
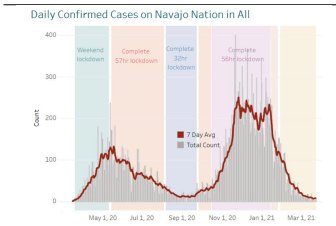
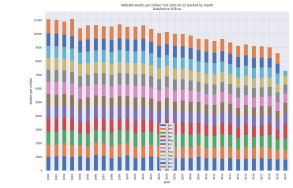
Visualization	Caption	Reasoning Error
	Respiratory deaths at 10 year low!	**Truncated axis**
	May 17 Update: US COVID-19 Test Results: Test-and-Trace Success for Smallpox	**Dual Axis**
	Corona Virus Interactive Map.	**Value as area or volume**
	Propaganda: RECORD NUMBER OF COVID POSITIVE CASES. Reality:	**Inverted axis**
	Interesting colour coding from the BBC	**Uneven binning**
	The Navajo Nation crushed the Covid curve. Success is possible.	**Unclear encoding**
	The worst pandemic of the most contagious disease we have seen for 100 years.	**Inappropriate encoding**

Table 9: Examples of misleading visualizations and their associated errors paired with captions.

A.4 Reasoning and Error Compositions across the Benchmark

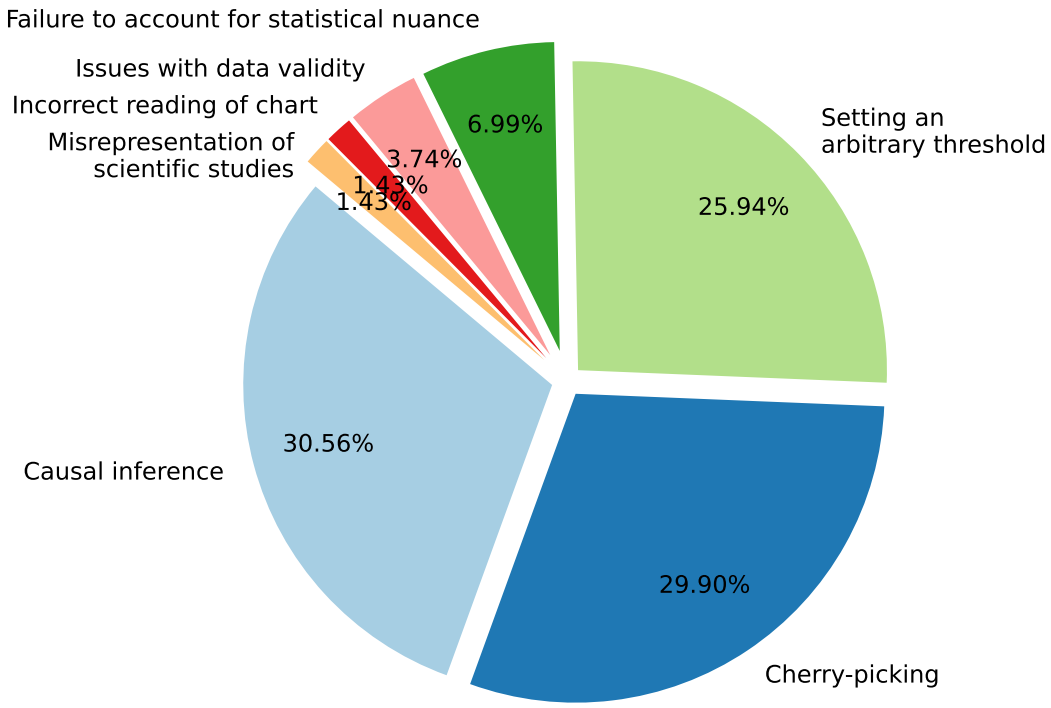


Figure 5: Reasoning Error Composition for the dataset.

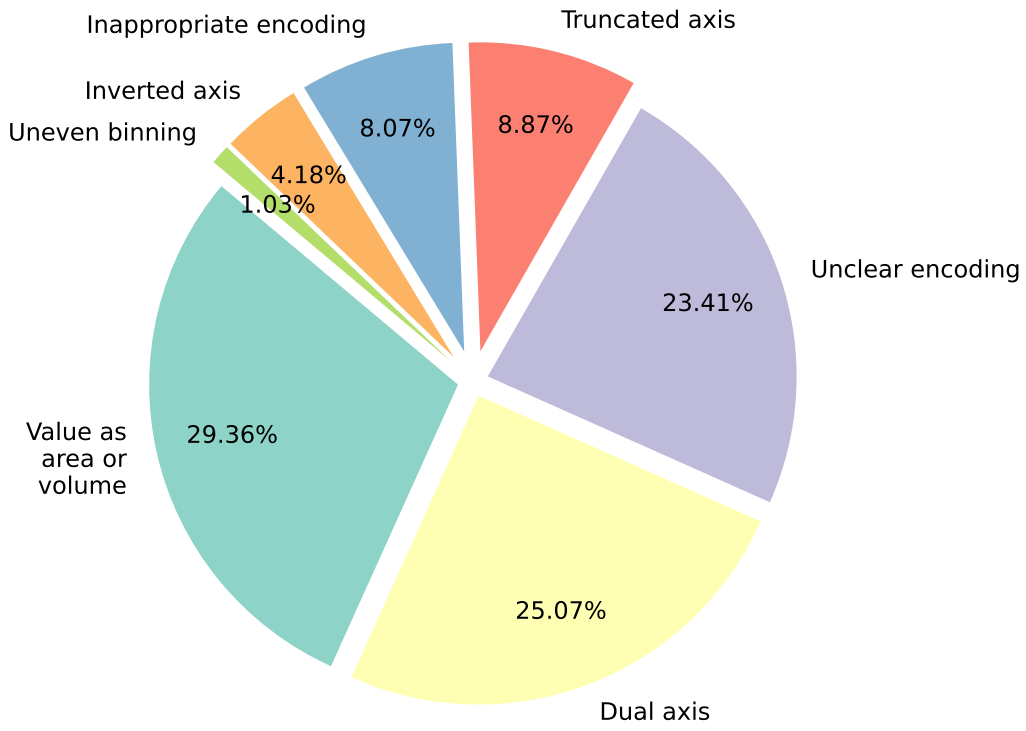


Figure 6: Visualization Error Composition for the dataset.

A.5 Inference Configuration, Retry Policy, Total Number of Runs

We report the inference configurations used across all experiments for reproducibility and clarity.

Temperature. For all models, we use the default temperature settings.

Maximum Tokens. We set the maximum token limit to 10000 tokens for all models. This budget includes both visible output tokens and internal reasoning tokens (for applicable models; GPT family and Gemini family), ensuring that generations are not truncated during multi-label classification and justification generation.

Retry Policy and Malformed Output Handling. For each classification call, we allow up to 5 retry attempts if the model output does not conform to the required JSON schema or if the predicted labels do not exactly match the predefined set of valid error categories provided in the prompt. If all retry attempts fail, the sample is excluded from metric computation. Across all experiments, 3.7% of samples were removed from evaluation due to invalid outputs.

Number of Runs. Our analysis includes eight models evaluated on a dataset of 3,015 samples. For each sample, we make two independent model calls: one for reasoning error classification and one for visualization error classification. This results in a total of 48,240 ($8 \times 3015 \times 2$) model inference calls. *Given typical academic resource constraints, this evaluation scale necessitates prioritization of experiments over exhaustive ablations.*

A.6 Annotation and Verification

Subset of our dataset taken from Lisnic et al. (2023). We reuse a subset of chart-caption pairs curated by Lisnic et al. (2023). In their work, the authors collected tweets containing visualizations and manually filtered them to remove non-visualizations and unrelated content. They developed a codebook through an open-coding process and iteratively refined it through independent annotation and discussion among the authors, resulting in a taxonomy of visualization design violations and caption-level reasoning errors. The finalized codebook was then applied to annotate the full dataset. To verify these inherited annotations, we randomly sampled 50 instances and had two authors independently review the presence of the reported visualization and reasoning errors. Only two disagreements were observed and were resolved through discussion, confirming consistency with the original annotations.

r/DataIsBeautiful. We randomly sampled 100 instances out of the 611 posts from *r/DataIsBeautiful* used in our dataset for verification. Two authors independently reviewed these samples to verify the absence of visualization and reasoning errors. Disagreements were observed in 6 cases and were resolved through discussion to ensure a consistent interpretation of the original labels.

Misleading Caption, Misleading Viz. For cases where both the visualization and caption are misleading (*N=501*), we reuse charts exhibiting visualization design errors from Lisnic et al. (2023) and author new captions that introduce specific reasoning errors. Three authors independently wrote the new captions and annotated the reasoning errors present in the captions. To assess consistency, we randomly sampled 50 such instances and had the same three authors independently review the reasoning-error annotations. Disagreements were observed, yielding a Krippendorff's α of 0.84. These disagreements were subsequently discussed jointly and resolved through group discussion.

A.7 Dataset statistics

# Errors (x)	Reasoning only	Visualization only	Reasoning + Visualization (y, z)					
			(1,1)	(1,2)	(2,1)	(2,2)	(3,1)	(4,1)
1	476	993	–	–	–	–	–	–
2	292	115	344	–	–	–	–	–
3	24	2	–	14	106	–	–	–
4	1	0	–	–	–	3	32	–
5	0	0	–	–	–	–	–	2

Table 10: Distribution of samples by total number of errors (x). **Reasoning-only** and **Visualization-only** denote samples containing exclusively reasoning and visualization errors respectively. **Reasoning + Visualization** reports compositional error cases with (y, z) indicating the number of reasoning and visualization errors, respectively, where $y + z = x$.

A.8 Image of our Annotation Interface

Figure 7 shows a screenshot of our annotation interface used by annotators to label visualization and reasoning errors in charts and captions.

A.9 Stability Across Multiple Runs

Since VLMs use stochastic decoding during inference, model outputs can vary slightly across runs

even when evaluated on the same dataset. To measure the stability of our results, we run ✦$_{2.5\text{-P}}$[2] three times with identical settings and report the mean and standard deviation of the evaluation metrics (F1, Partial Match, and Exact Match). Table 11 summarizes the results across runs.

Overall, we observe that the standard deviations are relatively small across the reported metrics, indicating that the model's performance is stable across repeated evaluations. This suggests that the results reported in the main paper are not driven by a single favorable run but instead reflect consistent behavior of the model under stochastic decoding.

Reasoning Error Classification Prompt

You will be provided with a visualization, its accompanying caption, and descriptions of reasoning errors. These reasoning errors represent ways in which people use captions to spread misinformation.

Your task is to carefully examine the image and its accompanying caption. Then, based on the information and the descriptions of reasoning errors, you need to identify which of the following error categories, if any, apply to this chart-caption pair.

If none of the reasoning errors apply, classify the reasoning error as "None."

Please classify which reasoning errors are present and explain your reasoning. If more than one classification applies, include all applicable classifications in a list. Even if only one classification applies, the "classification" field must still be a list.

Only provide output in the following JSON format:

{

"reason": "[Explanation]",

"classification": ["Cherry-picking/Causal inference/Setting an arbitrary threshold/Failure to account for statistical nuance/Incorrect reading of chart/Issues with data validity/Misrepresentation of scientific studies/None"]

}

Image: {image}

Accompanying Text: {caption}

Error Descriptions: {reasoning_error_descriptions}

[2]Note: ✦$_{3\text{-P}}$ will be deprecated as of March 26, 2026.

	F1	PM	EM
Combined	0.59 (± 0.04)	0.84 (± 0.03)	0.06 (± 0.01)
Reasoning Errors	0.56 (± 0.05)	0.58 (± 0.02)	0.25 (± 0.01)
Visualization Errors	0.62 (± 0.04)	0.58 (± 0.05)	0.24 (± 0.03)
Cherry Picking	0.57 (± 0.08)		
Causal Inference	0.67 (± 0.01)		
Arbitrary Threshold	0.62 (± 0.06)		
Failure to Account for Statistical Nuances	0.12 (± 0.00)	**N/A**	**N/A**
Incorrect Reading of the Chart	0.10 (± 0.03)		
Data Validity Issues	0.04 (± 0.04)		
Misrepresentation of Scientific Studies	0.15 (± 0.10)		
Truncated Axis	0.66 (± 0.04)		
Dual Axis	0.92 (± 0.04)		
Value as Area/Vol.	0.66 (± 0.04)		
Inverted Axis	0.44 (± 0.01)	**N/A**	**N/A**
Uneven Binning	0.13 (± 0.03)		
Unclear Encoding	0.33 (± 0.03)		
Inappropriate Encoding	0.16 (± 0.01)		

Table 11: Performance of ✦$_{2.5\text{-P}}$ across three independent runs. Each value corresponds to the mean and standard deviation (±).

A.10 Prompt Ablation

Visualization Error Classification Prompt

You will be provided with a visualization, its accompanying caption, and descriptions of the visualization error. These visualization errors represent ways in which people use visualization to spread misinformation.

Your task is to carefully examine the image and its accompanying caption. Then, based on the information and the descriptions of visualization errors, you need to identify which of the following error categories, if any, apply to this chart-caption pair.

If none of the visualization errors apply, you may classify the visualization error as "None."

Please classify which visualization errors are present and explain your reasoning. If more than one classification applies, include all applicable classifications in a list. Even if there is only one classification, the "classification" field must still be a list.

Only provide output in the following JSON format:

{

"reason": "[Explanation]",

"classification": ["Truncated axis/Dual axis/Value as area or volume/Inverted axis/Uneven binning/Unclear encoding/Inappropriate encoding/None"]

}

Image: {image}

Accompanying Text: {caption}

Error Descriptions: {visualizaton_error_descriptions}

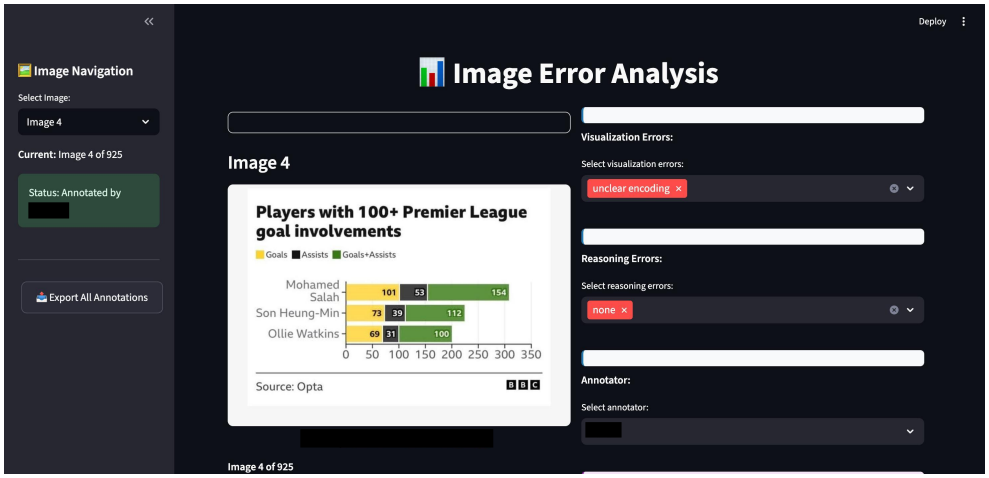


Figure 7: Example Picture of our Annotation Interface.

Prompt formulation can significantly influence the behavior of VLMs (Son and Lee, 2025). To assess the sensitivity of our results to prompt design, we conduct a prompt ablation using an alternative prompt variant. Specifically, we modify the task instructions to use more neutral wording and to slightly rephrase the classification objective. This change is intended to test whether the model's behavior is sensitive to minor variations in prompt phrasing and to evaluate the robustness of our results to alternative prompt formulations.

The results obtained for ✦ 2.5-P (Table 12) with this prompt variant closely match the performance observed with the prompt used in our main experiments (Table 11). This indicates that the overall results are largely insensitive to small changes in prompt wording, suggesting that the model's performance on our benchmark is robust to modest variations in prompt formulation.

A.11 Macro F1 scores

We report the combined Macro F1 scores along with Macro F1 for (i) reasoning error classification, (ii) visualization error classification on the whole benchmark (Table 13).

Because error categories are unevenly distributed and differ in intrinsic difficulty, this score provides a class-balanced perspective on performance. Unlike accuracy or weighted metrics, it shows whether models perform consistently across error types or disproportionately succeed on frequent or perceptually salient categories while failing on rarer ones. This metric, therefore, serves as a diagnostic indicator of robustness across the full error taxonomy.

	F1	PM	EM
Combined	0.61	0.86	0.08
Reasoning Errors	0.59	0.61	0.29
Visualization Errors	0.64	0.59	0.24
Cherry Picking	0.60		
Causal Inference	0.68		
Setting an Arbitrary Threshold	0.67		
Failure to Account for Statistical Nuances	0.13	**N/A**	**N/A**
Incorrect Reading of the Chart	0.11		
Data Validity Issues	0.01		
Misrepresentation of Scientific Studies	0.09		
Truncated Axis	0.68		
Dual Axis	0.94		
Value as Area/Vol.	0.65		
Inverted Axis	0.41	**N/A**	**N/A**
Uneven Binning	0.15		
Unclear Encoding	0.34		
Inappropriate Encoding	0.16		

Table 12: Ablation results for the prompt variant. Results are reported from a single run with ✦ 2.5-P

Model	Combined	Reasoning Errors	Visualization Errors
✦ 3-P	0.42	0.34	0.50
✦ 2.5-P	0.38	0.32	0.45
✦ 2.5-F	0.37	0.29	0.45
5	0.38	0.33	0.42
5-Mini	0.38	0.29	0.46
3	0.30	0.28	0.32
2.5	0.18	0.16	0.19
2.5-ChartQA	0.14	0.12	0.17

Table 13: Macro F1 scores of VLMs for reasoning and visualization errors separately and combined. Models consistently achieve higher scores on visualization error classification than reasoning errors, suggesting greater difficulty in identifying and reasoning about misinformation embedded in captions.

A.12 Grid-wise Results

To better understand model behavior under different misleadingness conditions, we present a de-

Model	△				○			
	F1$_m$	**F1$_w$**	**PM**	**EM**	**F1$_m$**	**F1$_w$**	**PM**	**EM**
3-P	0.16	0.50	0.79	0.08	0.30	0.69	0.91	0.28
2.5-P	0.18	0.54	0.75	0.02	0.28	0.67	0.90	0.07
2.5-F	0.16	0.49	0.75	0.00	0.28	0.68	0.84	0.02
5	0.17	0.53	0.80	0.01	0.26	0.67	0.90	0.16
5-Mini	0.16	0.50	0.81	0.00	0.28	0.68	0.88	0.06
3	0.15	0.49	0.62	0.01	0.22	0.58	0.87	0.03
2.5	0.08	0.22	0.75	0.00	0.11	0.28	0.81	0.04
2.5-ChartQA	0.05	0.17	0.69	0.01	0.10	0.28	0.75	0.08
	■				∅			
	F1$_m$	**F1$_w$**	**PM**	**EM**	**F1$_m$**	**F1$_w$**	**PM**	**EM**
3-P	0.45	0.74	0.95	0.10			0.91	0.47
2.5-P	0.43	0.74	0.98	0.01			0.57	0.09
2.5-F	0.44	0.74	0.98	0.02			0.37	0.04
5	0.42	0.74	0.98	0.00	**N/A**	**N/A**	0.75	0.30
5-Mini	0.42	0.74	0.98	0.01			0.60	0.16
3	0.34	0.67	0.99	0.00			0.65	0.08
2.5	0.22	0.42	0.75	0.01			0.96	0.73
2.5-ChartQA	0.21	0.41	0.74	0.01			0.93	0.62

Table 14: Performance of VLMs across the 2×2 grid. We report combined macro F1 (F1$_m$), combined weighted F1 (F1$_w$), combined Partial Match (PM), and combined Exact Match (EM) for each model. Across the grids, models perform poorly on this task, with F1 scores staying below 0.75, highlighting the difficulty of reliably identifying and classifying multimodal misinformation. For the *Non-Misleading Caption, Non-Misleading Viz* cell, we omit F1 scores as they are not applicable since there are no misinformation errors present, there are no true positives to evaluate per-error F1, and hence all per-error F1 scores are trivially zero and hence marked as **N/A**.

tailed breakdown across the four cells of the 2×2 misleadingness grid.

We first report combined performance within each cell (Table 14), followed by disaggregated results for reasoning and visualization error classification (Table 15). This grid-wise analysis clarifies whether performance varies systematically depending on whether misleadingness arises from the caption, the visualization, both, or neither.

To further characterize error-specific behavior, we report per-error F1 scores for each reasoning error type (Table 16) and visualization error type (Table 17) within each grid cell. These fine-grained results expose asymmetries in how models handle distinct categories of reasoning and visual distortions.

Precision, Recall, and False Positive Rates. Beyond F1, we provide per-error Precision and Recall scores across the full benchmark (Table 18) and across the 2 × 2 grid (Tables 19, 20) to disentangle over-prediction from under-detection. Given the deployment relevance of over-flagging benign content, we additionally report per-error False Positive Rates (FPR) on the whole benchmark (Table 22) and within each grid cell (Tables 21, 23). These metrics quantify calibration tendencies and help identify whether models systematically default toward predicting misleadingness in the absence of clear evidence.

	△								◯							
	Reasoning Errors				Visualization Errors				Reasoning Errors				Visualization Errors			
Model	$F1_m$	$F1_w$	PM	EM	$F1_m$	$F1_w$	PM	EM	$F1_m$	$F1_w$	PM	EM	$F1_m$	$F1_w$	PM	EM
3-P	0.32	0.50	0.54	0.11			0.61	0.61			0.61	0.61	0.59	0.69	0.76	0.43
2.5-P	0.36	0.54	0.71	0.06			0.20	0.20			0.30	0.30	0.57	0.67	0.84	0.25
2.5-F	0.32	0.49	0.68	0.02			0.30	0.30			0.07	0.07	0.57	0.68	0.82	0.28
5	0.33	0.53	0.66	0.01	**N/A**	**N/A**	0.52	0.52	**N/A**	**N/A**	0.47	0.47	0.53	0.67	0.81	0.31
5-Mini	0.33	0.50	0.68	0.00			0.51	0.51			0.20	0.20	0.57	0.68	0.85	0.30
3	0.31	0.49	0.60	0.12			0.04	0.04			0.40	0.40	0.44	0.58	0.75	0.11
2.5	0.16	0.22	0.28	0.01			0.64	0.64			0.73	0.73	0.22	0.28	0.26	0.05
2.5-ChartQA	0.10	0.17	0.24	0.02			0.58	0.58			0.65	0.65	0.20	0.28	0.27	0.12
	■								∅							
	Reasoning Errors				Visualization Errors				Reasoning Errors				Visualization Errors			
	$F1_m$	$F1_w$	PM	EM	$F1_m$	$F1_w$	PM	EM	$F1_m$	$F1_w$	PM	EM	$F1_m$	$F1_w$	PM	EM
3-P	0.39	0.66	0.80	0.21	0.51	0.84	0.82	0.48			0.83	0.83			0.55	0.55
2.5-P	0.39	0.68	0.91	0.08	0.48	0.81	0.79	0.19			0.52	0.52			0.15	0.15
2.5-F	0.39	0.67	0.94	0.03	0.49	0.82	0.81	0.29			0.19	0.19			0.22	0.22
5	0.43	0.68	0.95	0.03	0.41	0.81	0.79	0.31	**N/A**	**N/A**	0.62	0.62	**N/A**	**N/A**	0.42	0.42
5-Mini	0.36	0.67	0.94	0.01	0.48	0.84	0.83	0.30			0.35	0.35			0.42	0.42
3	0.37	0.68	0.97	0.05	0.30	0.67	0.67	0.05			0.62	0.62			0.11	0.11
2.5	0.21	0.41	0.56	0.13	0.24	0.44	0.42	0.15			0.87	0.87			0.82	0.82
2.5-ChartQA	0.21	0.41	0.57	0.11	0.20	0.42	0.40	0.17			0.77	0.77			0.77	0.77

Table 15: Performance breakdown across both reasoning and visualization errors, reported for the 2×2 grid. Each section shows macro F1 ($F1_m$), weighted F1 ($F1_w$), Partial Match (PM), and Exact Match (EM) scores separately for reasoning and visualization classification tasks. In most cases, models achieve lower F1 scores on reasoning error classification than on visualization error classification, indicating that reasoning-based deception is more challenging for current VLMs to detect. F1 scores are marked **N/A** in grid cells where the relevant modality contains no positive instances due to the absence of the corresponding error types.

	△							◯						
Model	Cher. Pick	Caus. Infer.	Arb. Thr.	Stat. Nu.	Chart Read	Data Val.	Mis. Sci.	Cher. Pick	Caus. Infer.	Arb. Thr.	Stat. Nu.	Chart Read	Data Val.	Mis. Sci.
3-P	0.48	0.80	0.16	0.08	0.00	0.37	0.38							
2.5-P	0.49	0.85	0.27	0.17	0.06	0.25	0.44							
2.5-F	0.44	0.81	0.18	0.16	0.04	0.23	0.39							
5	0.53	0.83	0.26	0.16	0.04	0.22	0.30				**N/A**			
5-Mini	0.51	0.80	0.16	0.15	0.03	0.25	0.40							
3	0.51	0.73	0.24	0.36	0.03	0.06	0.22							
2.5	0.47	0.13	0.14	0.18	0.03	0.08	0.06							
2.5-ChartQA	0.45	0.09	0.08	0.02	0.00	0.00	0.06							
	■							∅						
	Cher. Pick	Caus. Infer.	Arb. Thr.	Stat. Nu.	Chart Read	Data Val.	Mis. Sci.	Cher. Pick	Caus. Infer.	Arb. Thr.	Stat. Nu.	Chart Read	Data Val.	Mis. Sci.
3-P	0.64	0.71	0.86	0.28	0.09	0.14	0.00							
2.5-P	0.71	0.63	0.89	0.22	0.19	0.07	0.00							
2.5-F	0.72	0.53	0.91	0.23	0.13	0.09	0.14							
5	0.72	0.62	0.87	0.22	0.12	0.33	0.00				**N/A**			
5-Mini	0.72	0.52	0.92	0.22	0.12	0.07	0.00							
3	0.68	0.67	0.90	0.22	0.10	0.00	0.00							
2.5	0.65	0.38	0.30	0.08	0.03	0.00	0.04							
2.5-ChartQA	0.65	0.40	0.28	0.04	0.08	0.00	0.05							

Table 16: Per-error F1 scores for reasoning error classification across the 2×2 grid. VLMs perform relatively well on certain visualization error types such as *Cherry-picking*, *Setting an arbitrary threshold*, and *Causal Inference*, but struggle on others. Cells corresponding to *Non-Misleading Caption, Misleading Viz* and *Non-Misleading Caption, Non-Misleading Viz* do not contain any reasoning errors, so per-error F1 scores are marked as **N/A**.

Model	△ Trunc. Axis	△ Dual Axis	△ Area/ Vol.	△ Inv. Axis	△ Uneven Bin.	△ Unclr. Enc.	△ Inappr. Enc.	◯ Trunc. Axis	◯ Dual Axis	◯ Area/ Vol.	◯ Inv. Axis	◯ Uneven Bin.	◯ Unclr. Enc.	◯ Inappr. Enc.
3-P	N/A							0.75	0.97	0.74	0.49	0.23	0.60	0.36
2.5-P								0.71	0.97	0.75	0.47	0.20	0.58	0.32
2.5-F								0.74	0.95	0.76	0.47	0.17	0.60	0.30
5								0.65	0.97	0.77	0.16	0.23	0.60	0.32
5-Mini								0.70	0.98	0.73	0.31	0.31	0.62	0.31
3								0.44	0.94	0.62	0.15	0.17	0.52	0.25
2.5								0.27	0.80	0.11	0.07	0.00	0.18	0.10
2.5-ChartQA								0.22	0.77	0.17	0.02	0.00	0.15	0.11
	■ Trunc. Axis	■ Dual Axis	■ Area/ Vol.	■ Inv. Axis	■ Uneven Bin.	■ Unclr. Enc.	■ Inappr. Enc.	∅ Trunc. Axis	∅ Dual Axis	∅ Area/ Vol.	∅ Inv. Axis	∅ Uneven Bin.	∅ Unclr. Enc.	∅ Inappr. Enc.
3-P	0.86	0.98	0.83	0.67	0.04	0.08	0.12	N/A						
2.5-P	0.84	0.96	0.80	0.62	0.03	0.08	0.04							
2.5-F	0.83	0.95	0.83	0.65	0.02	0.09	0.05							
5	0.69	0.97	0.86	0.19	0.03	0.08	0.06							
5-Mini	0.78	0.98	0.86	0.55	0.04	0.08	0.03							
3	0.38	0.90	0.68	0.06	0.00	0.05	0.03							
2.5	0.35	0.87	0.19	0.16	0.00	0.07	0.04							
2.5-ChartQA	0.25	0.87	0.18	0.14	0.00	0.00	0.00							

Table 17: Per-error F1 scores for visualization error classification across the 2×2 grid. VLMs perform relatively well on certain visualization error types such as *Dual Axis*, *Truncated Axis*, and *Value as Area or Volume*, but struggle on others. Cells corresponding to *Misleading Caption, Non-Misleading Viz* and *Non-Misleading Caption, Non-Misleading Viz* do not contain any visualization errors, so per-error F1 scores are marked as **N/A**.

Model	Reasoning Errors														Visualization Errors											
	Cher. Pick		Caus. Infer.		Arb. Thr.		Stat. Nu.		Chart Read		Data Val.		Mis. Sci.		Trunc. Axis		Dual Axis		Area/ Vol.		Inv. Axis		Uneven Bin.		Unclr. Enc.		Inappr. Enc.	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
3-P	0.44	0.54	0.69	0.73	0.73	0.40	0.06	0.50	0.02	0.29	0.25	0.28	0.17	0.45	0.54	0.85	0.81	0.99	0.74	0.68	0.48	0.50	0.07	0.67	0.29	0.68	0.15	0.59
2.5-P	0.36	0.70	0.52	0.91	0.63	0.50	0.06	0.86	0.03	0.50	0.04	0.25	0.18	0.58	0.50	0.78	0.78	0.98	0.71	0.68	0.43	0.43	0.05	0.83	0.18	0.89	0.09	0.76
2.5-F	0.31	0.68	0.44	0.90	0.57	0.46	0.05	0.99	0.02	0.42	0.04	0.25	0.19	0.50	0.44	0.86	0.79	0.95	0.62	0.77	0.43	0.45	0.04	0.83	0.22	0.77	0.09	0.68
5	0.39	0.80	0.56	0.85	0.64	0.50	0.07	0.98	0.02	0.62	0.10	0.18	0.17	0.31	0.49	0.62	0.82	0.97	0.70	0.77	0.23	0.11	0.06	0.78	0.23	0.83	0.12	0.63
5-Mini	0.35	0.80	0.44	0.90	0.54	0.46	0.05	0.98	0.02	0.69	0.04	0.32	0.20	0.38	0.51	0.72	0.81	0.98	0.56	0.82	0.48	0.29	0.09	0.89	0.23	0.88	0.12	0.58
3	0.31	0.84	0.51	0.72	0.48	0.51	0.10	0.77	0.01	0.65	0.33	0.03	0.23	0.12	0.10	0.71	0.75	0.88	0.64	0.55	0.24	0.07	0.14	0.11	0.17	0.87	0.06	0.76
2.5	0.33	0.73	0.58	0.12	0.34	0.13	0.12	0.12	0.01	0.15	0.25	0.04	0.02	0.23	0.11	0.33	0.73	0.77	0.51	0.08	0.10	0.08	0.00	0.00	0.16	0.12	0.05	0.10
2.5-ChartQA	0.29	0.72	0.51	0.10	0.40	0.10	0.04	0.02	0.02	0.12	0.00	0.00	0.02	0.31	0.09	0.21	0.57	0.83	0.38	0.10	0.04	0.05	0.00	0.00	0.22	0.08	0.05	0.10

Table 18: Per-error Precision (P) and Recall (R) scores for reasoning and visualization error classification across the whole benchmark. Models show relatively high recall but low precision for certain error types such as *Failure to Account for Statistical Nuance* and *Unclear Encoding*, indicating a tendency to over-predict these categories even when not applicable. This behavior highlights challenges in accurately distinguishing subtle or context-dependent misinformation patterns.

Model	△ Cher. Pick P	R	Caus. Infer. P	R	Arb. Thr. P	R	Stat. Nu. P	R	Chart Read P	R	Data Val. P	R	Mis. Sci. P	R	○ Cher. Pick P	R	Caus. Infer. P	R	Arb. Thr. P	R	Stat. Nu. P	R	Chart Read P	R	Data Val. P	R	Mis. Sci. P	R
3-P	0.45	0.52	0.91	0.71	0.64	0.09	0.04	0.33	0.00	0.00	0.53	0.28	0.27	0.67														
2.5-P	0.40	0.62	0.81	0.90	0.76	0.16	0.09	0.94	0.03	0.28	0.25	0.25	0.31	0.79														
2.5-F	0.36	0.57	0.76	0.87	0.76	0.10	0.08	1.00	0.02	0.28	0.22	0.25	0.28	0.63														
5	0.40	0.75	0.82	0.83	0.74	0.16	0.08	0.95	0.02	0.43	0.34	0.16	0.26	0.37														
5-Mini	0.39	0.76	0.73	0.88	0.56	0.10	0.08	0.98	0.02	0.71	0.20	0.33	0.33	0.53							**N/A**							
3	0.38	0.74	0.82	0.66	0.45	0.17	0.22	0.88	0.02	1.00	0.50	0.03	0.37	0.16														
2.5	0.39	0.59	0.86	0.07	0.31	0.09	0.20	0.17	0.01	0.43	0.50	0.05	0.04	0.16														
2.5-ChartQA	0.37	0.59	0.66	0.05	0.50	0.04	0.04	0.02	0.00	0.00	0.00	0.00	0.03	0.21														
	■														⌀													
	Cher. Pick P	R	Caus. Infer. P	R	Arb. Thr. P	R	Stat. Nu. P	R	Chart Read P	R	Data Val. P	R	Mis. Sci. P	R	Cher. Pick P	R	Caus. Infer. P	R	Arb. Thr. P	R	Stat. Nu. P	R	Chart Read P	R	Data Val. P	R	Mis. Sci. P	R
3-P	0.74	0.56	0.64	0.79	0.95	0.79	0.18	0.66	0.05	0.37	0.10	0.25	0.00	0.00														
2.5-P	0.64	0.79	0.47	0.93	0.86	0.91	0.13	0.77	0.11	0.58	0.04	0.25	0.00	0.00														
2.5-F	0.64	0.82	0.36	0.98	0.91	0.91	0.13	0.98	0.08	0.47	0.05	0.25	0.14	0.14														
5	0.63	0.85	0.47	0.90	0.83	0.93	0.12	1.00	0.06	0.68	0.25	0.50	0.14	0.14														
5-Mini	0.61	0.86	0.36	0.95	0.93	0.91	0.12	0.98	0.06	0.68	0.04	0.25	0.00	0.00							**N/A**							
3	0.53	0.96	0.53	0.91	0.88	0.93	0.13	0.66	0.06	0.53	0.00	0.00	0.00	0.00														
2.5	0.52	0.88	0.62	0.28	0.81	0.18	0.11	0.06	0.02	0.05	0.00	0.00	0.02	0.43														
2.5-ChartQA	0.52	0.88	0.65	0.28	0.66	0.18	0.07	0.03	0.06	0.16	0.00	0.00	0.03	0.57														

Table 19: Per-error Precision (P) and Recall (R) for reasoning error classification across the 2×2 grid. Errors such as *Failure to Account for Statistical Nuance* exhibit high recall but very low precision, highlighting a tendency among models to over-predict this label even when it is not present. Cells corresponding to *Non-Misleading Caption, Misleading Viz* and *Non-Misleading Caption, Non-Misleading Viz* do not contain any reasoning errors, so values are marked as **N/A**.

Model	△ Trunc. Axis P	R	Dual Axis P	R	Area/ Vol. P	R	Inv. Axis P	R	Uneven Bin. P	R	Unclr. Enc. P	R	Inappr. Enc. P	R	○ Trunc. Axis P	R	Dual Axis P	R	Area/ Vol. P	R	Inv. Axis P	R	Uneven Bin. P	R	Unclr. Enc. P	R	Inappr. Enc. P	R
3-P															0.68	0.83	0.96	0.99	0.83	0.66	0.51	0.48	0.14	0.69	0.54	0.69	0.26	0.58
2.5-P															0.66	0.77	0.95	0.98	0.81	0.69	0.54	0.41	0.11	0.87	0.43	0.89	0.20	0.78
2.5-F															0.67	0.82	0.95	0.95	0.73	0.79	0.53	0.43	0.10	0.87	0.49	0.77	0.19	0.68
5															0.68	0.62	0.97	0.98	0.78	0.77	0.29	0.11	0.13	0.81	0.47	0.84	0.21	0.64
5-Mini							**N/A**								0.70	0.70	0.97	0.98	0.65	0.83	0.48	0.23	0.19	0.94	0.48	0.89	0.21	0.60
3															0.32	0.69	0.96	0.92	0.72	0.55	0.50	0.09	0.25	0.12	0.37	0.87	0.15	0.77
2.5															0.24	0.32	0.92	0.70	0.56	0.06	0.13	0.04	0.00	0.00	0.30	0.12	0.12	0.09
2.5-ChartQA															0.24	0.20	0.74	0.80	0.48	0.10	0.02	0.02	0.00	0.00	0.52	0.09	0.11	0.11
	■																				⌀							
	Trunc. Axis P	R	Dual Axis P	R	Area/ Vol. P	R	Inv. Axis P	R	Uneven Bin. P	R	Unclr. Enc. P	R	Inappr. Enc. P	R	Trunc. Axis P	R	Dual Axis P	R	Area/ Vol. P	R	Inv. Axis P	R	Uneven Bin. P	R	Unclr. Enc. P	R	Inappr. Enc. P	R
3-P	0.83	0.90	0.97	0.98	1.00	0.71	0.88	0.54	0.02	0.50	0.04	0.50	0.07	0.62														
2.5-P	0.85	0.82	0.96	0.97	0.98	0.67	1.00	0.45	0.02	0.50	0.04	0.81	0.02	0.50														
2.5-F	0.76	0.92	0.95	0.95	0.95	0.74	1.00	0.48	0.01	0.50	0.05	0.62	0.03	0.62														
5	0.77	0.62	0.97	0.97	0.99	0.76	1.00	0.10	0.02	0.50	0.04	0.69	0.03	0.50														
5-Mini	0.81	0.76	0.98	0.98	0.94	0.80	1.00	0.38	0.02	0.50	0.04	0.69	0.02	0.25							**N/A**							
3	0.25	0.74	0.97	0.84	0.93	0.54	0.25	0.03	0.00	0.00	0.03	0.75	0.01	0.62														
2.5	0.35	0.36	0.88	0.86	0.77	0.11	0.18	0.14	0.00	0.00	0.04	0.12	0.02	0.25														
2.5-ChartQA	0.29	0.22	0.88	0.86	0.62	0.10	0.21	0.10	0.00	0.00	0.00	0.00	0.00	0.00														

Table 20: Per-error Precision (P) and Recall (R) for visualization error classification across the 2×2 grid. Models often exhibit high recall but low precision for certain categories such as *Unclear Encoding*, suggesting over-prediction of these error types even when not warranted. Cells corresponding to *Misleading Text, Non-Misleading Viz* and *Non-Misleading Text, Non-Misleading Viz* do not contain visualization errors, so values are marked as **N/A**.

Model	△ Cher. Pick	△ Caus. Infer.	△ Arb. Thr.	△ Stat. Nu.	△ Chart Read	△ Data Val.	△ Mis. Sci.	○ Cher. Pick	○ Caus. Infer.	○ Arb. Thr.	○ Stat. Nu.	○ Chart Read	○ Data Val.	○ Mis. Sci.
3-P	0.38	0.09	0.02	0.62	0.07	0.02	0.04	0.09	0.06	0.03	0.20	0.09	0.02	0.02
2.5-P	0.53	0.25	0.02	0.83	0.07	0.06	0.04	0.21	0.15	0.06	0.46	0.18	0.15	0.02
2.5-F	0.58	0.32	0.02	0.93	0.12	0.08	0.04	0.25	0.18	0.08	0.70	0.32	0.16	0.01
5	0.64	0.21	0.03	0.91	0.20	0.03	0.02	0.15	0.10	0.04	0.38	0.30	0.05	0.01
5-Mini	0.69	0.38	0.04	0.96	0.38	0.12	0.03	0.21	0.18	0.07	0.68	0.50	0.21	0.01
3	0.70	0.18	0.10	0.27	0.58	0.00	0.01	0.28	0.14	0.10	0.22	0.49	0.00	0.00
2.5	0.54	0.01	0.10	0.06	0.26	0.00	0.09	0.22	0.01	0.05	0.02	0.16	0.00	0.02
2.5-ChartQA	0.59	0.03	0.02	0.03	0.07	0.01	0.15	0.31	0.02	0.02	0.02	0.05	0.01	0.07
	■ Cher. Pick	■ Caus. Infer.	■ Arb. Thr.	■ Stat. Nu.	■ Chart Read	■ Data Val.	■ Mis. Sci.	∅ Cher. Pick	∅ Caus. Infer.	∅ Arb. Thr.	∅ Stat. Nu.	∅ Chart Read	∅ Data Val.	∅ Mis. Sci.
3-P	0.19	0.16	0.03	0.46	0.28	0.02	0.01	0.05	0.04	0.03	0.08	0.02	0.01	0.00
2.5-P	0.44	0.35	0.10	0.75	0.18	0.05	0.01	0.11	0.12	0.05	0.31	0.03	0.19	0.01
2.5-F	0.46	0.58	0.06	0.95	0.23	0.04	0.01	0.19	0.17	0.08	0.58	0.10	0.18	0.00
5	0.51	0.34	0.14	0.98	0.39	0.01	0.01	0.08	0.08	0.06	0.31	0.11	0.04	0.00
5-Mini	0.55	0.58	0.05	0.99	0.39	0.05	0.01	0.16	0.15	0.10	0.62	0.17	0.21	0.00
3	0.84	0.27	0.10	0.62	0.34	0.00	0.00	0.24	0.11	0.10	0.20	0.25	0.00	0.00
2.5	0.83	0.06	0.03	0.07	0.09	0.01	0.25	0.10	0.01	0.00	0.01	0.06	0.00	0.02
2.5-ChartQA	0.83	0.05	0.06	0.06	0.10	0.03	0.30	0.21	0.01	0.02	0.01	0.03	0.00	0.03

Table 21: Per-error False Positive Rates (FPR) for reasoning error classification across the 2×2 grid. Models show relatively high FPR for *Failure to Account for Statistical Nuance*, indicating a tendency to over-predict this category even when not applicable.

	Reasoning Errors							**Visualization Errors**						
Model	Cher. Pick	Caus. Infer.	Arb. Thr.	Stat. Nu.	Chart Read	Data Val.	Mis. Sci.	Trunc. Axis	Dual Axis	Area/ Vol.	Inv. Axis	Uneven Bin..	Unclr. Enc.	Inappr. Enc.
3-P	0.15	0.07	0.03	0.32	0.10	0.02	0.02	0.04	0.04	0.04	0.01	0.06	0.27	0.17
2.5-P	0.28	0.19	0.05	0.57	0.12	0.12	0.02	0.04	0.04	0.06	0.01	0.10	0.62	0.38
2.5-F	0.32	0.26	0.06	0.77	0.20	0.12	0.02	0.06	0.04	0.10	0.02	0.11	0.42	0.34
5	0.27	0.15	0.05	0.59	0.25	0.04	0.01	0.03	0.04	0.07	0.01	0.07	0.44	0.24
5-Mini	0.33	0.26	0.07	0.79	0.38	0.16	0.01	0.04	0.04	0.13	0.01	0.05	0.45	0.24
3	0.41	0.16	0.10	0.29	0.44	0.00	0.00	0.34	0.05	0.06	0.01	0.00	0.67	0.54
2.5	0.32	0.02	0.05	0.04	0.16	0.00	0.08	0.15	0.05	0.02	0.02	0.00	0.10	0.08
2.5-ChartQA	0.40	0.02	0.03	0.03	0.06	0.01	0.12	0.12	0.11	0.03	0.03	0.00	0.04	0.09

Table 22: Per-error False Postiive Rates (FPR) for reasoning and visualization error classification on the whole dataset. Models show relatively high FPR for certain error types such as *Failure to Account for Statistical Nuance* and *Unclear Encoding*, indicating a tendency to over-predict these categories even when not applicable.

Model	△ Trunc. Axis	△ Dual Axis	△ Area/ Vol.	△ Inv. Axis	△ Uneven Bin.	△ Unclr. Enc.	△ Inappr. Enc.	○ Trunc. Axis	○ Dual Axis	○ Area/ Vol.	○ Inv. Axis	○ Uneven Bin.	○ Unclr. Enc.	○ Inappr. Enc.
3-P	0.03	0.09	0.04	0.01	0.03	0.16	0.14	0.04	0.01	0.04	0.02	0.06	0.33	0.23
2.5-P	0.04	0.10	0.05	0.01	0.07	0.57	0.40	0.04	0.01	0.06	0.01	0.10	0.63	0.42
2.5-F	0.06	0.09	0.07	0.01	0.07	0.38	0.33	0.04	0.01	0.10	0.02	0.12	0.43	0.40
5	0.04	0.08	0.06	0.01	0.04	0.33	0.18	0.03	0.01	0.08	0.01	0.08	0.52	0.32
5-Mini	0.03	0.09	0.11	0.01	0.04	0.35	0.21	0.03	0.01	0.16	0.01	0.06	0.52	0.30
3	0.64	0.11	0.06	0.00	0.00	0.47	0.37	0.15	0.01	0.08	0.00	0.01	0.79	0.58
2.5	0.28	0.08	0.02	0.02	0.00	0.10	0.06	0.10	0.02	0.02	0.01	0.00	0.16	0.09
2.5-ChartQA	0.22	0.16	0.03	0.04	0.01	0.04	0.05	0.07	0.08	0.04	0.03	0.00	0.04	0.12
	■ Trunc. Axis	■ Dual Axis	■ Area/ Vol.	■ Inv. Axis	■ Uneven Bin.	■ Unclr. Enc.	■ Inappr. Enc.	∅ Trunc. Axis	∅ Dual Axis	∅ Area/ Vol.	∅ Inv. Axis	∅ Uneven Bin.	∅ Unclr. Enc.	∅ Inappr. Enc.
3-P	0.02	0.02	0.00	0.00	0.09	0.33	0.15	0.07	0.02	0.07	0.01	0.06	0.30	0.11
2.5-P	0.02	0.02	0.01	0.00	0.13	0.60	0.38	0.07	0.02	0.09	0.02	0.11	0.70	0.30
2.5-F	0.03	0.03	0.03	0.00	0.16	0.41	0.36	0.10	0.03	0.15	0.03	0.11	0.47	0.27
5	0.02	0.02	0.00	0.00	0.12	0.50	0.24	0.05	0.02	0.10	0.02	0.07	0.43	0.17
5-Mini	0.02	0.01	0.04	0.00	0.08	0.51	0.27	0.07	0.02	0.16	0.01	0.05	0.44	0.17
3	0.24	0.02	0.03	0.01	0.00	0.84	0.68	0.35	0.04	0.06	0.01	0.00	0.66	0.59
2.5	0.08	0.07	0.02	0.04	0.00	0.09	0.17	0.11	0.04	0.01	0.01	0.00	0.06	0.03
2.5-ChartQA	0.06	0.07	0.05	0.02	0.00	0.07	0.15	0.10	0.09	0.02	0.03	0.00	0.03	0.04

Table 23: Per-error False Positive Rates (FPR) for visualization error classification across the 2×2 grid. Models show relatively high FPR for *Unclear Encoding*, indicating a tendency to over-predict this category even when not applicable.

A.13 Comparison to prior-work

	Task	Source of deception	Error Granularity	Chart-Caption Interaction	Error Attribution	Data Source	Evaluation Focus
(Kahou et al., 2017; Kafle et al., 2018; Methani et al., 2020; Masry et al., 2022)	Chart question answering	Assumed truthful charts	None	✗	✗	Synthetic/curated charts	Chart comprehension accuracy
(Huang et al., 2024)	Caption factuality checking and correction	Caption-chart factual mismatch	Low (value, label, trend)	✓	✗	Generated captions	Caption factual consistency
(Akhtar et al., 2024)	Fact-checking claims against charts	Explicit factual claims	Moderate (supported/refuted + explanation)	✓	✗	Real-world charts	Claim verification accuracy
(Chen et al., 2025)	MCQ-based misleading chart reasoning	Visualization design manipulation	Fine-grained (21 misleaders)	○	○	Synthetic + standardized charts	Answer correctness and reasoning
(Lo and Qu, 2024; Alexander et al., 2024)	Misleader detection via prompting	Visualization design errors	Medium (design issues)	✗	○	Social media charts	Prompt sensitivity and detection AUC
(Tonglet et al., 2025b)	Misleader classification	Visualization design violations	Fine-grained (12+ misleaders)	✗	✓(design only)	Real + synthetic charts	Multi-label misleader detection
(Mahbub et al., 2025)	Susceptibility analysis of VLMs	Visualization design distortions	Design-level	✗	✗	Synthetic chart pairs	Behavioral degradation analysis
Ours	**Misleading chart-caption detection**	**Caption-based reasoning errors & visualization errors**	**Fine-grained (reasoning + design taxonomy)**	✓	**✓(caption vs. visualization)**	**Real-world charts + human-authored captions**	**Diagnostic error attribution and over-flagging analysis**

Table 24: Comparison of our benchmark with prior work on misleading visualizations and chart understanding. Unlike existing benchmarks, our work explicitly disentangles misleadingness arising from caption-level reasoning errors versus visualization design errors and evaluates fine-grained error attribution by Vision-Language Models.