Union Is Strength! Unite the Power of LLMs and MLLMs for Chart Question Answering

Jiapeng Liu[1,2], **Liang Li**[1,*], **Shihao Rao**[1,2], **Xiyan Gao**[1], **Weixin Guan**[1,2], **Bing Li**[1], **Can Ma**[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{liujiapeng,liliang}@iie.ac.cn

Abstract

Chart Question Answering (CQA) requires models to perform chart perception and reasoning. Recent studies driven by Large Language Models (LLMs) have dominated CQA. These include employing more cognitively capable LLMs for indirectly reasoning over transformed charts, i.e., tables, and directly perceiving charts utilizing Multimodal Large Language Models (MLLMs) with a wider perceptual range. Yet, they often encounter bottlenecks due to the limitation of the receptive field of LLMs and the fragility of the complex reasoning of some MLLMs. To unite the strengths of LLMs and MLLMs to complement each other's limitations, we propose SYNERGY, a framework that unites the power of both models for CQA. SYNERGY first unites the chart with a table as the augmented perceptual signal. Next, it unites LLMs and MLLMs, scheduling the former to decompose a question into subquestions and the latter to answer these by perceiving the chart. Lastly, it operates LLMs to summarize the subquestion-answer pairs to refine the final answer. Extensive experimental results on popular CharQA and PlotQA benchmarks reveal that, with the power of union, SYNERGY outperforms strong competitors and achieves superior boosts over naive MLLMs by uniting them with a smaller LLM.

Code — https://github.com/liuJP2/Synergy

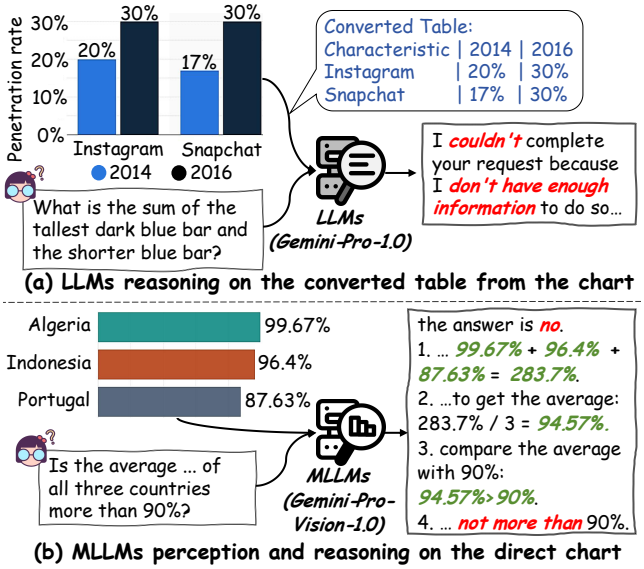

Figure 1: Prompting LLMs (a) and MLLMs (b) to perceive and reason on their respective sensory sources.

Introduction

Chart Question Answering (CQA) is a vital task in visual data analysis, serving a significant role in scientific research and business decision-making. It requires models to possess visual perception abilities to handle information-rich charts and excel in multimodal reasoning, including arithmetic and logical operations. Previous works (Masry et al. 2023; Cheng, Dai, and Hauptmann 2023) primarily focus on training or fine-tuning chart-specific models to achieve these abilities on charts. Recent studies (Baechler et al. 2024; Carbune et al. 2024) further enhance these abilities, effectively demonstrating reasoning on charts for more straightforward questions, but their performance significantly lags when addressing more complex questions.

Recently, Large Language Models (LLMs) (Touvron et al. 2023; Bai et al. 2023a; Dubey et al. 2024) have garnered substantial attention due to their exceptional capabilities in understanding and reasoning with textual data. These capabilities inspire many efforts to explore how LLMs can be utilized to overcome challenging problems in multimodal tasks, e.g., CQA. Some studies (Liu et al. 2022a; Xia et al. 2023) involve leveraging the concept of textual conversion, wherein visual charts are converted into textual tables that LLMs can process, as shown in Figure 1(a). Additionally, other studies (Team et al. 2023; Han et al. 2023; Lu et al. 2024) devise Multimodal Large Language Models (MLLMs), which use LLMs as the brain and incorporate visual cues processing capabilities to directly conduct reasoning on charts, as shown in Figure 1(b). These MLLMs generally equip an LLM with a visual encoder then bridge them with a mapping matrix. They undergo multi-stage training to align features across different modalities, endowing the LLM to perform reasoning on multimodal inputs.

However, the inherent limitations of LLMs and MLLMs may result in performance bottlenecks. Specifically, **LLMs' receptive field is limited to perceiving only textual tables**

*Corresponding author

converted from charts. Nevertheless, the representation of tables is limited and suffers from information loss. Because current chart-to-table conversions usually focus on extracting and organizing numerical information. Consequently, high-value information, e.g., color features and spatial relationships , is challenging to be expressed in a table. As seen in Figure 1 (a), the loss of critical visual cues constrains the inference performance of the LLM. **Some MLLMs built on LLMs are typically less capable of reasoning than the corresponding native LLMs** (Driess et al. 2023; Wang et al. 2023b). Specifically, there are significant differences in the training data distribution between MLLMs and the LLMs on which they are based. Moreover, most of them usually adopt a forced alignment strategy during training (Wang et al. 2023b). These two factors cause MLLMs to be more prone to potential misalignments and hallucinations when performing complex reasoning. Additionally, the reasoning chain of some MLLMs may be fragile (Fu et al. 2024), undermining the reliability of their reasoning processes. As shown in Figure 1 (b), despite having correct intermediate perceptions, MLLMs still draw wrong conclusions due to inconsistent reasoning. This observation is consistent with Fu et al. (2024).

LLMs excel in textual reasoning (cognition ability), while MLLMs have a broader perceptual scope (perception ability). Can they complement each other's strengths? Inspired by this, we propose SYNERGY, a framework that unites the power of cognitive and perceptive abilities from two kinds of models to achieve more accurate chart question answering. The critical insight of SYNERGY is to decompose the CQA into stages, each taking different signals to adopt the advantages of LLMs and MLLMs while avoiding their shortness. Specifically, SYNERGY first uses the existing chart2table method to extract table contents from the charts. A chart and its corresponding table are combined as an augmented perceptual signal. This is motivated by our observation that tables can provide more accurate numerical information than charts, while charts can provide more intuitive visual cues. This combination will make MLLMs perceive more accurately and robustly. Next, SYNERGY utilizes LLMs with more powerful cognitive ability to decompose a complex question into several simpler subquestions. The insight is that MLLMs are not as good as LLMs at handling complex questions. Then, SYNERGY employs MLLMs to answer simpler subquestions. It can not only avoid the loss of information caused by chart transitions (boosting) but also cover up the problem of its weak reasoning ability (avoiding shortness). In particular, augmented perceptual signals are input to MLLMs to obtain more accurate answers for subquestions in this stage. In the last stage, SYNERGY operates LLMs to summarize and refine the subquestion-answer pairs to arrive at the conclusive answer for the complex question.

Our contributions are summarized as follows:

- We propose a novel CQA framework SYNERGY, which unites the power of cognitive and perceptive abilities of LLMs and MLLMs to achieve better performance.
- We conduct extensive experiments on two popular benchmarks, ChartQA and PlotQA-sub, to verify the validity of the proposed SYNERGY. The experimental results show that SYNERGY performs substantially better than strong competitors.
- We show excellent compatibility and effectiveness in uniting the cognition ability of LLMs and the perceptual ability of MLLMs. Even though only a 8B LLM in SYNERGY, it significantly boosts the performance of the naïve MLLMs. That is, union is strength!

Related Work

Chart Question Answering (CQA) (Hoque, Kavehzadeh, and Masry 2022) aims to achieve understanding and analysis of charts by posing complex queries. Recent works (Masry et al. 2022; Chen et al. 2023; Masry et al. 2023) have incorporated transformers to capture complex visual and textual information for answering questions. These models provide efficient solutions but are typically smaller in scale.

Large Language Models (LLMs) (Brown et al. 2020; Ouyang et al. 2022) have shown impressive performance in few-shotscenarios. Researchers (Liu et al. 2022a; Xia et al. 2023) have explored applying LLMs to CQA by converting charts to tables and incorporating them into questions and in-context examples as prompts for LLM input. However, this conversion often results in the loss of visual cues, such as underlying trends and chart-specific features, which hinders the LLMs' reasoning capabilities. PROMPTCHART (Do et al. 2023) proposes adding extra textual visual cues, such as color, to the table. DOMINO (Wang et al. 2023a) leverages prompt engineering and fine-tuning strategies to enable LLMs to decompose questions and then prompts an additional fine-tuned DEPLOT to answer these subquestions. Compared with this, SYNERGY differs in two main aspects. First, SYNERGY requires no additional training or fine-tuning. Additionally, SYNERGY further refines and verifies the final answers obtained, which helps reduce the hallucination and improve the results' reliability.

Meanwhile, the emergence of Multimodal Large Language Models (MLLMs) offers new solutions for multimodal tasks like CQA. Some studies (Bai et al. 2023b; Zhu et al. 2023; Liu et al. 2023; Zhang et al. 2023) involve using vision encoders like ViT (Dosovitskiy et al. 2021) to process visual information and structures like Q-former (Li et al. 2023) to align visual and textual information, which are then input into pre-trained LLMs. Through multi-stage training, LLMs retain semantic understanding while gaining the ability to directly perceive information from multimodal inputs. Some closed-source MLLMs (Team et al. 2023; Achiam et al. 2023) have already developed comprehensive chart understanding abilities. Recent studies (Driess et al. 2023; Wang et al. 2023b) point out that some MLLMs exhibit weaker reasoning abilities compared to their underlying LLMs, especially with natural language questions. For instance, increasing the scale of PaLM-E (Driess et al. 2023) reduces catastrophic forgetting of language capabilities relative to PaLM (Chowdhery et al. 2023). Similarly, CogVLM (Wang et al. 2023b) shows that training an MLLM's language component on multimodal data can rapidly degrade its performance on pure text tasks.

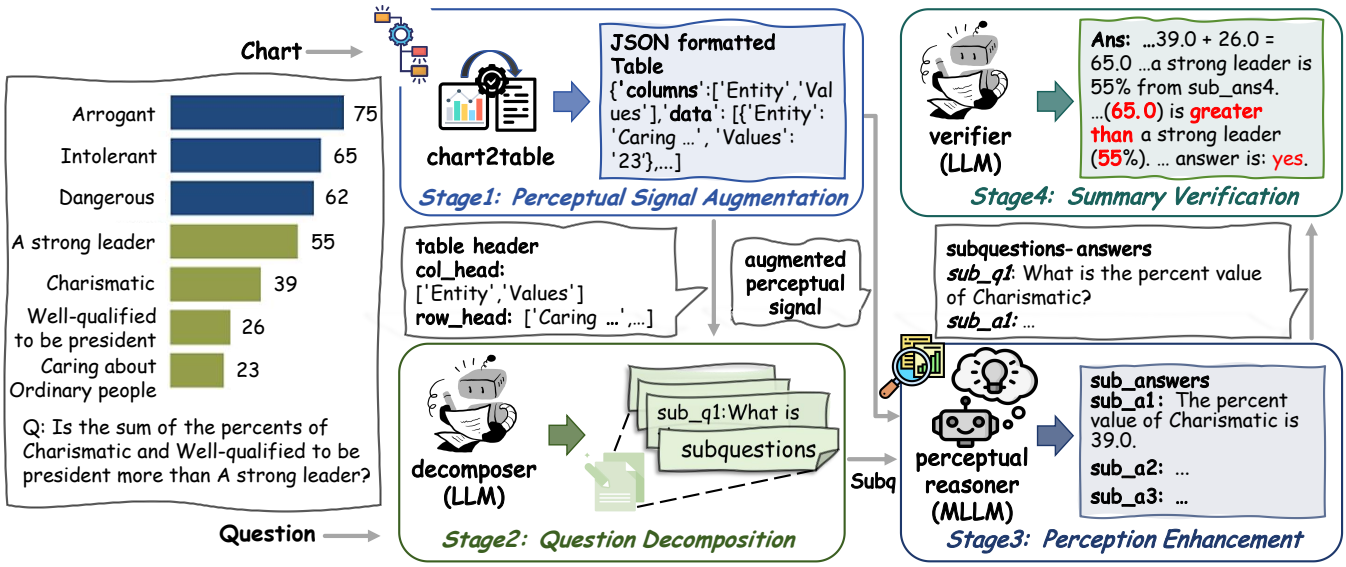

Figure 2: The architecture of our SYNERGY, which decomposes the chart question answering into four stages.

SYNERGY

The architecture of our SYNERGY is depicted in Figure 2. It unites the power of LLMs and MLLMs for reasoning and answering on the chart in four stages: Perceptual Signal Augmentation, Question Decomposition, Perception Enhancement and Summary Verification.

Perceptual Signal Augmentation

Compared with coarse-grained content perception, e.g., legend, precise and fine-grained content extraction, e.g., numerical value, is highly dependent on the perception ability of MLLMs. This makes the reasoning and answer accuracy for numerical values more sensitive to the perception ability of MLLMs. We notice that the chart contains intuitive visual cues, while the table provides more readily accessible, accurate numerical data. We hypothesize that tables and charts can complement each other's strengths. This is further confirmed in the Ablation Study section.

Therefore, we augment the single chart perceptual source with a corresponding table. Specifically, with chart $\mathcal{C}$, SYNERGY uses a chart2table model to transform $\mathcal{C}$ into table $\mathcal{T}$, which is then transformed into a JSON-formatted table $\mathcal{T}_{js}$. We then augment $\mathcal{C}$ with $\mathcal{T}_{js}$ to generate the augmented perceptual signal, abbreviated as au-signal.

Question Decomposition

Handling long reasoning chains for complex questions is more challenging for MLLMs. Therefore, we utilize the cognitive ability of LLMs to reduce the complexity of these questions.

Given a question $\mathcal{Q}$, SYNERGY prompts an LLM as a decomposer to parse $\mathcal{Q}$. While creating more examples via few-shot learning to encourage decomposition is feasible, we observed that LLMs tend to decompose questions into their most minor semantic units, which may not necessarily match the actual most minor semantic units. For example, "percentage" is broken down into "number" and "ratio," whereas the chart directly presents various percentage categories. Therefore, we use the row-column table headers $\mathcal{T}_{js}^{header}$ as the basis for decomposition, which provides a list of the minor units of the chart (e.g., x-labels):

$$Q_{sub} = \text{decomposer}(\mathcal{T}_{js}^{header}, \mathcal{Q}, Ct^{dec}) \quad (1)$$

where Ct^{dec} is a typical in-context learning paradigm that involves prepending a fixed one-shot to the current prompt, containing a manually crafted t_{js}^{header} and four (q, q_{sub}) pairs. Each sub-question in Q_{sub} can be separated by line breaks and finally formatted as a list $\{q_{sub}^1, q_{sub}^2, ..., q_{sub}^n\}$. Please refer to supplementary technical appendix for more details of Ct^{dec}.

Algorithm 1: Cross-Perceive

Input: au-signal and subquestions $\mathcal{Q}_{sub}$
Output: perception and reasoning path sub_{qa}

Initialize $index = 0$
Initialize a dictionary list sub_{qa}
for q_{sub}^j in $\mathcal{Q}_{sub}$ **do**
  if len($\mathcal{Q}_{sub}$) == 1 **then**
    $CurrentP = \mathrm{T}_{\mathrm{simpleQ}}$(au-signal, q_{sub}^j)
  else if $index$!= (len($\mathcal{Q}_{sub}$) -1) **then**
    $CurrentP = \mathrm{T}_{\mathrm{intermediate}}$(au-signal, q_{sub}^j)
  else
    $CurrentP = \mathrm{T}_{\mathrm{final}}$(au-signal, sub_{qa}, q_{sub}^j)
  end if
  ans_{sub}^j = perceptual_reasoner($CurrentP$)
  sub_{qa}.appened($\{q_{sub}^j, ans_{sub}^j\}$)
  $index$ += 1
end for
return au-signal and subquestions $\mathcal{Q}_{sub}$

Perception Enhancement

SYNERGY unites an MLLM with a broader perceptual scope, acting as a perceptual reasoner to robustnessly cross-perceive au-signal to address each simpler subquestion within $\mathcal{Q}_{sub}$.

As presented in Algorithm 1, SYNERGY first determines the complexity of $\mathcal{Q}_{sub}$ based on its length, thereby using different forms of prompting to activate the MLLM's perception ability of au-signal. Specifically, SYNERGY considers the $\mathcal{Q}_{sub}$ with only one element as a simple query and directly employs the $\mathrm{T}_{\text{simpleQ}}$ template with CoT (Wei et al. 2022) for perceptual reasoning:

$\mathrm{T}_{\text{simpleQ}}$

```
{au-signal_chart}
Here is the chart and its associated table
to answer this question. Please provide your
explanation first, then answer the question
step by step.
table in JSON format: {au-signal_Tjs}
Q: {Q}
the final short answer should start with "the
answer is ".
```

When Q_{sub} contains multiple decomposed subquestions, SYNERGY considers Q to be a complex query involving multi-step perception and reasoning. For each subquestion in $\{q^j_{sub} \mid j \in [1, n-1]\}$, SYNERGY prompts the perceptual reasoner to generate ans^j_{sub} where $j \in [1, n-1]$ by using $\mathrm{T}_{\text{intermediate}}$:

$\mathrm{T}_{\text{intermediate}}$

```
{au-signal_chart}
Analyze the chart and review the table, then
answer the following questions. Additionally,
if the table contains untrusted values, get the
relevant values from chart.
here is the table in JSON format:{au-signal_Tjs}
Q: {q^i_sub}
```

Then, SYNERGY uses $\mathrm{T}_{\text{final}}$ to prompt the perceptual reasoner to review the previous explicitly reasoning path $\{sub^j_{qa} \mid j \in [1, n-1]\}$ and to provide a preliminary answer for the n-th subquestion:

$\mathrm{T}_{\text{final}}$

```
{au-signal_chart}
Analyze the chart to answer the question no
more than three words.
Part of the analysis results of the chart are
provided below. You can refer to the chart,
table and some sub-QA pairs(the table and
sub-QA pairs are not guaranteed to be exactly
right) to get the final answer.
Additionally, if the table contains untrusted
or Nan values, get the relevant values from
chart.
This is a table in JSON format: {au-signal_Tjs}
sub-QA pairs for the chart:
{sub^j_qa | j ∈ [1, n − 1]}
Q: {q^j_sub}
a short answer is:
```

Summary Verification

When tackling complex queries, the long inference chain of MLLMs, is prone to producing inconsistent inference results. Even if there is a correct inference process, the MLLM may produce wrong conclusions. Considering that LLMs are better at cognizing and reasoning, SYNERGY schedules an LLM as the verifier to correct the preliminary answer of these complex questions based on the intermediate perceptual and reasoning path:

$$A_{final} = \text{Verifier}(sub_{qa}, Ct^{ver}) \tag{2}$$

where sub_{qa} represents the perception and reasoning path generated by the perceptual reasoner. Ct^{ver} refers to K-shot in-context examples, each of which comprises a set of artificially designed sub-qa pairs and a verification process tailored to the preliminary answer. Please refer to supplementary technical appendix for more details of Ct^{ver}.

Experimental Setup

Datasets and Metrics

We evaluate SYNERGY on the following two public datasets:
ChartQA (Masry et al. 2022) is divided into two sets: augmented and human. The former is synthetically generated, while the latter consists of queries written by humans, exhibiting greater complexity resembling real-world scenarios and requiring advanced reasoning abilities.
PlotQA (Methani et al. 2020) is an extensive dataset of machine-generated questions, divided into V1 and V2 sets. V1 focuses on chart-specific details, while V2 emphasizes numerical reasoning. Due to PlotQA containing millions of QA pairs for testing and the limitations on computing resources, conducting experiments on full PlotQA incurs time expenditure and monetary costs. Therefore, we randomly (seed=123) selected 1500 samples from V1 and 2500 samples from V2 of the test set to create the PlotQA-sub for experimentation.

Following the previous works (Masry et al. 2022; Liu et al. 2022b) evaluated on these datasets, we use standard accuracy with a relaxed correctness criterion that permits a maximum 5% tolerance on numerical error.

Baselines

We compare our approach with previous strong competitors, categorized as follows:
Chart-specific Models , methods involving training and fine-tuning models specifically for charts, including Unichart (Masry et al. 2023), Chart2Table PT PaLI-3 (Carbune et al. 2024) and ChartPaLI-5B (Carbune et al. 2024) .
LLMs-based Methods, approaches that use LLMs to complete tasks on tables. We use DEPLOT+FlanPaLM (540B) CoT (Liu et al. 2022a), PROMPTCHART (Do et al. 2023), DOMINO (Wang et al. 2023a) . Additionally, we also use DEPLOT+LLaMA3-Instruct (8B), DEPLOT+Gemini-Pro, and DEPLOT+GPT-4, which are implemented by us, employing 6-shot+CoT .
MLLMs-based Methods, methods that prompt MLLMs to perform reasoning on charts. We select several MLLMs with

MODEL	ChartQA			PlotQA-sub		
	human	augmented	avg.	V1$_{\text{(content perception)}}$	V2$_{\text{(numerical reasoning)}}$	avg.
Chart-specific Models						
UniChart (Masry et al. 2023)	43.92	88.56	66.24	-	-	-
Chart2Table PT PaLI-3 (Carbune et al. 2024)	48.96	92.72	70.84	-	-	-
ChartPaLI-5B (Carbune et al. 2024)	60.88	93.68	77.28	-	-	-
LLMs-based methods						
DEPLOT+FlanPaLM(540B) CoT (Liu et al. 2022a)	57.80	76.70	67.30	-	-	-
DEPLOT+LLaMA3-Instruct (8B, 6-shot+CoT) †	58.24	76.29	67.27	-	-	-
DEPLOT+Gemini-Pro (6-shot+CoT) †	58.64	81.29	69.97	-	-	-
DEPLOT+GPT-4 (6-shot+CoT) †	65.92	84.11	75.54	-	-	-
PROMPTCHART (Do et al. 2023)	63.20	81.44	72.32	-	-	-
DOMINO $_{\text{Fine-tuned DEPLOT+LLaMA2-70B}}$ (Wang et al. 2023a)	61.70	91.70	76.70	-	-	-
MLLMs-based methods						
LLaVA-V1.5(13B)† (Liu et al. 2024)	20.96	19.03	20.00	22.53	6.00	14.27
LLaVA-V1.5(13B, fine-tuned on chart) (Han et al. 2023)	37.68	72.96	55.32	-	-	-
ChartLLaMA(fine-tuned on chart) (Han et al. 2023)	48.96	90.36	69.66	-	-	-
Gemini-Pro-Vision† (Team et al. 2023)	62.27	77.90	70.09	42.71	16.96	29.83
Qwen-VL-Plus† (Bai et al. 2023b)	50.88	83.06	66.97	39.40	9.40	24.40
GPT-4V+CoT (Islam et al. 2024)	72.64	66.32	69.48	-	-	-
Ours (SYNERGY $_{\text{perceptual-reasoner + decomposer\&verifier}}$)						
SYNERGY $_{\text{LLaVA-V1.5(13B) + LLaMA3-8B}}$	48.96 ↑28.00	85.97 ↑66.94	67.46 ↑47.46	40.76 ↑18.23	38.32 ↑32.32	39.54 ↑25.27
SYNERGY $_{\text{LLaVA-V1.5(13B) + Gemini-Pro}}$	52.16 ↑31.20	87.42 ↑68.39	69.79 ↑49.76	41.75 ↑19.22	40.88 ↑34.88	41.31 ↑27.04
SYNERGY $_{\text{LLaVA-V1.5(13B) + GPT-4}}$	54.66 ↑33.70	87.10 ↑68.07	70.88 ↑50.88	44.07 ↑21.54	41.96 ↑35.96	43.01 ↑28.74
SYNERGY $_{\text{Gemini-Pro-Vision + LLaMA3-8B}}$	63.76 ↑1.49	85.24 ↑7.34	74.50 ↑4.41	53.12 ↑10.41	43.84 ↑26.88	48.48 ↑18.65
SYNERGY $_{\text{Gemini-Pro-Vision + Gemini-Pro}}$	66.64 ↑4.37	85.48 ↑7.58	76.06 ↑5.97	55.31 ↑12.6	45.22 ↑28.26	50.26 ↑20.43
SYNERGY $_{\text{Gemini-Pro-Vision + GPT-4}}$	67.20 ↑4.93	85.89 ↑7.99	76.54 ↑6.45	56.50 ↑13.79	45.20 ↑28.24	50.85 ↑21.02
SYNERGY $_{\text{Qwen-VL-Plus + GPT-4}}$	65.22 ↑14.32	90.81 ↑7.75	78.01 ↑11.04	54.74 ↑15.34	44.28 ↑34.88	49.51 ↑25.11

Table 1: Main experimental results on ChartQA and PlotQA-sub test benchmarks. "perceptual-reasoner + decomposer&verifier" refers to the MLLMs used for visual perception and the LLMs used for question decomposition and reasoning verification in SYNERGY. Results marked with "†" are our reimplements in this paper, with the detailed exposition in supplementary technical appendix. The different colors are used to indicate the performance improvement of these MLLMs incorporated in SYNERGY relative to the original performance.

varying levels of chart comprehension, including LLaVA-V1.5 (13B) (Liu et al. 2024), Gemini-Pro-Vision (Team et al. 2023), Qwen-VL-Plus (Bai et al. 2023b), GPT-4V+CoT (Islam et al. 2024), and a model specifically fine-tuned for charts, ChartLLaMA (Han et al. 2023).

Implementation Details

We utilize DEPLOT (Liu et al. 2022a) as the chart2table model. Additionally, we equip SYNERGY with a variety of combinations of MLLMs and LLMs. For perceptual reasoners, we select LLaVA-V1.5 (13B), Gemini-Pro-Vision, and Qwen-VL-Plus. For both decomposers and verifiers, we chose LLaMA3-Instruct (8B), Gemini-Pro (v1.0), and GPT-4 (gpt-4-1106-preview). We set the temperature to 0 for all models except for Qwen-VL-Plus, for which we set top_p to 0.001 and top_k to 1. By default, we use 6-shot for verifiers to perform in-context learning. Please refer to supplementary technical appendix for more details.

Experimental Results

Main Results

Table 1 summarizes the experimental results on CharQA and PlotQA-sub. We demonstrate SYNERGY's superiority by comparing and analyzing in the following three aspects.

Compared to naïve MLLMs SYNERGY takes full advantage of their wider perceptual range while utilizing LLMs to compensate for their weaker cognitive ability. Moreover, it shows surprising compatibility and generalization in the selection of LLMs. Specifically, for the LLaVA-V1.5, which is weak in chart perception, SYNERGY significantly improves its CQA accuracy, regardless of the cognitive level of the LLMs it collaborated with. Notably, SYNERGY improves LLaVA-V1.5 on the ChartQA and PlotQA datasets by $+50.88\%$ and $+28.74\%$, when taking GPT-4 as the decomposer and verifier.

Further, we observe that, with the MLLM's parameters unchanged, either fine-tuning LLaVA-V1.5 directly on Chart QA data or using more chart-related training data such as ChartLLaMA to enhance their chart perception, none have advanced beyond SYNERGY. Although SYNERGY only unites LLama3-Instruct (8B) and LLaVA-V1.5, it performs far better on the ChartQA (67.46%) than the fine-tuned LLaVA-1.5 (55.32%). It is also competitive with ChartLlama (69.66%), which spends more on computing and data resources.

Lastly, though the MLLMs have further improved in parameters and data, their chart question-answering capability still needs to catch up with the proposed SYNERGY. For example, SYNERGY exceeds Gemini-Pro-Vision on the ChartQA by $+4.41\% \sim +6.45\%$ and on the PlotQA-sub set by $+18.65\% \sim +21.02\%$. When uniting the powerful GPT-4, SYNERGY improves Qwen-VL-Plus by an astonishing $+11.04\%$ and $+25.11\%$ on both datasets, respectively.

To sum up, even though MLLMs enhance the chart per-

MODEL				ChartQA		
perceptual reasoner	au-signal	decomposer	verifier	human	augmented	avg.
LLaVA-V1.5 (13B)	-	-	-	22.40	20.46	21.43
	✓	-	-	47.29	77.74	62.52
	✓	✓	-	52.71	77.59	65.15
	✓	✓	✓	57.38	77.22	67.30
Gemini-Pro-Vision	-	-	-	61.15	71.94	66.54
	✓	-	-	63.85	77.32	70.59
	✓	✓	-	63.54	77.43	70.48
	✓	✓	✓	66.49	77.42	71.96

Table 2: Ablation study of SYNERGY on ChartQA validation set. GPT-4 is utilized as the decomposer and verifier.

ception by boosting the size and data, their cognitive ability is still a bottleneck, which may stem from current alignment technologies (Wang et al. 2023b).

Compared to LLM-based methods Thanks to the better perceptual ability, SYNERGY can more comprehensively complete complex reasoning on charts than LLMs. By uniting different MLLMs, SYNERGY can achieve comparable or better performance than LLM-based methods. For instance, we find SYNERGY $_{\text{LLaVA-V1.5(13B)+LLaMA3-8B}}$ only improves the LLaMA-based method (67.27%) by $+0.19\%$. This may be because, in this setup, the CQA bottleneck is not only due to a lack of cognitive ability but also perception. As we expected, when equipping LLaMA3-Instruct with the more powerful perceptual reasoner, Gemini-Pro-Vision, SYNERGY achieves an average accuracy improvement of $+7.23\%$. Even when compared to the stronger cognitive model GPT-4, SYNERGY improves its average accuracy on the ChartQA test set by up to 2.47%.

The above results also indicate that extracting information from tables with inadequate chart representation leads to performance bottlenecks in reasoning, even when additional visual cues are included (such as color) or when using more cognitive LLMs, e.g., Gemini-Pro and GPT-4.

On the other hand, DOMINO, which uses additional fine-tuned LLMs to guide fine-tuned DEPLOT in answering questions, performs better on the augmented set than our SYNERGY $_{\text{Gemini-Pro-Vision+LLaMA3-8B}}$. One primary reason is that DOMINO, which leverages prompt engineering and fine-tuning strategies, learns more effectively from the augmented set with simpler questions. In contrast, SYNERGY has yet to undergo similar supervised fine-tuning. However, on the human set, which includes more complex questions, our SYNERGY demonstrates stronger performance (63.76% vs. 61.7%). When incorporating larger LLMs like Gemini-Pro, SYNERGY shows even better results on the human set (66.64% vs. 61.7%).

Compared to Chart-specific Models Our SYNERGY demonstrates remarkable performance advantages on the ChartQA dataset, particularly in scenarios requiring complex multi-step reasoning, such as human set. In contrast, chart-specific models perform better on the augmented set, which contains more straightforward questions. We attribute this to the fact that these smaller-scale models excel at learning what is required for simple perceptual reasoning on charts but struggle with reasoning on charts involving complex queries that encompass intricate semantic information and require multi-step computational reasoning.

To sum up, SYNERGY has excellent flexibility, compatibility, and generalization, proving our motivation that union is strength.

Ablation Study

In this section, we conduct experiments to investigate the contributions of various components within SYNERGY. To mitigate the potential ablation bias introduced by a single model, we choose two MLLMs with different perceptual capabilities and regard their performance based on CoT (Wei et al. 2022) as baselines for comparison. The results are summarized in Table 2.

Specifically, we first explore the function of the augmented perceptual au-signal. As can be seen, the accuracy of both MLLMs is significantly improved with the table signal as an extra input besides the chart. Moreover, the argument signal boosts LLaVA-V1.5 (averaged $+41.09\%$) more significantly than Gemini-Pro-Visison (averaged $+4.05\%$), where the former is far less perceptive. The observation verifies our hypothesis that tables and charts can complement each other's strengths: the former provides more accurate numerical information, and the latter provides more intuitive visual cues. The unity of the two signals compensates for the shortcomings of perception by MLLMs.

After the introduction of the decomposer, LLaVA-V1.5 experiences a substantial performance boost, whereas the impact on Gemini-Pro-Vision is comparatively less pronounced. We attribute this to the fact that larger MLLMs are typically built upon LLMs with more advanced cognitive capabilities. These larger models benefit from their superior semantic understanding and reasoning abilities, which enable them to perform extended chains of reasoning more effectively and independently.

Finally, as the verifier is introduced, both MLLMs' performance is further enhanced. This improvement can be attributed to the cognitive capabilities of LLMs, which allow them to summarize the reasoning chain of MLLMs and identify, as well as correct, errors resulting from inconsistent or

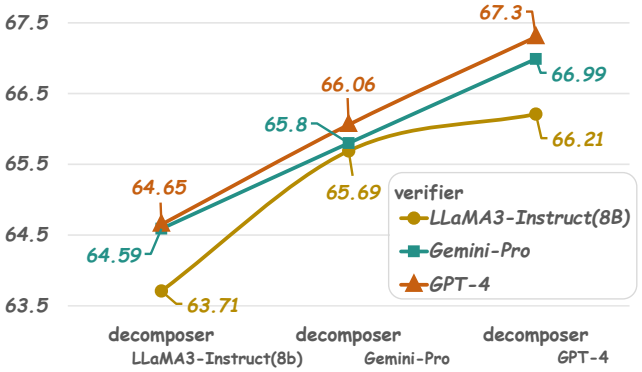

Figure 3: Performance variation on the ChartQA validation set by equipping LLaVA-V1.5(13B) with different combinations of LLMs as verifiers and decomposers.

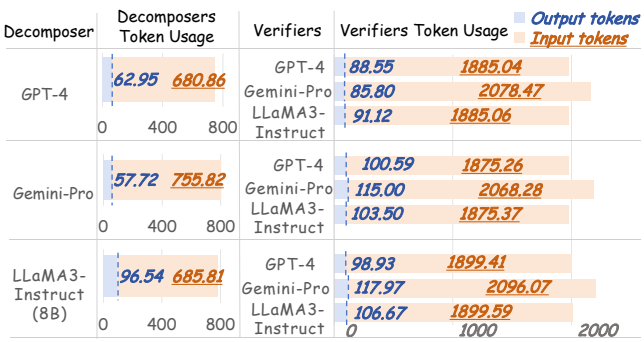

Figure 4: Token usage on ChartQA validation set with LLaVA-V1.5(13B) using different LLMs as decomposers and verifiers.

flawed reasoning processes.

Further Analysis

Effect of LLMs Across Roles We explore the decomposer's and the verifier's importance by changing the LLMs behind them. Considering the expensive costs of Gemini-Pro-Vision and Qwen-VL-Plus, we utilize the open-source LLaVA-V1.5 (13B) as the perceptual reasoner. Figure 3 shows the performance variation. It can be observed that the scale of LLMs is positively correlated with the SYNERGY performance, where the scale of LLMs driving decomposer makes a more significant impact on the result. Compared to using LLaMA3-Instruct (8B) and GPT-4 as the decomposer and verifier, switching their roles provides higher benefits for SYNERGY. We attribute this to the fact that the decomposer influences the verifier's summarization and verification capabilities. Smaller decomposers may generate low-quality subquestions, which lead the perceptual reasoner to produce incorrect sub-answers. Even if the verifier is large-scale, refining accurate answers from unreliable intermediate perceptual results is challenging.

Token Usage of LLMs Figure 4 shows the token usage of different LLMs. We find that taking smaller LLMs as the decomposer results in higher output token usage. In contrast, the disparity in output token usage among LLMs of different scales in the verifiers is negligible. This can be attributed to the limited ability of smaller LLMs to maintain output quality in contexts with limited information, resulting in the generation of unnecessary subquestions. Additionally, we note that the input token usage results of LLaMA3-Instruct and GPT-4 for identical inputs are highly similar. Lastly, utilizing GPT-4 as the decomposer paired with the open-source LLaMA3-Instruct as the verifier results in the lowest computational overhead (lowest GPT-4 API call) for the proposed SYNERGY and achieves performance comparable to Gemini-Pro-Vision (66.21% vs 66.54%).

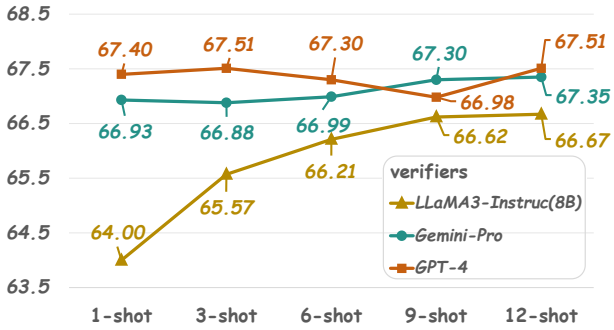

Figure 5: Performance trend on the ChartQA validation set with K-shot for the same decomposer (GPT-4) with different Verifiers.

Effect of Few-Shot Settings For Verifier We also explore the impact of different numbers of in-context examples on the prompt verifier's summarization and verification when using the powerful GPT-4 decomposer. We set LLaVA-V1.5 as the perceptual reasoner and the experimental results are shown in Figure 5. It can be seen that LLMs with stronger reasoning or cognitive ability require fewer examples to achieve performance convergence. Performance fluctuations occur as the number of in-context examples (K) increases, but the variation is never more than 0.5%. For the smaller LLaMA3-8B, performance improves with the number of shots increases, leveling off at 9-shot with an accuracy of 66.62%, which is only -0.68% lower than Gemini-Pro and GPT-4. This indicates that smaller open-source LLaMA3-Instruct can achieve high performance while reducing commercial API call costs by adding more in-context demonstrations. However, it also incurs the cost of manually designing contextual demonstrations.

Conclusion

This paper proposes a framework called SYNERGY for chart question-answering (CQA). This framework unites the cognitive capabilities of LLMs and the perception abilities of MLLMs by decomposing CQA into stages. The extensive experiments on ChartQA and PlotQA-sub datasets confirm that SYNERGY outperforms other strong competitors. Extensive experimental results and detailed analyses demonstrate that effectively uniting the power of LLMs and MLLMs can significantly improve the accuracy of CQA.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Baechler, G.; Sunkara, S.; Wang, M.; Zubach, F.; Mansoor, H.; Etter, V.; Cărbune, V.; Lin, J.; Chen, J.; and Sharma, A. 2024. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Carbune, V.; Mansoor, H.; Liu, F.; Aralikatte, R.; Baechler, G.; Chen, J.; and Sharma, A. 2024. Chart-based Reasoning: Transferring Capabilities from LLMs to VLMs. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 989–1004. Mexico City, Mexico: Association for Computational Linguistics.

Chen, X.; Wang, X.; Beyer, L.; Kolesnikov, A.; Wu, J.; Voigtlaender, P.; Mustafa, B.; Goodman, S.; Alabdulmohsin, I.; Padlewski, P.; et al. 2023. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Cheng, Z.-Q.; Dai, Q.; and Hauptmann, A. G. 2023. Chartreader: A unified framework for chart derendering and comprehension without heuristic rules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22202–22213.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Do, X. L.; Hassanpour, M.; Masry, A.; Kavehzadeh, P.; Hoque, E.; and Joty, S. 2023. Do LLMs Work on Charts? Designing Few-Shot Prompts for Chart Question Answering and Summarization. *arXiv preprint arXiv:2312.10610*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Han, Y.; Zhang, C.; Chen, X.; Yang, X.; Wang, Z.; Yu, G.; Fu, B.; and Zhang, H. 2023. ChartLlama: A Multimodal LLM for Chart Understanding and Generation. arXiv:2311.16483.

Hoque, E.; Kavehzadeh, P.; and Masry, A. 2022. Chart question answering: State of the art and future directions. In *Computer Graphics Forum*, volume 41, 555–572. Wiley Online Library.

Islam, M. S.; Rahman, R.; Masry, A.; Laskar, M. T. R.; Nayeem, M. T.; and Hoque, E. 2024. Are Large Vision Language Models up to the Challenge of Chart Comprehension and Reasoning? An Extensive Investigation into the Capabilities and Limitations of LVLMs. *arXiv preprint arXiv:2406.00257*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, F.; Eisenschlos, J. M.; Piccinno, F.; Krichene, S.; Pang, C.; Lee, K.; Joshi, M.; Chen, W.; Collier, N.; and Altun, Y. 2022a. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*.

Liu, F.; Piccinno, F.; Krichene, S.; Pang, C.; Lee, K.; Joshi, M.; Altun, Y.; Collier, N.; and Eisenschlos, J. M. 2022b. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Masry, A.; Kavehzadeh, P.; Do, X. L.; Hoque, E.; and Joty, S. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

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

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, P.; Golovneva, O.; Aghajanyan, A.; Ren, X.; Chen, M.; Celikyilmaz, A.; and Fazel-Zarandi, M. 2023a. DOMINO: A Dual-System for Multi-step Visual Language Reasoning. *arXiv preprint arXiv:2310.02804*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023b. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xia, R.; Zhang, B.; Peng, H.; Liao, N.; Ye, P.; Shi, B.; Yan, J.; and Qiao, Y. 2023. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*.

Zhang, P.; Wang, X. D. B.; Cao, Y.; Xu, C.; Ouyang, L.; Zhao, Z.; Ding, S.; Zhang, S.; Duan, H.; Yan, H.; et al. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.