This is the author's version of the article that has been published in the proceedings of the "1st Workshop on GenAI, Agents, and the Future of VIS", at IEEE VIS 2025, Vienna, Austria. https://visxgenai.github.io/

True (VIS) Lies: A Preliminary Analysis of How Generative AI is Capable of Recognizing Visualization Lies and their Rhetoric

Graziano Blasilli*
A.WA.RE – Advanced Visualization & Visual Analytics REsearch Group
Sapienza University of Rome, Rome, Italy

Marco Angelini†
Link Campus University, Rome, Italy

Abstract

This ongoing study analyzes the capability of multimodal Large Language Models (LLMs) to detect and interpret misleading data visualizations first, and then link them to root causes and intentionality using visualization rhetoric as a means to explain them. Three research questions are formulated and preliminarily analyzed experimentally using a dataset of 2,336 COVID-19-related tweets, half of which contain misleading visualizations using two models: Gemma3-27B and Qwen2.5-VL-72B. Results are additionally qualitatively evaluated on a small collection of "real-life" Vis Lies annotated by visualization researchers. The study's preliminary findings suggest that while these models have a baseline ability to detect visual deception, their effectiveness is highly dependent on the prompting strategy. Moreover, the results offer initial evidence that LLMs can begin to reason about the rhetorical dimension and intentionality of visual misinformation. The authors acknowledge the preliminary and non-generalizable nature of current findings, highlighting the need for further research.

Index Terms: LLMs, misleading visualizations, visualization rhetoric

1 Introduction

Generative Artificial Intelligence (genAI), and, in particular, Large Language Models (LLMs) and Vision Models (VLMs), permeated society and working processes regardless of the application domain in the last three years [24]. Promising assistance in solving and organizing tasks helps millions of users search, generate, and adapt information in multiple forms (e.g., text, images). It showed interesting behavior, showing pros and cons, in supporting decision-making tasks and the interpretation of information, with several works demonstrating good capabilities [3, 10]. Among these tasks, of particular interest is their usage for the interpretation of charts [4], which represent an aggregated visual form of analytical information from which to derive insights. This task can happen at different levels, ranging from scientific charts to infographics and communication images. The capability of correctly interpreting a chart can be affected by several factors, ranging from an adequate level of visualization literacy for the user to the knowledge of the domain from which data are represented, to the potential presence of bias or errors, implicit and explicit, in the chart itself. Unfortunately, these characteristics may not be present (in full or in part) in genAI models and their users, particularly lay users, worsened by a "positivist" attitude in the usage of these models and in the negligible capability of these models to persuade non-savvy users. The risk for lay users is to trust an incorrect chart simply because the genAI model misinterpreted it for the wrong reasons while being prompted to analyze it.

Several works coped with the problem of benchmarking and improving LLMs and VLMs to interpret charts correctly [1, 15, 14], most focused on detecting and reducing misinterpretation. In this work, we propose (i) a preliminary analysis of the exposure of lay users with respect to capability to recognize errors directly present in the charts, which we label according to the community as visualization lies (Vis lies) in two ways: free or hinted, but not specified directly, using long prompts, as this may not be representative of how lay and novice users interact with LLMs. (ii) an analysis of the detection capability of the root causes for misleading interpretation, to recognize if the error is voluntary (e.g., due to propaganda or bias) and which rhetorical means have been used, according to the visualization rhetorical taxonomy proposed by Hullman and Diakopoulos [8]. Preliminary results on `Gemma3:27B` and `Qwen2.5-VL:72B` models show interesting behavior in supporting these tasks, which can guide further analysis and refinement.

*e-mail: blasilli@diag.uniroma1.it
†e-mail: m.angelini@unilink.it

2 Related Work

Large Language Models and Generative AI capabilities in supporting visualization have been extensively tested in the last three years [23]. Several efforts can be identified for visualization generation [20, 5, 22, 2, 19], with a focus on generating a correct visualization following visualization literacy rules [6, 15]. While these works tested the LLMs' capability not to produce erroneous visualizations, their focus is not on detecting them and the eventual misinterpretations they provide to a lay user. Another part of the literature focused on supporting a user in the automatic or semi-automatic analysis and interpretation of charts [7, 21]. Podo et al. [16] created a dedicated model, distilled from ChatGPT-4, to support generation, explanation, and analysis suggestions for a lay user or domain expert coping with visual data representations. Islam et al. [9] extensively tested LLMs and vision models in chart understanding and reasoning, highlighting a non-negligible loss in accuracy due to hallucinations, factual errors, and data bias.

Finally, a few works directly targeted the problem of potential misinterpretations of generated charts and the detection of the possible causes of it. Pandey and Ottley [15] benchmarked multiple AI models, testing their capabilities to interpret charts and eventually discovering low performance in spotting misleading visualizations, with just 30% accuracy. Alexander et al. [1] show that GPT-4 models can detect misleading visualizations [13, 18] with varying accuracy depending on the used prompt-engineering technique. The authors share the goal with our work, but limited the testing only to ChatGPT-4 while we expand on this work by considering additional models. Moreover, we make a step forward by trying to detect the potentially malicious nature of the misleading visualization and the visual rhetoric strategy used. Lo and Qu [14] also delve into this problem, focusing on defining nine types of prompts, from simple to complex, to support the detection of misleading visualizations. The authors focus on improving detection capabilities, while in our work, we want to test more the case of a lay user potentially exposed to misinformation (looking more at the simple prompts than more complex ones) and the potential root causes and rhetoric used to convey the erroneous information. Additionally, Szafir [11] shows how, in most cases, misleading visualizations are not predominantly based on violations of visualization design guidelines, but other aspects play a role in it. Among them, we identified the usage of visualization rhetoric, which is not dissimilar to what can happen in natural language by using classic rhetorical constructs to deliver a message [17].

3 Method

This study aims to evaluate multimodal LLMs in two directions. First, we assess whether these models can identify misleading elements in

Table 1: Types of visualization rhetoric.

Rhetoric Type	Description and Example
Information Access Rhetoric	Refers to choices about what data to include, omit, filter, or aggregate. These choices shape what the viewer sees and which patterns or narratives are emphasized.
Provenance Rhetoric	Relates to how a visualization communicates trustworthiness, such as by citing sources, explaining methods, noting uncertainty, or including methodology or assumptions.
Mapping Rhetoric	Involves how data values are translated into visual features such as position, size, scale, metaphor, or color. This can create emphasis or shape perception beyond what the raw data conveys.
Linguistic-Based Rhetoric	Uses language elements like titles, labels, annotations, or captions to guide interpretation. Often includes metaphor, irony, rhetorical questions, or other narrative framing.
Procedural Rhetoric	Relates to interaction and navigation design, such as default views, filters, or animations. These features guide how users explore the data and can influence which conclusions they reach.

visualizations with and without explicit prompting instruction. Second, we test whether they can recognize rhetorical techniques in visualizations, following the framework of Hullman and Diakopoulos [8], and reason about whether such rhetoric contributes to intentional deception. To this end, we formulate three research questions (RQ1–RQ3), and we address these questions using two complementary experiments (EXP1 and EXP2).

3.1 Research Questions

- **RQ1**: Do multimodal LLMs possess the nuanced understanding required to identify and flag misleading elements in data visualizations? [14]
- **RQ2**: Are multimodal LLMs capable of identifying rhetorical techniques in misleading data visualizations?
- **RQ3**: Can multimodal LLMs understand how rhetorical techniques in data visualizations contribute to misleading visualizations, and determine whether such rhetoric is used intentionally to mislead?

3.2 Visualization Rhetoric

We adopt the five core types of visualization rhetoric proposed by Hullman and Diakopoulos [8]: Information Access, Provenance, Mapping, Linguistic-Based, and Procedural Rhetoric. They are summarized in Tab. 1.

3.3 Experiments

We designed two experiments (**EXP1**, **EXP2**) to address the three research questions. Each experiment uses a distinct prompt (**PRT1**, **PRT2**), fully provided as supplemental material and schematized in the following. Both experiments address **RQ1**, while **EXP2** also targets **RQ2** and **RQ3**.

3.3.1 Experiment 1

The goal of EXP1 is to assess whether a multimodal LLM can recognize misleading visualizations (**RQ1**) without being explicitly informed that such visualizations can contain deceptive elements. PRT1 asks the LLM to examine a visualization and its accompanying caption (if any), to provide a detailed description of its content, and extract insights. This setup simulates the behavior of a lay user who is unaware that the visualization might be misleading and unfamiliar with specific misleading errors (e.g., dual axes, cherry-picking [12]). Unlike Alexander et al. [1], where the authors instructed the models to identify misleading aspects, providing their definitions as well, EXP1 evaluates the *baseline* capability of the model without advanced prompt-engineering strategies, such as few-shot learning. Since PRT1 does not directly ask whether the visualization is misleading, we derive this information from the model's analysis by post-processing its output. For this purpose, we pass the output of PRT1 to an independent control model (`gpt-oss-120b`), prompted with PRT1bis, which classifies whether the analysis indicates that the visualization is misleading. This separation ensures that the control model is not influenced by the results given by the model under analysis.

3.3.2 Experiment 2

The EXP2 addresses all three research questions. In contrast to EXP1, PRT2 explicitly alerts the model that the visualization may be misleading and asks it to: perform a detailed analysis of the visualization; assess whether the visualization is misleading; determine whether any misleading aspects are likely intentional or unintentional; and identify any rhetorical techniques present, following the five core categories of visualization rhetoric. This design allows PRT2 to address **RQ1** by detecting misleading elements, **RQ2** by identifying rhetorical techniques, and **RQ3** by assessing whether those rhetorical techniques are intentionally used to mislead.

3.4 Prompts

In the following, we schematize the used prompts (PRT1, PRT1bis, PRT2). The full prompts are provided as supplemental material.
PRT1 Goal: Guide the LLM to interpret a visualization and its caption, describe its content and message, and self-assess confidence. In particular, it asks to: 1. Analyze image and caption. 2. Describe visualization elements (axes, trends, data). 3. Interpret the intended message. 4. Provide conclusion. 5. Rate confidence of the analysis with justification.
PRT1bis Goal: Ask a second LLM to evaluate whether an LLM-generated analysis of a visualization reveals misleading elements or rhetorical strategies. . In particular, it asks to: 1. Determine if the described visualization is misleading. 2. Assess whether any misleading aspects are intentional. 3. Identify present types of visual rhetoric (it has been provided with definitions Tab. 1) 4. Report confidence for all judgments.
PRT2bis Goal: Perform both interpretation and critical assessment of the visualization, identifying misleading elements and rhetorical techniques. 1. Describe and interpret the visualization. 2. Draw conclusions and self-assess confidence. 3. Evaluate if the chart is misleading, why, and whether intentional or not. 4. Identify present types of visual rhetoric (it has been provided with definitions Tab. 1) and if they are used to support misleading. 5. Report confidence for all judgments.

3.5 Models

To differentiate our study from existing work, we evaluated two multimodal LLMs: `Gemma3:27B` [1] and `Qwen2.5-VL:72B` [2]. We selected these models to represent distinct scenarios, covering a relatively compact model (Gemma3) suitable for execution on consumer hardware, and larger models (Qwen2.5) requiring higher computational resources. As an independent control model for experiment 1, we used `gpt-oss-120b`, a recently released high-performance reasoning model. Our work is still ongoing, and we plan to extend the evaluation to larger models, such as `Llama`, `Mistral`, and `DeepSeek-VL`, among others.

3.6 Dataset

We created the COVID-Tweets misleading dataset, derived from the original dataset collected by Lisnic et al. [12]. The original dataset includes 9,958 English-language posts shared on X (formerly Twitter) during the COVID-19 pandemic, each containing at least one data visualization. Lisnic et al. annotated these tweets to indicate whether the visualization was misleading or not. For misleading cases, the authors

[1] https://ollama.com/library/gemma3:27b
[2] https://ollama.com/library/qwen2.5vl:72b

also specified the type of visualization design violation (e.g., truncated axis, unclear encoding) and/or reasoning error (e.g., cherry-picking, causal inference) that contributed to the misinterpretation of data. Among them, 2,373 tweets contain at least one misleading element, while the remaining 7,585 were categorized as not misleading. However, due to X's content policies, Lisnic et al. only shared tweet IDs rather than the full content or images. To construct a balanced dataset, we first attempted to retrieve all 2,373 misleading tweets.

We successfully collected 1,168 misleading tweets; the rest were unavailable due to removal, privacy settings, or technical issues. We then collected a random sample of 1,168 tweets from the non-misleading tweets. In total, our dataset is composed of **2,336 tweets** (50% misleading, 50% not misleading) In accordance with X's API Terms of Service, we only release the tweet IDs as supplemental material to this paper, which allows for the retrieval of content if the tweets remain publicly accessible.

4 Results

4.1 Findings for RQ1

The first research question asked whether multimodal LLMs can identify misleading elements in data visualizations. Overall, the results suggest that the used models are sensitive to misleading cues, but their performances are unbalanced. In general, they tend to overestimate the presence of misleading elements, resulting in high recall for "Misleading" cases but poor recognition of "Not Misleading" ones. This pattern was evident in the baseline setting of Exp1 and the guided setting of Exp2, although the precision, recall, and overall quality varied across models. While this indicates that these models have a baseline capacity to detect deception in visualizations, their judgments are weak and highly dependent on prompt design. The following sections present the results of each experiment in detail.

4.1.1 Experiment 1

The results of EXP1 (Tab. 2) show that both models (`gemma3:27b` and `qwen2.5vl:72b`) have a similar behavior: they are much more effective at identifying misleading visualizations than correctly classifying non-misleading ones. There is a relatively high recall for the "Misleading" class (0.89 for `gemma3:27b` and 0.90 for `qwen2.5vl:72b`), contrasted with very low recall for the "Not Misleading" class (0.25 and 0.26, respectively). When visualizations are misleading, both models were able to capture them with high sensitivity. However, they tended to be careful and label many correct visualizations as misleading. The confusion matrices (Tab. 3) provide further insight into this behavior. For both models, the number of false positives (FP) is considerable: `gemma3:27b` incorrectly classified 879 visualizations as misleading, while `qwen2.5vl:72b` misclassified 868. This over-prediction of misleading visualizations explains the relatively low precision values (0.54–0.55 for the misleading class), indicating that a significant fraction of the detected "misleading" cases were accurate. Overall, both models achieve a comparable performance, with F1-scores of 0.57 (`gemma3:27b`) and 0.58 (`qwen2.5vl:72b`), while `qwen2.5vl:72b` shows slightly better balance between precision and recall. These findings highlight the limitations of those LLMs when asked to analyze visualizations without explicit cues about possible deception. Their analyses are strongly biased toward over-identification of misleading content, which may require a better prompting strategy to achieve better results.

4.1.2 Experiment 2

When models are informed that visualizations might be misleading (EXP2), their performance diverges. The results in Tab. 4 show that `gemma3:27b` achieved very high recall for the "Misleading" class (0.99), meaning it correctly classifies almost all instances of misleading visualizations. However, this came at the cost of a high reduced precision (0.51) and a nearly complete inability to identify "Not Misleading" cases (recall = 0.03). As confirmed by the confusion matrix (Tab. 5), `gemma3:27b` classified almost all inputs as misleading, resulting in too many false positives, and a low overall F1-score (0.51). In contrast, `qwen2.5vl:72b` exhibited a more balanced behavior. For the "Not Misleading" class, it reached a recall of 0.46 and a precision of 0.62, while for the "Misleading" class, it achieved a recall of 0.72 and a precision of 0.57. The confusion matrix (Tab. 5) shows a better distribution between true negatives and true positives, despite false negatives and false positives still remaining influential. The overall F1-score of 0.59 indicates that `qwen2.5vl:72b` performs better than `gemma3:27b`.

Table 2: EXP1: Results

Model	Class	Precision	Recall	F1-score
`gemma3:27b`	Not Misleading	0.70	0.25	0.37
	Misleading	0.54	0.89	0.68
				0.57
`qwen2.5vl:72b`	Not Misleading	0.73	0.26	0.38
	Misleading	0.55	0.90	0.68
				0.58

Table 3: EXP1: Confusion matrices for `gemma3:27b` and `qwen2.5vl:72b`. Rows: actual labels, columns: predicted labels.

Actual	`gemma3:27b`		`qwen2.5vl:72b`	
	Pred: Not	**Pred: Mis**	**Pred: Not**	**Pred: Mis**
Not	TN = 289	FP = 879	TN = 300	FP = 868
Mis	FN = 125	TP = 1043	FN = 112	TP = 1056

Table 4: EXP2: Results

Model	Class	Precision	Recall	F1-score
`gemma3:27b`	Not Misleading	0.75	0.03	0.06
	Misleading	0.51	0.99	0.67
				0.51
`qwen2.5vl:72b`	Not Misleading	0.62	0.46	0.53
	Misleading	0.57	0.72	0.64
				0.59

Table 5: EXP2: Confusion matrices for `gemma3:27b` and `qwen2.5vl:72b`. Rows: actual labels, columns: predicted labels.

Actual	`gemma3:27b`		`qwen2.5vl:72b`	
	Pred: Not	**Pred: Mis**	**Pred: Not**	**Pred: Mis**
Not	TN = 38	FP = 1130	TN = 542	FP = 626
Mis	FN = 13	TP = 1155	FN = 330	TP = 838

4.2 Findings for RQ2

Since `qwen2.5vl:72b` performed better in classifying misleading tweets, we focused our analysis of **RQ2** and **RQ3** on this model. This analysis considers only visualizations that are misleading and correctly classified as misleading by the model. As shown in Fig. 1, the model detected all five rhetorical types, finding cases where a single rhetorical type was employed and visualizations where different types were combined. Interestingly, both *Procedural Rhetoric* and *Provenance Rhetoric* never appeared in isolation: the model always detected their usage in conjunction with other rhetoric types. The most frequent identified usage pattern is the combination of *Mapping Rhetoric*, *Linguistic-Based Rhetoric*, and *Information Access Rhetoric*, which appeared in 294 misleading tweets.

Conversely, only a single tweet (1263011902424219648 Fig. 2) employed all five rhetoric types simultaneously. Among them, only

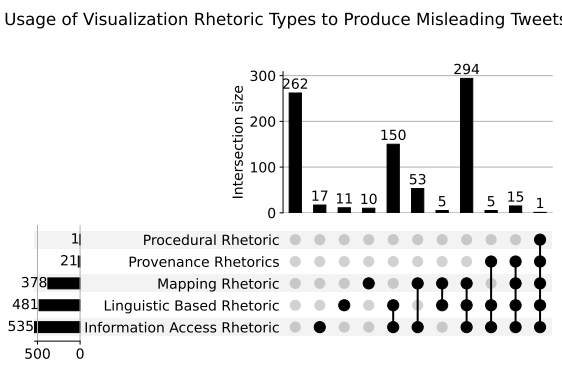

Figure 1: Usage of visualization rhetoric types identified in misleading tweets that have been correctly classified as misleading by `qwen2.5vl:72b`.

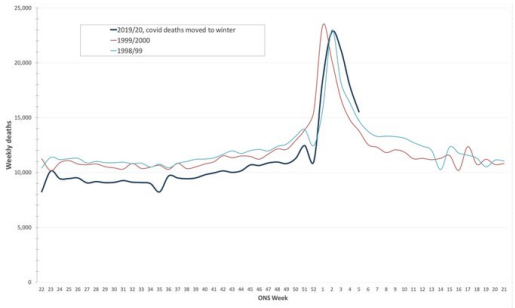

Figure 2: Tweet 1263011902424219648. `qwen2.5vl:72b` identified the usage of all five types of rhetoric. The model considers misleading this chart, providing this explanation: "*The graph may mislead by shifting COVID-19 deaths to the winter, which artificially inflates the peak and exaggerates the difference between the pandemic and other years*". The model also explained why it has found each type of rhetoric. Information Access Rhetoric: "*The graph includes COVID-19 deaths shifted to winter, which is a specific data manipulation choice*". Provenance Rhetoric: "*The graph does not provide sources, methods, or assumptions, which could enhance trustworthiness*". Mapping Rhetoric: "*The use of lines to represent weekly deaths and the manipulation of the COVID-19 data point to mapping rhetoric*". Linguistic-Based Rhetoric: "*The graph lacks titles, labels, or annotations that could guide interpretation*". Procedural Rhetoric: "*The graph is static and does not include interactive elements or default views*". Among them, the model considers only Mapping and Procedural Rhetoric as sources of the misleading content. In addition, the model considers as intentional the misleading: "*The manipulation of data to align with a specific narrative suggests an intentional effort to emphasize the pandemic's impact* ".

mapping and procedural rhetorics are considered to be generators of misleading visualization. While the motivation of Procedural Rhetorics seems adequate, it suggests that the model had to be informed that we are considering a static visualization.

These findings provide initial evidence that multimodal LLMs could go beyond binary classification of misleading visualization and cope with the rhetorical dimension of visual misinformation.

4.3 Findings for RQ3

In addition to detecting rhetorical strategies, the model (`qwen2.5vl:72b`) also provided short descriptions explaining why it identified a given rhetorical type in a visualization. This qualitative information is important, as it demonstrates that the model labels the visualizations and also attempts to justify its reasoning. For RQ3, the model has been asked to recognize the intentionality behind the use of

Figure 3: Tweet 1353286432656244737. `qwen2.5vl:72b` identified the usage of four types of rhetoric, all except Procedural Rhetoric. According to the model, the usage of the rhetoric to produce this misleading visualization **is intentional**. The model also explained why it has found each type of rhetoric. Information Access Rhetoric: "*The chart selectively includes data for specific companies and omits broader market data, which shapes the viewer's focus on these companies' performance. By omitting broader market data, the chart may mislead viewers about the relative performance of these companies*". Provenance Rhetoric: "*The chart does not provide information about data sources, methods, or assumptions, which is typical for such visualizations*". Mapping Rhetoric: "*The use of distinct colored lines and percentage change on the y-axis helps to visually compare the growth of different companies*". Linguistic Based Rhetoric: "*The caption 'COVID Misery Index, Winter Wonderland Edition' uses irony and playful language to frame the data in a particular way. The ironic caption may mislead viewers into thinking the chart represents overall economic misery rather than company growth*". Procedural Rhetoric: "*The chart does not include interactive elements or navigation features*".

rhetorical strategies in producing misleading visualizations. The results, presented in the upset plots Fig. 6 and Fig. 5 (see Appendix C), show a breakdown of the overall distribution previously illustrated in Fig. 1. It seems not to evidence a difference between the two cohorts, hinting at a similar usage of the combination of rhetorical means for both scenarios from a quantitative perspective. This also hints at the need to analyze these results qualitatively, looking at the confidence level and the textual reasoning provided to link the presence of misleading elements with the intent. On this line of thought, we preliminarily analyzed some cases. An illustrative example is reported in Appendix B (see Fig. 4), where four rhetorical types are present but, according to the model, their usage does not reflect an explicit intent to mislead. Conversely, Fig. 3 shows how the model interprets the rhetorical usage as deliberately attempting to conduct a misleading narrative. These findings indicate that the model can detect the presence of rhetorical strategies and their potential intent. While preliminary, this is an encouraging result for **RQ3**.

4.4 Qualitative analysis on Vis Lies data

To shed more light on qualitative aspects of the analysis on RQ3, we collected examples of real lies from the VIS Lies website (https://www.vislies.org/2024/gallery/) for the year 2024 (nine examples) and tested the `Qwen2.5-VL:72B` model on them. The complete analysis is reported in Table 6 in Appendix A. Among the examples, one returned an error, and eight executed correctly. Six were correctly labeled as misleading (True positives, TP), while two were considered not misleading (False negatives, FN). We also checked for both which rhetoric was used and how it related to the explanation provided on the website from visualization researchers initially proposing the lie. Interestingly, for the TP class, `Qwen2.5-VL:72B` was able to correctly identify 8 out of 11 (72.7% accuracy) cues for lying and mapping them to a visualization rhetoric aspect. The missing threes are primarily concerned with mapping issues (2 cases) and semantic relation to the message in the title (linguistics-based, 1 case). This spread was confirmed even for the FN cases. Of particular interest was Vis Lie 6,

which `Qwen2.5-VL:72B` labeled as a case of intentionally misleading information, citing *"The choice of time periods seems deliberate to emphasize the economic challenge, possibly to influence public perception"* relating it to linguistic and mapping-based rhetoric. Interestingly, the original researcher proposing it reports that *"Grocery prices were reported as a high contributing factor to the 2024 U.S. presidential election"*. In this case, the use of a model may have hinted at this aspect to a lay user, demonstrating a benefit in its usage. We plan to expand on this analysis by considering the whole set of Vis Lies examples.

5 Limitations and Conclusions

This work presented preliminary results of an evaluation of the capabilities of modern LLMs and VLMs to correctly interpret charts, being aware of potential errors when a layperson uses them to interpret charts. We further tested the capabilities of recognizing the intentionality of the errors and the rhetorical means used to produce the visual lie. Although some results are interesting, we report their preliminary nature as a limitation, in terms of both data and models tested. This makes the reported evidence anecdotal and still not generalizable. In fact, the reliance on a COVID-19 dataset imposes limitations, as such data could often reflect recurring rhetorical techniques specific to the pandemic context; using a broader, more diverse dataset could lead to more generalizable and robust findings. Additionally, the rhetorical means analysis needs further expansion to consider more subtle cases and test the capability to propose a more correct visual representation capable, by comparison, of visually conveying the misinforming parts. The work is ongoing, and we plan to continue to address these limitations.

Acknowledgments

This work was partially supported by the PRIN 2022ZLL7MW – "Conversational Agents: Mastering, Evaluating, Optimizing (CAMEO)" and by the MUR PRIN 2022 Project No. 202248FWFS "Discount quality for responsible data science: Human-in-the-Loop for quality data" within the NextGenerationEU Programme - M4C2.1.1.

References

[1] J. Alexander, P. Nanda, K.-C. Yang, and A. Sarvghad. Can gpt-4 models detect misleading visualizations? In *2024 IEEE Visualization and Visual Analytics (VIS)*, pp. 106–110, 2024. doi: 10.1109/VIS55277.2024.00029 1, 2

[2] N. Chen, Y. Zhang, J. Xu, K. Ren, and Y. Yang. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1301–1311, 2025. doi: 10.1109/TVCG.2024.3456320 1

[3] C.-W. Chiang, Z. Lu, Z. Li, and M. Yin. Enhancing ai-assisted group decision making through llm-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, p. 103–119. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3640543.3645199 1

[4] K. Choe, C. Lee, S. Lee, J. Song, A. Cho, N. W. Kim, and J. Seo. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):4712–4727, 2025. doi: 10.1109/TVCG.2024.3413195 1

[5] Y. Cui, L. W. Ge, Y. Ding, L. Harrison, F. Yang, and M. Kay. Promises and pitfalls: Using large language models to generate visualization items. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1094–1104, 2025. doi: 10.1109/TVCG.2024.3456309 1

[6] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do llms have visualization literacy? an evaluation on modified visualizations to test generalization in data interpretation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2025. doi: 10.1109/TVCG.2025.3536358 1

[7] K.-H. Huang, M. Zhou, H. P. Chan, Y. Fung, Z. Wang, L. Zhang, S.-F. Chang, and H. Ji. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In L.-W. Ku, A. Martins, and V. Srikumar, eds., *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 730–749. Association for Computational Linguistics, Bangkok, Thailand, Aug. 2024. doi: 10.18653/v1/2024.findings-acl.41 1

[8] J. Hullman and N. Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2231–2240, 2011. doi: 10.1109/TVCG.2011.255 1, 2

[9] M. S. Islam, R. Rahman, A. Masry, M. T. R. Laskar, M. T. Nayeem, and E. Hoque. Are large vision language models up to the challenge of chart comprehension and reasoning. In *EMNLP (Findings)*, pp. 3334–3368, 2024. 1

[10] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, and H. W. Park. Mdagents: An adaptive collaboration of llms for medical decision-making. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds., *Advances in Neural Information Processing Systems*, vol. 37, pp. 79410–79452. Curran Associates, Inc., 2024. 1

[11] M. Lisnic, C. Polychronis, A. Lex, and M. Kogan. Misleading beyond visual tricks: How people actually lie with charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3580910 1

[12] M. Lisnic, C. Polychronis, A. Lex, and M. Kogan. Misleading beyond visual tricks: How people actually lie with charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3580910 2

[13] L. Y.-H. Lo, A. Gupta, K. Shigyo, A. Wu, E. Bertini, and H. Qu. Misinformed by visualization: What do we learn from misinformative visualizations? *Computer Graphics Forum*, 41(3):515–525, 2022. doi: 10.1111/cgf.14559 1

[14] L. Y.-H. Lo and H. Qu. How good (or bad) are llms at detecting misleading visualizations? *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1116–1125, 2025. doi: 10.1109/TVCG.2024.3456333 1, 2

[15] S. Pandey and A. Ottley. Benchmarking visual language models on standardized visualization literacy tests. *Computer Graphics Forum*, 44(3):e70137, 2025. doi: 10.1111/cgf.70137 1

[16] L. Podo, M. Angelini, and P. Velardi. V-recs, a low-cost llm4vis recommender with explanations, captioning and suggestions, 2024. 1

[17] D. Raffini, A. Macori, L. Porcaro, T. Catarci, and M. Angelini. How persuasive could llms be? a first study combining linguistic-rhetorical analysis and user experiments. *arXiv preprint arXiv:2508.09614*, 2025. 1

[18] D. A. Szafir. The good, the bad, and the biased: five ways visualizations can mislead (and how to fix them). *Interactions*, 25(4):26–33, June 2018. doi: 10.1145/3231772 1

[19] Y. Tian, W. Cui, D. Deng, X. Yi, Y. Yang, H. Zhang, and Y. Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 31(3):1731–1745, 2025. doi: 10.1109/TVCG.2024.3368621 1

[20] P.-P. Vázquez. Are llms ready for visualization? In *2024 IEEE 17th Pacific Visualization Conference (PacificVis)*, pp. 343–352, 2024. doi: 10.1109/PacificVis60374.2024.00049 1

[21] F. Wang, B. Wang, X. Shu, Z. Liu, Z. Shao, C. Liu, and S. Chen. Chartinsighter: An approach for mitigating hallucination in time-series chart summary generation with a benchmark dataset. *IEEE Transactions on Visualization and Computer Graphics*, 31(6):3733–3745, 2025. doi: 10.1109/TVCG.2025.3567122 1

[22] Y. Wu, Y. Wan, H. Zhang, Y. Sui, W. Wei, W. Zhao, G. Xu, and H. Jin. Automated data visualization from natural language via large language models: An exploratory study. *Proc. ACM Manag. Data*, 2(3), May 2024. doi: 10.1145/3654992 1

[23] Y. Ye, J. Hao, Y. Hou, Z. Wang, S. Xiao, Y. Luo, and W. Zeng. Generative ai for visualization: State of the art and future directions. *Visual Informatics*, 8(2):43–66, 2024. doi: 10.1016/j.visinf.2024.04.003 1

[24] Z. Zhang, C. Shen, B. Yao, D. Wang, and T. Li. Secret use of large language model (llm). *Proc. ACM Hum.-Comput. Interact.*, 9(2), May 2025. doi: 10.1145/3711061 1

This is the author's version of the article that has been published in the proceedings of the "1st Workshop on GenAI, Agents, and the Future of VIS", at IEEE VIS 2025, Vienna, Austria. https://visxgenai.github.io/

APPENDIX A: DETAILED ANALYSIS OF VIS LIES 2024 EXAMPLES

Summary of Visual Lies Analysis (Info = Information access rhetoric; PRov. = Provenance rhetoric; Map. = Mapping rhetoric; Ling. = Linguistic-based rhetoric; Proc. = Procedural rhetoric).
The status column reports the decision of the Qwen model with its confidence between parentheses. Condition reports which elements of the visualization rhetoric supported misleading or not misleading decisions for all cases, with color coding representing the qualitative level of confidence (Dark green = confidence between 100% and 80%; Light green = confidence between 80% and 60%). Finally, the Recognized issues column reports details about the recognized explanations provided by the proposers of Vis Lies.

Table 6: Summary of Visual Lies Analysis (Info = Information access rhetoric; PRov. = Provenance rhetoric; Map. = Mapping rhetoric; Ling. = Linguistic-based rhetoric; Proc. = Procedural rhetoric. Dark green = conficence between 100% and 80%; Light green = confidence between 80% and 60%)

Visual Lie	Status	Condition	Info.	Prov.	Map.	Ling.	Proc.	Eval.	Recognized Issues
VL1: Inflated Bars	NOT recognized	supports misleading not misleading	- -	- -	- -	- -	- -	-	0/2 (returned an error)
VL2: Left Leaning	misleading (75%) NOT intentional (60%)	supports misleading not misleading	- -	- -	- -	- -	- -	TP	1/2 Capture y-axis, does not capture x-axis and relation to title
VL3: a plot of bubblies	misleading (80%) NOT intentional (70%)	supports misleading not misleading	- -	- -	- -	- -	- -	TP	2/2 Recognize both voronoi mapping errors
VL4: Lines outside drawing	NOT misleading (90%) NOT intentional (95%)	supports misleading not misleading	- -	- -	- -	- -	- -	FN	0/2 does not recognize rainbow colormap misuse and alignment to Gaussian curve
VL5: Mental health soup	NOT misleading (80%) NOT intentional (90%)	supports misleading not misleading	- -	- -	- -	- -	- -	FN	0/2 mostly mapping issues, but not definitely a lie. The model does not find it misleading.
VL6: Wage Imbalance	misleading (75%) intentional (65%)	supports misleading not misleading	- -	- -	- -	- -	- -	TP	1/1 recognizes the use of different time periods
VL7: Giants and Pixies	misleading (85%) NOT intentional (70%)	supports misleading not misleading	- -	- -	- -	- -	- -	TP	2/2 recognizes the confusing usage of silhouettes and the use of a relative scale
VL8: The Death of Music	misleading (75%) NOT intentional (60%)	supports misleading not misleading	- -	- -	- -	- -	- -	TP	1/3 recognizes effect of dates and sample size; does not recognize erroneous use of line chart and change in US life expectancy
VL9: Heat Share	misleading (90%) NOT intentional (80%)	supports misleading not misleading	- -	- -	- -	- -	- -	TP	1/1 recognizes erroneous usage of per capita measures

APPENDIX B: DETAILED EXAMPLES FOR RQ3

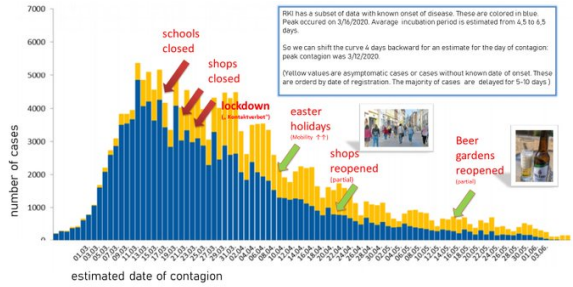

Figure 4: Tweet 1268617228468891648. `qwen2.5vl:72b` identified the usage of four types of rhetoric, all except Procedural Rhetoric. According to the model the usage of the rhetoric to produce this misleading visualization **is not intentional**. The model explained also why it has found each type of rhetoric. Information Access Rhetoric: "*The chart selectively includes data on symptomatic and asymptomatic cases, and marks specific events, shaping the viewer's focus on the impact of these events. The selective inclusion of data and events may oversimplify the complex factors influencing case numbers*". Provenance Rhetoric: "*The chart includes a note about the subset of data with known onset of disease and the estimated incubation period, providing some context for the data*". Mapping Rhetoric: "*The use of blue and yellow bars to differentiate between symptomatic and asymptomatic cases, and the placement of event markers, creates a visual narrative. The visual mapping may oversimplify the relationship between events and case numbers, leading to potential misinterpretation*". Linguistic Based Rhetoric: "*Annotations and labels guide the viewer's interpretation, emphasizing the impact of specific events on case numbers. The linguistic elements may lead viewers to overemphasize the direct impact of public health measures without considering other factors*". Procedural Rhetoric: "*The chart is static and does not include interactive elements or default views that could influence the viewer's exploration of the data*".

APPENDIX C: BREAKDOWN OF RHETORICAL MEANS

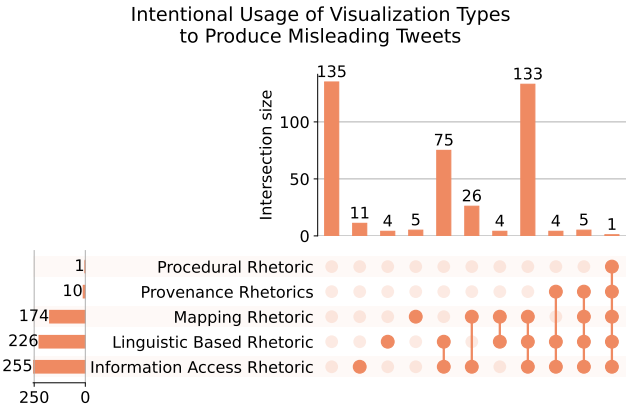

Figure 5: Usage of visualization rhetoric types identified in misleading tweets that have been correctly classified as misleading by `qwen2.5vl:72b`. According to the model, the usage of the rhetoric to produce the misleading visualization is intentional.

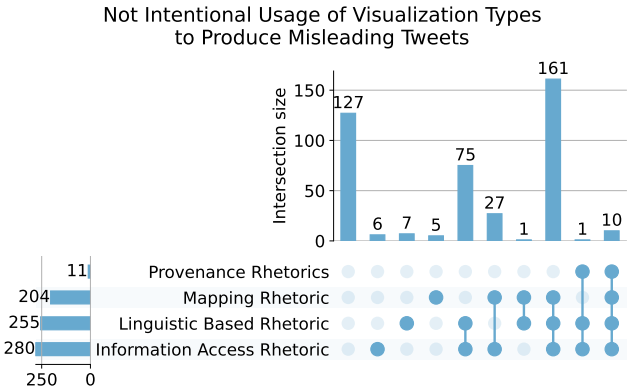

Figure 6: Usage of visualization rhetoric types identified in misleading tweets that have been correctly classified as misleading by `qwen2.5vl:72b`. According to the model the usage of the rhetoric to produce the misleading visualization is not intentional.