Chart Understanding with Large Language Model

Yaser James, Will Li, John Feng

1 Abstract

Chart visualization plays a pivotal role in conveying information efficiently, enabling humans to rapidly discern critical insights from well-crafted visuals. Yet, for complex charts, such as stock market trends [7], or those aiding understanding for low vision users [38], drawing meaningful interpretations—especially for novices—can be particularly challenging. While humans may struggle to assimilate historical patterns at a glance, large multimodal models (LMMs), trained on extensive text corpora and charts, have the potential to provide enhanced guidance. Leveraging these models can empower users to understand complex chart patterns with deeper contextual insights. While successful in the general domains, such open-source LMMs [39, 17, 37, 14, 6] are less effective for chart images because chart image understanding is tremendously different from natural scene image understanding. Contrasting with natural scene images, which primarily contain objects and reflect their spatial relationships, chart images contain unique abstract elements (such as flow diagrams, trend lines, color-coded legends, etc.) that convey specific data-related information. The behind of the their poor performance is because of lacking domain-specific training essential for chart understanding. In this project, we introduce a baseline multimodal model that integrates text and charts to enhance the chart comprehension capabilities of existing models, offering more pertinent insights and information related to the depicted charts. The input of our model will be the chart images and human questions. The output will be the answers generated by our model.

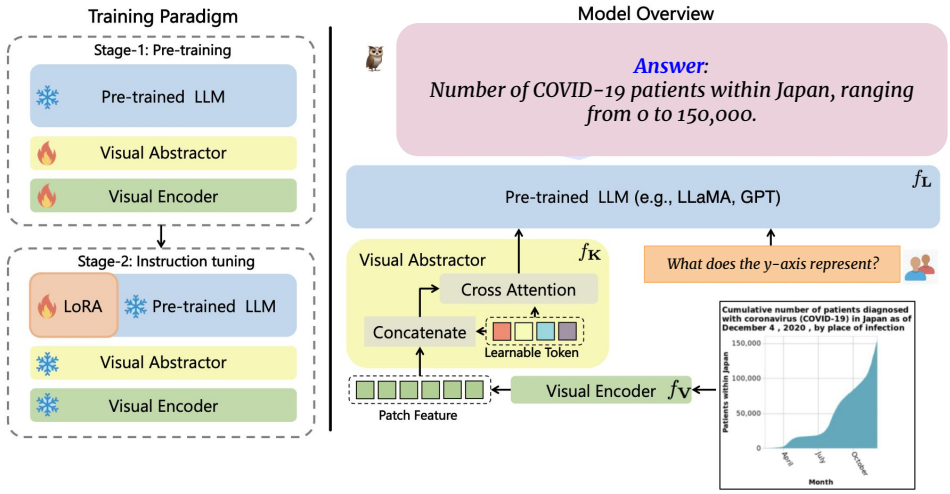


Figure 1: (Left) The training paradigm of our model. (Right) The architecture of our model. The input of our model will be the chart images and human questions. The output will be the answers generated by our model.

Preprint. Under review.

2 Introduction

Recent advancements in Large Language Models (LLMs) such as GPT-3, PaLM, ChatGPT, Bard, and LLaMA have demonstrated their remarkable ability to perform a wide array of tasks effectively. These developments have been pivotal in the field of natural language processing and understanding, as evidenced by works like [1], [4], [25], [18], and [30]. To augment these LLMs with visual processing capabilities, a new wave of open-source Large Multimodal Models (LMMs) has emerged. Models such as MiniGPT-4 [39], LLaVA [17], mPLUG-Owl [37], Multimodal-GPT [6], and LRV [14] have integrated sophisticated image understanding features into LLMs, enabling them to interpret and analyze visual data.

Despite their success in general domains, these open-source LMMs encounter specific challenges when applied to chart images. Chart image understanding significantly differs from natural scene image analysis. While the latter revolves around identifying objects and their spatial relationships, chart images comprise unique abstract elements like flow diagrams, trend lines, and color-coded legends that represent complex data-centric information.

The current generation of open-source LMMs often struggles to accurately interpret these intricate elements in chart images. This limitation stems primarily from a lack of domain-specific training, which is crucial for distinguishing between graph types, decoding axis labels and data points, and extracting relevant patterns and trends. Enhancing LMMs with advanced chart understanding capabilities would significantly improve their analytical and reasoning skills regarding chart information. Such an improvement would extend their utility across various sectors, including data analytics, academic research, and business intelligence, where precise interpretation of chart images is paramount. Our project involves the construction of chart comprehension question answering pairs and training of baseline chart understanding LMM.

3 Related Work

Large Multimdoal Models. Large Language Models (LLMs) excel in zero-shot tasks across multiple domains. Recent studies have ventured into using LLMs for multi-modal generation. One direction [32, 34, 33] uses ChatGPT as the intermediaries to choose the best tools or experts for visual interpretation according to users' inquiry. Another direction is end-to-end training [39, 17, 14, 37] utilize LLM and vision encoders to create integrated models for multi-modal tasks with few interconnected parameters to connect them. These works perform well on general visual and language tasks like image captioning, visual question answering with strong language skills. Yet, when it comes to intricate chart understanding, they often fall short due to a lack of specific training, struggling with diverse images, intricate text, and the connection between visual and textual content. Our work enhances visual chart comprehension by introducing a novel chart visual instruction-following corpus and chart understanding model.

Visualization/User Interaction. Initial advancements in creating natural language summaries were guided by planning-focused methods, encapsulated in Reiter's framework for converting data into text [26]. An illustration of this is the work by [24], who developed a system for generating captions. This system establishes caption structure by correlating data elements with visual components of charts, like text and graphical elements. It then employs a complexity metric to choose specific caption details and utilizes a textual planning tool for chart description. Similarly, the iGRAPH-Lite system [5] is designed to make charts comprehensible for visually impaired users through generated captions and chart navigation using a keyboard. Like the [24]. system, it relies on templates for concise chart descriptions. However, both systems focus more on chart interpretation instructions rather than delving into the broader insights that the charts communicate.

There has also been growing interest to automatically extract insights from a dataset and present them. [10, 20, 12] train a high-resolution image encoder on a large image-text pair corpus to learn text recognition during pretraining stage. However, these model rely on specific finetuning on different downstream datasets and can't achieve open-domain instruction understanding performance like Multimodal Large Language Models. Earlier datasets such as [9, 2, 23] primarily relied on synthetic data, with template-generated questions and answers selected from a fixed vocabulary. More recently, ChartQA [21] capitalized on real-world, web-crawled charts to develop its visual question-answering datasets, supplemented by human annotators. However, it mainly focus on compositional and visual

questions. [13] used Palm-2 to create question-answering data of academic charts. However, there exist hallucinations in their answers. The advantages of our dataset come from its larger size, more diverse topics, more rich language styles and good quality.

Visual Language Understanding/Low Vision Users Recent advancements in vision-language models have shown promising potential in enhancing the experiences of visually impaired individuals. The development of GPT-4V(ision) and Text-to-Speech (TTS) APIs by OpenAI has been a significant step in this direction, offering new possibilities for empowering the visually impaired through AI technologies [27]. In the realm of healthcare, AI and deep learning have been instrumental in diagnosing eye diseases, thereby aiding in the management of visual impairments [31]. Moreover, the research on assistive navigation systems like "SeeWay" emphasizes the significance of such technologies in improving the mobility and safety of blind or visually impaired individuals, which can have a profound impact on their personal and professional lives [36]. Comprehensive overviews and reviews, such as the one provided in ScienceDirect, highlight the range of current approaches and technologies aimed at assisting visually impaired individuals, underlining the diverse applications of vision-language models [19]. Additionally, the integration of visual features with language models, as seen in studies like "MiniGPT-4," demonstrates the continuous evolution of these technologies in enhancing vision-language understanding, further contributing to the field of assistive technology for the visually impaired [40]. Going beyond previous work, our baseline model enhances chart comprehension experience for low-vision users.

4 Methodology

Large Language models (LLMs) like ChatGPT, Bard, and LLaMA [30] have advanced quickly, showcasing notable zero-shot skills in various linguistic contexts. By incorporating the LLM as the language decoder, the resulting large multimodal models (LMMs) such as MiniGPT-4 [39], LLaVA [17], mPLUG-Owl [37] and other works [14, 6] exhibit strong zero-shot task completion performance on a variety of user-oriented vision-language tasks such as image understanding and reasoning. While successful in the general domains, such LMMs are less effective for chart images because chart image-text pairs are drastically different from general web content. Unlike general images, which are interpreted based on objects, positions, and interactions, chart images carry distinct abstract components such as trend lines and color-coded legends. Therefore, due to lacking specific training, current LMMs still face the challenge of comprehending intricate contents in charts, which requires strong OCR and complex reasoning skills.

Architecture. Our project is to build a Multimodal Chart Assistant (Mplug-Owl [37] as the backbone) with the help of current LLMs [30], which can not only comprehend human intents very clearly but also generate correct answers based on the chart. Specifically, as illustrated in Fig. 1, our model consists of a vision foundation model (CLIP vision encoder) f_V to encode the visual knowledge, a language foundation model (Vicuna [3]) f_L, and a visual abstractor module f_K. We first obtain dense image representations from the pre-trained visual foundation model f_V. However, such dense features would fragment the fine-grained image information and bring large computation due to the lengthy sequence when feeding into f_L. To mitigate this issue, we employ the visual abstractor module f_K to summarize visual information within several learnable tokens, thereby obtaining higher semantic visual representations and reducing computation, as illustrated in Fig. 1. The visual representations are combined with text queries and fed into the language model to generate the response.

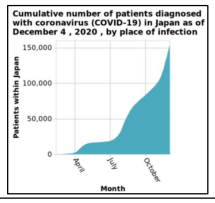
Training Paradigm. Previous approaches [17, 39] have employed a limited number of additional parameters to learn the alignment between visual data and language models for pretraining, constraining their capacity to comprehend complex visual information. To enhance the ability of large-scale language models to perceive visual information while integrating their internal abilities, we propose a novel paradigm as our first stage of pretraining that incorporates a trainable visual backbone f_V and an additional visual abstractor f_K, while maintaining the pre-trained language model f_L in a frozen state. This approach enables the model to effectively capture both low-level and higher semantic visual information and align it with the pre-trained language model without compromising its performance.

Upon completion of the prior phase, the model acquires the ability to retain a considerable amount of knowledge and provide reasonable answers to human queries. In the second stage of our training paradigm, given that the model can comprehend the visual concepts and knowledge depicted in

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.

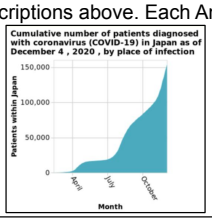
Figure 2: An example prompt for text-only GPT4 we use to generate instruction and answers for *Chart Information Extraction* task. The sentence in BLUE is the captions of the chart.

images through visual knowledge learning, we freeze the entire model and employ low-rank adaption [8] to adapt f_L by training multiple low-rank matrices for efficient alignment with human instructions. For each data record, we compute the loss on the response. During the training, we accumulate the gradient for text-only instruction data and multi-modal instruction data for multiple batches and updated the parameters. Therefore, by joint training with both language and multi-modal instructions, our model can better understand a wide range of instructions and respond with more natural and reliable output. In this stage, we use our chart instruction-following data for the training.

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:



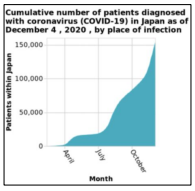
question1=>
answer1 =>
question2=>
answer2 =>
question3=>
answer3=>

GPT4 OUTPUT Example:

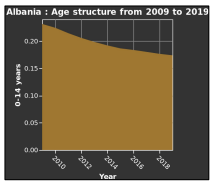
question1=> When was the first COVID-19 case diagnosed in Japan?
answer1 => March 2020.
question2=> How many COVID-19 cases were reported in Japan by December 4th, 2020?
answer2 => Approximately 160,000.
question3=> When did the greatest increase in COVID-19 cases in Japan occur?
answer3=> Between November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

Figure 3: An example prompt for text-only GPT4 we use to generate instruction and answers for *Chart Reasoning* task. The sentence in BLUE is the captions of the chart.

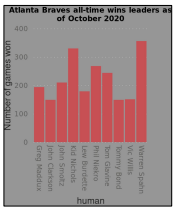
Training Objective. The training of the model revolves around the language modeling task. This involves teaching the model to produce following tokens by understanding the context that comes before them. The main goal during training is to increase the probability (log-likelihood) of these tokens being correct. It's essential to understand that the training loss calculation only includes discrete tokens, like those found in text.



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

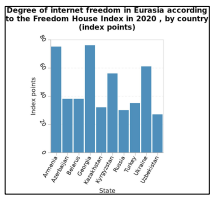


Question:
What is the name of the area diagram?
Answer:
Albania: Age Structure from 2009 to 2019.

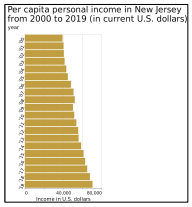


Question:
How many games did Warren Spahn win?
Choices: A) 250 games B) 350 games?
Answer:
B) 350 games.

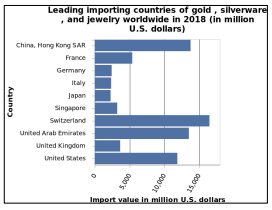
Figure 4: Examples of the *Chart Information Extraction Task*



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.

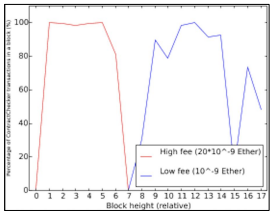


Question:
What has the per capita income in New Jersey shown from 2000 to 2019?
Answer:
A steady upward trend.

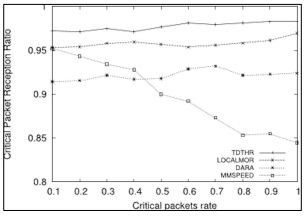


Question:
Which country is the top importer among Switzerland, UAE, and China?
Answer:
Switzerland

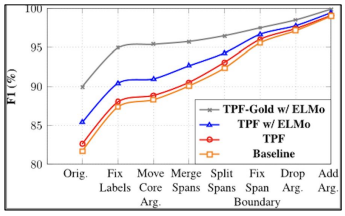
Figure 5: Examples of the *Chart Reasoning Task*



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 6: Examples of the *Arxiv Chart Understanding Task*

5 Datasets

This section introduces the construction of our instruction tuning data in detail. 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 shown in Figure **??**. 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. 2, 3. Our instruction-tuning data is formatted as: "Human: {question} AI: {answer}". Our dataset includes the following tasks: *chart information extraction* (Fig. 4), *chart reasoning* (Fig. 5), *scientific chart understanding* (Fig. 6). Given we use the chart description as the

input for GPT4, we have two sources for the chart descriptions: *Vistext* [29] and *MMC-Instruction* [15]. Finally, our dataset contains 100k chart question-answer pairs. The statistic is shown in Tab. 1.

Datasets	Question	Answer Type	Plot Type	Task Num
FigureQA	Template	Fixed Vocab	4	1
DVQA	Template	Fixed Vocab	1	1
PlotQA	Template	Fixed	1	1
ChartQA	Free-form	Open Vocab	Unbounded	2
SciGraphQA	Free-form	Open Ended	Unbounded	2
Ours	Free-form	Open Ended	Unbounded	3

Table 1: Comparison between our dataset with existing chart-based question-answering datasets.

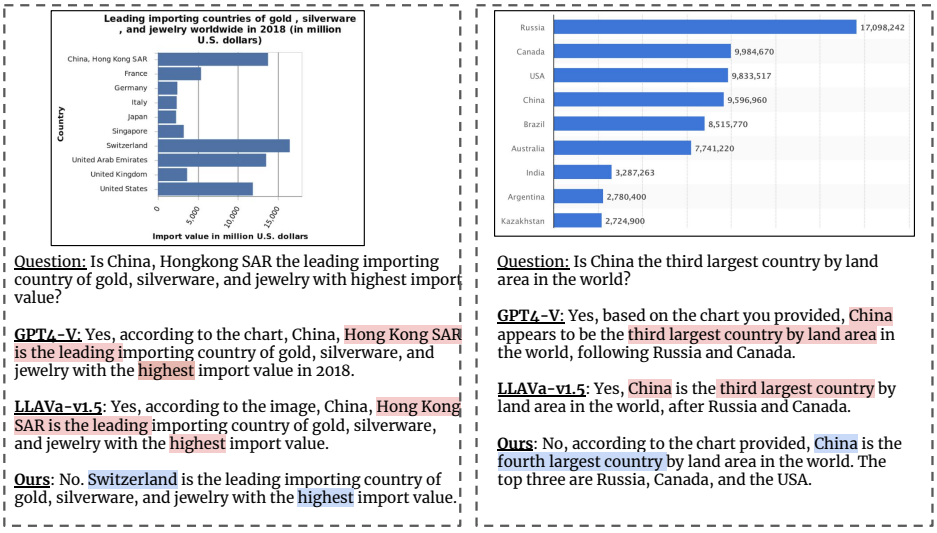


Figure 7: Failure examples of GPT4-V and LLAVa-v1.5. RED highlights the hallucinations from GPT-4V and LLAVa-v1.5. BLUE highlights the correct answers from our model.

6 Experiments

6.1 Experimental Setup

Our model is trained with 8 Nvidia Tesla V100 GPUs. We compare *MMCA* with existing models in three setups: (a) Open-source LMMs including MiniGPT4-7B [39], LRV-Instruction [14], LLaVA-7B [16] and Mplug-Owl-7B [37]. (b) Non-LLMs based models including Pix2Struct [12] and Donut [11].

6.2 Evaluation Datasets

We compare our model with baselines on several existing benchmarks in the zero-shot setting. (1) *MMC-Benchmark* [15], which utilizes ChatGPT to compare between reference answers and generated texts. In details, we provide ChatGPT with the question, the reference answer, and the model prediction and ask ChatGPT to assess the accuracy of the model prediction. One example is shown in Fig. 8 (2) ChartQA [21], DocVQA [22], and TextVQA [35], which use *Related-Accuracy* to evaluate the generated answers.

6.3 Results and Discussion

As indicated in Tab. 2 and Tab. 3, our model achieves better performance in in comparison with the existing open-source models. The improvement of our model demonstrates the effectiveness of our chart instruction tuning data in enabling the LMM to complete chart understanding tasks. What's

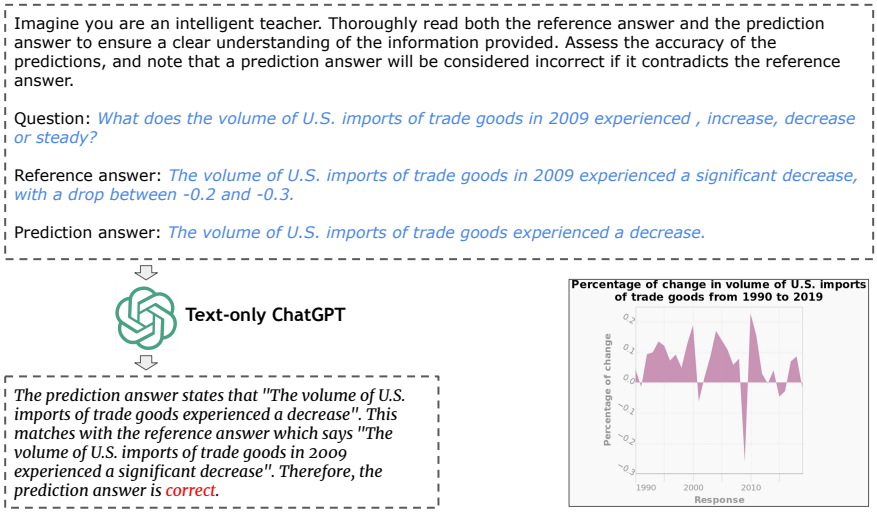


Figure 8: An example of using *text-only ChatGPT* for evaluation in the *MMC-Benchmark*. In the prompt, we provide *ChatGPT* with the question, reference answer, and predictions from models.

Free-form Evaluation	LLAVA	MiniGPT4	mPLUG-owl	LRV-Instruct	Ours
Chart Information Extraction	0.29	0.23	0.27	0.24	**0.35**
Chart Reasoning	0.28	0.20	0.22	0.19	**0.30**
Scientific Chart Understanding	0.30	0.21	0.28	0.23	**0.33**

Table 2: *MMC-Benchmark* evaluation results on LLaVA, MiniGPT4, mPLUG-Owl, LRC-Instruct and our model.

more, such results validate that with the help of open-domain large multi-modal models (LMMs), the performance of chart understanding can be significantly improved. 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. The weak performance of MiniGPT4 can be attributed to its limited instruction-tuning data with a single task in a small size. Though fine-tuned with instruction-tuning data from text-rich images, LLAVa 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. One visualization example is shown in Fig. 7.

7 Goal and Related Users

Our goal will be the following: (1) Introduce a comprehensive benchmark dataset encompassing charts paired with their relevant contexts, questions, and answers, designed specifically for fine-tuning vision-language (VL) models. (2) Deploy and assess a VL model, fine-tuned using our dataset, validated through human evaluations. (3) Develop an interactive online platform that seamlessly accepts charts/texts as inputs and delivers charts/texts as outputs, leveraging our trained model.

Individuals with Vision Impairments: Our chart-VL model offers a valuable tool for individuals with vision impairments who require assistance in interpreting charts or specific charts designed for visually impaired [28]. These users often face challenges in interpreting visual data representations, such as charts and images. Our tool can make data visualizations more accessible to them by translating visual data into descriptive insights that has the following purposes: **Increased Accessibility**: For the visually impaired, this tool offers a way to access and comprehend visual data that might otherwise be inaccessible. **Enhanced Data Comprehension**: The tool assists in providing a deeper understanding of charts and visual data, making complex data more digestible. **Efficient Data**

Model	ChartQA	DocVQA	TextVQA
Donut	41.8	67.5	43.5
Pix2Struct	56.0	72.1	-
MiniGPT4	46.5	42.3	-
LRV-Instruction	47.5	58.5	-
Ours	**57.4**	**72.5**	**57.6**

Table 3: Comparison with OCR-free methods and LMMs on existing public benchmarks.

Analysis: Instead of spending hours trying to discern patterns or insights from charts, users can utilize the tool to get quick interpretations.

8 Conclusion

This paper aims to tackle the challenge of chart understanding with Large Multimodal Models (LMMs). Firstly, we present a novel and large- scale chart instruction-tuning dataset called that includes diverse topics, language styles, chart types, and open-ended answers in line with human cognition. Second, we propose an instruction-tuned LMM called MMCA that outperforms existing open-source state-of-the-art (SoTA) methods on both well-established chart understanding datasets. Though we believe MMCA represents a significant step toward building a useful multimodal chart assistant, we notice MMCA is limited by a weak vision encoder and preliminary performance on chart-to-datatable and chart-to-json tasks. Future work is directed toward overcoming the challenges with advanced visual encoding and modeling.

References

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

[2] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 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, 2020.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

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

[5] Leo Ferres, Gitte Lindgaard, Livia Sumegi, and Bruce Tsuji. Evaluating a tool for improving accessibility to charts and graphs. *ACM Trans. Comput. Hum. Interact.*, 20:28:1–28:32, 2013.

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

[7] Trang-Thi Ho and Yennun Huang. Stock price movement prediction using sentiment analysis and candlestick chart representation. *Sensors*, 21(23), 2021.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

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

[11] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.

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

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

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

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

[16] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*, 2020.

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

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

[19] Maisha Mashiata, Tasmia Ali, Prangon Das, Zinat Tasneem, Md. Faisal Rahman Badal, Subrata Kumar Sarker, Md. Mehedi Hasan, Sarafat Hussain Abhi, Md. Robiul Islam, Md. Firoj Ali, Md. Hafiz Ahamed, Md. Manirul Islam, and Sajal Kumar Das. Towards assisting visually impaired individuals: A review on current status and future prospects. *Biosensors and Bioelectronics: X*, 12:100265, 2022.

[20] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023.

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

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

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

[24] Vibhu Mittal, Johanna D. Moore, Giuseppe Carenini, and Steven F. Roth. Describing complex charts in natural language: A caption generation system. *Comput. Linguistics*, 24:431–467, 1998.

[25] OpenAI. Introducing chatgpt. 2022.

[26] Ehud Reiter. An architecture for data-to-text systems. In *European Workshop on Natural Language Generation*, 2007.

[27] Luis Roque. Seeing with sound: Empowering the visually impaired with gpt-4v(ision) and text-to-speech. *Towards Data Science*, November 2023. Accessed: Date of Access.

[28] Ana Paula Scariot, Luis G. Montané-Jiménez, and Carmen Mezura-Godoy. Towards the creation of accessible charts for the visually impaired. In *Proceedings of the XXI International Conference on Human Computer Interaction*, Interacción '21, New York, NY, USA, 2021. Association for Computing Machinery.

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

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

[31] Jiaji Wang, Shuihua Wang, and Yudong Zhang. Artificial intelligence for visually impaired. *Displays*, 02 2023.

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

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

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

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

[36] Zongming Yang, Liang Yang, Liren Kong, Ailin Wei, Jesse Leaman, Johnell Brooks, and Bing Li. Seeway: Vision-language assistive navigation for the visually impaired. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 52–58, 2022.

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

[38] Wai Yu and Stephen Brewster. Evaluation of multimodal graphs for blind people. *Universal Access in the Information Society*, 2:105–124, 06 2003.

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

[40] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 04 2023.