CHARTHAL: A Fine-grained Framework Evaluating Hallucination of Large Vision Language Models in Chart Understanding

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have recently demonstrated remarkable progress, yet hallucination remains a critical barrier, particularly in chart understanding, which requires sophisticated perceptual and cognitive abilities as well as rigorous factual accuracy. While prior work has investigated hallucinations and chart comprehension independently, their intersection remains largely unexplored. To address this gap, we present CHARTHAL, a benchmark that features a fine-grained taxonomy of hallucination scenarios in chart understanding, along with a human-validated dataset of 1,062 samples. Our evaluation shows that state-of-the-art LVLMs suffer from severe hallucinations on CHARTHAL, including proprietary models such as GPT-5 and o4-mini, which achieve only 34.46% and 22.79% accuracy, respectively. Further analysis reveals that questions involving information absent from or contradictory to charts are especially likely to trigger hallucinations, underscoring the urgent need for more robust mitigation strategies.

1 Introduction

Since the emergence of GPT-4V (Achiam et al., 2023), Large Vision-Language Models (LVLMs) (Liu et al., 2023b; Zhu et al., 2023; Gemini et al., 2023; Wang et al., 2024b,a; Anthropic, 2024) have exhibited unprecedented capabilities in visual understanding and reasoning, enabling broad downstream image-to-text applications. Among them, chart understanding (Xu et al., 2023; Masry et al., 2022; Wang et al., 2024c) is particularly valuable yet highly challenging, as charts often contain dense information and intricate relationships among visual elements, thereby demanding precise visual perception and substantial cognitive reasoning efforts on LVLMs.

Despite the remarkable progress, LVLMs are prone to *hallucination*, which often manifests as generating information either inexistent in the real world or inconsistent with the input (Bai et al., 2024; Zhou et al., 2023). This issue can be especially pronounced and detrimental in the chart understanding task, where factual correctness and strict consistency with the visual input are essential.

However, though prior studies have extensively investigated hallucinations of LVLMs, most of them limited their scopes to hallucinations in object and scene understanding (Li et al., 2023; Sun et al., 2024), largely neglecting chart-specific settings. Similarly, existing chart understanding benchmarks mainly emphasize the final accuracy of answers to questions regarding input charts (Tang et al., 2025; Shen et al., 2024), but offer limited insights into whether and how hallucinations emerge. This leaves an important gap in systematically diagnosing hallucination behaviors in chart understanding.

To address this gap, we introduce CHARTHAL, the first fine-grained benchmark for evaluating hallucinations in LVLMs on chart understanding tasks to the best of our knowledge. In the form of chart question answering, CHARTHAL features a fine-grained taxonomy that organizes chart-question pairs into 12 hallucination-triggering scenarios according to question types and chart-question relations. Based on this framework, it further provides a human-validated dataset of 1,062 samples spanning diverse chart types and academic disciplines.

Our experiments on 15 state-of-the-art LVLMs reveal that severe hallucinations occur in chart understanding. Notably, even proprietary models such as GPT-5 and o4-mini, often regarded as highly capable, achieve only 34.46% and 22.79% overall accuracy, respectively. Errors are especially concentrated in unanswerable cases where questions involve inexistent or contradictory information. In such cases, models frequently fabricate content rather than correctly figure out the unanswerable nature of input chart-question pairs. These findings highlight critical vulnerabilities of current LVLMs and call for more robust strategies to miti-

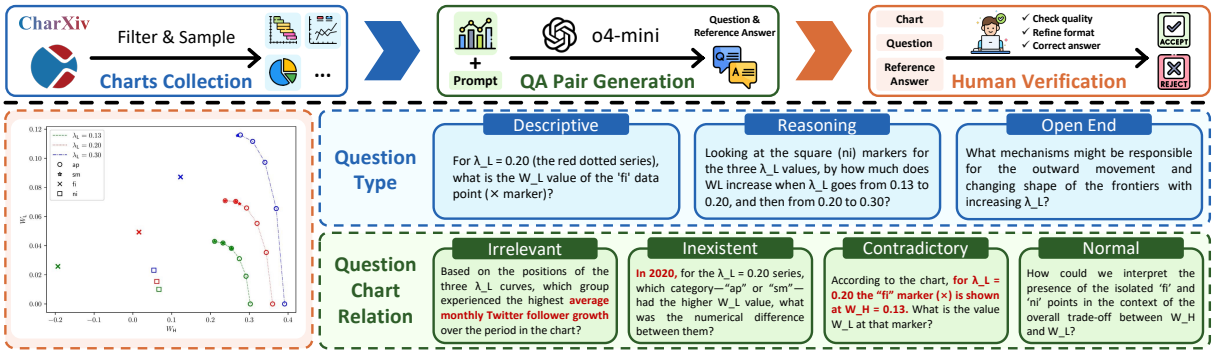

Figure 1: Overview of CHARTHAL. The top panel depicts the construction process of the dataset, while the bottom panel illustrates the taxonomy of questions in CHARTHAL with specific cases. Zoom in for better visual effects.

gate hallucination in chart understanding.

2 Related Works

Hallucination Evaluation of LVLMs. Hallucination, widely observed since the advent of large language models (LLMs) (Hurst et al., 2024; OpenAI, 2025b,a; Comanici et al., 2025), refers to generating information absent from the input or factually incorrect (Huang et al., 2025). With the extension of LLMs to visual modalities, hallucination in LVLMs emphasizes errors inconsistent with visual input (Bai et al., 2024). Given its severe impact on accuracy, many studies have explored hallucination evaluation. However, most works (Rohrbach et al., 2018; Li et al., 2023; Zhang et al., 2023; Fu et al., 2023; Zhang et al., 2024; Sun et al., 2024; Qiu et al., 2024) focus on objects and attributes in input images, while others (Zhou et al., 2023; Wang et al., 2023; Liu et al., 2023a; Jiang et al., 2024) address scene-level hallucinations. This leaves a major gap in evaluating hallucinations in charts. HallusionBench (Guan et al., 2023) is the only benchmark involving LVLM hallucination in charts to our knowledge, but it restricts questions to Yes/No format, oversimplifying scenarios of chart understanding.

Chart Understanding Benchmarks. Chart understanding has been studied through various benchmarks (Kahou et al., 2017; Kafle et al., 2018; Methani et al., 2020; Masry et al., 2022; Xu et al., 2023; Liu et al., 2024; Wang et al., 2024c; Xia et al., 2024; Masry et al., 2025; Tang et al., 2025; Lin et al., 2025), but most focus on answer correctness rather than hallucination analysis. CharXiv (Wang et al., 2024c) and ChartQAPro (Masry et al., 2025) include an unanswerable question subset, enabling limited hallucination evaluation, yet these subsets remain small and lack a fine-grained taxonomy of hallucination-triggering question types, constraining deeper analysis. Concurrently, ChartCap (Lim et al., 2025) explores hallucinations in chart understanding, but it is confined to the captioning task and does not cover interactive question answering.

3 The CHARTHAL Benchmark

3.1 Benchmark Design

CHARTHAL adopts chart question answering as its task form, where LVLMs are required to generate an answer a from given chart image c and corresponding question q. Unlike prior works employing Yes/No or multiple choice questions, we use *free-form* answering format to more faithfully reflect LVLM hallucination behaviors in realistic chart understanding scenarios (Li et al., 2024).

Moreover, to facilitate a fine-grained evaluation, we carefully design a systematic taxonomy based on 2 independent dimensions (Figure 1):

Question Type distinguishes: (i) *descriptive* questions, which can be answered by directly extracting information from the chart; (ii) *reasoning* questions, which demand computational or logical inference based on chart content; and (iii) *open-ended* questions, which involve analytical or predictive tasks without definitive correct answers.

Chart-Question Relation categorizes questions as (i) *irrelevant* (unrelated to chart content), (ii) *inexistent* (inquiring about missing information), (iii) *contradictory* (based on false premises inconsistent with chart data), or (iv) *normal* (valid questions aligned with chart content). Questions of the first three relations are considered *unanswerable*, as

	Desc.	Reason	Open	**Total**
Irrel.	56	99	114	269
Inexist.	151	105	88	344
Contra.	76	54	80	210
Normal	100	64	75	239
Total	383	322	357	1062

Table 1: Sample number per category of CHARTHAL.

Model	Question Type			Chart-Question Relation				Overall
	Desc.	Reason	Open	Irrel.	Inexist.	Contra.	Normal	
Proprietary Large Vision-Language Models								
Gemini-2.5-Pro	**60.31**	50.31	36.69	36.80	52.62	**49.52**	58.58	49.34
GPT-5 (high)	43.34	36.65	22.97	20.82	29.07	19.52	70.71	34.46
GPT-5	42.82	40.06	20.17	23.05	30.81	16.67	67.78	34.37
GPT-5-mini (high)	33.42	25.78	20.45	8.55	15.99	11.90	75.73	26.74
GPT-5-mini	34.20	24.53	21.57	8.18	15.99	9.05	**79.92**	27.02
GPT-5-nano (high)	33.68	20.19	19.89	10.41	17.73	5.24	69.04	24.95
GPT-5-nano	32.38	16.15	19.33	7.81	15.99	5.24	66.11	23.07
o4-mini (high)	34.46	23.60	9.52	10.78	14.53	3.33	65.27	22.79
o4-mini	32.90	22.67	10.92	11.52	12.21	1.90	67.36	22.41
GPT-4o-2024-11-20	40.47	36.02	20.73	27.88	36.92	6.67	53.97	32.49
Open-Source Large Vision-Language Models								
Llama-3.2-11B-Vision-Instruct	37.60	20.19	15.69	29.37	28.49	0.00	36.82	24.95
InternVL2.5-4B	19.58	9.63	8.12	4.83	11.34	0.00	34.73	12.71
InternVL2.5-8B	17.23	8.39	10.64	3.35	6.40	0.00	41.84	12.34
InternVL2.5-38B	28.72	18.32	10.36	10.04	21.80	0.00	43.51	19.40
InternVL2.5-78B	31.33	17.08	14.85	13.75	26.45	0.00	41.84	21.47
Qwen2.5-VL-3B-Instruct	13.05	4.04	7.84	0.74	0.87	0.00	35.98	8.57
Qwen2.5-VL-7B-Instruct	38.38	25.47	12.32	25.65	25.29	1.90	47.28	25.71
Qwen2.5-VL-32B-Instruct	36.81	25.47	8.68	21.56	25.29	1.43	44.35	23.92
Qwen2.5-VL-72B-Instruct	58.49	**59.63**	**44.82**	**65.06**	**69.77**	19.05	50.63	**54.24**

Table 2: Evaluation results on CHARTHAL. Scores are normalized to the range [0,100] for readability. The best and second-best results are marked in bold and underlined, respectively. "high" means setting reasoning effort as high.

they cannot be answered based on the given chart. They are more likely to trigger hallucinations since LVLMs are biased toward producing answers regardless of answerability (Kalai et al., 2025).

3.2 Dataset Construction

Based on our taxonomy, we build the dataset through a three-stage pipeline: (i) *image collection* by sampling representative charts from CharXiv (Wang et al., 2024c) to ensure diversity of types and domains; (ii) *QA pair generation* using o4-mini to generate questions and corresponding ground truth answers spanning all 12 subcategories for each selected chart; and (iii) *human verification* by expert annotators to refine answers and ensure quality. In total, we obtain 1,062 high-quality samples, with their distribution across all 12 categories presented in Table 1. Detailed sampling strategies, prompt designs, and verification procedures are provided in Appendix A.1.

3.3 Evaluation Criteria

To quantitatively evaluate the hallucination behavior of LVLM, we adopt a binary scoring system for CHARTHAL, where each response receives 1 if no hallucination occurs and 0 otherwise. In general, correct responses should either provide answers consistent with given ground truth or explicitly recognize when questions are unanswerable. Detailed evaluation criteria are included in Appendix A.2.

4 Experiments

4.1 Experimental Setups

We evaluate the performance of 15 state-of-the-art LVLMs on CHARTHAL. The proprietary models include Gemini-2.5-Pro (Comanici et al., 2025), GPT-5 series (OpenAI, 2025a), o4-mini (OpenAI, 2025b) and GPT-4o (Hurst et al., 2024). The open-source models include Llama 3.2-11B-Vision-Instruct (Meta, 2024), InternVL2.5 series (Chen et al., 2024) and Qwen2.5-VL series (Bai et al., 2025). The responses of these models are scored by GPT-4o according to the criteria in Sec 3.3. More detailed settings are provided in Appendix B.

4.2 Main Results and Analysis

Table 2 reports the results across all tested models.

Overall. Qwen2.5-VL-72B achieves the best overall score of 54.24%, outperforming Gemini-2.5-Pro by about 5 points. In contrast, o4-mini performs unexpectedly poorly, even trailing GPT-4o and only matching much smaller models like Qwen2.5-VL-7B. Its weakness mainly lies in unanswerable cases: only 14.53% of inexistent questions are correctly recognized as unanswerable (vs. 69.77% for Qwen2.5-VL-72B), and accuracy on contradictory questions drops to 3.33%. This suggests a training bias toward answer generation rather than answerability assessment.

Question Type Analysis. A clear difficulty hi-

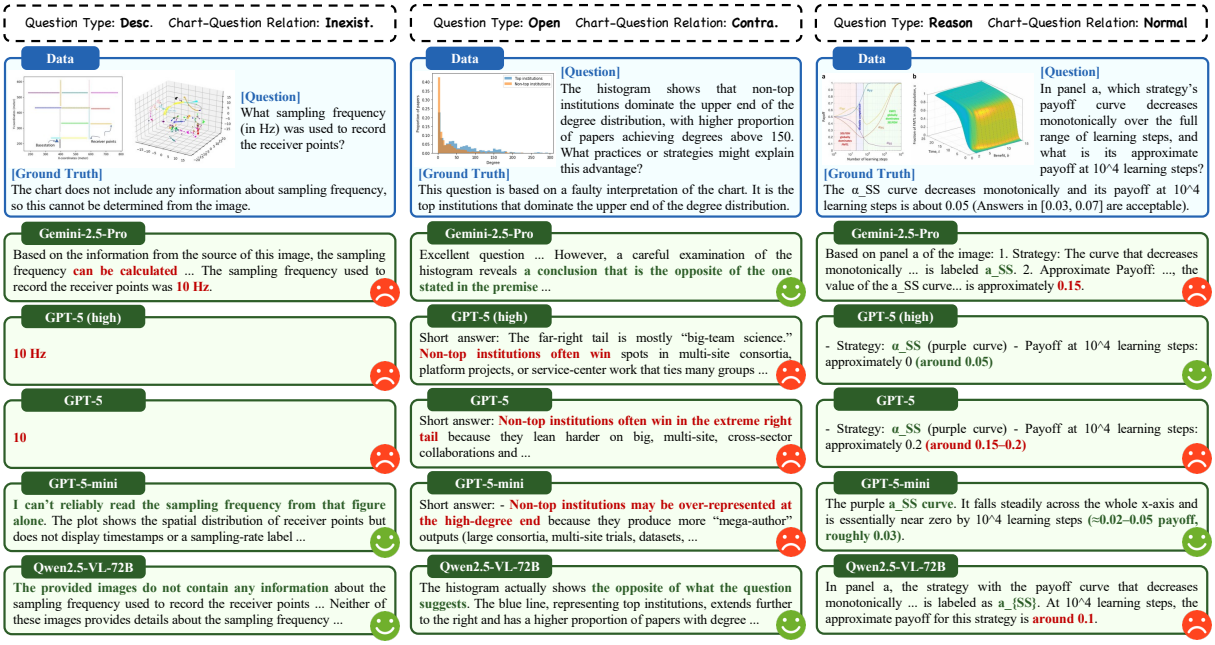

Figure 2: Sample cases of model responses on CHARTHAL. The correct parts are highlighted in green, while the incorrect or hallucinated parts are highlighted in red. Zoom in for better visual effects.

erarchy emerges: descriptive questions yield the highest scores (>60% for top models), reasoning questions are moderately challenging, while open-ended questions are the most error-prone, with even the strongest models struggling to exceed 45%.

Chart-Question Relation Analysis. Models perform reasonably on normal questions but hallucinate heavily when answerability issues arise. Contradictory questions are especially challenging, with many open-source models reaching 0% accuracy, showing that inconsistencies between charts and questions reliably induce fabricated responses.

Model Size Impact. Generally, larger models achieve better overall scores (Qwen2.5-VL and GPT-5 series), indicating that scale can enhance robustness to hallucinations comprehensively.

High Reasoning Impact. Increasing reasoning effort yields only small and inconsistent gains, and may even hurt performance, since correct reasoning relies on accurate visual perception. When the visual input is misinterpreted, extra reasoning often just amplifies the error rather than correcting it.

In summary, while LVLMs handle standard chart questions reasonably well, they remain highly vulnerable to hallucination in unanswerable scenarios. This exposes critical risks and highlights the need for strategies such as adversarial training to ensure trustworthy LVLM-based chart analysis.

5 Case Studies

We present several representative cases in Figure 2, from which we can observe that (i) increasing reasoning effort brings little benefit (GPT-5 in left and middle panels); (ii) the GPT-5 series are highly vulnerable to hallucination in unanswerable cases (left and middle panels); and (iii) while larger models generally perform better, smaller ones can occasionally outperform them in specific subcategories (GPT-5 vs. GPT-5-mini). These qualitative findings align well with the quantitative results in Table 2.

6 Conclusion

In this paper, we present CHARTHAL, the first systematic benchmark for evaluating hallucinations of LVLMs in chart understanding. With a carefully designed taxonomy of question types and chart-question relations, and a human-validated dataset covering 12 categories, our benchmark enables fine-grained analysis of LVLM hallucination behaviors in chart comprehension.

Experiments on 15 state-of-the-art LVLMs reveal critical vulnerabilities: while models handle answerable questions reasonably well, they often hallucinate in unanswerable cases, especially when confronted with contradictory premises or queried about inexistent information. Even advanced models like GPT-5 struggle to recognize unanswerability and tend to fabricate answers, underscoring the urgent need for robust mitigation strategies to make LVLMs more reliable in chart understanding.

Limitations

Although CHARTHAL offers the first fine-grained evaluation framework for hallucination, it still has several limitations:

Limited Dataset Scope. The dataset is built from scientific charts sampled from CharXiv, primarily originating from arXiv papers. Although these charts are complex and realistic, they do not capture the full diversity of chart types found in domains such as business, journalism, or education, which limits the benchmark's generalizability. Thus, one direction following this work could be expanding the dataset with charts from broader domains to enhance its coverage and applicability.

Insufficient Taxonomy Coverage. Our classification framework includes many scenarios that empirically trigger hallucinations, but it may not encompass all cases encountered in real-world usage. Users may pose more diverse, context-dependent, or adversarial questions beyond our benchmark's coverage. Therefore, future work could consider further expanding the taxonomy to better capture real-world scenarios.

Lack of Mitigation Evaluation. This work primarily diagnoses hallucination behaviors without systematically incorporating or comparing mitigation techniques. Future extensions could integrate and evaluate approaches such as adversarial training, calibration, or abstention to provide a more complete understanding of hallucination reduction.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

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

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

T Guan, F Liu, X Wu, R Xian, Z Li, X Liu, X Wang, L Chen, F Huang, Y Yacoob, and 1 others. 2023. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv.2310.14566*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 525–534.

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

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

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Junyoung Lim, Jaewoo Ahn, and Gunhee Kim. 2025. Chartcap: Mitigating hallucination of dense chart captioning. *arXiv preprint arXiv:2508.03164*.

Minzhi Lin, Tianchi Xie, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. 2025. Infochartqa: A benchmark for multimodal question answering on infographic charts. *arXiv preprint arXiv:2505.19028*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310, Mexico City, Mexico. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

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

OpenAI. 2025a. Gpt-5 system card.

OpenAI. 2025b. Openai o3 and o4-mini system card.

Han Qiu, Jiaxing Huang, Peng Gao, Qin Qi, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2024. Longhalqa: Long-context hallucination evaluation for multimodal large language models. *arXiv preprint arXiv:2410.09962*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Lingdong Shen, Kun Ding, Gaofeng Meng, Shiming Xiang, and 1 others. 2024. Rethinking comprehensive benchmark for chart understanding: A perspective from scientific literature. *arXiv preprint arXiv:2412.12150*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Sprague, and 1 others. 2025. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *arXiv preprint arXiv:2505.13444*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, and 1 others. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024b. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, and 1 others. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*.

Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2024. Quantity matters: Towards assessing and mitigating number hallucination in large vision-language models. *Preprint*, arXiv:2403.01373.

Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. 2023. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5718–5728, Singapore. Association for Computational Linguistics.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Details of the Benchmark

A.1 Dataset Construction Details

Based on our proposed taxonomy in Sec. 3.1, we construct the dataset through a three-stage pipeline that ensures both quality and diversity:

Image Collection. Given that the CharXiv dataset contains charts sourced from arXiv papers, which exhibit high complexity and present realistic challenges, we employ a carefully designed stratified sampling approach to extract a representative subset from CharXiv validation set. Specifically, since CharXiv charts contain varying numbers of subplots ranging from 1 to 120, and charts with excessive subplots tend to be overly complex for effective analysis and questioning, we selectively sample charts with 1 or 2 subplots at a 3:2 ratio. Subsequently, we apply proportional sampling across different academic disciplines to maintain domain diversity. Finally, we ensure comprehensive coverage of chart types (*e.g.*, bar charts, line graphs) to preserve the representativeness of the original CharXiv dataset.

Question-Answer Generation. We utilize the o4-mini model to automatically generate question-answer pairs spanning all 12 subcategories for each selected chart. The prompts employed for the generation process are provided in Table 3 to 14.

Human Verification. To ensure dataset quality, we recruit human annotators with undergraduate-level education to review and validate the generated question-answer pairs. Based on the corresponding chart content, annotators are instructed to discard pairs that clearly violate our design requirements and refine the remaining pairs through comprehensive improvements, including correcting factual errors, reassigning category labels, and refining ground truth answers (*e.g.*, providing acceptable ranges for questions where only approximate numerical values can be inferred from charts). This rigorous verification process ensures the high quality of our question-answer pairs.

A.2 Evaluation Criteria Details

Given the complex taxonomy of our benchmark, we establish specific evaluation criteria for each category to determine which responses meet expectations and which exhibit hallucination behaviors. Our evaluation framework employs a binary scoring system (1 for correct, 0 for incorrect) across six distinct evaluation scenarios:

Irrelevant Questions. For questions completely unrelated to chart content, correct responses must explicitly identify the irrelevance and refrain from attempting to answer. Responses that fail to recognize the mismatch or provide fabricated answers are marked as incorrect.

Inexistent Questions (Descriptive & Reasoning). When inexistent descriptive and reasoning questions seek information not present in the chart, appropriate responses should clearly state the absence of requested data and avoid speculation. Responses that ignore missing information or provide fabricated answers receive a score of 0.

Inexistent Questions (Open-ended). Similar to closed-ended cases, but acceptable responses may either declare the question unanswerable or provide appropriate speculative answers with explicitly uncertainty expressions. Definitive answers without acknowledging the nonexistence of the queried information are considered incorrect.

Contradictory Questions (Descriptive & Reasoning). For questions containing statements that contradict chart data, correct responses must identify the contradiction and either state the question is unanswerable or provide answers aligned with ground truth while ignoring the false premise. Numeric answers are evaluated against specified acceptable ranges when provided.

Contradictory Questions (Open-ended). Responses must recognize contradictions and either declare unanswerability or offer tentative answers with appropriately uncertain language. Overconfident or absolute responses are marked incorrect.

Normal Questions. For standard answerable questions, evaluation varies by answer type: (1) Numeric questions require exact value matches unless acceptable ranges are specified; (2) Questions with categorical answers expect model responses to match ground truth terms exactly, allowing different representations; (3) Open-ended questions are evaluated based on appropriate uncertainty expression and plausible reasoning rather than factual correctness; (4) Other questions are evaluated based on semantic alignment with ground truth without omissions or extraneous information.

B Details of Experiments

To better reflect real-world scenarios, we adopt a different setting than CharXiv (Wang et al., 2024c), which provides models only with chart images and questions, **without** any additional instructions (*e.g.*, "*If you believe the question is unanswerable, output 'not applicable'*" in CharXiv).

While generating model responses in evaluation, we sample one response for each question and set the sampling temperature to 0 to ensure reproducible results, except for reasoning models like o4-mini which do not permit temperature configuration. For the token generation limits, we use the default setting of each model to observe potential hallucination phenomena in model responses comprehensively. The open-source models adopted for evaluation are deployed on A6000 GPU with vLLM Library (Kwon et al., 2023).

During evaluation, we adopt GPT-4o-2024-11-20 with temperature set to 0 to score model responses according to the criteria in Appendix A.2. We'd like to note that using GPT-4o for evaluation and grading is a common practice in many recent benchmarks (Xu et al., 2023; Wang et al., 2024c; Guan et al., 2023; Sun et al., 2024), as it has demonstrated strong alignment with human judgments while being more efficient and scalable. The detailed prompts for guiding GPT-4o to perform the evaluation are provided in Table 15 to 21.

C Compliance

The chart images in CHARTHAL are sourced from the CharXiv (Wang et al., 2024c) dataset, which is open-sourced under CC BY-SA 4.0 License. We will also provide the arXiv paper preprint numbers corresponding to the charts used in our published dataset. The licenses of evaluated open-source models are provided as follows: (1) Llama 3.2-11B-Vision-Instruct (Meta, 2024) is under Llama 3.2 Community License; (2) InternVL2.5 series (Chen et al., 2024) are under Qwen and MIT License; (3) Qwen2.5-VL series (Bai et al., 2025) are under Qwen and Apache-2.0 License. The datasets and models used in this paper are intended solely for our research. Accordingly, we also expect the datasets we publicly release to be used strictly for research purposes.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a question that appears to be a simple, factual query about the chart, as if the answer could be directly extracted from it. However, make sure that the question is actually unrelated to the chart content (i.e. the question should refer to a topic that does not exist in the chart at all), so that no correct answer can be found in the image. The question should still sound plausible and specific, and it should be a closed question that expects a concrete answer. Answer: Please generate a short response to the question, pointing out that the topic or element asked about is not present in the chart, and therefore the question cannot be answered based on the chart. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 3: Prompt template for generating Descriptive-Irrelevant question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a question that appears to be a simple, factual question related to the chart. The question should be consistent with the chart's overall topic or domain, but it must ask about some specific detail—such as a category, label, year, or variable—that is not actually present in the chart. The question should seem plausible and answerable from the chart at first glance, but in reality the necessary information is missing. Answer: Please generate a short response that explains the question refers to a detail not shown in the chart (e.g., a year, category, or variable), and therefore it cannot be answered based on the chart data. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 4: Prompt template for generating Descriptive-Inexistent question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a question that begins with a factual statement that appears to be based on the chart but is actually contradicted by the chart. Then, ask a simple, descriptive question based on that incorrect premise. The question should sound plausible, as if it could be answered by referring to the chart, but it should be logically invalid due to the contradiction in the premise. Make sure the contradiction is in the premise statement (not in the actual question). Do not directly ask a false question. Instead, provide a false statement first, then follow it up with a concrete, factual question based on it. Answer: Please generate a short response that points out the contradiction in the premise compared to the actual chart data, clarifies what the correct chart information is, and provide the correct answer ignoring the false premise. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 5: Prompt template for generating Descriptive-Contradictory question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a simple, factual question that can be directly answered using the information shown in the chart. The question should be clearly related to the chart, refer to elements (such as labels, data points, categories, or values) that are actually present, and involve no inference or assumptions beyond what is visually provided. Make sure the question is correct and logically consistent with the chart. Answer: Please generate a short, accurate answer that directly uses the correct data from the chart to respond to the question. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 6: Prompt template for generating Descriptive-Normal question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a reasoning-based question that appears to require simple inference from the chart (such as comparing values, calculating differences, or identifying trends). However, the question must actually be unrelated to the content of the chart. That is, the question should refer to a topic that does not exist in the chart at all. Make it sound as if the chart could support reasoning to answer the question, but in reality, the necessary information is completely absent.Ensure that the reasoning required is straightforward and answerable in principle, but the problem is unanswerable here because the chart has nothing to do with the topic. Answer: Please generate a short response pointing out that the question refers to a topic that is not present in the chart, and that no inference can be made from the given data. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 7: Prompt template for generating Reason-Irrelevant question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a reasoning-based question that is clearly related to the chart's topic or domain and requires simple inference (such as computing a difference or trend). However, make sure the question asks about specific information—such as a year, group, variable, or category—that is not actually present in the chart. The question should sound plausible and as if it could be answered using the chart, but in reality, the key data needed for inference is missing. Answer: Please generate a response that explains the question cannot be answered because the required data (year, group, variable, etc.) is missing from the chart. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 8: Prompt template for generating Reason-Inexistent question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a question that begins with a statement about the chart, where the statement contradicts the actual data shown. Then, based on that false statement, ask a reasoning-based question that would normally require a simple inference (e.g., computing a difference or identifying a trend). The question should sound logical and answerable, but the initial premise must be factually incorrect according to the chart. Do not contradict the chart in the actual question; instead, embed the contradiction in the premise statement. Ensure that the reasoning expected is straightforward, even though the basis is flawed. Answer: Please generate a short response that points out the contradiction in the premise compared to the chart, and give the correct answer to the question if the question is answerable ignoring the false premise. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 9: Prompt template for generating Reason-Contradictory question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate a reasoning-based question that is logically consistent with the chart and can be answered by using the information provided. The question should involve simple inference, such as comparing values, identifying trends, computing differences or ratios, or summarizing grouped data. Make sure the question only refers to elements that are actually present in the chart, and that it is objectively answerable. Answer: Please generate a short, accurate answer that performs the correct inference or calculation using data from the chart, and presents the result clearly. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 10: Prompt template for generating Reason-Normal question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate an open-ended question that seems to ask for an explanation, reasoning, or opinion based on a chart, but in fact, the topic of the question is entirely unrelated to the content of the chart. The question should sound like it could be asked in response to a chart, but the actual theme (e.g., domain, topic, entities) should not appear in the chart at all. Avoid asking about specific information. Instead, ask for explanations, causes, interpretations, or predictions related to a topic that the chart does not cover. Answer: Please generate a short response pointing out that the topic mentioned in the question is not covered in the chart, so no meaningful explanation can be provided based on the chart data. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 11: Prompt template for generating Open-Irrelevant question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate an open-ended question that appears to ask for an explanation or interpretation of the chart, but it must be based on information—such as a year, category, variable, or region—that is not actually present in the chart. The question should be relevant to the general topic of the chart and sound plausible, but should include some missing or fabricated element that prevents it from being truly answerable. Do not include specific numerical queries or factual comparisons. Focus on asking for explanations, causes, interpretations, or predictions, but make sure they hinge on something not found in the chart. Answer: Please generate a response pointing out the specific element referred to is not shown in the chart, so any explanation would be purely hypothetical. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 12: Prompt template for generating Open-Inexistent question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate an open-ended question that begins with a statement about the chart, where the statement clearly contradicts the actual information shown. Then, based on this false premise, ask an analytical or interpretive question. The question should sound plausible and related to the chart's topic, but the initial assumption must be incorrect according to the chart. The question should not request a specific fact, but rather an explanation, a cause, an interpretation, or a prediction (e.g., why something happened, what it implies, etc.). The contradiction must be in the premise, not the question itself. Answer: Please generate an answer pointing out the false assumption in the question's premise, and suggests that any analysis based on it may be invalid or misleading. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 13: Prompt template for generating Open-Contradictory question-answer pairs.

Prompt
Given an image of a chart, your task is to generate a question and a corresponding reference answer, following the specific instructions below. Instruction for Generation: Question: Please generate an open-ended question that is logically consistent with the chart and invites explanation, interpretation, or speculation based on the data shown. The question should reference trends, patterns, comparisons, or anomalies that are truly present in the chart, and should encourage divergent thinking rather than factual retrieval and calculation. Avoid factual questions and focus on asking for explanations, causes, interpretations, or predictions. Answer: Please generate a thoughtful answer that responds to the question using evidence from the chart. The response should be tentative and interpretive, not definitive. Output Format: Question: <Insert the generated question here> Answer: <Insert the generated answer here>

Table 14: Prompt template for generating Open-Normal question-answer pairs.

Prompt
You are given a question about a chart and a corresponding response. The question is completely unrelated to the chart. Your task is to evaluate whether the response correctly identifies that the question cannot be answered from the chart. Guidelines: - The response should explicitly indicate that the question is unrelated to the chart. - It should not attempt to fabricate an answer. - If the response fails to point out the irrelevance or tries to answer the question, it should be considered incorrect. Question: <|QUESTION|> Response: <|RESPONSE|> Score the response above as follows: - 1 = Correctly identifies irrelevance and avoids answering. - 0 = Does not identify irrelevance OR tries to answer the question. DO NOT provide any explanation or additional text. Just return the score as a single integer. If the response is empty, give a score of 0. Output Format: Score: <Insert your score here, either 0 or 1>

Table 15: Prompt template for grading Irrelevant questions.

Prompt
You are given a question about a chart and a corresponding response. The question is related to the chart but asks about information that does not exist in the chart. Your task is to evaluate whether the response correctly identifies that the information is missing and that the question cannot be answered from the chart. Guidelines: - The response should clearly state that the requested information does not exist in the chart. - It should not attempt to guess or fabricate an answer. - If the response ignores the missing information issue or provides an answer, it should be considered incorrect. Question: <|QUESTION|> Response: <|RESPONSE|> Score the response above as follows: - 1 = Correctly identifies that the information is not in the chart and avoids guessing. - 0 = Does not identify missing information OR gives a fabricated answer. DO NOT provide any explanation or additional text. Just return the score as a single integer. If the response is empty, give a score of 0. Output Format: Score: <Insert your score here, either 0 or 1>

Table 16: Prompt template for grading Descriptive/Reasoning-Inexistent questions.

Prompt
You are given a question about a chart and a corresponding response. The question is related to the chart but asks about information that does not exist in the chart. Your task is to evaluate whether the response correctly identifies that the information is missing and and either states the question is unanswerable or provides a speculative answer. Guidelines: - The response should clearly state that the requested information does not exist in the chart. - It should either state that the question is unanswerable OR provide a speculative answer with an speculative tone. - If the response ignores the missing information issue and directly provides an definitive answer, it should be considered incorrect. Question: <|QUESTION|> Response: <|RESPONSE|> Score the response above as follows: - 1 = Correctly identifies that the information is not in the chart and avoids guessing. - 0 = Does not identify missing information OR gives a fabricated answer. DO NOT provide any explanation or additional text. Just return the score as a single integer. If the response is empty, give a score of 0. Output Format: Score: <Insert your score here, either 0 or 1>

Table 17: Prompt template for grading Open-Inexistent questions.

Prompt
You are given a question about a chart, a corresponding response, and a ground truth answer. The question contains a statement that contradicts the chart, and then asks a follow-up based on that contradiction. Your task is to evaluate whether the response correctly identifies the contradiction and either states the question is unanswerable or provides a correct answer ignoring the contradiction in the question. Guidelines: - The response should explicitly point out that the question contains incorrect or contradictory information relative to the chart. - It should either state that the question is unanswerable OR provide a correct answer that aligns with the given Ground Truth. - If the ground truth does not provide an answer, the response should also not provide an answer. - When the response tries to provide an answer, for questions expecting numerical answers, if the ground truth specifies an acceptable range (e.g., “Answers in the range of [35%, 40%] are ACCEPTABLE” or similar phrasing), then the answer is considered correct if it falls within this range. Otherwise, the semantic meaning of the answer should match the ground truth without omissions or additional unrelated information. - If the response fails to mention the contradiction OR answers the question based on the contradictory statement, it is incorrect. Question: <|QUESTION|> Ground Truth: <|REF_ANS|> Response: <|RESPONSE|> Score the response above as follows: - 1 = Correctly identifies the contradiction AND either states the question is unanswerable or provides a correct answer that matches the given Ground Truth. - 0 = Does not mention the contradiction OR answers the question as if it were valid. DO NOT provide any explanation or additional text. Just return the score as a single integer. If the response is empty, give a score of 0. Output Format: Score: <Insert your score here, either 0 or 1>

Table 18: Prompt template for grading Descriptive/Reasoning-Contradictory questions.

Prompt
You are given a question about a chart and a corresponding response. The question contains a statement that contradicts the chart, and then asks a follow-up based on that contradiction. Your task is to evaluate whether the response correctly identifies the contradiction and either states the question is unanswerable or provides a tentative answer ignoring the contradiction in the question. Guidelines: - The response should explicitly point out that the question contains incorrect or contradictory information relative to the chart. - It should either state that the question is unanswerable or provide a speculative answer with an appropriately uncertain tone. - If the response fails to mention the contradiction, answers the question based on the contradictory statement, sounds overconfident OR states an absolute answer, it is incorrect. Question: <|QUESTION|> Response: <|RESPONSE|> Score the response above as follows: - 1 = Correctly identifies the contradiction AND either states the question is unanswerable or provides a tentative answer with an appropriately uncertain tone. - 0 = Does not mention the contradiction OR answers the question as if it were valid OR sounds overconfident. DO NOT provide any explanation or additional text. Just return the score as a single integer. If the response is empty, give a score of 0. Output Format: Score: <Insert your score here, either 0 or 1>

Table 19: Prompt template for grading Open-Contradictory questions.

Prompt
You are given a question about a chart, a corresponding response, and a ground truth answer. The question is answerable from the chart and has a definitive correct answer. Your task is to evaluate how accurate the model response is compared to the ground truth. 1. **If the expected answer is related to numeric values:** - If the ground truth specifies an acceptable range (e.g., “Answers in the range of [35%, 40%] are ACCEPTABLE” or similar phrasing), then: - Give a score of 1 if the response contains a numeric value within this range. - Otherwise, give a score of 0. - If no range is specified: - Give a score of 1 if and only if the two numbers are exactly equal in value. Different notations (e.g., ‘0.01’ and ‘10^-2’, ‘1500’ and ‘1.5e3’) are acceptable. - Give a score of 0 if the numbers differ in value. 2. **If the expected answer is related to predefined terms:** - Give a score of 1 if and only if the final answer and the ground truth refer to the same term(s). Different forms or representations (e.g., α and alpha; $R^2_{t,h,v,m}$ and R^2_t,h,v,m) are acceptable. - If the question requires multiple terms, different order is acceptable as long as all terms match exactly. - Give a score of 0 if any required term is incorrect, missing, or replaced with a different term (e.g., ACC+ vs ACC; P-101 vs P=101). 3. **For all other cases:** - Give a score of 1 only if the semantic meaning of the model response matches the ground truth without omissions or additional unrelated information. - Any deviation in meaning should result in a score of 0. Question: <|QUESTION|> Ground Truth: <|REF_ANS|> Response: <|RESPONSE|> DO NOT provide any explanation or additional text. Just return the score as a single integer. If the response is empty, give a score of 0. Output Format: Score: <Insert your score here, either 0 or 1>

Table 20: Prompt template for grading Descriptive/Reasoning-Normal questions.

Prompt
You are given a question and a model response. The question is open-ended and does not have a definitive answer. Your task is to evaluate whether the response uses an appropriately uncertain tone and avoids giving a misleadingly definitive answer. Guidelines: - The response should acknowledge uncertainty, use hedging language (e.g., "possibly," "likely," "one reason could be"). - It should not present opinions as facts or give definitive causes. - If the response sounds overconfident or states an absolute answer, it is incorrect. Question: <|QUESTION|> Response: <|RESPONSE|> Score the response above as follows: - 1 = Maintains uncertainty and provides plausible reasoning. - 0 = Presents a definitive or overconfident answer without acknowledging uncertainty. DO NOT provide any explanation or additional text. Just return the score as a single integer. If the response is empty, give a score of 0. Output Format: Score: <Insert your score here, either 0 or 1>

Table 21: Prompt template for grading Open-Normal questions.