**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**
**TRƯỜNG ĐẠI HỌC BÁCH KHOA**
**KHOA KHOA HỌC VÀ KĨ THUẬT MÁY TÍNH**

# BÁO CÁO

## Đồ án Tổng hợp - Hướng trí tuệ nhân tạo (CO3101)
## Topic: Chart Visual Question and Anwering

NGÀNH: KHOA HỌC MÁY TÍNH

**GVHD:** ThS. Mai Xuân Toàn
TS. Trần Tuấn Anh
ThS. Trần Hồng Tài

**—o0o—**

**SVTH:** Lê Hoàng Khánh Vinh - 2213963
Cao Nguyễn Văn Trường - 2213732
Trần Đại Việt - 2213951

TP. HỒ CHÍ MINH, 12/2024

# Mục lục

# Acknowledgement

First of all, I would like to acknowledge Mr. Mai Xuan Toan, Mr. Tran Tuan Anh, Mr. Tran Hong Tai and the groups in the project class who have guided and supported us throughout the process of researching and implementing this project. In addition, we would like to extend our sincere thanks to the various research groups and organizations whose prior studies and developments on this topic have laid the groundwork for our article. Their contributions have served as the solid foundation upon which we have built our research, and we greatly appreciate their efforts in advancing this field.

Although there have been many efforts to complete the topic as completely as possible. However, due to limited time, low level of knowledge and awareness, the project cannot avoid shortcomings, I really hope to receive comments.

## Disclaimer

We hereby affirm that the entire content of this research project is the result of our serious and honest work under the careful guidance of our supervising professors. All information and reference materials used have been properly and accurately cited in accordance with the regulations.

We fully take responsibility for the truthfulness and accuracy of the contents in this project.

# Abstract

This work focuses on enhancing the capabilities of the UniChart model for Chart Visual Question Answering (Chart VQA) by integrating three key components: Data Extraction, Data Estimation, and Question Answering with Rationale Generation. First, we design a model to extract structured data from charts, generating a JSON representation that classifies visual elements (e.g., titles, colors) and reconstructs the underlying data table, inspired by DePlot. The goal is to improve the accuracy of extracted data, addressing common errors in numerical reasoning tasks caused by faulty data extraction. Second, we extend this approach with Data Estimation, training the model to predict data tables when chart elements are partially masked. This enables the model to approximate missing values, enhancing its robustness. Finally, for the Question Answering (QA) and Rationale task, we incorporate a Chain-of-Thoughts-inspired method to generate both answers and explanatory rationales, ensuring accuracy in multi-step reasoning for visual (e.g., color, trends) and numerical (e.g., calculations, comparisons) questions. Additionally, we explore a parameter-efficient fine-tuning (PEFT) strategy, freezing parameters after effective data extraction training and introducing new parameters for subsequent tasks. This method keeps the total trainable parameters consistent while scaling model capacity, potentially improving performance across tasks. Our approach aims to achieve a highly accurate and interpretable Chart VQA system.

# 1 Introduction

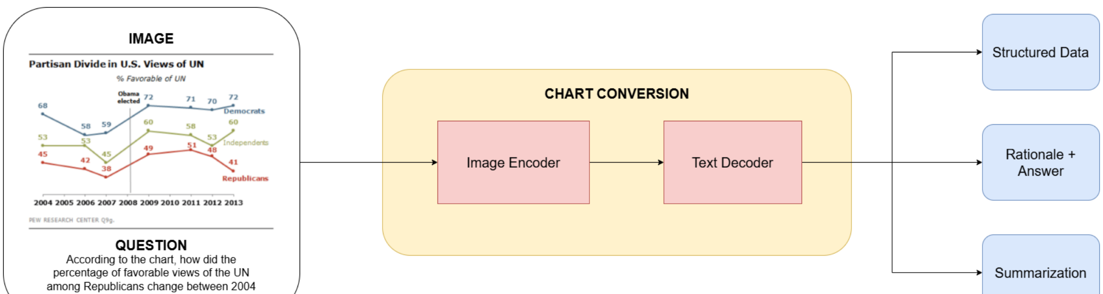
Multimodal reasoning, particularly on visual language such as plots and charts, is a highly complex task. For downstream tasks like question answering (QA) on plots and charts, a model must first extract relevant information from the image, organize it effectively, and then perform reasoning over the extracted entries. Previous studies have proposed end-to-end solutions for these tasks ([12], [15]). While these methods are effective, they require fine-tuning on large amounts of task-specific data and still struggle with queries that demand complex reasoning, even after fine-tuning. For instance, the MATCHA model ([15]) achieves only 38.2% accuracy on the ChartQA benchmark ([20]) for human-written queries.

Recent advances in vision-language models (VLMs) have largely stemmed from techniques that enhance their ability to represent visual information ([3]; [12]). These improvements allow the models to understand key visual elements, a crucial foundation for basic reasoning tasks. However, when it comes to more complex reasoning, which requires integrating the core image representation with the semantic understanding of a question to generate an answer, VLMs still face significant limitations. Many models struggle to effectively combine image and text representations in a contextual manner.

In the realm of large language models (LLMs), one notable approach to improving reasoning capabilities involves in-context learning, which encourages reasoning through techniques like chain-of-thought prompting ([24]), task decomposition ([26]), or embedding stored knowledge in model weights ([21]). Furthermore, fine-tuning on datasets that include rationales ([18]; [8]) has proven particularly effective for smaller models, enabling them to perform better on complex reasoning tasks.

Nowadays, researchers are also finding it more efficient to first train the model on extraction tasks like chart-to-table ([1], [14]) or chart-to-json ([16]), which are currently state-of-the-art models in Chart-QA benchmarks ([20]) with the best accuracy of 81.32% belongs to Chart-PaLI5B [1].

Inspired by all the methods above, especially Chart-PaLI5B ([1]), we aim to improve the accuracy of UniChart ([19]), a model for ChartVQA which was pre-trained in the same recipe and architecture as Donut ([11]). This work focuses on enhancing the capabilities of the UniChart ([19]) model for Chart Visual Question Answering (Chart VQA) by integrating three key components: Data Extraction, Data Estimation, and Question Answering with Rationale Generation. First, we design a model to extract structured data from charts, generating a JSON representation that classifies visual elements (e.g., titles, colors) and reconstructs the underlying data table, inspired by DePlot ([14]). The goal is to improve the accuracy of extracted data, addressing common errors in numerical reasoning tasks caused by faulty data extraction. Second, we extend this approach with Data Estimation, training the model to predict data tables when chart elements are partially masked. This enables the model to approximate missing values, enhancing its robustness. Finally, for the Question Answering (QA) and Rationale task, we incorporate a Chain-of-Thoughts-inspired method from [8], [18] and [1] to generate both answers and explanatory rationales, ensuring accuracy in multi-step reasoning for visual (e.g., color, trends) and numerical (e.g., calculations, comparisons) questions. Additionally, we explore a parameter-efficient fine-tuning ([5]) strategy, freezing parameters after effective data extraction training and introducing new parameters for subsequent tasks. This method keeps the total trainable parameters consistent while scaling model capacity, potentially improving performance across tasks. Our approach aims to achieve a highly accurate and interpretable Chart VQA system.




**Hình 1:** *Architecture.*

# 2 Dataset

## 2.1 Brief description of ChartQA

ChartQA is a widely used visual question-answering benchmark designed to evaluate the reasoning capabilities of vision-language models (VLMs).

The benchmark comprises two main components: (a) Human Set: Questions written by humans that require complex reasoning. (b) Augmented Set: Simpler, machine-generated questions designed to complement the human set.

The dataset includes charts sourced from four repositories: Statista, Pew, Our World in Data, and OECD. For three of these sources, ground-truth tables are provided. However, for charts from Pew, the corresponding tables are inferred using the ChartOCR model ([17]).

## 2.2 Data generation methods

### 2.2.1 Data table generation

We will first generate the underlying data table for each chart image using the `DePlot` model [14], which is currently available on Hugging Face. The `RNSS` and `RMS` metrics of `DePlot` are very high, with values of 97.1% and 94.2%, respectively. Given these results, we use `DePlot` as a teacher model for `UniChart` to perform knowledge distillation [7], which we predict will be more effective than constructing a dataset in the traditional manner.

Similar to the data mixture introduced in [14], the data source will be mainly real-world chart images from ChartQA dataset ([20]), chart-to-table pairs from DVQA ([10]), chart-to-table pairs from the train set of TaTa ([4]) and real-world charts with tables from `Wikipedia`.

### 2.2.2 QA + Rationale generation

From the charts and their corresponding underlying data tables, we will create tuples of the form (image, table) and input them into a large language model (LLM), such as GPT-4o, to generate question-answer (QA) pairs along with their associated rationales. The QA pairs will be categorized into two types:

- **Visual reasoning:** These questions will focus on the visual elements of the chart. They will range from simpler queries, such as identifying colors, chart types (e.g., bar, line), and layouts (e.g., vertical, horizontal), to more complex ones that require reasoning, such as recognizing trends (e.g., increasing, fluctuating).

- **Numeric reasoning:** The goal is to enhance the model's mathematical reasoning capabilities. This includes basic questions involving comparisons, minimum/maximum values, averages, and so on. Additionally, the model will tackle more advanced tasks, such as answering multi-step questions and addressing multiple questions within a single query, demonstrating the effectiveness of generating rationales.

### 2.2.3 Summarization generation

Image captioning is a fundamental task in AI, where machines are required to summarize the key content of an image in a textual format. This problem has been extensively explored in the literature ([22]; [6]; [9]; [13]). Building on previous work ([22]; [25]; [19]), we pretrain our model on this task to further improve its ability to generate accurate textual descriptions from chart images. Additionally, the summarization will be generated using a large language model (LLM), such as GPT-4o, like rationale generation.

# 3 Method

## 3.1 Data Table Generation

To improve the model's ability to accurately interpret and extract information from charts, we introduce a **Data Extraction** task as an initial training step. This task focuses on converting the visual information encoded in the chart into a structured, machine-readable format. For this purpose, we choose the **JSON format**, which provides a flexible and intuitive way to represent both the visual elements of the chart and the underlying data table. The visual elements may include components such as colors, axes, legends, labels, and data points, while the data table captures the structured numerical or categorical data represented by the chart. By requiring the model to extract and organize this information, we aim to create a detailed representation that bridges the gap between visual and textual modalities. This structured output not only makes the data easier to interpret but also serves as a critical foundation for downstream tasks, such as answering questions or generating rationales. Training the model on this task is predicted to significantly enhance its capability to extract and process chart data, which is the most essential step in achieving accurate and reliable results. Moreover, this approach ensures that even complex visual charts can be broken down into an interpretable format, paving the way for a more robust and generalizable performance across a variety of chart types.

## 3.2 Data Estimation

One of the unique challenges in **ChartVQA** compared to tasks like **DocVQA** is that not all the information required to answer questions is explicitly presented in the chart. In many cases, the model must infer or estimate missing values by analyzing the relationships between the visible elements, such as the spacing between bars, the scale of the axes, or other contextual clues. To address this complexity, we are developing a specialized dataset designed to improve the model's ability to perform such estimations accurately. This dataset builds upon the framework of the **Data Table Generation** task but introduces an additional layer of difficulty. Specifically, certain values on the chart are deliberately masked, and the model is tasked with predicting the masked values while generating the complete data table. By training the model on this augmented dataset, it learns to estimate missing values and infer relationships between visible data points, enhancing its reasoning and extrapolation capabilities. This method is critical for improving the model's performance on real-world scenarios, where partial or incomplete data often requires intelligent estimation for accurate interpretation.


**Hình 2:** *An example which requires the model to estimate the values.*

## 3.3 Visual/Numeric Reasoning with Rationale

After extracting the data, the model is tasked with performing mathematical operations on it, such as calculating the minimum, maximum, mean, or median values, as well as comparing data points. Additionally, the model needs to address challenges posed by visual elements, including bars, lines, layouts, or colors, which often play a crucial role in interpreting charts. To enhance the model's performance, each answer is accompanied by a detailed **rationale**, outlining the step-by-step reasoning process the model follows to arrive at the solution. This rationale generation is inspired by the **Chain-of-Thoughts (CoT)** technique ([24]), a powerful approach that improves both accuracy and reliability by encouraging the model to think systematically, reducing instances of "memory loss", where the model forgot the answer it calculated at the very steps before the current step, or hallucinated answers.

While CoT has been successfully demonstrated in large language models (LLMs) through few-shot prompting ([14], [16]), these models quickly learn reasoning patterns with minimal examples due to their vast pretraining. However, our model requires a more rigorous approach to achieve similar reasoning capabilities. Therefore, we adopt a **fully-supervised training setup** ([1]), incorporating the rationale directly into the target output alongside the answer. This approach, known as a **multi-task** setup, trains the model to simultaneously generate the answer and its accompanying rationale. The output is formatted as follows:

- **Original output**: "Answer: ..."
- **Multi-task output**: "Answer: ... Rationale: ..."

In this task, we plan to adopt a method called **Parameter-Efficient Fine-Tuning (PEFT)** ([5]), which was originally designed for fine-tuning, to train our model. Specifically, we will freeze all existing parameters of the model to preserve its performance on the data table generation task. To enhance the model's capabilities, we will expand its size by adding a suitable number of new parameters, enabling it to train effectively on the combined answer and rationale generation task. This approach not only increases the model's capability by increasing its size but also minimizes the risk of performance degradation in the data table generation task after completing the answer and rationale training. PEFT has demonstrated significant success in fine-tuning scenarios, and we aim to evaluate its effectiveness in the pretraining phase as well.

### 3.4 Summarization

Generating structured JSON data and detailed rationales are crucial foundational tasks that enable the model to perform effectively in summarization. In this task, the model is required to synthesize information from the chart and produce a coherent and concise paragraph that accurately summarizes the chart's content. This involves interpreting the extracted data, recognizing key patterns or trends, and integrating relevant insights into a human-readable summary. By training the model to summarize, we aim to enhance its ability to understand the overall context of the chart rather than focusing solely on individual data points. This task is critical for scenarios where users need quick, high-level insights into the chart's information. The summarization method leverages the outputs from previous tasks, such as JSON generation and rationales, to provide structured, context-aware, and accurate summaries, ensuring the model's performance remains reliable and interpretable across a range of chart types and complexities.

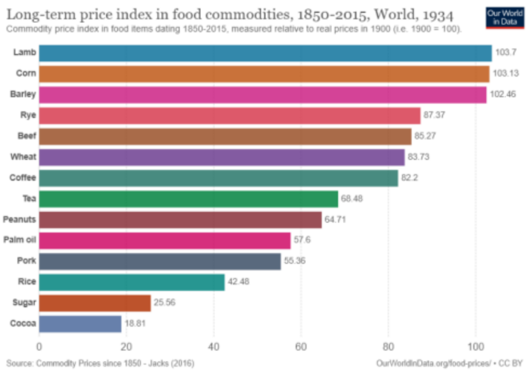
## 4 Experiment

In our experiments, we observed that our current model achieved only **34.48%** accuracy on the **ChartQA human dataset**, compared to the original UniChart ([19]) model's **44%** accuracy. A closer examination revealed that our model performed well on simple questions, such as those involving colors or basic mathematical operations, which do not require multi-step reasoning. However, the majority of incorrect responses stemmed from four key issues.

| | ChartQA human (Relaxed Accuracy) |
|---|---|
| UniChart (on paper) | 43.92 |
| UniChart (we tested ourselves) | 44 |
| UniChart + Rationale (ours experiment) | 34.48 |

**Hình 3:** *The accuracies of UniChart and Ours modified version of it.*

Firstly, the model exhibited a tendency to loop the generated rationale, repeatedly producing the same content until it exceeded the maximum output token limit. This behavior often resulted in no answer being generated, as seen in **134 questions** during our experiment. Some additional cases also displayed this looping issue but still managed to produce answers.

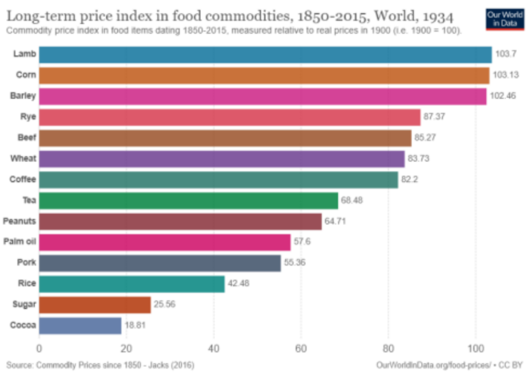



- Question: What is the difference in value between Lamb and Corn?
- Target: 0.57

**Rationale:** To find the difference in value between the largest and smallest values from the chart, follow these steps: 1. **Identify the Values**: - Lamb: 103.7 - Corn: 103.13 - Barley: 102.46 - Riye: 87.37 - Beef: 85.27 - Wheat: 83.73 - Coffee: 82.2 - Tea: 68.48 - Pearls: 57.6 - Pork: 55.36 - Rice: 42.48 - Sugar: 25.56 - Cocoa: 18.81 2. **Determine the Largest Value**: - Lamb: 103.7 - Corn: 103.13 - Barley: 102.46 - Riye: 87.37 - Beef: 85.27 - Wheat: 83.73 - Coffee: 82.2 - Tea: 68.48 - Pearls: 57.6 - Pork: 55.36 - Rice: 42.48 - Sugar: 25.56 - Cocoa: 18.81 3. **Determine the Largest Value**: - Lamb: 103.7 - Corn: 103.13 - Barley: 102.46 - Riye: 87.37 - Beef: 85.27 - Wheat: 83.73 - Coffee: 82.2 - Tea: 68.48 - Pearls: 57.6 - Pork: 55.36 - Rice: 42.48 - Sugar: 25.56 - Cocoa: 18.81 4. **Identify the Largest Value**: - Lamb: 103.7 - Corn: 103.13 - Barley: 102.46 - Riye: 87.37 - Beef: 85.27 - Wheat: 83.73 - Coffee: 82.2 - Tea: 68.48 - Pearls: 64.71 - Pork: 57.6 - Pork: 55.36 - Rice: 42.48 - Sugar: 25.56 - Cocoa: 18.81 5. **Calculate the Difference**: - To find the difference, subtract the smallest value from the largest value: \[ \text{Difference} = \text{Largest Value} - \text{Lowest Value} \] \[ \text{Difference} = 103.7 - \text{Maximum Value} \] \[ \] \[ \] \[ \] \[ \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \[ \] \approximately \( \( \( \quad \quad \[ \quad \text{( \text{( \)} = \) \]

**Hình 4:** *An example case about looping rationale.*

Secondly, the model frequently struggled to select the correct data in the rationale for inference, leading to confusion during reasoning. This issue was largely attributed to incorrect rationales generated during training using **GPT-4o-mini**, which impacted the quality of the model's reasoning process.

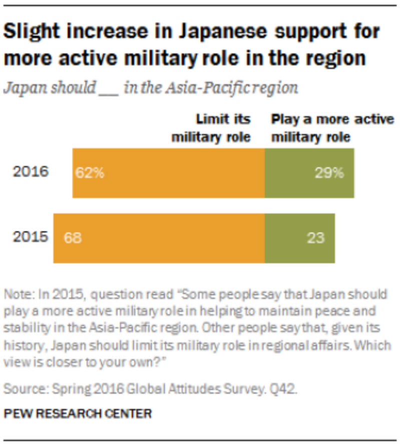



- Question: How many food item is shown in the bar graph?
- Target: 14

**Rationale:** To determine how many food items are shown in the bar graph, follow these steps:
1. **Identify the Graph Type**: The graph is a horizontal bar graph displaying the long-term price index in food commodities for various countries in the year 1934.
2. **Examine the Bars**: Look at the vertical axis, which lists the horizontal axis.
3. **Count the Bars**: - Lamb - Corn - Barley - Rye - Beef - Wheat - Coffee - Tea - Peanuts - Paim oil - Pork - Rice - Sugar - Coco
4. **Total Count**: Count the number of items that are shown in the bar graph.
5. **Conclusion**: There are a total of **7 food items** shown in the bar graph. Thus, the answer to the question "How many food item is?" is **7**.
**Answer:** 7

**Hình 5:** *The model's inference phase is correct but it failed at the conclusion step.*

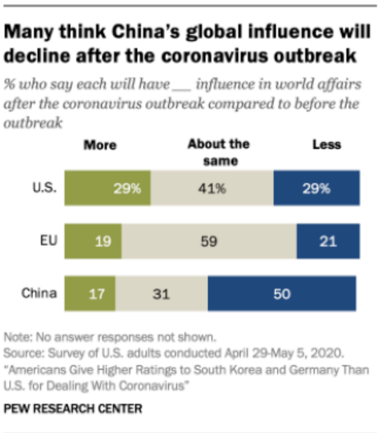
Thirdly, the model occasionally extracted either more or less data than required, or sometimes incorrect data, resulting in incorrect answers. This error also traces back to the flawed rationales produced by GPT-4o-mini during the training process, which propagated inaccuracies into the model's outputs.




- Question: What's the value of the lowest bar?
- Target: 23

**Rationale:** To determine the value of the lowest bar in the provided chart, follow these steps:
1. **Identify the Bars**: The chart has two bars representing two categories: "Limit its military role" and "Play a more active military role."
2. **Locate the Values for Each Year**: - **2016**: 62% - **2015**: 68% - **2016**: 29%
3. **Compare the Values**: Now, compare the two values: - 2016: 62% - 2015: 68%
4. **Determine the Lowest Value**: The lowest value among these is 29%.
5. **Conclusion**: The value of the lowest bar is **29%**.
**Answer:** 29

**Hình 6:** *The model extracted less data than expected.*




- Question: What's the average of all the values in the green bars (round to one decimal)?
- Label: 21.6

**Rationale:** To find the average of all the values for the green bars (representing "More," "About the same") from the provided chart, follow these steps:
1. **Identify the Values**: From the chart, extract the values for the green bars: - U.S.: 29% - EU: 19% - China: 17%
2. **Add the Values Together**: \[ 29\% + 19\% + 17\% + 29\% = 80\% \]
3. **Count the Number of Values**: There are 3 values.
4. **Calculate the Average**: \[ \text{Average} = \frac{\text{Total Sum}}{\text{Number of Values}} = \frac{80\%}{3} = 29\% \] Thus, the average of all the values in the green bars is **29%**.
**Answer:** 29

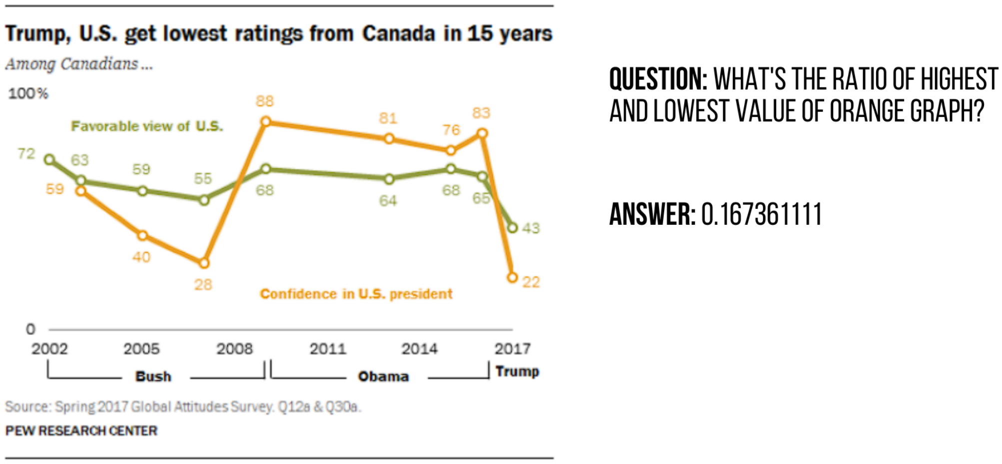
**Hình 7:** *The model extracted more data than expected.*

Finally, the fourth problem also originates from GPT-4o-mini's incorrect rationales. When provided with the tuple (image, question, answer) during training to generate rationales, GPT-4o-mini often produced outputs where the rationale did not align with the provided answer. In some cases, it explicitly included statements such as, *"There is something wrong in the provided tuple; please check the question and answer again,"* introducing abnormalities into the rationale. Consequently, the model learned these incorrect patterns during training, further degrading its ability to generate accurate outputs.

# 5 Limitations

While our approach demonstrates some promising results, it is not without limitations, which highlight areas for further investigation and improvement.

Firstly, our primary data source, the **ChartQA dataset** ([20]), contains several incorrect answers in both the human-annotated and augmented datasets. These errors introduce noise into the training process and can confuse the model, especially during reasoning tasks. Ensuring the accuracy of training data is crucial for model performance, and this limitation impacts both the reliability and validity of the results.




**Hình 8:** *One example of wrong data in ChartQA dataset.*

Secondly, **the size of our dataset** is relatively small, with only **7,158 examples**, which is significantly less than the large-scale datasets seen in other studies (e.g., 544.9k examples in [1]). This limited dataset size reduces the model's capacity to generalize effectively across diverse types of charts, questions, and reasoning processes. A larger dataset would allow the model to better learn the intricate relationships between chart elements and improve its ability to handle complex reasoning tasks.

Thirdly, our model struggles to handle charts with high visual complexity. Specifically, when charts contain a large number of values or when visual elements, such as lines or bars, overlap with each other, the model often hallucinates during the data extraction phase. These hallucinations, which originated from **UniChart** ([19]), lead to incorrect interpretations of the chart's structure and inaccurate answers. This limitation suggests that the model's understanding of visual elements and spatial relationships within charts requires further refinement.

Fourthly, while we leveraged **GPT-4o** and **GPT-4o-mini** to generate rationales and additional data, these models are not without flaws. They exhibit difficulties in processing images accurately and often generate incorrect or inconsistent data. In some cases, the rationale generated by GPT-4o models does not align with the provided answers, introducing further confusion during training. The flaws in these generated rationales propagate errors into the model, as the training process learns and reproduces these inaccuracies.

Finally, there are implementation challenges in our **multi-task setup**, which combines answer generation and rationale generation. Due to errors in our code, the setup appears to be misconfigured, resulting in artificially **high validation accuracy** but **low testing accuracy**. These errors likely stem from inconsistencies in task alignment or improper handling of training objectives, which require careful debugging and refinement.

# 6 Future Works

To overcome the limitations of our current approach, we propose several directions for future work.

First, we aim to train the model to generate underlying tables using the table structure produced by **DePlot** ([14]). By leveraging DePlot's outputs, we can improve the accuracy and consistency of the extracted tabular data.

Second, we will evaluate the accuracy and performance of the table generation model by using the generated tables as input for **GPT-4o**, testing its ability to answer questions and compute metrics based on this structured data.

Third, we will refine the rationale generation process using **GPT-4o** by training the model to generate rationales with a new dataset, where the input tuple consists of **(table, question, answer)** rather than **(image, question, answer)**. This change is expected to address the issue of incorrect data and rationales caused by flaws in the image-based inputs.

Fourth, we plan to enrich the dataset by increasing its size and diversity, incorporating more varied chart types, topics, and levels of complexity to improve the model's generalization capabilities.

Fifth, we will correct the misconfigured **multi-task setup** in our current implementation, ensuring alignment between training and testing objectives. This step is critical to resolving the discrepancy between high validation accuracy and low testing accuracy observed in our experiments.

Finally, we aim to adopt advanced reasoning techniques, such as **Self-Consistency** ([23]) and **Program-of-Thoughts** ([2]), to further enhance the model's performance. These techniques can improve the accuracy of multi-step reasoning tasks by introducing verification and programmatic reasoning processes.

# 7 Conclusion

In this study, we explored methods to enhance the performance of a Chart Visual Question Answering (ChartVQA) model by focusing on tasks such as data extraction, reasoning, and summarization. Our approach integrated multi-task learning to generate both answers and detailed rationales, drawing inspiration from techniques like Chain-of-Thoughts (COT) to improve step-by-step reasoning. Despite these efforts, our results reveal several challenges, including errors in data generation, limitations in the ChartQA dataset, and issues with multi-task training setup, which impacted the model's ability to handle complex reasoning tasks and visually dense charts.

To address these limitations, we proposed a set of future works aimed at improving the model's capabilities. These include leveraging table generation techniques like DePlot, refining rationale generation with more accurate inputs, enriching and expanding the dataset, and adopting advanced reasoning strategies such as Self-Consistency and Program-of-Thoughts. By implementing these steps, we aim to overcome current barriers and create a more robust and accurate ChartQA system.

This work highlights the complexity of reasoning over visual data and emphasizes the need for both high-quality datasets and well-aligned training objectives. While our current model shows promising results for simple tasks, significant opportunities remain to enhance its performance on more intricate and multi-step reasoning questions. We hope our research will contribute to advancing the field of visual question answering and provide a foundation for future innovations in chart-based reasoning tasks.

# 8 Appendix

# Tài liệu

[1] Victor Carbune, Hassan Mansoor, Fangyu Liu, Rahul Aralikatte, Gilles Baechler, Jindong Chen, and Abhanshu Sharma. Chart-based reasoning: Transferring capabilities from llms to vlms. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 989–1004, Mexico City, Mexico, 2024. Association for Computational Linguistics.

[2] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, 2023.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[4] Sebastian Gehrmann, Sebastian Ruder, Vitaly Nikolaev, Jan Botha, Michael Chavinda, Ankur Parikh, and Clara Rivera. TaTA: A multilingual table-to-text dataset for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1719–1740, Singapore, December 2023. Association for Computational Linguistics.

[5] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024.

[6] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[8] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023.

[9] Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1575–1583, 2021.

[10] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering, 2018.

[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, 2022.

[12] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

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

[15] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Nigel Collier, Yasemin Altun, and Julian Martin Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023.

[16] Yi-Cheng Liu and Wei-Ta Chu. Chart question answering based on modality conversion and large language models. In *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia*, AIQAM '24, pages 19–24, New York, NY, USA, 2024. Association for Computing Machinery.

[17] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1917–1925, January 2021.

[18] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason, 2023.

[19] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore, December 2023. Association for Computational Linguistics.

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

[21] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models, 2023.

[22] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2015.

[23] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[25] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 786–797. Springer, 2021.

[26] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.