Judging the Judges: Can Large Vision-Language Models Fairly Evaluate Chart Comprehension and Reasoning?

Md Tahmid Rahman Laskar[‡,*], Mohammed Saidul Islam[‡,†], Ridwan Mahbub[‡,†], Ahmed Masry[‡], Mizanur Rahman[‡], Md Amran Hossen Bhuiyan[‡], Mir Tafseer Nayeem[§], Shafiq Joty[¶], Enamul Hoque[‡,*], Jimmy Xiangji Huang[‡,*]
[‡]York University, [§]University of Alberta, [¶]Salesforce AI Research

Abstract

Charts are ubiquitous as they help people understand and reason with data. Recently, various downstream tasks, such as chart question answering, chart captioning, etc. have emerged. Large Vision-Language Models (LVLMs) show promise in tackling these tasks, but their qualitative evaluation is costly and time-consuming, limiting real-world deployment. While using LVLMs as judges to assess chart comprehension capabilities of other LVLMs could streamline evaluation processes, challenges like proprietary datasets, restricted access to powerful models, and evaluation costs hinder their adoption in industrial settings. To this end, we present a comprehensive evaluation of 13 open-source LVLMs as judges for diverse chart comprehension and reasoning tasks. We design both pairwise and pointwise evaluation tasks covering criteria like factual correctness, informativeness, and relevancy. Additionally, we analyze LVLM judges based on format adherence, positional consistency, length bias, and instruction-following. We focus on cost-effective LVLMs ($\leq$ 9B parameters) suitable for both research and commercial use, following a standardized evaluation protocol and rubric to measure the LVLM judge accuracy. Experimental results reveal notable variability: while some open LVLM judges achieve GPT-4-level evaluation performance (about 80% agreement with GPT-4 judgments), others struggle (below 10% agreement). Our findings highlight that state-of-the-art open-source LVLMs can serve as cost-effective automatic evaluators for chart-related tasks, though biases such as positional preference and length bias persist.

1 Introduction

Understanding data visualizations—such as bar and line charts—requires multimodal reasoning, as it involves integrating visual encodings with textual and

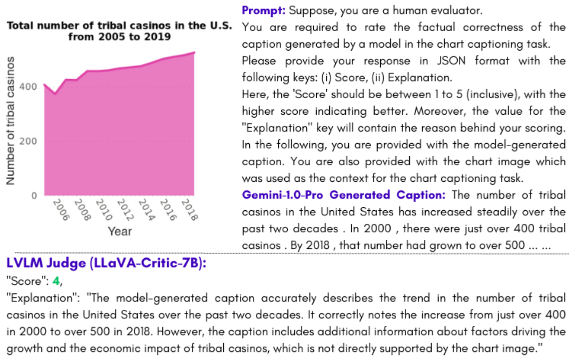


Figure 1: An example evaluation of *Gemini-1.0-Pro* model generated L2/L3 caption in the VisText dataset by an LVLM judge: the *LLaVA-Critic-7B* model.

contextual information (Hoque and Islam, 2024). Recent research has introduced various tasks (e.g., chart question answering, chart captioning, fact-checking with charts, etc.) to facilitate chart-based reasoning via natural language. These tasks demand the understanding of both the chart's visual content (data values, trends, visual encodings) and accompanying text or instructions.

Large Language Models (LLMs) have revolutionized NLP and vision-language tasks (Zhao et al., 2023), with growing interest in their use for chart comprehension and reasoning due to their strong multimodal capabilities. This progress can have a substantial impact on real-world industrial applications, where extracting insights from charts and graphs can drive critical business decisions (Obeid and Hoque, 2020; Masry et al., 2023; Meng et al., 2024). However, evaluating LLM performance in chart understanding presents notable challenges (Islam et al., 2024). For instance, traditional text-based metrics like BLEU fail to capture the quality of open-ended explanatory answers and also require human-annotated references. While human evaluation can address this problem, it is time-consuming and resource-intensive.

* Contact Emails: {tahmedge,enamulh,jhuang}@yorku.ca
† Equal Contributions.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1203–1216
July 28-30, 2025 ©2025 Association for Computational Linguistics

To address this, recent studies have proposed using LLMs themselves as automatic evaluators or judges (Gu et al., 2024; Li et al., 2024b). By employing LLMs to evaluate the chart comprehension abilities of other models (see Figure 1 for an example), the evaluation process can be streamlined, making the process more efficient and reproducible without human intervention. While this method accelerates development and reduces dependency on human annotations, its real-world adoption is hindered by privacy and scalability constraints. For example, organizations may be unwilling to share proprietary data with closed-source models from OpenAI, Google, or Anthropic. While closed-source models demonstrate impressive judging capabilities, their compatible open-source models are often large in size (e.g., 70B to 400B parameters). This requires high computing resources and usage costs. Therefore hinders real-world utilization.

To this end, this paper aims to investigate whether open-source smaller LVLMs (e.g., less than 10B parameters) can effectively evaluate answers about charts—assessing correctness, relevance, and other qualities—similarly to a human or a powerful LLM like GPT-4 (OpenAI et al., 2023). For this purpose, we conduct one of the first comprehensive evaluations of open-source LVLMs as evaluators on various chart benchmarks, consisting of diverse tasks like chart captioning and question answering. We focus on open-source, smaller VLMs (up to 10B parameters) to simulate realistic deployment scenarios where cost-effective or private models are preferred over large closed models. By benchmarking these models against high-quality reference judgments generated by closed-source LLMs like GPT-4 or 70B open-source LLM-Judge like LLaVA-Critic (Xiong et al., 2024), we aim to uncover to what extent current open models can serve as reliable judges, and when they fail.

Our major contributions to this paper are:

1. We establish an evaluation framework for chart comprehension using *"LVLM-as-a-Judge"*, with clear rubrics for pairwise and pointwise assessments over 100K judgments generated by GPT-4o and LLaVA-Critic-70B. Additionally, we introduce a new benchmark to assess the instruction-following abilities of LVLMs in chart-related tasks.

2. We evaluate a wide range of open-source multimodal LLMs as judges – 13 models ranging from 2B to 9B parameters – and analyze their performance against LLM-annotated (GPT-4 and LLaVA-Critic) and human-annotated reference judgments, across diverse chart benchmarks (OpenCQA and VisText) on answers generated by different LLMs to create challenging evaluation scenarios.

3. We provide an in-depth analysis of the judges' strengths and weaknesses, revealing issues like position bias and length bias, and discuss which models achieve substantially higher agreement with reference judgments, and which ones fail.

In addition, our code, judgment data, and our proposed instruction-following evaluation benchmark is released here: `https://github.com/tahmedge/chart_lvlm_judge`

2 Related Work

Earlier efforts in chart question answering include synthetic datasets like FigureQA (Kahou et al., 2017) and DVQA (Kafle et al., 2018), which generated templated QA pairs for simple charts but lacked real-world complexity. ChartQA (Masry et al., 2022) addressed this gap with real-world charts and more complex questions, while OpenCQA (Kantharaj et al., 2022) pushed further with open-ended, explanatory queries. Meanwhile, chart captioning has emerged as another avenue for summarizing chart content (Shankar et al., 2022; Rahman et al., 2023; Tang et al., 2023). Together, these datasets highlight the growing complexity of chart-based reasoning tasks and the need for more robust evaluation methods.

While the rise of multimodal LLMs offers potential for chart-related tasks, general vision-language models often struggle with chart-specific elements like axis text and precise data points (Islam et al., 2024). To address this, specialized models such as ChartLLaMA (Han et al., 2023), ChartInstruct (Masry et al., 2024), ChartGemma (Masry et al., 2025), and TinyChart (Zhang et al., 2024) have been developed, showing strong performance. However, evaluating these models is challenging, as many still depend on time-consuming human assessments for open-ended responses.

While using LLMs to evaluate other LLMs has gained a lot of attention, early efforts focused primarily on text-only tasks like summarization (Li et al., 2024b; Zheng et al., 2023). For multimodal tasks, models such as Prometheus-VL (Lee et al.,

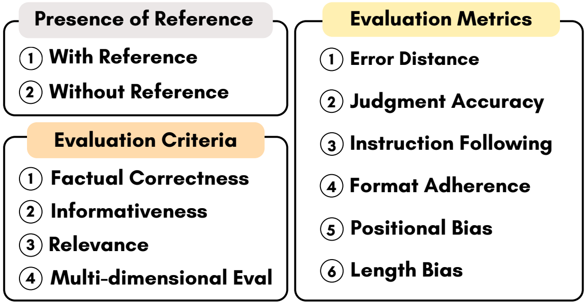


Figure 2: An overview of our evaluation methodology.

2024) and LLaVA-Critic (Xiong et al., 2024) introduced smaller open-source vision-language models (as small as 7B) fine-tuned to serve as general-purpose multimodal evaluators. Our work aligns with this direction, leveraging LVLMs as judges. Although concurrent studies explore similar capabilities (Chen et al., 2024), they report that early LVLMs like LLaVA-1.5 struggle with text-rich visuals such as charts and diagrams (Lee et al., 2024). Addressing the gap in evaluating recent LVLMs on chart-specific tasks, we present the first systematic study of state-of-the-art open-source LVLMs as judges across diverse chart comprehension and reasoning benchmarks.

3 Methodology

Given a chart and model generated response(s), we construct the prompt (see Appendix A.2 for some sample prompts) depending on the evaluation rubric. Following the prior work (Lee et al., 2024), we ask the LVLM judge to provide their answer along with an explanation, since adding an explanation during assessments ensured better judgment performance in early work. Below, we describe our evaluation method (also see Figure 2).

3.1 Evaluation Rubric Design

Following the prior work on LLM evaluation (Chen et al., 2024; Lee et al., 2024; Xiong et al., 2024), we define clear rubrics for the judges:

(i) Based on Evaluation Type:

- **Pairwise:** The judge must select the better answer between two given responses (e.g., Claude vs Gemini) about the chart for the given instruction.

- **Pointwise:** The judge must rate a single answer to the chart query on a Likert scale from 1 (very poor) to 5 (excellent).

(ii) Based on Reference Type:

- **With Reference:** The judge is also given the ground-truth answer or summary as a reference, and instructed to choose the response that better matches the reference as well as the chart context.

- **Without Reference:** The judge only sees the model response(s) and the chart image and must decide based on its own judgment.

(iii) Based on Evaluation Criteria:

- **Factual Correctness:** Focuses only on the factual accuracy of the response.

- **Informativeness:** Focuses on the amount of useful information in the response.

- **Relevance:** Focuses on measuring the relevancy of the response.

- **Multidimensional Evaluation:** Considers overall response quality based on factual correctness, informativeness, conciseness, and relevance.

(iv) Based on Evaluation Metrics:

- **Judgment Accuracy:** The percentage of instances where the answer picked by the judge same as the gold. It is relevant to the pairwise case.

- **Error Distance:** The average absolute difference between the judge's 1–5 rating and the reference's rating. It is relevant to the pointwise case.

- **Positional Bias Metric:** In the pairwise case, we swapped the order of answers and checked if the judge's decision changed.

- **Length Bias Metric:** Checked if the judge's wrong choice correlated with the answer length.

- **Instruction Following Evaluation Accuracy:** We analyzed whether the LVLM judge can effectively evaluate the instruction following capability of other LVLMs.

- **Format Adherence Accuracy:** This metric measures whether the judge's output followed the required JSON format.

3.2 Evaluation Data Construction

OpenCQA (Kantharaj et al., 2022)**:** This is an open-ended question-answering dataset on real charts. Each data point includes a chart and a question and expects an explanatory answer. We use its test set containing 1.1k QA instances.

VisText (Tang et al., 2023)**:** This is a chart captioning dataset with 12,441 charts, each paired with two types of captions: synthetic Level 1 (L1) captions that describe the chart's structural elements—such as chart type, title, axis labels, and scales—and human-generated Level 2/Level 3 (L2/L3) captions that provide insights into key statistics, trends, and patterns within the data. We use both L1 and L2/L3 captions with 1.2K test instances for each type.

For OpenCQA and VisText, we use the outputs generated by Islam et al. (2024) using *Gemini-1.0-*

Pro (Team et al., 2023) and *Claude-3-Haiku* (Anthropic, 2024) and compute the judgment scores using GPT-4o (OpenAI et al., 2023) and LLaVA-Critic-70B (Xiong et al., 2024) models and use as the judgment reference for diverse scenarios, as demonstrated in the previous section. This results in about 100K judgment data generated by GPT-4o and LLaVA-Critic-70B. We select these two models due to their impressive performance as a multimodal LLM-Judge (Xiong et al., 2024).

Chart-Instruct-Eval: We find that there are no datasets currently available in the chart domain that can assess the instruction-following capabilities of LVLMs. Therefore, we construct an instruction-following dataset (denoted as Chart-Instruct-Eval) to evaluate whether LVLM judges can evaluate the instruction-following capabilities of different models in chart-related tasks. For the dataset construction, we sample 400 charts from the ChartGemma (Masry et al., 2025) dataset. However, the original input instructions in the ChartGemma dataset lacked sufficient details. Hence, we could not use it for the instruction following purpose. Therefore, for each sample, we first create a detailed instruction containing specific requirements for the LLM response in terms of formatting, length, and structure to ensure instruction following. Then we manually prepare one good and one bad response corresponding to the instruction. Both responses convey similar content, but the good response fully adheres to all provided instructions, whereas the bad response disregards them. Finally, we assess the LLM judges whether they can reliably evaluate which response properly follows the instructions.

3.3 LVLM Judges

We evaluate **13** different open-source multimodal LLMs[1] as candidate judges, focusing on relatively smaller, publicly available models (2B–10B parameters). These include: **(i)** `XGen-MM-Phi-3-3.8B` (Xue et al., 2024) – *a multimodal model (3.8B) developed by Salesforce*, **(ii)** `MiniCPM-V-2.6-7B` (Yao et al., 2024) – *a 7B vision-language model by OpenBMB*, **(iii)** `Ph-3.5-3.8B-Vision-Instruct` (Abdin et al., 2024) – *a smaller vision model from Microsoft*, **(iv)** `Qwen2-VL-2B` - *Alibaba's Qwen (Wang et al., 2024) multimodal model with just 2B parameters*, **(v)** `Qwen2-VL-7B` – *The 7B version of the multimodal Qwen model*, **(vi)** `PaliGemma-3B` (Beyer et al., 2024) – *Google's multimodal open-source model*, **(vii)** `ChartGemma` (Masry et al., 2025) – *a chart-specialized model based on PaliGemma that is fine-tuned on chart tasks*, **(viii)** `Idefics-9B-Instruct`[2] – *an open multimodal model known for image understanding*, **(ix)** `InternLM-XComposer-7B` (Dong et al., 2024) – *a 7B vision model with composition abilities*, **(x)** `LLaVA-v1.6-Mistral-7B` – *A multimodal LVLM based on the LLaVA (Li et al., 2024a) architecture that also utilizes a 7B Mistral (Jiang et al., 2023) as the backbone*, **(xi)** `LLaVA-Critic-7B` – *a specialized evaluator model based on LLaVA and Qwen*, **(xii)** `mPLUG-Owl-3-7B` (Ye et al., 2023) – *a 7B multimodal model from Alibaba*, **(xiii)** `Janus-Pro-7B` (Chen et al., 2025) - *an open-source LVLM developed by Deepseek*. For more information about model selection, see Appendix A.1.

4 Experiments

In this section, we present the experimental results based on evaluating 13 LVLMs as judges across OpenCQA, VisText, and our proposed Chart-Instruct-Eval. The evaluation considers both pairwise and pointwise assessments, focusing on factual correctness, informativeness, relevance, positional bias, length bias, and instruction-following accuracy. We parse the LVLM-judge predicted judgments from their corresponding JSON-formatted responses using a parsing script (Laskar et al., 2023, 2024a,b). If the parsing script cannot properly parse the judgment from the response, we consider the LLM-generated answer as wrong for the pairwise case and error distance of 5 for the pointwise case. Note that we ran all our experiments using 1 A100 GPU with all the decoding parameters being set to the default values in HuggingFace (Wolf et al., 2020). Below, we demonstrate our findings:

4.1 Pairwise Evaluation Results

The pairwise evaluation measures how often the LVLM judges agree with GPT-4 or LLaVA-Critic-70B to select the better response in comparative assessments. We summarize the result in Table 1.

i. **Top-performing models:** LLaVA-Critic-7B achieved the highest agreement with reference judgments (above 75% average accuracy in each dataset). Another similar-sized (7B) LLM,

[1]We did not use the Prometheus-VL-7B (Lee et al., 2024) model since it requires a specific input format, making our prompts incompatible.

[2]HuggingFaceM4/idefics-9b-instruct

Model	Pairwise Evaluation: Judgment Accuracy (Higher is Better)									Pointwise Evaluation: Error Distance (Lower is Better)								
	OpenCQA			VisText L1			VisText L2/L3			OpenCQA			VisText L1			VisText L2/L3		
	GPT-4o	LC-70B	Avg.	GPT-4o	LC-70B	Avg.	GPT-4o	LC-70B	Avg.	GPT-4o	LC-70B	Avg.	GPT-4o	LC-70B	Avg.	GPT-4o	LC-70B	Avg.
Qwen2-VL-2B-Instruct	51.6	56.3	54.0	28.5	25.9	27.2	2.5	3.4	3.0	1.0	0.9	1.0	2.0	2.1	2.1	1.1	0.6	0.9
PaliGemma-3B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
ChartGemma-3B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
Phi-3.5-Vision-3.8B-Instruct	49.5	51.9	50.7	72.5	66.4	69.5	43.6	55.3	49.5	0.7	0.8	0.8	1.4	1.6	1.5	1.1	0.9	1.0
XGen-MM-Phi3-3.8B-Instruct	67.6	75.5	71.6	78.5	72.2	75.4	63.9	77.4	70.7	1.0	0.7	0.9	1.3	1.5	1.4	1.0	0.4	0.7
Janus-Pro-7B	46.6	48.7	47.7	48.6	45.6	47.1	52.6	57.0	54.8	1.0	0.7	0.9	1.0	1.2	1.1	1.0	0.4	0.7
Qwen2-VL-7B-Instruct	67.3	66.4	66.9	64.0	51.1	57.6	69.6	70.3	70.0	0.8	0.6	0.7	0.6	0.5	0.6	0.9	0.5	0.7
InternLM-Xcomposer2d5-7B	64.8	64.1	64.5	76.8	67.2	72.0	**69.7**	81.4	75.6	0.8	0.9	0.9	0.8	0.9	0.9	0.9	0.4	0.7
LLaVA-Next-v1.6-Mistral-7B	72.0	79.8	75.9	78.4	71.7	75.1	66.7	83.4	75.1	0.9	0.6	0.8	1.3	1.5	1.4	1.1	0.6	0.9
LLaVA-Critic-7B	**75.1**	**83.8**	**79.5**	**82.8**	**75.3**	**79.1**	69.0	**85.1**	**77.1**	**0.5**	**0.4**	**0.5**	**0.5**	**0.4**	**0.5**	**0.8**	**0.4**	**0.6**
mPLUG-Owl3-7B	60.8	59.4	60.1	72.2	64.0	68.1	46.1	39.2	42.7	0.8	0.6	0.7	1.0	1.0	1.0	0.9	0.4	0.7
MiniCPM-V-2.6-8B	64.3	68.6	66.5	49.2	42.9	46.1	44.8	39.1	42.0	1.0	0.8	0.9	1.3	1.3	1.3	1.7	1.5	1.6
Idefics-9B-Instruct	20.4	20.1	20.3	22.0	19.7	20.9	24.1	24.4	24.3	3.3	3.2	3.3	4.8	4.8	4.8	3.1	2.8	3.0

Table 1: Model performance based on average pointwise and pairwise scores across all reference types, as well as evaluation criteria (e.g., factual correctness, informativeness, etc.) in comparison to GPT-4o and LLaVA-Critic-70B (LC-70B) annotations (corresponding average score is also added). Bold values denote the best score in each case. Color coding for comparison: open-source models below 7B parameters , between 7-10B parameters .

the LLaVA-Next-v1.6-Mistral-7B model also performed competitively by exceeding 70% accuracy across each dataset. Interestingly, the XGen-MM model with just 3.8B parameters also achieved more than 70% accuracy, making it a very suitable judge in resource-constrained scenarios.

ii. **Lower-performing models:** Surprisingly, PaliGemma-3B and ChartGemma-3B achieved 0% agreement, indicating a poor alignment with reference judgments. Moreover, while the Qwen-2B model achieves decent performance in OpenCQA (above 50% accuracy), it achieves quite poor performance in VisText, especially in the L2/L3 scenario (below 10% accuracy). More surprisingly, the largest LVLM in our evaluation, the Idefics-9B-Instruct model achieves average accuracy below 25% in all datasets, highlighting its ineffectiveness as a judge. Our manual analysis revealed that these models failed due to not following instructions properly while also generating the response in the wrong format (improper JSON outputs). For PaliGemma, since it is not an instruction-tuned model, its poor performance could be related to the lack of understanding of instructions. The poor performance behind ChartGemma could be related to its training data lacking instructions related to judging tasks, therefore leading to poor generalization. We demonstrate some error examples of these LVLMs in Appendix A.3.

4.2 Pointwise Evaluation Results

This primarily measures the error distance between the ratings of the LVLM judge and the reference (GPT-4/LLaVA-Critic-70B) on a 1–5 Likert scale.

i. **Top-performing models:** Similar to the pairwise scenario, we find from Table 1 that LLaVA-Critic-7B again achieved the best performance in the pointwise scenario, achieving an error distance around 0.5. Other models like InternLM-Xcomposer-7B and Qwen2-VL-7B-Instruct that achieve quite good performance in pairwise scenarios, also demonstrate less error distance in pointwise scenarios (error distance below 1.0). Some other top-performing models in the pointwise scenario are LLaVA-Next-v1.5-Mistral-7B and MiniCPM-V-2.6-8B, which also achieve an error distance below 1.0 in 2 out of the 3 datasets.

ii. **Lower-performing models:** Similar to the pairwise scenario, PaliGemma-3B and ChartGemma-3B again produced irrelevant outputs resulting in the highest error distances (5.0). Moreover, despite being the largest model in our evaluation, the Idefics-9B-Instruct model performs quite poorly with a high error (on average, above 3).

4.3 Instruction and Format Adherence

We also assess the LVLM judges on their ability to maintain a standardized response format and whether they can evaluate the instruction following capabilities of other models. Based on the results in Table 2, we find that all 7B models achieve more than 90% format following capability. Smaller LVLMs like Qwen-2B and Phi-3.8B also achieve around 80% format adherence.

In terms of instruction following capability evaluation, we find that many LVLMs that could properly follow the format following requirement in their generated judgments for pairwise (§4.1) and pointwise (§4.2) evaluations, surprisingly generate the response in the wrong format in this evaluation. This makes our original parsing script penalize most of the LVLM-generated judgments as wrong. Therefore, we rewrite the parsing script to make it more flexible in terms of format adherence of the LVLM judge, since for this evaluation, our focus was to evaluate whether LVLM judges can properly assess instruction-following capabilities of different models in downstream chart-related tasks. Therefore, format adherence and other capabilities of the LVLM judges were not the focus of this evaluation. Our experiments reveal that mPLUG-Owl3-7B (93.5%) and Qwen2-VL-7B-Instruct (87.0%) achieve the top two results in terms of evaluating the instruction-following capability of different LVLM generated responses. Surprisingly, the LLaVA-Critic-7B model achieves only 45.5% accuracy in this task. This may indicate that the training data of the LLaVA-Critic-7B model may not contain such data, leading to a quite poor generalization in this dataset.

Moreover, PaliGemma-3B and ChartGemma-3B fail to follow the format requirements at all, and also unable to evaluate instruction following capability. Finally, the Idefics-9B-Instruct model, even with 9B parameters, achieves poor instruction and format following accuracy.

Model	Instruction Following	Format Adherence
Qwen2-VL-2B-Instruct	13.5	78.9
PaliGemma-3B	0.0	0.0
ChartGemma-3B	0.0	0.0
Phi-3.5-Vision-3.8B-Instruct	49.0	83.3
XGen-MM-Phi3-3.8B-Instruct	72.5	97.6
Janus-Pro-7B	73.0	96.7
Qwen2-VL-7B-Instruct	87.0	98.6
InternLM-Xcomposer2d5-7B	54.0	95.9
LLaVA- Next-v1.6-Mistral-7B	27.0	98.9
LLaVA-Critic-7B	45.5	**99.7**
mPLUG-Owl3-7B	**93.5**	98.9
MiniCPM-V-2.6-8B	54.5	90.3
Idefics-9B-Instruct	20.5	35.0

Table 2: Accuracy in terms of Instruction Following Evaluation (evaluated on Chart-Instruct-Eval) and Format Adherence (based on average across all datasets).

Model	Length Bias	Position Bias
Qwen2-VL-2B-Instruct	55.1	71.9
Phi-3.5-Vision-3.8B-Instruct	69.8	59.6
XGen-MM-Phi3-3.8B-Instruct	64.3	79.2
Janus-Pro-7B	27.2	50.6
Qwen2-VL-7B-Instruct	**21.5**	**35.8**
InternLM-Xcomposer2d5-7B	24.5	35.9
mPLUG-Owl3-7B	21.9	42.5
LLaVA-Next-v1.6-Mistral-7B	71.8	77.0
LLaVA-Critic-7B	76.4	39.6
MiniCPM-V-2.6-8B	37.4	45.5

Table 3: Length Bias and Position Bias for different models (results based on average across all datasets). Here, Lower values are better. Models achieving format following accuracy above 50% are only evaluated.

4.4 Bias Analysis

To assess potential biases in LVLM judges, we analyzed position bias (whether the order of the responses affects judgments) and length bias (whether longer responses are favored). Based on the result presented in Table 3, we find that the Qwen2-VL-7B-Instruct model exhibited the lowest positional bias and length bias. On the contrary, the LLaVA-Next-v1.6-Mistral-7B model showed very high bias in both scenarios, suggesting susceptibility to judge responses based on variations in the position of the responses as well as the length. Surprisingly, the LLaVA-Critic-7B model, which is the best-performing model in terms of judgment accuracy and error distance, demonstrates the highest length bias across all models, indicating a tendency to favor longer answers. We provide an example of the position bias in Figure 5, and an example of the length bias in Figure 6.

4.5 Human Evaluation

In this section, we conduct a human evaluation of the GPT-4o and the LLaVA-Critic-70B models which we used as the reference judge to evaluate the smaller open-source LVLMs. For this purpose, we randomly sample 100 responses generated by Islam et al. (2024) for the Claude-3-Haiku and the Gemini-1-Pro models in OpenCQA and VisText datasets. Then, we ask two human annotators having expertise in NLP and Computer Vision to rate these responses based on our evaluation criteria (e.g., informativeness, relevance, etc.) with references provided for 50% of the data and without any references for rest of the data.

Based on our human evaluation, we find that both annotators' judgments highly correlate with GPT-4o and LLaVA-Critic-70B, with an error distance below 1.0. Interestingly, we find that both annotators have a higher correlation with the open-source LLaVA-Critic-70B model (average error distance with LLaVA-Critic-70B: `0.81`, and with GPT-4o: `0.93`). Therefore, in real-world industrial scenarios where human annotation is costly and closed-source LLMs are not preferred due to privacy concerns in proprietary datasets, the open-source LLaVA-Critic-70B model could be a good alternative for data annotation.

4.6 Ablation Studies

(i) Effect of Reference Type: In this section, we compare the performance variation of different LVLMs in reference-based and reference-free scenarios (see Table 4). LVLMs that achieve more than

Model	With Reference	Without Reference
Qwen2-VL-2B-Instruct	47.4	55.7
Phi-3.5-Vision-3.8B-Instruct	51.6	47.3
XGen-MM-Phi3-3.8B-Instruct	66.8	68.4
Janus-Pro-7B	45.9	47.3
Qwen2-VL-7B-Instruct	66.7	67.8
InternLM-Xcomposer2d5-7B	62.1	67.5
LLaVA-Next-v1.6-Mistral-7B	71.0	73.0
LLaVA-Critic-7B	**74.9**	**75.3**
mPLUG-Owl3-7B	63.5	58.2
MiniCPM-V-2.6	63.2	65.4
Idefics-9B-Instruct	16.6	24.2

Table 4: Judgment Accuracy in comparison to GPT-4o in OpenCQA based on Reference-based (with reference) and Reference-free (without reference) evaluation.

Model	Factual Correctness	Informativeness	Relevancy
Qwen2-VL-2B-Instruct	2.6	1.6	2.0
Phi-3.5-Vision-3.8B-Instruct	1.6	1.3	1.4
XGen-MM-Phi3-3.8B-instruct-r-v1	1.7	1.4	1.6
Janus-Pro-7B	1.4	1.0	1.3
Qwen2-VL-7B-Instruct	0.7	0.4	0.5
InternLM-Xcomposer2d5-7B	1.0	0.8	0.9
LLaVA-Next-v1.6-Mistral-7B	1.7	1.4	1.5
LLaVA-Critic-7B	**0.6**	**0.3**	**0.4**
mPLUG-Owl3-7B	1.1	0.9	1.1
MiniCPM-V-2.6	1.7	1.3	0.9

Table 5: Average Error Distance (compared with LLaVA-70B-Critic) in VisText (L1) for different LVLMs based on various Evaluation Types. Here, lower values indicate better performance.

20% pairwise judgment accuracy in OpenCQA are selected for the analysis. While we find that different LVLMs have a slight change in performance with the presence and absence of references, the performance difference between them based on a paired t-test is not statistically significant ($p > 0.05$). This demonstrates that the open-source LVLMs are robust in both reference-based and reference-free evaluation.

(ii) Effect of Evaluation Criteria: In Table 5, we analyze the performance differences among various LVLMs with an error distance below 2.5 in VisText (L1) across multiple evaluation metrics: (i) informativeness, (ii) relevance, and (iii) factual correctness. While we observe slight performance variations based on the evaluation criteria, the paired t-test demonstrates that these differences are not statistically significant ($p > 0.05$), indicating robust performance across various evaluation measures.

5 Conclusion and Future Work

In this paper, we conducted a comprehensive evaluation of open-source LVLMs as automatic judges for chart comprehension and reasoning tasks. Our analyses revealed that while some open-source LVLMs (e.g., 7B models like LLaVA-Critic, Qwen2-VL, InternLM, and LLaVA-Next) can achieve judgment accuracy (with lower error rates) that is comparable to state-of-the-art closed-source models like GPT-4 or larger open-source models like LLaVA-Critic-70B; other models, such as ChartGemma and PaliGemma, struggle significantly, highlighting variability in their reliability. Despite the promising results of various models, issues like bias and lack of instruction following capability still persist. Therefore, future work should focus on mitigating biases, improving instruction following evaluation capability, alongside ensuring consistency across diverse evaluation criteria by developing a multimodal LLM judge using more recent models (Bai et al., 2025) for chart model evaluation.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the York Research Chairs (YRC) program, and Compute Canada.

Ethical Considerations

The models used for experiments are only used as the judge to evaluate other LVLM-generated responses. Therefore, the LVLM responses do not pose any ethical concerns. Additional compensation for human evaluation is not needed since it was conducted by two authors of this paper.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anthropic. 2024. Introducing the next generation of claude.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, pages 6562–6595.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation.

E. Hoque and M. Saidul Islam. 2024. Natural language generation for visualizations: State of the art, challenges and future directions. *Computer Graphics Forum*, n/a(n/a):e15266.

Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5648–5656.

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

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. Opencqa: Open-ended question answering with charts. In *Proceedings of EMNLP 2022*.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024a. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

Md Tahmid Rahman Laskar, Elena Khasanova, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2024b. Query-OPT: Optimizing inference of large language models via multi-query instructions in meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1140–1151, Miami, Florida, US.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

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

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

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10387–10409.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2025. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7775–7803.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. 2023. Gpt-4 technical report.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md. Tahmid Rahman Laskar, Md. Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Kantharaj Shankar, Leong Rixie Tiffany Ko, Lin Xiang, Masry Ahmed, Thakkar Megh, Hoque Enamul, and Joty Shafiq. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2022*.

Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, et al. 2023. Gemini: A family of highly capable multimodal models.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Appendix

A.1 Regarding Model and Dataset Selection

We selected popular LVLMs that were released by early 2025, with sizes less than 10B parameters. Although there are many other chart benchmarks currently available (Huang et al., 2024), we selected OpenCQA and VisText since qualitative evaluation is often required in these datasets (Islam et al., 2024).

A.2 Prompts for the LVLM Judge

OpenCQA Pointwise (With Reference)

Suppose, you are a human evaluator. You are required to rate the {Evaluation Criteria} of the answer generated by a model in comparison to the gold reference answer for a given question in the open-ended chart question answering task.

Please provide your response in JSON format with the following keys: (i) Score, (ii) Explanation.

Here, the 'Score' should be between 1 to 5 (inclusive), with the higher score indicating better. Moreover, the value for the "Explanation" key will contain the reason behind your scoring.

You should only provide the response in the required JSON format without any additional text.

In the following, you are first given the question, followed by the gold reference answer. Afterward, you are given the model-generated answer. You are also provided with the chart image as the context for the chart question-answering task.

[Question]

[Gold Reference Answer]

[Model Generated Answer]

`[Chart Image]`

OpenCQA Pairwise (Without Reference)

Suppose, you are a human evaluator. You are given the answers generated by two different models for a given question in the open-ended chart question answering task. Now, your task is to determine which model is better in terms of {Evaluation Criteria}.

Please provide your response in JSON format with the following keys: (i) Model, (ii) Explanation,

Here, the output value for the 'Model' key is the respective model that is better, could be either 'Model A' or 'Model B', or 'Tie' if both models are equally good. Moreover, the value for the "Explanation" key will contain the reason behind your preference.

You should only provide the response in the required JSON format without any additional text.

In the following, you are first given the question. Afterward, you are given the model-generated answers. You are also provided with the chart image as the context for the chart question-answering task.

[Question]

[Model 1 Generated Answer]

[Model 2 Generated Answer]

`[Chart Image]`

VisText L1 Pointwise (With Reference)

Suppose, you are an human evaluator. You are required to rate the {Evaluation Criteria} of the L1 caption describing the aspects of the chart's construction (e.g., chart type and axis labels) generated by a model in the chart captioning task.

Please provide your response in JSON format with the following keys: (i) Score, (ii) Explanation.

Here, the 'Score' should be between 1 to 5 (inclusive), with the higher score indicating better. Moreover, the value for the "Explanation" key will contain the reason behind your scoring.

You should only provide the response in the required JSON format without any additional text such that I can correctly parse the result from your JSON formatted response.

In the following, you are first provided with the gold reference caption. Afterward, you are given the model generated caption. You are also provided with the chart image which was used as the context for the chart captioning task.

[Gold Reference Caption]

[Model Generated Caption]

`[Chart Image]`

VisText L2/L3 Pairwise (No Reference)

Suppose, you are a human evaluator. You are given the captions generated by two different models in the chart captioning task. Now, your task is to determine which model is better based on {Evaluation Criteria}.

Please provide your response in JSON format with the following keys: (i) Model, (ii) Explanation.

Here, the output value for the 'Model' key is the respective model that is better, could be either 'Model A' or 'Model B', or 'Tie' if both models are equally good. Moreover, the value for the "Explanation" key will contain the reason behind your preference.

You should only provide the response in the required JSON format without any additional text such that I can correctly parse the result from your JSON formatted response.

In the following, you are provided with the model generated captions. You are also provided with the chart image which was used as the context for the chart captioning task.

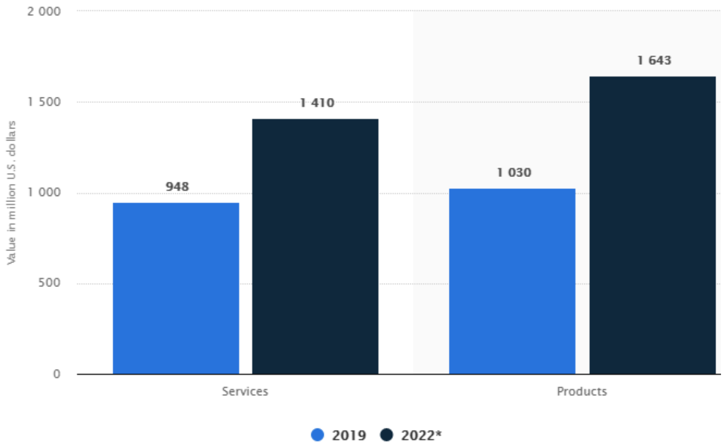
[Model 1 Generated Caption]

[Model 2 Generated Caption]

`[Chart Image]`

A.3 Error Analysis

Some example error cases are demonstrated below.



Prompt:
You are given the responses generated by two different models, each answering the same instruction about a specific chart.
Your goal is to identify which model generated response is better among them.
Below, you will be first provided with the task instruction, followed by the response generated by the Model A, and finally the Model B.
You are also provided with the chart image.
Please generate your response in the JSON format with the following keys: (i) Model, (ii) Explanation.
Here, the output value for the 'Model' key is the respective model that is better, could be either 'Model A' or 'Model B'.
The value for the 'Explanation' key will contain the reason behind your preference.
Task Instruction: After analyzing the chart and extracting the numeric values, generate only a Python code without any comments that prints a solution to the following question: Find the ratio between services and products in 2022 ?
Model A Response:

```
# Define the number of services provided in 2022
services_2022 = 1410
# Define the number of products sold in 2022
products_2022 = 1643
# Calculate the ratio of services to products
ratio = services_2022 / products_2022
# Print the calculated ratio
print(ratio)
```

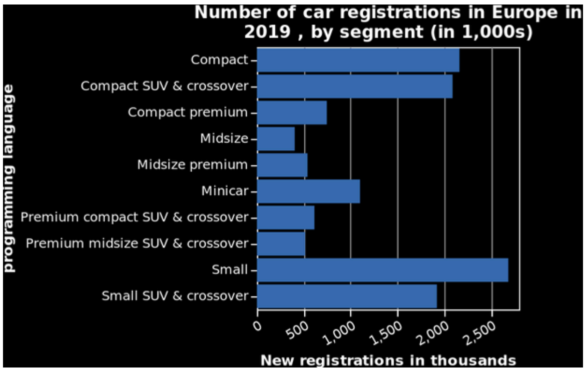
Model B Response:

```
services_2022 = 1410
products_2022 = 1643
ratio = services_2022 / products_2022
print(ratio)
```

ChartGemma:

```
print("Model A")
print("Model B")
```

Figure 3: An example of an error case involves the ChartGemma model being tasked with evaluating the Python code responses generated by two different models and providing a verdict on which one is better. However, instead of following the instructions, the model failed to complete the task correctly and simply returned two print statements as its output (highlighted in red text).



PaliGemma-3b:

The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in thousands in 2019, by segment, in Europe . The number of car registrations in

Prompt:

Suppose, you are a human evaluator.
You are required to rate the informativeness of the caption generated by a model in the chart captioning task.

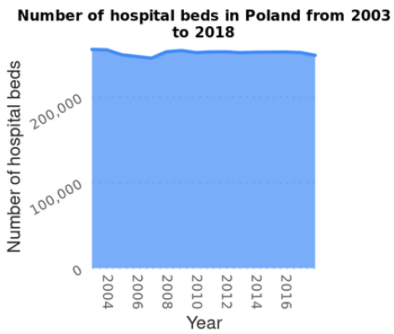
Please provide your response in JSON format with the following keys: (i) Score, (ii) Explanation.
Here, the 'Score' should be between 1 to 5 (inclusive), with the higher score indicating better.
The value for the "Explanation" key will contain the reason behind your scoring.
You should only provide the response in the required JSON format without any additional text such that I can correctly parse the result from your JSON formatted response.

In the following, you are provided with the model generated caption. You are also provided with the chart image which was used as the context for the chart captioning task.
Model Generated Answer: The chart shows the number of new car registrations in Europe by vehicle segment in 2019 , measured in thousands . The largest segment is Compact , with over 2,000 thousand new registrations , followed by Compact SUV & crossover with around 1,500 thousand . The Midsize and Midsize premium segments have the next highest registrations , around 800 thousand and 500 thousand respectively . Smaller segments such as Minicar , Premium compact SUV & crossover , and Premium midsize SUV & crossover range from around 200 thousand to 400 thousand new registrations . The smallest segments are Small and Small SUV & crossover , with under 500 thousand new registrations each . Overall , the chart highlights the dominance of compact and SUV/crossover models in the European car market in 2019 .

Figure 4: An example of an error case involves the PaliGemma-3b model being tasked with evaluating a chart caption generated by another model. Specifically, it was asked to rate the caption on a scale of 1 to 5 based on the 'Informativeness' criterion and to provide an explanation for the rating. However, instead of performing the evaluation correctly, the model hallucinated and repeatedly generated the same line without adhering to the required JSON format. (highlighted in red text).

Model A Generated Caption: The chart depicts the number of hospital beds in a given location over several years . It shows a steady increase in the number of hospital beds from 2004 to 2016 , with the number reaching a peak of around 200,000 beds in 2016 . The chart provides a clear visual representation of the growth in hospital bed capacity over the years covered in the data .

Model B Generated Caption: The number of hospital beds in the United States has remained relatively stable over the past decade . In 2004 , there were 250,000 hospital beds in the United States . By 2016 , this number had only decreased by 1.6 % to 246,000 . This decrease can likely be attributed to the increasing popularity of outpatient care and the use of home health services .



Prompt Version 1:

Suppose, you are an human evaluator.

You are given the captions generated by two different models in the chart captioning task. Now, your task is to determine which model is better by considering the relevancy, conciseness, informativeness, and factual correctness ... ...

Model A Generated Caption: The chart depicts the number of hospital beds ... ...

Model B Generated Caption: The number of hospital beds in the United States ... ...

Prompt Version 2:

Suppose, you are an human evaluator.

You are given the captions generated by two different models in the chart captioning task. Now, your task is to determine which model is better by considering the relevancy, conciseness, informativeness, and factual correctness ... ...

Model A Generated Caption The number of hospital beds in the United States ... ...

Model B Generated Caption: The chart depicts the number of hospital beds ... ...

LLaVA-Critic-7B for Prompt Version 1:

"Model": "Model A",

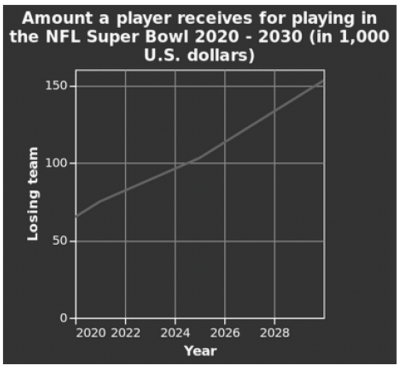
"Explanation": "Model A provides a more accurate and relevant description of the chart. It correctly identifies the steady increase in the number of hospital beds from 2004 to 2016, reaching a peak of around 200,000 beds in 2016. This aligns with the visual representation in the chart. Model B, on the other hand, inaccurately states that the number of hospital beds has remained relatively stable and even mentions a decrease, which is not supported by the chart. Therefore, Model A is more factually correct and informative."

LLaVA-Critic-7B for Prompt Version 2:

"Model": "Model A",

"Explanation": "Model A provides a more accurate and detailed description of the chart. It correctly identifies that the number of hospital beds in the United States has remained relatively stable over the past decade, with a slight decrease from 250,000 in 2004 to 246,000 in 2016. This aligns with the visual data presented in the chart. Model B, on the other hand, inaccurately describes the trend as a steady increase, which is not supported by the chart. Therefore, Model A is more factually correct and relevant to the chart."

Figure 5: An example of an error case for the LLaVA-Critic-7B model which demonstrates position bias by changing its selection of the better caption based on the change in the order of the model generated captions.



LLaVA-Critic-7B:
"Model": "Model A",
"Explanation": "Model A provides a detailed and accurate description of the chart, including specific amounts received by players on both winning and losing teams for each year from 2020 to 2030. This level of detail aligns closely with the information presented in the chart, making it highly informative and factually correct. Model B, while capturing the general trend of increasing compensation, lacks the specificity and accuracy found in Model A s response. Therefore, Model A is better in terms of relevancy, conciseness, informativeness, and factual correctness."

Prompt:
Suppose, you are an human evaluator.
You are given the captions generated by two different models in the chart captioning task. Now, your task is to determine which model is better by considering the relevancy, conciseness, informativeness, and factual correctness of the model generated caption. Ideally, a better quality caption should be concise, factually correct, and relevant to the chart, while also providing high level informative insights of the chart ... ...
Model A Generated Caption: The chart shows the amount of money a player receives for playing in the NFL Super Bowl from 2020 to 2030 . In 2020 , each player on the winning team received $ 150,000 , while each player on the losing team received $ 75,000 . In 2023 , the winning team received $ 157,000 per player , while the losing team received $ 78,500 per player . In 2024 , the winning team received $ 165,000 per player , while the losing team received $ 82,500 per player . In 2025 , the winning team received $ 175,000 per player , while the losing team received $ 87,500 per player . In 2026 , the winning team received $ 185,000 per player , while the losing team received $ 92,500 per player . In 2027 , the winning team received $ 195,000 per player , while the losing team received $ 97,500 per player . In 2028 , the winning team is expected to receive $ 210,000 per player , while the losing team is expected to receive $ 105,000 per player . In 2029 , the winning team is expected to receive $ 220,000 per player , while the losing team is expected to receive $ 110,000 per player . In 2030 , the winning team is expected to receive $ 230,000 per player , while the losing team is expected to receive $ 115,000 per player .
Model B Generated Caption: The chart depicts the amount a player receives for playing in the NFL Super Bowl from 2020 to 2030 . The data shows a steady increase in the compensation over the years , with the amount received starting at around $ 30,000 in 2020 and rising to approximately $ 150,000 by 2030 . The trend indicates a consistent growth in the financial rewards for participating in the prestigious NFL Super Bowl event over the next decade .

Figure 6: An example of an error case for the LLaVA-Critic-7B model which demonstrates length bias by selecting the lengthy caption even though in the given multidimensional prompt, conciseness was one of the criteria for a better caption.