Chart-HQA: A Benchmark for Hypothetical Question Answering in Charts

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have garnered significant attention for their strong visual-semantic understanding. Most existing chart benchmarks evaluate MLLMs' ability to parse information from charts to answer questions. However, they overlook the inherent output biases of MLLMs, where models rely on their parametric memory to answer questions rather than genuinely understanding the chart content. To address this limitation, we introduce a novel Chart Hypothetical Question Answering (HQA) task, which imposes assumptions on the same question to compel models to engage in counterfactual reasoning based on the chart content. Furthermore, we introduce HAI, a human-AI interactive data synthesis approach that leverages the efficient text-editing capabilities of LLMs alongside human expert knowledge to generate diverse and high-quality HQA data at a low cost. Using HAI, we construct Chart-HQA, a challenging benchmark synthesized from publicly available data sources. Evaluation results on 18 MLLMs of varying model sizes reveal that current models face significant generalization challenges and exhibit imbalanced reasoning performance on the HQA task. Our codebase and newly generated datasets are available at https://anonymous.4open.science/r/Chart-HQA-86BE.

1 Introduction

Multimodal Large Language Models (MLLMs) (Li et al., 2023a; Liu et al., 2023e) have demonstrated exceptional performance in visual-semantic understanding (OpenAI, 2023; Wang et al., 2023). Despite their success, existing MLLMs still face significant challenges in reading, understanding, and summarizing visual charts (Masry et al., 2022; Li and Tajbakhsh, 2023). Unlike natural images, which primarily rely on recognizable objects, relative positions, and interactive relationships to convey information, charts communicate complex semantic meanings through visual logic (Xu et al., 2024), such as trend lines, color-coded legends, and axis structures.

Most existing chart question answering benchmarks (Masry et al., 2022; Xia et al., 2024b) mainly focus on factoid question answering (FQA), where the model is required to directly extract information from the chart image to answer question, as shown in Figure 1.a. Although these benchmarks have made significant progress in expanding dataset scale (Xia et al., 2024a) and diversifying chart types (Xu et al., 2024), **they overlook the inherent output biases problem of MLLMs** (Huang et al., 2024; Guan et al., 2023), i.e., MLLMs tend to rely on their parametric memory to answer questions rather than interpreting the visual content of the chart. Taking the widely used multimodal model LLaVA-1.5 (Liu et al., 2023c) as an example, as shown in Figure 1, although LLaVA-1.5 correctly answered the question in the FQA task (Figure 1.a), the model still produced the same output when the counterfactual image was provided (Figure 1.b) or even the chart image was missing (Figure 1.c). This phenomenon indicates that introducing additional control conditions (e.g., missing images, counterfactual images) for the same chart question can effectively reveal the output bias of MLLMs, thereby reflecting their true understanding of charts. Unfortunately, to the best of our knowledge, no existing chart benchmarks have been designed to thoroughly investigate such problem.

To fill this gap, we propose an novel **hypothetical question answering** (HQA) task in the domain of chart understanding. Unlike directly modifying chart images as a control condition, we focus on imposing an assumption on the original chart question. As shown in Figure 1.d, the proposed HQA task requires models to independently *imagine* the corresponding counterfactual details based on the given assumption and original chart image, thereby establishing an accurate inference context. The

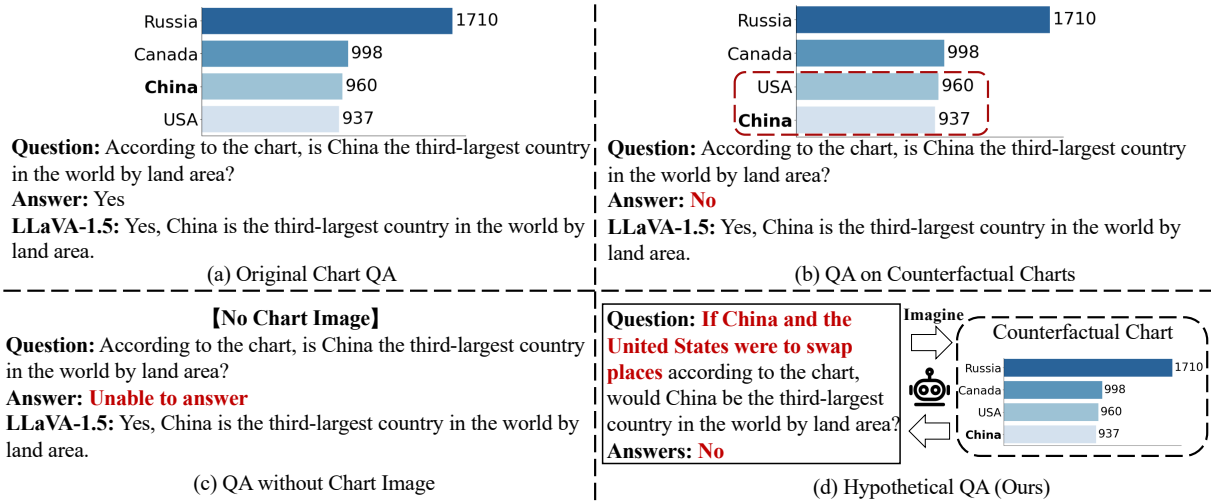


Figure 1: An example of biased output on charts from MLLMS and proposed hypothetical QA task. (a) Factoid QA results based on the original chart. (b) The response after counterfactual editing of the chart, where the land areas of "China" and "USA" are swapped. (c) The model's answers without the chart image input. (d) Illustration of hypothetical question and the corresponding counterfactual context to be imagined.

HQA task will undoubtedly enhance the practical use of MLLMs due to the universality of hypothetical questions in real-world scenarios.

However, constructing a high-quality chart HQA benchmark is not trivial. While a straightforward approach is to utilize human experts for data synthesis, existing research (Wang et al., 2022) has shown that human-generated data suffer from **limited diversity**. Specifically, most human-generated hypothetical questions tend to focus on common chart attributes such as specific data, falling short of covering a true variety of assumption types and different ways to describe them. Secondly, the same hypothetical scenario may not be applicable to different chart types and could even lead to **conflicting layout structures**. For example, the assumption "*suppose a specific value in the chart doubles*" is reasonable in a bar chart. However, in a pie chart, this assumption violates the structural constraint that all slices must sum to 100%. Undoubtedly, such structurally conflicting hypothetical questions significantly reduce the practical applicability of the HQA benchmark.

To overcome above challenges, we propose a human-machine interactive HQA data synthesis method named **HAI**. HAI combines the efficient text editing capabilities of LLMs with human expert knowledge to synthesize diverse and high-quality HQA data at a low cost. Specifically, HAI consists of two key components: **(1) Counterfactual proposal generator** (CIG). To diversify counterfactual assumptions, the CIG module randomly samples a subset of instructions from the seed instruction set (initially composed of limited manual instructions) and inputs them into the LLM along with the detailed description of charts to generate new instruction proposals and HQA instances. **(2) Human-feedback discriminator** (HFD). This module employs multiple human experts to review generated HQA instances from various perspectives, including answer accuracy, layout consistency, and question clarity. Subsequently, HQA instances validated by human experts are retained. Furthermore, leveraging the self-reflection capability (Shinn et al., 2023) of LLMs, the corresponding instruction proposals are revised based on human expert feedback, thereby expanding the seed instruction set. Based on the proposed method, we construct **Chart-HQA**, a challenging HQA benchmark derived from factoid QAs in ChartQA (Masry et al., 2022). There are ***900*** counterfactual instruction proposals and ***4*** answer types within Chart-HQA. We evaluate the zero-shot reasoning capabilities of **18** MLLMs of varying model sizes on Chart-HQA, including **8** specialist chart-based models and **10** generalist models. The results (as shown in Table 3) indicate that existing models generally exhibit limited reasoning capabilities in chart hypothetical question answering. For instance, the high-performing GPT-4o (OpenAI et al., 2024) experiences a significant performance drop, with relaxed accuracy decreasing from *85.7%* on ChartQA to *62.52%* on Chart-HQA. Additionally, we observe that most models demonstrate imbalanced performance across different answer types within Chart-HQA, highlighting potential avenues for op-

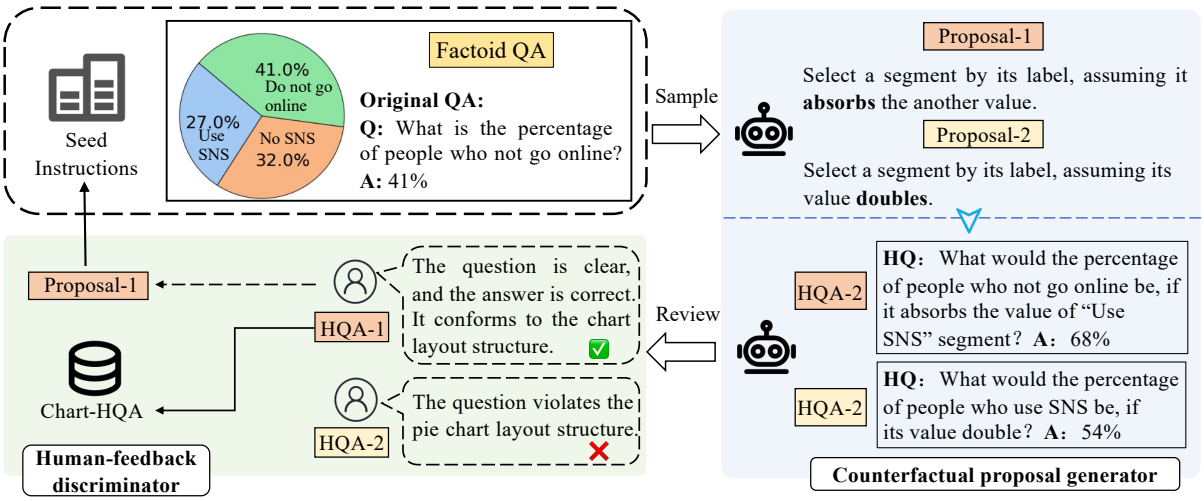


Figure 2: The illustration of our approach for synthesizing hypothetical questions, including two stages that synthesize new instruction proposals, and human verification.

timizing future MLLMs.

In summary, our main contributions are as follows:

- We propose a novel chart hypothetical question answering task to evaluate the true understanding capabilities of MLLMs on chart-based reasoning.
- We propose a human-machine interactive HQA data synthesis framework named HAI. It leverages LLMs to automatically generate diverse and high-quality HQA data under the guidance of human feedback.
- We unveil Chart-HQA, a challenging benchmark for chart hypothetical question answering. Extensive experiments demonstrate that existing MLLMs are generally inadequate in counterfactual chart comprehension abilities.

2 Human-machine Interactive Data Synthesis

In this section, we detail our unique method that synthesizes high-quality hypothetical QA for chart-based visual question answering, named HAI. As shown in Figure 2, our method consists of two interconnected modules, a Counterfactual proposal generator (CPG), and a Human-feedback discriminator (HFD). In section 2.1, we present how the CPG module iteratively generate new instruction proposals and HQA instances. In section 2.2, we introduce the HFD module to leverage human expert knowledge for validating HQA instances.

2.1 Counterfactual Proposal Generator

The primary goal of our method is to automatically generate diverse and high-quality Hypothetical Question Answering (HQA) instances using large language models (LLMs). However, directly leveraging LLMs to annotate large-scale HQA data is challenging, as it requires: (1) creatively formulating novel counterfactual operations based on the rich attributes of charts; (2) professionally composing logically consistent hypothetical questions aligned with the given context. Empirically, even when provided with detailed task descriptions and examples, LLMs tend to repetitively reference prior data, lacking valuable insights essential for designing diverse HQA data. Therefore, we propose the Counterfactual Proposal Generator to encourage the LLM to generate general instruction proposals that describe common counterfactual operations for different types of charts. Specifically, this component first initializes an instruction proposal pool, in which four seed proposals are manually crafted for each chart type. Then, given a set of general descriptions of chart attributes D_C, the module utilizes GPT-4 to generate counterfactual instruction proposals I_P, formulated as follows:

$$I_P = \textit{GPT-4}(I_S, P_I, D_C), \quad (1)$$

where I_S represents a sampled seed proposal from the instruction proposal pool as a contextual example, and P_I is a guiding prompt for GPT-4, with the specific prompt details presented in Appendix A.

After obtaining a diverse set of instruction proposals, the CPG module further utilizes generated proposals to generate specific hypotheti-

cal question-answering instances, ensuring their alignment with the context of particular factual chart-based questions. Notably, each hypothetical question corresponds to a factual question from ChartQA, but not every factual question in ChartQA has a corresponding hypothetical question. To generate specific HQA instances, we use the ChartQA annotation JSON file A_C, which contains structured metadata about the chart's content. Our empirical findings indicate that such structured textual representations provide LLMs with a more effective understanding of chart information compared to direct image inputs. Subsequently, we design an appropriate prompt to guide GPT-4 in generating specific HQA data based on the previously generated instruction proposals I_P, the original factual QA instance QA_O, and the chart annotations A_C, as follows:

$$QA_H = \textit{GPT-4}(I_P, QA_O, P_H, A_C) \quad (2)$$

where P_H represents a prompt template, its specific content detailed in Appendix A.

2.2 Human-feedback Discriminator

Although the counterfactual proposal generation component carefully utilizes the LLM to generate HQA data, it is inevitable that some low-quality instances may be produced, containing contradictions with the inherent structure of charts or unreasonable assumptions. To address this issue, we adopt a human validation process to ensure the quality of the generated HQA instances. Specifically, we recruit seven experts with professional knowledge of chart interpretation to review the generated HQA instances. Given the HQA instances produced by the counterfactual hypothesis proposal generation component, the reviewers are required to assess whether the generated counterfactual hypotheses align with the structural properties of the chart and whether the corresponding answers are correct. For instance, as illustrated in Figure 2, the reviewers evaluate the generated HQA instances from multiple perspectives, including the reasonableness of the question, accuracy of the answer, and complexity of the reasoning process. Based on human judgments, only validated HQA instances are retained, and the corresponding human feedback is incorporated into the instruction proposals before being added back to the instruction proposal pool. As a result, 63.4% of the generated HQA data successfully passed validation. Specific examples of HQA data review can be found in the appendix B.

Benchmarks	Chart	Question	QA formats
	Real-world	Hypothetical	Open-ended
Figure QA	✗	✗	✗
DVQA	✗	✗	✗
LEAF-QA++	✗	✗	✗
PlotQA	✗	✗	✓
ChartLlama	✗	✗	✓
MMC	✗	✗	✓
ChartQA	✓	✗	✓
ChartBench	✓	✗	✓
ChartX	✓	✗	✓
Chart-HQA (ours)	✓	✓	✓

Table 1: Comparison between existing benchmarks and our new Chart-HQA benchmark.

Statistic	**Number**
# of hypothetical questions	2173
# of instruction proposals	900
# of charts	947
# of answer types	4
Avg. Character per question	149.14
Avg. Character per assumption	82.10
Avg. Character per answer	6.29

Table 2: Key statistics for Chart-HQA.

3 Dataset Analysis

We apply the proposed data synthesis method to change questions from widely used benchmarks ChartQA (Masry et al., 2022) test-split to be hypothetical by adding a related assumption. The generated HQA benchmark is named Chart-HQA. We present the data analysis for Chart-HQA as below.

3.1 Comparison to Existing Benchmarks

As shown in Table 1, Chart-HQA differs from related benchmarks in various aspects: (1) Chart-HQA is the first benchmark to study hypothetical problems over chart context on open domains; (2) Questions in Chart-HQA are generated automatically by LLMs, which greatly reduces data construction costs. (3) Chart-HQA has an open-vocabulary QA format that requires applying counterfactual operations on the underlying chart data.

3.2 Key Statistics

The main statistics for Chart-HQA are shown in Table 2. The Chart-HQA benchmark contains 2172 hypothetical questions, which are all used to test zero-shot chart-based visual question answering. There are 900 instruction proposals, indicating that Chart-HQA has a rich diversity in the hypothetical problem distribution. The assumptions have an

Figure 3: Counterfactual operations in generated instruction proposals. The inner circle denotes noun objects in charts, the outer circle represents the action against the noun object.

average of 82.1 characters in length, showing that they have lexical richness. There are four answer types in Chart-HQA, and the answer could be a text span, an integer number, a decimal number, or a boolean answer. These statistics suggest that models need diverse symbolic reasoning abilities to answer the questions in Chart-HQA.

3.3 Proposal Diversity

We further demonstrate the diversity of generated instruction proposals. We identify the counterfactual operations in generated instruction proposals and then extract the verb-noun structure in the counterfactual operation using Berkeley Neural Parser (Kitaev and Klein, 2018; Kitaev et al., 2019). We randomly parse 10 generated instruction proposals for each chart style. As shown in Figure 3, we can see quite diverse intents and textual formats in these instruction proposals. Notably, the generated counterfactual operations adeptly capture the distinctive characteristics of data visualization. For instance, the counterfactual operation of "reversing trend" for the line chart introduces novel challenges to the model's domain knowledge.

4 Experiments

In this section, we conduct zero-shot transfer experiments on the proposed Chart-HQA (Section 4.2). In addition, we further perform a fine-grained analysis of MLLMs based on answer types (Section 4.3). Before discussing results, we provide details of the experimental setup below.

4.1 Experimental Settings

Models. We evaluate two types of models: (1) chart-oriented specialist models. This type of model specializes in pre-training on large amounts of chart data. we evaluate end-to-end models including Pix2struct, MatCha , Unichart, ChartLlama, ChartVLM, TinyChart and Docowl2.0. We also evaluate a tool-augmented model Deplot [1]. (2) generalist models, which are trained towards general capability for various vision-language tasks. The open-source models include Monkey, Qwen-VL-Chat, DeepSeek-VL, Qwen2.5-VL and InternVL-2.5, while the closed-source models contain Qwen-VL-Max, Gemini-Pro, GPT-4V, GPT-4o. For open-source models, we re-implement the results using the official codes. For closed-source models, we re-implement the results using the official APIs.

Metrics. To ensure a fair comparison with ChartQA results, we adopt the same evaluation method and metric used in ChartQA. Specifically, we choose the *relaxed accuracy* used in ChartQA as the evaluation metric, which means exact match accuracy with 5% tolerance on numerical error is used to report all QA results. In addition, we compute the *decline rate* to measure the performance difference of models between ChartQA and Chart-HQA, which is calculated as follows:

$$\text{Decline Rate} = \frac{|Acc_{QA} - Acc_{HQA}|}{Acc_{QA}} \times 100\%. \quad (3)$$

4.2 Zero-shot Transfer on Chart-HQA

Chart-HQA establishes a highly challenging benchmark for visual chart understanding. Table 3 compares various MLLMs on the ChartQA and our Chart-HQA. First, chart-specialist models exhibit severe generalization issues when handling counterfactual assumptions added to questions while keeping chart images unchanged. For example, large-scale specialist models with over 10B parameters, such as ChartVLM-L and ChartLlama, exhibit significant declines of 63.70% and 75.28% respectively on Chart-HQA compared to their strong performance on ChartQA. Second, generalist MLLMs also demonstrate limited capabil-

[1]We use the GPT3.5 (OpenAI, 2021) as the inference model of Deplot for our experiments.

Models	Type	#Params	ChartQA		Chart-HQA		Decline Rate (↓)
			Acc (↑)	Rank	Acc (↑)	Rank	
Specialist Models							
Pix2struct (Lee et al., 2023)	End-to-End	0.3B	56.00	#18	17.68	#17	68.43
MatCha (Liu et al., 2023b)	End-to-End	0.3B	64.20	#15	21.32	#14	66.79
Unichart (Masry et al., 2023)	End-to-End	0.2B	66.24	#13	18.69	#16	71.78
TinyChart (Zhang et al., 2024)	End-to-End	3B	83.60	#5	30.79	#11	63.17
DocOwl-v2.0 (Hu et al., 2024)	End-to-End	8B	70.00	#10	30.83	#10	55.96
ChartLlama (Han et al., 2023a)	End-to-End	13B	69.66	#11	17.22	#18	75.28
ChartVLM-L (Xia et al., 2024a)	End-to-End	14.3B	62.28	#16	20.15	#15	63.70
DePlot(GPT 3.5) PoT SC (Liu et al., 2023a)	Tool-augment	-	76.70	#8	49.49	#6	35.48
Generalist Models							
Qwen-VL-Chat (Bai et al., 2023)	End-to-End	7B	66.30	#12	28.60	#12	56.86
DeepSeek-VL-Chat (Lu et al., 2024)	End-to-End	7B	60.72	#17	27.93	#13	54.00
Qwen2.5-VL (Team, 2025)	End-to-End	7B	**87.30**	**#2**	**57.20**	**#3**	34.48
InternVL-v2.5-8B (Chen et al., 2025)	End-to-End	8B	84.80	#4	48.23	#7	43.13
Monkey (Li et al., 2023b)	End-to-End	9.8B	65.10	#14	31.45	#9	51.69
Qwen2.5-VL (Team, 2025)	End-to-End	72B	**89.50**	**#1**	**66.41**	**#1**	**25.80**
Gemini-Pro (Team et al., 2023)	End-to-End	-	74.10	#9	41.25	#8	44.33
Qwen-VL-Max (Team, 2024)	End-to-End	-	79.80	#6	53.41	#5	33.07
GPT-4V (OpenAI et al., 2024)	End-to-End	-	78.50	#7	56.49	#4	**28.01**
GPT-4o (OpenAI et al., 2024)	End-to-End	-	**85.70**	**#3**	**62.52**	**#2**	**27.05**

Table 3: Zero-shot transfer results with state-of-the-art generalist multi-modal language methods and chart-oriented specialist models on our proposed Chart-HQA. PoT denotes program-of-thought prompting. SC denotes self-consistency. We color each column as the best, second best, and third best.

ities in counterfactual reasoning over charts. For instance, the high-performing GPT-4o exhibits a notable 27.05% decline rate on Chart-HQA compared to its performance on ChartQA. Third, we further find that enhancing the model's symbolic reasoning ability is crucial for Chart-HQA. For example, with the same model size (∼7B), Qwen2.5-VL significantly outperforms Qwen-VL-Chat and InternVL-2.5 on Chart-HQA.

4.3 Fine-grained Evaluation Results

Most MLLMs exhibit imbalanced performance across different answer types within Chart-HQA. Table 4 presents the fine-grained evaluation results on Chart-HQA for different answer types. First, only GPT-4V demonstrates both high HQA performance across various answer types and balanced performance distribution. For example, GPT-4V achieves the best performance among all evaluated models on integer, decimal, and boolean answer types, and its performance variance (2.09) is the lowest among all evaluated models. Second, generalist models exhibit superior fine-grained performance and reasoning stability compared to chart-specialist models. This phenomenon indicates that pretraining on fundamental general-purpose abilities is beneficial for chart understanding.

4.4 Ablation Study

We further investigate the effectiveness of our HQA data synthesis method. Specifically, we synthesized 100 HQA instances respectively using three different approaches: Human expert-designed (Human), LLM-generated (Machine) and our proposed human-AI interactive method. Subsequently, we calculate the unit cost of data synthesis for each method. We then invite human experts to evaluate the synthesized data from three perspectives, each scored on a scale of 5: *Rationality*, which measures whether the generated questions conform to the intrinsic layout structure of the chart; *Complexity*, which assesses the difficulty of answering the questions; and *Diversity*, which evaluates the richness of counterfactual operations applied to the chart. The evaluation results are presented in Figure 4. First, in terms of rationality and complexity, our method performed comparably to human experts. For example, in complexity scoring, human-designed questions received a score of 4.5, while our method achieved a close score of 4.2. Second, regarding diversity, our method significantly outperform both human-designed and machine-generated approaches. Third, compared to the design cost of human experts, our method reduces costs by 90.7%, lowering the average unit cost to 0.12 CNY per sample. The primary reason for this cost effi-

Model	Type	#Params	Chart-HQA				Variance (↓)
			INT	DEC	BOOL	TEXT	
Specialist Models							
Pix2struct (Lee et al., 2023)	End-to-End	0.3B	17.81	18.48	20.21	14.60	**4.13**
MatCha (Liu et al., 2023b)	End-to-End	0.3B	23.01	21.83	4.26	20.94	59.06
Unichart (Masry et al., 2023)	End-to-End	0.2B	15.21	15.74	40.43	28.10	107.30
TinyChart (Zhang et al., 2024)	End-to-End	3B	35.71	26.61	57.45	39.39	125.60
DocOwl-v2.0 (Hu et al., 2024)	End-to-End	8B	30.36	27.84	54.26	37.47	106.30
ChartLlama (Han et al., 2023a)	End-to-End	13B	13.27	11.88	59.57	28.73	368.37
ChartVLM-L (Xia et al., 2024a)	End-to-End	14.3B	17.99	16.62	51.09	26.26	191.47
DePlot(GPT 3.5) PoT SC (Liu et al., 2023a)	Tool-augment	-	45.08	58.17	56.38	32.96	102.07
Generalist Models							
Qwen-VL-Chat (Bai et al., 2023)	End-to-End	7B	27.53	25.71	44.68	34.44	55.38
DeepSeek-VL-Chat (Lu et al., 2024)	End-to-End	7B	29.17	24.55	42.55	38.02	50.29
Qwen2.5-VL (Team, 2025)	End-to-End	7B	44.05	58.66	**67.02**	54.55	68.35
InternVL-v2.5-8B (Chen et al., 2025)	End-to-End	8B	36.90	48.71	**60.64**	48.21	70.50
Monkey (Li et al., 2023b)	End-to-End	9.8B	30.14	28.32	44.68	39.12	44.41
Qwen2.5-VL (Team, 2025)	End-to-End	72B	**51.55**	**68.02**	**72.34**	**64.54**	60.24
Gemini-Pro (Team et al., 2023)	End-to-End	-	38.71	43.55	55.32	36.46	53.06
Qwen-VL-Max (Team, 2024)	End-to-End	-	**50.34**	55.98	58.51	49.72	**13.86**
GPT-4V (OpenAI et al., 2024)	End-to-End	-	**54.58**	**58.38**	55.32	**55.52**	**2.09**
GPT-4o (OpenAI et al., 2024)	End-to-End	-	49.10	**64.79**	51.06	**61.98**	45.72

Table 4: Fine-grained Evaluation Results on Chart-HQA across different answer types. INT: Integer answers; DEC: Decimal answers; BOOL: Boolean text answers; TEXT: Text answers. We color each column as the best, second best, and third best.

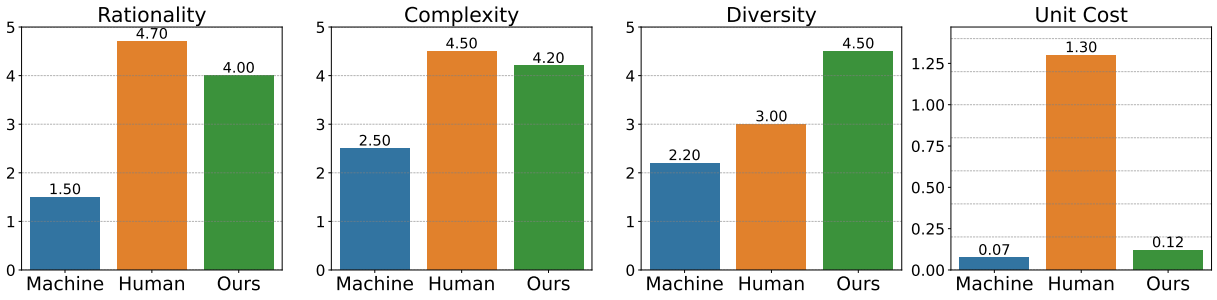


Figure 4: Human evaluation performance of three data synthesis methods, including human, machine, and our human-machine interaction approach. From left to right, the comparison includes question rationality, complexity, diversity, and synthesis cost (unit: CNY).

ciency is that, compared to the direct design HQA instances by human experts, our method effectively distributes the workload, where the LLM generate the questions and human experts review them, thereby significantly reducing the overall cost.

4.5 Case Study

To better illustrate the challenge of the proposed chart-based hypothetical question answering task, we conduct a specific case study on Chart-HQA using the powerful reasoning models GPT-4o, Gemini-2.0 (Team et al., 2023), Qwen-VL-Max, and InternVL2.5-78B (Chen et al., 2025) as shown in Figure 5. All models correctly answer the original factual question shown in Figure 5(a). However, when answering the hypothetical question proposed in Chart-HQA, the reasoning processes of models exhibit significant differences as shown in Figure 5(b). Specifically, GPT-4o first accurately reasons through the hypothetical scenario, correctly inferring that the maximum value of the green bar (78%) becomes the average value of the blue bar (34%). It then deduces that the hypothetical scenario does not affect the answer, thus maintaining the original response (51%). In contrast, Gemini-2.0 makes two critical errors. First, it confuses the colors in the chart, leading to an incorrect interpretation of the hypothetical scenario. Second, it fails to reason that the hypothetical scenario does not influence the question, ultimately replacing the correct answer with the erroneous assumption inference. Similar to Gemini-2.0, Qwen-VL-Max

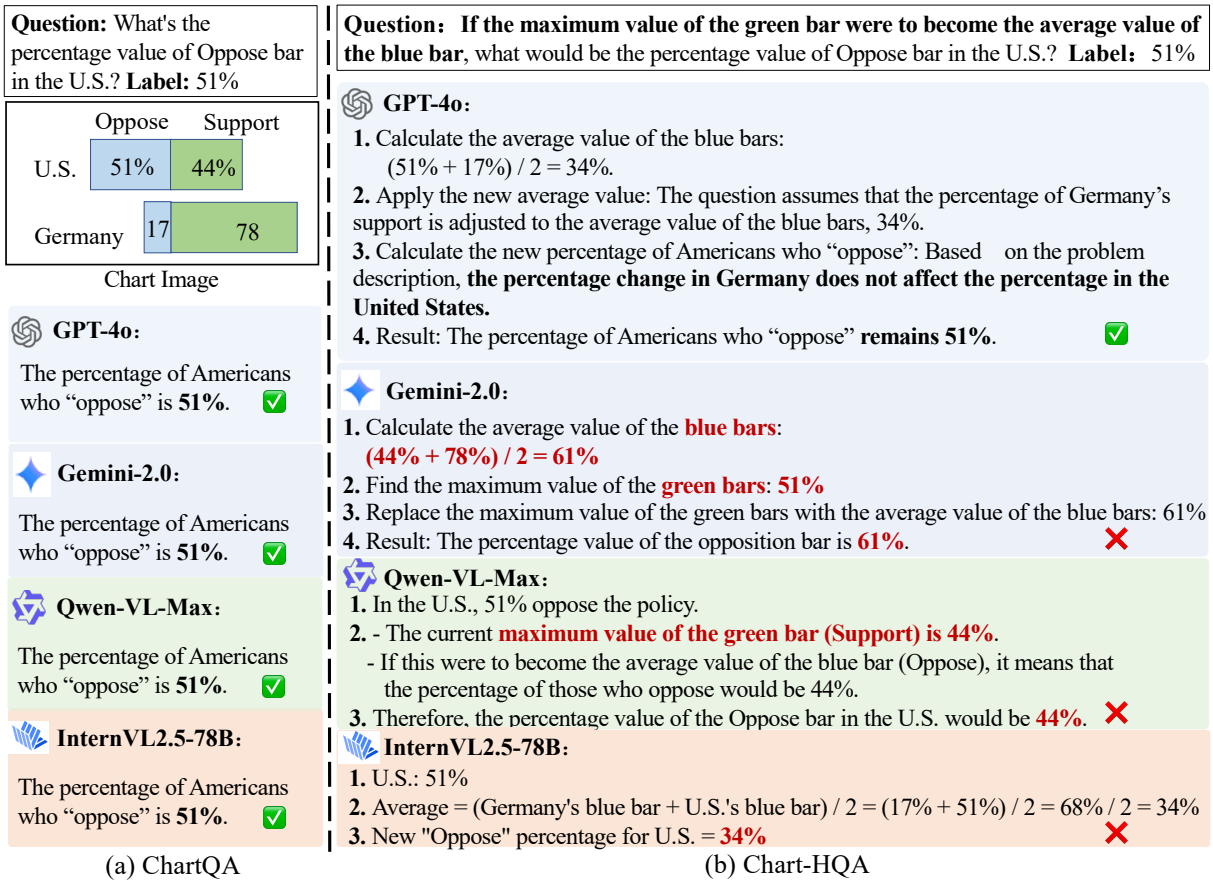


Figure 5: The visualization of examples in ChartQA and Chart-HQA(ours). We use **black bold** to highlight key reasoning steps of the model and red to mark incorrect reasoning steps.

and InternVL2.5-78B also fail to recognize that the hypothetical scenario does not affect the answer during the reasoning process, leading to incorrect responses. This phenomenon indicates that since MLLMs need to analyze the problem step by step, they tend to become deeply engaged in imaginative reasoning based on the assumed scenario, ultimately overlooking the actual content the question aims to query.

5 Related Works

5.1 Chart Benchmarks

HallusionBench (Guan et al., 2023) has revealed that state-of-the-art models, such as GPT-4V (OpenAI, 2023) and LLaVA-1.5 (Liu et al., 2023d), exhibit severe hallucinations when processing intricate chart-related queries. Additionally, several benchmarks including SciCap (Hsu et al., 2021), Chart2Text (Kantharaj et al., 2022), AutoChart (Zhu et al., 2021), and ChartSumm (Rahman et al., 2023), focus on chart-to-text summarization. For chart comprehension, ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020) serve as widely used evaluation datasets. ChartLlama (Han et al., 2023b), ChartX (Xia et al., 2024b), ChartY (Chen et al., 2024) and ChartBench (Xu et al., 2024) significantly increases the number of supported chart types and dataset scale. In contrast, we introduce Chart-HQA to systematically analyze the impact of inherent output biases in MLLMs on chart-based evaluations. By comparing its performance with the widely used ChartQA, we reveal the limitations of current MLLMs in visual chart understanding.

6 Conclusion

In this paper, we propose an novel chart hypothetical question answering task to reveal the inherent output bias problem of MLLMs. Subsequently, we present a human-machine interactive HQA data synthesis framework named HAI to synthesize diverse and high-quality HQA data at a low cost. We synthesized a challenging benchmark called Chart-HQA using publicly available data sources. Through a comprehensive analysis of 18 MLLMs of varying sizes, we reveal the shortcomings of current MLLMs in visual chart understanding.

Limitations

Due to the limited budget and computation resources, we only conducted zero-shot testing on Chart-HQA. In the future, we will increase the data scale to explore more experimental settings.

Ethical Statement

The dataset in this paper is constructed using publicly available sources and adheres to ethical guidelines for data collection and annotation. No personally identifiable or sensitive information is included. Efforts have been made to ensure fairness and minimize bias in data representation.

References

Jinze Bai, Shuai Bai, Shusheng Yang, et al. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Onechart: Purify the chart structural extraction via one auxiliary token. *arXiv preprint arXiv:2404.09987*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023a. Chartllama: A multimodal llm for chart understanding and generation.

Yucheng Han, Chi Zhang, Xin Chen, et al. 2023b. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. 2021. Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

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

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, et al. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912.

Junnan Li, Dongxu Li, Silvio Savarese, et al. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202, pages 19730–19742.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023b. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023b.

MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, et al. 2023d. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e. Visual instruction tuning. In *NeurIPS*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding.

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

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684.

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 (WACV)*.

OpenAI. 2021. Introducing chatgpt. OpenAI Blog.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner,

Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, et al. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2025. Qwen2.5-vl.

Qwen-VL Team. 2024. Qwen-vl-max.

Weihan Wang, Qingsong Lv, Wenmeng Yu, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, et al. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2024a. Chartx chartvlm: A versatile benchmark and foundation model for complicated chart reasoning.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. 2024b. 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. 2024. Chartbench: A benchmark for complex visual reasoning in charts.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*.

Jiawen Zhu, Jinye Ran, Roy Ka-wei Lee, et al. 2021. Autochart: A dataset for chart-to-text generation task. *arXiv preprint arXiv:2108.06897*.

A Prompts Templates for HQA

Within this section, we outline the prompt templates for automatically generating hypothetical questions. The prompt templates are shown in Table 5, 6.

B Illustrative Examples of Human Verification for HQA

Within this section, we show the examples of human verification processes for ensuring the quality of generated hypothetical questions. These examples are illustrated in Figures 6, 7.

Table 5: Prompt template of instruction proposal synthesis.

SYSTEM:
You are a creative prompt creator.
USER:
Given {CHART_DESCRIPTION}.
A series of data points contains a list of the following attributes (dictionary-style):
{FIELD_DESCRIPTION}
According to the chart description provided above, Your goal is to generate new instructions to guide the user in asking hypothetical questions based on information in the chart.
Your can draw inspiration from the #Given Instructions# to create a brand new instruction.
The new instruction must meet the following conditions:
1. It should only contains two parts: how to specify the elements and the assumed change to be applied on the elements.
2. The new instruction must be reasonable and must be understood and responded by humans.
3. Follow the sentence patterns in the examples.
4. Please replace specific concepts with general concepts.
5. Use attributes in charts to refer to specific elements.
#Given Instructions#:
1. {I1}
2. {I2}
3. {I3}
4. {I4}
Now please directly generate 3 new instructions without writing any other explanations:
<Output>:

Table 6: Prompt template for hypothetical question generation.

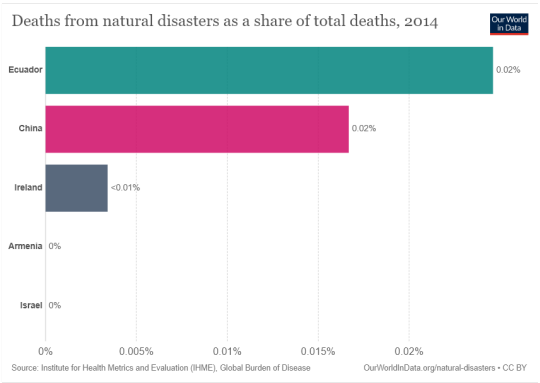
SYSTEM:
You are a Question Rewriter.
USER:
You are provided with metadata from {CHART_DESCRIPTION}. The chart's title and series of data points (models) are given in the metadata, with each model comprising attributes outlined in {FIELD_DESCRIPTION}.
Your role is to creatively rewrite original questions into Hypothetical Questions (HQ) based on the chart's information. Each original question should be rephrased into two different hypothetical questions.
Ensure:
1. Adhere to the ideas in #Feasible Rewrite Proposals#.
2. HQ should also adhere to the format in #Demonstration# and use specific details from the chart. It also needs to be as clear as possible.
3. Keep the original question as part of rewritten HQ.
4. The answer to the HQ should differ from the original answer.
5. Provide the name of the color in words, not any code like #FF0000.
6. When the answer is a percentage value, it needs to be answered as a percentage.
7. If the calculation process includes percentage values, you need to pay attention to the percent operation.

#Feasible Rewrite Proposals#
1. {I1}
2. {I2}
3. {I3}

#Demonstration#:
Original question: {Q_DEMON}
Hypothetical question examples:
1. {HQ_DEMON_1}
2. {HQ_DEMON_2}

#Chart Metadata#:
{CHART_METADATA}

Please directly complete HQs and produce the following text information. Note that the answers should not include any explanation or units.:
First Original Question:
Question: {Q1}
Answer: {A1}
HQ Rewrites:
Question_1:
Answer_1:
Question_2:
Answer_2:
Second Original Question:
Question: {Q2}
Answer: {A2}
HQ Rewrites:
Question_1:
Answer_1:
Question_2:
Answer_2:
<Output>:



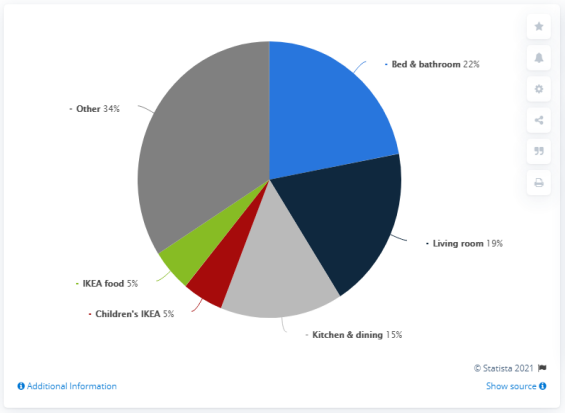
Hypothetical Question 1:
If the category of China with its pink bar **swapped places** with Israel, which would be the color of the bar representing the new position of China?

Validation result: Pass

Hypothetical Question 2:
If the category of China with its pink bar **swapped colors** with Israel, which would be the color of the bar representing China?

Validation result: Failed
Reason: Not answerable. Because the color of Israel bar does not show in the figure.

Figure 6: The first example of human verification.



Hypothetical Question 1:
Assuming the value of "Bed & Bathroom" doubles, how many segments have the same share?

Validation result: Pass

Hypothetical Question 2:
Assuming the value of "Bed & Bathroom" doubles, how does this affect the percentage of the "Kitchen & dining" share?

Validation result: Failed
Reason: The question is not clear. It is hard to find a unique and accurate answer.

Figure 7: The second example of human verification.