Eurographics Conference on Visualization (EuroVis) 2025
W. Aigner, N. Andrienko, and B. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 3

Benchmarking Visual Language Models on Standardized Visualization Literacy Tests

Saugat Pandey and Alvitta Ottley

Washington University in St. Louis, St. Louis, MO, USA

Abstract

The increasing integration of Visual Language Models (VLMs) into visualization systems demands a comprehensive understanding of their visual interpretation capabilities and constraints. While existing research has examined individual models, systematic comparisons of VLMs' visualization literacy remain unexplored. We bridge this gap through a rigorous, first-of-its-kind evaluation of four leading VLMs (GPT-4, Claude, Gemini, and Llama) using standardized assessments: the Visualization Literacy Assessment Test (VLAT) and Critical Thinking Assessment for Literacy in Visualizations (CALVI). Our methodology uniquely combines randomized trials with structured prompting techniques to control for order effects and response variability - a critical consideration overlooked in many VLM evaluations. Our analysis reveals that while specific models demonstrate competence in basic chart interpretation (Claude achieving 67.9% accuracy on VLAT), all models exhibit substantial difficulties in identifying misleading visualization elements (maximum 30.0% accuracy on CALVI). We uncover distinct performance patterns: strong capabilities in interpreting conventional charts like line charts (76-96% accuracy) and detecting hierarchical structures (80-100% accuracy), but consistent difficulties with data-dense visualizations involving multiple encodings (bubble charts: 18.6-61.4%) and anomaly detection (25-30% accuracy). Significantly, we observe distinct uncertainty management behavior across models, with Gemini displaying heightened caution (22.5% question omission) compared to others (7-8%). These findings provide crucial insights for the visualization community by establishing reliable VLM evaluation benchmarks, identifying areas where current models fall short, and highlighting the need for targeted improvements in VLM architectures for visualization tasks. To promote reproducibility, encourage further research, and facilitate benchmarking of future VLMs, our complete evaluation framework, including code, prompts, and analysis scripts, is available at `https://github.com/washuvis/VisLit-VLM-Eval`.

CCS Concepts

• ***Human-centered computing*** → ***Information visualization;***

1. Introduction

Large Language Models (LLMs) and Visual Language Models (VLMs) are emerging tools in data analysis and visualization, attracting considerable attention for their potential to address challenges faced by researchers and practitioners [MDW*23, HLL*24]. One of the most promising aspects of LLMs and VLMs is their ability to assist individuals with low vision, making visual data more accessible and understandable [LM22, SEN24]. They can also help mitigate information overload [TCD*24, WHB*24] or enable individuals who lack formal training to generate visualizations and engage with data through natural language, asking questions and receiving answers in a conversational manner [CLL*24, SS23, SEN24]. This democratization of data interaction can have a significant impact on how users interact with and interpret information.

Yet, while VLMs offer exciting prospects, there remain substantial questions about their capabilities, limitations, and reliability in visualization tasks, particularly when compared to human performance. Recent research activities in the visualization community are driven by the following questions: *To what extent can VLMs effectively interpret and reason about visual information? How do their visualization literacy and perception abilities compare to those of humans?* and *Can VLMs be trusted for use in complex data-driven environments?* These questions are essential because understanding VLMs' strengths and weaknesses could guide their application and help establish whether they are suitable for real-world use or require further development to meet needs.

To address these questions, recent work in the visualization community has examined VLM capabilities, including efforts to replicate foundational perceptual studies and explore applications such as natural language interaction with data, automatic captioning, visualization generation for research and educational purposes, and dataset creation for visualization tasks [GKS*24, Váz24, LQ24, CZX*24, WHB*24, KPR23, MS23]. These early studies offer valu-

able insights, but they also reveal significant limitations. First, some studies focus on single VLMs, overlooking the diversity in model architecture and training approaches. Second, while several methods have been proposed to assess people's visualization literacy [LKK16, GCK23, BMBH16, BRBF14, PO23, CLD*23], there is limited consensus on standardized benchmarks for VLM evaluation. Furthermore, VLMs exhibit unique challenges, such as a tendency to hallucinate or produce unreliable outputs that are highly sensitive to prompt phrasing and order effects, calling into question their robustness in visualization tasks [JLF*23, LYT*23].

To expand upon this crucial line of research, we present a comprehensive and detailed comparative analysis of four of the most recognized VLMs — GPT-4, Gemini, Claude, and Llama — using two standardized visualization literacy assessments: Visualization Literacy Assessment Test (VLAT) [LKK16] and Critical Thinking Assessment for Literacy in Visualizations (CALVI) [GCK23].

These assessments are meticulously designed to evaluate essential skills in reading, interpreting, and reasoning with various forms of visual representations, offering a nuanced understanding of each VLM's competencies. Additionally, we compare the performance of these VLMs against established human benchmarks to scrutinize how well these models respond to varying task types, adapt to different visualization styles, and interpret potentially misleading design elements. To ensure the robustness of our findings, we average the results over ten randomized trials to control for any order effects or prompt sensitivities. Through this detailed investigation, we aim to provide a clearer picture of VLMs' potential and limitations in advancing the data visualization field. We make the following contributions toward understanding the strengths and limitations of VLMs for visual data interpretation:

- We show that VLMs can accurately analyze a range of visual data formats. However, our findings also highlight that the effectiveness of these models varies significantly, influenced by the specific task, visualization type, and the models themselves. Among the models assessed, Claude stood out, demonstrating superior performance.
- Our comparative analysis with human performance reveals both promising capabilities and concerning limitations. VLMs approach or exceed human-level performance in specific tasks like trend identification (75-80% accuracy compared to 70% human average) and hierarchical structure detection (80-100% accuracy compared to 90% human average).
- However, while powerful, they require careful consideration before deployment in complex data-driven environments. Their strong performance in basic chart interpretation but poor reliability in detecting visualization deception (maximum 30.0% accuracy on CALVI compared to 39% human average) suggests they are better suited as assistive tools rather than autonomous systems.
- We provide a reproducible evaluation framework with randomized trials and structured prompting techniques to help assess future VLM capabilities.

2. Related Works

Large Language Models (LLMs) and Visual Language Models (VLMs) have transformed artificial intelligence by bridging textual and visual understanding. While LLMs excel in text processing, VLMs expand these capabilities through advanced transformer architectures [BPA*24], enabling sophisticated visual question answering [GLL*23, HXL*24], image captioning [LLWL24, CPG*23], and multimodal capabilities [LMX*22, Ope24]. These capabilities evolved from Vaswani et al.'s [Vas17] transformer architecture, progressing through GPT-1 [RNS*18] to GPT-3 [Bro20]. The field advanced further with frameworks like VisualBERT [LYY*19] and VilBERT [LBPL19], while models like CLIP [RKH*21] and Flamingo [ADL*22] demonstrated remarkable zero-shot capabilities in visual tasks.

However, these systems face limitations in visual reasoning and multi-modal learning. Key challenges include hallucinations — where models generate convincing but factually incorrect outputs — and sensitivity to prompt variations [JLF*23]. Training data bias presents another concern, as web-scale datasets can perpetuate biases and misinformation [BGMMS21]. Current research focuses on improving model transparency and reliability through adversarial training and debiasing algorithms [WDX*22, SS24].

2.1. VLMs in Visualization Research

Despite the limitations mentioned earlier, research at the intersection of VLMs and visualization has expanded rapidly in recent years, encompassing both the use of visualization techniques to understand and improve VLMs and the application of VLMs to advance visualization systems and tools. Our work contributes to this second research direction, where VLMs enable novel visualization capabilities and applications. By understanding these strengths and limitations, researchers and practitioners can better position LLMs and VLMs for real-world applications while addressing their inherent challenges.

Recent studies have shown that VLMs have a wide range of applications in visualization, from automated generation of visualization code and charts to sophisticated natural language interfaces for visual analytics systems [GKWK24, CZX*24, TCD*24]. These applications promise to enhance data literacy by making complex visualizations more accessible through natural language interaction and automated guidance [CLL*24]. However, this rapid adoption has paralleled an increasing focus on rigorous evaluation. Understanding their capabilities, limitations, and reliability becomes crucial as these models become more integrated into visualization systems. The scope and quick expansion of VLM applications in visualization underscore the importance of thorough, systematic evaluation approaches to ensure effective and responsible deployment in real-world scenarios.

2.2. Visualization Interpretation and Understanding in VLM Research

Most relevant to the current work, scholars have sought to explore the capabilities of VLMs for tasks related to visualization. There has been a series of recent studies addressing very similar research questions [BS24, GKS*24, Váz24, LQ24, CZX*24]. Most relevant to this work, Bendeck et al. [BS24] presented an empirical investigation of GPT-4's visualization literacy tasks using VLAT, examining performance across 8 different types of tasks across 12 vi-

sualization types. They demonstrated that while the model excels at trend identification and design best practices, it struggles with precise value retrieval and color discrimination, with GPT-4's overall performance in the 16th percentile compared to humans. Similarly, Guo et al. [GKS*24] investigated VLMs' perceptual capabilities through a series of graphical perception tasks, finding that VLMs can successfully replicate human perceptual judgments, particularly in tasks involving relative comparisons and trend analysis. While this might seem to contrast with Bendeck et al.'s findings, the difference lies in the type of tasks being evaluated - Guo et al. [GKS*24] focused specifically on perceptual judgment tasks. At the same time, VLAT encompasses a broader range of visualization interpretation skills.

Other work has sought to examine how well VLMs can identify misleading aspects of charts in addition to the basic ability to read and understand charts. Islam et al. [IRM*24] conducted a comprehensive evaluation of LVLMs across five major chart reasoning tasks, including chart question answering, summarization, and fact-checking. Their findings highlighted that while LVLMs demonstrate strong natural language generation capabilities, they also exhibit common issues such as hallucinations and data bias. In particular, Lo et al. [LQ24] investigated VLMs' ability to detect misleading visualizations through an exploratory evaluation approach using a dataset of 21 distinct chart issues. While their work showed promising potential for these models in supporting critical thinking during data interpretation, their open-ended methodology of explicitly asking about misleaders highlighted the need for more structured evaluation frameworks. In many real-world scenarios, identifying misleaders is not as straightforward. We take a different approach using CAVI [GCK23], which uses targeted questions addressing specific aspects of visual data with clearly defined response options, including an option to indicate when answers are impossible. This method facilitates a more natural identification of misleaders and accommodates the complexities inherent in interpreting visual information.

Our approach overcomes key limitations of earlier studies by controlling for order effects and prompt sensitivities through randomization. This establishes a reproducible framework for future evaluations as VLM technology continues to evolve. Our contribution is particularly timely, given the rapid advancement of VLMs and the growing need for reliable benchmarks in visualization research.

2.3. Assessment Frameworks for Visualization Literacy

The visualization community has developed several approaches for measuring and understanding visualization literacy. Early work by Börner et al. [BMBH16] pioneered the examination of public visualization literacy through familiarity-based questions about various data visualizations, revealing important insights about non-expert comprehension patterns. Boy et al. [BRBF14] established foundational methods using item response theory (IRT), though their work focused primarily on basic chart types like line graphs, bar charts, and scatterplots.

Building on these foundations, Lee et al. [LKK16] developed the Visualization Literacy Assessment Test (VLAT), offering comprehensive evaluation across 12 visualization types through 53 multiple-choice items. Mini-VLAT [PO23] later emerged as a more practical assessment tool while maintaining strong psychometric properties. Ge et al. [GCK23] advanced the field by developing CALVI specifically for assessing critical thinking about misleading visualizations, introducing systematic evaluation of "misleaders" through 45 validated items. Recent work by Cui et al. [CLD*23] has explored adaptive testing approaches, demonstrating comparable reliability with reduced question sets.

While these assessments have proven effective for evaluating human visualization literacy, there remains a critical gap in their application to VLMs. Despite the increasing use of VLMs in visualization tasks, there is a lack of standardized benchmarking and robust frameworks to assess these technologies across different visualization types and tasks. Our work addresses this gap by leveraging VLAT and CALVI as established benchmarks, extending their application to evaluate VLMs' visualization literacy.

3. Methodology

Our methodology addresses three fundamental questions raised in section 1 by investigating the extent of VLMs' visual interpretation abilities, their comparative performance against human benchmarks, and their reliability in real-world visualization tasks. This comprehensive evaluation framework employs standardized assessments to examine fundamental visualization comprehensive (ability to read and interpret) and critical thinking capabilities (ability to detect misleading visualizations) across different VLMs.

3.1. Model Selection and Configuration

Our study focuses on the state-of-the-art VLMs representing different visual-language understanding approaches. We selected these models based on several practical considerations: First, they offer stable, well-documented APIs that support consistent interaction patterns, which are crucial for reproducible research. Second, they can be deployed through cloud-based interfaces, eliminating the need for specialized hardware and making our evaluation framework accessible to a broader research community. Third, these models are widely used in practical scenarios, from data exploration to educational settings, making their evaluation particularly relevant for real-world applications. We excluded models such as Microsoft's CoPilot since it utilizes GPT-4 as its underlying model. Our final selection comprised four distinct VLMs, each representing different approaches to visual-language understanding:

- **GPT-4o**, developed by *OpenAI*, is an integrated visual and textual processing through a unified transformer architecture with attention mechanisms [Ope24].
- **Claude 3.5 Sonnet** is an AI assistant developed by *Anthropic* and employs constitutional AI principles to enhance reliability and minimize hallucinations [Ant].
- **Gemini 1.5 Pro**, previously known as *Bard* and developed by *Google DeepMind*, features end-to-end training on diverse visual-textual datasets, optimizing multimodal understanding [Dee].
- **Llama3.2-vision** is an open-source architecture with transparent visual processing capabilities developed by *Meta*. It offers models like the 11B and 90B variants that support image reasoning

tasks such as document-level understanding, captioning, and visual grounding [Met]. In this paper, we utilized the Llama 3.2 Vision 11B model.

We standardized several key configuration parameters across all models to ensure consistent and reproducible results. We set the *temperature* to 0, which minimizes response randomness by always selecting the most probable output tokens. This setting eliminates stochastic variation, making direct model comparisons more reliable. While alternative approaches—such as varying temperature settings—could provide insights into response variability, our goal was to establish a stable, deterministic benchmark. The *max_tokens* parameter was set to 300, limiting response length while ensuring sufficient detail for reasoning explanations. These settings prioritize deterministic behavior for our evaluation.

3.2. Assessment Framework

Our evaluation employs VLAT and CALVI, selected over alternatives like Mini-VLAT [PO23] and specialized tests [FDL20, FDWL22, CLD*23] for their comprehensive coverage of visualization tasks and chart types.

- **Visualization Literacy Assessment Test (VLAT)** provides a rigorous evaluation framework based on Classical Test Theory (CTT) [DeV06], which analyzes item performance through basic statistics like item difficulty and discrimination indices, and it comprises of 53 multiple-choice items across 12 visualization types. Originally validated with 200 participants, it includes an "Omit" option and employs a correction-for-guessing formula (Equation 1) [DE73, Fra88] to adjust scores based on incorrect responses. The test's complete-item approach enables direct comparison between human and VLM performance, with human scores typically ranging from 10.05 to 43.67 (M=28.82, SD=8.16) on the corrected scoring scale.
- **Critical thinking Assessment for Literacy in Visualizations (CALVI)** employs Item Response Theory (IRT) [ER13], which models the probability of correct responses based on both item properties and test-taker abilities to assess critical thinking through 45 items targeting various misleader types. The test consists of "trick" items using misleading and erroneous visualizations and "normal" items using well-formed visualizations inspired by VLAT. In its original validation study with 497 participants, each participant completed 30 items - 15 trick items randomly sampled from the bank of 45 items and 15 fixed normal items. The normal items serve as a baseline for assessing basic visualization interpretation abilities. Our study implements CALVI differently by having VLMs complete the full set of 45 trick items to enable comprehensive evaluation. In its original validation with human participants, performance on trick items ranged from 0% to 93% (M=39%, SD=16%). While this sampling approach complicates direct human-VLM comparisons, CALVI's systematic coverage of visualization deception makes it invaluable for assessing VLMs' critical analysis capabilities.

We chose these assessments' complementary strengths in evaluating different aspects of visualization literacy. Together, they provide a comprehensive framework for assessing both basic visualization interpretation skills and critical thinking abilities in the context of potentially misleading visualizations.

3.3. Prompt Engineering and Design

During prompt development, we began with a simple format: *"Please select the correct option(s) from the given choices. Respond with the chosen option number(s) followed by 'Why:' and then your explanation."* However, pilot testing revealed two key limitations: VLMs did not choose the "Omit" option unless explicitly prompted, and their unstructured responses made it challenging to extract choices and explanations systematically. These findings led to the development of more structured prompts, building upon work by Bendeck et al. [BS24], who evaluated GPT-4's visualization literacy capabilities.

As shown in Figure 1, we designed standardized prompts for both VLAT and CALVI assessments. The VLAT prompt explicitly instructs models to select answers based solely on chart information while discouraging guessing, with clear instructions about using the "Omit" option for uncertain responses. The CALVI [GCK23] prompt follows a similar format, emphasizing response organization and basing decisions purely on chart information.

These prompts were designed to elicit structured responses that facilitate systematic analysis while maintaining consistency with the original assessment frameworks. The explicit formatting instructions ensure that model responses can be automatically processed and evaluated across multiple runs. Additionally, the prompts emphasize the importance of basing answers solely on the provided visualizations, helping to isolate the models' visualization literacy capabilities from their broader knowledge base.

3.4. Evaluation Protocol

Each question is evaluated 10 times per model to account for potential variations in model responses. The questions are shown one at a time to prevent context contamination and to control the API rate limits. This also mitigates the potential interference between diverse visualization tasks while keeping consistent testing conditions across all models. The response collection follows a structured format, recording final answers and explanatory reasoning. For **VLAT**, we track the use of the "Omit" option as an indicator of model uncertainty. This data collection enables detailed analysis of model performance and decision-making patterns.

4. Visualization Literacy Assessment Results

Our evaluation results in Table 1 represent averaged performance across 10 independent runs for each model on both **VLAT** and **CALVI** assessment under the Random condition. We report average across 10 runs to provide stable performance estimates. This extends beyond prior work by Bendeck et al. [BS24], which used 3 runs. Our pilot studies comparing 5 versus 10 runs revealed significant variance in model outputs, justifying our choice of 10 runs for more reliable measurements. The mean values with standard deviation offer insights into both performance level and consistency, enabling meaning comparisons with prior works. This repeated evaluation approach is crucial because VLMs can produce varying responses to the same question across runs, even with temperature set to 0.

We first examine the corrected scores of both human and VLM

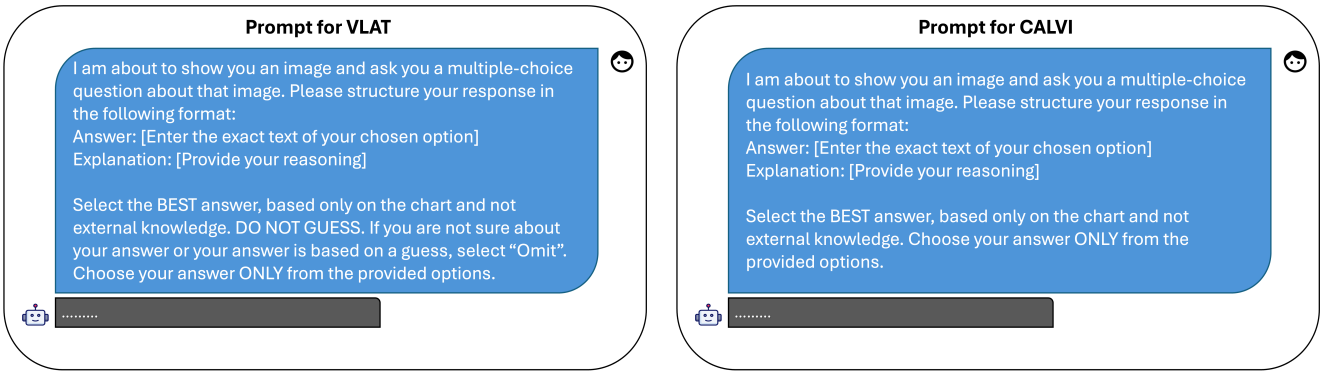

Figure 1: *VLAT (left) and CALVI (right) prompt templates used for VLM evaluation.*

performance on **VLAT**. The corrected score (*CS*) for each model was computed using the formula established by Lee et al. [LKK16]:

$$CS = R - \frac{W}{C-1} \tag{1}$$

where R represents the raw score (correct answers), W indicates incorrect answers, and C denotes number of choices available for each item. This formula adjusts for guessing by penalizing incorrect responses based on the number of possible choices. As shown in Table 2, our analysis encompasses 10 independent evaluation runs for each VLM, providing robust performance metrics compared against the original **VLAT** study's 191 human participants. Human participants achieved higher corrected scores (M=28.82) than most VLMs, though **CLAUDE** approached human-level performance with a corrected score of 28.96. Notably, VLMs showed consistent performance across runs, evidenced by lower standard deviations (0.77-2.67) than humans (8.16).

The results reveal variations in VLMs' ability to understand and interpret visualizations across both assessments. **CLAUDE** demonstrates superior performance in **VLAT** with an accuracy of 67.9%, outperforming other models (**GPT**: 49.8%, **GEMINI**: 42.5%, **LLAMA**: 43.8%). However, all models show notably lower performance on **CALVI**, with accuracies ranging from 21.8% (**CLAUDE**) to 30.0% (**GEMINI**). This considerable performance gap between **VLAT** and **CALVI** suggests that while VLMs have developed reasonable capabilities for basic visualization interpretation tasks, they struggle significantly with identifying and reasoning about misleading elements in visualizations.

Model	VLAT (%)		CALVI (%)	
	Mean	Std. Dev.	Mean	Std. Dev.
Human	65.5	13.3	39.0	16
CLAUDE	67.9	1.5	21.8	3.9
GPT	49.8	3.3	28.2	1.8
GEMINI	42.5	3.0	30.0	4.4
LLAMA	43.8	3.6	24.4	6.0

Table 1: *Average model performance and standard deviation across 10 runs compared to human baseline, under the Random condition where both questions and answer options were randomized.*

Table 2: *Comparison of Humans and VLMs on VLAT*

Model	Score Type	Mean (M)	Range	SD
Human [LKK16]	Regular	34.72	(14, 50)	7.05
	Corrected	28.82	(10.05, 43.67)	8.16
CLAUDE	Regular	36.00	(35, 37)	0.77
	Corrected	28.96	(27.55, 30.38)	1.10
GPT	Regular	26.40	(23, 29)	1.74
	Corrected	15.39	(10.58, 19.06)	2.47
GEMINI	Regular	22.50	(20, 26)	1.57
	Corrected	9.87	(6.34, 14.82)	2.21
LLAMA	Regular	23.20	(20, 26)	1.89
	Corrected	10.86	(6.34, 14.82)	2.67

4.1. Performance Across Chart Types

Model performance analysis across different chart types reveals distinct patterns in VLMs' visualization interpretation capabilities and their relationship to human performance, as shown in Figure 2. While humans maintain relatively consistent performance across visualization types, VLMs exhibit considerable variability. **CLAUDE** demonstrates superior performance across most chart types, achieving perfect accuracy (100%) on both 100% stacked bar charts and pie charts (compared to human performance of 80-90%), with exceptional performance on histograms (93.3%) and line charts (96.0%). All models show strong capabilities in line charts (**CLAUDE**: 96.0%, **GEMINI**: 80.0%, **GPT**: 62.0%, **LLAMA**: 76.0%).

However, performance degrades significantly with visualization complexity, particularly with bubble charts encoding multiple variables simultaneously (accuracy range: 18.6-61.4%). An interesting pattern emerges in spatial visualization interpretation: while models perform well on treemaps (**GPT**: 100%, **GEMINI**: 66.7%, **LLAMA**: 60.0%), they struggle with map-based visual-

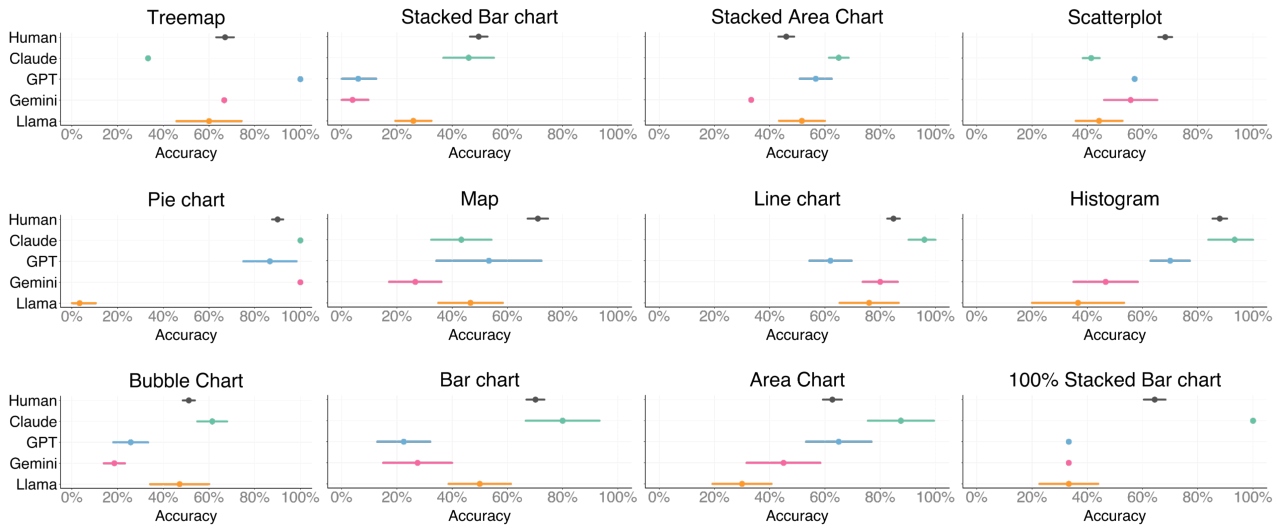

Figure 2: *Model performance comparison across different visualization types in* **VLAT**. *Each plot shows mean accuracy with 95% confidence intervals, comparing VLMs against human performance.*

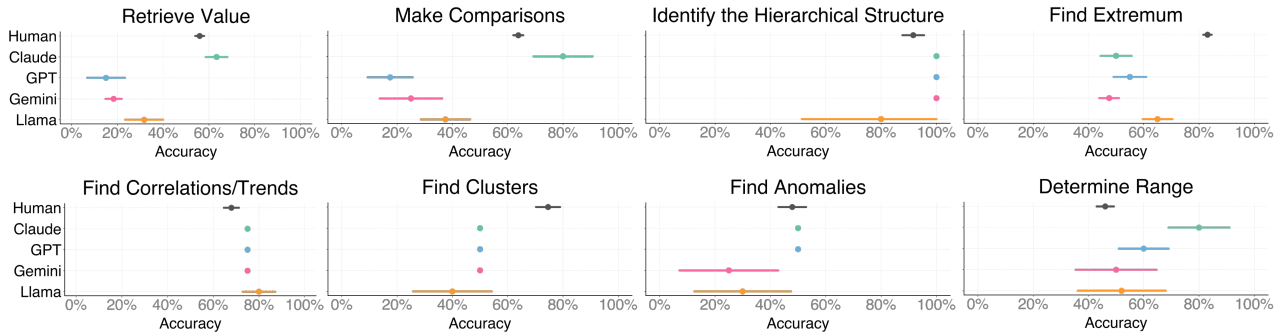

Figure 3: *Model performance comparison across different visualization tasks in* **VLAT**. *Each plot shows mean accuracy with 95% confidence intervals, comparing VLMs against human performance.*

izations (26.7-53.3%) compared to more consistent human performance. This disparity suggests stronger capabilities in processing hierarchical structures over geographic relationships. Basic visualization types like area charts and scatterplots show notable performance variation, highlighting inconsistencies in visual processing compared to the more uniform human performance patterns.

These patterns, i.e., excelling in structural analysis while struggling with precise numerical tasks and complex spatial relationships, align with broader findings in VLM research. Alnegheimish et al. [ANBEV24] found that deep learning models outperform VLMs by approximately 30% in anomaly detection. In contrast, Liu et al. [LHZ*24] demonstrated that VLMs require strategies like knowledge distillation for effective time series analysis.

4.2. Task-Specific Performance

The analysis of task-specific performance reveals nuanced patterns in how VLMs handle different visualization challenges, particularly compared to human benchmarks. Figure 3 gives an overview of these patterns on various visualization tasks and highlights surprising strengths of the VLM capabilities along with significant weaknesses.

VLMs have shown great ability in pattern recognition and understanding of structure in tasks. Hierarchical structure identification is a particular strength, with **CLAUDE**, **GEMINI**, and **GPT** achieving perfect accuracy (100%) and **LLAMA** maintaining strong performance (80%), all above typical human performance of 90%. Similarly, the models show consistent strength in finding correlations and trends at about 75-80%, which closely approaches human levels of 70%. This high performance in pattern-based tasks suggests that VLMs have developed robust capabilities in recognizing and interpreting systematic relationships within visualizations.

However, the performance terrain changes dramatically in tasks requiring precise numerical knowledge or more complex analytical reasoning. An analysis of value retrieval and evaluative tasks reveals significant disparities between models: while **CLAUDE** performs exceptionally well ($\approx$ 71%), even surpassing standard hu-

man performance, other models demonstrate considerably lower accuracy (LLAMA: 22.3%, GEMINI: 37.1%). This performance gap suggests that reasonable numerical interpretation is not inherent in VLMs but instead largely determined by specific architectural features or training methods.

Anomaly detection presents another revealing challenge, with performance ranging from 25% (GEMINI) to 50% (CLAUDE), compared to consistent human performance around 50%. This task, requiring both pattern recognition and deviation identification, seems to push the limits of current VLM capabilities. Interestingly, on range determination tasks, some VLMs even outperform humans, with CLAUDE achieving 80-90% accuracy, whereas human performance is 45%. This unexpected superiority in some quantitative tasks indicates possible advantages of computational approaches in specific analytical contexts.

The most pronounced human-VLM performance differences appear in tasks requiring the integration of multiple visual elements or precise numerical comprehension. While humans maintain relatively consistent performance across diverse task types, VLMs show marked variability, particularly in tasks demanding detailed analysis or complex inference. These findings align with broader research on VLM limitations in numerical reasoning and anomaly detection [ANBEV24], suggesting the need for specialized strategies like knowledge distillation to enhance performance [LHZ*24]. The observed patterns point to fundamental architectural limitations in processing complex numerical relationships, providing clear directions for future VLM development.

4.3. Uncertainty Analysis

Adding an "Omit" option in VLAT brings rich insights into how VLMs handle uncertainty in visualization interpretation tasks. Models exhibit distinct patterns in their uncertainty expression, as shown in Table 3. GEMINI has the highest propensity to acknowledge uncertainty, choosing to omit responses for 22.5% of questions (averaging 11.9 omissions per run). This conservative approach reflects a greater awareness of uncertainty compared to other models.

Table 3: *"Omit" response frequency by model across 10 runs, shown as counts and percentages of total VLAT questions.*

Model	Average Omits	Range	Percentage (%)
CLAUDE	4.3	4–6	8.1
GPT	3.8	2–6	7.2
GEMINI	11.9	11–14	22.5
LLAMA	4.3	3–6	8.1

5. Critical Thinking Assessment Results

A comparison of model performance across CALVI's misleader types reveals systematic patterns in VLMs' ability to detect visualization deceptions, as shown in Figure 4. All models detect missing normalization errors with 100% accuracy (except LLAMA at 10%). Models also demonstrate strong capabilities in detecting concealed uncertainty, with CLAUDE and GEMINI showing particularly high performance (100% and 90%, respectively). Interestingly, models show varied capabilities in specific misleader categories. GPT is good at detecting cherry picking (80%), while LLAMA shows superior performance in identifying misleading annotations (80%). However, all models struggle significantly with certain types of misleaders:

- **Missing Data**: All models consistently fail to identify missing data issues (0% accuracy across all models).
- **Scale Manipulations**: Performance is notably poor when dealing with inappropriate scale orders (12-18%) and inappropriate scale ranges (13.8-32.5%).
- **Overplotting**: Most models (CLAUDE, GPT, GEMINI) completely fail to detect overplotting issues (0% accuracy), with only LLAMA showing minimal capability (10%).

VLMs show an intriguing blind spot: while they excel at catching obvious chart errors like missing labels, they're surprisingly weak at spotting subtler forms of visual deception.

5.1. Model-Specific Capabilities

Analysis of the models reveals distinct patterns in their ability to detect specific visualization misleaders. CLAUDE demonstrates exceptional capabilities in identifying statistical and methodological flaws, achieving perfect accuracy (100%) in both concealed uncertainty and missing normalization categories. GPT exhibits particular strength in detecting cherry picking (80% accuracy), suggesting advanced capabilities in identifying selective data presentation bias. GEMINI shows balanced performance across categories with notable excellence in concealed uncertainty (90%) and missing normalization detection (100%). LLAMA , despite the lower overall performance, shows specialized expertise in detecting misleading annotations (80% accuracy), indicating potential in a contextual analysis of visualization elements. These diverse performance profiles suggest that architectural and training differences among models may foster specialized capabilities in detecting specific types of visualization deception. CLAUDE 's perfect accuracy in uncertainty detection likely stems from robust statistical reasoning capabilities, while GPT 's proficiency in identifying cherry-picking suggests advanced pattern recognition mechanisms.

5.2. Comparative Analysis with Human Performance

The comparative analysis of CALVI performance presents distinct methodological considerations that affect human-VLM comparisons. While our VLM assessment utilized CALVI's complete set of 45 misleader questions (achieving 21.8%-30.0% accuracy), the original human study followed a different protocol - approximately 500 participants each completed 30 items total: 15 trick items randomly sampled from the bank of 45 items (achieving M=39%, SD=16%) and 15 fixed normal items (achieving M=80%, SD=13%). This methodological distinction contrasts VLAT's approach, where all participants completed the whole set.

Analysis of performance across misleader types (Figure 4) reveals intriguing patterns in deception detection capabilities. VLMs show remarkable proficiency in certain areas, with CLAUDE, GPT,

Figure 4: *Comparison of VLM and human performance across different misleader types in* **CALVI** *assessment. Points show mean accuracy with 95% confidence intervals. MS = Manipulation of Scales.*

and **GEMINI** achieving perfect accuracy (100%) in missing normalization, significantly outperforming humans (10%). Similarly, for concealed uncertainty, **CLAUDE** and **GEMINI** demonstrate exceptional capabilities (100% and 90% respectively) compared to humans ($\approx$ 50%).

However, VLMs struggle with subtle manipulations. In detecting unconventional scale directions, humans maintain approximately 50% accuracy while VLMs perform poorly (8-18%). The challenges extend to missing data detection, where all models show 0% accuracy, and overplotting identification, where most VLMs fail. These limitations align with known constraints in processing fine-grained visual information [IRM*24] and numerical reasoning [ZGY*24].

Some misleader types reveal more nuanced patterns. In detecting misleading annotations, VLMs achieve varying success (50-80%), approaching human performance levels. **GPT**shows particular strength in cherry-picking detection (80% accuracy), exceeding human performance. These findings, supported by research on VLM capabilities [NJT24], suggest that while these models can effectively identify obvious visualization errors, they struggle with sophisticated forms of visual deception, requiring an integrated understanding of visualization principles and contextual reasoning. This performance gap emphasizes the continued importance of human oversight in critical visualization analysis.

5.3. Common Failure Patterns

Analysis across all models reveals systematic weaknesses in visualization deception detection, highlighting fundamental limitations in current VLM architectures. Most critically, all four VLMs demonstrate complete failure (0% accuracy) in identifying missing data issues, indicating a universal inability to detect omitted information in visualizations.

Scale manipulation detection emerges as another significant challenge. Models consistently struggle with identifying inappropriate scale ordering and ranges, with accuracy varying from 2.0% to 32.5%. This persistent weakness suggests fundamental limitations in evaluating visualization scaling decisions. Similarly, complex visualization issues, particularly overplotting detection, pose substantial challenges, with three models showing complete failure and **LLAMA** achieving minimal success (10% accuracy).

Function-related deceptions represent another area of universal difficulty, evidenced by consistently poor performance in detecting inappropriate scale functions (2.5 - 45.0%) and scale function manipulation (8.0 - 18.0%). These patterns point to three fundamental limitations in current VLM architectures: inadequate detail-oriented analysis capabilities [NJT24], limited mathematical reasoning [ZGY*24], and weak structural analysis skills [IRM*24]. These findings underscore critical areas requiring enhancement in future VLM development, particularly in strengthening capabilities for detailed visual analysis and mathematical relationship comprehension in data visualization.

6. Discussion

Our systematic evaluation of VLMs' visualization literacy offers insights for visualization researchers and practitioners. These findings illuminate the potential and limitations of current VLM architectures, providing a roadmap for future research and practical applications in visualization systems. Below, we expand on the key results, contextualize them within prior work, and propose avenues for development.

6.1. Comparing Claude's Visualization Literacy to Human Performance

Among the models we evaluated, **CLAUDE** demonstrated human-comparable performance on **VLAT**, marking a significant milestone in VLM capabilities. With an overall accuracy of 67.9%, **CLAUDE**

outperformed not only other VLMs but also approached human-level performance in several visualization types, including stacked area charts (85% vs 70% human accuracy), histograms (93.3% vs 75%), and line charts (96% vs 85%). However, without transparency regarding CLAUDE's architecture, attributing this success to specific technical innovations remains challenging.

Despite strong overall performance, CLAUDE showed consistent weaknesses in spatial encoding tasks (maps, scatterplots, treemaps), suggesting fundamental limitations in spatial reasoning common to current VLM architectures. This pattern indicates a hierarchical nature in VLMs' visualization literacy: strong performance in pattern recognition (75-80% accuracy in trend identification) but degrading capabilities as tasks require more complex reasoning or precise numerical analysis. Performance particularly declined in tasks requiring multiple variable interpretation, such as bubble charts (18.6-61.4% accuracy range across models).

This "complexity threshold" within current architectures—whether due to attention mechanisms, spatial encoding limitations, or training data—presents a crucial area for future research. Understanding these limitations could guide the development of hybrid models or specialized training approaches that better handle complex visualization tasks while maintaining the strong performance in basic pattern recognition.

6.2. Uncertainty Management Strategies

Our analysis reveals distinct patterns in how VLMs manage uncertainty during visualization tasks. Most notably, GEMINI's omission rate of 22.5% reflects a conservative approach to uncertainty, reducing false positions but potentially limiting insights in exploratory analysis where incomplete answers have value. This finding aligns with research on uncertainty quantification in AI [JLF*23] and suggests the need for models with dynamic risk tolerance adjustable to specific tasks and contexts. Future visualization systems could benefit from interfaces allowing users to calibrate model uncertainty thresholds based on their application needs.

6.3. Poor Performance in Detecting Misleading Visualizations

All models, including CLAUDE, demonstrated limited capabilities in identifying misleading visualizations, with CALVI accuracies ranging from 21.8% to 30.0%. This performance gap is particularly notable in detecting subtle manipulations such as inappropriate scale ordering (8-18% accuracy) and overplotting (0-10% accuracy). These findings contrast with prior work where VLMs showed better performance when explicitly prompted to evaluate misleading elements [LQ24], suggesting that task framing significantly influences performance.

The models showed striking disparities in their detection capabilities: while achieving near-perfect accuracy in identifying missing normalization (100% for most models) and concealed uncertainty (90-100% for top performers), they universally failed at detecting missing data (0% accuracy) and struggled with scale manipulations (13.8-32.5% accuracy). This pattern suggests that current VLMs excel at identifying obvious structural issues but struggle with more nuanced forms of visualization deception, highlighting a critical gap in their analytical capabilities that requires attention in future development.

6.4. VLM Strengths With Utility for Visualization Systems

The strong pattern recognition of VLMs make them valuable tools for initial data exploration and basic chart interpretation. Their high accuracy in trend identification (75-80%) and hierarchical structure detection (80-100%) suggests applications in educational settings, data journalism, or preliminary data analysis. However, their limitations in detecting misleading visualizations necessitate careful integration into practical applications.

We recommend a hybrid approach where VLMs handle initial analysis tasks while maintaining human oversight for critical decisions. Such an approach would employ VLMs for rapid initial pattern detection and trend summarization, while implementing confidence thresholds that trigger human review for complex or potentially misleading visualizations. Additionally, visualization systems should develop interfaces that clearly communicate model uncertainty and limitations to users. This strategy would leverage VLMs' computational efficiency while maintaining the critical thinking and context awareness that human analysts provide, particularly in scenarios where visualization misinterpretation could have significant consequences.

7. Limitations

Our study provides important insights into VLM visualization literacy capabilities, but several important limitations must be considered. The fundamental limitations begin with our assessment methodology. Although VLAT and CALVI are well-validated assessment instruments in the field of visualization literacy, their multiple-choice format may not capture the full range of VLM capabilities, potentially limiting the expression of more sophisticated reasoning processes [GCK23, Bro20]. These assessments also cannot evaluate important aspects of modern visualization practices, such as interactive visualization interpretation or dynamic data representations [HS12].

Our choices in model selection and configuration introduce another layer of constraints. We evaluated four prominent VLMs under specific parameter settings (temperature = 0), privileging consistency and reproducibility. However, this approach may not bring out the potential of these models under more flexible configurations, such as those employing stochastic sampling or incorporating human feedback mechanisms [Ope24, TLI*23]. Given the rapid development in VLM technology, newer architectures may mitigate some limitations we identified in visual processing capabilities.

A more fundamental challenge stems from the intrinsic differences between human and VLM approaches to visualization processing. While human analysts naturally incorporate domain expertise and holistic reasoning strategies, VLMs operate within more constrained computational frameworks [BCE*23]. This cognitive gap is particularly evident in CALVI tasks, where identifying subtle misleading elements requires critical thinking abilities that current VLM architectures struggle to replicate.

8. Future Work

Our findings reveal several promising research directions for advancing VLM capabilities in visualization interpretation:

Fine-tuning VLMs for Visualization Literacy Fine-tuning models on visualization-centric datasets could bridge the observed gaps in spatial reasoning and quantitative comprehension. The datasets should further include various types of visualizations, such as deceptive designs like overplotting, skewed baselines, and misleading scales, that would eventually train the model to recognize and respond to misleaders more precisely [PRS*15, LGS*22]. Techniques such as reinforcement learning with human feedback (RLHF) or curriculum learning could allow progressive improvements in handling complex visualizations [OWJ*22].

Leveraging Human-AI Synergy While VLMs exhibit clear strengths in pattern recognition and trend analysis, their weaknesses in tasks requiring numerical precision or critical evaluation necessitate hybrid frameworks. Such systems could use VLMs for preliminary tasks, such as identifying patterns or anomalies, and rely on human oversight for high-stakes decisions, such as identifying misleaders. This collaborative approach could leverage the respective strengths of both humans and VLMs to improve accuracy and reliability in visualization tasks [WKR*24].

Advancing Architectural Innovations Architectural advancements, such as vision transformers and multi-modal learning approaches, hold promise for addressing the limitations identified in our study [LZW*23]. Another promising avenue is the exploration of uncertainty quantification methods to make VLMs more robust in ambiguous or incomplete visualizations, as shown by recent research on probabilistic model outputs [APH*21].

Enhanced Prompt Engineering Strategies Our findings highlight how task framing impacts VLM performance in visualization analysis. Future research should explore sophisticated prompting strategies, such as chain-of-thought prompts or hierarchical task decompositions, to enhance VLMs' critical analysis [WWS*22]. This could include developing standardized prompt templates for different visualization tasks and investigating how varying levels of explicit instruction affect model performance in detecting misleading elements. While our study used standardized prompts to ensure fair comparisons, future work could explore more advanced strategies like multi-shot prompting, chain-of-thought reasoning, and model-specific prompt optimization [JMG*24]. These tailored approaches may provide deeper insights into model-specific strengths and enhance visualization literacy.

9. Conclusion

Our comprehensive evaluation of VLMs' visualization literacy capabilities reveals a complex landscape of advances and limitations. While some models approach human-level performance in basic visualization tasks (**CLAUDE** achieving 67.9% on **VLAT**), they struggle significantly with critical thinking and detecting visualization deception (21.8-30.0% on **CALVI**). These stark performance variations demonstrate that current VLMs are best suited as assistive tools rather than autonomous systems. Our reproducible framework provides a foundation for future VLM evaluations while emphasizing the importance of balanced human-AI collaboration in visualization analysis.

10. Acknowledgments

This project was partially supported by the National Science Foundation under OAC-2118201 and IIS-2142977. We thank Bum Chul Kwon and Sung-Hee Kim for sharing the data from the VLAT study. Lastly, we also thank the authors of CALVI for sharing the data in an open-source repository.

References

[ADL*22] ALAYRAC J.-B., DONAHUE J., LUC P., MIECH A., BARR I., HASSON Y., LENC K., MENSCH A., MILLICAN K., REYNOLDS M., ET AL.: Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems 35* (2022), 23716–23736. 2

[ANBEV24] ALNEGHEIMISH S., NGUYEN L., BERTI-EQUILLE L., VEERAMACHANENI K.: Large language models can be zero-shot anomaly detectors for time series? *arXiv preprint arXiv:2405.14755* (2024). 6, 7

[Ant] ANTHROPIC: claude 3.5 sonnet. URL: https://www.anthropic.com/claude/sonnet. 3

[APH*21] ABDAR M., POURPANAH F., HUSSAIN S., REZAZADEGAN D., LIU L., GHAVAMZADEH M., FIEGUTH P., CAO X., KHOSRAVI A., ACHARYA U. R., ET AL.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion 76* (2021), 243–297. 10

[BCE*23] BUBECK S., CHANDRASEKARAN V., ELDAN R., GEHRKE J., HORVITZ E., KAMAR E., LEE P., LEE Y. T., LI Y., LUNDBERG S., ET AL.: Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023). 9

[BGMMS21] BENDER E. M., GEBRU T., MCMILLAN-MAJOR A., SHMITCHELL S.: On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021), pp. 610–623. 2

[BMBH16] BÖRNER K., MALTESE A., BALLIET R. N., HEIMLICH J.: Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization 15*, 3 (2016), 198–213. 2, 3

[BPA*24] BORDES F., PANG R. Y., AJAY A., LI A. C., BARDES A., PETRYK S., MAÑAS O., LIN Z., MAHMOUD A., JAYARAMAN B., ET AL.: An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247* (2024). 2

[BRBF14] BOY J., RENSINK R. A., BERTINI E., FEKETE J.-D.: A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1963–1972. 2, 3

[Bro20] BROWN T. B.: Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020). 2, 9

[BS24] BENDECK A., STASKO J.: An empirical evaluation of the gpt-4 multimodal language model on visualization literacy tasks. *IEEE Transactions on Visualization and Computer Graphics* (2024). 2, 4

[CLD*23] CUI Y., LILY W. G., DING Y., YANG F., HARRISON L., KAY M.: Adaptive assessment of visualization literacy. *IEEE Transactions on Visualization and Computer Graphics* (2023). 2, 3, 4

[CLL*24] CHOE K., LEE C., LEE S., SONG J., CHO A., KIM N. W., SEO J.: Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. *IEEE Transactions on Visualization and Computer Graphics* (2024). 1, 2

[CPG*23] CHAN D., PETRYK S., GONZALEZ J. E., DARRELL T., CANNY J.: Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971* (2023). 2

[CZX*24] CHEN N., ZHANG Y., XU J., REN K., YANG Y.: Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics* (2024). 1, 2

[DE73] DIAMOND J., EVANS W.: The correction for guessing. *Review of educational research 43*, 2 (1973), 181–191. 4

[Dee] DEEPMIND G.: Google pro. URL: https://deepmind.google/technologies/gemini/pro/. 3

[DeV06] DEVELLIS R. F.: Classical test theory. *Medical care 44*, 11 (2006), S50–S59. 4

[ER13] EMBRETSON S. E., REISE S. P.: *Item response theory*. Psychology Press, 2013. 4

[FDL20] FIRAT E., DENISOVA A., LARAMEE R.: Treemap literacy: A classroom-based investigation. In *Eurographics Proceedings* (2020). 4

[FDWL22] FIRAT E. E., DENISOVA A., WILSON M. L., LARAMEE R. S.: P-lite: A study of parallel coordinate plot literacy. *Visual Informatics 6*, 3 (2022), 81–99. 4

[Fra88] FRARY R. B.: Formula scoring of multiple-choice tests (correction for guessing). *Educational measurement: Issues and practice 7*, 2 (1988), 33–38. 4

[GCK23] GE L. W., CUI Y., KAY M.: Calvi: Critical thinking assessment for literacy in visualizations. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (2023), pp. 1–18. 2, 3, 4, 9

[GKS*24] GUO G., KANG J. J., SHAH R. S., PFISTER H., VARMA S.: Understanding graphical perception in data visualization through zero-shot prompting of vision-language models. *arXiv preprint arXiv:2411.00257* (2024). 1, 2, 3

[GKWK24] GORNIAK J., KIM Y., WEI D., KIM N. W.: Vizability: Enhancing chart accessibility with llm-based conversational interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (2024), pp. 1–19. 2

[GLL*23] GUO J., LI J., LI D., TIONG A. M. H., LI B., TAO D., HOI S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023), pp. 10867–10877. 2

[HLL*24] HONG S., LIN Y., LIU B., WU B., LI D., CHEN J., ZHANG J., WANG J., ZHANG L., ZHUGE M., ET AL.: Data interpreter: An llm agent for data science. *arXiv preprint arXiv:2402.18679* (2024). 1

[HS12] HEER J., SHNEIDERMAN B.: Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue 10*, 2 (2012), 30–55. 9

[HXL*24] HU W., XU Y., LI Y., LI W., CHEN Z., TU Z.: Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), vol. 38, pp. 2256–2264. 2

[IRM*24] ISLAM M. S., RAHMAN R., MASRY A., LASKAR M. T. R., NAYEEM M. T., HOQUE E.: 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* (2024). 3, 8

[JLF*23] JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A., FUNG P.: Survey of hallucination in natural language generation. *ACM Computing Surveys 55*, 12 (2023), 1–38. 2, 9

[JMG*24] JEONG D. P., MANI P., GARG S., LIPTON Z. C., OBERST M.: The limited impact of medical adaptation of large language and vision-language models. *arXiv preprint arXiv:2411.08870* (2024). 10

[KPR23] KAVAZ E., PUIG A., RODRÍGUEZ I.: Chatbot-based natural language interfaces for data visualisation: A scoping review. *Applied Sciences 13*, 12 (2023), 7025. 1

[LBPL19] LU J., BATRA D., PARIKH D., LEE S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems 32* (2019). 2

[LGS*22] LO L. Y.-H., GUPTA A., SHIGYO K., WU A., BERTINI E., QU H.: Misinformed by visualization: What do we learn from misinformative visualizations? In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 515–525. 10

[LHZ*24] LIU C., HE S., ZHOU Q., LI S., MENG W.: Large language model guided knowledge distillation for time series anomaly detection. *arXiv preprint arXiv:2401.15123* (2024). 6, 7

[LKK16] LEE S., KIM S.-H., KWON B. C.: Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics 23*, 1 (2016), 551–560. 2, 3, 5

[LLWL24] LIU H., LI C., WU Q., LEE Y. J.: Visual instruction tuning. *Advances in neural information processing systems 36* (2024). 2

[LM22] LIEW A., MUELLER K.: Using large language models to generate engaging captions for data visualizations. *arXiv preprint arXiv:2212.14047* (2022). 1

[LMX*22] LU P., MISHRA S., XIA T., QIU L., CHANG K.-W., ZHU S.-C., TAFJORD O., CLARK P., KALYAN A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems 35* (2022), 2507–2521. 2

[LQ24] LO L. Y.-H., QU H.: How good (or bad) are llms at detecting misleading visualizations? *IEEE Transactions on Visualization and Computer Graphics* (2024). 1, 2, 3, 9

[LYT*23] LIU Y., YAO Y., TON J.-F., ZHANG X., CHENG R. G. H., KLOCHKOV Y., TAUFIQ M. F., LI H.: Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374* (2023). 2

[LYY*19] LI L. H., YATSKAR M., YIN D., HSIEH C.-J., CHANG K.-W.: Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019). 2

[LZW*23] LIU Y., ZHANG Y., WANG Y., HOU F., YUAN J., TIAN J., ZHANG Y., SHI Z., FAN J., HE Z.: A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems* (2023). 10

[MDW*23] MA P., DING R., WANG S., HAN S., ZHANG D.: Insightpilot: An llm-empowered automated data exploration system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (2023), pp. 346–352. 1

[Met] META: Llama can now see and run on your device - welcome llama 3.2. URL: https://huggingface.co/blog/llama32. 4

[MS23] MADDIGAN P., SUSNJAK T.: Chat2vis: generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *Ieee Access 11* (2023), 45181–45193. 1

[NJT24] NAGAR A., JAISWAL S., TAN C.: Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)* (2024), IEEE, pp. 1–8. 8

[Ope24] OPENAI: Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774, arXiv:2303.08774. 2, 3, 9

[OWJ*22] OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., ET AL.: Training language models to follow instructions with human feedback. *Advances in neural information processing systems 35* (2022), 27730–27744. 10

[PO23] PANDEY S., OTTLEY A.: Mini-vlat: A short and effective measure of visualization literacy. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, pp. 1–11. 2, 3, 4

[PRS*15] PANDEY A. V., RALL K., SATTERTHWAITE M. L., NOV O., BERTINI E.: 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 (2015), pp. 1469–1478. 10

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763. 2

[RNS*18] RADFORD A., NARASIMHAN K., SALIMANS T., SUTSKEVER I., ET AL.: Improving language understanding by generative pre-training. 2

[SEN24] STRÖBEL M., ECKERT K., NAGEL T.: Hey chatgpt, can you visualize my data?–a multi-dimensional study on using an llm for constructing data visualizations. 1

[SS23] SULTANUM N., SRINIVASAN A.: Datatales: Investigating the use of large language models for authoring data-driven articles. In *2023 IEEE Visualization and Visual Analytics (VIS)* (2023), IEEE, pp. 231–235. 1

[SS24] SHAH M., SUREJA N.: A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. *Archives of Computational Methods in Engineering* (2024), 1–13. 2

[TCD*24] TIAN Y., CUI W., DENG D., YI X., YANG Y., ZHANG H., WU Y.: Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics* (2024). 1, 2

[TLI*23] TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., ET AL.: Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023). 9

[Vas17] VASWANI A.: Attention is all you need. *Advances in Neural Information Processing Systems* (2017). 2

[Váz24] VÁZQUEZ P.-P.: Are llms ready for visualization? In *2024 IEEE 17th Pacific Visualization Conference (PacificVis)* (2024), IEEE, pp. 343–352. 1, 2

[WDX*22] WANG Z., DONG X., XUE H., ZHANG Z., CHIU W., WEI T., REN K.: Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10379–10388. 2

[WHB*24] WANG H. W., HOFFSWELL J., BURSZTYN V. S., BEARFIELD C. X., ET AL.: How aligned are human chart takeaways and llm predictions? a case study on bar charts with varying layouts. *IEEE Transactions on Visualization and Computer Graphics* (2024). 1

[WKR*24] WANG X., KIM H., RAHMAN S., MITRA K., MIAO Z.: Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–21. 10

[WWS*22] WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D., ET AL.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35* (2022), 24824–24837. 10

[ZGY*24] ZOU C., GUO X., YANG R., ZHANG J., HU B., ZHANG H.: Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836* (2024). 8