

Deploying Tiny LVLM Judges for Real-World Evaluation of Chart Models: Lessons Learned and Best Practices

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

Abstract

Large Vision-Language Models (LVLMs) with only 7B parameters have shown promise as automated judges in chart comprehension tasks. However, tiny models (≤ 2 B parameters) still perform poorly as judges, limiting their real-world use in resource-constrained settings. To address this, we propose two approaches to ensure cost-efficient evaluation: (i) multi-criteria prompting, which combines separate evaluation criteria into a single instruction, and (ii) domain-adaptive transfer learning, in which we address the low-resource problem by fine-tuning a 2B-parameter LVLM on synthetic judgments in a chart dataset to create the **ChartJudge-2B** model. Experiments show that multi-criteria prompting exposes robustness gaps, which led to a huge drop in performance for 7B models, including specialized LVLM judges like LLaVA-Critic. In addition, we find that by fine-tuning a tiny LVLM (i.e., ChartJudge-2B), we can effectively transfer knowledge from one dataset to another to make it a more specialized model. Our fine-grained analysis across chart types and query complexities offers actionable insights into trade-offs between model size, prompt design, and transferability—enabling scalable, low-cost evaluation for chart reasoning tasks. Our code and the data will be made publicly available.

1 Introduction

Charts are widely used to communicate data, but understanding them requires multimodal reasoning that integrates visual and textual cues (Hoque and Islam, 2024). While LVLMs show strong performance on such tasks across automatic metrics, real-world deployment often requires qualitative evaluation, which typically relies on costly and time-consuming human judgment (Islam et al., 2024).

Recently, Laskar et al. (2025a) proposed the utilization of LVLMs as the judge to evaluate other

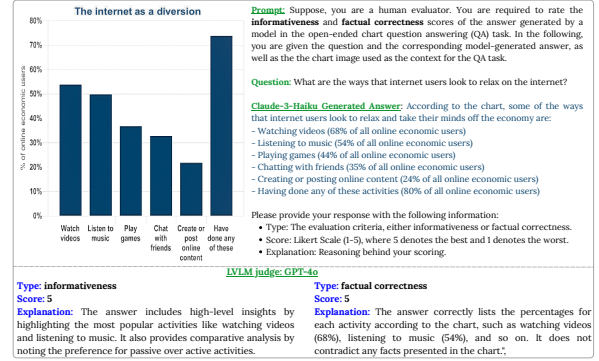


Figure 1: GPT-4o as the judge evaluating *Claude-3-Haiku* model answer in the OpenCQA dataset across multiple criteria (informativeness, factual correctness).

LVLMs in downstream chart tasks (see Figure 1 for an example of an LVLM judge for charts), with certain 7B LVLMs achieving judgment performance closer to powerful LVLMs like GPT-4o (OpenAI et al., 2023) or LLaVA-Critic-70B (Xiong et al., 2024). This creates opportunities to use open-source LVLM judges in real-world scenarios where industries do not share sensitive proprietary charts with closed-source LLM providers. However, tiny LVLMs (≤ 2 B parameters) still underperform, limiting their applicability in resource-constrained settings, since available tools for optimizing LVLMs for inference are much less available than text-only LLMs (Shinde et al., 2025).

To address these challenges, we explore two strategies for cost-efficient LVLM-based evaluation considering real-world industrial settings: (i) *multi-criteria prompting*, which fuses multiple evaluation dimensions into a single query to reduce inference latency & cost, and (ii) *domain-adaptive transfer learning*, where a 2B-parameter LVLM is fine-tuned on synthetic judgments from a stronger model. The result is **ChartJudge-2B**, a lightweight evaluator tailored for chart tasks (Figure 2). Our experiments show that multi-criteria prompting exposes significant performance drops in existing

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

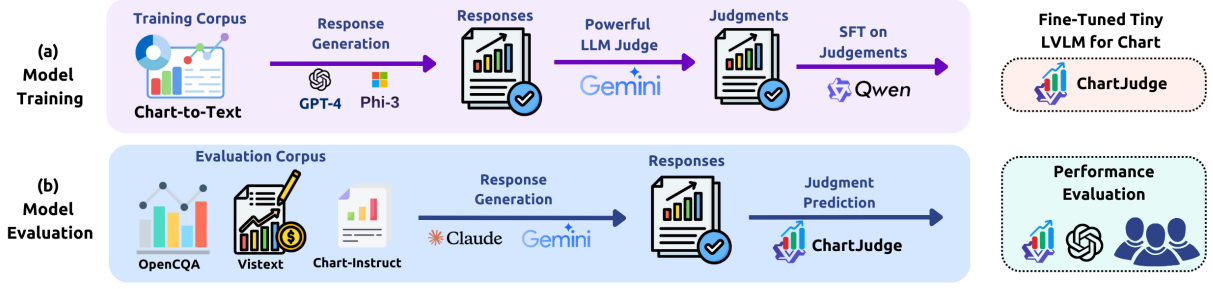


Figure 2: Our Fine-Tuning Approach. (a) At first, responses generated by various LVLMs (e.g., GPT-4, Phi-3) for chart captioning (e.g., Chart-to-Text dataset) are judged on diverse criteria by a different LVLM (e.g., Gemini-1.5-Pro). A small LVLM (e.g., Qwen2-VL-2B-Instruct) is then fine-tuned on these judgments to create **ChartJudge-2B**. (b) For evaluation, responses from LVLMs (e.g., Claude, Gemini) on chart benchmarks are judged by **ChartJudge-2B** and compared with a larger LVLM (e.g., GPT-4o) or human ratings.

LVLMs—not only in 7B models like the specialized LLaVA-Critic-7B, but also in its larger variant, LLaVA-Critic-70B (Xiong et al., 2024). In contrast, ChartJudge-2B, with only 2B parameters, generalizes effectively across chart datasets, offering a cost-effective alternative for evaluation. Our work makes *three* key contributions:

- We demonstrate that ChartJudge-2B can effectively transfer knowledge across chart datasets, even when distilled from a different LVLM judge than the one used for evaluation, making it a robust, cost-efficient evaluator in low-resource settings.
- We show that multi-criteria prompting reveals significant robustness issues and performance drops in open-source LVLMs, including specialized judges like LLaVA-Critic. However, a tiny LVLM can achieve strong performance in multi-criteria prompting when fine-tuned even on limited amount of data.
- We provide practical insights on model selection, prompt design, model errors, and transfer strategies, helping guide scalable deployment of LVLM judges for real-world chart reasoning tasks.

As a secondary contribution, our code and the data will be made publicly available here: https://github.com/tahmedge/chart_lvlm_judge.

2 Related Work

The complexity of chart understanding tasks like question answering (Hoque et al., 2022; Masry et al., 2022, 2025a; Kantharaj et al., 2022) and captioning (Shankar et al., 2022; Tang et al., 2023)

has led to the creation of various specialized chart-specific models (Han et al., 2023; Masry et al., 2025b). However, evaluating these specialized models is a significant hurdle due to the dependence on humans. To overcome this, recent research explores using LVLMs as judges for the chart domain (Laskar et al., 2025a).

Although early LVLM judges like Prometheus-VL (Lee et al., 2024) found text-rich visuals like charts challenging for evaluation, Laskar et al. (2025a) demonstrate that many recently proposed open-source LVLMs with only 7B parameters can achieve performance comparable to GPT-4o. This makes it possible to even use smaller open-source LVLMs in real-world scenarios where industries restrict the use of closed-source API models for evaluation on proprietary charts. However, tiny LVLMs (under 2B) still lag behind, limiting their viability in resource-constrained environments.

To address this gap, in this paper, we investigate various strategies to minimize the inference cost of using open-source LVLMs in real-world scenarios. Due to the effectiveness of techniques like prompt engineering (Islam et al., 2024) and fine-tuning of smaller models (Masry et al., 2025b) that help achieve performance gain while reducing inference cost, we seek to find the best strategy to utilize these techniques in the context of the LVLM judge.

3 Methodology

3.1 Proposed Approach

To achieve cost-effective inference in real-world industrial settings, we proposed two approaches.

(i) Fine-Tuning Approach: The lack of human-annotated judgment dataset in the chart domain prohibits the training of any LVLMs-as-the-judge for

chart-related tasks. Since early research shows that annotations generated by powerful closed-source LVLMs correlate well with humans (Lee et al., 2024; Chen et al., 2024), we propose a *domain-adaptive fine-tuning approach* (Laskar et al., 2020, 2022, 2025b) that leverages knowledge distillation by first generating judgments in a given chart dataset on different LVLM-generated responses using a larger LVLM. Then, the smaller LVLM is fine-tuned on this judgment data for better adaptation in the chart domain (see Figure 2).

(ii) Prompting-Based Approach: Since LVLM fine-tuning requires good computing resources, it is often difficult to fine-tune LVLMs in resource-constrained scenarios (Shinde et al., 2025). In such cases, prompt engineering can help improve the performance of LVLM without any additional training. However, evaluating with separate prompts for different evaluation criteria increases both inference latency and cost (Laskar et al., 2023, 2024b). In the prompting-based approach, our main objective is to reduce evaluation cost by minimizing the overall inference tokens and GPU time. To this end, we propose the *multi-criteria prompting* approach that combines multiple evaluation criteria within a single prompt (see Appendix A.7 for a sample prompt) for the LVLM judges to generate judgments for all criteria in their response (see Appendix A.8 for a sample response).

3.2 Evaluation Protocol

Given a chart and model-generated response(s), we follow the work of Laskar et al. (2025a) to construct prompts, evaluation rubric, and evaluation metrics.

Evaluation Rubric and Metrics: We construct prompts for both *pairwise* and *pointwise* evaluation. The constructed prompts contain samples *with reference* as well as *without reference*, for diverse evaluation criteria: *factual correctness*, *informativeness*, *relevance*, and *multidimensional criteria* (e.g., overall response quality). In line with the prior work, we also use the following evaluation metrics: *Judgment Accuracy*, *Error Distance*, *Bias Evaluation*, *Instruction Following Evaluation Capability*, and *Format Adherence*. We refer readers to the Appendix A.2 for the definitions of these metrics.

Evaluation Datasets: For evaluation, we use the following three datasets: (i) **OpenCQA** (Shankar et al., 2022) for open-ended chart QA.

Type	# Single-Criterion	# Multi-Criteria
<i>Source-wise</i>		
Statista	6,898	—
Pew	2,827	2,827
<i>Pointwise Labels</i>		
1	801	510
2	1,000	548
3	1,000	414
4	1,000	179
5	1,000	113
<i>Pairwise Labels</i>		
Tie	2,000	268
Model A	1,500	568
Model B	1,424	226

Table 1: Fine-tuning dataset statistics for single-criterion and multi-criteria settings. Here, ‘#’ denotes total number of samples.

(ii) **VisText** (Tang et al., 2023) for chart captioning consisting of L1 (describing chart’s structural elements) and L2/L3 captions (providing key insights in the data).

(iii) **Chart-Instruct-Eval** (Laskar et al., 2025a) that assesses LVLM judges on evaluating the instruction following capabilities of other LVLMs in chart-related tasks.

The judgment annotations in these datasets are collected from (Laskar et al., 2025a), with OpenCQA and VisText containing 100K judgment data, and Chart-Instruct-Eval containing 400 samples. More specifically, for OpenCQA and VisText, we collect the outputs generated by Islam et al. (2024) using *Gemini-1.0-Pro* (Team et al., 2023) and *Claude-3-Haiku* (Anthropic, 2024), and use the judgment scores annotated by (Laskar et al., 2025a) using *GPT-4o* (OpenAI et al., 2023) and *LLaVA-Critic-70B* (Xiong et al., 2024) models for diverse evaluation rubrics across 100K judgment data. For Chart-Instruct-Eval, we use the existing annotation in the dataset. For the multi-criteria evaluation, we use the OpenCQA dataset.

3.3 ChartJudge-2B

We fine-tune the tiny LVLM Qwen2-VL-2B-Instruct¹ (Wang et al., 2024) to build the **ChartJudge-2B** model. We describe our fine-tuning data construction procedure below.

We constructed a training dataset comprising both pointwise and pairwise evaluations of chart image summaries sourced from the *Chart-to-Text* benchmark (Shankar et al., 2022). However, the Chart-to-Text dataset is a chart captioning bench-

¹The Qwen2-VL-2B-Instruct model was selected due to its strong performance as a 2B parameter model across diverse tasks (Wang et al., 2024).

Model	OpenCQA		VisText L1		VisText L2/L3	
	Pair (↑)	Point (↓)	Pair (↑)	Point (↓)	Pair (↑)	Point (↓)
ChartJudge-2B	61.7	1.3	64.6	1.6	52.3	1.3
Qwen2-VL-2B-Instruct	54.0	1.0	27.2	2.1	3.0	0.9
Phi-3.5-Vision-3.8B-Instruct	50.7	0.8	69.5	1.5	49.5	1.0
XGen-MM-Phi3-3.8B-Instruct	71.6	0.9	75.4	1.4	70.7	0.7
Qwen2-VL-7B-Instruct	66.9	0.7	57.6	0.6	70.0	0.7
InternLM-Xcomposer2d5-7B	64.5	0.9	72.0	0.9	75.6	0.7
LLaVA-Next-v1.6-Mistral-7B	75.9	0.8	75.1	1.4	75.1	0.9
LLaVA-Critic-7B	79.5	0.5	79.1	0.5	77.1	0.6

Table 2: Average results (judgment accuracy for pairwise, error distance for pointwise) of different models (below 4B , above 4B) from Laskar et al. (2025a) compared to our ChartJudge-2B model. ‘↑’ = higher better; ‘↓’ = lower better.

mark, whereas our evaluation datasets contain both question answering (QA) and chart captioning. Since Chart-to-Text contains data from Statista and Pew, we generate synthetic questions on its Pew split to use it as an open-ended QA task, while using the Statista split for chart captioning without any modification. For question generation, we prompt the **Gemini-1.5-Pro** (Georgiev et al., 2024) model to generate a question by providing the chart and its corresponding gold reference as input (see Appendix A.5 for the prompt).

To mitigate potential *teacher model* bias, we intentionally decoupled the models used at each stage of our pipeline. For the fine-tuning data, responses were generated by *GPT-4V* (OpenAI et al., 2023) and *Phi-3* (Abdin et al., 2024). The judgments on these responses, which serve as training labels, were then created by a different model, *Gemini-1.5-Pro* (Georgiev et al., 2024). For the final evaluation of our *ChartJudge-2B* model, we again used a distinct set of models: responses were generated by *Claude-3-Haiku* (Anthropic, 2024) and *Gemini-1.0-Pro*, and the reference judgments were provided by *GPT-4o* and *LLaVA-Critic-70B*. This is done to ensure robustness across different LVLM-generated responses and judgments by using different models for training and evaluation.

In the pointwise setting, each summary was assigned a rating ranging from 1 to 5, accompanied by a textual justification for the rating. In the pairwise setting, the model was prompted to compare two responses (Model A and Model B) and select the better one, or indicate a tie if no clear preference could be determined, along with an explanatory rationale for the decision. We present our fine-tuning dataset statistics in Table 1 (more information on the dataset is also provided in Appendix A.3). Below, we describe the dataset based on the *single-criterion* and the *multi-criteria* settings.

Model	Instruction (↑)	Format (↑)	Bias (↓)
ChartJudge-2B	36.5	95.9	79.5
Qwen2-VL-2B-Instruct	13.5	78.9	63.5
Phi-3.5-Vision-3.8B-Instruct	49.0	83.3	64.7
XGen-MM-Phi3-3.8B-Instruct	72.5	97.6	71.8
Qwen2-VL-7B-Instruct	87.0	98.6	28.7
InternLM-Xcomposer2d5-7B	24.5	35.9	30.2
LLaVA-Next-v1.6-Mistral-7B	27.0	98.9	74.4
LLaVA-Critic-7B	45.5	99.7	58.0

Table 3: Performance of our fine-tuned ChartJudge-2B and other baselines for Instruction Following evaluation capability on Chart-Instruct-Eval, Format Adherence (average across all datasets), and Bias (average of Length and Position Bias across all datasets). ‘↑’ = higher is better; ‘↓’ = lower is better.

Single-Criterion: For the single-criterion scenario, our fine-tuning dataset consists of 9,725 samples. While the instances in the dataset are randomly sampled, we avoid any bias towards a particular type of evaluation type/criteria by defining a data distribution schema such that the dataset does not contain too many samples of a particular evaluation type/criteria.

Multi-Criteria: We construct a dataset of 2,827 samples by utilizing *informativeness* and *factual correctness* as the evaluation criteria in the multi-criteria prompt. This dataset also comprises both pointwise and pairwise evaluations of chart images and sourced from the Pew split of our single-criterion evaluation dataset. Moreover, similar to the single-criterion scenario, the annotations were generated using the *Gemini-1.5-Pro* model.

4 Experiments

4.1 Experimental Settings

For inference, we set the temperature to 1.0 to ensure diversity in LLM responses. All other decoding parameters were set to their default values, as provided by the respective API providers, OpenAI, Google Vertex AI, and Anthropic for closed-source models, and HuggingFace (Wolf et al., 2020) for open-source models. Given that the task did not require generating lengthy outputs, we limited the maximum output length to 300 tokens. For fine-tuning, we trained the Qwen2-VL-2B-Instruct for three epochs using a batch size of 2, with the learning being tuned between $1e^{-4}$ to $2e^{-5}$. All experiments were run using two NVIDIA H100 GPUs.

4.2 Results and Discussion

We design a parsing script (Laskar et al., 2024a) to extract the target outputs from the LVLM-judge-

Model	Single-Criterion							Multi-Criteria						
	Pairwise ($\uparrow$)			Pointwise ($\downarrow$)			Format Following ($\uparrow$)	Pairwise ($\uparrow$)			Pointwise ($\downarrow$)			Format Following ($\uparrow$)
	FC	I	Avg.	FC	I	Avg.		FC	I	Avg.	FC	I	Avg.	
ChartJudge-2B (Multi-Criteria)	57.70	74.99	66.35	2.15	2.00	2.08	95.83	30.87	62.85	46.86	1.13	2.08	1.61	86.60
Qwen2-VL-2B-Instruct	29.14	69.93	49.54	1.19	0.95	1.07	81.10	5.15	39.17	22.16	3.61	1.89	2.75	44.29
Phi-3.5-vision-3.8B-instruct	29.66	64.34	47.00	0.80	0.83	0.82	79.99	21.65	71.83	46.74	1.03	0.99	1.01	92.28
XGen-MM-Phi3-3.8B-Instruct	57.13	76.63	66.88	1.09	0.99	1.04	96.33	19.80	68.33	44.07	1.46	1.07	1.27	85.54
Qwen2-VL-7B-Instruct	60.11	73.37	66.74	0.89	0.88	0.89	97.78	10.36	70.88	40.62	1.18	1.17	1.18	74.98
InternLM-Xcomposer2d5-7B	60.38	68.80	64.59	0.74	0.73	0.74	94.01	7.78	27.48	17.63	3.41	3.54	3.48	34.88
LLaVA-Next-v1.6-Mistral-7B	62.85	81.41	72.13	0.98	0.85	0.92	98.57	0.51	1.34	0.93	4.86	4.88	4.87	2.29
LLaVA-Critic-7B	69.65	83.90	76.78	0.70	0.58	0.64	99.57	0.00	0.00	0.00	5.00	5.00	5.00	0.00

Table 4: Results in OpenCQA for Single and Multi-Criteria (judgment accuracy for pairwise, error distance for pointwise). Here, ‘FC’ denotes ‘Factual Correctness’ and ‘I’ denotes ‘Informativeness’. ‘ $\uparrow$ ’ = higher is better; ‘ $\downarrow$ ’ = lower is better.

generated judgments. Below, we demonstrate our findings.

4.2.1 Fine-Tuning Results

As shown in Table 2, our fine-tuned model, **ChartJudge-2B**, significantly improves the performance of Qwen2-VL-2B-Instruct in pairwise evaluation across all datasets. It also outperforms larger models in several settings, including Qwen-7B on Vistext L1, and Phi-3.8B on OpenCQA and Vistext L2/L3.

However, it does not necessarily improve the performance in pointwise scenarios (except in Vistext L1). One possible reason for this discrepancy could be because of using different LVLM judgment scores as the reference during training (Gemini-1.5-Pro) and evaluation (GPT-4o/LLaVA-Critic-70B), leading to the alignment variation on the judgment scores. Despite these variations in absolute scores, we found a positive Spearman’s rank correlation (Schober et al., 2018) between ChartJudge-2B and larger models in how they score examples in pointwise evaluation. This indicates a similar alignment in scoring behavior and relative ranking, even if the exact scores differ.

Moreover, we find in Table 3 that the instruction following evaluation capability and format following capability are also increased by a large margin for the fine-tuned Qwen2-VL-2B-Instruct model, although the overall bias also increased. This could be a limitation of fine-tuned LVLM judges, since the LLaVA-Critic-7B model, which achieves the best result in Table 2, also demonstrates high bias.

4.2.2 Performance Comparison: Multi-Criteria vs Single-Criterion

We construct multi-criteria prompts in the Pew split of Chart-to-Text and fine-tune the Qwen2-VL-2B-Instruct model only on this data. From the results

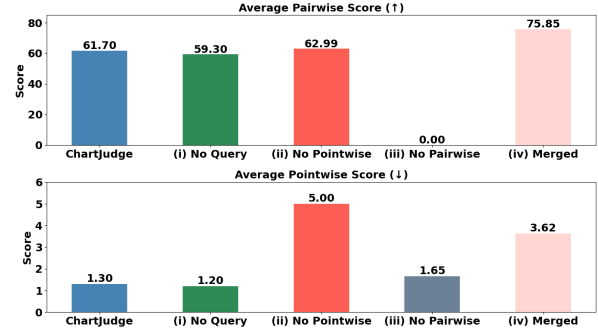


Figure 3: Ablation results for the ChartJudge-2B model in OpenCQA via training (i) without query, (ii) only pointwise samples, (iii) only pairwise samples, and (iv) merged version of only pairwise and pointwise models.

presented in Table 4, we find that most existing LVLMs perform quite poorly in multi-criteria settings. However, our fine-tuned model outperforms most other baselines in this setting, especially in pairwise evaluation. This demonstrates that by fine-tuning on limited multi-criteria judgment data, we can achieve better adaptation using a tiny LVLM, offering further optimization of the inference cost. Surprisingly, the best-performing LLaVA-Critic-7B model achieves the worst result in the multi-query setting. Our analysis shows that LLaVA-based models, LLaVA-Next, and LLaVA-Critic (both 7B & 70B), fail to follow the multi-criteria instruction (see Table 8), leading to poor results. This may happen due to the training data of LLaVA-based models missing multi-criteria judgments.

4.2.3 Ablation Test Results

We conduct ablation tests using ChartJudge-2B to study the following:

- (a) What is the effect of removing synthetic queries in our fine-tuning dataset?
- (b) If we train the model using only pointwise or pairwise data, can model ensure robustness

when evaluated on the data in which they were not trained?

(c) Is it useful to train multiple models, one for pointwise and another for pairwise, and then apply Model Merging (Yang et al., 2024) on each model?

For model merging, the judge LVLM is created using two individual Qwen2-VL-2B-Instruct (Bai et al., 2023) models, one fine-tuned for pair-based judgement and the other fine-tuned for point-based judgement. The implementation was done using the *mergekit* (Goddard et al., 2024) library, and the linear model merging method was used (Wortsman et al., 2022), considering its simplicity. An additional `visual_merger.mlp` layer with all weights initialized to 0 was inserted into both base models to ensure compatibility with the *mergekit* implementation. The weight parameter, representing the contribution of each model to the averaging process, was set to 1.0 for both models.

From our ablation test results presented in Figure 3), we observe the following:

(a) Removing query relevance drops model performance in the pairwise setting.

(b) The model loses its judgment capability in different settings when trained only on one type of data (e.g., model trained only on pairwise cannot do judgments in pointwise and vice versa).

(c) Model merging helps the models trained on individual settings (only pairwise or pointwise) to boost their performance (for instance, the best result in pairwise is achieved using a merged model).

4.2.4 Performance on Human-Annotated Complex Charts and Queries

Following the work of Mukhopadhyay et al. (2024), we annotate complex queries and charts in OpenCQA (see Appendix A.4 for annotation details) by two human experts in data science and computer vision to investigate the performance of our fine-tuned model in complex scenarios. Both annotators label 896 samples as having complex charts and queries, where the ChartJudge-2B model achieves 63.3% judgment accuracy, with its zero-shot counterpart, the Qwen2-VL-2B-Instruct model achieving 54.5% accuracy.

4.2.5 Robustness and Generalization

Robustness on Weak Base Models: To test the robustness of our domain adaptation approach, we applied it to a base model known to be a very weak LVLM judge. We selected PaliGemma-3B, which achieves 0% accuracy as a zero-shot chart judge in

Model	Judgment Accuracy
ChartJudge-2B	49.53
Qwen2-VL-2B-Instruct	80.26
Qwen2-VL-7B-Instruct	90.64

Table 5: Results based on pair-wise comparison across 3K samples in the ChEBI-20-MM dataset for *out-of-domain generalization* evaluation for different models.

the work of (Laskar et al., 2025a). After fine-tuning the model with our proposed method, its pairwise accuracy on the Vistext dataset dramatically improved from 0% to 55.9%. Moreover, the format adherence accuracy is also increased to 77.0% from 0%. This experiment demonstrates that our fine-tuning technique can overcome significant inherent weaknesses in diverse base models and that the performance gains are attributable to our domain adaptation strategy.

Out-of-Domain Generalization Performance:

In our prior experiments, we observe the effectiveness of domain adaptive transfer learning using a 2B parameter model by fine-tuning on a dataset different than our evaluation dataset. Since fine-tuning may also lead to catastrophic forgetting (Luo et al., 2023), in this section, we investigate the performance on a completely different image understanding tasks. For our evaluation, we use the ChEBI-20-MM (Liu et al., 2024) dataset for *caption generation from molecular images* and evaluate the responses generated by Jahan et al. (2025) using the Claude-3-Haiku model and the Gemini-1.5-Flash model. We then evaluate the responses based on pair-wise comparison to select the better model by considering the relevancy, conciseness, informativeness, and factual correctness of the response (see Appendix A.9 for the evaluation prompt). Similar to our prior experiments, we consider GPT-4o annotated judgments as the reference. From the results presented in Table 5, we find that the performance of the ChartJudge-2B model significantly deteriorates after fine-tuning, in comparison to its backbone Qwen2-VL-2B-Instruct model, with the 7B parameter Qwen2-VL performing the best. Therefore, while ChartJudge-2B improves performance in chart benchmarks, it lacks out-of-domain generalization.

4.2.6 Inference Cost and Latency Analysis

In this section, we do some cost and latency analyses. While our ChartJudge model could be deployed in a machine with just 8GB VRAM, other

powerful baselines like Qwen-2-VL-7B-Instruct or LLaVA-Critic-7B require GPUs having at least 24GB VRAM. Therefore, ChartJudge-2B can significantly reduce the deployment cost when used as the LVLM judge to evaluate other LVLMs in real-world industrial settings. Moreover, to ensure efficiency in the CI/CD pipeline, it would also be useful to have a judge model having faster inference. In this case, ChartJudge-2B again demonstrates effectiveness. Based on our evaluation on a machine with 1 A30 GPU, we find that the throughput of ChartJudge-2B is approximately 110 tok/sec (8.1 ms/token latency). On the contrary, the throughputs for the 7B models (e.g., Qwen-2-VL-7B-Instruct, LLaVA-Critic-7B) are approximately 63 tok/sec (15.9 ms/token latency). Moreover, in terms of the inference cost, 7B models can be deployed in GCP² on a machine with 1 L4 GPU with 24GB VRAM, which costs \$0.7 to \$0.9 per hour; whereas 2B models can be deployed using T4 GPU that costs \$0.3 to \$0.5 per hour. This makes ChartJudge-2B twice as fast (while being **2x** cheaper) than the 7B parameter LVLM judges. We also provide further cost and latency benchmarking in Appendix A.6.

4.3 Deployment Considerations

Despite its tiny size, ChartJudge-2B emerges as a pragmatic industrial judge, especially in pairwise scenarios. Since its pointwise error distance remains comparatively high, we recommend deploying it as an efficient pairwise judge that compares the existing deployed model against the newly developed model for pairwise comparison. Moreover, the model’s ability to perform evaluation based on multi-criteria prompting enables cost-efficient inference, which can significantly reduce the usage cost. While the model ChartJudge-2B loses its zero-shot ability when evaluated on an out-of-domain molecular dataset, the *ChEBI-20-MM* benchmark, a brief fine-tune on synthetic judgments following our proposed domain adaptive transfer learning technique can be effective for adaptation across more tasks rather than being a deployment blocker.

5 Conclusion and Future Work

This paper proposes two cost-efficient strategies to improve the usability of tiny LVLMs as evaluation judges in chart comprehension tasks: (i) multi-criteria prompting, and (ii) fine-tuning via knowl-

edge distillation from larger LVLMs to compact 2B-parameter models for better domain adaptation. Our experiments reveal that while various 7B LVLM judges suffer significant performance degradation under multi-criteria settings, our proposed ChartJudge-2B achieves superior performance with just 2B parameters than many larger LVLMs in both single and multi-criteria settings. However, increased bias, with limited improvements in the pointwise evaluation imply the constraints of training on synthetic judgments. In the future, we plan to develop more flexible and modular prompting strategies that can be programmatically adjusted to different application requirements, allowing for dynamic evaluation across diverse chart-related tasks (Rahman et al., 2025; Mahbub et al., 2025), as well as other domains (Laskar et al., 2025b). Furthermore, to address the increased bias observed in the fine-tuned model, we will conduct experiments to mitigate length bias. Finally, we will continue to study how to improve synthetic judgment data to further improve overall performance.

Limitations

While our proposed fine-tuned model may not always outperform the larger LVLMs that are used as the baseline, the main focus of this work is to demonstrate the effectiveness of knowledge distillation in the low-resource setting in the chart domain for LVLM-judges. Meanwhile, our proposed technique helps the model to outperform other models of similar size, alongside demonstrating robustness. Moreover, due to the rapid progress in LLM and LVLM development, there may have other tiny models that could be studied instead of Qwen2-VL-2B-Instruct model. However, we select Qwen2-VL due to the effectiveness of this model in prior research (Laskar et al., 2025a) (see Appendix A.1 for our explanation on model selection). Since fine-tuning LVLMs require huge computing resources, we also focused to demonstrate the effectiveness of a 2B parameter model considering real-world scenarios. Moreover, for multi-criteria prompting, we only tested with factual correctness and informativeness since our focus was just to investigate how the LVLMs perform in the multi-criteria scenario in terms of the two commonly used qualitative evaluation metrics (Masry et al., 2023). Nonetheless, future work may investigate the performance in other cases (e.g., relevance, conciseness) in the multi-criteria prompting approach.

²<https://cloud.google.com/compute/gpus-pricing>, last accessed: July 4, 2025

Acknowledgments

We thank all the anonymous reviewers of the EMNLP Industry Track for their excellent review comments. This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the York Research Chairs (YRC) program, Canada Foundation for Innovation (CFI), Google’s Gemini Academic Program for the API Credits, and Compute Canada for the computing resources.

Ethical Statements

The models used for experiments do not pose any ethical concerns since they are only used as the judge to evaluate other LVLm-generated responses. Additional compensation for human evaluation is not needed since it was conducted by the authors of this paper. Licensing requirements were maintained accordingly while using different artifacts (OpenAI, Gemini, Anthropic, HuggingFace).

References

- Marah Abidin, 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](#).
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#).
- 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.
- 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*.
- Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, and et al. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#).
- Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. [Arcee’s MergeKit: A toolkit for merging large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.
- 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.
- Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. 2022. Chart question answering: State of the art and future directions. In *Computer Graphics Forum*, volume 41, pages 555–572. Wiley Online Library.
- 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.
- Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2025. Evaluating the effectiveness of cost-efficient large language models in benchmark biomedical tasks. *arXiv preprint arXiv:2507.14045*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- 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, Toronto, Canada. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.
- Md Tahmid Rahman Laskar, Enamul Hoque, and Xiangji Huang. 2020. Wsl-ds: Weakly supervised learning with distant supervision for query focused multi-document abstractive summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5647–5654.
- Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Ridwan Mahbub, Ahmed Masry, Mizanur Rahman, Amran Bhuiyan, Mir Tafseer Nayeem, Shafiq Joty, Enamul Hoque, and Jimmy Huang. 2025a. [Judging the judges: Can large vision-language models fairly evaluate chart comprehension and reasoning?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1203–1216, Vienna, Austria. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Israt Jahan, Elham Dolatabadi, Chun Peng, Enamul Hoque, and Jimmy Huang. 2025b. [Improving automatic evaluation of large language models \(LLMs\) in biomedical relation extraction via LLMs-as-the-judge](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25483–25497, Vienna, Austria. Association for Computational Linguistics.
- 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. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.
- Pengfei Liu et al. 2024. Scientific language modeling: A quantitative review of large language models in molecular science. *arXiv preprint arXiv:2402.04119*.
- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.
- Ridwan Mahbub, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, and Enamul Hoque. 2025. The perils of chart deception: How misleading visualizations affect vision-language models. *arXiv preprint arXiv:2508.09716*.
- Ahmed Masry, Xuan Long Do, 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.
- Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025a. [ChartQAPro: A more diverse and challenging benchmark for chart question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria. Association for Computational Linguistics.
- 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, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2025b. 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.
- Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024. Unraveling the truth: Do vlms really understand charts? a deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16696–16717.
- OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. 2023. [Gpt-4 technical report](#).
- Mizanur Rahman, Md Tahmid Rahman Laskar, Shafiq Joty, and Enamul Hoque. 2025. Text2vis: A challenging and diverse benchmark for generating multimodal visualizations from text. *arXiv preprint arXiv:2507.19969*.

- Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.
- Kanharaj 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.
- Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. 2025. A survey on efficient vision-language models. *arXiv preprint arXiv:2504.09724*.
- 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.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- 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*.
- Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

A Appendix

A.1 Regarding Model and Dataset Selection

For the model and dataset selection, we follow the work of (Laskar et al., 2025a). However, we remove the models for our experiments that perform poorly in their work (e.g., larger models with at least 7B parameters that achieve below 50% judgment performance). Based on this criterion, we selected the following models as the baseline: (i) XGen-MM-Phi-3-3.8B (Xue et al., 2024) – a multimodal model (3.8B) developed by Salesforce, (ii) Phi-3.5-3.8B-Vision-Instruct (Abdin et al., 2024) – a smaller vision model from Microsoft, (iii) Qwen2-VL-2B - Alibaba’s Qwen (Wang et al., 2024) multimodal model with just 2B parameters, (iv) Qwen2-VL-7B – The 7B version of the multimodal Qwen model, (v) InternLM-XComposer-7B (Dong et al., 2024) – a 7B vision model with composition abilities, (vi) LLaVA-v1.6-Mistral-7B – A multimodal LVM based on the LLaVA (Li et al., 2024) architecture that also utilizes a 7B Mistral (Jiang et al., 2023) as the backbone, and the (vii) LLaVA-Critic-7B – a specialized evaluator model based on LLaVA and Qwen.

A.2 Evaluation Process

In this section, we describe our evaluation process in detail. Note that we compute the results for different LVLMs across various evaluation metrics by comparing with GPT-4o and LLaVA-Critic-70B annotated judgments, and use the average score as reference to compare the performance of the LVLMs. Below, we illustrate the evaluation rubric and metrics.

(i) Based on Evaluation Type:

- **Pairwise:** Judge selects the better of the two model-generated responses.
- **Pointwise:** Judge rates a single response on a 1–5 Likert scale.

(ii) Based on Reference Type:

- **With Reference:** Ground-truth answer is provided with the chart and the response for judgment.
- **Without Reference:** Judge relies only on the chart and response(s).

(iii) Based on Evaluation Criteria:

Type	# Single-Criterion	# Multi-Criteria
Bar	7737	1857
Line	1622	792
Pie	148	98
Table	138	-
Area	51	51
Scatter	29	29

Table 6: Fine-tuning dataset statistics for single-criterion and multi-criteria settings based on Chart Type. Here, ‘#’ denotes total number of samples.

- **Factual Correctness:** Assesses factual accuracy of the response.
- **Informativeness:** Measures the usefulness of the information in the response.
- **Relevance:** Evaluates whether the response is relevant to the chart, and the query (in QA tasks).
- **Multidimensional Evaluation:** Considers overall response quality by considering factual correctness, informativeness, and relevance.

(iv) Based on Evaluation Metrics:

- **Judgment Accuracy:** Check for exact match between the model-predicted judgment and the gold judgment (pairwise).
- **Error Distance:** Measures the average deviation from the reference rating (pointwise).
- **Bias Metric:** Checks if judgment changes with response order or judge mistakenly prefers the wrong answer due to bias towards lengthy responses (pairwise).
- **Instruction Following Evaluation Accuracy:** Measures the judge’s ability to assess instruction following capabilities of other LVLMs in chart-related tasks.
- **Format Adherence Accuracy:** Evaluates whether the response generated by the judge follows the required output format (e.g., JSON).

A.3 Fine-Tuning Dataset Statistics

We demonstrate the statistics of our fine-tuning dataset based on chart type in Table 6.

Model Size	# Tokens		Avg. Time	
	CPU	GPU	CPU	GPU
7B Parameters	0.06	1.5	15.9	0.63
3.8B Parameters	0.13	3.3	7.8	0.30
2B Parameters	0.40	3.7	2.5	0.25

Table 7: Inference Benchmarking of different-sized LVLMs. Here, ‘# Tokens’ denotes the average number of tokens per second, while ‘Avg. Time’ denotes the average time (in seconds) to generate each token.

A.4 Annotation of Complex Charts & Queries

Human Annotation Guidelines

Chart and Question Categorization: To evaluate model performance, you are required to annotate chart images and questions based on complexity:

- **Simple Charts:** Contain two columns (representing dependent and independent variables).
- **Complex Charts:** Contain more than two columns, depicting multiple variables and visual elements.
- **Simple Questions:** Focus on single-step data extraction.
- **Complex Questions:** Require multi-step reasoning, comparisons, and logical inferences.

A.5 Question Generation Prompt

Chart-to-Text Dataset (Pew Split)

Generate a concise question related to the summary of the given chart.

[Model Generated Summary]

[Chart Image]

A.6 Benchmarking Inference Cost and Latency

We also benchmarked the inference speed of models of varying sizes on an AMD machine with a 64-core processor, testing both GPU (1 A100) and CPU-only modes. As shown in Table 7, smaller models exhibit significantly faster inference, especially in resource-constrained, CPU-only environments. Our 2B model achieves an average time per token of 2.5 seconds on a CPU, whereas a 7B model takes 15.9 seconds. This highlights the real-world value of using a fine-tuned tiny model for efficient, low-latency evaluation.

A.7 Multi-Criteria Prompts for LVLM Judge

OpenCQA Pointwise (With Reference)

Suppose, you are a human evaluator. You are required to rate the **Informativeness** and **Factual Correctness** 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.

Provide your response in an Array of JSON with the following keys: (i) Score, (ii) Explanation, (iii) Type.

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 reasoning behind your scoring, while the value for the 'Type' key will contain either 'informativeness' or 'factual correctness', depending on the type for which you assess the answer.

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 **Informativeness** and **Factual Correctness**.

Provide your response in an Array of JSON with the following keys: (i) Model, (ii) Explanation, (iii) Type.

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, while the value for the 'Type' key will contain either 'informativeness' or 'factual correctness', depending on the type for which you assess the answer.

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 A Generated Answer]

[Model B Generated Answer]

[Chart Image]

A.8 Sample Response in Multi-Criteria Prompting

Response in OpenCQA (Pairwise)

```
[
  {
    "Model": "Model B",
    "Explanation": "Model B provides a thorough breakdown of the Mexican-American population growth, with the numbers accurately matching the chart.",
    "Type": "Informativeness"
  },
  {
    "Model": "Model A",
    "Explanation": "Model A is factually correct as it accurately reports the figures for both new immigrants and births during the decade 2000-2010.",
    "Type": "Factual Correctness"
  }
]
```

A.9 LLM-Judge Prompt for Out-of-Domain Generalization Evaluation

Pairwise Evaluation on ChEBI-20-MM for Molecular Image Captioning

You are given the summaries generated by two different models in the molecular image 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 provide high level informative insights of the molecular image.

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.

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 given the model generated captions. You are also provided with the molecular image which was used as the context for the molecular image captioning task.

[Model A Generated Caption]

[Model B Generated Caption]

[Chart Image]

A.10 Error Examples

Some error examples in multi-criteria settings for different LVLMs are shown in Table 8.

