MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning

Fuxiao Liu[1,2*]**, Xiaoyang Wang**[2]**, Wenlin Yao**[2]**, Jianshu Chen**[2]**,**
Kaiqiang Song[2]**, Sangwoo Cho**[2]**, Yaser Yacoob**[1]**, Dong Yu**[2]

[1]University of Maryland, College Park [2]Tencent AI Lab, Bellevue, USA

{fl3es, yaser}@umd.edu, {shawnxywang, wenlinyao, jianshuchen, riversong, swcho, dyu}@global.tencent.com

Abstract

With the rapid development of large language models (LLMs) and their integration into large multimodal models (LMMs), there has been impressive progress in zero-shot completion of user-oriented vision-language tasks. However, a gap remains in the domain of chart image understanding due to the distinct abstract components in charts. To address this, we introduce a large-scale MultiModal Chart Instruction (**MMC-Instruction**) dataset comprising 600k instances supporting diverse tasks and chart types. Leveraging this data, we develop MultiModal Chart Assistant (**MMCA**), an LMM that achieves state-of-the-art performance on existing chart QA benchmarks. Recognizing the need for a comprehensive evaluation of LMM chart understanding, we also propose a MultiModal Chart Benchmark (**MMC-Benchmark**), a comprehensive human-annotated benchmark with nine distinct tasks evaluating reasoning capabilities over charts. Extensive experiments on MMC-Benchmark reveal the limitations of existing LMMs on correctly interpreting charts, even for the most recent GPT-4V model. Our work provides an instruction-tuning methodology and benchmark to advance multimodal understanding of charts. Code and data are available at `https://github.com/FuxiaoLiu/MMC`.

1 Introduction

Large Language models (LLMs) such as GPT-3, PaLM, ChatGPT, Bard, and LLaMA (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2022; Manyika, 2023; Touvron et al., 2023) have undergone rapid development, demonstrating significant capabilities in performing a wide range of tasks effectively. To enable LLMs with vision ability, open-source large multimodal models (LMMs) such as MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023e), mPLUG-Owl (Ye et al.,

2023), Multimodal-GPT (Gong et al., 2023), and LRV (Liu et al., 2023b) have been developed, incorporating advanced image understanding capabilities into LLMs to interpret and analyze visual inputs. While successful in the general domains, such open-source LMMs are less effective for chart images because chart understanding differs tremendously from natural scene image understanding. In contrast with natural scene images, which primarily contain objects and reflect their spatial relationships, chart images contain unique abstract elements, including trend lines and color-coded legends that convey specific data-related information.

Current open-source LMMs are limited in their ability to accurately interpret complex chart contents, as they often lack domain-specific training essential for tasks such as differentiating between various types of graphs, interpreting axis labels and data points, and extracting meaningful patterns and trends. Integrating advanced chart understanding capabilities could further refine the LMMs' ability to analyze contextually and reason about the information presented in charts, thereby broadening their applicability in fields like data analytics, academic research, and business intelligence.

In this paper, we introduce MultiModal Chart Instruction (**MMC-Instruction**), a 600k chart understanding dataset consisting of both chart-text alignment data and chart instruction-tuning data. *MMC-Instruction* is not only much larger but also more diverse compared to existing public datasets (Kahou et al., 2017; Masry et al., 2022; Methani et al., 2020; Kafle et al., 2018). Unlike previous work with templated-based questions, *MMC-Instruction* is constructed by prompting GPT-4 (OpenAI, 2023a) to generate instructions with diverse language styles and tasks (Tab. 1). Furthermore, our *MMC-Instruction* considers a variety of chart types, including but not limited to histograms, scatter plots, area charts, and more complex graphical representations. By performing unified instruction tuning
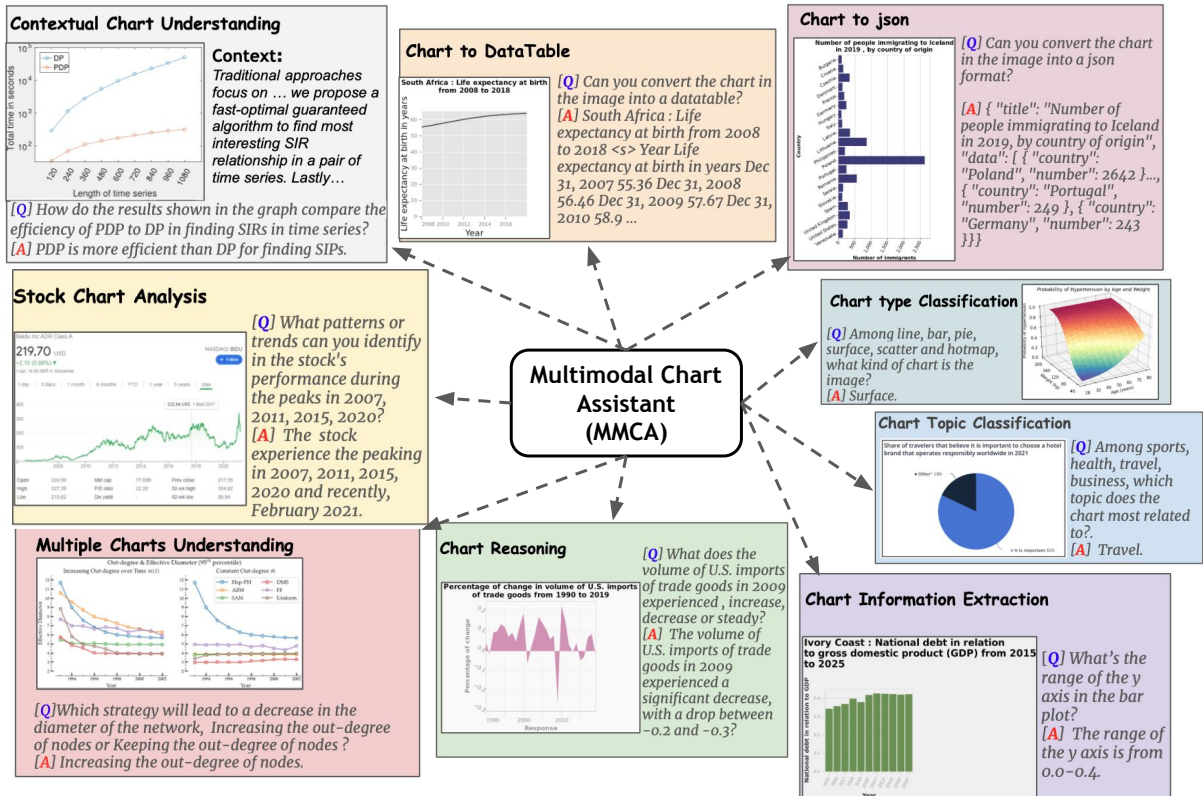

1287

Figure 1: Diagram of our human-annotated *MMC*, consisting of **nine** distinct tasks, various topics (***business, health, biology, engineering, etc***), various chart types (***bar, histograms, line, scatter, heatmap, etc***), free-form questions and open-ended answers. More examples are shown in the Appendix.

Datasets	Fig. Num	Question	Ans. Type	Ans. Length	Plot Type	Task Num	Benchmark
FigureQA	180k	Template	Fixed Vocab	1.0	4	1	✘
DVQA	300k	Template	Fixed Vocab	1.0	1	1	✘
PlotQA	224k	Template	Fixed	1.0	1	1	✘
ChartQA	21.9k	Free-form	Open Vocab	1.2	Unbounded	2	Human Check
SciGraphQA	295k	Free-form	Open Ended	-	Unbounded	2	✘
MMC-Instruction	600k	Free-form	Open Ended/MQA	23.7	Unbounded	9	Human Check

Table 1: Comparison between *MMC-Instruction* with existing **chart** question-answering datasets. *MQA* means multiple-choice question answering. *MMC-Instruction* is **larger** and **more diverse**. "Ans." stands for "Answer".

upon current LMMs with *MMC-Instruction*, we further propose a modularized LMM, namely Multimodal Chart Assistant (**MMCA**), jointly finetuned on a wide range of visually situated language understanding tasks. *MMCA* achieves state-of-the-art performance on current chart question-answer benchmarks compared with existing open-source LMMs.

To accurately assess the capabilities of current Large Multimodal Models (LMMs) for chart understanding, we introduce a novel comprehensive evaluation tool: the MultiModal Chart Benchmark (**MMC-Benchmark**). First, *MMC-Benchmark* is the first human-annotated benchmark in line with human cognition to evaluate LMM's ability to comprehend visual charts. Second, it contains a wide range of tasks, including *chart information extraction*, *chart reasoning*, *contextual chart understanding*, *chart topic classification*, *stock chart analysis*, *multiple chart understanding*, *chart type classification*, *chart-to-datatable* and *chart-to-json*. Third, *MMC-Benchmark* offers two quantitative evaluation methods, including free-format Generation Ability Evaluation using GPT-4 and multiple-choice QA format Chart Understanding Ability Evaluation without the requirement of GPT-4. Our evaluation highlights the limitations of existing open-source LMMs. In addition, we further broaden our analysis through experiments with GPT-4V (OpenAI, 2023b; Yang et al., 2023b; Liu

et al., 2023a), the latest multimodal version of GPT-4. Our experiments indicate that *MMC-Benchmark* also poses significant challenges to GPT-4V, especially in *Chart to Datatable* and *Chart to Json* tasks. It indicates the importance of *MMC-Instruction* corpus and *MMC-Benchmark* in advancing multimodal understanding.

Our main contributions are as follows:

- **MMC-Instruction** dataset. We present a novel large-scale instruction-tuning dataset for chart understanding. It includes diverse topics, language styles, chart types, and open-ended answers in line with human cognition.
- **MMC-Benchmark**. We present a manually annotated benchmark specifically designed to assess the capability of LMMs in chart understanding across nine distinct sub-tasks to ensure a comprehensive evaluation.
- **MMCA** model. We propose an instruction-tuned LMM model that outperforms existing open-source state-of-the-art LMMs for chart understanding on both existing chart understanding benchmarks and our benchmark.

2 Related Work

Multimodal Large Language Model. Recently, Large Language Models (LLMs) have shown strong performances in zero-shot tasks across multiple domains. Recent studies explore using LLMs for multi-modal task completion. One direction (Wu et al., 2023a; Yang et al., 2023c,a) uses ChatGPT as the intermediary to choose the best tools or experts for visual interpretation according to user's inquiries. Another direction is end-to-end training (Zhu et al., 2023; Liu et al., 2023e,b; Ye et al., 2023; Yin et al., 2023; Wu et al., 2023b; Zhang et al., 2023; Cao et al., 2023; Zhai et al., 2023) utilizing LLMs and visual encoders to create integrated models for multimodal tasks with inter-connected parameters to relate them. These existing approaches perform well on general visual and language tasks like image captioning and visual question answering with strong language skills. However, when it comes to chart understanding, they often fall short due to a lack of specific training to bridge the chart information with the textual content. Our work enhances chart understanding by introducing a novel chart visual instruction-tuning corpus and chart understanding model.

Chart Text Understanding. Another line of research (Kantharaj et al., 2022; Masry et al., 2023;

Lee et al., 2023) is to train a high-resolution image encoder on a large image-text pair corpus to learn text recognition during the pretraining stage. However, these models rely on specific finetuning on different downstream datasets and cannot achieve open-domain multi-task understanding like LLMs or LMMs do. Earlier datasets such as (Kahou et al., 2017; Chaudhry et al., 2020; Methani et al., 2020; Masry et al., 2023; Liu et al., 2020, 2023c) primarily rely on synthetic data, with template-generated questions and answers selected from a fixed vocabulary. More recently, ChartQA (Masry et al., 2022) utilizes real-world, web-crawled charts to develop its visual question-answering datasets, supplemented by human annotators. However, it mainly focuses on compositional and visual questions. (Li and Tajbakhsh, 2023) uses Palm-2 to generate question-answering data for academic charts. However, the answers generated by Palm-2 contain hallucinations. Comparatively, the advantages of our dataset come from its larger size, more diverse topics, richer language styles, and good quality.

3 MMC-Instruction

3.1 Chart-Text Alignment Data

To build a large training corpus for chart-text alignment with a diverse range of styles and topics, we aim to collect chart and text data from online sources. We first collect the *Scientific Chart-Caption* corpus with both chart and text crawled from arXiv. In addition, we filter several existing public datasets that are suitable for chart-text alignment. The collected charts can be categorized into multiple topics, including (*computer science, business, health, biology, agriculture, etc.*), and a variety of chart types, including but not limited to *(histograms, scatter plots, area charts, and heatmap)*. More statistic is shown in Tab. 1 and Tab. 2.

Scientific Chart-Caption data collected by us. We first download the academic articles (2010-2020) through an official dump from the arXiv website. It is licensed under CC-0, which grants remake and republish rights. Unlike (Hsu et al., 2021) using PDFs, we utilize the source files containing the original LaTeX and figure files. In order to improve the dataset quality, we removed the source files without LaTeX or figure files and the source files that are hard to parse. We only keep the chart figures with rich text information by deleting the pairs whose caption length is less than 25 tokens. Finally, we collect 210k chart-text pairs in total.

Benchmark	Size	Images	Source	Answer
VQA	>1M	General	Annotated	Open
GQA	>1M	General	Synthesized	Open
MME	1.5k	General	Annotated	Y/N
Lynx-Bench	0.5k	Video	Annotated	Open
MMBench	3k	General	Repurposed	MQA
MM-Vet	0.2k	General	Repurposed	MQA
MathVista	1.4k	Math	Synthesized	MQA
MMC-Benchmark	2k	Chart/Plot	Internet, Annotated	Open/MQA

Statistic	Num
MMC-Instruction	600k
– Scientific Chart-Caption	210k
– Filtered Existing Datasets	190k
– GPT-Generated Instructions	200k
MMC-Benchmark	2k
– Unique number of images	1,063
– Multiple-choice questions	1,275
– Free-form questions	851
– Average question length	15.6

Table 2: Comparison between *MMC-Benchmark* with existing vision-language benchmarks. *MQA* means multiple-choice question answering. *Repurposed* means the benchmark is a compilation of prior datasets. *Y/N* means yes/no questions. *MMC-Benchmark* is the only existing benchmark with high-quality images for chart understanding.

Tasks	Image Source	Question Source	Question Type	Number	Human Check
Chart Information Extraction	Statista.com	GPT-4	Free-form/MQA	330	✔
Chart Reasoning	Statista.com	GPT-4	Free-form/MQA	256	✔
Contextual Chart Understanding	arxiv	GPT-4, human	Free-form/MQA	56	✔
Multiple Chart Understanding	arxiv	GPT-4, human	Free-form/MQA	52	✔
Chart Type Classification	Web Crawl	Groundtruth label	Free-form/MQA	360	✔
Chart Topic Classification	Web Crawl	Groundtruth label	Free-form/MQA	536	✔
Chart To DataTable	VisText	Source Article	Free-form/MQA	400	✔
Chart To Json	VisText	GPT-4	Free-form/MQA	96	✔
Stock Chart Analysis	Google Bard	Source Article	Free-form/MQA	40	✔

Table 3: Compositions of *MMC-Benchmark*. The distributions of topics and types are shown in Fig. 6 and Fig. 7.

Leveraging Existing Datasets. For chart-text alignment training with diverse chart caption data, we further include the following five public chart datasets for which the underlying data tables are available: (i) Statista (Kantharaj et al., 2022), (ii) PlotQA (Methani et al., 2020), (iii) VisText (Tang et al., 2023), (iv) ChartInfo (Lal et al., 2023), (v) Unichart (Masry et al., 2023). We randomly picked approximately 190k image-text pairs from these public datasets to increase the diversity.

3.2 Chart Instruction-Tuning Data

This section introduces the construction of our instruction tuning data with 200k instances. To align the model to follow a variety of instructions, we construct diverse instruction-tuning instances about the provided chart images by prompting the language-only GPT-4 (OpenAI, 2023a). Specifically, given a chart description, we design instructions in a prompt that asks GPT-4 to generate questions and answers in a style as if it could see the image (even though it only has access to the text). The prompt examples for GPT-4 are shown in Fig. 23, 24, 25, 26. Our instruction-tuning format is: "Human: {question} AI: {answer}". *MMC-Instruction* includes the following tasks: *chart information extraction*, *chart reasoning*, *scientific chart understanding*, *chart-to-datatable*, and *chart-to-json*.

Chart Information Extraction requires the

model to extract from the input chart detailed information such as title, coordinate value, scope, etc. To achieve this goal, we collect the generated L1 captions from (Tang et al., 2023), whose content enumerates aspects of the chart's construction. Then, we ask GPT-4 to generate question-answer pairs about the detailed construction information about the chart given descriptions (Fig. 23). Additionally, we require the generated answers to be less than 20 words to address hallucination.

Chart Reasoning requires the model to analyze and identify data patterns, relationships, and anomalies of the input chart. To achieve this goal, we collect the generated L2/L3 captions from (Tang et al., 2023), which summarize the statistics and synthesize the cognitive phenomena of the chart. Then, we ask GPT-4 to generate question-answer pairs that require analysis skills in Fig. 24.

Scientific Chart Understanding is a challenging task that needs scientific background knowledge. To create instruction-tuning data, we combine the abstract, title, and image captions of arXiv papers to construct the comprehensive textual context. Sometimes, the image caption is too short for GPT-4 to generate meaningful questions and answers regarding the image. To provide more context regarding the image, we also created a prompt that included paragraphs mentioning the figure in the paper. From our observation, we find a portion

of the questions are not graph-related but a follow-up on the textual context in previous answers. We use heuristic rules to delete the non-chart-related questions. The prompt is shown in Fig. 25.

Chart-to-DataTable and **Chart-to-Json** are the tasks of transforming the visual information represented in the chart into the structured data format of a table or a JSON. This process typically requires interpreting the graphical elements of the chart, such as bars, lines, or pie segments, quantifying their values, and then organizing these values into a tabular format that accurately reflects the original chart. As shown in Fig. 26, we transform the groundtruth data table from (Tang et al., 2023) to create the JSON format into our *MMC-Instruction*.

Further Quality Control. We first remove instances with answers longer than 20 words. We remove the instances mentioning unneeded content like "given caption" and "existing descriptions". As for the *Chart-to-Json* task, we remove the instances without mentioning *"title"* as the key. To examine the quality of our dataset, we randomly sample 500 instances and ask expert annotators to determine whether the output answers from GPT-4 are correct or not, with regard to the instruction and the image content. We find that 91% of the instructions are appropriate for the image inputs. Furthermore, 85% of outputs are acceptable responses to the instructions. Though some responses may contain errors, most generations conform to the correct structure, serving as applicable instruction-tuning guidelines.

4 MMC-Benchmark

The recent progress of LMMs has enabled the open-ended zero-shot completion of user-oriented vision-language tasks such as open-ended chart understanding. As a result, a comprehensive evaluation benchmark is necessary to evaluate the performances of different LMMs on these tasks and provide quantitative guidance for future research and development. However, for chart understanding, existing benchmarks often fall short of evaluating open-ended questions and unbounded chart types. Our dataset, *MMC-Benchmark*, is therefore motivated to bridge this gap, offering three unique characteristics for chart understanding:

(i) *MMC-Benchmark* is the first benchmark with human annotations to evaluate LMM's ability to comprehend visual charts.

(ii) *MMC-Benchmark* is more diverse with various sources and nine different tasks, including *chart information extraction*, *chart reasoning*, *contextual chart understanding*, *multiple chart understanding*, *chart type classification*, *chart topic classification*, *chart-to-datatable*, *chart-to-json*, and *stock chart analysis*, with examples shown in Fig. 1.

(iii) *MMC-Benchmark* provides two evaluation methods for convenient quantitative analysis, including free-format Generation Ability Evaluation using GPT-4 and multiple-choice QA format Chart Understanding Ability Evaluation without the requirement of GPT-4. The statistic of *MMC-Benchmark* is shown in Tab. 2 and Tab. 3.

4.1 Data Annotation and Quality Control

For *chart information extraction* and *chart reasoning* tasks, the images are samples from (Masry et al., 2022), but the instruction-answer pairs are all manually constructed by us rather than from existing public annotations. For *contextual chart understanding* and *multiple chart understanding*, we collect the source images from scientific charts of arXiv that are not presented in our training sets. Contextual chart understanding requires the models to read the context information to answer the questions of the charts. We utilize the abstract of the scientific paper as the context information. We manually design the questions for the multiple chart understanding, which evaluates the model's complex reasoning ability to compare between multiple charts. The *chart type classification* task contains seven types: line, bar, pie, scatter, heatmap, histogram, and Radar. The images of line, bar, and pie chart are from (Methani et al., 2020) while others are collected by us from Google Bard. The *chart topic classification* task includes health, business, science, travel, biology, engineering, and sports, whose images are crawled from Google. As for the *chart-to-datatable* and *chart-to-json* tasks, we use the images and data tables from (Wu et al., 2023a). The json data is generated by prompting GPT-4 with the datatable as the input. Finally, for *stock chart analysis*, we collect the chart images of stock from Google Bard without including corresponding captions due to hallucination concerns. Instead, we look through the source article and manually construct the questions about the stock trend, predictions, and corresponding background knowledge. We adhere to copyright and license regulations, avoiding data from sites prohibiting copy and redistribution. More examples are shown in Fig. 1, 9, 10, 11, 12, 13, 14, 15, 16. The topic

and type distributions are shown in Fig. 7 and 6.

4.2 Evaluation Protocols

In order to evaluate LMMs' generation ability and chart understanding ability, the instructions in *MMC-Benchmark* consist of two parts.

Generation Ability Evaluation utilizes GPT-4 "gpt-4-32k-0314" to assess the accuracy of prediction given question and reference answers using prompts shown in Fig. 8. We randomly select 300 samples from our testing set and manually evaluate the model predictions. We find GPT-4 assisted evaluation can achieve 0.90 agreement (Cohen's kappa agreement) with human evaluation.

Understanding Ability Evaluation (MQA), which aims to let the model select the correct answer from multiple-choice questions (MQA) given the chart. For each image, we manually design choices for each question. *Understanding Ability Evaluation* does not require the utilization of GPT-4. We adopt micro-averaged accuracy as the evaluation metric in (Yu et al., 2023) with the help of systematic, rule-based evaluation pipelines.

5 MultiModal Chart Assistant (MMCA)

Architecture. Our model *MMCA* (Fig. 2) is built on mPLUG-Owl (Ye et al., 2023)) that guides LLMs to follow multimodal instructions. In order to improve the existing LLMs to perform better on chart understanding tasks, we further fine-tune mPLUG-Owl 7B (Ye et al., 2023)) on our proposed *MMC-Instruction* corpus consisting of *Chart-Text Alignment Data* and *Chart Instruction-Tuning Data*. mPLUG-Owl contains a pre-trained visual foundation model (CLIP vision encoder), a visual abstractor, and a language foundation model (Vicuna). The visual foundation model is responsible for extracting visual features from the input images, and the visual abstractor distills these features using a set of learnable tokens. The resulting visual features are combined with the word embeddings of the input sentence and fed into the language model to generate the response. We incorporate a two-stage training paradigm.

Stage-1: Chart Text Alignment. In this stage, we freeze the language decoder and train the visual parts with our *Chart-Text Alignment Data* for one epoch. This stage enables the mapping of visual features of charts to LLM's word embedding space.

Stage-2: Chart Instruction Tuning. In the second stage, we freeze the visual abstractor, visual encoder, and language decoder and use the low-
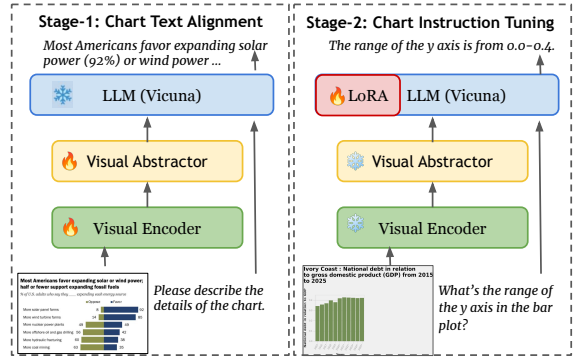

Figure 2: The overall architecture of MMCA, which is continuously trained in two stages.

rank adaptation approach (LoRA) (Ye et al., 2023) to fine-tune the language model. Specifically, we train the language model with LoRA on our *Chart Instruction-Tuning Data* for three epochs. This stage enables LLM's instruction following capabilities for chart understanding.

6 Experiments

6.1 Experimental Setup

Our model training and inference are conducted with Tesla V100 GPUs. The evaluation is conducted under a zero-shot setting. More implementation details are discussed in the Appendix.

6.2 Baselines

We compare *MMCA* with existing models in three setups: (a) Open-source LMMs including MiniGPT-v2-7B (Chen et al., 2023a), mPLUG-owl-7B (Ye et al., 2023), LRV-Instruction-7B (Liu et al., 2023b), LLaVA1.5-7B (Liu et al., 2023d), and Multimodal-GPT-9B (Gong et al., 2023). (b) GPT-4V (OpenAI, 2023b) by OpenAI. (c) Non-LLMs based models including Pix2Struct (Lee et al., 2023) and Donut (Kim et al., 2022).

6.3 Experiment Results

6.3.1 Evaluation Results on *MMC-Benchmark*

As indicated in Tab. 4, Tab. 5 and Tab. 8, *MMCA* achieves better performance in all nine tasks in comparison with the existing open-source models. The improvement of *MMCA* demonstrates the effectiveness of our *MMC-Instruction* data in enabling the LMM to complete chart understanding tasks. In addition, we find that current LMMs are better at understanding cross-modality relationships in the image but weaker at comprehending text layout information. This can be attributed to their lack of text recognition, scientific knowledge, and math reasoning abilities. Though fine-

Free-form Evaluation	LLAVA1.5	MiniGPT-v2	mPLUG-Owl	LRV-Instruct	MMCA (Ours)	GPT-4V
Chart Information Extraction	0.32	0.29	0.27	0.24	**0.35**	**0.63**
Chart Reasoning	0.30	0.23	0.22	0.19	**0.30**	**0.57**
Contextual Chart Understanding	0.33	0.29	0.28	0.23	**0.33**	**0.55**
Multiple Chart Understanding	0.27	0.20	0.23	0.21	**0.29**	**0.39**
Chart Type Classification	0.30	0.27	0.25	0.22	**0.31**	**0.79**
Chart Topic Classification	0.31	0.23	0.24	0.21	**0.32**	**0.82**
Stock Chart Analysis	0.27	0.28	0.25	0.23	**0.32**	**0.70**
Chart to Datatable	0.00	0.00	0.05	0.00	**0.08**	<u>0.05</u>
Chart to Json	0.01	0.00	0.00	0.00	**0.05**	<u>0.04</u>
Overall	0.24	0.21	0.20	0.17	**0.26**	<u>0.51</u>

Table 4: *MMC-Benchmark* evaluation results of LLaVA1.5, MiniGPT-v2, mPLUG-Owl, LRC-Instruct, MMCA, and the recent GPT-4V regarding the Generation Ability Evaluation. Given the reference response, we apply *GPT-4* to determine the correctness/incorrectness (as in Fig. 8) of the response for each test sample. The ratio of correct responses out of responses for all test samples in each task is used for evaluation. Tab. 9 shows the sizes of models.

MQA Evaluation	LLAVA1.5	MiniGPT-v2	mPLUG-Owl	LRV-Instruct	MMCA (Ours)	GPT-4V
Chart Information Extraction	0.47	0.43	0.45	0.45	**0.49**	**0.76**
Chart Reasoning	0.45	0.39	0.41	0.41	**0.47**	**0.74**
Contextual Chart Understanding	0.49	0.51	0.50	0.42	**0.55**	**0.79**
Multiple Chart Understanding	0.42	0.41	0.43	0.45	**0.47**	**0.65**
Chart Type Classification	0.55	0.52	0.55	0.50	**0.59**	**0.85**
Chart Topic Classification	0.59	0.56	0.54	0.51	**0.64**	**0.87**
Stock Chart Analysis	0.52	0.49	0.45	0.45	**0.57**	**0.81**
Chart to Datatable	0.57	0.46	0.44	0.35	**0.64**	**0.71**
Chart to Json	0.51	0.44	0.41	0.39	**0.59**	**0.69**
Overall	0.51	0.47	0.45	0.43	**0.56**	<u>0.76</u>

Table 5: *MMC-Benchmark* evaluation results of LLaVA1.5, MiniGPT-v2, mPLUG-Owl, LRC-Instruct, MMCA and the recnet GPT-4V regarding the Understanding Ability Evaluation via *Multichoice QA* (MQA) task. We calculate the accuracy of the model predictions in the MQA setting. There is no need to call *GPT-4* for this evaluation.

tuned with instruction-tuning data from text-rich images, LLAVa1.5 and mPLUG-Owl do not perform well, indicating that strong text recognition abilities in images do not guarantee high performance on *MMC-Benchmark*, which requires comprehensive visual perception and chart reasoning capability. Additionally, current LMMs perform badly on the *chart-to-datatable* and *chart-to-json* tasks. We speculate it is because these two tasks require strong OCR ability to output all the data values in the chart correctly. If one value is missing, the prediction will be regarded as incorrect. Besides, we also find that the overall performance of *Multiple Chart Understanding* is lower than *Contextual Chart Understanding*. This phenomenon may be attributed to the lack of training data with multiple images as input.

6.3.2 Results on Public Benchmarks

We compare our *MMCA* with the state-of-the-art methods on existing public benchmarks including ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), and TextVQA (Yang et al., 2021). As shown in Tab. 6, our *MMCA* outperforms existing LMMs, including MiniGPT4 and LRV-Instruction,

Model	ChartQA	DocVQA	TextVQA
Donut	41.8	67.5	43.5
Pix2Struct	56.0	72.1	-
MiniGPT-v2	49.5	61.3	50.7
LLaVA1.5	52.5	66.5	58.2
Mplug-Owl	51.6	62.2	54.3
MMCA (Ours)	**57.4**	**72.5**	**59.6**

Table 6: Comparison with OCR-free methods and LMMs on existing public benchmarks.

on the three public benchmarks. Moreover, without including any finetuning on the corresponding datasets, our proposed *MMCA* model outperforms both Pix2Struct and Donut that are finetuned on these datasets. Such results validate that with the help of LMMs, chart understanding performance can be significantly improved.

6.3.3 Comparison with GPT-4V(ision)

We further evaluate and benchmark GPT-4V (OpenAI, 2023b) via the "gpt-4-vision-preview" model on our proposed *MMC-Benchmark*. The quantitative results are shown in Tab. 4 and Tab. 5. Remarkably, GPT-4V surpasses all open-source LMMs by
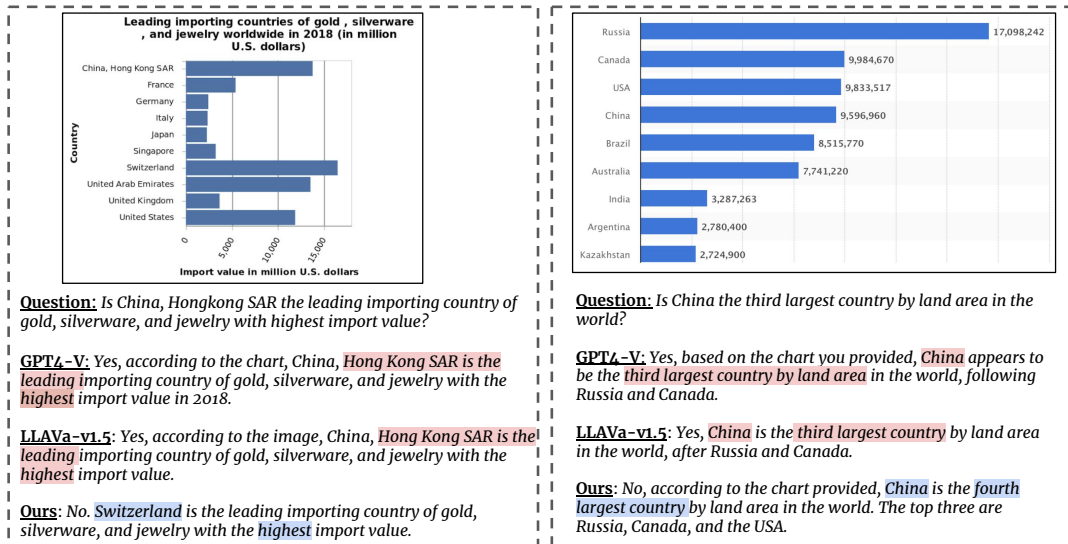
Figure 3: RED highlights incorrect answers while BLUE highlights the correct ones. (Left): the failure of GPT-4V and LLaVA1.5 can be attributed to *vision perception error*. (Right): the failure can be attributed to *language bias*.

substantial margins on seven out of the nine tasks of *MMC-Benchmark*. Such results prove GPT-4V's strong skills in text recognition, scientific knowledge, and math reasoning. ***The only two tasks GPT-4V performs poorly are Chart to Datatable and Chart to Json***. As shown in Fig.20, GPT-4V misrecognizes the data value from the charts. GPT-4V also predicts incorrectly on the *Multiple Charts Understanding* tasks such as Fig. 18, 19.

We examine 100 randomly sampled error instances from GPT-4V's predictions. The instances are analyzed by expert annotators who identify the root causes. The distribution of errors is in Fig. 4. **Language Bias (35%)**: As indicated in Fig. 3 (right), the strong language prior or parametric memory misleads GPT-4V to answer *"China appears to be the third largest country by land area in the world"*, which conflicts with the information mentioned in the chart *"USA appears to be the third largest country by land area"*. **Perception Error (39%)**: As in Fig. 3 (left) and Fig. 18, the perception error occurs when GPT-4V fails to interpret the chart (Liu et al., 2023a). The remaining errors include **Reasoning Error (15%)** in Fig. 19 and **Lack of Knowledge (11%)** in Fig. 21. These errors are attributed to various factors such as complex text interpretation, lack of domain-specific knowledge, or failure to extract answers from long context. More cases are shown in Fig. 20, and 22.

6.3.4 Error Analysis of Open-Source Models

Not Following Instructions. Even with a very concise instruction design, there are LMMs that do not follow the user's instructions. For example, in Fig. 27b, when asked *"Please identify the proportion of Americans who favor the coal mining."*, PixsStruct and MiniGPT-v2 answer *"Yes"* and *"Most Americans favor exporting or expanding solar and wind powers."*, respectively. In our opinion, a good chart understanding model should be able to follow instructions. However, to the best of our knowledge, most of the existing LLM-based or LMM-based models, except for GPT-4V, are not able to follow human instructions well. More examples are shown in Fig. 27a, 27c, and 28.

Vision Encoder is Weak. Existing LMMs typically use CLIP as the vision encoder and do not update its parameters during training. However, as CLIP is trained to align visual embeddings with short captions, its capability of modeling the spatial interactions of chart elements like trend lines and color-coded legends is limited. The potential method is to add segmentation (Kirillov et al., 2023) and project the segments into the LLM token embedding space. Instead, in our proposed *MMCA* approach, we finetune LMMs on our *MMC-Instruction* data by updating the vision parts during training and improving the integration of visual elements into the LLM input domain. The result improvements prove the effectiveness of *MMC-Instruction* and the training strategy in *MMCA*. Fig. 5 shows the distributions of failure causes.

7 Conclusion

This paper aims to tackle the challenge of chart understanding with Large Multimodal Models (LMMs). Firstly, we present a large-scale chart

instruction-tuning dataset *MMC-Instruction*, including diverse topics, language styles, chart types, and open-ended answers in line with human cognition. Secondly, we introduce a human-annotated benchmark called *MMC-Benchmark* to evaluate LLMs' abilities for chart understanding quantitatively. Finally, we propose an instruction-tuned LMM called *MMCA* that outperforms existing open-source SoTA methods.

8 Limitations

Our study innovatively utilizes a large multimodal model with 7 billion parameters, showcasing substantial capabilities within the constraints of our current computational resources. While we recognize that employing even larger models, such as the 13 billion parameter variants, could further enhance our findings, lacking access to high-end computing resources like A100 limits our current scope. This presents an exciting avenue for future research, where we aim to expand our model's complexity and depth as more advanced computational means become available.

9 Ethical Considerations

Copyright and Licensing: Strict adherence to copyright and licensing regulations is mandatory. Data from sources that prohibit copying or redistribution will be explicitly avoided. **Data Privacy**: Compliance with privacy laws and ethical standards in data handling is paramount. The annotators should avoid collecting questions that contain any private information.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. 2023. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

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

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

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

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

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 *European Conference on Computer Vision*, pages 498–517. Springer.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Jay Lal, Aditya Mitkari, Mahesh Bhosale, and David Doermann. 2023. Lineformer: Line chart data extraction using instance segmentation. In *International Conference on Document Analysis and Recognition*, pages 387–400. Springer.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

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

Fuxiao Liu, Hao Tan, and Chris Tensmeyer. 2023c. Documentclip: Linking figures and main body text in reflowed documents. *arXiv preprint arXiv:2306.06306*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visualnews : Benchmark and challenges in entity-aware image captioning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023d. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

James Manyika. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*.

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. *arXiv preprint arXiv:2305.14761*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

OpenAI. 2022. Introducing chatgpt.

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

OpenAI. 2023b. Gpt-4v(ision) system card.

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023a. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023c. Mmreact: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. Tap: Text-aware pretraining for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2023. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*.

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

A Appendix

A.1 MMC-Benchmark

In this section, we discuss more about our *MMC-Benchmark*.

Generation Ability Evaluation utilizes GPT-4 to assess the accuracy of the model prediction given the question and reference answers in Fig. 8. Then we ask GPT-4 to assess the prediction accuracy.

Distriutions of Plot Types and Topics. Fig. 7 and Fig. 6 present the distributions of chart topic and plot types in *MMC-Benchmark*. Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15 and Fig. 16 show the data examples of different tasks in our *MMC-Benchmark*.

A.2 Experiment

A.2.1 More Experiments Results

We further compare *MMCA* with Donut (Kim et al., 2022), BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023) and Shikra (Chen et al., 2023b). From Tab. 8, we observe that non-LLM based models like Donut work well on the *Chart Information Extraction* and *Chart Reasoning* tasks. However, the performance drops a lot when facing other tasks, including *Multiple Chart Understanding*, *Chart Type Classification*, and *Chart to Json*. There could be two reasons. First, the language decoder of non-LLM can not understand the questions correctly. Second, Donut's training set is not diverse enough to cover various topics and plot types. It demonstrates the value of our *MMC-Instruction*.

A.2.2 Implementation Details

Our *MMCA* model is trained with 8 Nvidia Tesla V100 GPUs. Based on the second-stage checkpoint of mPLUG-Owl, we conduct Chart Text Alignment training for one epoch with a batch size of 8. We use the same data augmentation strategy as in BLIP-2 (Li et al., 2023), including random resized cropping and horizontal flipping with a probability of 0.5. The number of learnable queries is set to 64. We use the AdamW optimizer. The cosine learning rate decay scheduler is used with a peak learning rate of $1e^{-4}$ and 1,000 warmup steps. For the learning rate of the vision encoder, we employ layer-wise learning rate decay with a factor of 0.9 to retain the low-level visual representation. For Chart Instruction Turning, we train the language model for three epochs with a learning rate of $2e^{-5}$ and a batch size of 8.

A.2.3 Multiple-Choice Questions Evaluation

For multiple-choice questions, we design systematic, rule-based evaluation pipelines. Specifically, we construct robust regular expressions and develop response-processing workflows to mitigate the potential influence of any intermediate generations (e.g., reasoning steps, calculations) in the long response. These are employed to extract key phrases, such as numbers and conclusion phrases, from the long responses for accurate answer matching. If there is no valid answer in the model's response, we perform random selection as a remedy for multiple-choice questions or consider the response incorrect for open questions.

A.2.4 Error Analysis of GPT-4V(ision)

We examine 100 randomly sampled error instances from GPT-4V's predictions. The instances are analyzed by expert annotators who identify the root causes. The distribution of errors is in Fig. 4.

Language Bias (35%). *Language Bias* refers to perceptions formed without relevant visual input. As indicated in Fig. 3 (right), the strong language prior or parametric memory misleads GPT-4V to answer *"China appears to be the third largest country by land area in the world"*, which conflicts with the information mentioned in the chart *"USA appears to be the third largest country by land area"*.

Perception Error (39%). *Perception Error* denotes the misinterpretation of accurate visual information. As depicted in Fig. 3 (left), the perception error occurs when GPT-4V fails to detect the trend in the chart (Fig. 18).

Other Errors. The remaining errors include **Reasoning Error (15%)** in Fig. 19 and **Lack of Knowledge (11%)** in Fig. 21. These errors are attributed to various factors, such as complex text interpretation challenges, lack of domain-specific knowledge, or failure to extract precise answers from long context. More cases are shown in Fig. 20 and Fig. 22.

A.2.5 Error Analysis of Open-Source Models

We examine 100 randomly sampled error instances from open-source models. The instances are analyzed by expert annotators who identify the root causes. The distribution of errors is in Fig. 5. Different from GPT-4V, one key issue of the open-source model is *Not Following Instructions (27%)*. Even with a very concise instruction design, there are LMMs that do not follow the user's instructions. For example, in Fig. 27b, when asked *"Please iden-*
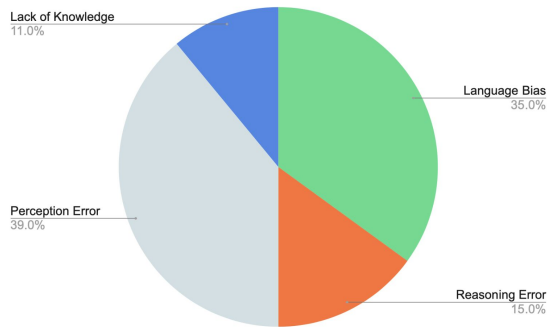
Figure 4: Error distribution of GPT-4V over 100 randomly sampled error instances.
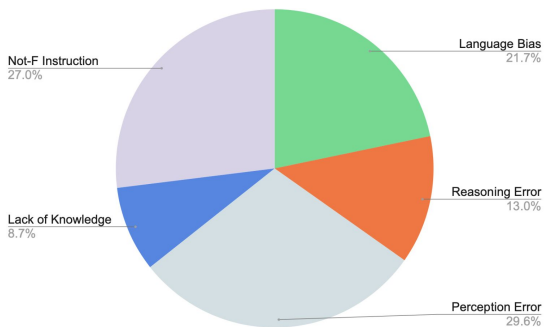

Figure 5: Error distribution of open-source models over 100 randomly sampled error instances. **Not-F Instruction** means *"Not Following Instructions"*.

Figure 6: Distributions of chart types in *MMC-Benchmark*.

Model	ChartQA	DocVQA	TextVQA
MMCA (Ours)	**57.4**	**72.5**	**59.6**
-w/o FT Vision Encoder	54.2	67.8	57.2

Table 7: Ablation experiments without fine-tuning vision encoder in MMCA.

tify the proportion of Americans who favor the coal mining.", PixsStruct and MiniGPT-v2 answer *"Yes"* and *"Most Americans favor exporting or expanding solar and wind powers."*, respectively. In our opinion, a good chart understanding model should be able to follow instructions. However, to the best of our knowledge, most of the existing LLM-based or LMM-based models, except for GPT-4V, are not able to follow human instructions well. More examples are shown in Fig. 27a, 27c, and 28.

Another key issue is **Vision Encoder is Weak (29.6%)**. Existing LMMs typically use CLIP as the vision encoder and do not update its parameters during training. However, as CLIP is trained to align visual embeddings with short captions, its capability of modeling the spatial interactions of chart elements like trend lines and color-coded legends is limited. One potential method is to add segmentation (Kirillov et al., 2023) and project the segments into the LLM token embedding space. Instead, in our proposed *MMCA* approach, we fine-tune LMMs on our *MMC-Instruction* data by updating the vision parts during training and improving the integration of visual elements into the LLM input domain. As shown in Tab. 7, the model with-
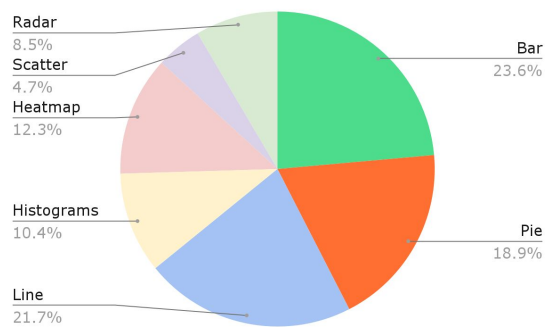
out fine-tuning the vision encoder under-performs our proposed MMCA model. It indicates that fine-tuning the vision encoder part of the model is necessary. The improvements in our experiments also demonstrate the effectiveness of our proposed *MMC-Instruction* dataset and the training strategy in *MMCA*. Please refer to Fig. 27a, Fig. 27b, Fig. 27c, and Fig. 28 for more examples.

A.2.6 More Discussions

Chart-to-DataTable and Chart-to-Json are extremely Difficult. As shown in Tab. 4, all current LMMs, including **GPT-4V**, perform badly on these two tasks. It is probably due to the fact that these two tasks require strong OCR skills to output all the data values in the chart correctly. If one value is missing, the prediction will be regarded as incorrect. Compared to the baselines in Fig. 27a, our *MMCA* model is able to produce more accurate responses in correct output formats.

MMC-Benchmark is more Challenging than Previous Benchmarks. From Tab. 5, we find that the overall scores for existing models on *MMC-Benchmark* are lower than those on the current benchmarks like ChartQA. Such results are expected since the questions in *MMC-Benchmark* are more diverse, and the answers are open-ended. Additionally, *MMC-Benchmark* contains more topics that require both a comprehensive understanding of charts and proficient language skills.

1299

MQA Evaluation	Donut	Shikra	BLIP2	InstructBLIP	MMCA (Ours)
Chart Information Extraction	0.46	0.38	0.36	0.41	**0.49**
Chart Reasoning	0.42	0.39	0.38	0.40	**0.47**
Contextual Chart Understanding	0.37	0.43	0.42	0.45	**0.55**
Multiple Chart Understanding	0.38	0.41	0.40	0.42	**0.47**
Chart Type Classification	0.42	0.48	0.50	0.52	**0.59**
Chart Topic Classification	0.45	0.56	0.51	0.55	**0.64**
Stock Chart Analysis	0.41	0.47	0.44	0.48	**0.57**
Chart to Datatable	0.32	0.39	0.40	0.41	**0.64**
Chart to Json	0.38	0.41	0.39	0.48	**0.59**
Overall	0.51	0.47	0.42	0.45	**0.56**

Table 8: *MMC-Benchmark* evaluation results on Donut, Shikra, BLIP-2, InstructBLIP, and our MMCA regarding the Understanding Ability Evaluation via *Multichoice QA* (MQA) task. We calculate the accuracy of the model predictions in the MQA setting. There is no need to call *GPT-4* for this evaluation.

Method	Vision Encoder	Language Model
Donut	ViT-g (1.3B)	Bert (0.34B)
Pix2Struct	ViT-g (1.3B)	BART (1.3B)
BLIP-2	ViT-g (1.3B)	Vicuna (7B)
MiniGPT-v2	ViT-g (1.3B)	Vicuna (7B)
LLaVA1.5	ViT-L (0.3B)	Vicuna (7B)
mPLUG-Owl	ViT-L (0.3B)	LLaMA (7B)
InstructBLIP	ViT-g (1.3B)	Vicuna (7B)
LRV-Instruction	ViT-L (0.3B)	Vicuna (7B)
MMCA (Ours)	ViT-L (0.3B)	Vicuna (7B)

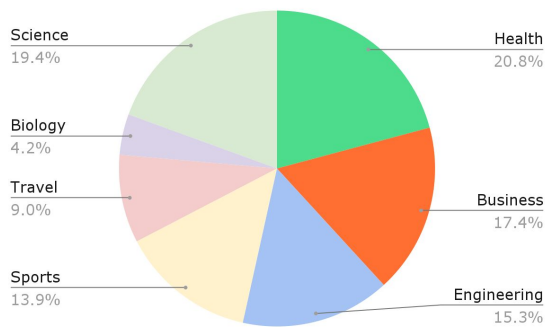Table 9: The backbones for the vision encoder and language model of the baselines and our MMCA model.

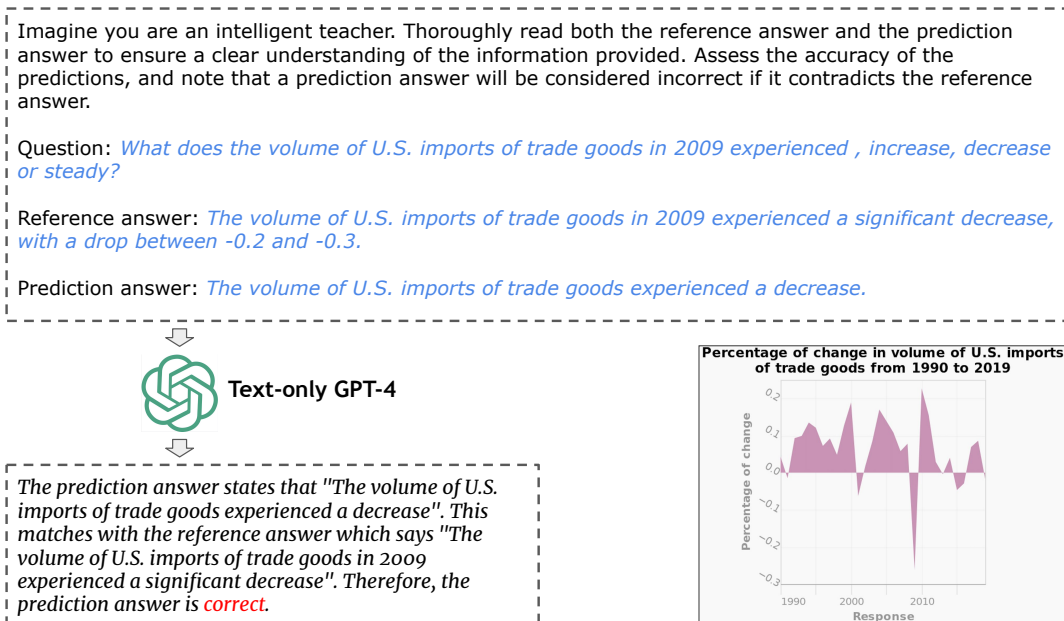Figure 7: Distributions of chart topic in *MMC-Benchmark*.

Imagine you are an intelligent teacher. Thoroughly read both the reference answer and the prediction answer to ensure a clear understanding of the information provided. Assess the accuracy of the predictions, and note that a prediction answer will be considered incorrect if it contradicts the reference answer.

Question: *What does the volume of U.S. imports of trade goods in 2009 experienced , increase, decrease or steady?*

Reference answer: *The volume of U.S. imports of trade goods in 2009 experienced a significant decrease, with a drop between -0.2 and -0.3.*

Prediction answer: *The volume of U.S. imports of trade goods experienced a decrease.*

Text-only GPT-4

The prediction answer states that "The volume of U.S. imports of trade goods experienced a decrease". This matches with the reference answer which says "The volume of U.S. imports of trade goods in 2009 experienced a significant decrease". Therefore, the prediction answer is correct.

Figure 8: An example of Generative Ability Evaluation by *text-only* GPT-4. In the prompt, we provide GPT-4 with the question, reference answer, and predictions from models. Then, GPT-4 accesses the accuracy of the model prediction following our instruction.

Question:
Among line, bar, pie, surface, scatter and heatmap, what kind of chart is the image?
Answer:
Surface

Question:
Among line, bar, pie, surface, scatter and heatmap, what kind of chart is the image?
Answer:
Scatter

Question:
Among line, bar, pie, surface, scatter and heatmap, what kind of chart is the image?
Answer:
Heatmap
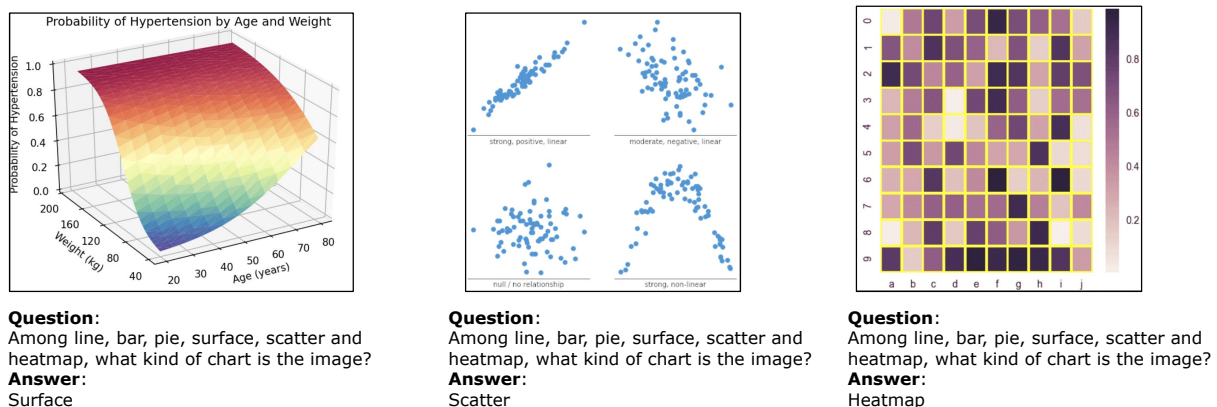
Figure 9: Examples of the *Chart Type Classification* task.

Question:
What does the y-axis represent?
Answer:
Number of COVID-19 patients within Japan, ranging from 0 to 150,000.

Question:
What is the name of the area diagram?
Answer:
Albania: Age Structure from 2009 to 2019.

Question:
How many games did Warren Spahn win? Choices: A) 250 games B) 350 games?
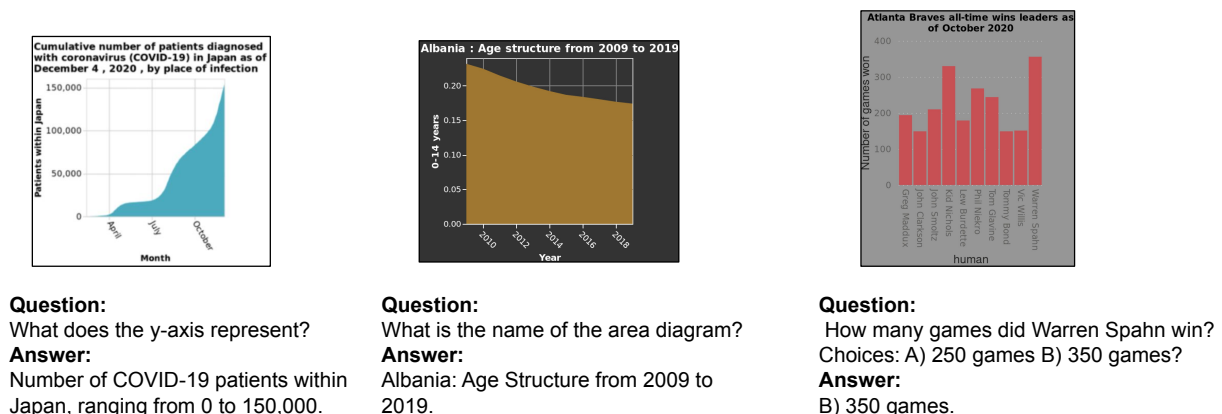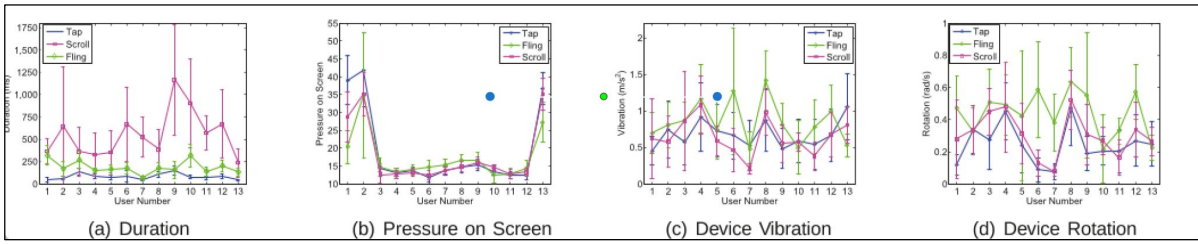**Answer:**
B) 350 games.

Figure 10: Examples of the *Chart Information Extraction* task.

1301

Question:
What is the purpose of the graph that highlights the variations in interacting features among different user numbers in the static scenario? a) To illustrate the uniqueness and diversity of behavior biometric b) To showcase the similarities and commonalities of behavior biometric c) To analyze the impact of user numbers on behavior biometric"
Answer:
To illustrate the uniqueness and diversity of behavior biometric

Figure 11: Examples of the *Multiple Charts Understanding* task.
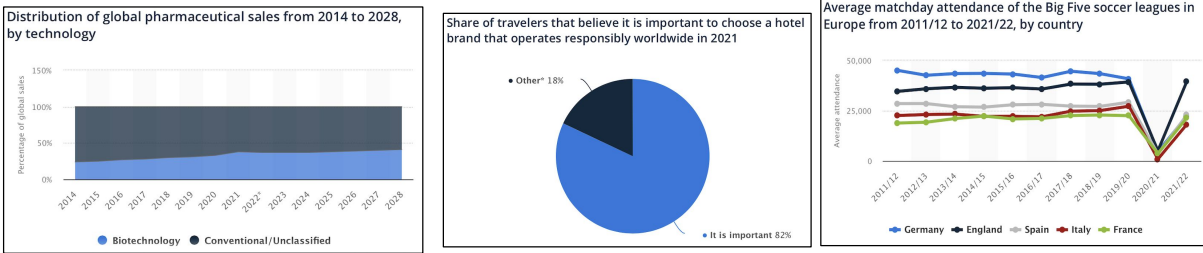
Question:
Among sports, health, travel, business, which topic is the chart most related to?
Answer:
Health

Question:
Among sports, health, travel, business, which topic is the chart most related to?
Answer:
Travel

Question:
Among sports, health, travel, business, which topic is the chart most related to?
Answer:
Sports

Figure 12: Examples of the *Chart Topic Classification* task.
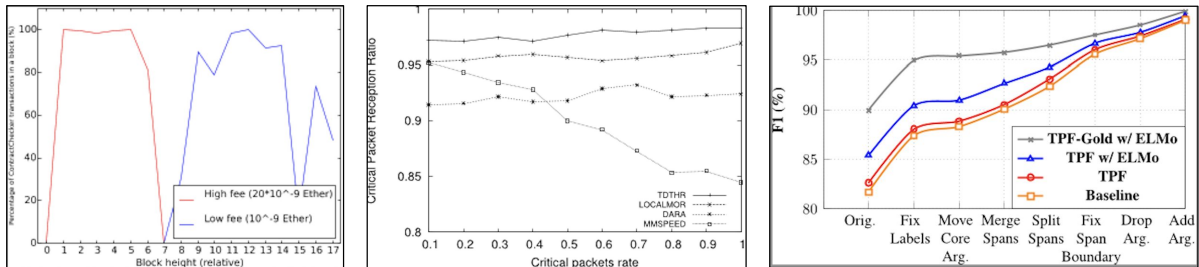
Question:
What is the purpose of this graph?
Answer:
To illustrate the relationship between transaction fees and blockchain availability

Question:
What does the graph compare in terms of packet reception ratio for critical packets?
Answer:
Four different algorithms

Question:
How do the oracle transformations enhance the CoNLL-2005 models?
Answer:
By applying operations that correct errors in predicted arguments

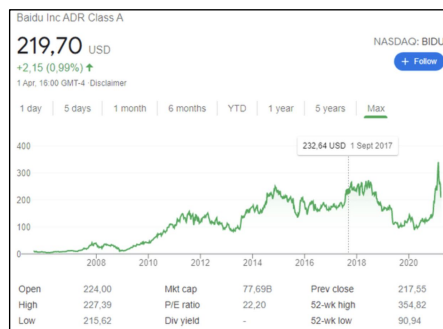Figure 13: Examples of the *Arxiv Chart Understanding* task.

Question:
What has been the trend of the stock price from January 2021 until 2022?
Answer:
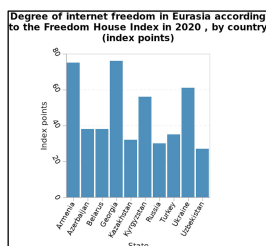The trend of the stock price from January 2021 until 10/01/22 is upward

Question:
BAIDU stock has also been very volatile. You see the stock peaking in 2007, 2011, 2015, 2020 and recently, February 2021, only to crash after and stay down for a few years. Is this statement true or false?
Answer:
Yes. The statement is correct.
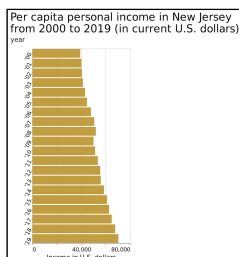
Figure 14: Examples of the *Stock Charts Analysis* task.

Question:
Is the level of freedom in Russia higher, lower, or equal to that of Georgia?
Answer:
The level of freedom in Russia is lower than Georgia.

Question:
What has the per capita income in New Jersey shown from 2000 to 2019?
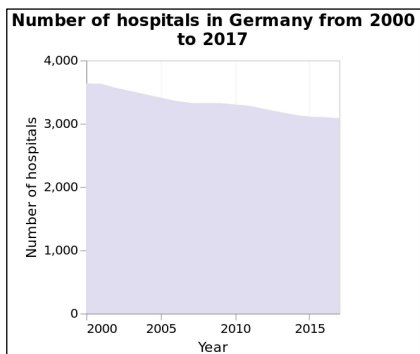**Answer:**
A steady upward trend.
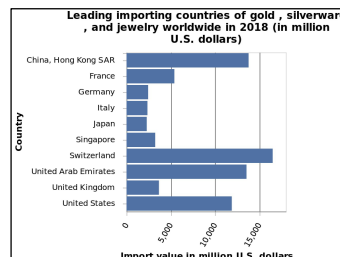
Question:
Which country is the top importer among Switzerland, UAE, and China?
Answer:
Switzerland

Figure 15: Examples of the *Chart Reasoning* task.

Question:
Can you convert the chart in the image into a json format?
Answer:
```
{
 "title": "Number of hospitals in Germany from 2000 to 2017",
 "data": [
   {"date": "Dec 31, 1999", "number_of_hospitals": 3635},
   {"date": "Dec 31, 2000", "number_of_hospitals": 3628},
   {"date": "Dec 31, 2001", "number_of_hospitals": 3564},
   {"date": "Dec 31, 2002", "number_of_hospitals": 3513},
   {"date": "Dec 31, 2003", "number_of_hospitals": 3460},
   {"date": "Dec 31, 2004", "number_of_hospitals": 3409},
   {"date": "Dec 31, 2005", "number_of_hospitals": 3359},
   {"date": "Dec 31, 2006", "number_of_hospitals": 3326},
   {"date": "Dec 31, 2007", "number_of_hospitals": 3322},
   {"date": "Dec 31, 2008", "number_of_hospitals": 3324},
   {"date": "Dec 31, 2009", "number_of_hospitals": 3301},
   {"date": "Dec 31, 2010", "number_of_hospitals": 3278},
   {"date": "Dec 31, 2011", "number_of_hospitals": 3229},
   {"date": "Dec 31, 2012", "number_of_hospitals": 3183},
   {"date": "Dec 31, 2013", "number_of_hospitals": 3138},
   {"date": "Dec 31, 2014", "number_of_hospitals": 3108},
   {"date": "Dec 31, 2015", "number_of_hospitals": 3100},
   {"date": "Dec 31, 2016", "number_of_hospitals": 3084}
 ]
}
```

Figure 16: Examples of the *Chart to Json* task.

Question:
The stock price is trading within a narrow range, which is a neutral signal. Is it true or false?

GPT-4V Output:
In technical analysis, trading within a narrow range indicates that the market is consolidating, and it doesn't give a clear signal about a bullish (positive) or bearish (negative) trend until a breakout occurs above the resistance or below the support level. Therefore, saying that the stock price is trading within a narrow range is a neutral signal is true. It suggests that investors are waiting for more definitive information that could push the stock out of the range in either direction.

Question:
The stock price is trading within a narrow range, which is a neutral signal. Is it true or false?

GPT-4V Output:
The statement you've provided about the stock price trading within a narrow range being a neutral signal is not supported by the image you've uploaded. The image actually shows a chart with a descending trendline, indicating that the stock has been experiencing a downtrend. The trendline is drawn above the price bars, showing that the highs of the price bars are getting lower over time. This is typically considered a bearish signal, not a neutral one.
So, the statement is false based on the information provided in the image.
The chart depicts a downtrend, not a stock trading within a narrow range.

Figure 17: GPT-4V works well on *Stock Chart Analysis* task.
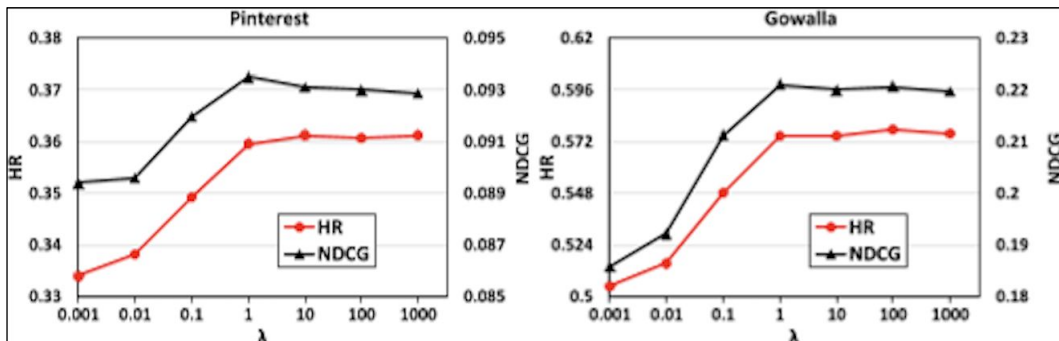
Question:
What happens when λ is smaller than 1 in terms of the performance of AMF? Choices:A) Decreasing λ leads to gradual improvements in the performance of AMF.B) Increasing λ leads to gradual improvements in the performance of AMF.C) The performance of AMF remains the same when λ is smaller than 1.
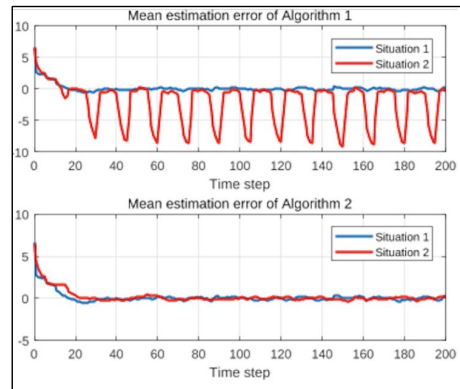
GPT-4V Output:
From the graphs, we can observe the following trends for values of λ smaller than 1:
For the Pinterest dataset, as λ decreases from 1 to 0.001, both HR and NDCG metrics increase, which indicates that the performance of AMF improves.
For the Gowalla dataset, similarly, as λ decreases from 1 to 0.001, the HR and NDCG also increase, suggesting improvement in performance.
Based on these observations, the correct answer is:
A) Decreasing λ leads to gradual improvements in the performance of AMF.

Figure 18: Failure cases of GPT-4V on *Multiple Chart Understanding* task. RED means the incorrect answers from GPT-4V. **Error category**: *Vision Perception Error*.
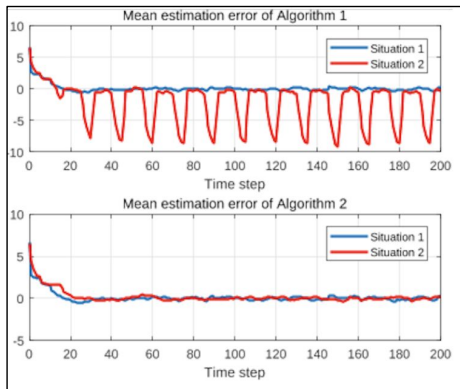
Question:
According to the graph, which of the following statements accurately describes the bias of Algorithm 1 and Algorithm 2 under Situation 1? Choices:A) Algorithm 1 is asymptotically biased while Algorithm 2 is unbiased.B) Both Algorithm 1 and Algorithm 2 are asymptotically biased.C) Both Algorithm 1 and Algorithm 2 are asymptotically unbiased.
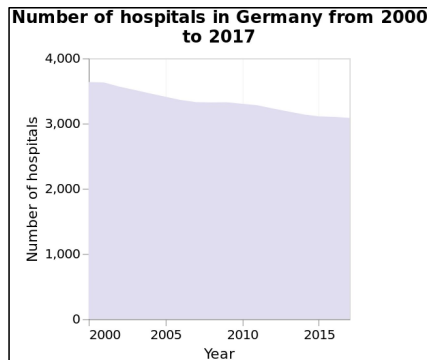
GPT-4V Output:
Given the observations, the correct choice would be: C) Both Algorithm 1 and Algorithm 2 are asymptotically unbiased.

Question:
According to the graph, which of the following statements accurately describes the bias of Algorithm 1 and Algorithm 2 under Situation 1? Choices: A) Algorithm 1 is asymptotically biased while Algorithm 2 is unbiased. B) Both Algorithm 1 and Algorithm 2 are asymptotically unbiased. C) Both Algorithm 1 and Algorithm 2 are asymptotically biased.

GPT-4V Output:
Based on this analysis, the correct answer would be: A) Algorithm 1 is asymptotically biased while Algorithm 2 is unbiased.

Figure 19: Failure cases of GPT-4V on *Multiple Chart Understanding* task. RED means the incorrect answers from GPT-4V. GREEN denotes the section of the question where the order of choices is switched. **Error category**: *Reasoning Error*.

Question:
Can you convert the chart in the image into a json format?

GPT-4V Output:
{"title": "Number of hospitals in Germany from 2000 to 2017",
  "years": [2000, 2005, 2010, 2015, 2017],
  "values": [4000,3000,2000,1000]
}
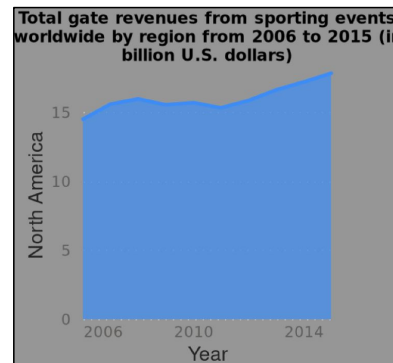
Question:
Can you convert the chart in the image into a json format?

GPT-4V Output:
{"title": "Total gate revenues from sporting events worldwide by region from 2006 to 2015 (in billion U.S. dollars)",
  "years": [2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015],
   "values": [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]}

Figure 20: Failure cases of GPT-4V on *Chart to Json* task. RED means the incorrect answers from GPT-4V. **Error category**: *Vision Perception Error*.

	Oxygen saturation (%)	Pressure (mm Hg)
Right atrium	75	3
Right ventricle	75	25/4
Pulmonary trunk	75	22/10
Left atrium	99	10
Left ventricle	99	165/45
Ascending Aorta	99	155/70

Question:
A 38-year-old man comes to the physician because of recurrent palpitations over the past 2 months. He says that he was diagnosed with a minor heart defect during childhood but does not remember the details and has not had regular medical follow-up examinations since. Physical examination shows a heart murmur. Cardiac catheterization is performed, and hemodynamic results show <image 1>. Which of the following is the most likely diagnosis? Option: (A) Mitral stenosis (B) Mitral regurgitation. (C) Aortic value regurgitation (D) Aortic stenosis

GPT-4V Output:
Based on the provided hemodynamic results: The oxygen saturation of 75% in the right atrium, right ventricle, and pulmonary trunk, followed by a jump to 99% in the left atrium, left ventricle, and ascending aorta, suggests the presence of a left-to-right shunt, where oxygenated blood is mixing with deoxygenated blood. The pressure readings are also indicative. The normal pressure in the left atrium is usually around 10 mm Hg, while the pressure in the left ventricle during systole (contraction) can go up to 120 mm Hg or more and during diastole (relaxation) drops significantly. The given pressure for the left ventricle is 165/45 mm Hg, which is significantly high during systole. Meanwhile, the pressure in the ascending aorta is 155/70 mm Hg, which is closer to normal systolic pressure but has an elevated diastolic pressure. Thus, the most likely diagnosis based on the provided data is: (B) Mitral regurgitation.

Figure 21: Failure cases of GPT-4V on *Chart to Json* task. RED means the incorrect answers from GPT-4V. **Error category**: *Lack of Knowledge*.

	Gold	Silver	Bronze
China	23	21	28
United States	36	38	36
Russian Fed.	23	21	28
Great Britain	19	13	15
Germany	16	10	15
Australia	14	15	17
Korea	13	10	8
Japan	9	6	10
Italy	8	10	10
France	7	16	17

Question:
According to the metal number chart of the 2008 Beijing Olympics, Does China have the most gold metals, followed by the USA and Russia?

GPT-4V Output:
Based on the chart provided, China has 36 gold medals. The United States has 19 gold medals. The Russian Federation (Russian Fed.) has 23 gold medals. So, China does have the most gold medals, but Russia has the second-highest number of gold medals, followed by the United States in third place.

Figure 22: Failure cases of GPT-4V on *Chart to Json* task. RED means the incorrect answers from GPT-4V. **Error category**: *Vision Perception Error* and *Language Bias Error*.

Prompt:

Here is the description of a chart "*Here is a area chart is labeled Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection. On the x-axis, Month is measured with a categorical scale starting with April and ending with October. There is a linear scale with a minimum of 0 and a maximum of 150,000 along the y-axis, labeled Patients within Japan* "

Please generate 3 different questions and answers pairs about title, x-axis, y-axis, data range or data pattern of the chart. The answers should come from the descriptions above. Each Answer must be less than 20 words. The output format should be as follows:

question1=>
answer1 =>
question2=>
answer2 =>
question3=>
answer3=>
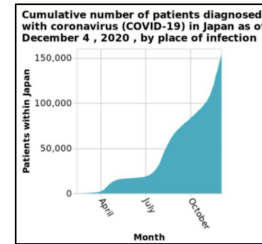
GPT4 OUTPUT Example:

question1=> What does the area chart represent?
answer1 => Cumulative COVID-19 cases in Japan by place of infection from April to October 2020.
question2=> What does the x-axis represent?
answer2 => Months from April to October 2020
question3=> When did the greatest increase in COVID-19 cases in Japan occur?
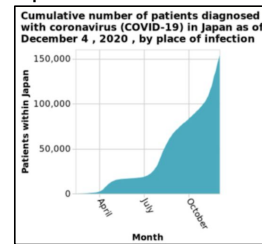answer3=> Between November and December 2020.

Figure 23: An example prompt for text-only GPT-4 we use to generate instruction and answers for *Chart Information Extraction* task. The sentence in BLUE is the captions of the chart.

Prompt:

Here is the description of a chart "*Here is a area chart is labeled Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection. On the x-axis, Month is measured with a categorical scale starting with April and ending with October. There is a linear scale with a minimum of 0 and a maximum of 150,000 along the y-axis, labeled Patients within Japan* "

Please generate 3 different questions and answers pairs about the trend, data pattern and other insightful analysis of the chart. The answers should come from the descriptions above. Each Answer must be less than 20 words. The output format should be as follows:

question1=>
answer1 =>
question2=>
answer2 =>
question3=>
answer3=>

GPT4 OUTPUT Example:

question1=> When was the first COVID-19 case diagnosed in Japan?
answer1 => March 2020.
question2=> How many COVID-19 cases were reported in Japan by December 4th, 2020?
answer2 => Approximately 160,000.
question3=> When did the greatest increase in COVID-19 cases in Japan occur?
answer3=> Between November and December 2020.

Figure 24: An example prompt for text-only GPT-4 we use to generate instruction and answers for *Chart Reasoning* task. The sentence in BLUE is the captions of the chart.

Prompt:

Here is the description of a chart "*Displays vary widely in brightness, hue, white balance, contrast and many other parameters that will influence the appearance of light. To affirm this hypothesis, an SLR camera with fixed parameters observes 3 displays and models the CDTF for each one. See Samsung in Fig. 4(a), LG in Fig. 4(b), and iMac 4(c). Although each display is tuned to the same parameters, including contrast and RGB values, each display produces a unique CDTF. Variance of Light Output among Displays. An SLR camera captured a range of grayscale [0,255] intensity values produced by 3 different LCDs. These 3 CDTF curves highlight the dramatic difference in the light emmitance function for different displays, particularly the LG.* "

Please generate questions and answers pairs about title, data pattern and purpose of the chart. The answers should come from the descriptions above. The output format should be as follows:

question=>
answer =>

GPT4 OUTPUT Example:

question=> What's the main purpose of the image?

answer => The main purpose of the graph is to illustrate the variance in light output among displays. This is done by plotting the CDTF curves for three different LCDs: Samsung. LG. and ; iMac. The CDTF curve is a function that describes the relationship between the displayed intensity and the emitted light intensity. The shows that the three displays produce significantly different CDTF curves, which indicates that they have different light output characteristics.

Figure 25: An example prompt for text-only GPT-4 we use to generate instruction and answers for *Multiple Chart Understanding* task. The sentence in BLUE is the captions of the chart.


Prompt:

Here is the datatable of a chart "*Apparent steel consumption per capita in Mexico from 2003 to 2018 (in kilograms of finished steel products) <s> Year Per capita use in kilograms 2018 194.4 2017 204.8 2016 199.8 2015 197.6 2014 189 2013 167.9 2012* 173.1 2011 166.3 2010 149.7 2009 129.9 2008 154.3 2007 164 2006 158.9 2005 143.9 2004 152.6 2003 144*"

Please convert the data table into the json format with the title and data as the key..

GPT4 OUTPUT Example:

{Title: "Apparent steel consumption per capita in Mexico",
Data: [{"Year": 2018, "Per capita use in kilograms": 194.4},
    {"Year": 2017, "Per capita use in kilograms": 204.8},
    {"Year": 2016, "Per capita use in kilograms": 199.8},
    {"Year": 2015, "Per capita use in kilograms": 197.6},
    {"Year": 2014, "Per capita use in kilograms": 189.0},
    {"Year": 2013, "Per capita use in kilograms": 167.9},
    {"Year": 2012, "Per capita use in kilograms": 173.1},
    {"Year": 2011, "Per capita use in kilograms": 166.3},
    {"Year": 2010, "Per capita use in kilograms": 149.7},
    {"Year": 2009, "Per capita use in kilograms": 129.9},
    {"Year": 2008, "Per capita use in kilograms": 154.3},
    {"Year": 2007, "Per capita use in kilograms": 164.0},
    {"Year": 2006, "Per capita use in kilograms": 158.9},
    {"Year": 2005, "Per capita use in kilograms": 143.9},
    {"Year": 2004, "Per capita use in kilograms": 152.6},
   {"Year": 2003, "Per capita use in kilograms": 144.0}
]}

Figure 26: An example prompt for text-only GPT-4 we use to generate instruction and answers for *Chart-to-Json* task. The sentence in BLUE is the captions of the chart.
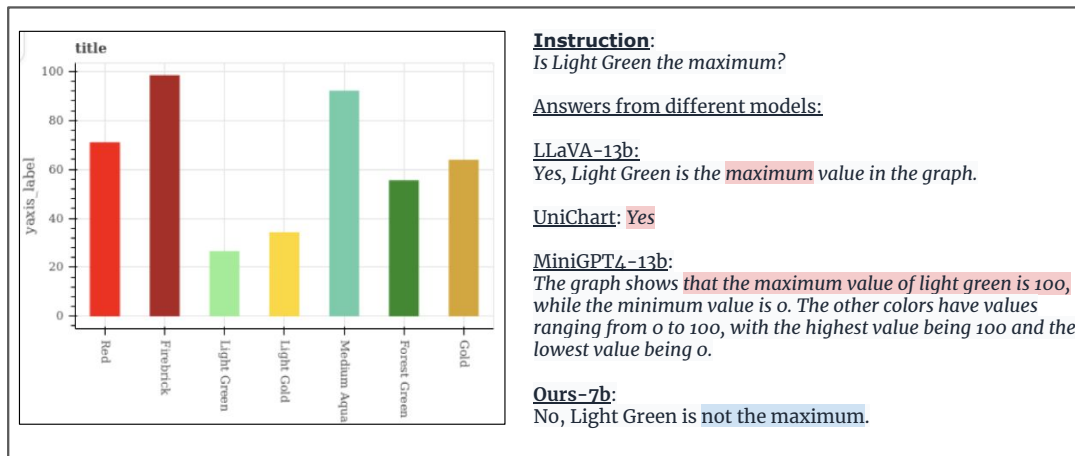
(a) Examples of *Chart Reasoning* task results from our model and other open-source models.
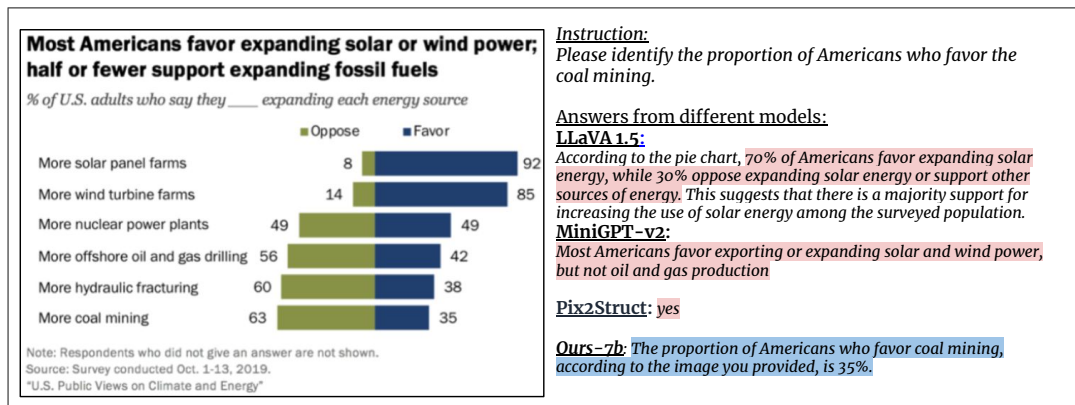
(b) Examples of *Chart Reasoning* task results from our model and other open-source models.
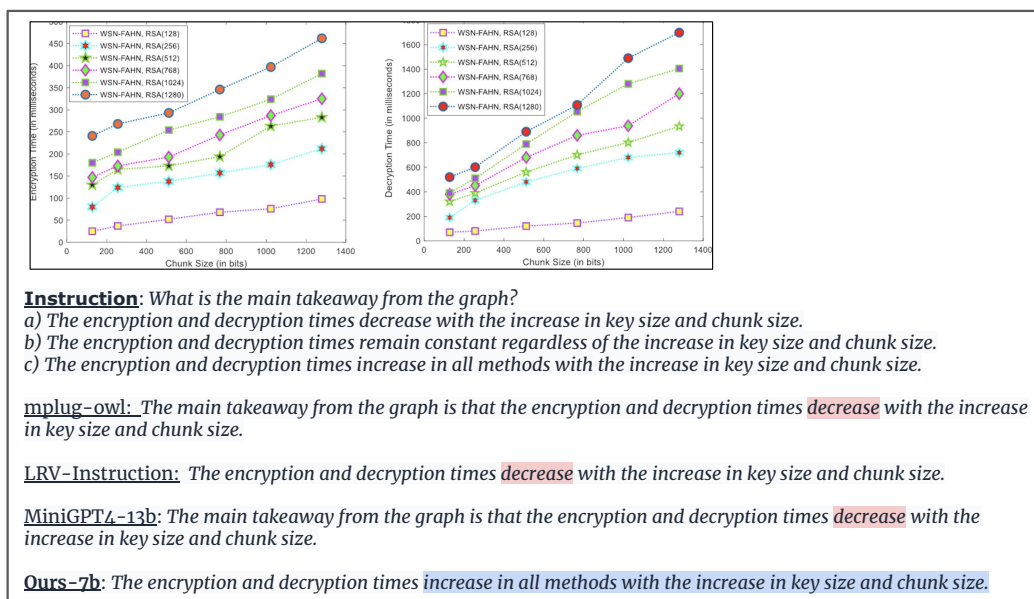
(c) Examples of *Multiple Chart understanding* task results from our model and other open-source models.

Figure 27: Result examples of our model and other open-source models for three types tasks in *MMC-Benchmark*. RED means incorrect answers, and BLUE means correct answers.

Instruction:
What does the graph compare in terms of packet reception ratio for critical packets?

Answers from different models:

mplug-owl:
The graph compares the packet reception ratio for critical packets in different scenarios. The scenarios are represented by different colored lines, with each line indicating the performance of a particular combination of network conditions.

LRV-Instruction: *Four different packets*

MiniGPT4-13b:
The graph compares the packet reception ratio for critical packets in two different scenarios. The first scenario is a wireless network with a single access point, while the second scenario is a wireless network with multiple access points.

Ours-7b:
The graph compares the performance of four different algorithms in terms of critical packet reception rate.
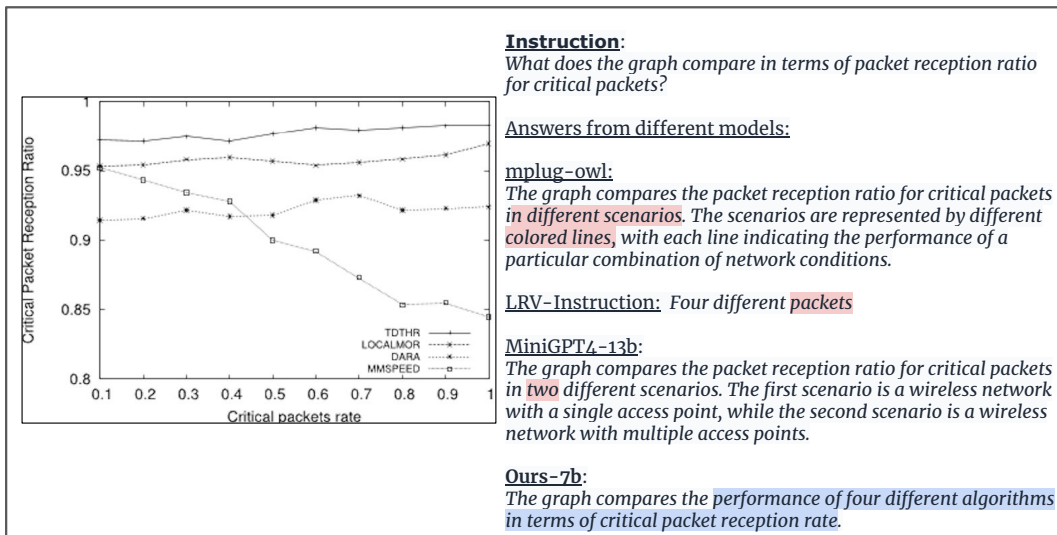
Figure 28: Examples of *Scientific Chart Understanding* task results from our model and other open-source models in *MMC-Benchmark*. RED means incorrect answers, and BLUE means correct answers.