

From Benchmarks to Real-World Use: Evaluating the Robustness of Vision-Language Models for Chart Understanding

Awal Ahmed Fime^{1*} Mohammed Saidul Islam^{2*} Md Tahmid Rahman Laskar³
Md Zarif Hossain¹ Ridwan Mahbub³ Mir Tafseer Nayeem⁴ Mizanur Rahman³
Ahmed Imteaj¹ Enamul Hoque Prince³

¹Florida Atlantic University ²Vector Institute ³York University ⁴University of Alberta

Abstract

Recent advances in large Vision-Language Models (VLMs) have led to impressive performance on chart understanding tasks. However, existing evaluations are largely limited to clean, canonical charts and do not adequately measure robustness under realistic conditions. To address this gap, we introduce ROBUSTCHART360, a comprehensive benchmark for evaluating chart understanding robustness under realistic real-world variations. ROBUSTCHART360 spans six robustness dimensions covering textual and cross-modal interference, rendering and display degradation, accessibility-oriented transformations, and visual style variation. Using this benchmark, we conduct a large-scale evaluation of 14 proprietary and open-source VLMs across 2,505 chart instances and multiple robustness conditions designed to simulate practical deployment scenarios. Our analysis on 35,070 model responses reveals that current VLMs are highly sensitive to seemingly minor variations such as low-resolution rendering, and mobile layouts, with performance dropping by as much as 33.9% under certain conditions, such as modality gap. We further find that models often over-rely on textual priors when visual and textual signals conflict, indicating weak visual grounding despite strong performance on standard benchmarks. Our findings highlight a substantial gap between benchmark accuracy and real-world robustness, motivating the development of more reliable and visually grounded multimodal systems for chart understanding. We will release our code and dataset at <redacted>.

1 Introduction

Natural language and data visualization are two complementary modalities for communicating insights effectively (Hoque et al., 2022). Visualizations enable users to identify trends, patterns, and anomalies, while natural language complements

them by summarizing key findings, explaining relationships, and answering data-related queries. This synergy has inspired a growing body of research on chart comprehension and reasoning tasks, including chart question answering (Masry et al., 2022; Kantharaj et al., 2022), chart-to-text generation (Obeid and Hoque, 2020; Shankar et al., 2022), fact-checking with charts (Akhtar et al., 2023b), and data storytelling (Islam et al., 2024a).

Recent advances in vision-language models (VLMs) have enabled their adoption in chart-centric applications such as business intelligence assistants, accessibility tools, automated reporting systems, and data-analysis agents (Hoque and Islam, 2024). Despite strong performance on benchmarks such as ChartQA, benchmark accuracy alone does not guarantee reliable behavior in practice. Real-world chart understanding often involves noisy or conflicting textual context, diverse user prompts, low-resolution screenshots, mobile displays, and varying presentation styles—conditions that are rarely captured by existing evaluations. Moreover, VLMs remain susceptible to hallucinations, language biases, and failures on more challenging chart reasoning tasks (Cui et al., 2023; Guan et al., 2024; Islam et al., 2024b; Masry et al., 2025a). Although chart robustness has recently received attention (Mukhopadhyay et al., 2024), little is known about how current models behave under the variations commonly encountered in practical analytical workflows. This raises a fundamental question: *Do modern VLMs actually make predictions based on chart evidence, or are their predictions driven by surrounding textual cues and presentation artifacts?*

To answer this question, we introduce ROBUSTCHART360, a benchmark for evaluating chart understanding under realistic visualization variations. The benchmark spans two classes of challenges commonly encountered in chart-centric applications (see Fig. 1). *Textual and cross-*

* Equal contributions.



Figure 1: Overview of ROBUSTCHART360. The benchmark comprises six robustness tasks spanning textual and cross-modal robustness (TC1–TC3) and display and presentation robustness (DP1–DP3).

modal robustness evaluates whether models remain grounded in visual evidence when prompts, captions, or contextual information vary or conflict with the chart. *Display and presentation robustness* evaluates reliability under realistic viewing conditions, including low-resolution charts, mobile layouts, and presentation-style variations. Using ROBUSTCHART360, we conduct a large-scale evaluation of 14 leading proprietary and open-source VLMs to quantify robustness, identify the factors of failure, and characterize the gap between benchmark performance and real-world reliability.

Our study makes three main contributions. (1) We introduce ROBUSTCHART360, a benchmark for evaluating chart understanding under six realistic robustness challenges spanning textual and cross-modal interference, display degradation, mobile viewing conditions, and presentation variations. (2) We conduct a large-scale evaluation of 14 proprietary and open-source VLMs across 2,505 chart instances, revealing how various visual and contextual variations affect model behavior. (3) Through comprehensive analysis, we reveal a sub-

stantial gap between benchmark performance and real-world reliability. We found that when visual and textual signals disagree, many leading VLMs follow textual cues rather than chart evidence, exposing a substantial gap in visual grounding and deployment readiness.

2 Related Work

Robustness in Vision-Language Models: Recent closed-source models (OpenAI, 2026; Comanici et al., 2025; Anthropic, 2025) achieve state-of-the-art performance on chart understanding benchmarks (Masry et al., 2022; Wang et al., 2024), while open-source alternatives are rapidly improving (Bai et al., 2025; Wang et al., 2025; Abdin et al., 2024). Despite these advances, VLMs remain vulnerable to hallucinations and failures under visual and multimodal variations (Guan et al., 2024; Rahmzadehgeri et al., 2024; Leng et al., 2024). Prior work shows that model predictions can be highly sensitive to prompt phrasing, instruction style, and language formality (Jin et al., 2022; Zhang et al., 2026; Ramu et al., 2025; Zhang et al., 2025; Jiang

et al., 2026; Sarch et al., 2024), where semantically equivalent prompts can yield substantially different predictions. Other works investigate modality-gap robustness and multimodal consistency, showing that models often over-rely on textual context when visual and textual signals conflict (Leng et al., 2024; Huang et al., 2024b,a; Hsu et al., 2026). Robustness issues have also been studied under adversarial perturbations, OCR corruption, background and contextual shifts, and visual degradations (Xiao et al., 2021; Li et al., 2023; Ryali et al., 2021). While robustness has been extensively studied for general VLMs, its implications for chart understanding remain less explored.

Robust Chart Understanding: Chart understanding has attracted attention through the development of many evaluation benchmarks (Tang et al., 2023; Shankar et al., 2022; Akhtar et al., 2023a; Masry et al., 2022, 2025a). Recent work has further improved chart reasoning through instruction tuning and reinforcement learning (Masry et al., 2024, 2025b). Despite this progress, existing evaluations largely assume clean charts with consistent layouts, high-resolution rendering, and reliable textual context, while studies have shown that chart-understanding models remain susceptible to factual errors, hallucinations, bias and deceptions in more realistic settings (Huang et al., 2023; Islam et al., 2024b; Mahbub et al., 2025b,a). Most closely related, Mukhopadhyay et al. (2024) studied robustness under visual perturbations, evaluating whether models remain invariant to alternative renderings of the same chart. In contrast, we extend robustness evaluation beyond visual rendering to include prompt variation, cross-modal conflicts, low-resolution displays, mobile layouts, and presentation variations commonly encountered in real-world chart consumption.

3 The ROBUSTCHART360 Benchmark

In this section, we describe the ROBUSTCHART360 construction pipeline (Figure 4), consisting of four stages: (i) defining the robustness taxonomy, (ii) collecting chart corpora, (iii) creating verified task-specific Q&A pairs, and (iv) evaluating VLMs on the resulting benchmark. Additional details are provided in Appendix §A.

3.1 Taxonomy Construction

We organize ROBUSTCHART360 around two common sources of robustness failures in real-world

chart understanding systems (Figure 1 & Table 3): changes in surrounding textual context and changes in how charts are rendered and viewed.

Textual and Cross-Modal Robustness (TC) evaluates whether models remain grounded in chart evidence when textual information changes or conflicts with the visualization. Prior work shows that chart captions and visualizations frequently emphasize different aspects of the data in real-world documents (Kim et al., 2023), creating opportunities for inconsistency between visual and textual channels. Motivated by these deployment scenarios, we evaluate three conditions **TC1: Caption/Annotation Mismatch**, where accompanying text contradicts chart content; **TC2: Input Prompt Variation**, where semantically equivalent questions are expressed using different styles and expertise levels; and **TC3: Modality Gap**, which tests whether models rely on chart evidence or textual priors when visual and textual signals disagree. **Display and Presentation Robustness (DP)** evaluates reliability under realistic viewing and rendering conditions that do not alter the underlying data. This category includes: **DP1: Low-Resolution Plots**, which simulate compressed screenshots and degraded displays; **DP2: Mobile Variants**, which test chart interpretation under small-screen layouts; and **DP3: Background Theme Variation**, which evaluates robustness to presentation changes such as light, dark, and gray themes. These conditions are important because charts are frequently consumed through screenshots, reports, and mobile devices, where resolution, layout, and presentation can vary substantially (Wu et al., 2020).

3.2 Chart Corpus Collection

For each task, we construct realistic variants of reference charts while preserving their semantics and intended answer. This enables controlled evaluation under real-world variations. The only exception is *input prompt variation* (TC2), where the chart is fixed and robustness is evaluated through semantically equivalent question reformulations.

Textual and Cross-Modal Variants. We use ChartQA (Masry et al., 2022), a widely-used real world benchmark as the source for charts. These tasks isolate robustness to textual inconsistencies, linguistic variation, conflicting multimodal signals, and rendering degradation while preserving the underlying chart content and ground-truth answers. For TC1 and TC3, we introduce controlled text-based inconsistencies: TC1 targets caption, legend,

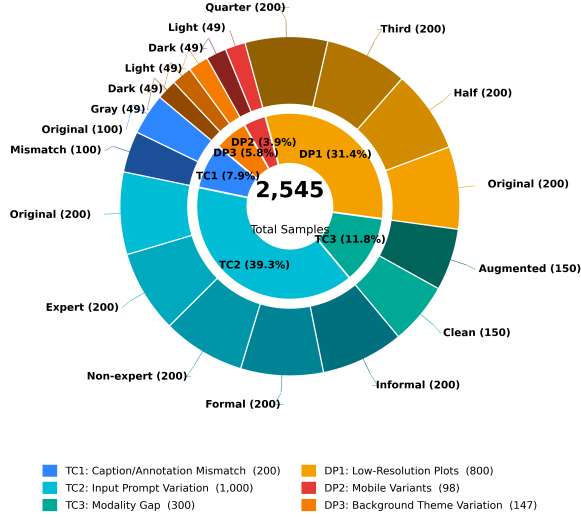


Figure 2: Dataset summary by task and subcategory.

and annotation mismatches, while TC3 introduces conflicts between textual and visual cues.

Display and Presentation Variants. For DP1, we downscale collected charts using bilinear interpolation to 50%, 33.3%, and 25% of the original width and height, corresponding to $\frac{1}{4}$, $\frac{1}{9}$, and $\frac{1}{16}$ of the original pixel count. For DP2 and DP3, we build a separate chart corpus because ChartQA’s rendered images and limited chart-type diversity do not support controlled layout or theme manipulation. Using an iterative human-in-the-loop process with *Claude-Opus-4.7* (Anthropic, 2026), we generate paired desktop/mobile layouts and light/dark/gray themes while preserving the underlying data, chart semantics, and answers; all variants are manually verified and refined rigorously.

3.3 Q&A Pair Creation

We construct Q&A pairs with task-specific procedures, separating ChartQA-derived tasks from programmatically generated presentation variants.

Textual and Cross-Modal Variants. For TC1, Q&A pairs are created through human–LLM collaboration and manually verified to ensure that answers are grounded in the chart rather than inconsistent text. For TC2, we use *Claude-Opus-4.7* to generate four semantically equivalent variants of each original ChartQA question: *expert*, *non-expert*, *formal*, and *informal*. These variants capture technical, lay, professional, and conversational phrasing, respectively. Since each selected chart contains two original questions, this yields 10 samples per chart. For TC3, we add a subtle but misleading caption and formulate questions around its

claim, while deriving ground-truth answers solely from the chart. Models that follow the caption are penalized, whereas those that rely on the chart are considered visually grounded.

Display and Presentation Variants. For DP1, we reuse the original ChartQA questions, while for DP2 and DP3, which are not ChartQA-derived, we generate new Q&A pairs through an LLM-assisted pipeline. To reduce generator-specific bias and increase linguistic and reasoning diversity, we prompt GPT-5.4 (OpenAI, 2026), Gemini-3.1-Pro (Google DeepMind, 2026), and Claude-Opus-4.6 (Anthropic, 2026) to generate 5 candidate questions per chart, and manually select the best one.

3.4 Dataset Composition & Diversity

ROBUSTCHART360 is designed to capture the robustness challenges that chart-understanding systems face in real-world deployment. It comprises **2,505 samples** spanning six tasks and 23 subcategories across two complementary dimensions: *Textual and Cross-Modal* (58.9%), which evaluates robustness to linguistic variation and conflicting multimodal signals, and *Display and Presentation* (41.1%), which evaluates robustness to changes in rendering quality, viewing conditions, and visual presentation (Figure 2). Beyond these robustness settings, ROBUSTCHART360 provides substantial visual and linguistic diversity. It covers a wide range of chart types, including bar, line, pie, donut, scatter, bubble, area, grouped-bar, and stacked-bar charts (Table 4), as well as diverse question formulations, vocabulary, and reasoning styles (Table 5). Together, these characteristics make ROBUSTCHART360 a realistic testbed for evaluating chart-understanding systems under deployment-time semantic and perceptual variations. More details are provided in §A.4.

4 Method

Models: We evaluate robustness of both closed and open-source VLMs under chart presentation shifts. Proprietary models include GPT-5 and GPT-5-mini (OpenAI, 2025), Claude-Sonnet-4.5 (Anthropic, 2025), and Gemini-2.5-Pro/Flash (Cohan et al., 2025). In addition, we evaluate recent open-source VLM families across varied parameter scales, including Qwen3-VL (2B, 4B, 8B), InternVL3.5 (2B, 4B, 8B), and Gemma3 (4B, 12B, 27B) (Bai et al., 2025; Wang et al., 2025; Gemma Team, 2025). This setup enables compar-

Model	TC1: Caption Mismatch			TC2: Input Prompt Variation							TC3: Modality Gap		
	Orig.	Mismatch	Δ_{rob}	Orig.	Expert	Non-exp.	$\Delta_{\text{rob}}^{\text{E/N}}$	Formal	Informal	$\Delta_{\text{rob}}^{\text{F/I}}$	Clean	Aug.	Δ_{rob}
GPT-5	80	51	29	85.0	85.0	80.5	4.5	84.5	80.0	4.5	32.67	26.00	6.67
GPT-5-mini	78	47	31	85.0	84.0	86.5	0.0	87.0	84.0	3.0	41.33	29.33	12.00
Claude-Sonnet-4.5	73	45	28	69.0	74.0	69.5	4.5	73.5	70.5	3.0	62.54	41.34	21.20
Gemini-2.5-Pro	77	46	31	87.0	88.0	85.5	2.5	87.0	87.0	0.0	55.44	34.65	20.79
Gemini-2.5-Flash	69	42	27	78.0	80.0	80.0	0.0	81.0	78.5	2.5	46.80	32.65	14.15
Gemma3-27B	56	36	20	58.0	62.0	59.0	3.0	59.5	58.0	1.5	43.30	26.67	16.63
Gemma3-12B	50	29	21	54.0	58.0	52.5	5.5	55.0	55.5	0.0	30.00	24.00	6.00
Gemma3-4B	35	13	22	40.5	40.0	40.0	0.0	42.5	37.5	5.0	50.67	41.33	9.34
InternVL3.5-14B	65	38	27	82.0	78.5	77.5	1.0	79.5	80.0	0.0	62.67	50.00	12.67
InternVL3.5-8B	62	36	26	81.0	79.0	77.0	2.0	81.0	79.5	1.5	48.00	40.00	8.00
InternVL3.5-4B	63	34	29	80.0	78.5	76.0	2.5	79.5	78.0	1.5	55.33	42.67	12.66
Qwen3-VL-32B	72	41	31	83.5	81.0	82.5	0.0	83.0	80.5	2.5	58.00	51.20	6.80
Qwen3-VL-8B	65	38	27	75.5	77.5	73.5	4.0	77.0	76.5	0.5	63.33	42.67	20.66
Qwen3-VL-4B	57	27	30	73.5	74.5	72.5	2.0	73.5	74.0	0.0	60.00	45.33	14.67

Table 1: Textual and cross-modal robustness results on ROBUSTCHART360. All values are reported as percentages. Lower Δ_{rob} indicates stronger robustness. Green and red highlight the best and worst robustness gaps within each model group.

isons across access type, model family, and scale.

Evaluation Metric: Following Masry et al. (2022), we use relaxed exact match accuracy for all tasks to determine whether a model prediction is equivalent to the reference answer.

Robustness Gap. To quantify performance degradation under perturbations, we report a unified robustness gap:

$\Delta_{\text{rob}}(m) = A_{\text{ref}}(m) - \min(A_{\text{ref}}(m), A_{\text{min}}(m))$, where $A_{\text{ref}}(m)$ denotes accuracy under the reference condition and $A_{\text{min}}(m)$ is the lowest accuracy among its corresponding variants. This metric provides a compact summary of a model’s vulnerability to presentation shifts, with smaller values indicating stronger robustness ($\Delta_{\text{rob}} = 0$ indicating no degradation).

5 Results and Analysis

5.1 Key findings

Benchmark success does not imply deployment readiness. Tables 1 and 2 reveal substantial robustness gaps across all evaluated models despite strong performance on clean benchmark settings. Every model experiences notable degradation under at least one realistic robustness condition, including frontier systems such as GPT-5, Gemini-2.5-Pro, and Claude-Sonnet-4.5. Performance drops often exceed 20 points and, in some cases reach 53.5 points in the worst case (Qwen3-VL-8B in DP1). Notably, no model consistently demonstrates robustness across all six evaluation conditions, suggesting that current chart-understanding capabilities remain fragile under realistic variations.

VLMs are relatively robust to prompt phrasing but vulnerable to misleading contextual in-

formation. Table 1 shows that not all language variation affects chart understanding equally. In TC2, semantically equivalent prompt reformulations produce measurable but relatively modest performance changes, with average robustness gaps of 2.25 points for expert-to-non-expert phrasing and 1.82 points for formal-to-informal phrasing. Although 10 of 14 models exhibit some prompt sensitivity, the largest observed drop is only 5.5 points. In contrast, changes to the surrounding textual context while leaving the chart unchanged lead to substantially larger failures. In TC1, caption and annotation mismatches degrade all 14 models, with an average robustness gap of 27.1 points; GPT-5 drops from 80% to 51%, while Qwen3-VL-32B falls from 72% to 41%. Similarly, TC3 yields an average modality-gap drop of 13.0 points, with Claude-Sonnet-4.5 and Qwen3-VL-8B losing more than 20 points. These results indicate that current VLMs are far more vulnerable to misleading contextual information than to differences in user phrasing, suggesting persistent reliance on textual context when it conflicts with chart evidence.

No model is consistently robust across deployment conditions. Table 2 shows that robustness varies substantially across display-side perturbations and models. Low-resolution rendering (DP1) is consistently the most challenging condition, degrading every model as resolution decreases. At quarter resolution, Claude-Sonnet-4.5 experiences the largest drop among closed models ($\Delta_{\text{rob}} = 41.0$), while Qwen3-VL-8B suffers the largest drop overall ($\Delta_{\text{rob}} = 53.5$). In contrast, mobile layouts (DP2) and background-theme variation (DP3) have a smaller average effect but expose distinct model-

Model	DP1: Low-Resolution Plots					DP2: Mobile Variants			DP3: Background Theme Variation				
	Orig.	Half	Third	Quarter	Δ_{rob}	Light	Dark	Δ_{rob}	Light	Dark	$\Delta_{\text{Dark}_{\text{rob}}}$	Gray	$\Delta_{\text{Gray}_{\text{rob}}}$
GPT-5	43	39	36	22	21	79.6	73.5	6.1	77.6	79.6	0.0	81.6	0.0
GPT-5-mini	41	36.5	34	17	24	79.6	67.3	12.3	79.6	69.4	10.2	79.6	0.0
Claude-Sonnet-4.5	54	32.5	15.5	13	41	69.4	69.4	0.0	67.3	69.4	0.0	71.4	0.0
Gemini-2.5-Pro	32.5	32	28	22.5	10	77.6	67.3	10.3	79.6	65.3	14.3	79.6	0.0
Gemini-2.5-Flash	26.5	24.5	21	16	10.5	73.5	71.4	2.1	67.3	73.5	0.0	77.6	0.0
Gemma3-27B	32.5	33.0	22.5	7.0	25.5	55.1	57.1	0.0	49.0	51.0	0.0	53.1	0.0
Gemma3-12B	37.5	36.5	28.0	11.5	26.0	42.9	44.9	0.0	51.0	51.0	0.0	42.9	8.1
Gemma3-4B	38.5	34.5	22.5	10.5	28.0	34.7	40.8	0.0	32.7	34.7	0.0	32.7	0.0
InternVL3.5-14B	65.5	62.0	35.0	16.5	49.0	57.1	55.1	2.0	46.9	44.9	2.0	42.9	4.0
InternVL3.5-8B	64.0	58.5	37.5	13.0	51.0	51.0	49.0	2.0	51.0	57.1	0.0	55.1	0.0
InternVL3.5-4B	54.0	56.0	36.0	12.5	41.5	51.0	53.1	0.0	59.2	51.0	8.2	57.1	2.1
Qwen3-VL-32B	68.0	52.0	33.0	20.5	47.5	59.2	61.2	0.0	59.2	55.1	4.1	59.2	0.0
Qwen3-VL-8B	68.5	52.0	29.0	15.0	53.5	46.9	51.0	0.0	51.0	51.0	0.0	55.1	0.0
Qwen3-VL-4B	66.5	48.0	26.5	15.5	51.0	38.8	38.8	0.0	49.0	49.0	0.0	51.0	0.0

Table 2: Textual and cross-modal robustness results on ROBUSTCHART360. All values are reported as percentages. Lower Δ_{rob} indicates stronger robustness. For DP1 and DP2, Δ_{rob} denotes the maximum performance drop from the baseline. For DP3, $\Delta_{\text{Dark}_{\text{rob}}}$ and $\Delta_{\text{Gray}_{\text{rob}}}$ are measured relative to the light-theme baseline. Green and red highlight the best and worst robustness gaps within each model group.

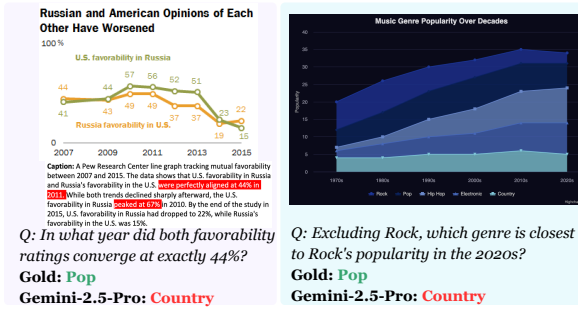


Figure 3: Representative robustness failures. Left: A misleading caption shifts the model from the correct chart-grounded answer to an incorrect text-grounded answer. Right: A presentation-only change (background) leads to wrong answer despite unchanged data.

specific weaknesses, particularly for GPT-5-mini and Gemini-2.5-Pro. Importantly, robustness under one condition does not imply robustness under another. For example, Claude-Sonnet-4.5 is nearly invariant to mobile and theme changes yet exhibits one of the largest low-resolution failures. These findings indicate that chart-understanding robustness is multi-dimensional and cannot be adequately characterized by a single benchmark score.

5.2 Qualitative Analysis

To complement our quantitative evaluation, we manually analyzed 60 failure cases across the six robustness conditions and identified recurring weaknesses in multimodal grounding and visual reasoning. Figure 3 highlights two representative examples. In the modality-gap setting (left), sev-

eral models switch from the correct answer (2009) to an incorrect answer (2011) after reading a misleading caption claiming that the two favorability ratings converged at 44% in 2011, despite the chart clearly indicating that this occurred in 2009. This behavior suggests that models can rely on surrounding textual cues rather than chart evidence when the two conflict. In the background-theme setting (right), Gemini-2.5-Pro incorrectly predicts *Country* instead of *Pop* even though only the chart’s visual presentation changes while the underlying data remain identical. Together, these examples illustrate the two dominant failure modes observed throughout ROBUSTCHART360: over-reliance on contextual text and sensitivity to presentation-level visual variation.

6 Conclusion and Future Work

This paper introduced ROBUSTCHART360, a comprehensive benchmark for evaluating the robustness of VLMs under realistic chart understanding conditions. Through large-scale evaluation across 14 proprietary and open-source VLMs, we showed that current models remain substantially brittle under practical deployment variations, despite strong performance on standard chart benchmarks. Models remain highly vulnerable to misleading textual context, modality conflicts, and degraded visual quality, with performance drops of up to 31% under caption mismatches, 21.2% under modality gaps, and over 50% under severe low-resolution conditions. In contrast, linguistic phrasing varia-

tions produce comparatively modest degradation, indicating that failures primarily stem from weaknesses in multimodal grounding and visual fidelity rather than surface-level language variation. We hope ROBUSTCHART360 serves as a foundation for developing more robust, visually grounded, and deployment-ready multimodal systems for chart understanding. Future extensions will explore additional robustness dimensions, including multilingual charts, OCR corruption, adversarial perturbations, and interactive dashboards.

Limitations

This work provides a broad robustness benchmark for chart understanding, but it does not cover all possible real-world visualization settings. ROBUSTCHART360 focuses on static chart images and question-answering-style evaluation. It does not fully capture interactive dashboards, animated visualizations, multi-step analytical workflows, or multilingual chart understanding. This benchmark only emphasizes selected perturbation types, such as textual mismatch, prompt variation, modality gaps, low resolution, mobile layouts, and background theme changes, leaving other practical factors such as OCR corruption, compression artifacts, extreme cropping, adversarial chart manipulation, and domain-specific visualization conventions for future work.

Our evaluation is limited to a fixed set of proprietary and open-source VLMs available at the time of experimentation. Since closed-source models may change over time and their training data, architectures, and inference procedures are not fully transparent, our results should be interpreted as empirical observations rather than causal explanations of model behavior. In addition, although we use manual verification and automated judging to support reliable evaluation, both generated question-answer pairs and LLM-as-a-judge scoring may still contain occasional errors. Future work can expand the benchmark with larger human validation, more diverse chart domains, and additional robustness dimensions.

Ethics Statement

This work is intended solely for scientific research on the robustness of vision-language models for chart understanding. Our benchmark is built from publicly available chart resources and controlled perturbations, and does not intentionally include

private, personally identifiable, or sensitive information. The introduced mismatches and visual variations are designed only for diagnostic evaluation, not to promote deceptive visualization practices. Open-source models are used according to their licenses, and proprietary models are accessed through official APIs. We encourage users of this benchmark to report limitations clearly and avoid using the results as evidence of reliability in high-stakes decision-making settings.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council (NSERC), Canada, the Canada Foundation for Innovation, CUPE 3903 Research Grant Fund, and the Digital Research Alliance of Canada. We thank all the anonymous reviewers for their excellent review comments.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023a. [Reading and reasoning over chart images for evidence-based automated fact-checking](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.
- Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2023b. [Chartcheck: An evidence-based fact-checking dataset over real-world chart images](#). *Preprint*, arXiv:2311.07453.
- Anthropic. 2025. [Claude sonnet 4.5 system card](#). Technical report.
- Anthropic. 2026. [Claude opus 4.6 system card](#). System card. Published February 2026.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. [Qwen3-vl technical report](#). arXiv preprint arXiv:2511.21631.
- George Comanici and 1 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). arXiv preprint arXiv:2507.06261.

- Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. [Holistic analysis of hallucination in gpt-4v\(ision\): Bias and interference challenges](#). *Preprint*, arXiv:2311.03287.
- Gemma Team. 2025. [Gemma 3 technical report](#). arXiv preprint arXiv:2503.19786.
- Google DeepMind. 2026. [Gemini 3.1 pro model card](#). Model card. Published February 2026.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.
- Highcharts. 2026. Highcharts: Highcharts JS, the JavaScript charting framework. <https://github.com/highcharts/highcharts>. Accessed: 2026-05-19.
- E. Hoque and M. Saidul Islam. 2024. [Natural language generation for visualizations: State of the art, challenges and future directions](#). *Computer Graphics Forum*, n/a(n/a):e15266.
- E. Hoque, P. Kavehzadeh, and A. Masry. 2022. [Chart question answering: State of the art and future directions](#). *Computer Graphics Forum*, 41(3):555–572.
- Ting-Yao ‘Edward’ Hsu, Yi-Li Hsu, Shaurya Rohatgi, Chieh-Yang Huang, Ho Yin Sam Ng, Ryan Rossi, Sungchul Kim, Tong Yu, Lun-Wei Ku, Clyde Lee Giles, and 1 others. 2026. Do large multimodal models solve caption generation for scientific figures? lessons learned from scicap challenge 2023. *Transactions of the Association for Computational Linguistics*, 14:233–252.
- Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024a. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749.
- Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R. Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2023. [Do lvlms understand charts? analyzing and correcting factual errors in chart captioning](#). *Preprint*, arXiv:2312.10160.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024b. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.
- Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024a. [DataNarrative: Automated data-driven storytelling with visualizations and texts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19253–19286, Miami, Florida, USA. Association for Computational Linguistics.
- Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024b. 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*.
- Chaoya Jiang, Yongrui Heng, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2026. Vlm-r³: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38:63841–63869.
- Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2763–2775.
- Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. Opencqa: Open-ended question answering with charts. In *Proceedings of EMNLP (to appear)*.
- Dae Hyun Kim, Seulgi Choi, Juho Kim, Vidya Setlur, and Maneesh Agrawala. 2023. Emphasis checker: A tool for guiding chart and caption emphasis. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):120–130.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. 2023. [Imagenet-e: Benchmarking neural network robustness via attribute editing](#). *Preprint*, arXiv:2303.17096.
- Ridwan Mahbub, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, and Enamul Hoque. 2025a. The perils of chart deception: How misleading visualizations affect vision-language models. In *2025 IEEE Visualization and Visual Analytics (VIS)*, pages 6–10. IEEE.
- Ridwan Mahbub, Mohammed Saidul Islam, Mir Tafseer Nayeem, Md Tahmid Rahman Laskar, Mizanur Rahman, Shafiq Joty, and Enamul Hoque. 2025b. [From](#)

- charts to fair narratives: Uncovering and mitigating geo-economic biases in chart-to-text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28929–28947, Suzhou, China. Association for Computational Linguistics.
- Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025a. [ChartQAPro: A more diverse and challenging benchmark for chart question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria. Association for Computational Linguistics.
- Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [ChartQA: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- Ahmed Masry, Abhay Puri, Masoud Hashemi, Juan A Rodriguez, Megh Thakkar, Khyati Mahajan, Vikas Yadav, Sathwik Tejaswi Madhusudhan, Alexandre Piché, Dzmitry Bahdanau, and 1 others. 2025b. Bigcharts-r1: Enhanced chart reasoning with visual reinforcement finetuning. *arXiv preprint arXiv:2508.09804*.
- Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10387–10409.
- Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024. [Unraveling the truth: Do VLMs really understand charts? a deep dive into consistency and robustness](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16696–16717, Miami, Florida, USA. Association for Computational Linguistics.
- Jason Obeid and Enamul Hoque. 2020. [Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.
- OpenAI. 2025. [Openai gpt-5 system card](#). *arXiv preprint arXiv:2601.03267*.
- OpenAI. 2026. [Introducing gpt-5.4](#). OpenAI Blog. Released March 5, 2026.
- Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34.
- Pritika Ramu, Apoorv Saxena, Varsha Sankar, Debraj Basu, and 1 others. 2025. Styleadaptedlm: Enhancing instruction following models with efficient stylistic transfer. *arXiv preprint arXiv:2507.18294*.
- Chaitanya K. Ryali, David J. Schwab, and Ari S. Morcos. 2021. [Characterizing and improving the robustness of self-supervised learning through background augmentations](#). *Preprint*, arXiv:2103.12719.
- Gabriel Sarch, Lawrence Jang, Michael J Tarr, William W Cohen, Kenneth Marino, and Katerina Fragkiadaki. 2024. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems*, 37:75942–75985.
- Kantharaj Shankar, Leong Rixie Tiffany Ko, Lin Xiang, Masry Ahmed, Thakkar Megh, Hoque Enamul, and Joty Shafiq. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.
- Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. 2023. [VisText: A Benchmark for Semantically Rich Chart Captioning](#). In *The Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. [InternV3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *arXiv preprint arXiv:2508.18265*.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Hao-tian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. [Charxiv: Charting gaps in realistic chart understanding in multimodal llms](#). *Preprint*, arXiv:2406.18521.
- World Wide Web Consortium. 2025. Web content accessibility guidelines (WCAG) 2.1. <https://www.w3.org/TR/WCAG21/>. W3C Recommendation.
- Aoyu Wu, Wai Tong, Tim Dwyer, Bongshin Lee, Petra Isenberg, and Huamin Qu. 2020. Mobilevisfixer: Tailoring web visualizations for mobile phones leveraging an explainable reinforcement learning framework. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):464–474.
- Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2021. [Noise or signal: The role of image backgrounds in object recognition](#). In *International Conference on Learning Representations*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and 1 others. 2026. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36.

A The ROBUSTCHART360 Benchmark

A.1 Taxonomy Construction

Robustness in chart understanding is not a single failure mode. Existing chart QA benchmarks have advanced visual and logical reasoning over charts, but they often evaluate models under relatively clean settings (Masry et al., 2022). Recent work further shows that more diverse, real-world chart settings can expose substantial gaps in LVLM performance (Masry et al., 2025a). In deployment, charts appear in news articles, dashboards, slide decks, PDF reports, social media posts, and accessibility tools, where visual and textual signals may be compressed, reformatted, incomplete, or rhetorically framed.

We therefore design the ROBUSTCHART360 taxonomy around two broad channels through which chart understanding can break. *Textual and cross-modal robustness* tests whether models remain grounded when captions, legends, annotations, or user questions are misleading, stylistically varied, or inconsistent with the visual evidence. This group captures failures caused by changes in the linguistic or multimodal context surrounding the chart, including cases where models may over-rely on textual priors rather than visually verifying the plotted evidence.

Display and presentation robustness tests whether models remain reliable under realistic rendering, viewing, and presentation shifts. This includes low-resolution screenshots, mobile layouts, dark mobile interfaces, and background theme variation. These conditions are practically important because charts are often consumed under constrained devices and interface settings, and accessibility guidelines emphasize that meaningful visual information should remain distinguishable from adjacent colors, including non-text graphical elements (World Wide Web Consortium, 2025). Background and theme changes further test whether models preserve chart semantics under presentation-level variation that should not alter the underlying data interpretation.

Together, these groups convert robustness from a broad desideratum into testable deployment-relevant questions: Does the model privilege visual evidence over misleading text? Does it remain reliable under constrained display and presentation settings? Does it preserve chart reasoning when visual styling changes but the underlying chart semantics remain fixed? Table 3 summarizes the

resulting taxonomy of seven task families.

A.2 Chart Corpus Collection

TC1. For the caption/legend mismatch task, we use charts from the popular ChartQA dataset, which provides a diverse collection of real-world charts. We select charts containing captions, legends, or annotations so that controlled textual mismatches can be introduced without modifying the underlying visual data.

TC2. Real-world users ask chart questions in varied styles, from technical to informal or incomplete phrasing. We test VLM robustness by keeping the chart and answer fixed while perturbing only the question, thereby isolating linguistic robustness from visual understanding. We randomly selected 100 chart images from the ChartQA (Masry et al., 2022) for this analysis.

TC3. For the modality-gap robustness task, we select chart images from ChartQA dataset and keep the original visualization unchanged, including the chart structure, color scheme and labels. We then add a misleading caption that contradicts the visual evidence, allowing us to evaluate whether the model grounds its answer in the chart or is influenced by misleading textual context.

DP1. For low-resolution plots, we use charts from ChartQA, a dataset of charts paired with human-annotated and machine-generated question-answer pairs. We preserve the original chart structure, color scheme, and textual content, and perturb only the image resolution.

DP2. For the mobile variants, we synthetically generated chart pairs using Claude Opus 4.7 (Anthropic, 2026) through an iterative human-in-the-loop process. The model was instructed to create charts of specific types using HTML along with Highcharts.js (Highcharts, 2026), which were then rendered as screenshots in a headless browser. For each chart, the model generated both the desktop website version and its corresponding mobile version. The chart pairs were manually verified for correctness and, when necessary, iteratively refined using the LLM.

DP3. For the dark variants, we followed a similar approach to the mobile variants in DP2, using Claude Opus 4.7 (Anthropic, 2026) through an iterative human-in-the-loop process. In this case, the model was instructed to generate a dark-mode

Robustness Group	Tasks	What It Measures
Textual and Cross-Modal Robustness	TC1: Caption/Annotation Mismatch TC2: Input Prompt Variation TC3: Modality Gap	Tests whether VLMs remain visually grounded when the textual channel changes, becomes misleading, or conflicts with the visual evidence in the chart.
Display and Presentation Robustness	DP1: Low-Resolution Plots DP2: Mobile Variants DP3: Background Theme Variation	Tests whether VLMs remain reliable under realistic rendering, viewing, and presentation shifts, including compression, small screens, dark mobile layouts, and background changes.

Table 3: Taxonomy of robustness groups in ROBUSTCHART360. The benchmark groups seven task families according to how real-world variation enters chart understanding: through textual or cross-modal interference, or through display and presentation shifts.

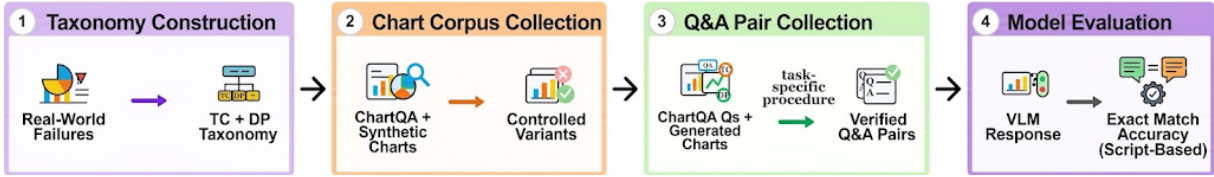


Figure 4: Overview of the ROBUSTCHART360 construction and evaluation pipeline. ROBUSTCHART360 is constructed by defining a TC/DP robustness taxonomy, collecting ChartQA-derived and synthetic chart corpora, generating controlled chart variants, and creating verified task-specific Q&A pairs. We perform automatic script based VLM evaluation.

variant for each synthetically generated chart. The resulting chart pairs were manually evaluated for correctness and refined iteratively when needed.

A.3 Q&A Pair Creation

TC1. For the caption/legend mismatch task, for each selected chart, we created a perturbed chart with a manually introduced caption, legend, or annotation mismatch. The question–answer pairs were created through human and multiple-LLM collaboration, followed by manual verification to ensure that the correct answer was grounded in the visual chart rather than the mismatched text.

TC2. For the 100 selected charts from ChartQA, we used *Claude-4.7-Opus* to generate four semantically equivalent question variants per chart: *expert*, *non-expert*, *formal*, and *informal*. These variants capture technical, lay, professional, and conversational phrasing, respectively. Since each of the selected charts has two original questions, our dataset contains 10 samples per selected chart.

TC3. To evaluate whether VLMs rely on visual evidence or textual priors when interpreting charts, we construct modality-gap Q&A pairs by introducing controlled inconsistencies between the chart image and its accompanying textual context. For each selected chart, we create an augmented version that includes a caption containing a subtle but

misleading claim about the visualized data. The corresponding question is then designed around this misleading textual claim, while the ground-truth answer is determined strictly from the chart itself. Therefore, a model that answers based on the caption is considered to be relying on textual cues and is marked incorrect, whereas a model that answers based on the chart is considered visually grounded.

DP1. To construct the low-resolution variants used in DP1, we collected 200 samples from the ChartQA dataset. Each original chart image was systematically downsampled to three reduced resolutions using bilinear interpolation. Specifically, the half-resolution variant was produced by scaling both the width and height of the original image to 50% of their original dimensions, resulting in an image with one-quarter the total pixel count. The third-resolution variant was obtained by scaling each spatial dimension to 33.3%, yielding approximately one-ninth the original pixel area, while the quarter-resolution variant was created by reducing each dimension to 25%, retaining only one-sixteenth of the original pixel count.

DP2. & DP3. To generate diverse chart-grounded QA pairs for the background-theme and mobile-variant settings, we use an LLM-assisted question–answer generation process. To reduce

Chart Type	TC1: Caption/Annotation	TC2: Prompt Var.	TC3: Modality Gap	DP1: Low Resolution	DP2: Mobile	DP3: Background
 Bar Chart	118	35	6	472	6	24
 Line Chart	64	19	254	128	12	18
 Pie Chart	14	11	0	24	4	6
 Donut Chart	0	0	0	0	6	9
 Scatter Plot	0	0	0	0	4	3
 Bubble Chart	0	0	0	0	4	6
 Area Chart	0	0	0	0	6	12
 Grouped Bar	4	15	0	72	16	12
 Stacked Bar	10	20	0	104	14	18
 Other	0	0	0	0	26	39
Total	200	100	260	800	98	147

Table 4: Chart-type coverage across benchmark tasks.

Task	# Samples	Unique Q. Tokens	Q. TTR	Avg. Q. Tokens	Avg. Q. Chars.	Avg. Ans. Tokens	Avg. Ans. Chars.
TC1: Caption/Annotation Mismatch	200	349	0.15	11.54	67.42	1.36	5.43
TC2: Input Prompt Variation	1000	869	0.08	11.25	62.57	1.47	5.60
TC3: Modality Gap	260	298	0.08	12.21	70.79	1.09	3.23
DP1: Low-Resolution Plots	800	738	0.07	12.12	66.09	1.56	5.37
DP2: Mobile Variants	98	299	0.21	14.69	85.53	1.18	5.98
DP3: Background Theme Variation	147	299	0.14	14.69	85.53	1.18	5.98

Table 5: Linguistic statistics of the benchmark by task. Q. denotes question, Ans. denotes answer, and TTR denotes type-token ratio.

model-specific bias in the generation process, we prompt three state-of-the-art proprietary multi-modal models, such as GPT-5.4 (OpenAI, 2026), Gemini-3.1-Pro (Google DeepMind, 2026), and Claude-Opus-4.6 (Anthropic, 2026), with each chart image and ask each model to generate five questions. We explicitly encourage diversity across question types, including *factoid* questions that require precise visual extraction, *fact-checking* questions that assess the validity of chart-grounded claims, and *hypothetical* questions that require reasoning over the visualized data. This QA-pair generation strategy helped diversify the linguistic form, reasoning pattern, and visual evidence required by the questions, while reducing reliance on the biases or preferences of any single generator model.

A.4 Dataset Stats.

Figure 2 presents the dataset distribution across two broad evaluation dimensions, namely Textual and Cross-Modal (TC) and Display and Presentation (DP), comprising a total of 2505 samples, respectively. The TC dimension encompasses three tasks including Caption/Annotation Mismatch (200 samples), Input Prompt Variation (1,000 samples across five subcategories of Original, Expert, Non-expert,

Formal, and Informal with 200 samples each), and Modality Gap (260 samples evenly split between Clean and Augmented). The DP dimension covers four tasks including Low-Resolution Plots (800 samples across Original, Half, Third, and Quarter subcategories with 200 samples each), Mobile Variants (98 samples split evenly between Light and Dark) and Background Theme Variation (147 samples distributed equally across Light, Dark, and Gray subcategories). In total, the dataset comprises **2,505 samples** spanning 23 subcategories across seven tasks, ensuring broad coverage of real-world visual and linguistic perturbations.

A.4.1 Chart Type Diversity

Table 4 summarizes the chart-type diversity across benchmark tasks. The benchmark contains 1,645 chart instances across six tasks, covering both common and specialized visualization types. Bar charts form the largest group, with 661 instances (40.2%), followed by line charts with 495 instances (30.1%). The benchmark also includes substantial coverage of stacked bar charts (166; 10.1%) and grouped bar charts (119; 7.2%), enabling evaluation on multi-series visual reasoning. Less frequent but important chart types, including pie, donut, scatter, bubble,

and area charts, are also represented, particularly in the mobile and background-theme settings. This distribution reflects both the prevalence of bar and line charts in real-world chart corpora and our effort to include diverse visual encodings across tasks.

A.4.2 Linguistic Diversity

Table 5 reports the linguistic diversity of the benchmark across 2,505 samples. The tasks vary in lexical diversity, with unique question tokens ranging from 298 in TC3 to 869 in TC2, and question TTR ranging from 0.07 to 0.21. Question length also varies across tasks: average question length ranges from 11.25 to 14.69 tokens and 62.57 to 85.53 characters, with DP2 and DP3 containing the longest questions on average. In contrast, answers remain concise across all tasks, averaging 1.09–1.56 tokens and 3.23–5.98 characters. These statistics indicate that the benchmark evaluates models under diverse question formulations while maintaining short, factoid-style answers suitable for controlled automatic evaluation.

B Additional Results and Analysis

B.1 TC1. Caption/Annotation Mismatch

Table 1 shows that mismatched captions and annotations substantially reduce model accuracy across both closed-source and open-source VLMs. For example, GPT-5 drops from 80 on the original charts to 51 when the caption, legend, or annotation is mismatched, even though the visual chart content remains unchanged. Similar drops are observed for open-source models, such as Qwen3-VL-32B, which decreases from 72 to 41. The consistent negative delta across both model groups indicates that VLMs often rely on misleading textual cues instead of fully verifying the visual evidence in the chart.

B.2 TC2. Input Prompt Variation

Table 1 shows that prompt phrasing has a modest but consistent effect on chart QA performance. Unlike caption mismatch or modality gap, the drops are relatively small: the average robustness gap is 2.25 points for expert-to-non-expert and 1.82 points for formal-to-informal phrasing. This indicates that VLMs are generally more robust to linguistic variation than to misleading chart context.

Closed-source models achieve the highest absolute accuracy, with Gemini-2.5-Pro performing best and remaining highly stable across variants. However, strong models are not always

the most phrasing-invariant: GPT-5 and Claude-Sonnet-4.5 both show 4.5-point drops from expert to non-expert prompts. Among open models, InternVL3.5 and Qwen3-VL variants often show small robustness gaps despite lower absolute accuracy, while Gemma3 models are more sensitive, especially Gemma3-12B under non-expert phrasing and Gemma3-4B under informal phrasing.

Overall, TC2 suggests that prompt variation can be a meaningful robustness challenge. Non-expert or informal wording does not uniformly hurt performance, but single-prompt evaluation can obscure model-specific sensitivity to user phrasing, as illustrated by representative failure cases in Table 6.

B.3 TC3. Modality-gap

Table 1 presents the clean and augmented cross-modal consistency results across closed-source and open-source VLMs. In the augmented setting, each chart is paired with a misleading caption, creating a controlled modality gap between the visual evidence and the accompanying textual information. Therefore, a performance drop from the clean to the augmented setting indicates that the model relies more on the misleading textual information than on the actual visual evidence in the chart. From the results, we observe that most models show clear degradation under this setting. Among closed-source models, GPT-5 drops from 32.67% to 26.00%, while GPT-5-mini and GPT-5-nano decrease by 12.00 and 12.21 points, respectively. Similar trends are observed for open-source models, where Qwen3-8B drops from 63.33% to 42.67%, and InternVL3.5-14B decreases from 62.67% to 50.00%. Although Qwen3-32B shows a smaller reduction, its performance still decreases under the misleading caption setting. These results suggest that most VLMs tend to over-rely on textual information, even when it conflicts with the visual evidence in the chart.

B.4 DP1. Low-resolution Plots

Table 2 demonstrates that the resolution degradation consistently imposes a performance cost across all evaluated models, yet the magnitude of this effect varies substantially. Among closed-source models, the Gemini family demonstrates the greatest resilience, with Gemini-2.5-Pro and Gemini-2.5-Flash achieving robustness drops (Δ_{rob}) of only 10.0 and 10.5, respectively. By contrast, Claude-Sonnet-4.5 exhibits the steepest degradation among closed-source models ($\Delta_{rob} = 41$), col-

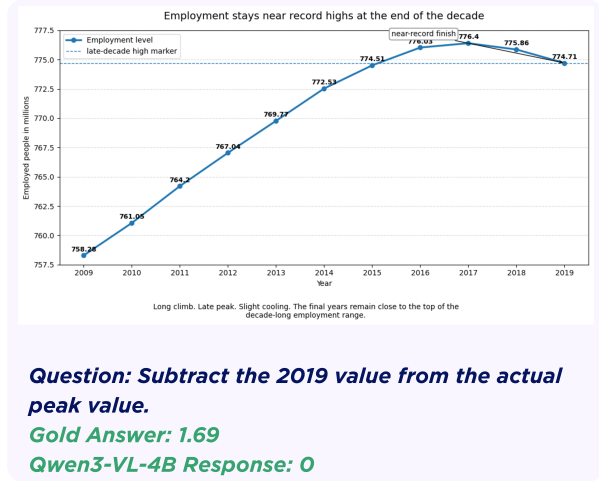
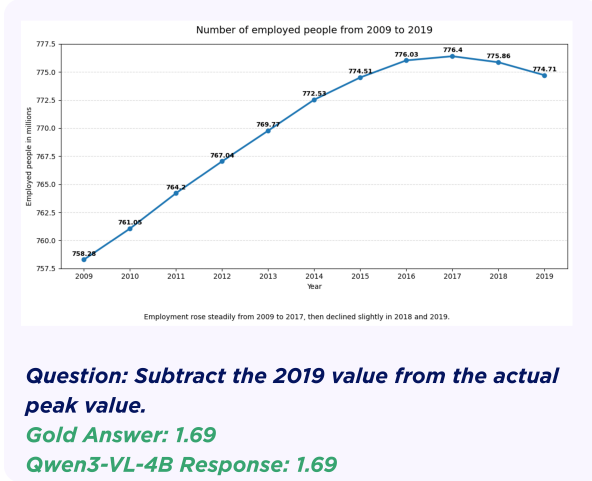


Figure 5: Example of caption/annotation mismatch. Although the chart values support the correct answer of 1.69, the mismatched annotation causes the model to follow the misleading textual cue and return 0, showing weak grounding in the visual chart evidence.

lapsing from 54.0 at full resolution to just 13.0 at quarter resolution, suggesting that its visual parsing strategy is strongly reliant on fine-grained pixel detail. A similar pattern holds across open-source models. The highest-performing families at full resolution suffer the most severe drops, with Qwen3-VL-8B and InternVL3.5-8B recording Δ_{rob} values of 53.5 and 51.0, respectively. The Gemma3 family, while performing more modestly overall, exhibits comparatively higher robustness ($\Delta_{rob} \leq 28.0$), echoing the closed-source Gemini trend. Taken together, these results indicate a systematic tension between high-resolution accuracy and low-resolution robustness. The models optimized for strong visual feature extraction tend to be disproportionately sensitive to image quality degradation, which has important practical implications for deployment in bandwidth-constrained or low-fidelity display environments.

B.5 DP2. Mobile Variants

Table 2 shows that mobile rendering produces a relatively mild shift for open models but a clearer degradation for closed models. Closed models achieve higher absolute accuracy under the light-mobile condition, but their average performance drops from 75.9 to 69.8 under dark mobile rendering, with the largest robustness gaps observed for GPT-5-mini (12.3) and Gemini-2.5-Pro (10.3). In contrast, open models are largely stable: their average accuracy slightly increases from 48.5 to 50.1, and most models have a robustness gap of 0.0, indicating no degradation relative to the light-mobile

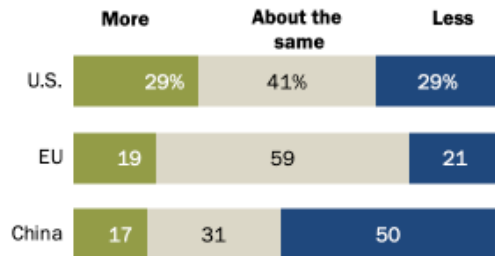
baseline. The only open models with non-zero gaps are InternVL3.5-14B and InternVL3.5-8B, both with a small drop of 2.0 points. Overall, the results suggest that mobile aspect ratio changes do not uniformly degrade chart understanding; instead, robustness depends strongly on model family, with several open models showing greater invariance despite lower absolute accuracy.

B.6 DP3. Background Theme Variation

Table 2 shows that background perturbations affect model performance, but the effect is model-dependent rather than uniformly harmful. Closed models obtain an average accuracy of 74.3 under the light background, drop to 71.4 under the dark background, and increase to 78.0 under the gray background. Their gray-background robustness gap is 0.0 for all models, indicating that no closed model is harmed by the gray variant relative to the light baseline. However, the dark background exposes substantial weaknesses in some otherwise strong models: GPT-5-mini and Gemini-2.5-Pro show large robustness gaps of 10.2 and 14.3 points, respectively, while GPT-5, Claude-Sonnet-4.5, and Gemini-2.5-Flash remain unaffected. Open models show a more heterogeneous pattern: several models are stable or improve under both dark and gray backgrounds, but InternVL3.5-4B and Qwen3-VL-32B degrade under dark, while Gemma3-12B and InternVL3.5-14B degrade under gray. Overall, the experimental results suggest that background robustness is highly model-specific; gray themes are often safe or beneficial, whereas dark themes

Many think China's global influence will decline after the coronavirus outbreak

% who say each will have ___ influence in world affairs after the coronavirus outbreak compared to before the outbreak



Note: No answer responses not shown.

Source: Survey of U.S. adults conducted April 29-May 5, 2020.

"Americans Give Higher Ratings to South Korea and Germany Than U.S. for Dealing With Coronavirus"

PEW RESEARCH CENTER

Type	Question
Original	What's the leftmost value of the bar in China?
Expert	What is the value of the leftmost bar associated with China?
Non-expert	What number is on the bar farthest to the left for China?
Formal	What is the value of the leftmost bar for China?
Informal	What's the value of China's bar on the far left?
Gold Answer	17

Figure 6: Question variations for the same ChartQA example.

can cause notable failures even in high-performing proprietary models.

C ROBUSTCHART360 Examples

In this section, we provide representative samples from ROBUSTCHART360 across each categories.

TC1: Caption/Annotation Mismatch. This setting evaluates whether VLMs remain grounded in the visual chart evidence when accompanying textual elements, such as captions, legends, or annotations, conflict with the underlying data. Figure 5 shows a representative failure case: the chart itself indicates that the actual peak employment value is 776.40 and the 2019 value is 774.71, so the correct subtraction result is 1.69. However, the mismatched annotation leads Qwen3-VL-4B to return 0, suggesting that the model follows the misleading textual cue rather than verifying the plotted values. This example illustrates how caption or annotation mismatches can expose weak visual grounding in chart question answering.

TC2. Input Prompt Variation: Prompt-variation robustness in Figure 6 shows how the same chart question can be phrased in multiple semantically equivalent ways while still requiring the

same answer. The chart on the left is a bar chart about views on whether China's global influence will change after the coronavirus outbreak. For China, the leftmost bar segment represents the percentage who say China will have more influence, and the value shown is 17. On the right, the same question is rewritten in different styles, including original, expert, non-expert, formal, and informal versions. Although the wording changes, all versions ask for the same visual information, so the gold answer remains 17.

Table 6 expands this idea by showing a table of model performance across equivalent prompt pairs. It highlights cases where a model answers one version of a question correctly but fails when the question is rephrased in a non-expert or informal style. The green columns show the successful query and correct prediction, while the red columns show the failed reformulation and the incorrect prediction. Examples include models such as GPT-5, Qwen3-VL-32B, Gemini-2.5-Flash, and InternVL3.5-4B making mistakes even when the rewritten question has the same meaning as the original.

TC3. Modality-gap robustness Modality-gap robustness setting in Figure 7 tests whether vision-

Model	Pair	Gold	Correct Query	Correct Pred.	Failed Query	Failed Pred.
GPT-5	Orig.→Non-exp.	Spain	Which country data is consistently above 1.5 kg?	Spain	Which country stays above 1.5 kg the whole time?	None
GPT-5	Formal→Informal	2	How many “Fairly well” bars are repeated at the same value?	2	How many “Fairly well” bars are the same?	4
Qwen3-VL-32B	Orig.→Informal	18	What is the timespan measured in the graph in years?	18	How many years are shown on the graph?	9
Qwen3-VL-32B	Expert→Non-exp.	7.53	Subtract the minimum male life expectancy from the median female life expectancy over the years.	7.53	Take the lowest male life expectancy and subtract it from the middle female life expectancy value across all years.	6.12
Gemini-2.5-Flash	Orig.→Non-exp.	No	Is the median of green graph from 2002 to 2006 greater than smallest value of orange graph?	No	Is the middle value of the green graph from 2002 to 2006 bigger than the smallest value on the orange graph?	Yes
Gemini-2.5-Flash	Formal→Informal	87.5	What is the sum of the median values of the blue and green graphs?	87.5	What’s the blue median plus the green median?	44
InternVL3.5-4B	Orig.→Informal	Yes	Is the starting value of China in the graph greater than 40%?	Yes	Does China start off over 40% in the graph?	No
InternVL3.5-4B	Expert→Non-exp.	4	How many shows in the chart have a higher median viewer age than The Daily Show?	4	How many shows have viewers who are older than the people watching The Daily Show?	5

Table 6: Representative examples where a model correctly answers one prompt formulation but fails a semantically equivalent non-expert or informal variant. Green columns indicate successful predictions, while red columns highlight failures caused by prompt reformulation.

language models rely more on the actual chart image or on misleading textual context added around it. Each example shows an original chart alongside an augmented version that includes extra descriptive text, which can contradict or distract from the visual evidence. The questions ask for specific values from the charts, such as the year when U.S. and Russian favorability ratings both reached 44%, or the percentage for “U.S. respects personal freedoms” in 2014. The gold answers are shown in green, while incorrect model responses are shown in red. The results demonstrate that some models correctly extract information from the chart, while others are misled by the added text and give answers that are inconsistent with the visual data.

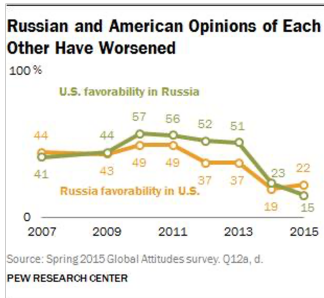
DP1: Low-Resolution Plots Figure 8 illustrates how reducing chart resolution can significantly affect model performance. For each plot, the models are able to produce the correct answer when the chart is shown at its original size. In that case axis labels, data values, legends, and category names are clearly visible. However, when the same charts are presented in quarter resolution, important visual details become blurred or difficult to read. As a result, the models misinterpret numerical values, confuse categories, or fail to identify the correct trend, leading to incorrect answers.

DP2. Mobile Variants In the mobile variants category (Figure 9), charts are shown in a compact mobile-style layout and models are asked to answer visual reasoning questions. The examples in-

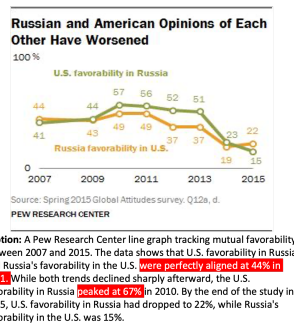
clude different chart types, such as line charts, area charts, horizontal bar charts, donut charts, stacked bar charts, and grouped bar charts. Each panel compares the gold answer with responses from a closed model and an open model. The results show that mobile-style chart layouts can create challenges for model interpretation, especially when the task requires precise comparison, identifying category changes, estimating relationships between values, or verifying whether a statement is true or false. While some questions are answered correctly by both models, other examples show errors caused by misreading chart values, confusing nearby categories, or failing to reason over multiple visual elements. Overall, the figure highlights that mobile chart variants can affect model reliability even when the visual information needed to answer the question is present in the chart.

DP3. Background Theme Variation Figure 10 presents examples from the background theme variation category, where charts are rendered with different visual themes such as dark backgrounds, light backgrounds, and varied color palettes. The panels include several chart types, including line charts, area charts, horizontal bar charts, donut charts, stacked bar charts, and grouped bar charts. Each example shows the chart, the question, the gold answer, and responses from both GPT-5 and Qwen3-VL-32B-Instruct. The results illustrate that changes in background style and color contrast can influence model performance, especially when the task requires comparing values, identifying the

Question: In what year did both favorability ratings converge at exactly 44%?

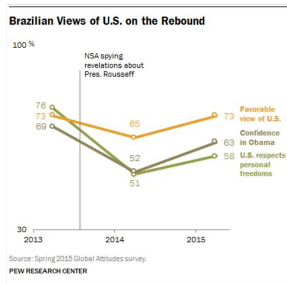


Gold: 2009
GPT-5-mini: 2009
Gemini-2.5-Pro: 2009
InternVL3.5-14B: 2007
Qwen3-VL-32B: 2009

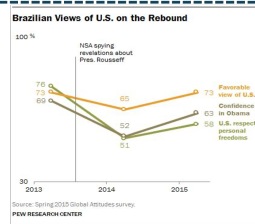


Gold: 2009
GPT-5-mini: 2011
Gemini-2.5-Pro: 2011
InternVL3.5-14B: 2011
Qwen3-VL-32B: 2009

Question: What percentage is shown for 'U.S. respects personal freedoms' in 2014?



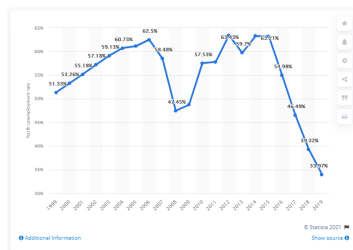
Gold: 51%
GPT-5-mini: 48%
Gemini-2.5-Pro: 51%
InternVL3.5-14B: 51%
Qwen3-VL-32B: 51%



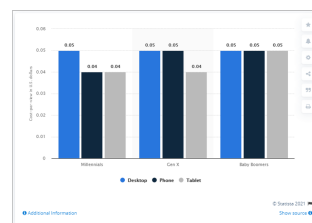
Gold: 51%
GPT-5-mini: 30%
Gemini-2.5-Pro: 51%
InternVL3.5-14B: 30%
Qwen3-VL-32B: 51%

Figure 7: Examples from the modality-gap robustness category. Each example presents the original chart and its augmented version with misleading textual context. For each case, we show the question, gold answer, and responses from multiple closed-source (e.g., GPT-5, Gemini) and open-source (e.g., InternVL, Qwen3-VL) VLMs. Text highlighted in red marks the injected claims that contradict the chart. For example, in the second sample it states that "U.S. respects personal freedoms" **remained the lowest throughout** and **fell to 30%** in 2014, whereas this line is in fact the highest in 2013 (76%) and its 2014 value is 51%.

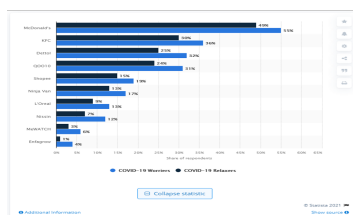
closest category, or reasoning across multiple chart elements. In some cases, both models correctly match the gold answer, while in others the open model produces incorrect responses, such as selecting the wrong month, category, or truth value.



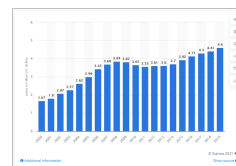
Q: What was the youth unemployment rate in Bosnia and Herzegovina in 2019?
Gold: 33.97
GPT-5-mini: 33.8
Gemma3-4B: 13.6
Internvl3.5-8B: -11.0



Q: What was the average CPV of youtube ads viewed by Millennials on desktop?
Gold: 0.05
GPT-5-mini: 0.05
Gemma3-4B: 0.08
Internvl3-8B: 0.01

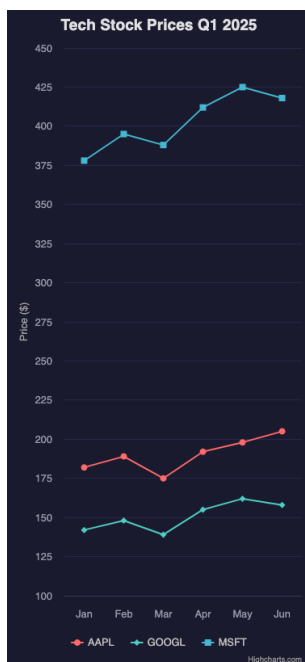


Q: What is the most popular disinfectant brand in Singapore?
Gold: Dettol
GPT-5-mini: Dettol
Gemma3-4B: Lysol
Internvl3.5-8B: Dettol

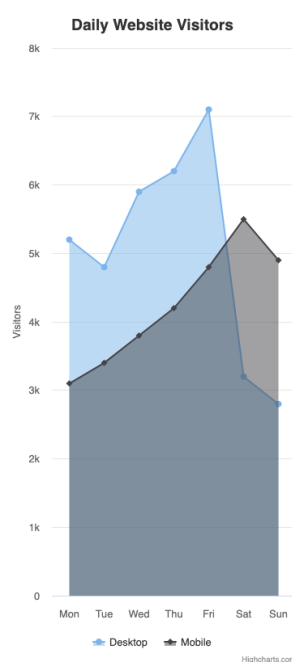


Q: What was the value of loans of FDIC-insured commercial banks secured by real estate in dollars in 2019?
Gold: 4.6
GPT-5-mini: 4.6
Gemma3-4B: 1.3
Internvl3.5-8B: 1.65

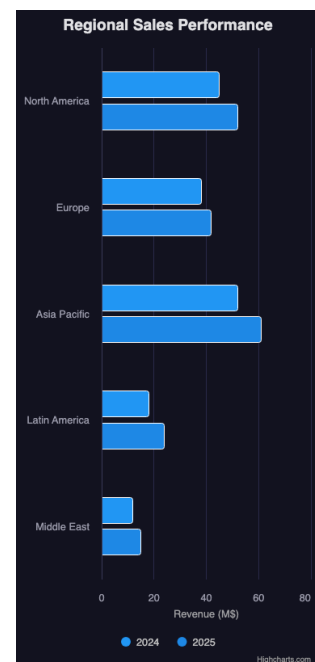
Figure 8: Examples of low resolution. Top row show the response to the charts downsampled to 25% of the original resolution, while the bottom row show the response to the charts downsampled to 50%. Each panel shows the chart, question, gold answer, GPT-5-mini, Gemma3-4B, and Internvl3.5-8B response.



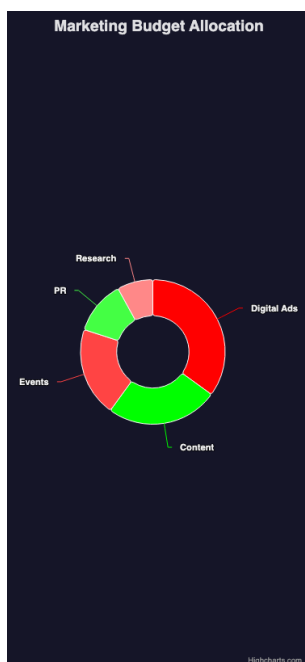
Q: In which month is the gap between MSFT and AAPL the largest?
Gold: May
Closed (GPT-5): May
Open (Qwen3-VL-32B-Instruct): Jun



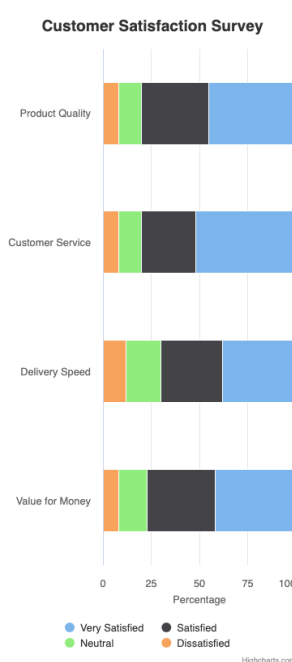
Q: Which series leads for more days across the week?
Gold: Desktop
Closed (GPT-5): Desktop
Open (Qwen3-VL-32B-Instruct): Desktop



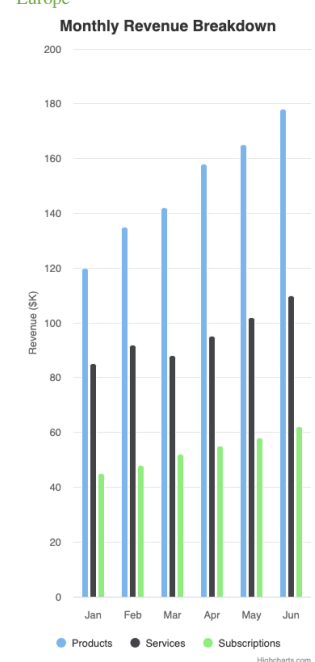
Q: In 2025, which region's revenue is closest to the midpoint between the highest and lowest 2025 regional revenues?
Gold: Europe
Closed (GPT-5): Europe
Open (Qwen3-VL-32B-Instruct): Europe



Q: Which category would become the second-largest allocation if Digital Ads were removed from the chart?
Gold: Content
Closed (GPT-5): Events
Open (Qwen3-VL-32B-Instruct): Content

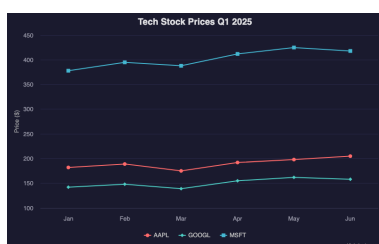


Q: For which category is the Satisfied segment closest in size to the combined Neutral and Dissatisfied segments?
Gold: Customer Service
Closed (GPT-5): Delivery Speed
Open (Qwen3-VL-32B-Instruct): Delivery Speed

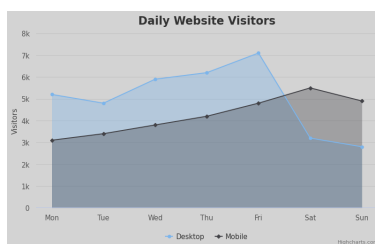


Q: True or False: the month with the highest Services revenue is also the month with the highest total revenue across all three categories.
Gold: TRUE
Closed (GPT-5): True
Open (Qwen3-VL-32B-Instruct): False

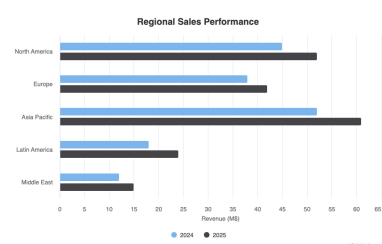
Figure 9: Examples from the Mobile variants category. Each panel shows the chart, question, gold answer, GPT-5 response, and Qwen3-VL-32B response.



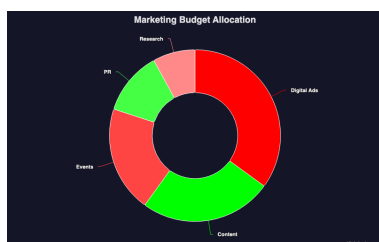
Q: In which month is the gap between MSFT and AAPL the largest?
Gold: May
Closed (GPT-5): May
Open (Qwen3-VL-32B-Instruct): Jun



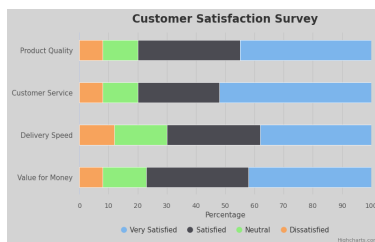
Q: Which series leads for more days across the week?
Gold: Desktop
Closed (GPT-5): Desktop
Open (Qwen3-VL-32B-Instruct): Desktop



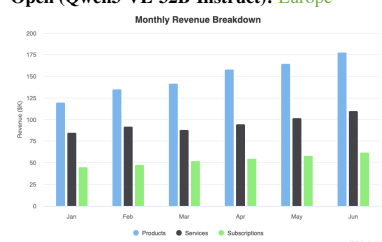
Q: In 2025, which region's revenue is closest to the midpoint between the highest and lowest 2025 regional revenues?
Gold: Europe
Closed (GPT-5): Europe
Open (Qwen3-VL-32B-Instruct): Europe



Q: Which category would become the second-largest allocation if Digital Ads were removed from the chart?
Gold: Content
Closed (GPT-5): Events
Open (Qwen3-VL-32B-Instruct): Content



Q: For which category is the Satisfied segment closest in size to the combined Neutral and Dissatisfied segments?
Gold: Customer Service
Closed (GPT-5): Delivery Speed
Open (Qwen3-VL-32B-Instruct): Delivery Speed



Q: True or False: the month with the highest Services revenue is also the month with the highest total revenue across all three categories.
Gold: TRUE
Closed (GPT-5): True
Open (Qwen3-VL-32B-Instruct): False

Figure 10: Examples from the Background Theme variation category. Each panel shows the chart, question, gold answer, GPT-5 response, and Qwen3-VL-32B response.