ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering

Yifan Wu[1*], **Lutao Yan**[3*], **Leixian Shen**[2], **Yunhai Wang**[4], **Nan Tang**[1,2], **Yuyu Luo**[1,2†]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]South China University of Technology, [4]Renmin University of China
ywu012@connect.hkust-gz.edu.cn, yuyuluo@hkust-gz.edu.cn
https://chartinsight.github.io/

Abstract

Chart question answering (ChartQA) tasks play a critical role in interpreting and extracting insights from visualization charts. While recent advancements in multimodal large language models (MLLMs) like GPT-4o have shown promise in high-level ChartQA tasks, such as chart captioning, their effectiveness in low-level ChartQA tasks (*e.g.,* identifying correlations) remains underexplored. In this paper, we address this gap by evaluating MLLMs on low-level ChartQA using a newly curated dataset, *ChartInsights*, which consists of 22,347 (chart, task, query, answer) covering 10 data analysis tasks across 7 chart types. We systematically evaluate 19 advanced MLLMs, including 12 open-source and 7 closed-source models. The average accuracy rate across these models is 39.8%, with GPT-4o achieving the highest accuracy at 69.17%. To further explore the limitations of MLLMs in low-level ChartQA, we conduct experiments that alter visual elements of charts (*e.g.,* changing color schemes, adding image noise) to assess their impact on the task effectiveness. Furthermore, we propose a new textual prompt strategy, *Chain-of-Charts*, tailored for low-level ChartQA tasks, which boosts performance by 14.41%, achieving an accuracy of 83.58%. Finally, incorporating a visual prompt strategy that directs attention to relevant visual elements further improves accuracy to 84.32%.

1 Introduction

Visualization charts can effectively convey data insights, but the abundance of information they provide makes it challenging for users to efficiently and accurately extract desired information (Li et al., 2024b; Luo et al., 2023; Xie et al., 2024). Automated chart question answering (ChartQA) (Ye et al., 2024; Zeng and Battle, 2023) is crucial to help users pinpoint relevant information based on their intents (Amar et al., 2005; Saket et al., 2019).

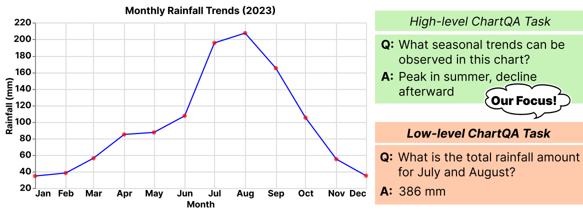


Figure 1: Examples of Two Types of ChartQA Tasks

High-Level and Low-Level ChartQA Tasks. ChartQA tasks can generally be categorized into two types: **high-level** tasks and **low-level** tasks. High-level tasks focus on questions that require understanding the overall context or summary of the chart, involving broader, goal-oriented inquiries that seek to understand overarching trends or patterns. In contrast, low-level tasks focus on specific, detail-oriented inquiries that seek precise data points or comparisons within the chart, involving straightforward, factual information retrieval (Amar et al., 2005; Saket et al., 2019). For example, as shown in Figure 1, given the same line chart showing rainfall in different months of the same year, a high-level task might ask about the cyclical or seasonal trends in the chart, while a low-level task would be more focused on the data itself in the chart, such as asking how much total rainfall there was in July and August.

Traditionally, ChartQA has been a challenging problem due to the limited capabilities in natural language understanding and the high complexity of chart reasoning (Li et al., 2024a; Liu et al., 2024c; Tang et al., 2024). Fortunately, recent advancements in multimodal large language models (MLLMs) have made it possible for users to interact with systems using natural language to extract specific information from data across various modalities. This progress has illuminated new possibilities for ChartQA on different levels of tasks.

*Equal contribution
†Yuyu Luo is the corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12174–12200
November 12-16, 2024

Table 1: Comparison with Existing Datasets

Task Levels	Datasets	#-Task Types	#-Chart Types	#-Charts	#-Queries	#-Queries/#-Charts	Metadata?
High-level ChartQA	DVQA (Kafle et al., 2018)	3	1	300K	3.5M	11.7	✓
	ChartQA (Masry et al., 2022)	4	3	4.8K	9.6K	2	✓
Mix	FigureQA (Kahou et al., 2018)	6	5	120K	1.5M	12.9	✗
	ChartLlama (Han et al., 2023)	7	10	11K	160K	14.5	✗
	ChartBench (Xu et al., 2023)	4	11	2.1K	16.8K	8	✗
Low-level ChartQA	**ChartInsights (ours)**	**10**	7	2K	22K	11.2	✓

Prior Art: High-Level ChartQA with MLLMs. Recent studies have explored the capabilities of MLLMs in performing high-level ChartQA (Cheng et al., 2023; Han et al., 2023; Huang et al., 2023; Masry et al., 2023; Xia et al., 2024; Xu et al., 2023; Zhou et al., 2023). The findings reveal that state-of-the-art MLLMs like GPT-4o have demonstrated promising results in addressing high-level tasks, and have outlined future research directions.

Our Focus: Low-Level ChartQA with MLLMs. Existing evaluations of ChartQA primarily focus on high-level tasks, such as chart captioning, while overlooking low-level ChartQA tasks (*e.g.,* characterizing distributions) that humans frequently encounter in daily life. Specifically, studies from the visualization and visual analysis community have well-defined 10 widely used low-level ChartQA tasks (Amar et al., 2005; Saket et al., 2019). Thus, our study seeks to systematically evaluate the effectiveness of MLLMs in addressing these 10 low-level ChartQA tasks.

Contributions. The key contributions are:

(1) **ChartInsights Dataset.** We curate ChartInsights, *the first* dataset for evaluating low-level data analysis tasks on charts. ChartInsights includes diverse chart variants, textual and visual prompts, and comprehensive metadata, enabling the investigation of MLLMs' performance across various low-level ChartQA scenarios.
(2) **Comprehensive Evaluations.** Our study establishes benchmarks by evaluating 19 MLLMs on 10 low-level ChartQA tasks, providing valuable insights into the current capabilities of MLLMs in processing and analyzing chart information.
(3) **New Experimental Findings.** We summarize 12 experimental findings, highlighting the importance of visual prompts, chart elements, and image quality in performing low-level ChartQA tasks.
(4) **Chain-of-Charts.** We introduce the *Chain-of-Charts* strategy, a new textual prompt designed to enhance MLLMs' reasoning capabilities in

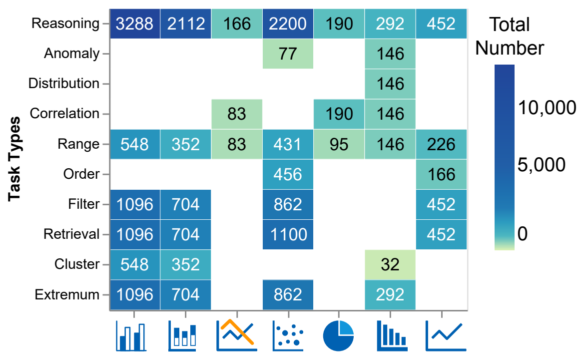


Figure 2: Low-level Tasks vs. Chart Types

ChartQA tasks by leveraging a series of interconnected question-answer pairs to guide the model.

2 ChartInsights Dataset

Since no dataset exists for low-level ChartQA tasks, we construct a large-scale dataset, ChartInsights, to systematically evaluate the performance of MLLMs in these tasks. The construction details of ChartInsights are provided in Appendix A.2.

Overview of ChartInsights. ChartInsights contains 22,347 (chart, task, query, answer) samples across 7 chart types for 10 low-level data analysis tasks on charts (Amar et al., 2005). Figure 2 shows the distribution of 10 low-level tasks and 7 chart types. *Please refer to Appendix B for more details.*

Comparison with Existing Datasets. As shown in Table 1, ChartInsights differs from existing ChartQA datasets by emphasizing low-level data analysis tasks. It covers 10 distinct tasks across 7 chart types, including 2000 charts with an average of 11.2 queries each. Additionally, all relevant metadata are available, making ChartInsights a valuable resource for future research in ChartQA. *Refer to Appendix A.1 for further discussions.*

3 Experiments

3.1 Experimental Design

As shown in Figure 3, we utilize our ChartInsights to systematically evaluate the effectiveness

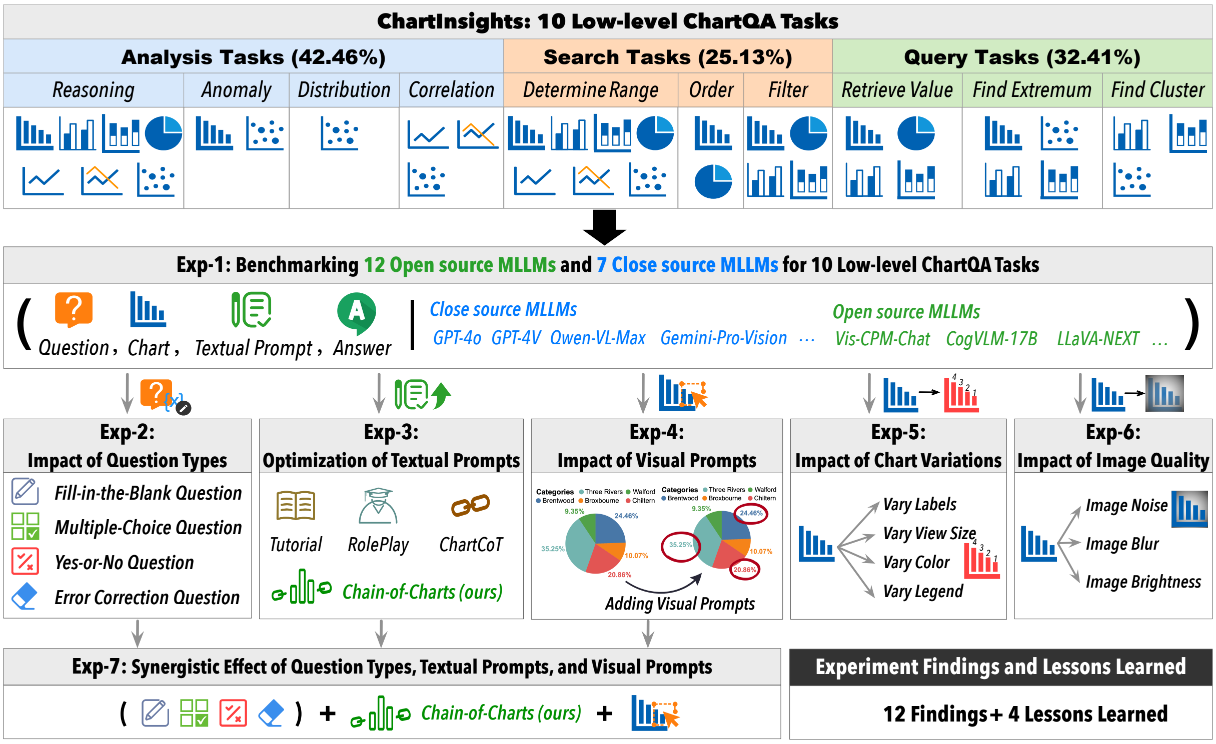


Figure 3: An Overview of Experimental Settings

of MLLMs in 10 low-level ChartQA from different angles. The experiments are designed as follows:

Exp-1 *Benchmarking MLLMs*: We start by benchmarking the performance of widely used MLLMs across 10 low-level ChartQA tasks involving 7 different types of charts. This experiment establishes a baseline for understanding the capabilities and limitations of MLLMs in low-level ChartQA tasks.

Exp-2 *Impact of Question Types*: We analyze how different question types influence MLLMs interactions, helping to identify which types elicit the most accurate and informative responses.

Exp-3 *Textual Prompt Strategies*: We investigate the effect of various textual prompt strategies, such as Chain-of-Thoughts, on MLLMs performance.

Rethink. Prior evaluations on high-level ChartQA tasks (Cheng et al., 2023; Han et al., 2023; Huang et al., 2023; Xu et al., 2023; Zhou et al., 2023) primarily focused on optimizing *textual* prompts. Given that ChartQA involves both "reading" and "understanding" chart images, we conduct an in-depth exploration of the impact of *visual modification* on charts and prompts.

Exp-4 *Impact of Visual Prompts*: We conduct an in-depth exploration of the impact of visual prompts on MLLMs performance to understand how guiding the MLLMs' attention to specific visual elements can enhance its analytical capabilities.

Exp-5 *Impact of Chart Variations*: We vary chart elements to analyze how changes in color schemes, view sizes, and legends affect the performance.

Exp-6 *Impact of Image Quality*: We evaluate the effect of image quality by introducing various levels of image perturbations, such as noise and resolution changes, to understand the robustness of MLLMs in handling low-quality charts.

Exp-7 *Synergistic Effects of Different Strategies*: Finally, we explore the synergistic effects of combining different question types, textual prompts, and visual prompts to enhance the overall performance of MLLMs in low-level ChartQA tasks.

Next, we discuss the main results and findings in Section 3.2 and lessons learned in Appendix D.

3.2 Experimental Results and Findings

◆ Exp-1: Evaluation and Benchmarking

Experimental Settings. We evaluate 19 widely used models (see Table 2) from both academia and industry, including 12 open-source and 7 closed-source MLLMs. Inspired by existing evaluation strategies (Fu et al., 2024; Lu et al., 2024), we randomly selected 20% of the ChartInsights dataset as our test set to reduce testing costs while ensuring the reliability of experimental results.

The test set comprises 400 charts, spanning 10

Table 2: Performance of 19 MLLMs across 10 Low-level ChartQA Tasks.

Models	Analysis				Search			Query			Overall (%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
Open Source MLLMs											
VisCPM-Chat-v1.1 (?)	28.4	**46.1**	33.3	51.9	23.0	6.4	25.1	15.8	32.0	29.6	26.2
BLIP2 (Li et al., 2023b)	24.8	23.4	25.0	15.1	25.3	20.2	39.8	27.8	30.3	30.1	28.3
CogVLM-17B (Wang et al., 2024)	20.3	23.1	43.6	29.6	**37.7**	10.8	9.1	37.9	56.6	26.7	29.4
OmniLMM-12B (Modelbest, 2024)	24.7	19.9	27.0	34.9	35.7	28.3	30.0	33.0	39.9	33.1	31.1
LLaVA1.5 (Liu et al., 2024b)	32.4	6.3	30.9	23.1	21.7	32.7	35.6	32.6	35.8	43.5	32.2
ChartAssistant (Meng et al., 2024)	24.6	27.7	35.8	28.1	30.5	22.5	14.7	39.4	63.0	26.4	32.4
MiniCPM-v2 (Hu et al., 2024)	19.5	55.1	33.3	56.5	24.9	16.7	36.3	37.9	52.4	32.0	33.0
mPLUG-Owl2 (Ye et al., 2023)	31.0	27.0	29.4	35.3	28.4	22.5	40.3	30.9	41.1	27.3	33.3
Qwen-VL-Chat (Bai et al., 2023)	27.8	36.3	**45.1**	**55.8**	33.8	20.0	28.7	31.3	50.2	27.1	33.4
ViP-LLaVA (Cai et al., 2023)	28.8	6.6	34.8	30.3	21.9	**35.8**	**40.4**	42.2	38.3	33.8	33.8
LLaVA-NEXT (Liu et al., 2024a)	**30.6**	7.4	26.5	38.0	29.5	33.3	23.4	**53.5**	59.8	**52.3**	38.5
Sphinx-v2 (Lin et al., 2023)	30.0	28.9	37.8	36.1	25.8	23.5	36.7	49.7	**66.3**	45.3	**40.2**
Closed Source MLLMs											
Qwen-VL-Plus (Bai et al., 2023)	30.8	27.3	47.1	47.1	43.0	34.6	20.7	58.7	65.5	62.5	42.6
Gemini-Pro-Vision (Team et al., 2024)	25.6	30.1	45.6	58.7	75.3	32.9	30.1	60.4	80.9	55.3	48.4
ChatGLM-4V (Zeng et al., 2023)	34.1	28.9	39.2	42.3	55.5	18.9	43.4	58.1	69.3	71.4	48.4
Claude3-Haiku (Anthropic, 2024)	33.0	9.0	42.7	46.2	60.4	26.2	40.0	62.3	75.1	66.8	49.5
Qwen-VL-Max (Bai et al., 2023)	28.8	25.8	62.3	63.0	66.1	40.2	38.9	67.0	79.6	66.8	51.7
GPT-4V (OpenAI, 2023)	35.2	19.5	53.4	59.6	70.0	41.9	44.3	67.6	88.7	72.9	56.1
GPT-4o (OpenAI et al., 2024)	**55.9**	**34.0**	**70.1**	**68.5**	**80.6**	**68.9**	**49.9**	**82.6**	**93.9**	**74.3**	**69.2**

Table 3: Performance of 19 MLLMs across 7 Chart Types and 4 Question Types. (FB: Fill-in-the-Blank Question; MC: Multiple-Choice Question; YN: Yes-or-No Question; EC: Error Correction Question)

Models	Chart Types							Question Types			
	Grouped Bar	Stacked Bar	Grouped Line	Basic Bar	Basic Line	Scatter Plot	Pie	FB	MC	YN	EC
Open Source MLLMs											
VisCPM-Chat-v1.1 (?)	24.8	21.5	24.4	30.3	25.3	**34.9**	20.4	7.4	39.8	49.3	8.3
BLIP2 (Li et al., 2023b)	32.2	31.1	24.7	34.2	15.0	18.2	8.7	3.1	46.9	57.7	5.5
CogVLM-17B (Wang et al., 2024)	26.2	24.2	**32.1**	40.1	35.3	30.2	19.6	27.6	49.7	31.3	9.2
OmniLMM-12B (Modelbest, 2024)	28.8	25.9	26.5	37.8	43.4	27.5	34.4	10.1	46.3	57.0	10.8
LLaVA1.5 (Liu et al., 2024b)	31.3	30.5	22.6	35.7	40.6	26.8	36.3	9.6	41.4	67.9	9.9
ChartAssistant (Meng et al., 2024)	32.9	27.8	24.1	41.2	35.6	28.2	23.3	**30.2**	46.3	40.9	12.2
MiniCPM-v2 (Hu et al., 2024)	33.4	28.7	18.8	40.4	36.3	29.9	28.0	21.0	42.0	60.8	8.3
mPLUG-Owl2 (Ye et al., 2023)	32.4	31.8	29.2	37.0	38.4	29.2	34.3	13.2	49.9	54.6	15.5
Qwen-VL-Chat (Bai et al., 2023)	31.2	26.3	25.9	**39.3**	42.8	38.5	34.7	22.5	52.3	48.6	10.4
ViP-LLaVA (Cai et al., 2023)	32.2	31.2	16.1	39.8	34.4	28.1	39.1	7.9	40.1	**73.4**	13.7
LLaVA-NEXT (Liu et al., 2024a)	37.1	33.7	18.8	50.4	**48.4**	27.4	37.1	23.1	42.2	63.2	**25.6**
Sphinx-v2 (Lin et al., 2023)	**39.4**	**35.7**	26.5	**51.2**	42.5	31.5	36.7	28.2	**56.1**	60.0	16.7
Closed Source MLLMs											
Qwen-VL-Plus (Bai et al., 2023)	36.5	35.8	26.2	57.0	32.8	44.2	42.5	32.9	59.2	54.0	24.2
Gemini-Pro-Vision (Team et al., 2024)	45.0	43.3	35.1	57.7	43.8	47.3	51.2	40.9	55.4	43.9	53.3
ChatGLM-4V (Zeng et al., 2023)	46.0	43.9	36.9	58.7	56.3	39.7	49.9	32.1	53.7	72.2	35.8
Claude3-Haiku (Anthropic, 2024)	50.7	48.1	32.1	55.6	46.9	40.4	48.3	39.4	51.2	61.8	45.9
Qwen-VL-Max (Bai et al., 2023)	47.1	46.3	38.4	66.4	43.1	47.7	48.1	44.2	75.3	48.1	39.0
GPT-4V (OpenAI, 2023)	52.0	48.2	47.3	67.2	49.4	62.7	53.6	44.1	64.4	66.4	49.7
GPT-4o (OpenAI et al., 2024)	**62.7**	**59.8**	**53.9**	**84.1**	**56.9**	**74.6**	**70.9**	**61.0**	**77.8**	**74.8**	**63.1**

low-level analysis tasks and encompassing 4,388 (chart, task, query, and answer) samples. This sizable validation dataset allows for a comprehensive evaluation of MLLMs' capabilities in low-level ChartQA tasks. Then, we analyze the answers of MLLMs, compare them with Ground Truth, and calculate the accuracy. For each model, we evaluate through four question types in textual prompts, *i.e.,* Fill-in-the-Blank, Multiple-Choice, Yes-or-No, and Error Correction questions.

Overall Results. Table 2 presents the performance of the 19 models across 10 low-level tasks, while Table 3 showcases their performance across 7 chart types and 4 question types. The findings indicate that closed-source models outperform open-source models by a significant margin among the 19 models evaluated. Notably, GPT-4o demonstrates exceptional performance across all tasks.

Finding-1: *Closed-Source models exhibit far superior generalization performance in low-level analysis tasks compared to open-source models.*

In addition, our analysis of open-source models reveals that the ability to comprehend low-level ChartQA tasks is not directly proportional to the number of model parameters. For example, MiniCPM (with 2B parameters) outperforms larger models like CogVLM (with 17B parameters). This suggests that factors beyond scale significantly influence a model's capacity.

Finding-2: *The ability of open-source models to understand low-level charts is not directly proportional to their number of model parameters.*

Moreover, Table 3 shows that although some open-source models perform poorly overall, they have high accuracy on Yes-or-No tasks, with ViP-LLaVA's accuracy even being close to that of GPT-4o. We analyze the recall rates of ViP-LLaVA and GPT-4o on Yes-or-No tasks based on the confusion matrix and found that ViP-LLaVA's recall rate for the "No" label is only 25%, which is quite different from the final accuracy of 73.4%; while GPT-4o's recall rate for the "No" label is 72%, which is relatively close to the final accuracy of 74.8%. We believe that open-source models like ViP-LLaVA have a certain tendency towards the "No" label, and it is because there are more "No" labels in our data that these open-source models have a high accuracy on Yes-or-No questions.

Finding-3: *We find open-source models like ViP-LLaVA show higher accuracy on Yes-or-No questions because of a possible bias towards "No" labels.*

GPT-4o as a representative research object. Since GPT-4o significantly outperforms other MLLMs, we utilize it as a representative MLLM to systematically evaluate the effectiveness of low-level ChartQA tasks under different scenarios. Specifically, we will conduct a series of in-depth evaluations (*i.e.,* **Exp-3** to **Exp-7**) using GPT-4o.

Overall Results of GTP-4o. The overall result of GPT-4o is shown in Figure 5. These heatmaps visualize the performance of GPT-4o on various low-level ChartQA tasks under different prompt conditions. The progression from subfigures (a) to (d) clearly indicates the incremental benefits of incorporating visual prompts, optimization strategies, and their combination, culminating in the most effective approach for improving GPT-4o's performance in low-level ChartQA tasks.

As depicted in the bar chart at the top of Figure 4, GPT-4o's accuracy reaches near 90% for basic bar charts, but hovers around 66% for similar tasks involving other chart types. This performance

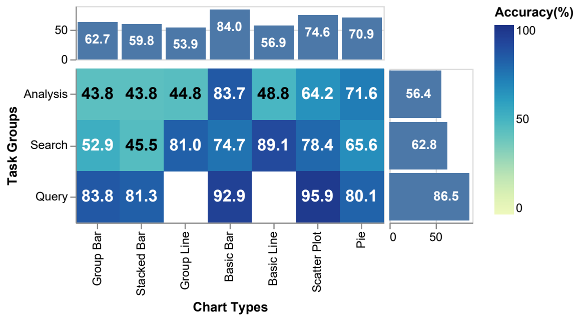


Figure 4: Effectiveness of GPT-4o

gap suggests that, despite the dataset containing a significant number of reasoning tasks that are generally straightforward for humans, the effectiveness of GPT-4o in ChartQA tasks has not yet reached that of the average human.

Finding-4: *The performance of GPT-4o declines as task complexity increases, mirroring human performance, but it has not yet matched the analytical capabilities of average humans.*

◆ Exp-2: Impact of Question Types

The experimental settings are the same as in Exp-1.

The Effectiveness of Question Types. Figure 5 (a) provides an overview of GPT-4o's performance across 10 low-level tasks using four different textual prompts. Notably, GPT-4o achieves the highest overall accuracy of 77.8% in "Multiple-Choice" questions. It also demonstrates a strong performance in "Yes or No" questions, with an accuracy of 74.8%. Comparatively, GPT-4o performs relatively better in "Multiple-Choice" and "Yes-or-No" questions compared to "Fill-in-the-Blank" and "Error Correction" questions. The former prompt types inherently provide GPT-4o with answer options to select or evaluate, while the latter two require direct answer generation from GPT-4o.

Finding-5: *Structured textual prompts and candidate answers significantly enhance GPT-4o's ability to reason out correct responses.*

◆ Exp-3: Optimization of Textual Prompts

This experiment will examine the influence of commonly used textual prompt optimization strategies on MLLMs, including RolePlay (Shanahan et al., 2023), Tutorial, and ChartCoT (Xu et al., 2023), which is based on the Chain-of-Thought (CoT) prompt strategy (Wei et al., 2022). The Chain-of-Thought (CoT) prompt strategy has proven ef-

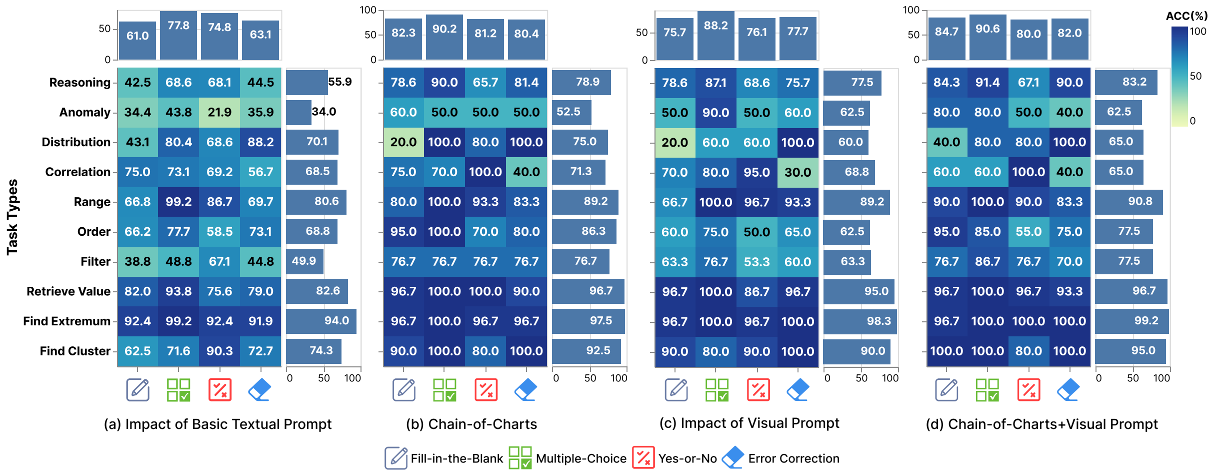


Figure 5: The Effectiveness of GPT-4o across 10 Low-level Tasks and 4 Question Types

fective in various scenarios (Wei et al., 2022). CoT aims to guide the model by mimicking the step-by-step reasoning process humans use to solve problems. Recently, Xu et al. (Xu et al., 2023) implemented the CoT strategy for ChartQA tasks, namely ChartCoT (Xu et al., 2023). ChartCoT poses a series of questions to guide the model in understanding the chart's details before formulating an answer. However, ChartCoT faces challenges in ensuring the accuracy of GPT-4o's responses to guiding questions, especially with complex charts.

Chain-of-Charts Prompts. Therefore, we introduce a novel prompting strategy, termed *Chain-of-Charts*, which builds based on the Chain-of-Thoughts approach (Wei et al., 2022), as shown in Appendix F.6. The core of *Chain-of-Charts* lies in orchestrating a sequence of questions and their corresponding answers $((q_1, a_1), (q_2, a_2), \ldots (q_m, a_m))$ to progressively guide the model towards a deeper understanding of the chart's details, thereby enhancing its ability to formulate accurate responses.

Experimental Settings. Shanahan et al. (Shanahan et al., 2023) suggested that giving large models specific roles could enhance their performance on particular tasks. Inspired by this, we assigned GPT-4o the role of a visualization expert to see if it would enhance its performance. Our observations showed that more detailed prompts resulted in more precise and accurate responses from GPT-4o. As a result, we created a detailed ChartQA tutorial called the Tutorial prompt. Examples of these two prompts can be found in Appendix F.3 and Appendix F.4 respectively.

Overall Results. Figure 5(b) reports the performance of GPT-4o with Chain-of-Charts. Compared with Figure 5(a), we can observe a significant enhancement in GPT-4o's capabilities across 10 tasks and four question types. Table 4(a) shows the overall accuracy of GPT-4o on 10 low-level tasks under 5 Textual Prompt strategies. Overall, Chain-of-Charts leads in average accuracy across all tasks with 83.5%, outperforming ChartCoT's accuracy of 76.1% by 6.9%. Specifically, Chain-of-Charts achieves the highest accuracy on five tasks including Reasoning, Determine Range, Order, Filter, Retrieve Value, Find Extremumm and Find Cluster, with accuracy of 78.9%, 89.2%, 86.3%, 76.7%, 96.7%, 97.5% and 92.5%. These tasks demand precise reasoning from GPT-4o, based on accurate identification of element coordinates and values. The Chain-of-Charts prompt framework effectively provides GPT-4o with the correct value and coordinate references, significantly aiding in the accurate positioning of different elements.

Finding-6: *Chain-of-Charts supplies GPT-4o with accurate chart reference information, enhancing the model's comprehension and detailed reasoning of chart structures and elements.*

◆ Exp-4: Impact of Visual Prompts

Experimental Settings. In this experiment, we design three types of visual prompts based on graphical overlay strategies (Kong and Agrawala, 2012), namely, handwriting, regular shape, and special design, as shown in Figure 9. The design of visual prompts are detailed in Appendix F.2. Specifically, we generate 255 visual prompts for 35 charts as the testing samples. These 255 visual prompts are associated with 1,020 test samples and cover 10 low-level tasks. Please refer to **Step-5** in the Appendix A.2 for details.

Overall Results. Figure 5(c) demonstrates the strong performance of GPT-4o with visual prompts across 10 tasks and 4 question types. Additionally, Figure 22(a)-(c) showcases the performance of visual prompts under different textual prompts, low-level ChartQA tasks, and chart types. In general, visual prompts enhance GPT-4o's performance. Particularly, Figure 22(b) reveals significant improvements in *Reasoning* and *Anomaly Detection* tasks, indicating the model's ability to accurately analyze and reason with relevant data.

Finding-7: *Visual prompts greatly improve GPT-4o's performance, showing the value of visual information for comprehension and reasoning.*

However, GPT-4o does not exhibit significant benefits from visual prompts in *Correlation* and *Order* tasks. These tasks often challenge GPT-4o to discern complex relationships among more than three distinct elements. In such scenarios, visual prompts may lose their specificity and lead to confusion due to the introduction of multiple new visual elements, especially in tasks like Order, where the added visual information can be misleading.

Finding-8: *Different tasks require tailored visual prompts for effective chart comprehension. Using the same style of visual prompts across tasks can have a negative impact on certain tasks.*

◆ Exp-5: Impact of Chart Variations

Experimental Settings. A chart can exhibit visual differences by varying its constituent elements (e.g., data labels), as illustrated in Appendix A.2 and Figure 8(a). Intuitively, these visual variations may influence GPT-4o's performance on low-level ChartQA tasks. To investigate this hypothesis, we explore how varying chart elements affect the performance of GPT-4o. To this end, we develop 356 visual variants of 35 base charts as test samples. These variants are associated with 17,972 textual prompts, spanning 10 low-level tasks.

Overall Results. Overall, Figure 6(a) indicates that most chart variants have a minor negative impact on GPT-4o's performance. However, the absence of data labels significantly impairs its performance across seven chart types (Figure 6(a)). This is understandable, as data labels aid GPT-4o in comprehending the underlying insights conveyed by the charts. Interestingly, GPT-4o shows a performance boost of 17.5% in anomaly detection and 5.5% in filtering tasks when data labels are not present (Figure 6(b)). This implies that data labels may sometimes hinder GPT-4o's ability to identify anomalies and filter values effectively.

Figure 6(b) reveals that certain chart variants, such as larger x/y/data labels, positively impact GPT-4o's performance in tasks like anomaly detection, filtering, and clustering. These tasks inherently involve comparisons between elements. We hypothesize that alterations in chart elements can shift GPT-4o's focus towards visual comparisons rather than numerical ones.

The bar chart on the right in Figure 6 illustrates the varied impacts of 15 chart variants on GPT-4o's performance. We posit that data labels play a crucial role in GPT-4o's low-level data analysis capabilities, as removing or reducing them tends to diminish its effectiveness. Furthermore, adding marks to the legend or eliminating the legend's color significantly affects GPT-4o's performance on specific tasks by introducing visual clutter or removing essential visual cues, respectively. For instance, changes in color and legend can greatly assist GPT-4o in solving Correlation tasks.

Finding-9: *While most chart variants have a minimal impact on GPT-4o's performance, the absence of data labels significantly affects its accuracy. Additionally, larger labels and the removal of data labels can actually enhance GPT-4o's performance in anomaly detection and filtering tasks, as it shifts its focus to visual comparisons.*

◆ Exp-6: Impact of Image Quality

Experimental Settings. In this experiment, we evaluate the robustness of GPT-4o in low-level ChartQA tasks by introducing six types of noise, as shown in Figure 8 (b). We use 245 visual variants of 35 charts as testing samples, along with 8,456 textual prompts covering 10 tasks.

Overall Results. Figure 21 reports the experimental results. Generally, six methods of degrading image quality tend to negatively impact GPT-4o across a broad range of tasks and chart types. Among these, Median Blur stands out as the most detrimental, causing an average performance decline of 16.8%. We consider that median blurring makes numerical labels unreadable, resulting in a

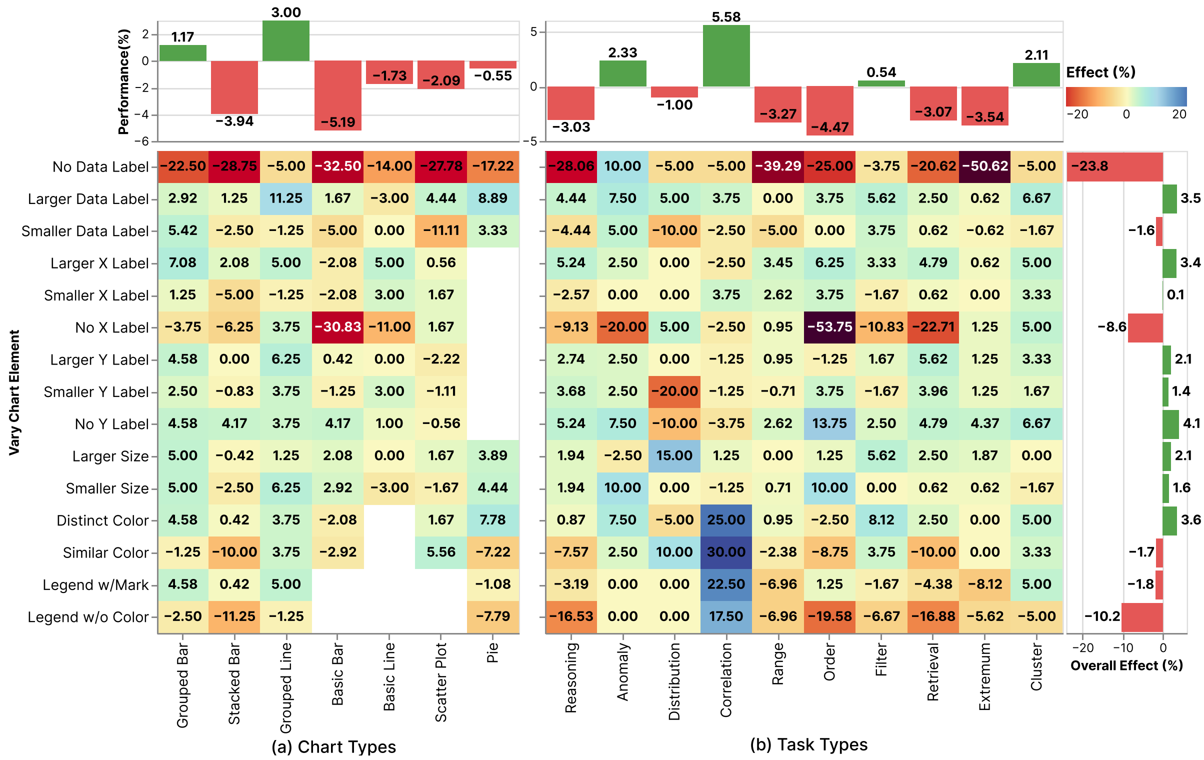


Figure 6: The Impact of Varying Chart Elements on GPT-4o's Performance.

significant decrease in the performance of tasks directly related to numerical values. Interestingly, both increasing and decreasing the brightness show positive effects on the majority of tasks, with an average improvement of 0.6% and 1.3% respectively. In addition, Figure 21(b) reveals a unique finding where the distribution and cluster tasks did not experience any negative impact under six types of noise. In fact, the performance of the distribution task improved by an average of 9.17%, while the cluster task improved by an average of 5.28%.

Finding-10: *GPT-4o's accuracy varies under the influence of different types of noise on various charts. Interestingly, there are instances where the accuracy improves, particularly in visually semantic tasks. This suggests that GPT-4o can rely more on visual information when the textual information is compromised.*

◆ Exp-7: Synergistic Effect of Question Types, Textual Prompts, and Visual Prompts

Experimental Settings. As previously mentioned, using Visual Prompts alone cannot robustly improve the accuracy of GPT-4o. By comparing the bar charts on the right side of Figure 5(a) and (c), we can see that the accuracy of GPT-4o using Visual Prompts actually decreases in Distribution and Order tasks. In contrast, by comparing the bar charts on the right side of Figure 5(a) and (c), we find that under the influence of Chain-of-Charts, GPT-4o has made significant improvements in all 10 tasks. Therefore, we ask: *Can the combination of visual and text prompts enhance performance in low-level ChartQA tasks with GPT-4o?* We use the same samples as in Exp-4 for this experiment.

Overall Results. Figure 5(d) and shows GPT-4o's accuracy following the integration of Chain-of-Charts and Visual Prompt, demonstrating a clear enhancement over the outcomes depicted in Figures 5(a), (b), and (c), which demonstrates the combined strategy's effectiveness.

Table 4 reports the performance improvements in GPT-4o after integrating various textual prompts with visual prompts. Furthermore, this combination attained the highest accuracy in six tasks.

Finding-11: *Combining visual prompts with the Chain-of-Charts strategy significantly improves the performance, suggesting that integrating multiple types of prompts can leverage their respective strengths.*

Discussions about Textual and Visual Prompts. As depicted in Figure 22, the integration of Chain-of-Charts and visual prompts enables GPT-4o to outperform other settings. However, the improve-

Table 4: GPT-4o: The Synergistic Effect of Visual and Textual Prompts (Overall Accuracy (%))

Prompts	Analysis				Search			Query			Overall(%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
(a) The Effectiveness of Textual Prompts											
Basic Textual Prompts	55.9	34.0	70.1	68.5	80.6	68.9	49.9	82.6	93.9	74.3	69.2
ChartCoT	75.7	72.5	65.0	63.7	80.0	62.5	60.0	89.2	92.5	87.5	76.1
Role-Play	61.8	72.5	60.0	71.2	85.8	71.2	64.2	87.5	96.7	80.0	74.6
Tutorial	75.0	**77.5**	**80.0**	**81.2**	86.7	37.5	75.8	86.7	94.2	90.0	78.4
Chain-of-Charts (ours)	**78.9**	52.5	75.0	71.3	**89.2**	**86.3**	**76.7**	**96.7**	**97.5**	**92.5**	**83.5**
(b) Synergistic Effect of Visual and Textual Prompts											
Basic Textual Prompts	77.5	62.5	60.0	68.8	89.2	62.5	63.3	95.0	98.3	90.0	79.4
ChartCoT	80.7	65.0	55.0	71.2	85.0	56.2	60.8	91.7	92.5	87.5	78.0
Role-Play	77.9	70.0	55.0	75.0	90.8	55.0	61.7	94.2	99.2	85.0	79.4
Tutorial	80.0	**80.0**	**75.0**	**85.0**	**92.5**	32.5	75.8	90.8	99.2	92.5	81.6
Chain-of-Charts (ours)	**83.2**	62.5	65.0	65.0	90.8	**77.5**	**77.5**	**96.7**	**99.2**	**95.0**	**84.3**

ment over using Chain-of-Charts alone is slight. We discuss the possible reasons behind:

First, after carefully analyzing the experimental results, we discover that GPT-4o exhibits a certain degree of hallucination in chart understanding. For example, even if the calculation process is accurate, GPT-4o may provide answers that do not match any of the multiple-choice options, leading to incorrect results. This indicates that the model's accuracy is significantly affected by hallucination. Moreover, we also observe that GPT-4o might output numerical information unrelated to the chart even when explicitly recognizing values, further evidencing the hallucination phenomenon in chart reading.

Finding-12: *Adding a Visual Prompt improves performance, but its impact is limited when applied to the Chain-of-Charts strategy.*

4 Related Work

Low-Level ChartQA Tasks. Low-level data analysis tasks in chart involve activities such as data retrieval and correlation determination. These tasks were defined by Amar et al. (2005) and later evaluated by Saket et al. (2019) in a crowdsourced experiment. In this paper, we use these ten tasks as a framework to assess the effectiveness of MLLMs in low-level ChartQA.

Evaluating MLLMs in ChartQA Tasks. Recent studies (Cheng et al., 2023; Han et al., 2023; Huang et al., 2023; Masry et al., 2023; Xia et al., 2024; Xu et al., 2023; Zhou et al., 2023) have leveraged MLLMs to perform high-level ChartQA tasks, such as chart captioning. For example, Huang et al. (2023) evaluated the capabilities of representative MLLMs, such as GPT-4V and Gemini (Team et al., 2024), on chart captioning tasks, finding challenges in accurately reflecting factual chart information. Diverging from this focus on high-level tasks, our research uniquely targets low-level ChartQA tasks (Saket et al., 2019).

ChartQA Datasets. In the last decade, several ChartQA datasets have been presented, as shown in Table 1. For example, ChartBench (Xu et al., 2023) includes 2.1K charts for four types of ChartQA tasks. However, a gap remains - no existing dataset comprehensively evaluates the 10 critical low-level ChartQA tasks (Amar et al., 2005). In addition, to conduct customized evaluations, we need access to chart metadata (*e.g.,* underlying data), not just images. Therefore, we curate a large-scale dataset, ChartInsights, consisting of 22,347 quartets - each with a chart, a query, and its answer.

We also include more detailed discussion about the related work in Appendix E.

5 Conclusion

In this paper, we curate a large-scale dataset, ChartInsights, specifically designed for low-level ChartQA tasks. To evaluate the capabilities of MLLMs on these tasks, we conduct a series of experiments using 19 widely used MLLMs from multiple perspectives. Specifically, we investigate the impact of chart variants and visual prompts on performance, demonstrating the importance of chart quality and visual attention. We also propose a new textual prompt strategy, named Chain-of-Charts, to harness the capabilities of MLLMs for low-level ChartQA. By incorporating visual prompts, we achieve the best average accuracy of 84.32% using GPT-4o. Future work can explore incorporating data prompts and multi-agent frameworks to further enhance the effectiveness of MLLMs in diverse ChartQA tasks.

Limitations

Limited Chart Types. Our experiments set benchmarks for the performance of GPT-4o across seven widely used chart types, providing valuable insights into the model's capabilities in low-level ChartQA tasks. However, this focus inherently excludes a range of more complex chart types, such as heatmaps, radar charts, and others, which present unique analytical challenges and opportunities for data representation. Therefore, including a more diverse chart type, especially those with complex structure and interpretation such as heat maps and radar charts, will provide a more comprehensive perspective on ChartQA for MLLM. This extension is critical for assessing the adaptability and effectiveness of MLLM in a wider range of graph interpretation tasks.

Limited Visual Prompts Design Space . Our exploration of visual prompts in facilitating ChartQA tasks with GPT-4o has shown their potential to enhance model performance. However, our investigation into the design space of visual prompts has been preliminary, lacking a comprehensive and systematic exploration of the full spectrum of possibilities. This limitation narrows the scope of our findings and potentially overlooks more effective strategies that could further improve the accuracy and efficiency of MLLMs in interpreting and analyzing charts.

Lacking of Considering the Data Prompts. Our approach primarily relied on chart images, neglecting the underlying data that generated these charts. This omission could hinder the model's ability to perform more complex analysis and reasoning based on the actual data points. Future work could explore integrating the underlying data as part of the prompt, potentially through multimodal inputs, to provide a richer context for the model's analyses.

Without Fine-tuning MLLMs. We only use the "off-the-shelf" GPT-4o to conduct evaluation, without considering other MLLMs because GPT-4o is known as one of the best models in the visual question-answering task. In addition, we don't perform task-specific fine-tuning because we want to benchmark GPT-4o in low-level tasks and investigate the impact of textual and visual prompts, which is orthogonal to fine-tuning the MLLMs. Future work can fine-tune MLLMs using our dataset to investigate their effectiveness.

Acknowledgements

This paper is supported by NSF of China (62402409, 62132017), Guangdong Basic and Applied Basic Research Foundation (2023A1515110545), CCF-Huawei Populus Grove Fund (CCF-HuaweiDB202403), Shandong Provincial Natural Science Foundation under Grant (ZQ2022JQ32), Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (24XNKJO3).

References

Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pages 111–117. IEEE.

Anthropic. 2024. Claude 3.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. Preprint, arXiv:2308.12966.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2023. Making large multimodal models understand arbitrary visual prompts. arXiv preprint arXiv:2312.00784.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. Preprint, arXiv:2102.10407.

Zhi-Qi Cheng, Qi Dai, and Alexander G. Hauptmann. 2023. Chartreader: A unified framework for chart derendering and comprehension without heuristic rules. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 22202–22213.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. Preprint, arXiv:2306.13394.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. Preprint, arXiv:2311.16483.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. Preprint, arXiv:2404.06395.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do large language models know about facts? Preprint, arXiv:2310.05177.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024a. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. Preprint, arXiv:2403.12027.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R. Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2023. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. Preprint, arXiv:2312.10160.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. 2024b. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. Preprint, arXiv:2401.08276.

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 (CVPR).

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. Preprint, arXiv:1710.07300.

Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim. 2023. Exploring chart question answering for blind and low vision users. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (ACM CHI), pages 1–15.

Nicholas Kong and Maneesh Agrawala. 2012. Graphical overlays: Using layered elements to aid chart reading. IEEE Trans. Vis. Comput. Graph., 18(12):2631–2638.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. Preprint, arXiv:2307.16125.

Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. 2024a. The dawn of natural language to SQL: are we fully ready? Proc. VLDB Endow., 17(11):3318–3331.

Guozheng Li, Runfei Li, Yunshan Feng, Yu Zhang, Yuyu Luo, and Chi Harold Liu. 2024b. Coinsight: Visual storytelling for hierarchical tables with connected insights. IEEE Trans. Vis. Comput. Graph., 30(6):3049–3061.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning (ICML), pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Preprint, arXiv:2201.12086.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. Preprint, arXiv:2311.17005.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. Preprint, arXiv:2311.10774.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. Advances in neural information processing systems (NeurIPS), 36.

Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuyu Luo, Yuxin Zhang, Ju Fan, Guoliang Li, and Nan Tang. 2024c. A survey of NL2SQL with large language models: Where are we, and where are we going? CoRR, abs/2408.05109.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. Preprint, arXiv:2310.02255.

Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020a. cogvlmlean: Interactive cleaning for progressive visualization. Proc. VLDB Endow., 13(12):2821–2824.

Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020b. Interactive cleaning for progressive visualization through composite questions. In ICDE, pages 733–744. IEEE.

Yuyu Luo, Wenbo Li, Tianyu Zhao, Xiang Yu, Lixi Zhang, Guoliang Li, and Nan Tang. 2020c. Deeptrack: Monitoring and exploring spatio-temporal data - A case of tracking COVID-19 -. Proc. VLDB Endow., 13(12):2841–2844.

Yuyu Luo, Xuedi Qin, Chengliang Chai, Nan Tang, Guoliang Li, and Wenbo Li. 2022a. Steerable self-driving data visualization. IEEE Trans. Knowl. Data Eng., 34(1):475–490.

Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018a. Deepeye: Towards automatic data visualization. In ICDE, pages 101–112. IEEE Computer Society.

Yuyu Luo, Xuedi Qin, Nan Tang, Guoliang Li, and Xinran Wang. 2018b. Deepeye: Creating good data visualizations by keyword search. In SIGMOD Conference, pages 1733–1736. ACM.

Yuyu Luo, Jiawei Tang, and Guoliang Li. 2021a. nvbench: A large-scale synthesized dataset for cross-domain natural language to visualization task. arXiv preprint arXiv:2112.12926.

Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. 2021b. Synthesizing natural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. In SIGMOD Conference, pages 1235–1247. ACM.

Yuyu Luo, Nan Tang, Guoliang Li, Wenbo Li, Tianyu Zhao, and Xiang Yu. 2020d. Deepeye: A data science system for monitoring and exploring COVID-19 data. IEEE Data Eng. Bull., 43(2):121–132.

Yuyu Luo, Nan Tang, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. 2022b. Natural language to visualization by neural machine translation. IEEE Trans. Vis. Comput. Graph., 28(1):217–226.

Yuyu Luo, Yihui Zhou, Nan Tang, Guoliang Li, Chengliang Chai, and Leixian Shen. 2023. Learned data-aware image representations of line charts for similarity search. Proc. ACM Manag. Data, 1(1):88:1–88:29.

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. Preprint, arXiv:2305.14761.

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

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. arXiv preprint arXiv:2401.02384.

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).

Modelbest. 2024. Openbmb omnilmm-12b. https://github.com/OpenBMB/OmniLMM.

Tamara Munzner. 2014. Visualization analysis and design. CRC press.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. Preprint, arXiv:2311.16103.

OpenAI. 2023. GPT-4V(ision) system card.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and et al. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2020. Making data visualization more efficient and effective: a survey. VLDB J., 29(1):93–117.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. Preprint, arXiv:2103.00020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. Preprint, arXiv:2204.06125.

Bahador Saket, Alex Endert, and Çagatay Demiralp. 2019. Task-based effectiveness of basic visualizations. IEEE Trans. Vis. Comput. Graph., 25(7):2505–2512.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. Nat., 623(7987):493–498.

Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2023. Towards natural language interfaces for data visualization: A survey. IEEE Trans. Vis. Comput. Graph., 29(6):3121–3144.

Leixian Shen, Enya Shen, Zhiwei Tai, Yun Wang, Yuyu Luo, and Jianmin Wang. 2022. GALVIS: Visualization Construction through Example-Powered Declarative Programming. In CIKM Conference, pages 4975–4979. ACM.

Jiawei Tang, Yuyu Luo, Mourad Ouzzani, Guoliang Li, and Hongyang Chen. 2022. Sevi: Speech-to-visualization through neural machine translation. In SIGMOD Conference, pages 2353–2356. ACM.

Nan Tang, Chenyu Yang, Ju Fan, Lei Cao, Yuyu Luo, and Alon Y. Halevy. 2024. Verifai: Verified generative AI. In CIDR. www.cidrdb.org.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2024. Gemini: a family of highly capable multimodal models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. Preprint, arXiv:1706.03762.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogvlm: Visual expert for pretrained language models. Preprint, arXiv:2311.03079.

Yun Wang, Zhitao Hou, Leixian Shen, Tongshuang Wu, Jiaqi Wang, He Huang, Haidong Zhang, and Dongmei Zhang. 2023. Towards Natural Language-Based Visualization Authoring. IEEE Trans. Vis. Comput. Graph., 29(1):1222 – 1232.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS.

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

Yupeng Xie, Yuyu Luo, Guoliang Li, and Nan Tang. 2024. Haichart: Human and AI paired visualization system. Proc. VLDB Endow., 17(11):3178–3191.

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.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. Preprint, arXiv:2311.04257.

Yilin Ye, Jianing Hao, Yihan Hou, Zhan Wang, Shishi Xiao, Yuyu Luo, and Wei Zeng. 2024. Generative ai for visualization: State of the art and future directions. Preprint, arXiv:2404.18144.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model. Preprint, arXiv:2210.02414.

Zehua Zeng and Leilani Battle. 2023. A review and collation of graphical perception knowledge for visualization recommendation. In CHI, pages 820:1–820:16. ACM.

Mingyang Zhou, Yi R. Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding in vision and language task via cross-modal pre-training on plot table pairs. Preprint, arXiv:2305.18641.

Yizhang Zhu, Shiyin Du, Boyan Li, Yuyu Luo, and Nan Tang. 2024. Are large language models good statisticians? In NeurIPS.

Appendices

A ChartInsights Construction

In this section, we will first discuss the design goals for curating datasets for low-level tasks (Section A.1). We will then provide details of constructing ChartInsights (Section A.2).

A.1 Design Goals for ChartInsights

G1: Supporting Low-level Data Analysis Tasks. Our first goal is to facilitate the support of 10 low-level data analysis tasks (Amar et al., 2005; Saket et al., 2019). This focus addresses a critical gap in existing ChartQA datasets, which often overlook the granularity required to fully understand and interact with the data presented in charts.

G2: Evaluating Visual and Textual Variants on Charts. We highlight the critical role of visual variants (*e.g.* color, size, shape) in data visualizations, which are key to conveying and interpreting information effectively. Despite their importance, these variants are often neglected in existing ChartQA datasets and evaluations. Our goal is to address this issue by incorporating a diverse array of visual variants, including varying chart elements, image quality, and visual prompts. In addition, we also want to investigate the impact of different textual prompts on the low-level analysis task.

G3: Making Metadata Available. The third goal tackles the prevalent issue of inaccessible data and metadata in current ChartQA datasets. By offering comprehensive access to each chart's metadata, such as the source data, chart type, and visual element specifics (like color schemes and labels), our dataset enhances analytical depth into chart design's impact on ChartQA performance.

A.2 Construction Pipeline for ChartInsights

To fulfill our three design goals, our construction process begins with the collection of charts with metadata from existing datasets. After collecting and reviewing a large number of datasets, we decided to extract charts from nvBench and ChartQA. The reason is that most charts in these two datasets contain numerical information of elements, which can meet the requirements for 10 low-level ChartQA tasks. We extracted approximately 900 charts from ChartQA (Masry et al., 2022) and about 1100 charts from nvBench (Luo et al., 2021a). Next, we meticulously assign specific low-level data analysis tasks to appropriate chart types. Lastly, we develop diverse textual prompt strategies, along with visual variants and prompts, tailored to each chart. Note that we save all metadata during the construction process, which can make the users customize their dataset based

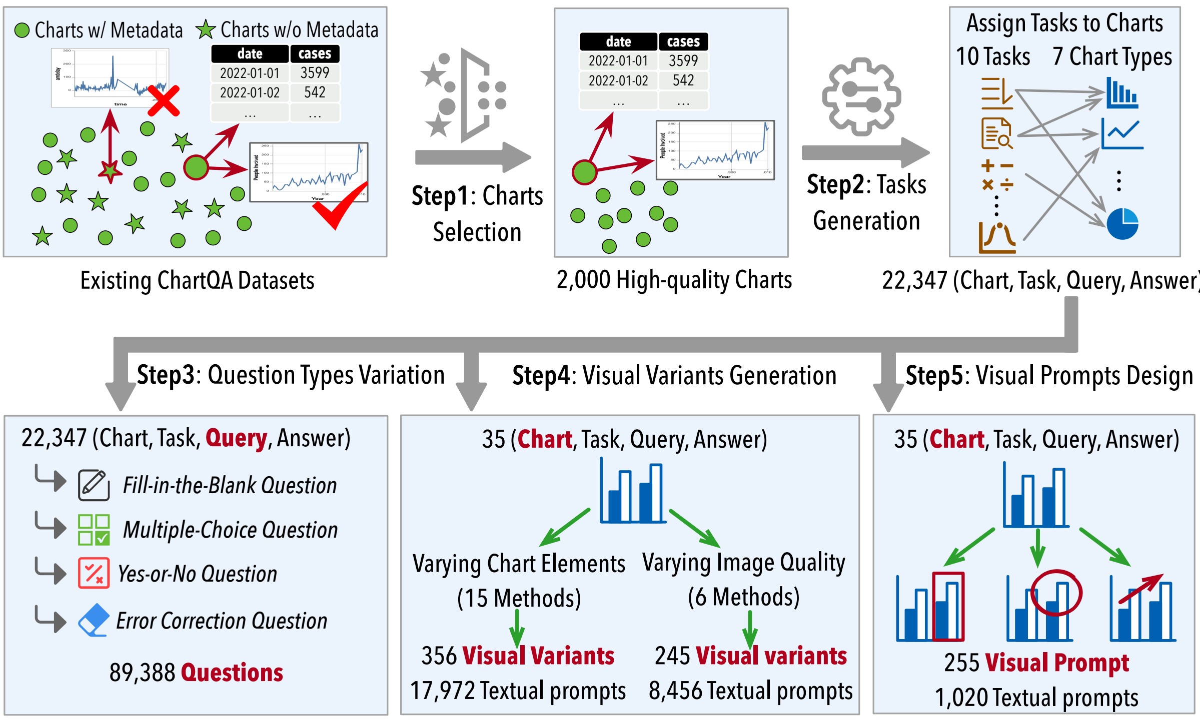


Figure 7: The Pipeline for ChartInsights Construction

on ChartInsights easily.

As shown in Figure 7, the construction of our ChartInsights consists of five steps: Candidate Charts Selection, Low-Level Tasks Generation, Textual Prompts Design, Visual Variants Generation, and Visual Prompts Design.

Step 1: Candidate Charts Selection. In order to more comprehensively evaluate the ability of MLLMs on low-level data analysis tasks, and to conduct more detailed and extended experiments, the datasets (tabular data) and visualization charts we collected need to meet the following three requirements: First, these datasets should contain the original metadata of the chart such as the underlying data for rendering, allowing us to create customized reasoning tasks based on the metadata data. Second, the charts in these datasets should contain data labels, because the lack of data labels will greatly limit the types of low-level tasks. Taking Figure 19 as an example, the chart shows data for different days of the week, and the data labels are as follows: the data label corresponding to Fri is 92020 and the data label corresponding to Mon is 75806. Third, these datasets should contain both simple and complex charts so that the difficulty of the charts is reasonable.

Then, we get a total of 2K high-quality charts as well as their metadata as our initial dataset. The initial dataset contains a total of 7 types of charts, namely stacked bar charts, grouped bar charts, basic bar charts, line charts, grouped line charts, scatterplots, and pie charts.

Step 2: Low-level Tasks Generation. Next, we design a set of low-level tasks for the collected charts. We follow the approach of previous works on designing low-level tasks for charts (Amar et al., 2005; Munzner, 2014; Saket et al., 2019), resulting in 10 low-level tasks in this paper, as shown in top of Figure 3. We group the 10 low-level tasks into three categories, namely Analysis, Search, and Query, based on their purpose and required reasoning abilities (Munzner, 2014).

Next, we should decide which tasks are applicable to which types of charts. We will follow the recommendations on the task-based effectiveness of humans to assign the tasks to each chart type (Saket et al., 2019). Finally, we have 22,347 (chart, task, question, answer).

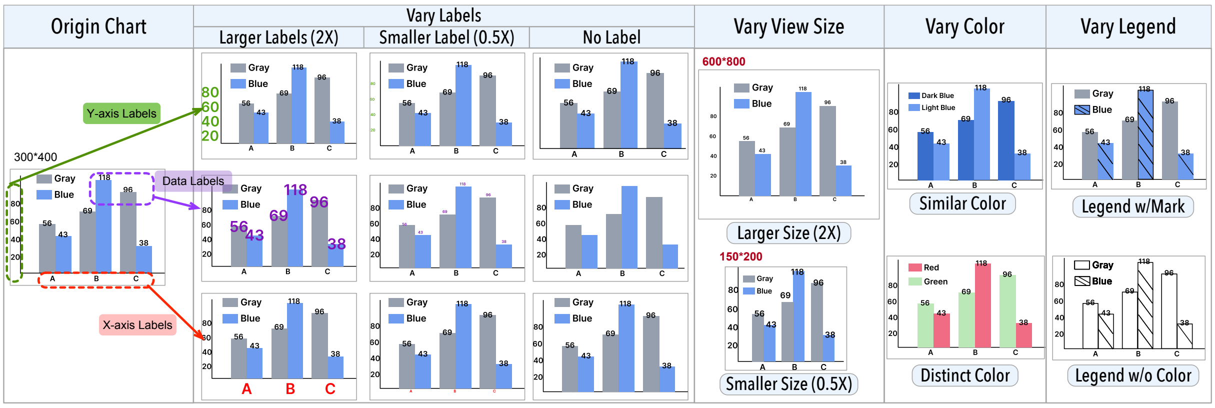
Step 3: Question Type Variation. In order to better explore the impact of different types of questions influence the interaction with MLLMs. We have designed 4 question types, namely Fill-in-the-Blank, Multiple-choice, Yes-or-No, and Error Correction questions. 1) For Fill-in-the-Blank prompt, we maintain the asking method of the initial question and set the answer format for Fill-in-the-Blank prompt; 2) For Multiple-choice prompt, we still maintain the asking method of the initial question, but at this time we will provide a list of choices for MLLMs, which usually contains one correct answer and two wrong answers, and tells MLLMs to choose the answer from the options; 3) For Yes-or-No prompt, we first change the initial question to a true or false question and tell MLLMs whether it needs to be answered correctly or Wrong; and 4) For Error Correction prompt, we put the wrong answer into the original question with a certain probability and change it into a statement.

We vary the 22,347 quartets (chart, task, query, answer) by the four question types mentioned above, resulting in 89,388 quartets (chart, task, question, answer).

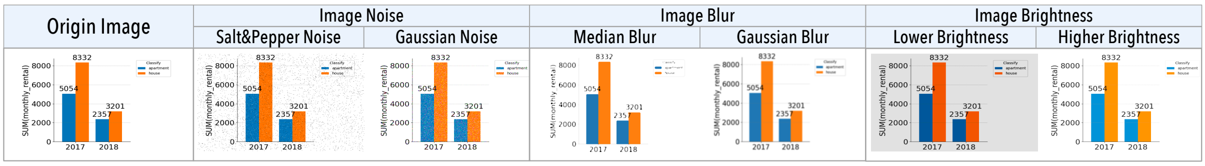
Step 4: Visual Variants Generation. Visual variants (*e.g.* color, size, shape) of a chart play a key role in delivering insights, but these variants are often overlooked in existing ChartQA datasets and evaluations, and thus we aim to bridge this gap. To this end, we vary the chart elements and add image noise to vary the chart quality.

Step 4.1: Varying the Chart Elements. As shown in Figure 8(a), we change the visual elements of these charts from four aspects, namely labels, chart scale, element color, and legend. To achieve this, we sample 5 charts from each category of charts as seeds, resulting in 35 charts. For varying labels, we enlarge, reduce, and remove the x-axis, y-axis, and data labels, respectively. For varying view sizes, we enlarge and reduce the chart, respectively. For varying element color, we change the elements in the chart to the same color or a higher contrast color; For varying legend, we first add marks to different types of categories, and then delete the colors. Finally, we generate 356 visual variants for 35 charts. These 356 visual variants (charts) are associated with 17,972 textual prompts and cover 10 low-level tasks.

Step 4.2: Varying the Image Quality. We add image noise, apply image blur, and adjust the brightness to vary the chart image quality, as shown in Figure 8(b). To achieve this, we sample 5 charts from each category of charts as seeds, resulting in



(a) Vary Chart Elements



(b) Vary Image Quality

Figure 8: Vary Visual Elements on Charts. (a) We vary chart labels, view size, color, and legend in a total of 15 ways. (b) We alter the image quality by adding noise, applying blur, and adjusting brightness.

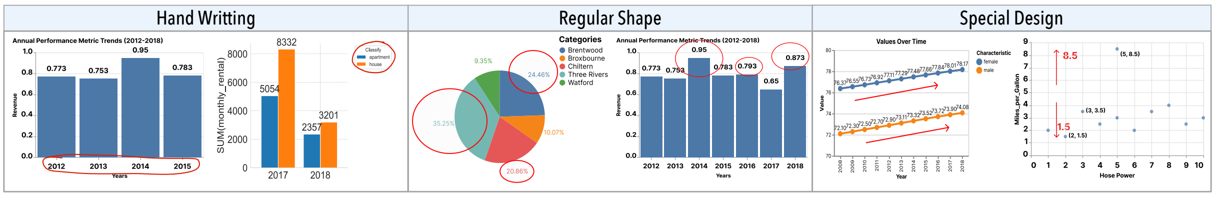


Figure 9: Three Types of Visual Prompts.

35 charts. For adding image noise, we choose Gaussian noise and salt and pepper noise; For applying image blue, we use median blur and Gaussian blur; For adjusting image brightness, we choose to make the brightness of the chart higher and lower. Finally, we generate 245 visual variants for 35 charts. These 245 visual variants (charts) are associated with 8,456 textual prompts and cover 10 low-level tasks.

Step 5: Visual Prompts Design. Kong et al. (Kong and Agrawala, 2012) presented five types of graphical overlays to enhance users' capabilities in performing data analysis tasks such as extraction and comparison of numerical values. Intuitively, we want to verify whether overlays would have a positive impact on MLLMs's performance. Therefore, we design three types of visual prompts (*i.e.* graphical overlays) for the charts.

We consider three types of visual prompts, as shown in Figure 9. The first is to directly circle the content in the chart that is highly relevant to the question in handwriting, such as circling the values of the two elements mentioned in the reasoning question. The second method is regular shapes, which uses regular shapes (such as circles or rectangles) to label elements in the diagram. This makes it easier to use the size of a shape to imply the sequential relationship of elements. For example, use three circles of different sizes to correspond to the three values in the ordering task. The third way is special design. We design effective visual prompts tailored for different low-level tasks. For example, we use arrows to represent the monotonicity of the trend, for the correlation task. To generate the visual prompts, we first sample 35 charts from seven chart types, then apply various visual prompt strategies to them, resulting in 255 charts with different visual prompts. These 255 charts are associated with 1020 questions for 10 low-level tasks.

B Ten Low-Level ChartQA Tasks

Our ChartInsights include 10 low-level data analysis tasks on charts, as shown in Figure 10. These tasks are well-defined by the visualization and vi-

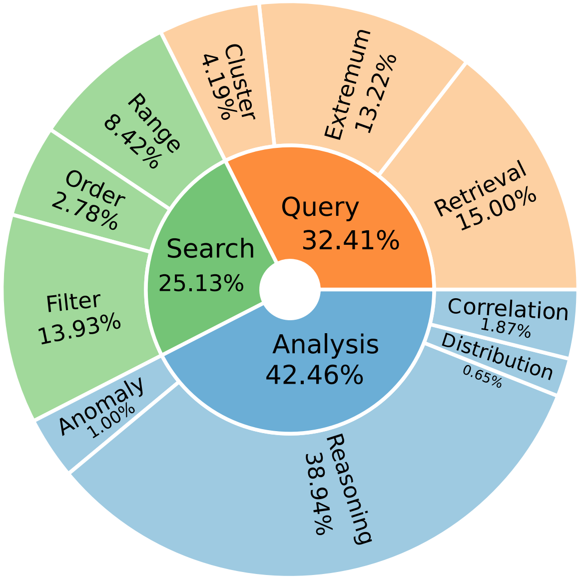


Figure 10: Low-level Tasks Distribution

sual analysis community (Amar et al., 2005; Saket et al., 2019).

T1: Data Retrieval. Users will be asked to locate the value of an element based on some information, or they may be asked to answer structural questions such as how many elements there are in the chart. Figure 11 shows an example.

T2: Reasoning. Users need to calculate the aggregate value of multiple elements based on the data in the chart and the requirements of the question. Figure 12 shows an example.

T3: Filter. This task will randomly select a value of an element in the current chart as a benchmark, and users need to filter the remaining elements according to this benchmark and the requirements of the question. Figure 13 shows an example.

T4: Determine Range. This task requires users to determine the value range of the chart based on the values of the elements in the chart. Figure 14 shows an example.

T5: Cluster. As shown in Figure 15, this task requires users to return the number of categories of elements in the chart.

T6: Find Extreme. This question requires users to find the maximum and minimum values in the chart and return them, as depicted in Figure 16.

T7: Correlation. Users need to determine the relationship between the changes in the elements in the chart and the axes. Some questions can be

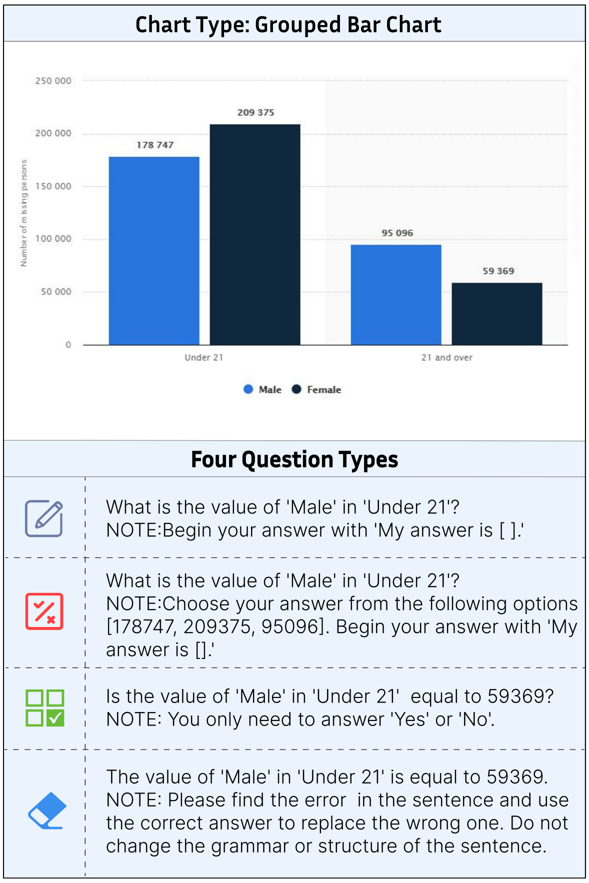


Figure 11: An Example of Data Retrieval Tasks

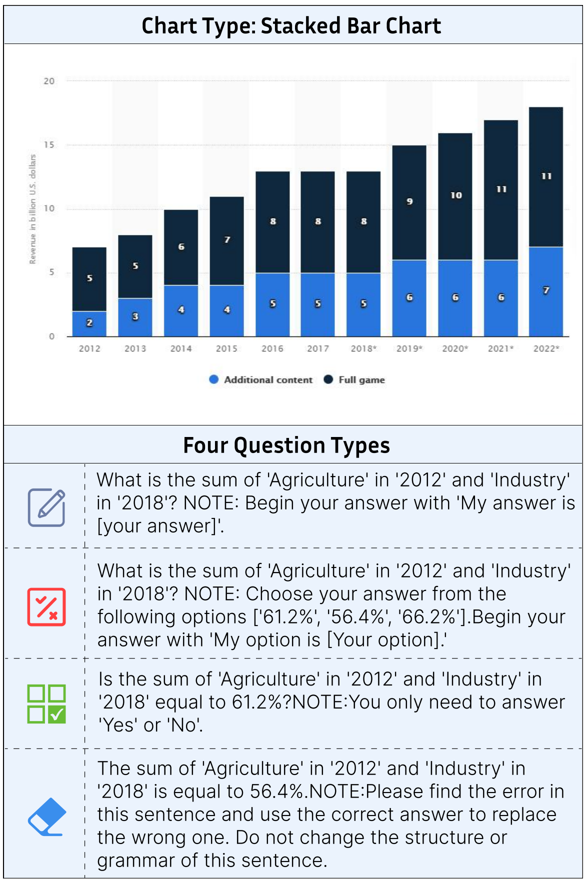
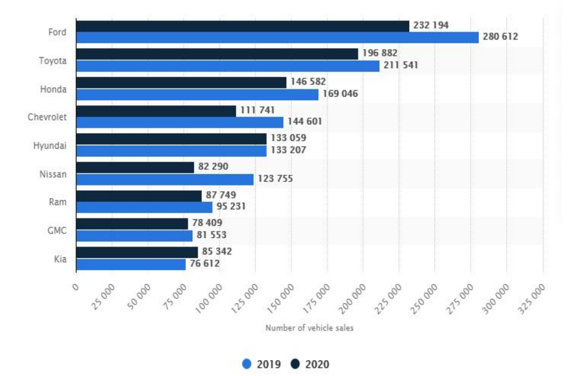


Figure 12: An Example of Reasoning Tasks

Chart: Grouped Bar Chart

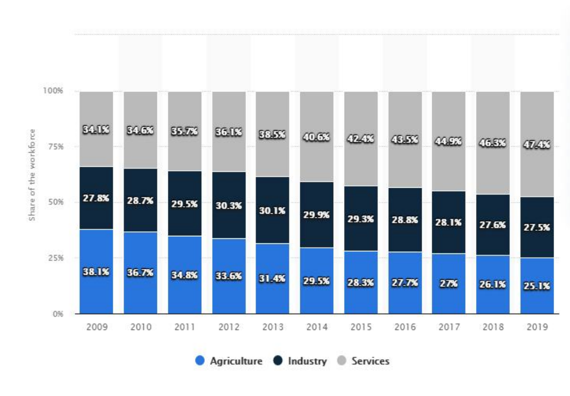


Four Question Types

	How many bars have smaller values than the bar of '2019' in 'GMC'? NOTE:Begin your answer with 'My answer is [your answer]'.
	How many bars have smaller values than the bar of '2019' in 'GMC'? NOTE: Choose your answer from the following options [2, 3, 4].Begin your answer with 'My option is [Your option].'
	Is the number of bars have smaller values than the bar of '2019' in 'GMC' equal to 4? NOTE: You only need to answer 'Yes' or 'No'.
	The number of bars have smaller values than the bar of '2019' in 'GMC' is equal to 3. NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one.

Figure 13: An Example of Filter Tasks

Chart: Stacked Bar Chart

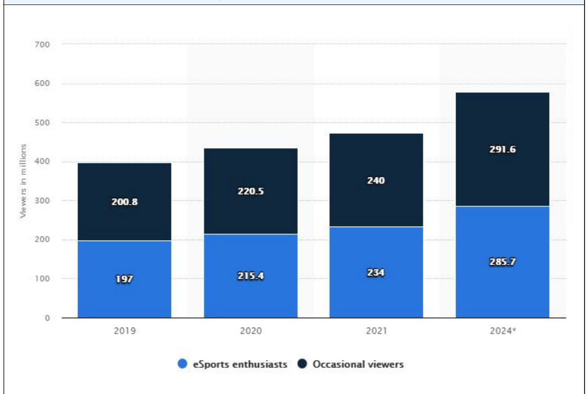


Four Question Types

	How many categories are there in the chart? NOTE: Begin your answer with 'My answer is [your answer]'.
	How many categories are there in the chart? NOTE: Choose your answer from the following options [4, 3, 5].Begin your answer with 'My option is [Your option].'
	Is the number of bars have smaller values than the bar of '2019' in 'Industry' equal to 4?NOTE:You only need to answer 'Yes' or 'No'.
	The number of categories in the chart is equal to 4.NOTE:Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 15: An Example of Cluster Tasks

Chart Type: Basic Bar Chart

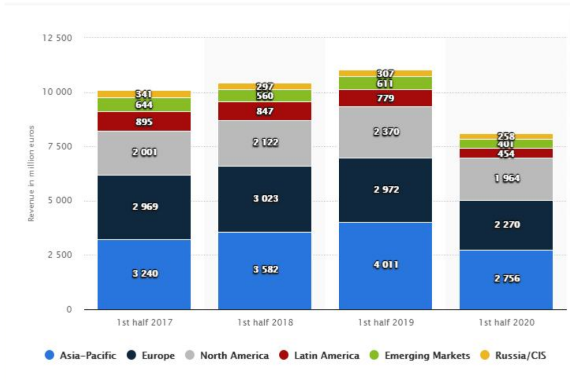


Four Question Types

	What is the range of bar values? NOTE: Begin your answer with 'My answer is [your answer]'.
	What is the range of bar values? NOTE: Choose your answer from the following options [(197.0, 220.5), (220.5, 291.6), (197.0, 291.6)]. Begin your answer with 'My option is [Your option].'
	Is the range of bar values from 197.0 to 291.6? NOTE: You only need to answer 'Yes' or 'No'.
	The range of bar values is from 197.0 to 220.5. NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 14: An Example of Determine Range Tasks

Chart Type: Stacked Bar Chart



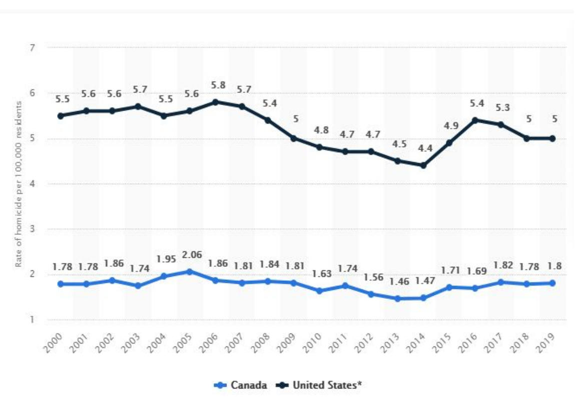
Four Question Types

	What is the maximum value in this chart? NOTE: Begin your answer with 'My answer is [your answer]'.
	What is the maximum value in this chart? NOTE: Choose your answer from the following options [4011, 2969, 2270].Begin your answer with 'My option is [Your option].'
	Is the maximum value in this chart 4011? NOTE: You only need to answer 'Yes' or 'No'.
	The maximum value in this chart is 2969. NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 16: An Example of Find Extreme Tasks

answered visually, but many require a reasoning process. Figure 17 shows an example.

Chart Type: Grouped Line Chart



Four Question Types

How many lines in this chart are monotonic? NOTE: Begin your answer with 'My answer is [your answer]'.

How many lines in this chart are monotonic? NOTE: Choose your answer from the following options [1, 0, 2].Begin your answer with 'My option is [Your option].'

2 lines in this chart are monotonic. NOTE: You only need to answer 'Yes' or 'No'.

1 line in this chart are monotonic. NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

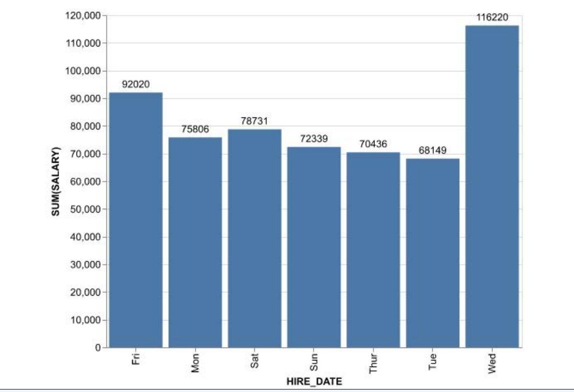
Figure 17: An Example of Correlation Tasks

T8: Find Anomaly. This task requires users to identify values that appear to be anomalies based on their different values, either visually or through calculation. Figure 18 shows an example.

T9: Order. This task involves sorting the elements in the chart. Users will be asked to sort the elements in ascending or descending order of value and output the names of the top three elements for each type of sorting. Figure 19 shows an example.

T10: Distribution. This task is mainly for scatter plots, where users need to determine the distribution range of the dots. Figure 20 shows an example.

Chart Type: Basic Bar Chart



Four Question Types

Which bar seems to have abnormal length? NOTE: Begin your answer with 'My answer is [your answer]'.

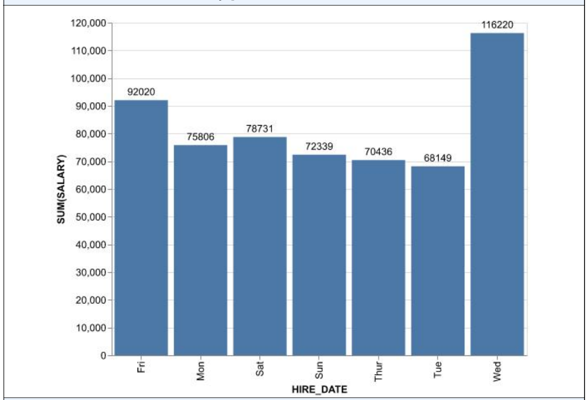
Which bar seems to have abnormal length? NOTE: Choose your answer from the following options ['Wed', 'Mon', 'Sun'].Begin your answer with 'My option is [Your option].'

Dose the bar of 'Wed' seem to have abnormal length? NOTE: You only need to answer 'Yes' or 'No'.

The bar of 'Mon' seem to have abnormal length. NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 18: An Example of Find Anomaly Tasks

Chart Type: Basic Bar Chart



Four Question Types

What are the x-axis names of the top three bars in descending order?NOTE:Begin your answer with 'My answer is [your answer]'.

What are the x-axis names of the top three bars in descending order?NOTE:Choose your answer from the following options [('Wed', 'Fri', 'Sat'), ('Wed', 'Sat', 'Fri'), ('Fri', 'Wed', 'Sat')].Begin your answer with 'My option is [Your option].'

Are these the x-axis names of the top three bars in descending order ('Wed', 'Fri', 'Sat')?NOTE:You only need to answer 'Yes' or 'No'.

These the x-axis names of the top three bars in descending order ('Wed', 'Sat', 'Fri').NOTE:Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 19: An Example of Order Tasks

Table 5: The Average Accuracy of Different Chart Types vs.Ten Low-level Tasks ("–" means "N/A") on GPT-4o

Chart Types	Analysis				Search			Query			Overall (%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
Grouped Bar	43.75	–	–	–	83.68	–	37.5	80.79	96.97	63.42	62.72
Stacked Bar	43.75	–	–	–	58.46	–	38.97	79.41	81.07	85.66	59.83
Grouped Line	31.55	–	–	71.43	80.95	–	–	–	–	–	53.87
Basic Bar	**82.98**	**96.15**	–	–	**95.17**	**72.34**	**65.77**	**87.38**	**99.43**	–	**84.05**
Basic Line	60.94	–	–	36.72	89.06	–	–	–	–	–	56.87
Scatter Plot	72.79	18.14	**70.10**	**87.25**	78.43	–	–	–	96.08	**94.23**	74.58
Pie	71.59	–	–	–	78.41	59.72	61.65	80.11	–	–	70.86

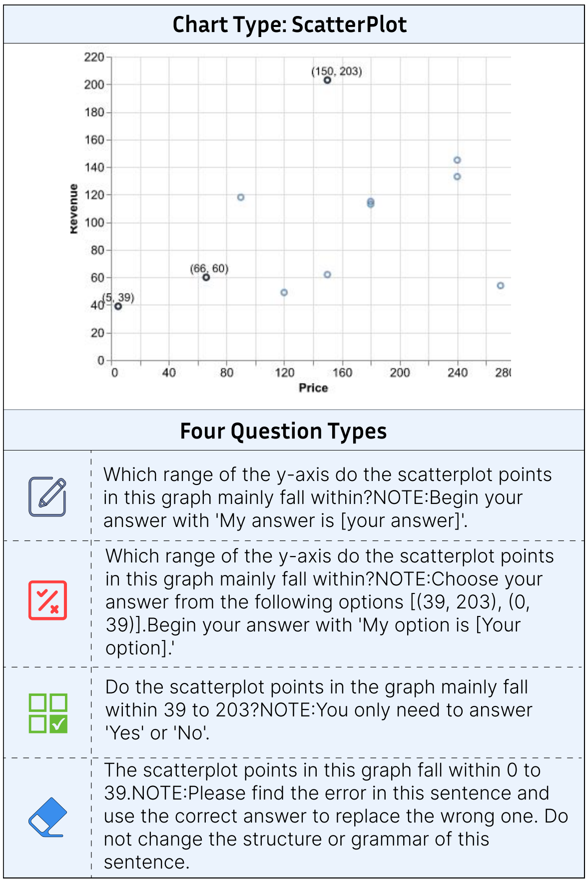


Figure 20: An Example of Distribution Tasks

C More Experimental Details on GPT-4o

The results include more details about evaluation on GPT-4o.

More Discussions on Exp-1. Table 5 shows the overall accuracy of GPT-4o for different chart types in 10 low-level tasks. Overall, GPT-4o has the highest performance on basic bar charts, reaching an average accuracy of 84.05%. The main reason is that the chart structure of the basic bar chart is relatively simple. Similarly, GPT-4o achieves better results on charts with simple structures such as scatter plots and pie charts. For charts with complex structures, such as stacked bar charts, grouped bar charts, and grouped line charts, the average accuracy of GPT-4o is close to 50%.

More Discussions on Exp-2. Table 6 presents the overall performance of GPT-4o across 10 low-level tasks with four Question Types. Specifically, GPT-4o exhibits the highest overall accuracy with the Multiple-Choice prompt, achieving 74.79%. In addition, it also performs well with the Yes-or-No prompt, with 77.78% accuracy.

More Discussions on Exp-5. Heatmaps (Figure 6) demonstrate the percentage change in performance across different chart types and task types when chart elements are varied. Bar charts summarize the overall effect, indicating that data labels significantly influence performance. It consists of two main parts:

Chart Types (a): This heatmap shows the change in performance (in percentages) for various chart types when different elements (e.g., data labels, X/Y labels, size, color) are modified. Performance varies widely, with removing data labels and X labels often result in a significant drop in performance, while larger data labels and different colors often improve performance.

Task Types (b): This heatmap shows how the performance of different task types changes when similar chart elements change. Tasks such as reasoning and finding extremum are particularly sensitive to change, and performance can be significantly degraded when data labels are removed or reduced.

More Discussions on Exp-6. Heatmaps (Figure 21) show how different types of image quality changes affect GPT-4o's performance on various chart types and tasks. Median blur has the most negative impact, making it hard for GPT-4o to read numbers and lowering its performance significantly. On the other hand, changing brightness (either higher or lower) slightly improves performance on most tasks. Overall, poor image quality generally hurts GPT-4o's recognition capability.

Table 6: The Effectiveness of Question Types vs. Ten Low-level Tasks on GPT-4o

Question Types	Analysis				Search			Query			Overall (%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
Fill-in-the-Blank	42.45	34.38	43.14	**75.00**	66.84	66.15	38.81	82.05	92.38	62.50	61.03
Multiple-Choice	**68.56**	**43.75**	**80.39**	73.08	**99.22**	**77.69**	48.81	**93.75**	**99.17**	71.59	**77.78**
Yes-or-No	68.12	21.88	68.63	69.23	86.68	58.46	**67.12**	75.64	92.38	**90.34**	74.79
Error Correction	44.49	35.94	88.24	56.73	69.71	73.08	44.75	79.01	91.89	72.73	63.08
Overall Accuracy (%)	55.91	33.99	70.10	68.51	80.61	68.84	49.87	82.61	**93.96**	74.29	69.17

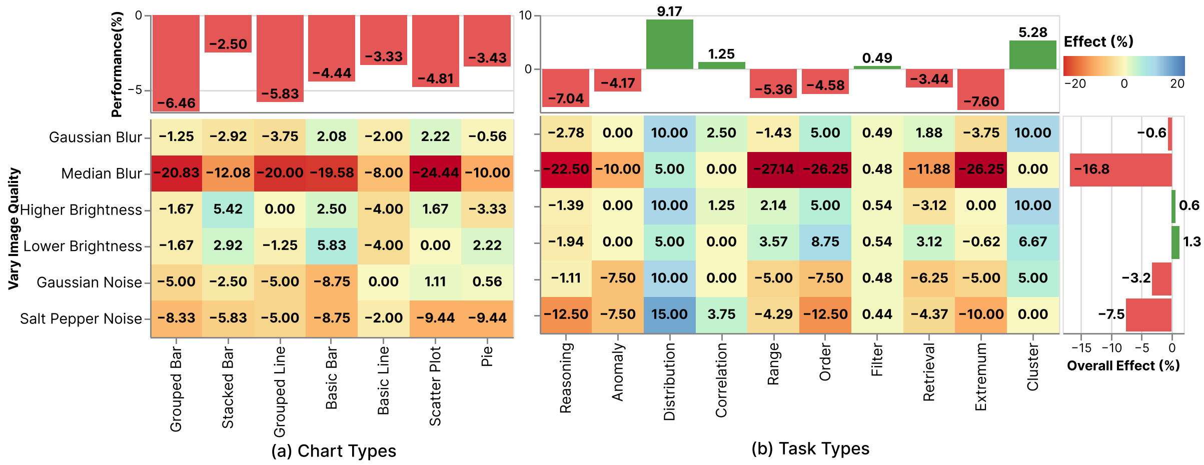
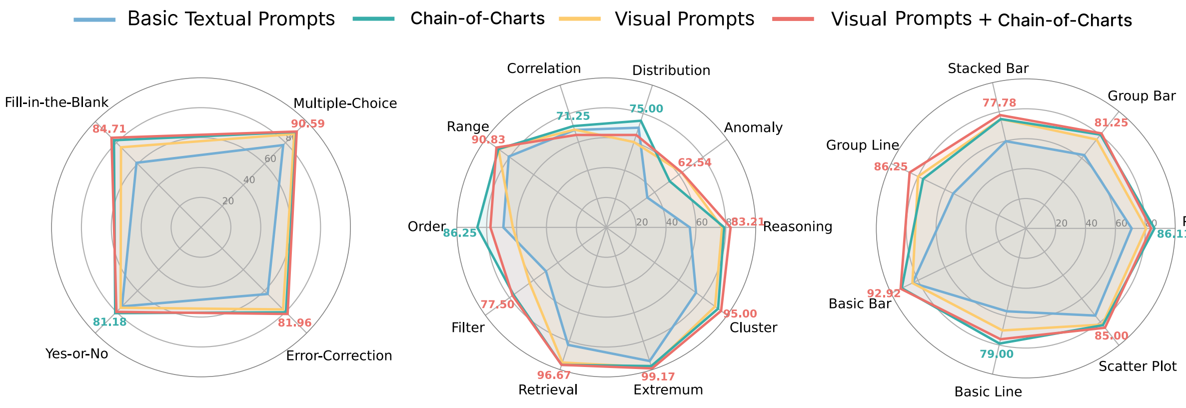


Figure 21: The Impact of Image Quality on GPT-4o's Performance.

Optimizations on Textual and Visual Prompts. The radar charts (Figure 22) clearly demonstrate that the effectiveness of different prompt types varies depending on the question, task and chart types. The accuracy of GPT-4o in Exp-1, Exp-3, Exp-4, and Exp-7 is concatenated in series. The combination of visual prompts and chain-of-chart prompts notably enhances performance in various aspects, showcasing their effectiveness in improving GPT-4o's analytical capabilities.



(a) Accuracy(%) vs. Question Types

(b) Accuracy (%) vs. Low-level Tasks

(c) Accuracy (%) vs. Chart Types

Figure 22: Comparing Different Prompting Methods.

D Lessons Learned

Effectiveness of Textual Prompts, Visual Prompts and Their Combination. Incorporating various prompt strategies, including textual and visual prompts, significantly impacts MLLMs (*e.g.,* GPT-4o) accuracy. Textual prompts with structured candidate answers enhance reasoning capabilities, while visual prompts enhance the chart understanding through visual attention, particularly in anomaly detection and filtering tasks.

Importance of Chart Elements and Image Quality. Alterations in chart elements and the quality of chart images influence MLLMs's performance. Specifically, certain modifications like larger labels or the absence of data labels can improve the model's efficiency in specific tasks by focusing its attention on visual comparisons. However, image quality degradation, especially median blurring, negatively affects the model's ability to process numerical values accurately.

Strengths and Weaknesses of MLLMs in Low-level ChartQA Tasks. MLLMs perform well in tasks requiring direct data retrieval and basic comparisons, showing high accuracy in Query and Search task categories. However, MLLMs face challenges in more complex reasoning, anomaly detection, and correlation tasks, indicating a need for further optimization of prompting strategies and model training to overcome these limitations.

Potential for Future Application and Development. The experiments demonstrate a promising direction for enhancing MLLMs' performance in visual data analysis through the development of specialized prompting strategies and the careful manipulation of visual elements.

E Additional Related Work

E.1 Low-Level Analysis Tasks on Charts

Visualization charts offer numerous insights that aid users in performing data analysis tasks (Luo et al., 2020a,b,c, 2022a, 2018a,b, 2021b, 2020d, 2022b; Qin et al., 2020; Shen et al., 2023, 2022; Tang et al., 2022; Wang et al., 2023). Low-level data analysis tasks typically involve activities requiring direct interpretation and processing of specific visual elements within a chart, such as data retrieval, outlier identification, and correlation determination (Amar et al., 2005; Kim et al., 2023; Munzner, 2014). Amar et al. (Amar et al., 2005) identified ten low-level tasks, highlighting the real-world activities users undertake with visualization tools to understand their data. Subsequently, Saket et al. (Saket et al., 2019) evaluated the effectiveness of five basic charts across ten low-level analysis tasks using two datasets through a crowdsourced experiment. In this paper, we aim to evaluate how effectively GPT-4o can interpret charts by using these ten low-level data analysis tasks as a framework.

E.2 Multimodal Large Language Models

The field of Multimodal Large Language Models (MLLMs) is experiencing rapid advancements, with efforts concentrated on developing artificial intelligence systems capable of processing and producing multi-modal content, including text, images, videos, and more (Zhu et al., 2024). Early research such as CLIP (Radford et al., 2021) demonstrated the effective combination of visual and linguistic information through contrastive learning, while subsequent work like DALL-E (Ramesh et al., 2022) further showcased the potential of Trans-

An Example of Basic Textual Prompt

Initial Question: What is the minimum value in this chart?
Fill-in-the-Blank: What is the minimum value in this chart? Begin your answer with 'My answer is [].'
Multiple-Choice: What is the minimum value in this chart? NOTE: Choose your answer from the following options [A, B ,C]. Begin your answer with 'My option is [Your option].'
Yes-or-No: Is the minimum value in this chart equal to ......? NOTE: You only need to answer 'Yes' or 'No'.
Error-Correction: The minimum value in this chart is equal to ...... NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 23: An Example of Basic Textual Prompt

former (Vaswani et al., 2023) architecture in generating images that match text descriptions. Building on these foundational successes, the research community has ventured into refining these models for diverse multi-modal applications, employing strategies like fine-tuning and prompt-based learning. For example, VisualGPT (Chen et al., 2022) and BLIP (Li et al., 2022) have been adapted for Visual Question Answering (VQA) tasks, significantly enhancing their multi-modal task performance. Concurrently, the development of various benchmarks (Hu et al., 2023; Huang et al., 2024b; Li et al., 2023a, 2024c; Ning et al., 2023), including MME (Fu et al., 2024), has been crucial. These benchmarks provide a wide array of tasks and datasets, facilitating a comprehensive evaluation of MLLMs' abilities across different contexts. In this paper, we try to harness the off-the-shelf MLLMs for low-level data analysis tasks on charts.

E.3 MLLMs for Chart Question Answering

With the advancements in MLLMs, such as GPT-4o, it becomes increasingly promising to automatically comprehend charts and extract insights according to user queries (Huang et al., 2024a; Masry et al., 2022; Shen et al., 2023; Zeng and Battle, 2023). This process is known as chart question answering, *i.e.* ChartQA for short. Recent research efforts have focused on understanding the capabilities of MLLMs in performing ChartQA tasks. These studies can be categorized into two groups: evaluation studies and the construction of datasets for ChartQA.

Evaluating MLLMs on ChartQA Tasks. Several recent studies (Cheng et al., 2023; Han et al., 2023; Huang et al., 2023; Masry et al., 2023; Xia et al., 2024; Xu et al., 2023; Zhou et al., 2023) have attempted to leverage the capabilities of MLLMs to perform high-level ChartQA tasks such as chart captioning and chart-to-text. For example, Huang et al. evaluated the capabilities of representative MLLMs, such as GPT-4V and Bard (*i.e.* Gemini) (Team et al., 2024), on chart captioning tasks. Their findings indicated that GPT-4o faces challenges in generating captions that accurately reflect the factual information presented in charts. Moreover, these studies have highlighted various promising directions for future research in this field. Diverging from the emphasis on high-level tasks in previous works, our research uniquely targets *low-level* ChartQA tasks (Amar et al., 2005; Saket et al., 2019).

ChartQA Datasets. In the last decade, several ChartQA datasets have been presented (Cheng et al., 2023; Han et al., 2023; Kafle et al., 2018; Kahou et al., 2018; Liu et al., 2023; Masry et al., 2022; Methani et al., 2020; Xia et al., 2024; Xu et al., 2023), as shown in Section A Table 1. For example, ChartBench (Xu et al., 2023) includes 2.1K charts for four types of ChartQA tasks. However, a gap remains evident in the landscape of existing ChartQA datasets: none are tailored to comprehensively evaluate the 10 *low-level* tasks identified as critical to the ChartQA task. Moreover, to conduct more customized evaluations, such as modifying the visual elements or adding a visual prompt, we need access to the metadata (*e.g.* the underlying data) of the charts, not just the chart images. Therefore, we curate a large-scale dataset ChartInsights, which consists of a total of 89,388 quartets, each including a chart, a specified task, a corresponding

An Example of RolePlay Prompt

Initial Question: What is the minimum value in this chart?
Fill-in-the-Blank With RolePlay: You are an expert on chart understanding with specialized skills in numerical analysis. Your keen eye for detail allows you to accurately identify and extract numerical values from various chart elements, such as the x-axis/y-axis categories and the legend keys. Your role is to analyze charts, promptly determine the sum or average of specified elements, and communicate your findings in an accessible manner. What is the minimum value in this chart? Begin your answer with 'My answer is [].'
Multiple-Choice With RolePlay: You are an expert on chart understanding with specialized skills in numerical analysis. Your keen eye for detail allows you to accurately identify and extract numerical values from various chart elements, such as the x-axis/y-axis categories and the legend keys. Your role is to analyze charts, promptly determine the sum or average of specified elements, and communicate your findings in an accessible manner. What is the minimum value in this chart? NOTE: Choose your answer from the following options [A, B ,C]. Begin your answer with 'My option is [Your option].'
Yes-or-No With RolePlay: You are an expert on chart understanding with specialized skills in numerical analysis. Your keen eye for detail allows you to accurately identify and extract numerical values from various chart elements, such as the x-axis/y-axis categories and the legend keys. Your role is to analyze charts, promptly determine the sum or average of specified elements, and communicate your findings in an accessible manner. Is the minimum value in this chart equal to ......? NOTE: You only need to answer 'Yes' or 'No'.
Error-Correction With RolePlay: You are an expert on chart understanding with specialized skills in numerical analysis. Your keen eye for detail allows you to accurately identify and extract numerical values from various chart elements, such as the x-axis/y-axis categories and the legend keys. Your role is to analyze charts, promptly determine the sum or average of specified elements, and communicate your findings in an accessible manner. The minimum value in this chart is equal to ...... NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 24: An Example of RolePlay Prompt

query, and its answer.

F Prompts

In this section, we provide detailed descriptions and examples of the prompt strategies used in our evaluations.

F.1 Basic Textual Prompts

Each ChartQA task can be framed in four different question types: Fill-in-the-Blank, Error Correction, Multiple-Choice, and Yes-or-No questions. While the core meaning remains the same across these types, Fill-in-the-Blank and Error Correction questions are more open-ended, Multiple-Choice questions require selecting the correct option from several choices, and Yes-or-No questions involve determining the truthfulness of a statement. Figure 23 shows examples of basic textual prompts for the four question types mentioned above.

F.2 Visual Prompts

In this paper, we design three types of visual prompts: handwriting, regular shape, and special design. Figure 9 shows examples of these visual prompts in our ChartInsights and evaluations.

- *Handwriting Visual Prompts*: These prompts involve manually annotating the relevant visual elements directly on the chart, simulating handwritten notes. This style is particularly useful for tasks like Find Extreme and Data Retrieval, where specific elements need to be located. The handwritten annotations guide the MLLMs to focus on the pertinent parts of the chart.
- *Regular Shape Visual Prompts*: These prompts use simple geometric shapes, such as circles, rectangles, and arrows, to highlight key areas of the chart. This method provides clear and precise indications of important elements and re-

An Example of Tutorial Prompt

Initial Question: What is the minimum value in this chart?
Fill-in-the-Blank With Tutorial: What is the minimum value in this chart? NOTE: Begin your answer with ‘My answer is [your answer]’. Firstly, identify the chart’s basic structure and type to understand the visual elements used in the chart and how these elements represent data. Subsequently, observing the chart title, legend, and axes, which provide essential information about the data’s theme and measurement units. Next, identify key data points, such as significant highs, lows, or trends. Further steps include comparing relationships between different data series and interpreting the proportions of the data. Finally, summarize the information gathered.
Multiple-Choice With Tutorial: What is the minimum value in this chart? NOTE: Choose your answer from the following options [A, B ,C]. Begin your answer with ‘My option is [Your option].’ Firstly, identify the chart’s basic structure and type to understand the visual elements used in the chart and how these elements represent data. Subsequently, observing the chart title, legend, and axes, which provide essential information about the data’s theme and measurement units. Next, identify key data points, such as significant highs, lows, or trends. Further steps include comparing relationships between different data series and interpreting the proportions of the data. Finally, summarize the information gathered.
Yes-or-No With Tutorial: Is the minimum value in this chart equal to ......? NOTE: You only need to answer ‘Yes’ or ‘No’. Firstly, identify the chart’s basic structure and type to understand the visual elements used in the chart and how these elements represent data. Subsequently, observing the chart title, legend, and axes, which provide essential information about the data’s theme and measurement units. Next, identify key data points, such as significant highs, lows, or trends. Further steps include comparing relationships between different data series and interpreting the proportions of the data. Finally, summarize the information gathered.
Error-Correction With Tutorial: The minimum value in this chart is equal to ...... NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence. Firstly, identify the chart’s basic structure and type to understand the visual elements used in the chart and how these elements represent data. Subsequently, observing the chart title, legend, and axes, which provide essential information about the data’s theme and measurement units. Next, identify key data points, such as significant highs, lows, or trends. Further steps include comparing relationships between different data series and interpreting the proportions of the data. Finally, summarize the information gathered.

Figure 25: An Example of Tutorial Prompt

gions, aiding the MLLMs in understanding the chart structure and data distribution.

- *Specially Designed Visual Prompts*: These prompts are tailored to specific low-level chart analysis tasks and incorporate customized visual elements that align with the unique requirements of each task. For instance, color-coded overlays or patterned highlights might be used to draw attention to particular data trends or anomalies.

We manually annotate these visual prompts to assist MLLMs in understanding the specific requirements of low-level chart analysis tasks. By providing clear and targeted visual cues, we aim to enhance the models’ ability to accurately interpret and analyze chart data.

F.3 RolePlay Prompts

RolePlay prompts guide (multimodal) large language models to adopt specific roles, allowing them to perform tasks in accordance with the behaviors and expertise of those roles (Shanahan et al., 2023). In this paper, we assign the role of a data visualization expert to the MLLMs. By simulating the thought processes and actions of an expert, the model can better interpret and analyze chart data. This approach helps the model generate more accurate and contextually relevant responses.

Figure 24 shows examples of RolePlay prompts for the four question types mentioned above, demonstrating how the model, acting as a data visualization expert, addresses the low-level ChartQA

An Example of ChartCoT Prompt

Initial Question: What is the minimum value in this chart? **Fill-in-the-Blank With ChartCoT:** Let's answer following questions one by one: 1. What type is this chart? 2. What are the labels of x-axis? 3. What are the data labels of each element? 4. What is the minimum value in this chart? NOTE: Begin your answer with 'My answer is [your answer]'.
Multiplt-Choice With ChartCoT: Let's answer following questions one by one: 1. What type is this chart? 2. What are the labels of x-axis? 3. What are the data labels of each element? 4. What is the minimum value in this chart? NOTE: Choose your answer from the following options [178747, 95096, 59369]. Begin your answer with 'My option is [Your option].'
Yes-or-No With ChartCoT: Let's answer following questions one by one: 1. What type is this chart? 2. What are the labels of x-axis? 3. What are the data labels of each element? 4. Is the minimum value in this chart equal to ......? NOTE: You only need to answer 'Yes' or 'No'.
Error-Correction With ChartCoT: Let's answer following questions one by one: 1. What type is this chart? 2. What are the labels of x-axis? 3. What are the data labels of each element? The minimum value in this chart is equal to ...... NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 26: An Example of ChartCoT Prompt

tasks.

F.4 Tutorial Prompts

Through our experiments, we have found that the more details provided in the input, the more comprehensive and often more accurate the output from MLLMs becomes. Based on this discovery, we propose the Tutorial Textual Prompt. This approach involves breaking down the steps for reading and understanding visualization charts to guide MLLMs through the analysis process. The Tutorial Prompt provides a detailed, step-by-step explanation of how to interpret various elements of a chart. By explicitly outlining these steps, we aim to enhance the model's ability to process and analyze chart data accurately. This method helps the model to follow a structured approach, ensuring that it considers all relevant aspects of the chart in its analysis.

Figure 25 shows examples of Tutorial prompts for the four question types mentioned above. Specifically, the tutorial might start by instructing the model to identify the type of chart and its key components, such as axes, labels, and legends. It then guides the model through interpreting the data presented in the chart, noting trends, outliers, and significant data points. By providing this structured guidance, the model can generate more precise and contextually relevant responses.

F.5 ChartCoT Prompts

The design of ChartCoT is based on the concept of Chain of Thought (Wei et al., 2022), which involves crafting a series of guiding questions to enable MLLMs to produce high-quality responses. This method encourages the model to think step-by-step, enhancing its reasoning capabilities and ensuring a thorough understanding of the chart data.

In our evaluations, we set up three progressively detailed questions to guide the model's thought process:

- *Chart Type Identification*: The first question typically pertains to identifying the type of chart. This step ensures that the model correctly understands the basic structure and purpose of the chart, whether it is a bar chart, line chart, pie chart, etc.
- *Coordinate Information*: The second question relates to the coordinate information of the chart. Here, the model is prompted to recognize and interpret the axes, scales, and any legends or labels that provide context for the data points. This step is crucial for understanding how the data is organized and presented.
- *Numerical Information*: The third question concerns the numerical information of the elements within the chart. This includes extracting specific data values, identifying trends, and making comparisons between different data points. This

An Example of Chain-of-Charts Prompt

Initial Question: What is the minimum value in this chart?
Fill-in-the-Blank With Chain-of-Charts: Learn from the previous three questions and answers first, and then answer the last question.1. Q: What type is this chart? A: ...... 2. Q: What are the labels of x-axis? A: ...... 3. Q: What are the data labels of each element? A: ...... 4. Q: What is the minimum value in this chart? NOTE: Begin your answer with 'My answer is [your answer]'. A: ......
Multiplt-Choice With Chain-of-Charts: Learn from the previous three questions and answers first, and then answer the last question. 1. Q: What type is this chart? A: ...... 2. Q: What are the labels of x-axis? A: ...... 3. Q: What are the data labels of each element? A: ...... 4. Q: What is the minimum value in this chart? NOTE: Choose your answer from the following options [A, B, C]. Begin your answer with 'My option is [Your option].'
Yes-or-No With Chain-of-Charts: Learn from the previous three questions and answers first, and then answer the last question.1. Q: What type is this chart? A: ...... 2. Q: What are the labels of x-axis? A: ...... 3. Q: What are the data labels of each element? A: ...... 4. Is the minimum value in this chart equal to ......? NOTE: You only need to answer 'Yes' or 'No'.
Error-Correction With Chain-of-Charts: Learn from the previous three questions and answers first, and then answer the last question.1. Q: What type is this chart? A: ...... 2. Q: What are the labels of x-axis? A: ...... 3. Q: What are the data labels of each element? A: ...... 4. The minimum value in this chart is equal to ...... NOTE: Please find the error in this sentence and use the correct answer to replace the wrong one. Do not change the structure or grammar of this sentence.

Figure 27: An Example of Chain-of-Chats Prompt

step ensures that the model can accurately read and analyze the quantitative aspects of the chart.

By structuring the prompts in this way, ChartCoT guides the model through a logical progression of understanding, from basic chart recognition to detailed data analysis. This approach helps in generating more accurate and contextually relevant responses.

Figure 26 shows examples of ChartCoT prompts for the four question types mentioned above. These examples illustrate how the model, guided by a chain of thought, addresses Fill-in-the-Blank, Error Correction, Multiple-Choice, and Yes-or-No questions effectively.

F.6 Chain-of-Charts Prompts

Chain-of-Charts is a new textual prompt we have developed based on the ChartCoT method in Appendix F.5. In our experiments, we observed that due to hallucinations in MLLMs, merely setting up guiding questions does not always ensure that the model correctly grasps the chart information.

To address this issue, Chain-of-Charts also includes the answers to each guiding question, providing the model with immediate feedback and reinforcement. By inputting both the guiding questions and their answers, as shown in Figure 27, Chain-of-Charts aims to mitigate the risk of hallucinations and improve the model's comprehension and accuracy. This approach helps the model build a more reliable understanding of the chart data, as it can cross-reference its responses with the provided answers.

The structured processes of Chain-of-Charts are:

- *Guiding Questions*: Similar to ChartCoT, we begin with a series of progressively detailed guiding questions, covering chart type identification, coordinate information, and numerical information.
- *Provided Answers*: For each guiding question, we input the corresponding answer. This step ensures that the model receives immediate clarification and can adjust its understanding based on accurate information.
- *Enhanced Responses*: By continuously referencing the answers to the guiding questions, the model can generate more accurate and contextually relevant responses for the final task.