ChartMind: A Comprehensive Benchmark for Complex Real-world Multimodal Chart Question Answering

Anonymous ACL submission

Abstract

Chart question answering (CQA) has become a critical multimodal task for evaluating the reasoning capabilities of vision-language models. While early approaches have shown promising performance by focusing on visual features or leveraging large-scale pre-training, most existing evaluations rely on rigid output formats and objective metrics, thus ignoring the complex, real-world demands of practical chart analysis. In this paper, we introduce ChartMind, a new benchmark designed for complex CQA tasks in real-world settings. ChartMind covers seven task categories, incorporates multilingual contexts, supports open-domain textual outputs, and accommodates diverse chart formats, bridging the gap between real-world applications and traditional academic benchmarks. Furthermore, we propose a context-aware yet model-agnostic framework, ChartLLM, that focuses on extracting key contextual elements, reducing noise, and enhancing the reasoning accuracy of multimodal large language models. Extensive evaluations on ChartMind and three representative public benchmarks with 14 mainstream multimodal models show our framework significantly outperforms the previous three common CQA paradigms: instruction-following, OCR-enhanced, and chain-of-thought, highlighting the importance of flexible chart understanding for real-world CQA. These findings suggest new directions for developing more robust chart reasoning in future research.

1 Introduction

Chart question answering (Ma et al., 2024; Qin et al., 2022) is a prominent multimodal task designed to evaluate the reasoning capabilities of vision-language models, especially their multimodal perception ability and local reasoning ability. Early studies treat CQA as a discriminative task, focusing on directly modeling visual elements to answer questions (Kafle et al., 2018; Chang et al., 2022). However, these methods often struggle with generalization due to their inability to capture the semantic and visual richness of charts. Hence, researchers introduce more visual semantic information (e.g., OCR) to enhance the multimodal perception ability (Liu et al., 2023; Wang et al., 2023a). Recent studies have shown the potential of multimodal large language models (LLMs) on the CQA task by adopting large-scale multimodal pre-training (Kim et al., 2022; Lee et al., 2023) or chain-of-thought (COT) reasoning (Li et al., 2024b; Wei et al., 2024), suggesting that leveraging large-scale datasets and supervised fine-tuning improves the interpretation of multimodal charts.

Several benchmarks (Zaib et al., 2022; Bajić and Job, 2023; Huang et al., 2024) have been proposed to better understand the strengths and weaknesses of multi-modal LLMs for CQA. However, human evaluations often suffer from high variability and instability due to individual and cultural differences, leading many existing benchmarks (Kafle et al., 2018; Mahinpei et al., 2022) to rely predominantly on automatic metrics (*e.g.*, F1 scores). While such approaches effectively evaluate the accuracy of a single answer (*e.g.*, "*2024*" for "*What is the largest value in column X?*"), they do not fully capture the need for complex and multi-step reasoning commonly required in real-world scenarios. Many professional data analysis tasks demand advanced inference, such as multi-hop reasoning or synthesizing information from multiple charts. Consequently, most existing benchmarks have widely ignored the logical steps involved in such inferencing, focusing instead on whether the answer includes the correct keyword or value.

In addition, as shown in Figure 1, we summarize three main challenges in existing benchmarks: multilingual charts, diverse formats, and questions lacking a single definitive answer, such as chart summarization. Models need to handle both visual comprehension and logical reasoning. To extract meaningful information, they must first recognize

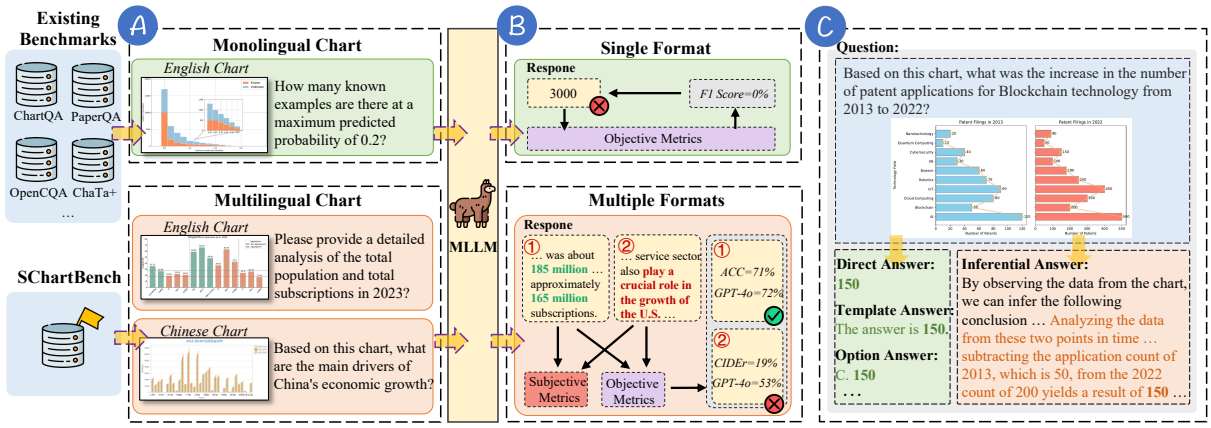

Figure 1: Key Challenges in CQA Benchmarks: (A) Predominantly monolingual, limiting multilingual applicability in chart question answering; (B) Fixed formats and metrics, restricting adaptability to diverse charts; (C) Emphasis on deterministic answers, overlooking complex reasoning, such as trend analysis, and summarization.

visual elements, such as colors, structures, and spatial relationships. Then, they must analyze the logical connections between elements and answer complex queries, such as performing calculations, identifying trends, and finding relationships within the data. Moreover, the wide range of real-world chart types (*e.g.*, bar charts, line charts, scatter plots) creates higher demands for models to generalize and perform well on new and unseen formats.

To address the above challenges, in this paper, we comprehensively review current CQA studies and introduce a new benchmark, ChartMind, specifically designed to evaluate complex CQA tasks in real-world settings. ChartMind covers seven task categories: Chart Conversion, Chart OCR Recognition, Suggestions, Chart Classification Analysis, Chart Summarization, Chart Assistance, and Information Positioning. Unlike existing benchmarks, ChartMind bridges the gap between the industrial scenarios and the academic benchmark, including the multilingual context, emphasizing open-domain textual outputs and enabling a broader evaluation across diverse formats. We further propose a novel context-aware chart understanding framework, ChartLLM, which is model-agnostic and can be applied to any MLLM to selectively extract key contextual elements, thus reducing noise and improving generalization.

To validate our benchmark, we conduct a comprehensive study of 14 mainstream multimodal models, comparing ChartLLM-based approaches with three widely used CQA paradigms: (1) instruction-following methods driven by predefined prompts, (2) OCR-enhanced methods that prioritize text extraction, and (3) COT-based methods emphasizing step-by-step reasoning.

Our contributions are as follows: (1) We introduce **ChartMind**, the first benchmark for complex CQA tasks in real-world settings. Covering seven task categories, multilingual contexts, and diverse chart formats, it bridges the gap between real-world applications and traditional academic benchmarks. (2) We propose **ChartLLM**, a context-aware yet model-agnostic framework that focuses on extracting key contextual elements, reducing noise, and enhancing the reasoning accuracy of MLLMs. (3) Through experiments across seven task categories, two languages, and seven chart formats, we show that ChartLLM outperforms prevalent CQA paradigms. These findings highlight the need for flexible chart understanding and foster advanced research on real-world chart analysis.

2 Related Work

CQA Methods. The development of CQA methods (Zeng et al., 2024; Li et al., 2024b; Xu et al., 2023) has evolved from early discriminative approaches to structured reasoning and large-scale pretraining (Zhou et al., 2023; Li et al., 2023; Huang et al., 2024; Tan et al., 2024). Early models like IMG+QUESS (Kafle et al., 2018) and V-MODEQA (Chang et al., 2022) use CNNs for visual encoding and RNNs for query processing but struggle with generalization due to limited reasoning and OOV issues. OCR-enhanced methods (Liu et al., 2023; Wang et al., 2023a) convert chart data into structured text to facilitate numerical and textual extraction but often fail to capture spatial relationships and visual attributes, making them sensitive to noise. COT-based models such as MATCHA (Li et al., 2024b) and mChartQA (Wei et al., 2024) improve accuracy by stepwise de-

Dataset	Avg. Ans. Length	Instances Number	Language Format	Diverse Format	Task Format	Topic Format	Chart Format	Pie	Scatter	Common Bar	Grouped Bar	Stacked Bar	Complex Line	Common Line
ChartQA (Masry et al., 2022)	1.15	2,500	English	1	1	3	3	✓	✗	✓	✗	✗	✗	✓
MMC-Benchmark (Liu et al., 2024a)	1.08	2,126	English	1	4	5	2	✗	✓	✗	✗	✗	✗	✓
PaperQA (Lu et al., 2023)	1.26	107	English	1	1	2	4	✓	✓	✓	✗	✗	✗	✓
OpenCQA (Kantharaj et al., 2022a)	55.73	1,159	English	1	1	4	4	✓	✓	✓	✗	✗	✗	✓
Chart-to-Text (Kantharaj et al., 2022b)	73.49	3,474	English	1	1	3	4	✓	✓	✓	✗	✗	✗	✓
LineCap (Mahinpei et al., 2022)	13.63	1,930	English	1	1	1	2	✗	✗	✗	✗	✗	✓	✓
ChartMind	**119.69**	**757**	**EN&ZH**	**2**	**7**	**6**	**7**	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of ChartMind with Existing Chart QA Datasets.

composition but rely on structured input, limiting adaptability to diverse chart formats. Other multimodal models like Donut (Kim et al., 2022) and Pix2Struct (Lee et al., 2023) eliminate OCR dependency, enabling end-to-end pretraining. Instruction-following models, such as Qwen-VL (Bai et al., 2023) and GPT-4o (Achiam et al., 2023), leverage large-scale multimodal pretraining for CQA reasoning but still struggle with multilingual charts, format diversity, and complex real-world scenarios.

CQA Benchmarks. The development of CQA models necessitates reliable benchmarks to evaluate performance across diverse tasks (Zaib et al., 2022; Bajić and Job, 2023). Existing datasets fall into Factoid Question Answering (FQA), Open-Domain Question Answering (OQA), and Captioning (CAP) categories (Huang et al., 2024). FQA datasets, such as ChartQA (Kafle et al., 2018), MMC-Bench (Liu et al., 2024a), and PaperQA (Lu et al., 2023), assess factual queries, including numerical extractions, trend identification, and relational interpretations, relying on predefined chart types for objective reasoning. OQA datasets like OpenCQA (Kantharaj et al., 2022a) introduce open-ended questions but enforce rigid output structures and rely on automated metrics like BLEU, limiting adaptability to complex reasoning. CAP datasets, including Chart-to-Text (Kantharaj et al., 2022b) and LineCap (Mahinpei et al., 2022), generate textual chart descriptions but remain constrained by structured evaluation metrics. Despite advancements, most benchmarks are monolingual, rigid, and fact-centric, restricting applicability to real-world scenarios requiring flexible reasoning, open-ended outputs, and diverse chart formats. Table 1 compares representative CQA benchmarks.

3 Construction of ChartMind

3.1 Data Processing

The dataset is curated from multiple open sources, including public datasets, GitHub repositories, and Overleaf academic projects, following OpenCQA (Kantharaj et al., 2022a) and ChartQA (Masry et al., 2022). All sources comply with open licenses (e.g., CC BY 4.0, MIT) for legal accessibility. The dataset comprises over 1,200 charts in English and Chinese, spanning various formats (e.g., Pie, Stacked Bar, etc.). To maintain quality, a manual filtering process removes charts with missing labels, unreadable fonts, or poor clarity, ensuring relevance for complex reasoning tasks. This process underpins Stage I of Figure 2.

3.2 Generated Questions & Answers

Based on the collected charts, we use GPT-4o (Achiam et al., 2023) to generate question-answer pairs across seven task categories. Few-shot prompting ensures diverse, context-aware question-answer pairs for these categories (e.g., suggestions, trend summaries, etc.). For each chart, GPT-4o generates task-specific questions and corresponding answers aligned with chart context and task objectives, such as converting visual data into structured formats or providing recommendations. All QA pairs undergo initial review for logical consistency, coherence, and factual accuracy, forming a high-quality dataset, as shown in Stage II of Figure 2.

3.3 Human Check

To ensure dataset reliability, three expert annotators with over two years of CQA research experience manually review all triplets for accuracy, clarity, and contextual relevance. Questions must align with chart content, and answers must be logically consistent and complete. Triplets with ambiguity, factual errors, or misalignment are flagged for revision or removal. This ensures a high-quality dataset for evaluating complex CQA tasks, as illustrated in Stage III of Figure 2.

3.4 Quality Assurance

To enhance ChartMind's reliability and consistency, we apply additional quality assurance mechanisms beyond standard human review. Inspired by TableBench (Wu et al., 2024) and ArXivQA (Li et al., 2024a), we refine annotation processes and align human and model evaluations. Annotators continuously provide feedback to improve task def-

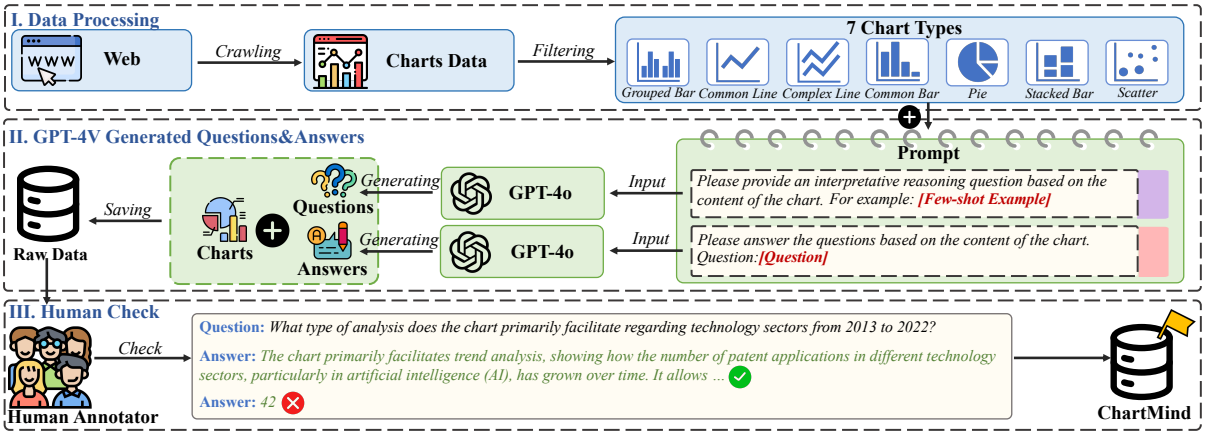

Figure 2: Data Construction Pipeline for the ChartMind.

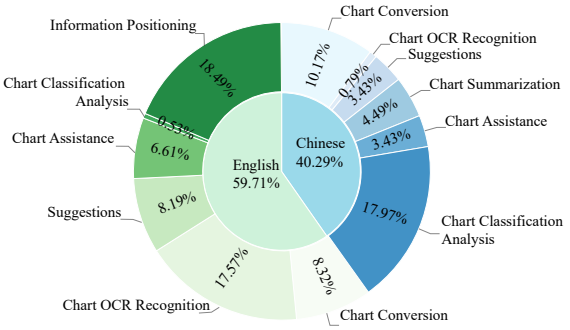

Figure 3: Language and task distribution in ChartMind.

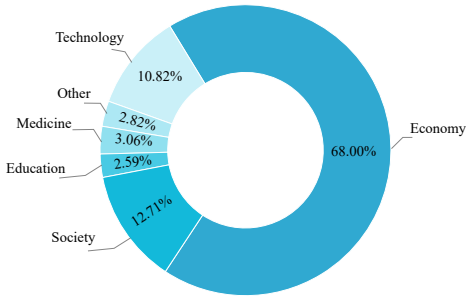

Figure 4: Topic distribution in ChartMind.

initions and guidelines, ensuring clarity and consistency. We further validate answers by retaining only triplets where model-generated answers align with human evaluations, removing ambiguous or inconsistent data. This ensures that the dataset accurately reflects logical reasoning.

3.5 Data Statistics

ChartMind evaluates models on complex reasoning abilities in CQA. Key characteristics are summarized below.

Language and Topic Distribution Figure 3 shows the dataset's language distribution: 59.71% English and 40.29% Chinese, ensuring balanced bilingual evaluation. This enables the assessment of bilingual reasoning consistency. Figure 4 illustrates topic diversity, with Economy comprising 68.00%, followed by Education and Technology, ensuring broad real-world applicability.

Task	Samples	Query Length (Min / Max)	Answer Length (Min / Max)
Chart Conversion	140	11 / 477	5 / 55
Chart OCR Recognition	139	13 / 351	8 / 59
Suggestions	88	17 / 492	13 / 53
Chart Classification Analysis	37	360 / 503	72 / 79
chart Summarization	34	76 / 335	12 / 113
Chart Assistance	76	9 / 276	12 / 41
Information Positioning	140	11 / 208	11 / 35
Total	**757**	**9 / 503**	**5 / 113**

Table 2: Task Type Statistics in ChartMind.

Task and Answer Variability Table 2 provides detailed statistics on task distribution and answer complexity. Tasks vary in difficulty, with *Chart Classification Analysis* requiring the longest queries (up to 503 tokens), while *Chart Assistance* involves the shortest queries (9–276 tokens). Answer length also varies significantly, ranging from 5 to 113 tokens, with *Chart Summarization* generating the longest responses. This highlights the benchmark's complexity, requiring models to handle both concise and complex reasoning.

4 ChartLLM

4.1 Problem Definition

CQA is a task that involves providing an answer A to a natural language question Q, based on the information contained in a chart C. The answer A may take various forms, depending on the type of question. Specifically, A could be a numerical value, a categorical label, an entity set, or an open-domain sentence. These different answer types require distinct reasoning capabilities, ranging from retrieval-based reasoning (e.g., extracting numeri-

cal values) to analytical reasoning (e.g., identifying patterns and trends in the chart). Formally, the answer A is represented as a collection of values or entities $\{a_1, a_2, \ldots, a_k\}$, where $k \in \mathbb{N}^+$.

4.2 Reasoning Methods

Instruction-following (Wei et al., 2021) and In-context learning (Dong et al., 2024) refer to strategies that optimize input for LLMs to generate practical outputs based on task-specific instructions and context. These methods enable models to leverage the provided task instructions to guide reasoning and output generation. To fully assess the reasoning capabilities of LLMs for CQA, we propose three distinct reasoning methods that aim to evaluate the model's reasoning performance.

Instruction-following-based methods Such methods (Wei et al., 2021) leverage task-specific instructions to guide LLMs in reasoning tasks. The model utilizes a prompt to interpret chart data and generate answers. The prompt P provides additional contextual guidance for the natural language question Q, specifying how the model should reason over the chart data. The reasoning process can be expressed as:

$$M(C, Q, P) \rightarrow A \tag{1}$$

where M represents the model, C is the chart, Q is the natural language question, P is the instruction prompt, and A is the answer. This approach can be applied in both fine-tuning and zero-shot settings, allowing the model to adapt to tasks based on the provided instructions.

OCR-enhanced methods OCR-enhanced methods (Liu et al., 2023) augment reasoning by incorporating textual content extracted from charts using OCR tools. These tools provide the model with additional information embedded in the chart, which may not be directly accessible through its visual content. The reasoning process is formulated as:

$$M(C, Q, O(C)) \rightarrow A \tag{2}$$

where $O(C)$ denotes the OCR-extracted content from the chart C. OCR tools offer essential support in understanding chart-based queries by enhancing the model's input with relevant textual data.

COT-based methods COT-based methods (Wei et al., 2022) break down the reasoning process into intermediate steps to improve both the accuracy and interpretability of the model's responses. This approach decomposes the reasoning into a sequence of logical steps, which enhances the model's ability to solve complex tasks. The process is represented as:

$$M(C, Q) \rightarrow \{r_1, r_2, \ldots, r_k\} \rightarrow A \tag{3}$$

where $r_1, r_2, \ldots, r_k$ represent intermediate reasoning steps, and A is the final answer. CoT is particularly useful for tasks requiring step-by-step reasoning, such as analyzing trends, identifying patterns, or extracting structured insights from complex chart data.

4.3 ChartLLM: Context Extraction for CQA

The ChartLLM is designed to enhance CQA by extracting and structuring relevant contextual information from a chart. Given a chart C, the context $C_{\text{context}} = \{T, L, X, Y\}$, where T is the title, L is the legend, X is the X-axis label, and Y is the Y-axis label, is generated to represent the essential elements of the chart. This approach minimizes irrelevant data and focuses solely on the components required for accurate reasoning in CQA tasks. To extract C_{context}, predefined prompts, such as "Extract key information from the chart, including title, legend, and X and Y-axis information," guide the model in identifying the necessary elements of the chart. This ensures the extracted context is concise, relevant, and foundational for reasoning.

The reasoning objective for ChartLLM is to predict the answer A that maximizes the conditional probability given the question Q and the extracted context C_{context}. This can be expressed as:

$$A = \text{argmax}_{a \in \mathcal{A}} \sum_{i=1}^{n} \mathbb{E}_{C_{\text{context}}, Q} \left[\log P(a_i \mid C_{\text{context}}, Q; \Theta)\right] \tag{4}$$

Here, A is the predicted answer, $\mathcal{A}$ represents the candidate answer space, C_{context} is the extracted context from the chart C, Q is the natural language question, a_i is the i-th candidate answer, and Θ denotes the model parameters.

5 Experiments

5.1 Experimental Setup

We evaluate four paradigms for CQA tasks, including instruction-following, COT-based reasoning, OCR-enhanced methods, and our proposed ChartLLM framework. These methods are tested on 14 MLLMs from three categories: specialized CQA models, general-purpose open-source models, and general-purpose closed-source models. The

Models	Size	ChartMind			ChartQA			Chart-to-Text			OpenCQA
		ACC	Avg.CIDEr	Avg.GPT-4o Score	Aug. ACC	Hum. ACC	Avg. ACC	Pew. BLEU	Statista. BLEU	Avg. BLEU	Avg. BLEU
Instruction-Following-Based (Wei et al., 2021)											
TinyChart† (Zhang et al., 2024a)	3B	5.36	18.45	16.81	93.60	72.16	82.88	10.84	27.04	18.94	19.62
ChartInstruct† (Masry et al., 2024)	7B	9.82	24.55	15.05	82.40	40.64	61.52	12.81	39.39	26.10	14.78
ChartLlama† (Han et al., 2023)	7B	20.54	21.34	12.72	90.36	48.96	69.66	14.23	40.71	27.47	4.70
Sphinx-v2 (Lin et al., 2023)	7B	9.82	25.95	13.69	60.96	43.92	52.44	3.43	4.94	4.19	3.10
LLaVA1.5 (Liu et al., 2024c)	7B	34.82	39.50	15.58	20.12	25.20	22.66	15.70	11.07	13.39	15.17
ViP-LLaVA (Cai et al., 2024)	7B	20.54	37.01	15.56	17.60	26.16	21.88	1.36	2.59	1.98	15.04
LLaVA-NEXT (Liu et al., 2024b)	7B	20.54	47.37	31.09	74.26	46.30	60.28	13.85	6.63	10.24	8.07
IXC-2.5 (Zhang et al., 2024b)	7B	47.30	40.10	43.31	92.40	74.32	83.36	17.69	11.86	14.78	9.39
Qwen2-VL (Bai et al., 2023)	7B	57.14	37.32	47.89	94.10	72.00	83.05	11.07	22.98	17.03	8.26
mPLUG-Owl2 (Ye et al., 2024)	8B	25.00	36.17	14.22	24.13	27.34	25.74	12.83	5.97	9.40	5.34
MiniCPM-v2 (Hu et al., 2024)	8B	22.32	28.48	10.63	91.12	69.02	80.07	22.17	11.01	16.59	20.05
CogVLM (Wang et al., 2023b)	17B	23.21	40.20	29.35	23.95	39.53	31.74	16.38	11.84	14.11	1.75
GLM-4V-plus (GLM et al., 2024)	-	59.83	38.36	21.52	16.80	12.80	14.80	5.69	5.71	5.70	7.41
GPT-4o (Achiam et al., 2023)	-	**61.89**	47.25	68.81	95.34	76.06	85.70	17.75	8.70	13.23	13.92
OCR-Enhanced (Liu et al., 2023)											
TinyChart† (Zhang et al., 2024a)	3B	6.71 (+1.35)	13.91 (-4.54)	17.91 (+1.10)	94.86 (+1.26)	73.95 (+1.79)	84.41 (+1.53)	13.85 (+3.01)	28.27 (+1.23)	21.06 (+2.12)	20.15 (+0.53)
ChartInstruct† (Masry et al., 2024)	7B	10.01 (+0.19)	32.80 (+8.25)	23.42 (+8.37)	83.74 (+1.34)	42.17 (+1.53)	62.96 (+1.44)	14.95 (+2.14)	40.83 (+1.44)	27.89 (+1.79)	16.01 (+1.23)
ChartLlama† (Han et al., 2023)	7B	22.03 (+1.49)	21.07 (-0.27)	26.70 (+13.97)	90.85 (+0.49)	49.26 (+0.30)	70.06 (+0.40)	16.02 (+1.79)	39.97 (-0.74)	28.00 (+0.53)	5.89 (+1.19)
Sphinx-v2 (Lin et al., 2023)	7B	11.54 (+1.72)	24.14 (-1.81)	17.21 (+3.52)	64.08 (+3.12)	45.49 (+1.57)	54.79 (+2.35)	8.81 (+5.38)	2.39 (-2.55)	5.60 (+1.41)	3.16 (+0.06)
LLaVA1.5 (Liu et al., 2024c)	7B	36.15 (+1.33)	33.49 (-6.01)	21.03 (+5.45)	19.73 (-0.39)	25.95 (+0.75)	22.84 (+0.18)	15.94 (+0.24)	12.67 (+1.60)	14.30 (+0.91)	16.31 (+1.14)
ViP-LLaVA (Cai et al., 2024)	7B	25.38 (+4.84)	36.77 (-0.24)	26.45 (+10.89)	27.12 (+9.52)	24.94 (-1.22)	26.03 (+4.15)	14.13 (+12.77)	14.37 (+11.78)	14.25 (+12.27)	18.08 (+3.04)
LLaVA-NEXT (Liu et al., 2024b)	7B	41.15 (+20.61)	47.83 (+0.46)	31.51 (+0.42)	70.47 (-3.79)	52.68 (+6.38)	61.58 (+1.30)	15.16 (+1.31)	8.82 (+2.19)	11.99 (+1.75)	8.25 (+0.18)
IXC-2.5 (Zhang et al., 2024b)	7B	42.31 (-4.99)	40.35 (+0.24)	45.38 (+2.06)	94.23 (+1.83)	73.40 (-0.92)	83.82 (+0.46)	17.03 (-0.66)	12.34 (+0.48)	14.68 (-0.10)	14.53 (+5.14)
Qwen2-VL (Bai et al., 2023)	7B	42.31 (-14.83)	36.04 (-1.27)	49.28 (+1.39)	94.23 (+0.13)	75.96 (+3.96)	85.10 (+2.05)	11.08 (+0.01)	23.21 (+0.23)	17.15 (+0.12)	11.75 (+3.49)
mPLUG-Owl2 (Ye et al., 2024)	8B	27.62 (+2.62)	30.60 (-5.57)	24.67 (+10.44)	35.58 (+11.45)	37.18 (+9.84)	36.38 (+10.65)	11.82 (-1.01)	7.30 (+1.33)	9.56 (+0.16)	4.45 (-0.89)
MiniCPM-v2 (Hu et al., 2024)	8B	23.04 (+0.72	19.73 (-8.75)	18.10 (+7.47)	92.36 (+1.24)	73.21 (+4.19)	82.79 (+2.72)	20.93 (-1.24)	5.75 (-5.26)	13.34 (-3.25)	20.60 (+0.55)
CogVLM (Wang et al., 2023b)	17B	25.54 (+2.33)	39.00 (-1.20)	36.80 (+7.45)	29.81 (+5.86)	48.72 (+9.19)	39.27 (+7.53)	20.85 (+4.47)	13.88 (+2.04)	17.37 (+3.26)	1.79 (+0.04)
GLM-4V-plus (GLM et al., 2024)	-	44.64 (-15.19)	44.83 (+6.47)	35.79 (+14.27)	17.95 (+1.15)	16.87 (+4.07)	17.41 (+2.61)	7.91 (+2.22)	7.63 (+1.92)	7.77 (+2.07)	8.72 (+1.31)
GPT-4o (Achiam et al., 2023)	-	49.31 (-12.58)	46.48 (-0.76)	71.79 (+2.98)	96.20 (+0.86)	78.04 (+1.98)	87.12 (+1.42)	20.13 (+2.38)	9.86 (+1.16)	15.00 (+1.77)	14.85 (+0.93)
COT-Based (Wei et al., 2022)											
TinyChart† (Zhang et al., 2024a)	3B	6.01 (+0.65)	13.58 (-4.87)	19.30 (+2.49)	94.84 (+1.24)	74.46 (+2.30)	84.65 (+1.77)	12.31 (+1.47)	28.53 (+1.49)	20.42 (+1.48)	20.74 (+1.12)
ChartInstruct† (Masry et al., 2024)	7B	9.96 (+0.14)	31.95 (+7.40)	22.44 (+7.39)	83.35 (+0.95)	42.74 (+2.10)	63.05 (+1.53)	14.34 (+1.53)	41.32 (+1.93)	27.83 (+1.73)	15.25 (+0.47)
ChartLlama† (Han et al., 2023)	7B	21.44 (+0.90)	18.99 (-2.36)	21.77 (+9.04)	91.63 (+1.27)	50.04 (+1.08)	70.84 (+1.18)	15.76 (+1.53)	41.42 (+0.71)	28.59 (+1.12)	6.32 (+1.62)
Sphinx-v2 (Lin et al., 2023)	7B	9.91 (+0.09)	25.03 (-0.92)	16.26 (+2.57)	61.86 (+0.90)	46.79 (+2.87)	54.33 (+1.89)	3.53 (+0.10)	5.09 (+0.15)	4.31 (+0.12)	3.13 (+0.03)
LLaVA1.5 (Liu et al., 2024c)	7B	35.77 (+0.95)	35.61 (-3.89)	19.68 (+4.10)	16.90 (-3.22)	28.57 (+3.37)	22.74 (+0.08)	15.20 (-0.50)	11.66 (+0.59)	13.43 (+0.04)	15.93 (+0.76)
ViP-LLaVA (Cai et al., 2024)	7B	23.31 (+2.77)	36.13 (-0.88)	22.24 (+6.68)	22.12 (+4.52)	28.21 (+2.05)	25.17 (+3.29)	15.48 (+14.12)	12.20 (+9.61)	13.84 (+11.86)	15.67 (+0.63)
LLaVA-NEXT (Liu et al., 2024b)	7B	40.23 (+19.69)	47.44 (+0.07)	27.34 (-3.75)	68.49 (-5.77)	52.13 (+5.83)	60.31 (+0.03)	14.81 (+0.96)	6.29 (-0.34)	10.55 (+0.31)	8.09 (+0.02)
IXC-2.5 (Zhang et al., 2024b)	7B	41.15 (-6.15)	41.23 (+1.13)	46.73 (+3.42)	93.91 (+1.51)	72.82 (-1.50)	83.37 (+0.01)	17.36 (+0.23)	11.92 (+0.06)	14.64 (-0.14)	14.39 (+5.00)
Qwen2-VL (Bai et al., 2023)	7B	40.69 (-16.45)	44.72 (+7.41)	55.12 (+7.24)	94.87 (+0.77)	77.88 (+5.88)	86.38 (+3.33)	16.70 (+5.63)	23.91 (+0.93)	20.30 (+3.27)	10.32 (+2.06)
mPLUG-Owl2 (Ye et al., 2024)	8B	25.89 (+0.89)	35.10 (-1.08)	21.27 (+7.04)	27.56 (+3.43)	31.09 (+3.75)	29.33 (+3.59)	14.00 (+1.17)	7.84 (+1.87)	10.92 (+1.52)	7.88 (+2.54)
MiniCPM-v2 (Hu et al., 2024)	8B	22.78 (+0.46)	28.81 (+0.33)	18.18 (+7.54)	92.37 (+1.25)	71.47 (+2.45)	81.92 (+1.85)	26.56 (+4.39)	12.53 (+1.52)	19.54 (+2.95)	20.30 (+0.25)
CogVLM (Wang et al., 2023b)	17B	24.01 (+0.80)	40.04 (-0.16)	37.14 (+7.79)	27.31 (+3.36)	44.93 (+5.40)	36.12 (+4.38)	17.94 (+1.56)	12.57 (+0.73)	15.26 (+1.15)	3.41 (+1.66)
GLM-4V-plus (GLM et al., 2024)	-	41.00 (-18.83)	39.55 (+1.19)	21.68 (+0.16)	18.63 (+1.83)	15.96 (+3.16)	17.30 (+2.50)	6.86 (+1.17)	7.72 (+2.01)	7.29 (+1.59)	8.83 (+1.42)
GPT-4o (Achiam et al., 2023)	-	46.15 (-15.74)	48.19 (+0.95)	69.00 (+0.19)	95.39 (+0.05)	77.23 (+1.17)	86.31 (+0.61)	19.20 (+1.45)	9.31 (+0.61)	14.26 (+1.03)	15.42 (+1.50)
ChartLLM-Based											
TinyChart† (Zhang et al., 2024a)	3B	7.69 (+2.33)	20.07 (+1.62)	23.21 (+6.40)	95.04 (+1.44)	74.41 (+2.25)	84.73 (+1.85)	14.68 (+3.84)	34.22 (+7.18)	24.45 (+5.51)	**21.84** (+2.22)
ChartInstruct† (Masry et al., 2024)	7B	11.54 (+1.72)	34.79 (+10.24)	26.43 (+11.39)	85.93 (+3.53)	43.52 (+2.88)	64.73 (+3.20)	15.52 (+2.71)	41.42 (+2.03)	28.47 (+2.37)	18.53 (+3.75)
ChartLlama† (Han et al., 2023)	7B	22.67 (+2.13)	22.54 (+1.19)	27.58 (+14.85)	91.42 (+1.06)	51.72 (+2.76)	71.57 (+1.91)	17.94 (+3.71)	**40.47** (-0.24)	**29.21** (+1.74)	7.40 (+2.70)
Sphinx-v2 (Lin et al., 2023)	7B	13.85 (+4.03)	30.11 (+4.16)	23.68 (+9.99)	62.80 (+1.84)	48.00 (+4.08)	55.40 (+2.96)	7.90 (+4.47)	7.35 (+2.41)	7.63 (+3.44)	6.88 (+3.78)
LLaVA1.5 (Liu et al., 2024c)	7B	36.92 (+2.10)	38.39 (-1.11)	26.95 (+11.37)	25.44 (+5.32)	31.68 (+6.48)	28.56 (+5.90)	18.21 (+2.51)	17.83 (+6.76)	18.02 (+4.63)	17.40 (+2.23)
ViP-LLaVA (Cai et al., 2024)	7B	26.23 (+5.69)	41.98 (+4.97)	28.79 (+13.23)	23.96 (+6.36)	29.04 (+2.88)	26.50 (+4.62)	14.31 (+12.95)	14.38 (+11.79)	14.35 (+12.37)	18.72 (+3.68)
LLaVA-NEXT (Liu et al., 2024b)	7B	42.31 (+21.77)	49.40 (+2.03)	34.40 (+3.32)	75.82 (+1.56)	47.68 (+1.38)	61.75 (+1.47)	15.26 (+1.41)	8.93 (+2.30)	12.10 (+1.86)	9.02 (+0.95)
IXC-2.5 (Zhang et al., 2024b)	7B	47.31 (+0.01)	43.38 (+3.28)	51.88 (+8.56)	94.88 (+2.48)	76.24 (+1.92)	85.56 (+2.20)	19.82 (+2.13)	14.70 (+2.84)	17.26 (+2.48)	16.83 (+7.44)
Qwen2-VL (Bai et al., 2023)	7B	57.66 (+0.52)	45.54 (+8.22)	56.10 (+8.21)	94.40 (+0.30)	77.44 (+5.44)	85.92 (+2.87)	20.96 (+9.89)	24.45 (+1.47)	22.71 (+5.68)	18.53 (+10.27)
mPLUG-Owl2 (Ye et al., 2024)	8B	29.38 (+4.38)	40.46 (+4.29)	29.15 (+14.93)	38.76 (+14.63)	40.34 (+13.00)	39.55 (+13.82)	13.01 (+0.18)	8.91 (+2.94)	10.96 (+1.56)	6.26 (+0.92)
MiniCPM-v2 (Hu et al., 2024)	8B	24.21 (+1.89)	38.65 (+10.17)	23.73 (+13.09)	93.84 (+2.72)	71.86 (+2.84)	82.85 (+2.78)	**27.68** (+5.51)	24.55 (+13.54)	26.12 (+9.53)	20.88 (+0.83)
CogVLM (Wang et al., 2023b)	17B	26.38 (+3.17)	41.05 (+0.85)	41.85 (+12.50)	33.41 (+9.46)	51.73 (+12.20)	42.57 (+10.83)	21.46 (+5.08)	14.74 (+2.90)	18.10 (+3.99)	2.48 (+0.73)
GLM-4V-plus (GLM et al., 2024)	-	60.18 (+0.35)	47.00 (+8.64)	37.19 (+15.67)	19.74 (+2.94)	18.04 (+5.24)	19.66 (+4.86)	9.75 (+4.06)	8.97 (+3.26)	9.36 (+3.66)	9.74 (+2.34)
GPT-4o (Achiam et al., 2023)	-	**61.89** (+0.00)	**50.42** (+3.17)	**73.89** (+5.08)	**98.63** (+3.29)	**79.49** (+3.43)	**89.06** (+3.36)	23.65 (+5.90)	11.07 (+2.37)	17.36 (+4.14)	16.04 (+2.12)

Table 3: Performance of multimodal models on ChartMind and three structured-output CQA datasets. The best results are highlighted in **bold**, and the second-best results are underlined. †Specialized CQA models.

evaluation spans four datasets, including our proposed ChartMind and three structured-output CQA datasets—ChartQA (Masry et al., 2022), Chart-to-Text (Kantharaj et al., 2022b), and OpenCQA (Kantharaj et al., 2022a)—which primarily rely on predefined answer formats and automated scoring metrics. In contrast, ChartMind introduces diverse chart formats and open-domain textual outputs, enabling a more comprehensive assessment of real-world CQA scenarios. Further implementation details, model descriptions, and benchmark specifications are provided in Appendix B.

5.2 Main Results

To evaluate the effectiveness and robustness of ChartLLM-based methods over OCR-enhanced (Liu et al., 2023) and COT-based (Wei et al., 2022) approaches in open-ended and structured-output reasoning, Table 3 compares their performance across various benchmarks. Both OCR-enhanced and COT-based methods yield significant improvements (blue text), but

Models	Size	Avg. GPT-4o Score	Avg. Human Score
ChartInstruct (Masry et al., 2024)	7B	26.43	22.52
ChartLlama (Han et al., 2023)	7B	27.58	23.11
TinyChart (Zhang et al., 2024a)	3B	23.21	21.97
mPLUG-Owl2 (Ye et al., 2024)	8B	29.15	29.31
Sphinx-v2 (Lin et al., 2023)	7B	23.68	22.31
CogVLM (Wang et al., 2023b)	17B	41.85	34.96
LLaVA1.5 (Liu et al., 2024c)	7B	26.95	22.93
MiniCPM-v2 (Hu et al., 2024)	8B	23.73	24.01
ViP-LLaVA (Cai et al., 2024)	7B	28.79	30.75
LLaVA-NEXT (Liu et al., 2024b)	7B	34.40	32.31
IXC-2.5 (Zhang et al., 2024b)	7B	51.88	36.61
Qwen2-VL (Bai et al., 2023)	7B	56.10	40.39
GLM-4V-plus (GLM et al., 2024)	-	37.19	39.35
GPT-4o (Achiam et al., 2023)	-	73.89	50.73
PCC (Cohen et al., 2009)	-	**93.09**	

Table 4: Correlation of GPT4o and Human Eval.

their effectiveness varies by task. OCR-enhanced methods often degrade performance (red text), particularly in open-ended reasoning, where redundancy and noise from textual extraction disrupt holistic reasoning. For instance, GPT-4o's (Achiam et al., 2023) ACC in open-ended tasks drops by -12.58 with OCR-enhanced methods, reflecting their sensitivity to flexible reasoning. COT-based methods enhance structured-output reasoning but struggle in open-ended tasks, reducing GPT-4o's ACC by -15.74 due to difficulties in integrating contextual and visual elements. ChartLLM-based methods address these challenges by strategically extracting key contextual information and minimizing redundancy, reducing external noise in reasoning. By focusing on essential chart elements and preserving relevant semantic relationships, they achieve superior performance with consistent adaptability across both reasoning types. Their ability to balance context extraction and noise reduction underscores their robustness in handling complex chart reasoning.

5.3 Correlation Analysis of Metrics

To assess the consistency between automated and human evaluation in open-ended CQA, Table 4 analyzes the correlation between GPT-4o Score and Human Score across 14 multimodal models. The Pearson Correlation Coefficient (PCC) (Cohen et al., 2009) is 93.09, indicating a strong linear relationship. High-performing models like GPT-4o (Achiam et al., 2023) and Qwen2-VL (Bai et al., 2023) show strong alignment between GPT-4o and human scores, validating automated evaluation reliability. Notably, models like mPLUG-Owl2 (Ye et al., 2024) and ViP-LLaVA (Cai et al., 2024) exhibit slight deviations, where human scores marginally exceed automated ones, possibly reflecting nuanced human judgment in open-ended reasoning. The high PCC confirms GPT-4o Score as a robust proxy for human evaluation, reinforcing its applicability in open-ended CQA.

5.4 Sensitivity Analysis

Language-Level Analysis. To evaluate the sensitivity of different paradigms to multilingual challenges in CQA tasks, we analyze model performance on English and Chinese charts in ChartMind. Figure 5 reveals that models perform significantly better on English charts compared to Chinese charts, highlighting challenges in multilingual scenarios. Instruction-following methods, such as

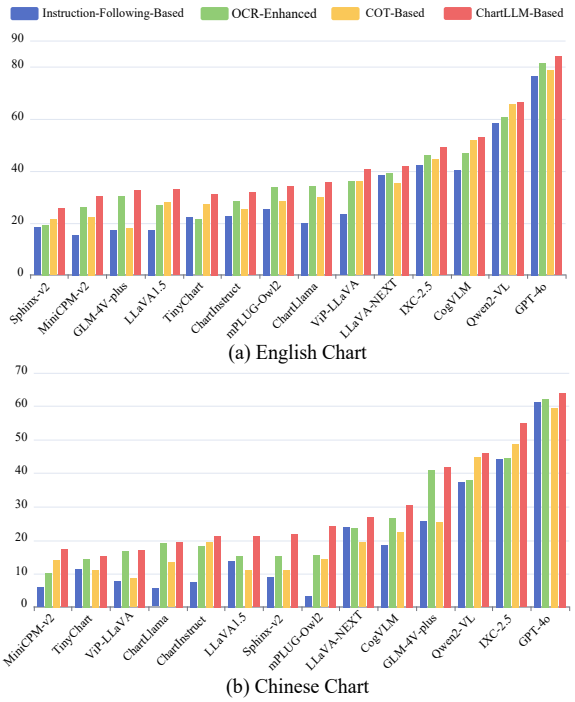

Figure 5: Performance of multimodal models across Chinese and English datasets in ChartMind.

GPT-4o (Achiam et al., 2023) and LLAVA1.5 (Liu et al., 2024c), show severe performance degradation on Chinese charts due to limited multilingual capabilities. In contrast, IXC-2.5 (Zhang et al., 2024b) and GLM-4V-plus (Wang et al., 2023b) handle Chinese data more effectively. OCR-enhanced and COT-based methods mitigate performance declines to some extent, with OCR often outperforming COT in Chinese scenarios by leveraging visual-semantic elements. ChartLLM-based methods enhance multilingual robustness by extracting context, achieving the best performance across diverse multilingual charts types.

Task-Level Analysis. In order to explore how different paradigms handle diverse CQA tasks, we evaluate model performance across seven tasks in ChartMind. As shown in Figure 6, the tasks exhibit varying levels of difficulty, with *Chart Conversion* and *Chart Summarization* being the most challenging due to their reliance on complex semantic integration and reasoning, while tasks like *Suggestions* and *Information Positioning* are relatively easier, requiring localized data extraction and straightforward analysis. Performance disparities are particularly evident in high-difficulty tasks, where instruction-following methods struggle significantly, revealing their limitations in integrating multimodal information. OCR-enhanced methods

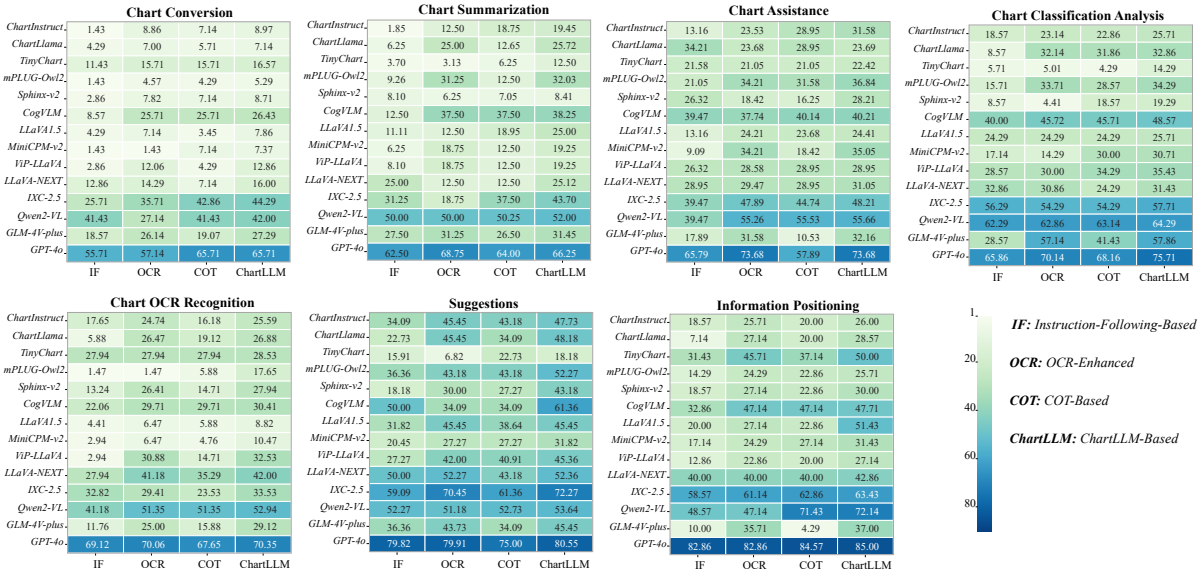

Figure 6: Performance of multimodal models on seven tasks in ChartMind.

perform well in text-heavy tasks such as *Chart OCR Recognition*, but often introduce noise in tasks requiring holistic understanding, such as *Chart Summarization*. COT-based methods improve on logical reasoning in tasks like *Suggestions*, yet falter in capturing complex dependencies in tasks such as *Chart Assistance*. In contrast, ChartLLM-based methods consistently demonstrate superior adaptability and performance across tasks, excelling in high-difficulty scenarios by effectively integrating contextual and visual features while maintaining robust results in simpler tasks.

Chart-Type-Level Analysis. We analyze the sensitivity of different paradigms to various chart types in ChartMind, examining their strengths and weaknesses. Chart types vary in complexity, with *Pie* and *Stacked Bar* requiring high-context reasoning, while *Complex Line* mainly involves direct data extraction. Instruction-following models, such as GPT-4o (Achiam et al., 2023) and LLAVA1.5 (Liu et al., 2024c), struggle with complex charts, while OCR-enhanced methods perform well on text-heavy types but fail in visual-reliant tasks. COT-based methods exhibit stable performance but lack high-context reasoning. ChartLLM achieves the highest overall performance, effectively utilizing contextual elements. A comprehensive breakdown of model performance across chart types is provided in Appendix D.

5.5 Error Analysis

Specific error types observed in the ChartMind benchmark include value recognition errors, judgment errors, calculation errors, and color recognition errors. Value recognition errors stem from inaccuracies in extracting or interpreting numerical values from charts. Judgment errors result from flawed reasoning or misinterpreting task requirements, such as misunderstanding the question's context. Calculation errors reflect weaknesses in performing arithmetic operations, while color recognition errors occur when models fail to associate chart elements with their corresponding colors, particularly in complex legends or bar segments. A detailed breakdown of these errors is shown in Appendix E.

6 Conclusion

We introduce *ChartMind*, the first benchmark designed to evaluate complex CQA tasks in real-world settings. ChartMind addresses critical limitations in existing benchmarks by supporting multilingual charts, diverse output formats, and seven diverse CQA tasks. Through experiments on 14 multimodal models across four paradigms, we demonstrate the effectiveness of ChartLLM, a model-agnostic framework that leverages context-aware chart understanding to significantly enhance reasoning accuracy. ChartLLM consistently outperforms OCR-enhanced and COT-based methods, setting a new standard for evaluating complex CQA in real-world scenarios. Future work will extend ChartMind to multi-turn, multi-chart dialogues, cross-chart reasoning, and mixed chart-text contexts for complex queries, further advancing multimodal chart understanding for real-world CQA.

Limitations

ChartMind provides a benchmark for complex CQA evaluation, yet several limitations remain. First, the dataset primarily relies on publicly available charts, potentially introducing biases in data distribution and task complexity. Ensuring broader representativeness requires further dataset expansion and diversification. Second, although ChartMind defines seven reasoning tasks, real-world chart analysis often involves more advanced reasoning, such as multi-turn interactions, cross-chart comparisons, and textual-visual information integration, which remain underexplored. Third, the reliance on automated evaluation methods, such as GPT-4 ratings, introduces challenges in capturing nuanced human judgment in complex reasoning. Addressing these issues requires refining evaluation methodologies and incorporating more human annotations. Future improvements may focus on expanding the dataset, enhancing evaluation metrics, and integrating multi-turn reasoning and cross-chart analysis to better reflect real-world scenarios.

References

Josh Achiam, Steven Adler, et al. 2023. Gpt-4 technical report. In *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. In *arXiv preprint arXiv:2308.12966*.

Filip Bajić and Josip Job. 2023. Review of chart image detection and classification. *IJDAR*, 26(4):453–474.

Mu Cai, Haotian Liu, and others. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *CVPR*, pages 12914–12923.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. In *arXiv preprint arXiv:2211.08545*.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, et al. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, 16(4):1–4.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A survey on in-context learning. In *EMNLP*, pages 1107–1128.

Team GLM, Aohan Zeng, Bin Xu, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. In *arXiv preprint arXiv:2406.12793*.

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. In *arXiv preprint arXiv:2311.16483*.

Shengding Hu, Yuge Tu, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. In *COLM*, pages 1–33.

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

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022a. Opencqa: Open-ended question answering with charts. In *EMNLP*, pages 11817–11837.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022b. Chart-to-text: A large-scale benchmark for chart summarization. In *ACL*, pages 4005–4023.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *ECCV*, volume 13688, pages 498–517.

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

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *arXiv preprint arXiv:2403.00231*.

Zhe Li, Xinyu Wang, Yuliang Liu, Lianwen Jin, et al. 2023. Improving handwritten mathematical expression recognition via similar symbol distinguishing. *TMM*, 26:90–102.

Zhuowan Li, Bhavan Jasani, et al. 2024b. Synthesize step-by-step: Tools templates and llms as data generators for reasoning-based chart vqa. In *CVPR*, pages 13613–13623.

Ziyi Lin, Chris Liu, Renrui Zhang, et al. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. In *arXiv preprint arXiv:2311.07575*.

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

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *NAACL*, pages 1287–1310.

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

Haotian Liu, Chunyuan Li, et al. 2024c. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *arXiv preprint arXiv:2310.02255*.

Linfeng Ma, Han Fang, Zehua Ma, Zhaoyang Jia, Weiming Zhang, and Nenghai Yu. 2024. C 3 hartmark: A chart watermarking scheme with consecutive-encoding and concurrent-decoding. *TCSVT*, 34(10):4005–4018.

Anita Mahinpei, Zona Kostic, and Chris Tanner. 2022. Linecap: Line charts for data visualization captioning models. In *2022 IEEE VIS*, pages 35–39.

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

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension and reasoning. In *arXiv preprint arXiv:2403.09028*.

Bosheng Qin, Haoji Hu, and Yueting Zhuang. 2022. Deep residual weight-sharing attention network with low-rank attention for visual question answering. *TMM*, 25:4282–4295.

Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. 2024. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In *ECCV*, pages 305–322. Springer.

Peifang Wang, Olga Golovneva, Armen Aghajanyan, Xiang Ren, Muhao Chen, Asli Celikyilmaz, and Maryam Fazel-Zarandi. 2023a. Domino: A dual-system for multi-step visual language reasoning. In *arXiv preprint arXiv:2310.02804*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. Cogvlm: Visual expert for pretrained language models. In *arXiv preprint arXiv:2311.03079*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Jingxuan Wei, Nan Xu, Guiyong Chang, Yin Luo, Bi-Hui Yu, and Ruifeng Guo. 2024. mchartqa: A universal benchmark for multimodal chart question answer based on vision-language alignment and reasoning. In *arXiv preprint arXiv:2404.01548*.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. 2024. Tablebench: A comprehensive and complex benchmark for table question answering. In *arXiv preprint arXiv:2408.09174*.

Jie Xu, Xiaoqian Zhang, Changming Zhao, et al. 2023. Improving fine-grained image classification with multimodal information. *TMM*, 25(8):2082 – 2095.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, pages 13040–13051.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *KIS*, 64(12):3151–3195.

Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. 2024. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *TVCG*, 30(11):1–11.

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

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. 2024b. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. In *arXiv preprint arXiv:2407.03320*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. In *arXiv preprint arXiv:2311.07911*.

A Chart Types and Tasks in ChartMind

ChartMind supports a diverse range of chart types and reasoning tasks, ensuring a comprehensive evaluation of complex reasoning in CQA. As shown in Figure 7 The dataset includes seven distinct chart types—Pie, Common Bar, Scatter, Grouped Bar, Complex Line, Stacked Bar, and Common Line—capturing varied visual structures and data representations. Additionally, ChartMind defines seven reasoning tasks: Chart Conversion, Chart OCR Recognition, Suggestions, Chart Assistance, Chart Classification, Chart Summarization, and Information Positioning, covering key aspects of multimodal chart understanding. These distributions illustrate ChartMind's ability to comprehensively assess complex multimodal reasoning, spanning diverse chart types and reasoning paradigms. Compared to prior benchmarks, ChartMind provides a broader evaluation scope, capturing the complexity of real-world CQA tasks.

B Experimental Setup Details

B.1 Implementation Details

To assess the performance of models on complex CQA tasks in real-world settings, we experiment with four types of paradigms. First, we test MLLMs in the instruction-following setting (Zhou et al., 2023), where we use prompts to evaluate their ability to answer chart-related questions. Second, we apply COT-based methods (Wei et al., 2022), which break down reasoning processes into intermediate steps to generate answers. Third, we adopt OCR-enhanced methods inspired by DePlot (Liu et al., 2023), which extract chart content as text and use it as input for multimodal reasoning models. Finally, we propose the ChartLLM method, which enhances reasoning performance by extracting structured contextual information, such as chart titles, legends, and axes, using Qwen2-VL (Bai et al., 2023), and feeding this information into models for further analysis.

B.2 Models

We evaluate 14 MLLMs across three categories: specialized CQA models, general-purpose open-source multimodal models, and general-purpose closed-source multimodal models. The majority of the models have a parameter size of approximately 7B, with a few exceptions, including smaller models such as TinyChart (Zhang et al., 2024a) with 3B parameters and larger models like CogVLM (Wang et al., 2023b) with 17B parameters. For specialized CQA models, we include ChartInstruct (Masry et al., 2024), ChartLlama (Han et al., 2023), and TinyChart (Zhang et al., 2024a). These models are specifically trained on CQA datasets, making them particularly suited for tasks requiring precise understanding of chart-related queries. Among open-source general-purpose multimodal models, we evaluate mPLUG-Owl2 (Ye et al., 2024), Sphinx-v2 (Lin et al., 2023), CogVLM (Wang et al., 2023b), LLaVA1.5 (Liu et al., 2024c), MiniCPM-v2 (Hu et al., 2024), ViP-LLaVA (Cai et al., 2024), LLaVA-NEXT (Liu et al., 2024b), IXC-2.5 (Zhang et al., 2024b), and Qwen2-VL (Bai et al., 2023). These models leverage extensive multimodal training datasets, including CQA data, and exhibit strong performance on chart-related tasks. Finally, closed-source general multimodal models, including GPT-4o (Achiam et al., 2023) and GLM-4V-plus (GLM et al., 2024), are state-of-the-art models with advanced multimodal reasoning capacities, providing strong competition to existing open-source systems.

B.3 Benchmarks and Metrics

To comprehensively evaluate multimodal CQA tasks, we adopt three representative structured-output reasoning datasets—ChartQA (Masry et al., 2022), Chart-to-Text (Kantharaj et al., 2022b), and OpenCQA (Kantharaj et al., 2022a)—alongside our proposed benchmark, ChartMind. ChartQA and Chart-to-Text primarily take a chart and a natural language question as input and generate structured textual answers, such as numerical values, categorical labels, or predefined captions, making them well-suited for factual extraction tasks. OpenCQA, despite allowing open-ended queries, constrains responses to structured formats evaluated by automated metrics like BLEU, limiting its ability to assess flexible reasoning. To address these constraints, ChartMind introduces a more comprehensive evaluation by supporting diverse chart types, open-ended textual outputs, and seven complex reasoning tasks, enabling a broader assessment of multimodal reasoning. Models are evaluated using Accuracy and CIDEr for structured assessments, while GPT-4o score and Human score serve as open-ended evaluation metrics, with GPT-4o score as the primary metric, as detailed in Appendix C. The structured-output datasets are evaluated using Accuracy and BLEU score.

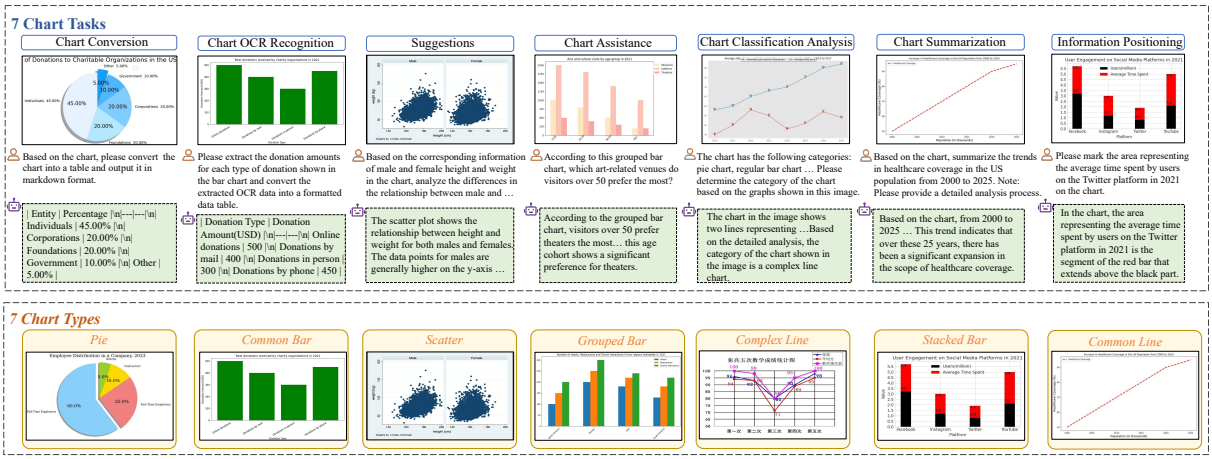

Figure 7: Overview of the seven chart types and seven reasoning tasks included in ChartMind.

C GPT-4o Scoring Prompt Design

The GPT-4o score prompt evaluates the performance of models on CQA tasks by assessing two key dimensions: output quality and output correctness. Output quality focuses on the fluency of the model's answer, the completeness of its reasoning process, and its ability to follow instructions accurately. Output correctness measures the overall accuracy of the reasoning, the correctness of the data, and the logical alignment with the human reference answer or chart content. The input to the prompt includes a JSON object containing the question, the human reference answer, and the model-generated answer. The output is also formatted as a JSON object, which includes a detailed explanation of the scoring rationale along with scores for both dimensions. The full design of the scoring prompt is visualized in Figure 8.

D Chart-Type-Level Analysis

To evaluate the sensitivity of different paradigms to diverse chart types in CQA tasks, we analyze their performance across seven chart types in ChartMind. Figure 9 presents a detailed breakdown of model performance. Chart types exhibit varying complexity, with *Pie* and *Stacked Bar* being the most challenging due to their reliance on integrated contextual reasoning, while simpler types like *Complex Line* primarily require straightforward data extraction. Instruction-following methods (Wei et al., 2021), such as GPT-4o (Achiam et al., 2023) and LLAVA1.5 (Liu et al., 2024c), show significant performance drops in high-complexity charts, underscoring their limitations in managing holistic reasoning tasks. OCR-enhanced methods (Liu et al., 2023) excel in text-heavy charts such as *Grouped Bar*, leveraging their ability to extract textual information, but struggle with tasks like *Scatter* that demand comprehensive visual-semantic integration. COT-based methods (Wei et al., 2022) demonstrate moderate performance across most chart types, performing relatively well in structured charts like *Common Line*, yet falling short in tasks requiring high-contextual reasoning. ChartLLM-based methods achieve the highest overall performance, excelling in high-difficulty charts by effectively using critical contextual elements and showcasing adaptability to diverse chart types. These results highlight the necessity of contextual reasoning for high-performance chart understanding.

E Error Analysis

Figure 10 illustrates specific examples of the four major error types observed in the ChartMind: value recognition errors, judgment errors, calculation errors, and color recognition errors. These examples highlight typical failure cases, such as incorrect identification of numerical values in bar segments (value recognition), flawed logical reasoning or mismatched context interpretation (judgment), inaccurate arithmetic operations (calculation), and misassociation of chart elements with their respective colors in legends or overlapping areas (color recognition). The figure provides detailed scenarios, such as errors in identifying peak values, interpreting differences in chart segments, and miscalculating relationships between visual elements. These cases emphasize the challenges faced by models in aligning visual interpretation with reasoning accuracy.

*You are a professional **chart-based question-answering evaluation expert**. You need to evaluate the model's performance based on **charts**, **questions**, **human reference answers**, and **model answers**. In your evaluation, please analyze the performance in two dimensions in detail:*

*1. **Output Quality (0-1 points)**: Evaluate whether the model's answer is fluent, whether the reasoning process is complete, and whether the instructions are accurately followed.*
*2. **Output Correctness (0-1 points)**: Assess whether the reasoning is correct overall, whether most of the data is accurate, and whether the model's answer aligns with the logic of the human reference answer.*

Input Format
The input is a JSON object with the following fields:

"question": "string, the question description",
"human_reference": "string, the human reference answer",
"model_answer": "string, the model's generated answer"

Scoring Criteria
*- **Output Quality Score (0-1 points)**:*
* - **0 points**: The expression is not fluent, the reasoning process is lacking, or the instructions are not followed.*
* - **1 point**: The expression is generally clear and fluent, the logic is reasonable, and it adheres to the instructions.*

*- **Output Correctness Score (0-1 points)**:*
* - **0 points**: The reasoning process is incorrect, the data is inaccurate, or the key elements such as labels, colors, etc., are not correctly identified.*
* - **1 point**: The reasoning process is generally reasonable, the key data in the model's answer is mostly consistent with the reference answer or the chart content, and it aligns with the question requirements, even if it is not 100% consistent with the human reference answer.*

Output Format
The output should be a JSON object, including a detailed analysis and score:

```json
{
* "reason": "string, please use Chinese to describe in detail the quality and correctness of the model's output, including the reasoning process and data accuracy. Specially compare the data with human reference answers and chart content, and explain the basis for the score.",*
* "quality_score": "int, the output quality score (0 or 1)",*
* "correctness_score": "int, the output correctness score (0 or 1)"*
}

Task Requirements
Based on the chart information and model responses, conduct a detailed analysis of the model's reasoning logic and data accuracy, providing specific reasons for scoring. Note that the reference answers are only examples; the model's response should be generally reasonable and consistent with the question in terms of logic and data.

Figure 8: Prompt design for GPT-4o score.

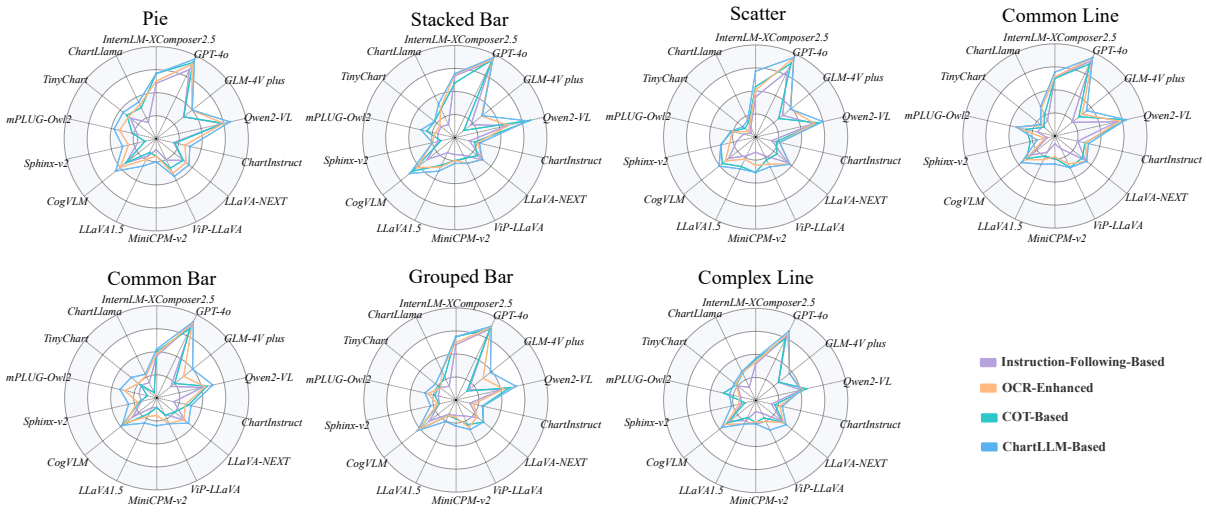

Figure 9: Performance of multimodal models across chart types, categorized by four paradigms.

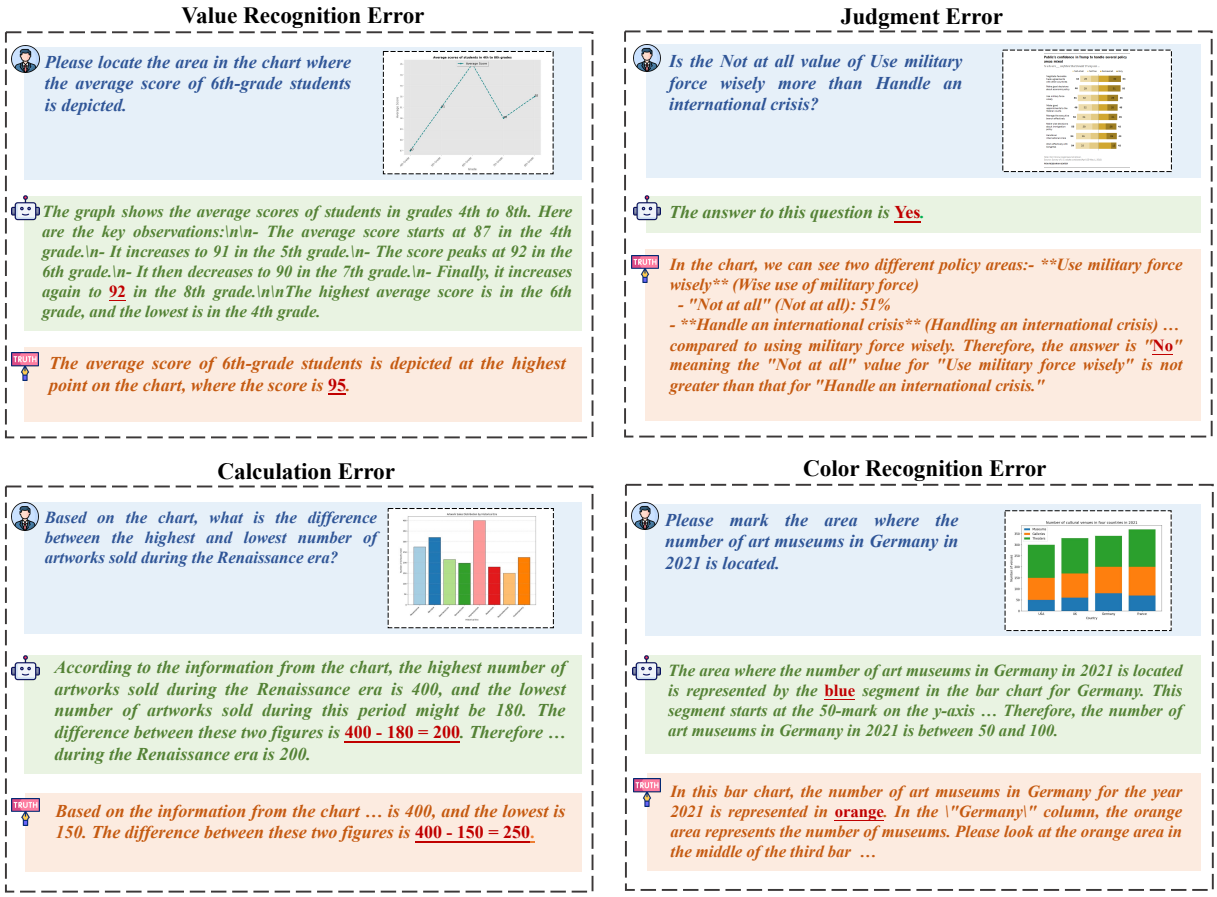

Figure 10: The four major error types in ChartMind.