

A. Related work

Multi-modal Large Language Model Recently, there has been a significant increase in the development of multi-modal large language models (MLLM), which allow LLMs to perceive other modalities. In particular, many studies have been introduced to enhance LLMs with the ability to understand visual input. These studies have generally been conducted by combining vision encoders (*e.g.*, CLIP [52]) with LLMs [8, 61, 63] by a vision-to-language projection modules, such as a linear layer, Q-Former [31], and Resampler [2]. Using this approach, a range of MLLMs including LLaVA [36, 37], MiniGPT-v2 [5], MiniGPT-4 [67], BLIP-2 [31], InstructBLIP [11], Qwen-VL [3], mPLUG-Owl [64], LLaVA-NeXT [29], and LLaVA-NeXT-Interleave [30] have been developed, showing robust capabilities in vision understanding for natural images.

Similarly, various studies have been conducted on chart-specific models capable of chart understanding and reasoning. Matcha [35] and Unichart [45] were developed based on a pre-training approach for specific chart tasks. Although this pre-training approach significantly enhances chart comprehension ability, it requires task-specific fine-tuning, which limits its generalization capability. These limitations were improved by using a model with a structure that connects a vision encoder to LLM, and performing instruction tuning on this model. ChartLlama [15], ChartInstruct [46], and ChartAst [47] improved chart understanding performance using this approach.

However, while these MLLMs perform well in of each modality with text, such as images or charts independently, these models cannot process images, charts, and text simultaneously. In other words, they cannot handle the multi-modal contexts of documents that include these three modalities. Therefore, in this study, we propose IC-tuning method to align image and chart modalities to LLM, enhancing chart comprehension performance while preventing catastrophic forgetting of image and language understanding performance.

Chart instruction-following data ChartLlama [15], ChartInstruct [46], and ChartAst [47] proposed chart instruction-following data for various tasks using GPT and improved chart comprehension performance by instruction tuning MLLMs based on this data. ChartLlama constructed instruction-following data for tasks such as data extraction, detailed description, chart code generation, and chart editing, based on the description of chart and the raw data (table). ChartInstruct developed instruction-following data for 9 tasks, including chart summarization and CoT reasoning QA, based on tables that describe charts and metadata such as the chart title. ChartAst proposed chart-specific instruction-tuning benchmark, referred to as ChartSFT. ChartSFT consists of data for five tasks, including chart summarization and open-ended QA, where open-ended QA data is created using GPT based on tables, captions and paragraphs describing the table.

These existing chart instruction-following data were generated exclusively based on tables without using the chart images. Therefore, these data lacks the ability to understand the visual information that charts inherently possess. In contrast, we newly propose the visual-aware chart instruction-following data based on chart images, enhancing chart understanding performance.

Multi-modal QA datasets In recent years, various research efforts on multi-modal QA datasets have been proposed. A significant portion of multi-modal QA datasets targets Visual Question Answering (VQA) task using image and text modalities [4, 12, 13, 19, 58]. And there are some studies focusing on video and text [27, 54] or table and text [6] modalities.

On the other side, ManyModalQA [16] and MultiModalQA [60] include three modalities: image, table, and text. In ManyModalQA, answers to questions can be derived from a single modality without the need for cross-modal reasoning. In contrast, MultiModalQA advanced this by creating multi-hop questions that require joint reasoning across three different modalities.

However, there is currently no dataset for question answering in multi-modal contexts including image, chart, and text modalities, which is the challenge addressed by this study. Therefore, we propose ICT-QA dataset by generating diverse sets of single and cross-modal QA datasets based on GPT-4V.

B. Qualitative results

B.1. IC-tuning

Figures 3 and 4 show the qualitative results of *Vicuna (IC-tuning)* and *Vicuna (IC-tuning w/o stage 3)* in chart summarization and OpenCQA, respectively. The parts highlighted in red indicate errors, while those highlighted in blue indicate where performance surpasses that of the other model.

In Figure 3, “Gold table” represents the chart in tabular format and is included as a reference for qualitative analysis. These tables are used in G-Eval. As evident from D1, when stage 3 tuning is not performed, the result includes factual errors about certain values of the chart, and provides for only a few specific values of chart. In contrast, after undergoing stage 3 tuning, the result has no factual errors and covers almost all values of chart. As shown in D2, while *Vicuna (IC-tuning w/o stage 3)* accurately generates the numerical values of the chart, it provides incorrect legend. On the other hand, *Vicuna (IC-tuning)* not only generates the numerical values correctly, but also provides useful information within the chart (e.g., title, color, source). These results show that training visual-aware chart instruction-following data in IC-tuning enhances the factual accuracy and informativeness in the chart summarization task.

We also conducted a qualitative evaluation of OpenCQA (Figure 4). In the case of Q1, *Vicuna (IC-tuning w/o stage 3)* incorrectly maps the values to each legend. In contrast, *Vicuna (IC-tuning)* accurately provides the values for each legend and generates summary of the comparison results for the two legends as requested by the question. As observed in Q2, the question asks for opinion on both Maduro and the opposition party. However, *Vicuna (IC-tuning w/o stage 3)* only mentions Maduro and provides incorrect values. On the other hand, *Vicuna (IC-tuning)* generates answer that is relevant to the question with correct values. In conclusion, these results demonstrates that stage 3 tuning improves factual accuracy and relevance in the open-ended QA task.

B.2. ICT-QA

Figure 5 shows the example of single-modal question that requires referring to the chart in multi-modal contexts. LLaVA generates incorrect answers by referring number in text. On the other hand, *Vicuna (IC-tuning)* generates a correct answering by referring chart. Figure 6 shows the example of cross-modal question that requires referring to both image and text in the multi-modal contexts. *Vicuna (IC-tuning)* generates factual errors and statements that are not related to the question. On the other hand, *Vicuna (IC-tuning + ICT-QA)* generates a correct answer.

Figures 7 to 9 display the qualitative results of *Vicuna (IC-tuning + ICT-QA)* on ICT-QA test dataset, showing examples of results for image QA, chart QA, and text QA, respectively. Figure 7 is the case of the image QA data belonging to *Purpose* type (Table 1), which requires referencing both image and text modalities for answering. *Vicuna (IC-tuning + ICT-QA)* effectively understands the text of the document, interprets the image accurately, and provides the correct response. This result demonstrates that *Vicuna (IC-tuning + ICT-QA)* possesses multi-modal reasoning capabilities. Figure 8 is the case of chart QA data belonging to *Comparison* type (Table 1), and the answer is accurately derived from the information presented in the chart. Lastly, Figure 9 is the case of the text QA data belonging to *Identification* type (Table 1), which is an example of successfully extracting the necessary information from the text and generating a response. Upon qualitative analysis, it is observed that *Vicuna (IC-tuning + ICT-QA)* performs QA tasks effectively within a multi-modal contexts, understanding specific or multiple modalities as we expected.

Additionally, as shown in Figures 10 and 11, we perform a comparison of qualitative results between our models and comparative models. ChartLlama did not follow the instructions of ICT-QA well. Therefore, we select cases when ChartLlama most closely followed the instructions for the comparison of qualitative results. In Figure 10, showing a chart QA example, LLaVA and ChartLlama use only chart and text as inputs, whereas *Vicuna (IC-tuning)* and *Vicuna (IC-tuning + ICT-QA)* utilize image, chart, and text modalities as inputs to generate responses. Figure 11, an example of Image QA, is the result when LLaVA and ChartLlama use only image and text as inputs, while our models use the entire modality as input. The results show that LLaVA and ChartLlama, unlike our models, often provide answers that are irrelevant to the questions and generate hallucinations. Furthermore, by comparing *Vicuna (IC-tuning)* and *Vicuna (IC-tuning + ICT-QA)*, it demonstrates that ICT-QA enhances the ability to perform question answering in multi-modal contexts and generates more accurate and higher quality answers.

C. Error analysis

Figure 12 shows the results of *Vicuna (IC-tuning + ICT-QA)* on challenging examples from ICT-QA. E1, E2, and E3 are examples of chart QA, and E4 is an example of image QA.

D. Details of IC-tuning

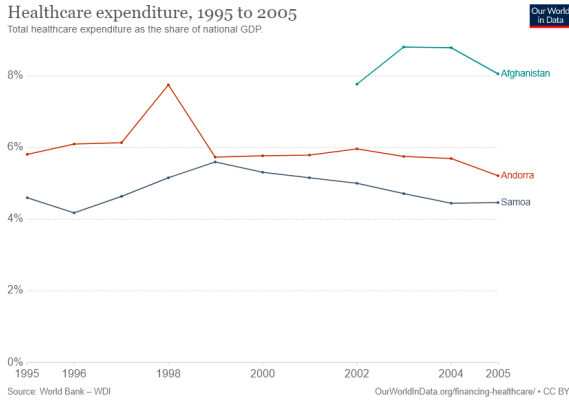
In the first stage of IC-tuning, we train only the projection layers of the modality encoders using a mixture of modality feature alignment data, as shown in Table 7, for one epoch, with a learning rate of $2e-3$ and a batch size of 128.

In the second stage of IC-tuning, we fine-tune Vicuna and projection layers using a mixture of multi-modal instruction-following data, as shown in Table 8, for 3 epoch with a learning rate of $2e-5$ and a batch size of 128.

D1

Healthcare expenditure, 1995 to 2005

Total healthcare expenditure as the share of national GDP.



Gold table

Characteristic	Afghanistan	Andorra	Samoa
1995	-	5.94	4.73
1996	-	6.23	4.31
1998	-	7.88	5.29
2000	-	5.9	5.44
2002	7.9	6.1	5.13
2004	8.92	5.83	4.58
2005	8.19	5.35	4.6

Vicuna (IC-tuning) w/o stage 3 :

The line chart shows the healthcare expenditure of Afghanistan, Andorra, and Samoa from 1995 to 2005. Afghanistan had the highest healthcare expenditure, increasing from 7.93 in 2002 to 8.93 in 2003. Andorra had the second highest expenditure, ranging from 5.94 in 2002 to **6.1 in 2003**. Samoa had the lowest expenditure, ranging from **4.32 in 2002** to 4.6 in 2005. Overall, **the trend shows an increase in healthcare expenditure for all three regions from 1995 to 2005**.

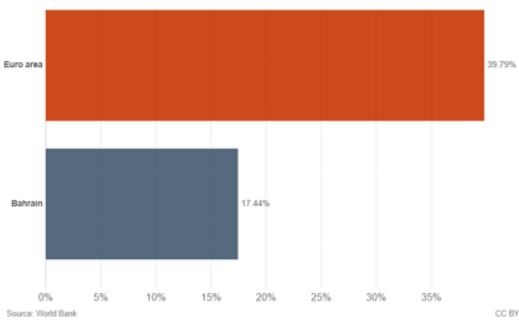
Vicuna (IC-tuning) :

The chart titled "Healthcare expenditure, 1995 to 2005" displays the **total healthcare expenditure as a share of national GDP** for three different countries: Afghanistan, Andorra, and Samoa. The data is sourced from the World Bank – WDI and is visualized over a 10-year period. For Afghanistan, the chart shows a significant increase in healthcare expenditure from 2002 onwards. In 2002, the expenditure was **about 7.93%** of the country's GDP. This figure rose sharply to **about 8.95% in 2004** and then slightly decreased to **about 8.2% in 2005**. Andorra's healthcare expenditure as a share of GDP started at **about 5.96% in 1995**. It increased to **about 6.25% in 1996** and continued to rise to **about 7.91% in 1998**. After a slight decrease to **about 5.92% in 2000**, there was a small increase to **about 6.11% in 2002**. The expenditure then decreased to **about 5.83% in 2004** and further to **about 5.36% in 2005**. Samoa's healthcare expenditure remained relatively stable over the 10-year period. It started at **about 4.75% in 1995**, decreased slightly to **about 4.32% in 1996**, and then increased to **about 5.3% in 1998**. The expenditure continued to rise to **about 5.45% in 2000** and then decreased to **about 5.15% in 2002**. It further decreased to **about 4.59% in 2004** and saw a minor increase to **about 4.61% in 2005**. The chart is marked with a CC BY license from Our World in Data.

D2

Proportion of labor force who are women, 1990

Female labor force as a percentage of total labor force. Labor force comprises people ages 15 and older who are economically active.



Gold table

Characteristic	Value
Euro area	39.79%
Bahrain	17.44%

Vicuna (IC-tuning) w/o stage 3 :

The bar chart compares the proportion of the labor force who are women in Euro area and Bahrain in 1990. The x-axis shows the proportion of the labor force who are women in Euro area and Bahrain, while **the y-axis shows the proportion of the labor force who are women in Bahrain**. Euro area has the highest proportion of women in the labor force, at 39.79%, while Bahrain has the lowest proportion at 17.44%. This chart indicates that Euro area has a higher proportion of women in the labor force than Bahrain, indicating that Euro area has a higher rate of female participation in the labor force.

Vicuna (IC-tuning) :

The chart is titled "Proportion of labor force who are women, 1990." It illustrates **the female labor force as a percentage of the total labor force, where the labor force comprises people ages 15 and older who are economically active**. The data is represented in a horizontal bar graph format with two bars, each corresponding to a different region.

The first bar represents the **Euro area and is colored in a shade of red**. It indicates that the proportion of the labor force that is female in the Euro area was **39.79%** in 1990. The second bar **represents Bahrain and is colored in a shade of blue**. It shows that the proportion of the labor force that is female in Bahrain was **17.44%** in 1990.

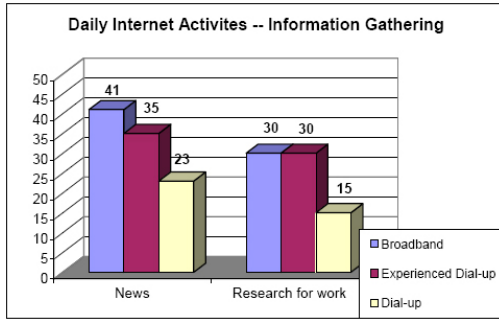
The chart includes a source attribution at the bottom, stating "**Source: OECD (2017)**" and has a "CC BY" license notation, indicating that the chart is licensed under a Creative Commons Attribution license. The chart is also marked with the logo of "Our World in Data" in the top right corner.

Figure 3. The example result on Chart Summarization.

In the third stage of IC-tuning, we further train the Vicuna and projection layers using visual-aware chart instruction-following data, as shown in Table 9, along with each 40K image and 40K language instruction-following data sampled from instruction tuning data in stage 2. In this stage, we also train the model for 3 epochs with a learning rate of 2e-5 and a batch size of 128.

For each stage, we used 16 A100 GPUs with 40GB memory, taking 5 days for stage 2 and 1 day for stage 3. Hyperparameters such as batch size and learning rate are the same as those used in LLaVA.

Q1



Question : Compare the proportion of dial-up users and broadband users who get their news on the internet.

Vicuna (IC-tuning) w/o stage 3 :

41 % of dial - up users get their news online , compared with **35 % of broadband** users .

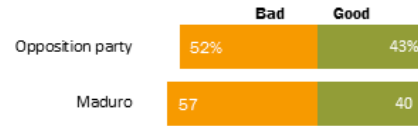
Vicuna (IC-tuning) :

To compare the proportion of dial-up users and broadband users who get their news on the internet, we look at the "News" category for both types of internet users. The chart shows that about **23% of dial-up** users and about **41% of broadband** users get their news on the internet. **Therefore, a higher proportion of broadband users compared to dial-up users get their news on the internet.**

Q2

Many View Maduro and Opposition Negatively

Influence on the way things are going in Venezuela



Source: Spring 2014 Global Attitudes survey. Q38b, Q38r.

PEW RESEARCH CENTER

Question : What is the opinion of Venezuelans towards President Maduro and the opposition party?

Vicuna (IC-tuning) w/o stage 3 :

A **52 %** majority of Venezuelans say they have a bad opinion of **Maduro** , while **43 %** say they have a good opinion .

Vicuna (IC-tuning) :

The chart shows that **a majority of Venezuelans have a negative view of both President Maduro and the opposition party**. Specifically, **57% view Maduro negatively**, and **52% view the opposition party negatively**. This indicates that, at the time of the survey, **neither Maduro nor the opposition party enjoyed the support of a majority of Venezuelans.**

Figure 4. The example result on OpenCQA.

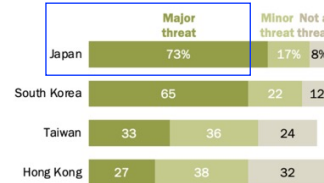
Image



Chart

North Korea's nuclear weapons program considered a serious threat in Japan, South Korea

% who say North Korea's nuclear weapons program is a ... to them



Note: Those who did not answer are not shown.
Source: Survey conducted June 2-Sept. 17, 2023, among adults in five East Asian publics.

PEW RESEARCH CENTER

Text

...
In parts of East Asia, roughly half of adults or more view China's power and influence as a major threat, according to a Pew Research Center survey conducted between June and September 2023.

...
And large majorities in all places surveyed call China at least a *minor* threat.

In Japan, **76%** of adults consider China a major threat.

...

Question : What percentage of people in Japan consider North Korea's nuclear weapons program to be a major threat?

LLaVA :

In Japan, **76%** of adults consider China's power and influence as a major threat, according to a Pew Research Center survey conducted between June and September 2023.

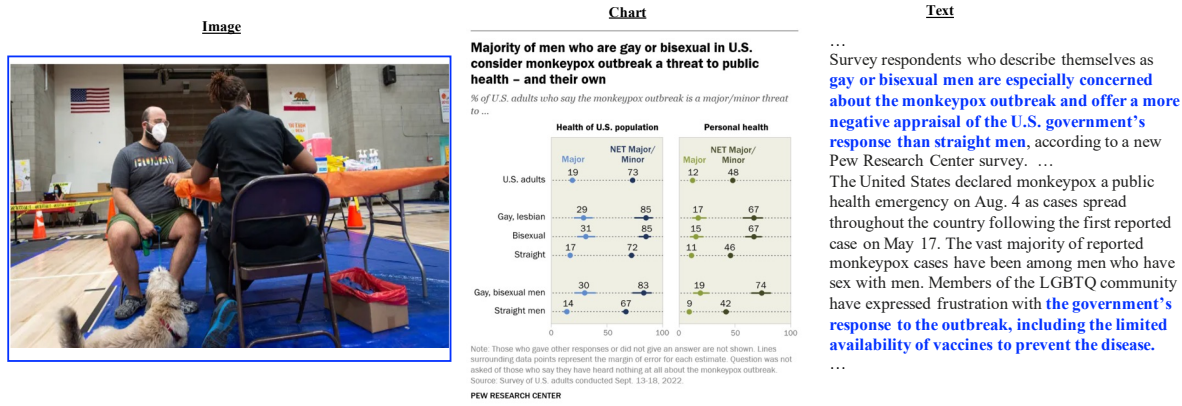
Vicuna (IC-tuning) :

To find the specific percentage of people in Japan who consider North Korea's nuclear weapons program to be a major threat, we look directly at the data provided for Japan. According to the chart, **73%** of adults in Japan say North Korea's nuclear weapons program is a major threat to their country. Therefore, the answer is **73%**.

Figure 5. The example result on single-modal QA of ICT-QA

E. Details of ICT-QA

We crawled articles and filtered them to include only those composed of 500 to 3000 tokens, in order to eliminate articles that are too short or too long. And then we used these data for our multi-modal contexts. The statistics of ICT-QA dataset are presented in Table 10. The detailed statistics for ICT-QA dataset are presented in Table 11. And Figure 13 shows the details



Question : How does this image relate to the concerns expressed by gay or bisexual men regarding the monkeypox vaccine response?

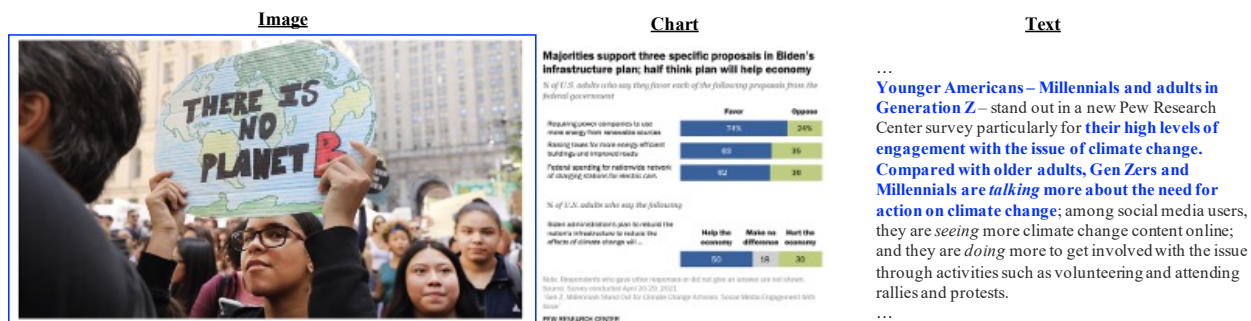
Vicuna (IC-tuning) :

The image shows a man and **a dog sitting in chairs**, with the man being gay or bisexual, **as indicated by the labels next to the man and dog. This image is related to the concerns expressed by gay or bisexual men regarding the monkeypox vaccine response in that it visually represents a segment of the population that is more likely to be concerned about the outbreak and their personal health threat perception.** The chart provided in the document shows that gay or bisexual men are more likely to say the monkeypox outbreak is a threat to U.S. public health and to their personal health, and they are also more likely to say they will get a monkeypox vaccine. This suggests that the concerns expressed by gay or bisexual men, as highlighted in the image, are reflected in the data provided.

Vicuna (IC-tuning + ICT-QA) :

The image illustrates a **healthcare worker administering a vaccine to a man**, which directly relates to the context of the text discussing the **monkeypox vaccine response**. The image visually represents the availability of vaccines and the ongoing efforts to vaccinate the population, **particularly emphasizing the importance of the vaccine in the context of the outbreak and the concerns expressed by gay or bisexual men about the response to the outbreak.**

Figure 6. The example result on cross-modal QA of ICT-QA



Question : Why might this image have been included in the document discussing the generational analysis of climate change activism?

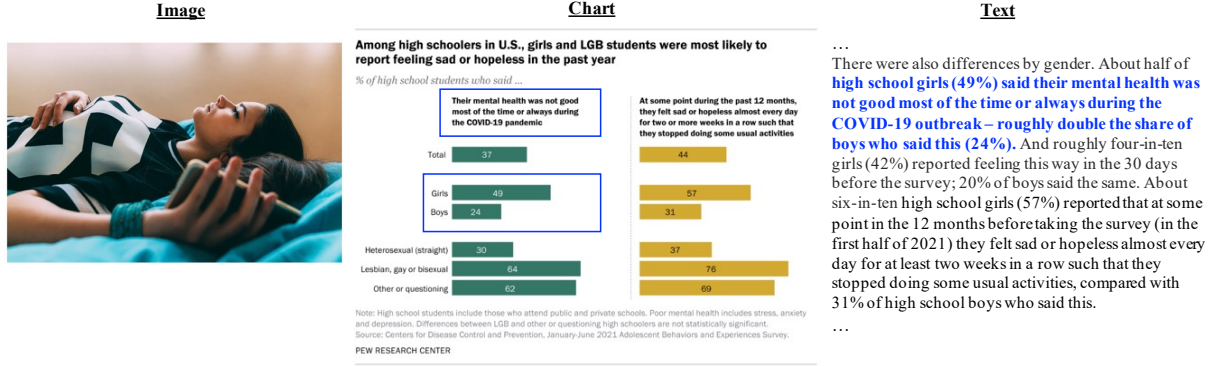
Vicuna (IC-tuning + ICT-QA) :

This image might have been included in the document to visually represent **the active engagement of younger generations in climate change activism, as described in the text.** It serves as a real-world example of **how younger Americans, including those from Generation Z, are becoming more involved in addressing climate change issues.** The image helps to underscore the text's point about **the varying levels of support for climate action among different age groups and highlights the role of younger individuals as leaders in the climate debate.**

Figure 7. The example result on image QA of ICT-QA

of ICT-QA and provides examples of both single-modal and cross-modal questions within multi-modal contexts.

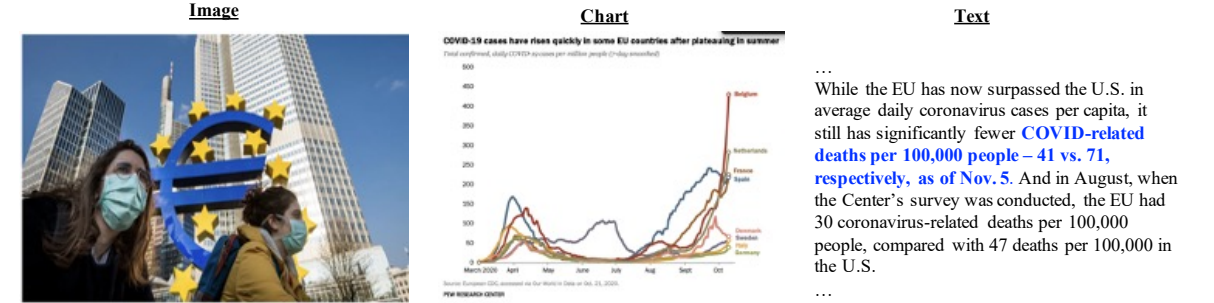
We trained LLM with ICT-QA dataset for 3 epochs with a batch size of 64, using 16 A100 GPUs 40GB, and took about 3 hours.



Question : How does the percentage of high school girls reporting poor mental health during the pandemic compare to that of high school boys?

Vicuna (IC-tuning + ICT-QA) :
 The chart shows that **a higher percentage of high school girls reported poor mental health during the pandemic (49%) compared to high school boys (24%)**, indicating that girls were more likely to report this issue.

Figure 8. The example result on chart QA of ICT-QA



Question : What were the COVID-19 related death rates per 100,000 people in the EU and the U.S. as of November 5, according to the text?

Vicuna (IC-tuning + ICT-QA) :
As of November 5, the COVID-19 related death rate per 100,000 people in the EU was 41, which is significantly lower than the rate in the U.S., which was 71. This information **is derived from the text**, which compares the death rates between the EU and the U.S. at a later stage in the pandemic.

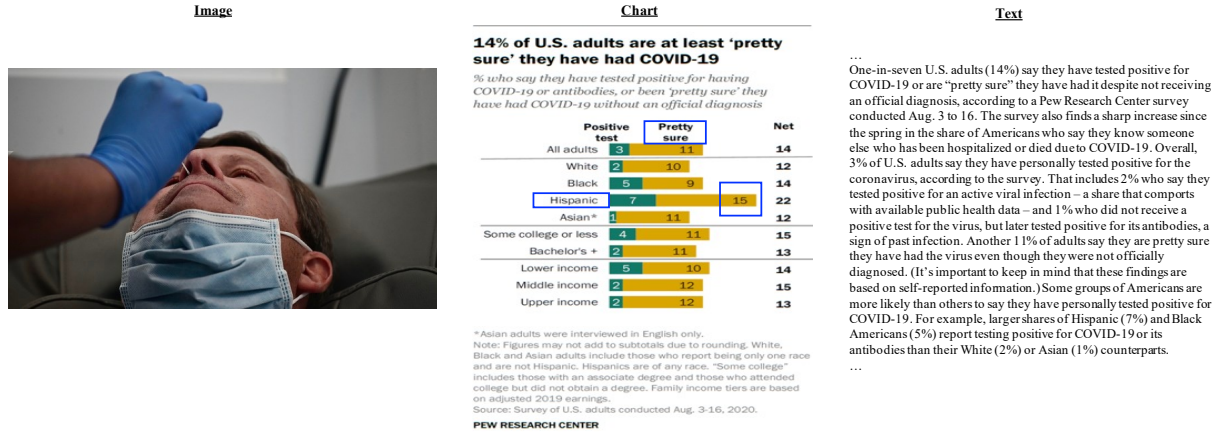
Figure 9. The example result on text QA of ICT-QA

Modality	Data	Size
Image	CC3M [37]	595K
Chart	Chart Summarization [45]	480K
	Data Table Generation [45]	480K
Total		1555K

Table 7. A mixture of modality feature alignment data in Stage 1

F. Benchmarks

Chart Benchmarks we evaluated our model using traditional chart benchmarks: ChartQA [44], Chart-to-text [21], Chart-to-table [44] and OpenCQA [20]. The metrics we use are relaxed accuracy for ChartQA, BLEU [51] for Chart-to-text,



Question : What percentage of Hispanic U.S. adults are 'pretty sure' they have had COVID-19 according to the survey?

LLaVA :

According to the survey, **7% of Hispanic U.S. adults say they have tested positive for COVID-19 or are "pretty sure"** they have had it despite not receiving an official diagnosis.

ChartLlama :

14,14,14 ...

Vicuna (IC-tuning) :

To find the specific percentage of Hispanic U.S. adults who are 'pretty sure' they have had COVID-19, we look directly at the data provided for the Hispanic demographic. According to the chart, **9% of Hispanic U.S. adults are 'pretty sure'** they have had COVID-19. Therefore, the answer is **9%**.

Vicuna (IC-tuning + ICT-QA) :

According to the survey, **15% of Hispanic U.S. adults are 'pretty sure'** they have had COVID-19. This information is derived from the specific section of the chart that details the responses of Hispanic individuals, which shows **15% falling into the 'Pretty sure' category**.

Figure 10. Comparison of ICT-QA results by models. Red text indicates factual errors or incorrect parts.

Relative Number Set Similarity (RNSS) [44] and Relative Mapping Similarity (RMS) [34] for Chart-to-table, and BLEU for OpenCQA. We compare our model with six other models: Pix2Struct [26], Matcha [35], UniChart [45], ChartInstruct-Llama [46], ChartLlama [15], and ChartAst [47]. Pix2Struct is a pre-trained image-to-text model, and Matcha extends Pix2Struct by integrating chart derendering and mathematical reasoning. UniChart is a model trained on chart-related task data, using a chart encoder and text decoder. ChartInstruct, ChartLlama, and ChartAst are LLMs trained on the various chart instruction-following data they propose.

Vision benchmarks We evaluated the performance of image modality across 12 benchmarks, commonly used to assess visual understanding ability in recent vision-language models [3, 36, 37]: TextVQA [57], LLaVA-Bench (In-the-Wild) [37], VQA-v2 [13], GQA [18], VisWiz [14], ScienceQA-IMG [41], POPE [32], MME [12], MMBench [39], MMBench-Chinese [39], SEED-Bench [28], and MM-Vet [65]. In Table 12, the names of benchmarks are abbreviated as follows: LLaVA-B refers to LLaVA-Bench (In-the-Wild); SQA to ScienceQA; MMB to MMBench; MMB-C to MMBench-Chinese; and SEED to SEED-Bench.

Language benchmarks We assessed the performance of text modality on six benchmarks selected from Hugging Face Open LLM Leaderboard¹⁶, primarily used for official evaluation of various LLMs: ARC [9], HellaSwag [66], MMLU [17], TruthfulQA [33], WinoGrande [53], and GSM8k [10]. The results of this assessment are presented in Table 13).

G. Prompts for generating visual-aware chart instruction-following data

G.1. Chart Summarization data

Figure 14 shows the prompt designed for generating chart summarization data using GPT-4V. **Bold** text indicates the inputs used in data generation. Underlined text indicate the parts that need to be modified in the following case: when using only charts without tables, the prompt excludes underlined parts.

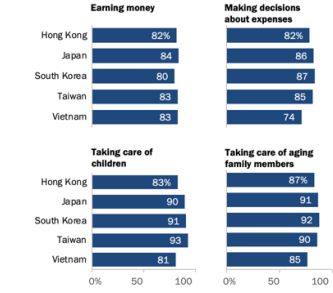
¹⁶https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Image



Chart

In East Asia, most adults view both men and women as responsible for finances and caring for others
% who say both men and women should be primarily responsible for ____ in a family



Source: Survey conducted June 2-Sept. 17, 2023, in five East Asian publics.
PEW RESEARCH CENTER

Text

...
In parts of East Asia, large majorities say men and women should share familial responsibilities, according to a Pew Research Center survey conducted June 2-Sept. 17, 2023. The responsibilities we asked about are earning money, making decisions about expenses, taking care of children and taking care of aging family members. Around three-quarters or more in Hong Kong, Japan, South Korea, Taiwan and Vietnam say men and women should share responsibility for every task asked about. On some measures, like earning money, 80% or more in each place say this.
...

Question : How does the image relate to the text's discussion of shared familial responsibilities in East Asia?

LLaVA :

The image shows a man and a woman standing next to a baby in a stroller, with both of them looking at their cell phones. This scene relates to the text's discussion of shared familial responsibilities in East Asia as it highlights the prevalence of smartphones and technology use in daily life, even during moments that should be focused on family bonding and care. The image can be interpreted as a metaphor for the challenges faced by families in East Asia, where the pressure to balance traditional family roles with modern technology and societal expectations can lead to a lack of attention and care for each other. The man and woman in the image might be representatives of the larger East Asian society, where the emphasis on individualism and achievement can sometimes overshadow the importance of family and interpersonal relationships.

ChartLLama :

The image shows a family of four, consisting of a man, a woman, and two children, standing on a city street. They are all looking at their cell phones, which may suggest that they are sharing responsibilities in terms of managing their daily lives and staying connected with others. The text discusses the views of adults in East Asia on shared familial responsibilities, specifically in relation to earning money, making decisions about expenses, taking care of children, and taking care of aging family members. The image of the family in the city street may represent a modern East Asian family that is actively involved in managing their daily lives and responsibilities, which aligns with the findings of the Pew Research Center survey. The image also highlights the prevalence of smartphones in modern society, as the family is seen using their cell phones while standing together. This could indicate that technology plays a significant role in their daily lives, and they rely on it to stay connected and manage their responsibilities. In conclusion, the image of the family of four looking at their cell phones on the city street relates to the text's discussion of shared familial responsibilities in East Asia by showcasing a modern family that is actively involved in managing their daily lives and responsibilities, which aligns with the findings of the Pew Research Center survey. The image also highlights the role of technology in modern society and how it may impact the way families manage their responsibilities.

Vicuna (IC-tuning) :

The image shows a group of people, including both men and women, engaging in different activities such as using a cell phone, pushing a stroller, and carrying a handbag. The presence of both men and women in the image suggests that the shared responsibilities in East Asia, as discussed in the text, are not limited to one gender but are rather distributed between both men and women. The image visually represents the idea that in East Asian societies, adults from both genders are expected to contribute equally to the care of children and the aging, as well as managing finances and making decisions about expenses. The diversity in the activities performed by the individuals in the image also reflects the variety of roles and responsibilities that are shared between men and women in these societies.

Vicuna (IC-tuning + ICT-QA) :

The image illustrates a scenario where both parents are involved in the upbringing and care of their child, which aligns with the text's emphasis on the shared responsibilities of men and women in East Asian families. The image visually represents the concept that in East Asia, as described in the survey, most adults believe that men and women should contribute equally to child care and other familial tasks.

Figure 11. Comparison of ICT-QA results by models. Red text indicates factual errors or incorrect parts.

G.2. Open-ended QA data

Figure 15 shows the prompt designed for generating open-ended QA data using GPT-4. Following the generation of chart summarization data (Appendix G.1), the generated summaries of charts are used as input to create open-ended QA data. {Description of chart} denotes the summary of the chart data generated via Appendix G.1.

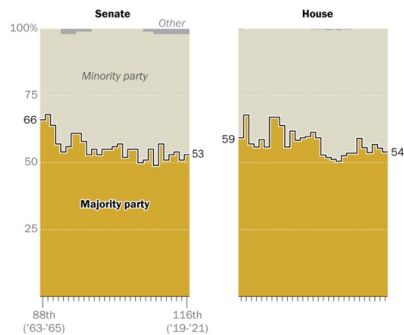
G.3. Reasoning QA data

Figure 16 shows the prompt for generating Reasoning QA data using GPT-4. This prompt is designed referencing to the prompt used in Deplot [34].

E1

Majorities in the House and Senate have slimmed down over time

% of seats held by majority/minority party, by Congress



Note: The data for which party held a majority in a Congress is from the beginning of each session – it does not take into account any shifts during that session.
Source: U.S. House of Representatives and U.S. Senate historical records.

PEW RESEARCH CENTER

Question : What was the peak percentage of seats held by the majority party in the Senate during the range of Congresses shown?

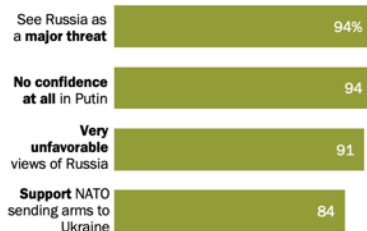
Vicuna (IC-tuning + ICT-QA) :

The peak percentage of seats held by the majority party in the Senate during the range of Congresses shown was 66%, as indicated by the highest point **on the orange line** at the beginning of the timeline.

E3

Poles view Russia as a major threat and support NATO sending arms to Ukraine

% who (have) ...



Source: Spring 2022 Global Attitudes Survey. Q5f, 18c, 43c & 46a.
"Spotlight on Poland: Negative Views of Russia Surge, but Ratings for U.S., NATO, EU Improve"

PEW RESEARCH CENTER

Question : How does the support for NATO sending arms to Ukraine compare to the no confidence level in Putin among Poles?

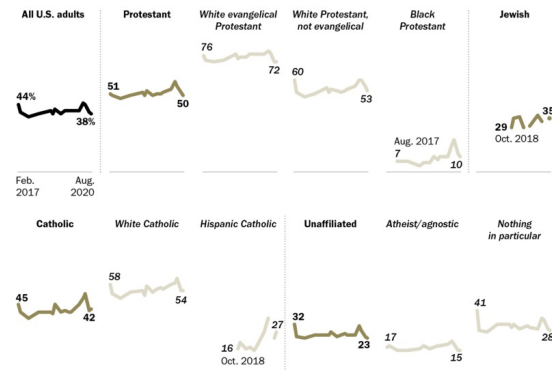
Vicuna (IC-tuning + ICT-QA) :

The support for NATO sending arms to Ukraine is **significantly higher than** the no confidence level in Putin among Poles. The chart shows that 84% support this action, while 94% have no confidence in Putin, indicating a much smaller percentage in support of the action compared to the overwhelming lack of confidence in Putin.

E2

Trump's job approval rating consistently highest among White Christians, far lower among other groups

% of U.S. adults who approve of Donald Trump's handling of his job



Note: Gaps in line charts indicate an insufficient sample size for the group in question on the ATP wave indicated.
White and Black adults include those who report being only one race and are not Hispanic. Hispanics are of any race.
Source: Pew Research Center surveys conducted 2017-2020.

PEW RESEARCH CENTER

Question : How do the approval ratings of White evangelical Protestants compare with those of Hispanic Catholics as depicted in the chart?

Vicuna (IC-tuning + ICT-QA) :

As depicted in the chart, the approval rating of White evangelical Protestants is consistently higher than that of Hispanic Catholics throughout the time period shown. White evangelical Protestants start with an approval rating of **50%** in February 2017 and end with 72% in October 2018, while Hispanic Catholics begin at **45%** in February 2017 and fluctuate slightly, ending at **54%** in October 2018.

E4



Question : What are the major themes depicted in each part of the composite image?

Vicuna (IC-tuning + ICT-QA) :

The left part of the image depicts a scene of voting, with individuals standing at voting booths and **a sign that reads "I Voted!"** The right part of the image shows **a scene of a voting location with people voting, and a sign that reads "VOTE HERE."** The bottom part of the image shows a scene of a voting location with people voting and a sign that reads **"VOTE HERE."** The top part of the image shows a scene of a voting location with people voting and a sign that reads **"VOTE HERE."**

Figure 12. The examples of error analysis from ICT-QA. Red text indicates errors

Task	Data	Size	All
Chart modality instruction-following data			
Chart Question Answering	ChartQA [44]	30k	207k
	PlotQA [48]	85k	
	Unichart-pretrain [45]	85k	
	OpenCQA [20]	7k	7k
Chart Description	Unichart-pretrain	163k	200k
	Chart-to-text [21]	37k	
Data table generation	Unichart-pretrain	181k	200k
	ChartQA	19k	
Image & language modality instruction-following data			
LLaVA 1.5 Mixture [36]	LLaVA [37]	158K	665K
	ShareGPT [1]	40K	
	VQAv2 [13]	83K	
	GQA [18]	72K	
	OKVQA [43]	9K	
	OCRVQA [49]	80K	
	A-OKVQA [55]	66K	
	TextCaps [56]	22K	
	RefCOCO [23, 42]	48K	
	VG [25]	66K	
Total			1272K

Table 8. A mixture of multi-modal instruction-following data in Stage 2

Data source	#Chart	Chart Summarization	Open ended QA	Reasoning QA	Total
Unichart-pretrain	27,499	27,499	20,930	6,976	55,405
ChartQA	1,848	-	-	1,901	1,901
Total	29,347	27,499	20,930	8,877	57,306

Table 9. The statistics of visual-aware chart instruction-following data.

	Single-modal QA	Cross-modal QA	Total
Image QA	544	1,088	1,632
Chart QA	1,872	936	2,808
Text QA	2,889	-	2,889
Total	5,305	2,024	7,329

Table 10. The statistics of ICT-QA

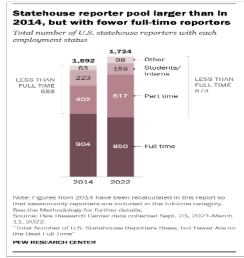
H. Prompts for generating ICT-QA

Figure 17 shows the prompt designed for generating ICT-QA dataset using GPT-4V. In image and chart QA, **{Context}** refers to the paragraphs near the corresponding image and chart, respectively, while for text QA, **{Context}** represents the

Image



Chart



Text

A new Pew Research Center study finds that the total number of reporters assigned to the 50 state capitols to inform citizens about legislative and administrative activity has increased by 8% since 2014, the last time this study was conducted. The gain comes largely from two main developments: new nonprofit news outlets that are employing statehouse reporters, and a shift to more part-time statehouse reporting.

Single modal question example

- What does the image imply about the characteristics of the statehouse reporters present?
- What are the primary sources of funding for the report on the state of news, information, and journalism in the digital age?
- How many full-time statehouse reporters were recorded in 2014 compared to 2022?

Cross-modal question example

- How is the scene in the image related to the context provided by the text?
- Why is the chart included in the document?

Figure 13. The example of ICT-QA dataset.

	# QA data				# Article
	Image QA	Chart QA	Text QA	Total	
Train set	1,467	2,529	2,598	6,594	866
Test set	165	279	291	735	97
Total	1,632	2,808	2,889	7,329	963

Table 11. The statistics of train and test sets for ICT-QA

	TextVQA	LLaVA-B	VQA-v2	GQA	VisWiz	SQA	POPE	MME	MMB	MMB-C	SEED	MM-Vet	Avg.
Continual learning													
LLaVA	58.26	66.10	78.50	61.97	50.10	68.12	<u>85.97</u>	1507.88(75.39)	65.03	58.42	60.91	<u>32.10</u>	63.41
- CLIP post-tuning	52.13	56.40	74.60	60.45	51.77	66.73	82.34	1419.73(70.99)	61.34	55.15	58.31	28.20	59.87
- Unichart tuning	55.91	65.40	77.17	<u>61.33</u>	51.40	68.02	85.36	1447.54(72.38)	64.18	<u>56.36</u>	60.37	28.80	62.22
End-to-end													
Vicuna	-	-	-	-	-	-	-	-	-	-	-	-	-
- CLIP(image, chart) tuning	56.77	64.50	76.90	59.68	<u>52.00</u>	69.91	84.05	1416.84(70.84)	62.46	53.61	58.97	32.90	61.88
- CLIP(image), Unichart(chart) tuning	<u>57.62</u>	66.10	<u>77.96</u>	60.63	53.42	<u>69.16</u>	84.85	<u>1465.25(73.26)</u>	<u>64.60</u>	55.32	<u>60.57</u>	32.00	<u>62.96</u>

Table 12. The ablation studies on IC-tuning for Vision benchmarks. In the case of MME, the number in parentheses indicates normalized performance and the average performance (Avg.) was calculated using it.

remaining text within the full text, excluding the paragraphs used in generating image and chart QA data.

I. Criteria for G-Eval

For G-Eval on the chart summarization, we adopt the following three evaluation criteria. *Factual Correctness* measures whether the generated description do not factually contradict the charts. *Informativeness* measures how well the generated descriptions cover the information contained in the chart and how informative they are about the chart. Finally, *Fluency* evaluates the quality of the generated answers in terms of grammar, spelling, punctuation, word choice, and sentence structure.

For G-Eval on OpenCQA and ICT-QA, we adopt the following three evaluation criteria: *Factual Correctness*, *Relevance*, and *Fluency*. *Factual Correctness* measures whether the generated answers factually align with the given context. In the case

	ARC	HellaSwag	TruthfulQA	MMLU	Winogrande	GSM8k	Avg.
Continual learning							
LLaVA	<u>52.82</u>	76.07	45.88	51.77	71.43	15.62	52.27
- <i>CLIP post-tuning</i>	49.23	73.23	40.93	47.92	70.72	9.78	48.64
- <i>Unichart tuning</i>	50.60	75.01	42.26	50.80	<u>72.53</u>	10.16	50.23
End-to-end							
Vicuna	53.16	77.47	50.37	51.14	71.82	18.12	53.68
- <i>CLIP(image, chart) tuning</i>	51.96	76.43	<u>50.31</u>	<u>51.87</u>	72.38	15.16	53.02
- <i>CLIP(image), Unichart(chart) tuning</i>	51.62	<u>76.50</u>	48.59	52.76	73.32	<u>16.68</u>	<u>53.25</u>

Table 13. The ablation studies on IC-tuning for Language benchmarks

Chart: {Chart}

Table: {Table}

Given chart and its table, your goal is to generate detailed description for the chart.

Your description must not mention "table" and as if it were generated from chart alone.

If the chart contains text, you must copy that text.

The description should include all numeric values shown in the chart.

If you estimate numeric value from a chart where the numbers are not displayed, you must use the term "about" to indicate the estimated value in the description.

Figure 14. The prompt for generating Chart summarization data

Chart: {Description of chart}

Given chart, generate 4 questions and answers.

You should start with "Question_1: " and "Answer_1: "

Question_1 requires identifying a specific target or describing the characteristics of the specific target.

Question_2 requires comparisons between specified targets.

Question_3 requires mathematical or logical operations such as average, sum, difference and so on.

Question_4 requires describing the trends shown in the chart.

Question must be simple enough.

For Question_1, Question_2 and Question_3, answer includes step by step thinking and a final answer.

For Question_4, answer must be detailed enough but do not use external knowledge to explain the trends in the chart.

Figure 15. The prompt for generating Open-ended QA data

of OpenCQA, the context refers to charts and titles, whereas in ICT-QA, it refers the entire multi-modal contexts including images, charts, and text. *Relevance* measures how consistent the generated answers are with the question and the given context. *Fluency* is the same criteria for chart summarization task.

J. G-Eval results on Chart Summarization and OpenCQA

We present the G-Eval results for chart summarization and OpenCQA in Table 14, including scores for each evaluation criteria.

You are responsible to generate answer for given chart and question.
 Note that chart is converted to table. Assume that given table is chart, not table.
 Let's think step by step. And then generate answer after "The answer is ".

[Chart]

Year	Democrats	Republicans	Independents
2004	68.1%	45.0%	53.0%
2006	58.0%	42.0%	53.0%
2007	59.0%	38.0%	45.0%
2009	72.0%	49.0%	60.0%
2011	71.0%	51.2%	58.0%
2012	70.0%	48.0%	53.0%
2013	72.0%	41.0%	60.0%

Q: In which year republicans have the lowest favor rate?

A: Let's find the column of republicans. Then let's extract the favor rates, they [45.0, 42.0, 38.0, 49.0, 51.2, 48.0, 41.0]. The smallest number is 38.0, that's Row 3. Row 3 is year 2007. The answer is 2007.

Q: What is the sum of Democrats' favor rates of 2004, 2012, and 2013?

A: Let's find the rows of years 2004, 2012, and 2013. We find Row 1, 6, 7. The favor rates of Democrats on that 3 rows are 68.1, 70.0, and 72.0. $68.1+70.0+72=210.1$. The answer is 210.1.

Q: By how many points do Independents surpass Republicans in the year of 2011?

A: Let's find the row with year = 2011. We find Row 5. We extract Independents and Republicans' numbers. They are 58.0 and 51.2. $58.0-51.2=6.8$. The answer is 6.8.

Q: Which group has the overall worst performance?

A: Let's sample a couple of years. In Row 1, year 2004, we find Republicans having the lowest favor rate 45.0 (since $45.0 < 68.1$, $45.0 < 53.0$). In year 2006, Row 2, we find Republicans having the lowest favor rate 42.0 ($42.0 < 58.0$, $42.0 < 53.0$). The trend continues to other years. The answer is Republicans.

Q: Which party has the second highest favor rates in 2007?

A: Let's find the row of year 2007, that's Row 3. Let's extract the numbers on Row 3: [59.0, 38.0, 45.0]. 45.0 is the second highest. 45.0 is the number of Independents. The answer is Independents.

[Chart]

{Table}

Q: {Question}

Figure 16. The prompt for generating reasoning QA data

K. G-Eval results on ICT-QA

Table 15 shows the G-Eval results on ICT-QA test data, including scores for each evaluation criteria. The *w/o chart* category represents the case in image QA where only image and text modalities are given as input, without chart modality. The *Total* indicates the performance averaged across all categories, including image QA (excluding *w/o chart*), chart QA, and text QA.

{Image}

Title: **{Article title}**

Context: **{Context}**

First, generate detailed description of given chart. (starting with “Detailed description of the chart:”)

Second, generate 3 questions and answers about the chart in the context of given text. Generated question and answer should start with “Question_1: “ and “Answer_1”:

Question should ask about information described in the chart. Question should not contain phrases that refer to charts, such as “on this chart.” Question_1 requires identifying a specific target or describing the characteristics of the specific target. Question_2 requires comparisons between specified targets. Question_3 why the chart is included in document. Answer should include only the information described in the chart and text. Answer should be detailed. Answer should explain where the evidence was derived from chart or text.

(a) Image QA

{Chart}

Title: **{Article title}**

Context: **{Context}**

First, generate detailed description of given image. (starting with “Detailed description of the image:”)

Second, generate 3 questions and answers about the image in the context of given text. Generated question and answer should start with “Question_1: “ and “Answer_1”:

Question should ask about information described in the image in the context of text. Question_1 ask identifying the target or describing the characteristics of the specific target. Question_2 ask how the specific target of image relates to the surrounding text. Question_3 ask why the image is included in document.

Answer should include only the information described in the image and text. Answer should be detailed. Answer should explain where the evidence was derived from image or text.

(b) Chart QA

Title: **{Article title}**

Context: **{Context}**

Generate 3 questions and answers. Generated question and answer should start with “Question_1: “ and “Answer_1”:

Question should ask about information described in the given text. Question_1 requires identifying a specific target or describing the characteristics of the specific target. Question_2 requires reasoning. Question_3 requires summarizing. Answer should include only the information described in the text. Answer should be detailed. Answer should explain where the evidence was derived from text.

(c) Text QA

Figure 17. The prompts for generating ICT-QA

L. Prompts for G-Eval

We used GPT-4o, which can handle both visual and textual inputs, for the evaluation of chart summarization, OpenCQA, and ICT-QA tasks. In Figures 18 to 20, **bold** text indicates the input conditions for the evaluation.

L.1. Chart Summarization

Figure 18 is the prompts for G-Eval on chart summarization task. Underlined text indicate the parts that need to be modified in the following case: for charts without a corresponding table, the underlined parts are excluded from the prompt.

L.2. OpenCQA

Figure 19 is the prompts for G-Eval on OpenCQA.

Model	Chart Summarization				OpenCQA			
	Factual correctness	Informativeness	Fluency	Avg.	Factual correctness	Relevance	Fluency	Avg.
ChartLlama	1.53	2.53	3.55	2.54	1.89	2.29	3.51	2.56
ChartAst	2.91	3.17	3.76	3.28	3.55	3.76	4.13	3.81
Vicuna (IC-tuning)	3.48	3.93	4.15	3.85	3.84	4.10	4.38	4.11
- w/o unichart	2.68	3.11	3.62	3.14	3.21	3.46	3.94	3.54
- w/o stage 2	<u>3.45</u>	<u>3.88</u>	<u>4.12</u>	<u>3.82</u>	<u>3.65</u>	<u>4.05</u>	<u>4.32</u>	<u>4.01</u>
- w/o stage 3	3.06	3.35	3.84	3.41	3.36	3.54	4.07	3.65

Table 14. Detailed G-Eval results on Chart Summarization and OpenCQA

Model	Task	Input			GPT-4			
		Image	Chart	Text	Factual correctness	Relevance	Fluency	Avg.
LLaVA	Image QA	✓		✓	3.65	3.74	4.38	3.92
	Chart QA		✓	✓	3.36	3.68	4.35	3.80
	Text QA	✓		✓	3.85	3.90	4.42	4.05
	Total				3.62	3.78	4.38	3.93
LLaVA-NeXT-Interleave	Image QA	✓	✓	✓	3.35	3.58	4.19	3.71
	- w/o chart	✓		✓	3.90	3.96	4.44	4.10
	Chart QA		✓	✓	3.29	3.75	4.20	3.75
	Text QA		✓	✓	3.52	3.82	4.19	3.84
	Total				3.39	3.74	4.19	3.78
Vicuna (IC-Tuning)	Image QA	✓	✓	✓	2.77	3.16	4.09	3.34
	- w/o chart	✓		✓	3.70	3.82	4.40	3.97
	Chart QA	✓	✓	✓	3.53	4.00	4.28	3.93
	Text QA	✓	✓	✓	3.77	4.02	4.27	4.02
	Total				3.45	3.82	4.23	3.83
Vicuna (IC-Tuning + ICT-QA)	Image QA	✓	✓	✓	4.04	4.05	4.40	4.17
	Chart QA	✓	✓	✓	4.02	4.23	4.44	4.23
	Text QA	✓	✓	✓	4.56	4.46	4.50	4.51
	Total				4.24	4.28	4.46	4.33
Vicuna (IC-Tuning + ICT-QA) - w/o sp token, interleave	Image QA	✓	✓	✓	3.85	3.94	4.30	4.03
	- w/o chart	✓		✓	4.05	4.02	4.45	4.17
	Chart QA	✓	✓	✓	3.88	4.15	4.33	4.12
	Text QA	✓	✓	✓	4.39	4.36	4.44	4.40
	Total				4.08	4.18	4.37	4.21

Table 15. Detailed G-Eval results on ICT-QA

L.3. ICT-QA

Figure 20 is the prompts for G-Eval on ICT-QA. For evaluation, each description generated during the ICT-QA data construction process, as well as images and charts, are used as prompt inputs. In cases such as LLaVA and ChartLlama, which can only use either an image or a chart as an input, the prompt slightly differs from Figure 20. Underlined text indicate the

You are an expert evaluator assessing the performance of detailed description of the chart and its table. Your task is to evaluate the detailed description. You will receive the chart and its table for reference.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. You must evaluate each evaluation criteria independently, making sure that the scores between the criteria are not influenced by each other.

Evaluation Criteria:

Factual correctness(1-5)- This metric measures the extent of the factual alignment between the description and the chart, and the extent to which the description contains statements entailed by the chart.

Informativeness(1-5)- This metric measures how much the description covers the information contained in the chart. Ideally, an informative description should contain high-level insights from the chart.

Fluency(1-5) - This metric measures the quality of the description in terms of grammar, spelling, punctuation, word choice, and sentence structure.

Evaluation Steps:

Factual correctness:

1. Look at the chart and its table carefully and identify the main facts and details it presents.
2. Read the detailed description and compare it to the chart and its table. Check if the description contains any factual errors that are not supported by the chart and its table.
3. Assign a score for Factual correctness from 1 to 5 based on the Evaluation Criteria.

Informativeness:

1. Look at the chart and its table carefully and identify the information and the main insights it presents.
2. Check the detailed description to ensure it sufficiently covers the information presented in the chart and its table.
3. Consider whether the description provides insightful analysis or context that enhances understanding of the data presented in the chart.
4. Assign a Informativeness score from 1 to 5.

Fluency:

1. Read the detailed description and identify any grammatical errors, misspellings, punctuation mistakes, awkward phrasing, or unclear expressions within the description.
2. Evaluate the overall quality of the writing in terms of its readability, considering how effectively the description conveys its message to the reader.
3. Assign a Fluency score from 1 to 5.

Example:

Chart:

{Chart}

Table:

{Table}

Detailed description:

{Generated description}

The output should first give the average score based on three criteria, then output scores for each criteria with your reason about scores. The output should follow this format:

Average: number

Factual correctness: number

Reason for Factual correctness: text

Informativeness: number

Reason for Informativeness: text

Fluency: number

Reason for Fluency: text

Figure 18. The prompt for G-Eval on Chart Summarization

parts that need to be modified in the following cases: when utilizing only images and text as input modalities for the model, the prompt excludes underlined parts related to charts; conversely, when utilizing only charts and text, it excludes underlined parts related to images.

You are an expert evaluator assessing the performance of multi-modal LLM's response to questions based on the chart. Your task is to evaluate the response generated by multi-modal LLM. You will receive the question, the chart and its title for reference. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. You must evaluate each evaluation criteria independently, making sure that the scores between the criteria are not influenced by each other.

Evaluation Criteria:

Factual correctness(1-5)- This metric measures the extent of the factual alignment between the response and the chart, and the extent to which the response contains statements entailed by the chart.

Relevance(1-5) - This metric measures how consistent the response is with respect to the question and the chart.

Fluency(1-5) - This metric measures the quality of the response in terms of grammar, spelling, punctuation, word choice, and sentence structure.

Evaluation Steps:

Factual correctness:

1. Look at the chart and its title carefully and identify the main facts and details it presents.
2. Read the response and compare it to the chart and its title. Check if the response contains any factual errors that are not supported by the chart and its title.
3. Assign a score for Factual correctness from 1 to 5 based on the Evaluation Criteria.

Relevance:

1. Read the question, response and the chart carefully.
2. Compare the response to the chart with question and identify the main points of the chart and its title.
3. Assess how well the response covers the main points of the chart, and how much irrelevant or redundant information it contains.
4. Assign a Relevance score from 1 to 5.

Fluency:

1. Read the response and identify any grammatical errors, misspellings, punctuation mistakes, awkward phrasing, or unclear expressions within the response.
2. Evaluate the overall quality of the writing in terms of its readability, considering how effectively the response conveys its message to the reader.
3. Assign a Fluency score from 1 to 5.

Example:

Chart:

{Chart}

Title:

{Title}

Question:

{Question}

Response generated by the multi-modal LLM:

{Generated answer}

The output should first give the average score based on three criteria, then output scores for each criteria with your reason about scores. The output should follow this format:

Average: number

Factual correctness: number

Reason for Factual correctness: text

Relevance: number

Reason for Relevance: text

Fluency: number

Reason for Fluency: text

Figure 19. The prompt for G-Eval on OpenCQA

You are an expert evaluator assessing the performance of multi-modal LLM's response to questions based on documents consisting of images, charts and text. Your task is to evaluate the response generated by multi-modal LLM. You will receive the question, the image, the chart, the detailed description of image, the detailed description of chart and full text of the document for reference.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. You must evaluate each evaluation criteria independently, making sure that the scores between the criteria are not influenced by each other.

Evaluation Criteria:

Factual correctness(1-5) - This metric measures the extent of the factual alignment between the response and the document, and the extent to which the response contains statements entailed by the document(Image, Chart, Detailed description of image/chart and Full text of document).

Relevance(1-5) - This metric measures how consistent the response is with respect to the question and the document(Image, Chart, Detailed description of image/chart and Full text of document).

Fluency(1-5) - This metric measures the quality of the response in terms of grammar, spelling, punctuation, word choice, and sentence structure.

Evaluation Steps:

Factual correctness:

1. Look at the image, the chart, the detailed description of image/chart and the full text of document carefully and identify the main facts and details it presents.
2. Read the response and compare it to the main facts and details. Check if the response contains any factual errors that are not supported by the document.
3. Assign a score for Factual correctness from 1 to 5 based on the Evaluation Criteria.

Relevance:

1. Read the question, response and the document carefully.
2. Compare the response to the document with question and identify the main points of the document.
3. Assess how well the response covers the main points of the document, and how much irrelevant or redundant information it contains.
4. Assign a Relevance score from 1 to 5.

Fluency:

1. Read the response and identify any grammatical errors, misspellings, punctuation mistakes, awkward phrasing, or unclear expressions within the response.
2. Evaluate the overall quality of the writing in terms of its readability, considering how effectively the response conveys its message to the reader.
3. Assign a Fluency score from 1 to 5.

Example:

Document:

Image:

{Image}

Chart:

{Chart}

Detailed description of image:

{Description of the image}

Detailed description of chart:

{Description of the chart}

Full text of document:

{Full text of article}

Question:

{Question}

Response generated by the multi-modal LLM:

{Generated answer}

The output should first give the average score based on three criteria, then output scores for each criteria with your reason about scores. The output should follow this format:

Average: number

Factual correctness: number

Reason for Factual correctness: text

Relevance: number

Reason for Relevance: text

Fluency: number

Reason for Fluency: text

Figure 20. The prompt for G-Eval on ICT-QA