Instruction-tuned QwenChart for Chart Question Answering

Viviana Ventura, Lukas Kleybolte, Alessandra Zarcone
Technische Hochschule Augsburg
viviana.ventura, lukas.kleybolte, alessandra.zarcone@tha.de

Abstract

Charts, where information is delivered jointly by visual and textual features, represent a challenge when it comes to downstream tasks such as chart question answering, where both kinds of information contribute to the task. The standard approach is to decouple the task in two steps, first extracting information from the charts, or representing it as a table, text or code, and then a second reasoning step to output the answers. Today, the advancements in visual encoding of Visual Large Language Models (VLLM) have shown their capabilities to solve such complex tasks without using in-between representations of the charts or massive in-domain training. We propose a solution for the Scientific Visual Question Answering (SciVQA) Shared Task, on which our team THAii_LAB scored the second position in the final leaderboard. Our new instruction fine-tuned and Chain-of-Thought (CoT) model QwenChart-7B showed that even in a complex new benchmark general models can achieve great performances with low-cost training, matching the capabilities that LLMs have showed in unimodal downstream tasks. An out-of-domain evaluation showed satisfactory results, albeit with an expected drop in performance.

1 Introduction

Everything in a chart conveys information: besides labels such as numbers or text, they feature shapes, colors and complex visual elements such as bars, lines or points that contribute to the delivery of their meaning. Understanding complex texts such as scientific articles also requires chart comprehension, including answering questions about charts in natural language (QA over charts or chart QA). To tackle this task, previous work has focused on two main aspects: information extraction from charts and complex, often logical or arithmetic, reasoning over that information.

Early approaches would identify and extract information to feed into a classifier (Kafle et al., 2018; Chaudhry et al., 2020). Since the rise of Visual Large Language Models (VLLMs), many approaches convert charts into a format suitable for a language model, such as text descriptions, (Liu et al., 2023a), tables, or code (Lee et al., 2023; Liu et al., 2023b; He et al., 2025), due to the limited resolution capability of the visual encoders, and the conversely great capabilities of the LLMs. While using tables instead of images leads to some information loss, this approach still remains preferable.

Despite reaching satisfactory performance in general visual understanding tasks, VLLMs have struggled with downstream chart understanding tasks (Huang et al., 2024; Islam et al., 2024; Li et al., 2024a; Lu et al., 2024; Xu et al., 2025a,b). VLLMs usually consist of a visual encoder and a language decoder. The complexity of the visual features of charts represents a bottleneck for visual encoders, whereas the language decoder struggles to extract the necessary information from the visual representations due to the complexity of the relations between visual and linguistic elements (Liu et al., 2025). A common approach today involves augmenting data with task-specific instructions and fine-tuning a pre-trained model accordingly (Han et al., 2023; Islam et al., 2024; Liu et al., 2024; Masry et al., 2024, 2025). Some researchers have also opted to train the visual encoder using chart–table pairs to enhance its representational capabilities (Han et al., 2023; Islam et al., 2024; Liu et al., 2024; Masry et al., 2024, 2025; Xu et al., 2025b).

Borisova et al. (2025) introduced the Scientific Visual Question Answering (SciVQA) shared task[1], designed to evaluate multi-modal QA systems on real-world scientific figures through a diverse set of

[1] https://www.codabench.org/competitions/5904/#/pages-tab

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 240–251
July 31, 2025 ©2025 Association for Computational Linguistics

both finite and infinite questions. The task emphasizes reasoning over complex visualizations and includes chart types rarely represented in earlier datasets, such as architecture diagrams, confusion matrices, and compound figures. In this context, we propose QwenChart-7B, a vision-language model specifically designed for the SciVQA task which achieved second place in the competition.

QwenChart-7B has been instruction-tuned with Low Rank Adaptation (LoRa, Hu et al. 2021) and exploits Chain-of-Thought (CoT, Wei et al., 2022) to improve its reasoning capabilities. We show that QwenChart-7B is capable of achieving good performance in chart QA without pretraining on domain data or using in-between representations of charts, such as tables. Furthermore, we show that scaling size of the model does not have a great impact on performance and identify what parameters mostly contribute to the performance of the model. Our model is one of the first visual models reaching high performance on a challenging benchmarks such as SciVQA without using intermediate representations of charts.

Our contributions are the following:

- a new instruction-tuned VLLM (QwenChart-7B) that achieves high scores in a challenging benchmark such as SciVQA, reaching the second place in the SciVQA shared task;
- several experiments, showing the influence of parameters, size of the model and additional information in the training data during fine-tuning.

2 Related work

2.1 Data and Benchmarks

Early benchmarks for chart QA included a limited variety of charts, more often synthetically generated than derived from real-world sources. DVQA (Kafle et al., 2018) and FigureQA (Kahou et al., 2018) are the first datasets for factoid QA over synthetically generated line, bar, and pie charts. Early datasets provided an alignment with structured auxiliary data such as numerical data or tables (Kahou et al., 2018; Masry et al., 2022), which was necessary to compensate for the lack of sufficiently robust methods to directly extract graph components (Luo et al., 2021; Rane et al., 2021; Kato et al., 2022).

ChartQA (Masry et al., 2022) is one of the most widely used benchmarks for chart understanding and features both synthetically generated and real-world graphs.

SciVQA[2] (Borisova et al., 2025) is a new chart corpus built from two pre-existing datasets, ACL-Fig (Karishma et al., 2023) and SciGraphQA (Li and Tajbakhsh, 2023). The 3000 figures are from English scientific publications from the ACL Anthology[3] and arXiv[4]. Unlike other datasets, it is composed exclusively of real-world figures, rather than synthetic data and features a wide variety of figure types, including trees, architecture diagrams, neural networks, confusion matrices, scatter plots, and box plots. In addition, it is annotated both with finite and infinite questions, as well as unanswerable questions. The figures are paired with captions and chart types as additional metadata. An additional challenge in SciVQA are figures with more than one chart.

2.2 Limitations of VLLMs in chart QA

Despite recent advancements in tasks such as image understanding brought forward by the emergence of VLLMs, QA over charts remains challenging. Typical approaches focus on two different aspects: (1) understanding the chart, that is extracting its meaningful components, such as numbers, labels but also shape, colors and position of points and (2) reasoning over the extracted information, for example, to compute mathematical operations based on numbers extracted from the figures.

Early approaches used encoder-only classification-based models to encode chart and question separately, and combining them later with attention blocks (Kafle et al., 2018; Chaudhry et al., 2020; Singh and Shekhar, 2020), but were often limited as they had a fixed output vocabulary (Santoro et al., 2017; Kafle et al., 2018; Kahou et al., 2018).

Recently, VLLMs have demonstrated remarkable capabilities in various chart comprehension tasks, outperforming specialized models (Huang et al., 2024), such as ChartBERT (Akhtar et al., 2023), MatCha (Liu et al., 2023b) or UniChart (Masry et al., 2023). However, VLLMs are not as good at chart understanding as they are in other visual tasks (Huang et al., 2024; Islam et al., 2024; Li et al., 2024a; Lu et al., 2024; Xu et al., 2025a,b).

[2] https://huggingface.co/datasets/katebor/SciVQA

[3] https://aclanthology.org

[4] https://arxiv.org

Proprietary models such as GPT-V[5], Gemini (Team et al., 2025) and Claude[6] currently achieve the best results in zero-shot scenarios in most of vision-language benchmarks, showing strong zero/few shot inference capabilities. Models such as GPT-4o[7] have shown unprecedented performance in chart understanding compared to open-source models (Islam et al., 2024; Wang et al., 2024), such as Phi-3 (Abdin et al., 2024) or LlaVA (Liu et al., 2023c). However, the performance is not comparable to that achieved in non-visual tasks.

There are two major bottlenecks in chart understanding: the perception capabilities of existing VLLMs are limited (Razeghi et al., 2024; Zhang et al., 2024b), and they fail in extracting the necessary information from the provided visual representations (Liu et al., 2025). Therefore, the state of the art approach separate a vision encoder and a text decoder stage, with a stronger focus on the former or the latter. Common approaches are to transform charts into structured formats, such as tables, code, and text (Lee et al., 2023; Liu et al., 2023a,b; Zhou et al., 2023), as a bridge to a text decoder to leverage the power of LLM in reasoning.

Some authors have stressed the impact of input resolution on pre-training and fine-tuning (Zhang et al., 2024a). The standard procedure would be to resize images into fixed resolution to reduce the length of the visual feature sequence. However, high and native resolution are essentials for chart understanding. Models such as Tinychart (Zhang et al., 2024a) tried to solve this issue merging visual tokens inside each vision transformer layer.

2.3 Instruction-tuned VLLMs

Some authors point out that VLLMs still struggle in analyzing charts due to the weak alignment between vision and language caused by the lack of charts in pre-trained model data (Xu et al., 2025b).

Recently, many VLLMs models have been trained on charts to improve their representations including ChartLLaMa (Han et al., 2023), ChartAssistant (Islam et al., 2024), MMC (Liu et al., 2024), ChartInstruct (Masry et al., 2024), and ChartGemma (Masry et al., 2025).

Besides being trained on charts, all these models follow the same methodology: they use chart-specific instruction tuning[8] to enhance the extraction capability of the language decoder (Liu et al., 2023c; Islam et al., 2024; Liu et al., 2024; Masry et al., 2024, 2025). With instruction tuning the model should learn to understand and internally represent the components of a chart, such as axes, labels, bars, trends. Hence, the first step is to augment dataset of charts with instructions, rationales or CoT data (Wang et al., 2023; Carbune et al., 2024; Huang et al., 2024; Jia et al., 2024; Li et al., 2024b; Kim et al., 2025; Wang et al., 2025).

While showing promising results, models which are fine-tuned on task-specific datasets show their limits when it comes to generalizing on unseen data.

3 QwenChart

3.1 Model Architecture

To develop our model[9], **QwenChart**, we fine-tuned Qwen2.5-VL (Bai et al., 2025) using LoRa (Hu et al., 2021) on an instruction-based chart dataset generated via dynamic CoT prompting (Wei et al., 2022). With LoRa, the original model weights are frozen and only a few new parameters are trained. Instead of updating all the weights in a large matrix, LoRa inserts small trainable matrices that approximate the change, thus maintaining the capabilities of the original model intact while reducing the computational cost. Our dataset comprises chart images and associated metadata from the SciVQA dataset (Borisova et al., 2025). Section 3.2 describes the process we followed to augment SciVQA.

Our model is particularly suited for chart tasks thanks to the dynamic encoding, i.e., the ability to receive images with different sizes as input without the need for normalization. As discussed in Section 2.2, native and high resolution are two important features for chart understanding. Bai et al. (2025) trained a Vision Transformer (ViT, Dosovitskiy et al. 2021) from scratch with native dynamic encoding to maintain images (or videos) with native resolution. They also incorporate a Window Attention in the ViT. The model comes in 4 sizes: 3B, 7B, 32B and 72B. We used the 7B model and compared it with the 72B. The model is composed by a visual encoder, a cross-modal projector and a text decoder.

[5] https://openai.com/index/gpt-4v-system-card/

[6] https://www.anthropic.com/news/claude-3-family

[7] https://openai.com/index/hello-gpt-4o/

[8] Llava (Liu et al., 2023c) is the first attempt to use instruction tuning with multi-modal models.

[9] https://github.com/tha-atlas/QwenChart

3.2 Pre-Processing

3.2.1 Data augmentation with dynamic prompting and Chain-of-Thought

To prepare the SciVQA dataset for fine-tuning we built a dynamic prompting pipeline with instructions and CoT.

For each question-chart pair a different prompt was generated. Figure 2 (in the Appendix A.1), presents two example prompts for a single question–chart pair. The first prompt is specifically designed to match the format of questions found in the SciVQA dataset. The second is a more generic prompt we developed to facilitate experiments on other benchmarks, allowing for prompt adaptation based on the target dataset.

The prompt is built using the metadata from SciVQA. The first information provided in the prompt is the type of chart, then the caption. Then the question is provided, followed by some clues about the information the model should focus on. This can change based on the type of question that is provided. The model was instructed to provide concise answers, as Qwen2.5-VL tends to generate overly verbose responses.

To support this claim, we conducted a controlled comparison using two prompting strategies:

Simple Question Prompt: the prompt contains only the question;

Dynamic Prompt (ours): a structured prompt instructing the model to provide a concise answer.

We observed a significant difference in response length. On average, answers generated using the Simple Question Prompt were approximately 31.18 words, while responses using our Dynamic Prompt averaged just 1.35 words, closely aligning with the gold standard answers (1.32 words on average). An example of answers generated by the model with the two different prompts can be found in Appendix A.2. This experiment confirms that explicit prompting for brevity is essential to prevent unnecessarily long and redundant answers from Qwen2.5-VL. Given that we use ROUGE-1 and ROUGE-L metrics (Lin, 2004) for evaluation, it was essential to produce outputs that closely matched the gold standard. For this reason, we also specified the use of digits only and the inclusion of appropriate suffixes. Moreover, the instruction on how to respond when a question was unanswerable was included to ensure consistency with the format of the gold standard. Additional instructions were adapted based on the nature of the question. For example, whether it involved multiple-choice or binary-choice formats, or if addressed six visual attributes or not (shape, size, position, height, direction or colour). The final section of the prompt, labeled <thinking>, represents the CoT component. We observed that including this step encourages the model to engage in self-reflection, resulting in more reasoned and coherent responses. The CoT prompting leads to a substantial improvement across all evaluated metrics, with ROUGE-1 F1 increasing from 72.41% to 79.23% and ROUGE-L F1 from 72.30% to 79.06% - reflecting a gain of nearly 7 points in both cases.

3.2.2 Image Pre-processing

As an additional preprocessing step prior to fine-tuning, we applied a 10% white padding uniformly around each image in the dataset. This modification was introduced after observing that the model exhibited difficulties in accurately recognizing objects located near the image boundaries. Two human annotators manually checked the results from first experiments on 100 QA pairs and identify this tendency in the model.

3.2.3 Conversation-Based Queries

We converted every dataset entry from SciVQA in conversation-based queries that contained the prompt as described in Section 3.2.1, with the goal of using the queries as training data. Each entry in the SciVQA dataset consists of an image paired with a corresponding question, along with additional metadata (figure type, figure caption, and question category). The question type is classified as unanswerable, infinite, or finite (e.g., multiple choice or binary), and is further annotated as either visual or non-visual depending on whether it involves any of six predefined visual attributes: shape, size, position, height, direction, or color. In the conversation query we added this system message: "You are a Vision Language Model specialized in interpreting visual data from chart images. Your task is to analyze the provided chart image and respond to queries with concise answers, usually a single word, number, or short phrase. The charts include a variety of types (e.g., line charts, bar charts) and contain colors, labels, and text. Focus on delivering accurate, succinct answers based on the visual information. Avoid additional explanation unless absolutely necessary".

During the fine-tuning process, the gold (ground-truth) answer was included at the end of each conversational query, in order to provide the model

with supervised learning signals. This information was excluded in the testing phase.

4 Experimental setup

4.1 Instruction-tuned QwenChart with dynamic prompting

We performed supervised fine-tuning of Qwen2.5-VL (Bai et al., 2025) on the training set of SciVQA prepared as described in Section 3.2. Specifically, we set up a rank of r = 64, an alpha (scaling factor) of 32. The dropout rate is set to 5%. We applied LoRA to the query, key, value and output projection layers of the attention modules of the text decoder and to the gate, up, down projectors of the Multi-Layer Perceptron. All other parameters, including the visual encoder, remained frozen during fine-tuning.

The total number of Qwen2.5-VL is 9,537,950,720, we trained the 13.0912% of them. Training was conducted for 2 epochs with an effective batch size of 24 (batch size = 6, gradient accumulation = 4), using a learning rate of 2e-4 and bfloat16 precision. Experiments were run on 8× H100 (80GB) GPUs. This version of the model, called **QwenChart-7B**, is the one used for the final submission on the leaderboard of the SciVQA shared task (Borisova et al., 2025). Furthermore, we fine-tuned the 72B Qwen2.5-VL version following the same configuration to see how it copes with the scaling up of the model. This version is called **QwenChart-72B**.

4.2 Instruction-tuned QwenChart with general prompting

We developed a different version of the prompt that can be adapted to other datasets, as illustrated in Section 3.2.1 and in Figure 2 in Appendix A.1. We fine-tuned Qwen2.5-VL on the the training set of SciVQA, prepared as discussed in Section 3.2, but using the adapted prompt version. For this model, we use the same configuration detailed in Section 4.1. This version of the model, **QwenChart2-7B**, does not include captions in the training data.

4.3 Evaluation

We evaluate the performance of our proposed models —**QwenChart-7B**, **QwenChart2-7B**, and **QwenChart-72B**— on both the development and test sets of the SciVQA benchmark (Table 1). To assess generalization capabilities, we also evaluate QwenChart-7B on ChartQA (Masry et al., 2022) (Table 1, last row), a widely adopted benchmark for chart question answering.

For comparison, we report zero-shot performance of two strong baseline models: the original **Qwen2.5-VL** and **Gemma3-12B-IT**[10]. For this results we used the dynamic prompt as in Section 3.2.1. These results, presented in Table 1, serve as a reference point to quantify the impact of fine-tuning and instruction design in our models.

We conduct our evaluation across different scenarios using ROUGE-1, ROUGE-L, and BERTScore (Zhang et al., 2020).

5 Analysis and Discussions

Table 1 shows that QwenChart-7B is the top performer on SciVQA across all metrics (highest ROUGE-1, ROUGE-L and BERTScore), indicating both lexical and semantic closeness to the ground truth. It slightly outperforms the larger QwenChart-72B, suggesting size alone does not guarantee better performance. QwenChart2-7B shows a performance drop on ChartQA. This implies that our model is not robust enough for generalization on out-of-domain data. Qwen2.5-VL on zero-shot performs well on SciVQA especially if compared to Gemma3.12b-it. The QwenChart models (7B, 72B, and 2-7B) show consistently high performance, but we notice a significant increase in performance with version QwenChart-7B. We also observe that figure captions have limited impact on results, as QwenChart2-7B achieves strong performance despite not being trained with caption information.

One of the key contributions of this work is the demonstration that high performance on chart understanding can be achieved using a visual model that does not rely on intermediate representations such as tables or code. This is particularly significant in the context of the SciVQA benchmark, which features a diverse set of real-world charts. The strong performance of QwenChart-7B, which surpasses even its larger counterpart (QwenChart-72B), suggests that model architecture and prompt engineering may have a more substantial impact on downstream performance than model size.

Another advantage lies in the efficient training process enabled by LoRA. By fine-tuning only 13% of the model's parameters, we achieve competitive results while significantly reducing computational cost and preserving the core capabilities of the pre-

[10] https://huggingface.co/google/gemma-3-12b-it

	ROUGE-1			ROUGE-L			BERTScore		
Model	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Qwen2.5-VL (dev set)	71.57%	72.96%	71.72%	71.52%	72.88%	71.67%	97.29%	97.38%	97.25%
Gemma3-12b-it (dev set)	60.96%	62.83%	60.43%	60.92%	62.78%	60.41%	96.61%	96.75%	96.52%
QwenChart-7B (test set)	78.99%	79.60%	79.49%	78.92%	79.53%	79.42%	98.39%	98.41%	98.40%
QwenChart-7B (dev set)	79.23%	80.24%	79.25%	79.06%	80.05%	79.08%	98.40%	98.50%	98.33%
QwenChart-72B (dev set)	77.54%	78.29%	77.93%	77.40%	78.16%	77.79%	98.23%	98.29%	98.19%
QwenChart2-7B (dev set)	76.62%	77.25%	77.16%	76.50%	77.13%	77.03%	98.19%	98.22%	98.19%
QwenChart2-7B (ChartQA)	66.38%	66.46%	67.20%	66.28%	66.27%	67.10%	94.69%	94.19%	95.23%

Table 1: Evaluation metrics across models on development and test set of SciVQA and ChartQA (validation set) (last row).

	ROUGE-1			ROUGE-L			BERTScore		
QA type	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
finite binary non-visual	79.86%	79.86%	79.86%	79.86%	79.86%	79.86%	100.0%	100.0%	100.0%
finite binary visual	78.93%	78.93%	78.93%	78.93%	78.93%	78.93%	100.0%	100.0%	100.0%
finite non-binary non-visual	75.68%	75.79%	77.96%	74.5%	74.61%	76.79%	98.25%	98.07%	98.43%
finite non-binary visual	65.36%	65.0%	67.0%	65.36%	65.0%	67.0%	98.5%	98.21%	98.79%
infinite non-visual	74.46%	76.0%	75.5%	74.43%	75.96%	75.5%	96.39%	96.43%	96.54%
infinite visual	62.36%	63.36%	62.57%	62.04%	62.96%	62.21%	96.79%	96.82%	96.86%
unanswerable	95.0%	95.0%	95.0%	95.0%	95.0%	95.0%	99.11%	99.14%	99.07%

Table 2: Evaluation metrics of QwenChart-7B on the development set of SciVQA by QA type.

trained model. Dynamic prompting, combined with CoT rationales, further enhances the model's reasoning capabilities. This strategy allows the model to decompose complex questions into intermediate logical steps, resulting in more coherent and contextually accurate responses.

Despite promising results on SciVQA, our experiments reveal a performance drop on the ChartQA benchmark, indicating that the model's generalization capability to out-of-domain data is limited. This suggests potential overfitting to the prompt format or chart types seen during fine-tuning. Further efforts are needed to enhance the robustness of instruction-tuned models across datasets.

We observed that the ROUGE-1, ROUGE-L, and BERTScore metrics exhibit certain limitations when applied to this type of task. Compared to BERTScore, ROUGE proves to be more sensitive, as it is better able to highlight performance differences. ROUGE, in fact, imposes a heavier penalty on responses that do not exactly match the gold standard, making it more suitable for this task. However, this can also lead to an underestimation of model performance when responses are correct but differ in form from the reference answers. Table 3 shows some illustrative examples.

Answer	**Gold Answer**
RANDOM, SSID	RANDOM and SSID
0.4	0.32–0.52
Three	3
IT	Italian
A B C D	A,B,C,D

Table 3: Examples of QwenChart-7B answers vs gold answers from SciVQA development set.

5.1 Error Analysis

To gain deeper insights into model performance across different chart and question types, we conducted a quantitative analysis of the performance of QwenChart-7B on the development set of SciVQA (Table 2 and Table 4). The results reveal several notable patterns in how QwenChart-7B handles various categories of questions within the figure type.

First, we observe that binary (yes/no, true/false) answer set questions—both visual[11] and non-visual —yield the highest performance across all metrics. This suggests that the model excels when the answer space is limited and well-structured. Similarly, multiple choice visual questions also perform

[11]A visual question in SciVQA dataset is a question that addresses six designated features of the image: shape, size, position, height, direction or color.

	ROUGE-1			ROUGE-L			BERTScore		
Figure Type	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Line Chart	68.79%	69.78%	68.86%	68.72%	69.72%	68.86%	97.64%	97.5%	97.57%
Line Chart, Table	85.71%	85.71%	85.71%	85.71%	85.71%	85.71%	98.29%	98.14%	98.57%
Tree	71.28%	72.57%	70.72%	71.28%	72.57%	70.72%	98.42%	98.58%	98.5%
Scatter Plot	71.5%	71.79%	71.36%	71.5%	71.79%	71.36%	98.42%	98.78%	98.15%
Pie Chart	84.29%	83.71%	85.71%	84.29%	83.71%	85.71%	99.29%	99.14%	99.43%
Architecture Diagram	91.15%	91.5%	90.93%	90.93%	91.22%	90.72%	99.57%	99.65%	99.5%
Box Plot	79.71%	78.57%	82.14%	79.71%	78.57%	82.14%	98.71%	98.43%	99.14%
Neural Networks	83.14%	83.14%	83.28%	83.14%	83.14%	83.28%	99.5%	99.57%	99.65%
Confusion Matrix	81.71%	81.43%	83.57%	81.71%	81.43%	83.57%	97.57%	97.29%	97.43%
Graph	76.5%	76.86%	77.85%	76.07%	76.36%	77.35%	98.15%	98.08%	98.28%
Bar Chart	73.0%	73.86%	73.43%	73.0%	73.86%	73.43%	97.71%	97.57%	98.0%
Histogram	83.35%	85.71%	82.14%	83.35%	85.71%	82.14%	99.35%	99.5%	99.22%
Venn Diagram	85.71%	85.71%	85.71%	85.71%	85.71%	85.71%	100.0%	100.0%	100.0%
Vector Plot	95.29%	100.0%	92.86%	95.29%	100.0%	92.86%	97.86%	98.0%	97.71%
Other	35.14%	32.86%	42.86%	35.14%	32.86%	42.86%	98.29%	97.71%	98.86%
Line Chart, Bar Chart	42.86%	42.86%	42.86%	42.86%	42.86%	42.86%	97.29%	97.43%	97.14%
Flow Chart	85.71%	85.71%	85.71%	85.71%	85.71%	85.71%	98.0%	98.57%	97.57%
Tree, Graph	62.86%	61.86%	64.29%	58.14%	57.14%	59.57%	95.57%	94.86%	96.57%
Illustrative Diagram	74.57%	74.29%	75.0%	74.57%	74.29%	75.0%	98.14%	97.86%	98.29%
Line Chart, Scatter Plot	71.43%	71.43%	71.43%	71.43%	71.43%	71.43%	100.0%	100.0%	100.0%
Heat Map	77.14%	75.0%	85.71%	77.14%	75.0%	85.71%	97.29%	96.43%	98.29%

Table 4: Evaluation metrics of QwenChart-7B on the development set of SciVQA by figure type.

strongly, indicating that the model handles moderate complexity well.

On the other hand, performance drops for visually-anchored queries. Specifically, infinite visual questions scored the lowest. This may be due to the model's difficulty in generating precise free-form answers from ambiguous or densely visual inputs without clearly bounded outputs.

Table 4 demonstrates that the type of figure significantly impacts model performance. "Vector Plot" yielded the highest overall performance with scores of 95.29% ROUGE-1 F1 and 97.86% BERTScore F1, indicating the model's strong ability to extract and interpret information from this format. "Pie Chart", "Architecture Diagram", and "Neural Networks" also demonstrated consistently strong results, suggesting that these figure types offer more visually consistent and interpretable structures for the model. In contrast, "Other" and hybrid types like "Line Chart, Bar Chart" significantly underperformed, with ROUGE-1 F1 scores as low as 35.14% and 42.86%, respectively. This disparity indicates that composite visualizations or less conventional diagrams introduce ambiguity or complexity that current models struggle to resolve effectively. This aligns with findings by Zhu et al. (2025), who highlight that VLMMs are still not robust when it comes to multi-chart reasoning. Conversely, we observed that other multi-chart figures, such as "Line Chart, Table", or "Line Chart, Scatter Plot" yield acceptable scores (85.71% and 71.43% with ROUGE-1 F1). Overall, these results underscore the importance of figure type in influencing model performance and reveal that chart complexity and visual composition remain critical challenges for VLMMs.

Notably, the model performs almost perfectly on unanswerable questions, indicating that it reliably recognizes when the provided visual information is insufficient to answer the question.

These findings support the broader observation that structured question formats (e.g., yes/no answers) better align with the model's reasoning capabilities, while open or unconstrained queries involving visual reasoning are more challenging. It should also be noted that the proportion of chart types and questions in the training dataset was not balanced. Future improvements may involve training on more varied chart types to improve generalization.

6 Conclusions

In this work, we introduced QwenChart-7B, an instruction-tuned VLLM built on Qwen2.5-VL for the shared task SciVQA. Our approach leverages dynamic CoT prompting and LoRA-

based parameter-efficient fine-tuning. Despite its relatively small size, QwenChart-7B demonstrates state-of-the-art performance on the challenging SciVQA benchmark, outperforming even larger models like QwenChart-72B. This suggests that architecture-specific optimization and well-designed prompts can surpass gains from model scaling alone. However, we also observed limitations in out-of-domain generalization, particularly on the ChartQA benchmark, indicating room for improvement. Future work will explore richer multimodal alignment, broader datasets, and more generalized instruction strategies to address these challenges and further improve performance across diverse chart types and QA formats.

Limitations

Despite the strong performance of QwenChart-7B on SciVQA, several limitations remain. First, the model struggles with generalization when evaluated on out-of-domain benchmarks such as ChartQA. This suggests a sensitivity to dataset-specific features and prompt formulations, potentially limiting its broader applicability without additional fine-tuning. Second, the relatively small amount of fine-tuning data used may not adequately capture the diversity of real-world chart formats and question styles, further constraining generalization in unseen tasks and out-of-domain data. Another limitation concerns the evaluation methodology. While automatic metrics such as ROUGE-1, ROUGE-L, and BERTScore are standard in natural language generation tasks, they are not ideally suited for assessing short, factual responses typical in chart QA. These metrics may fail to penalize near-miss answers or reward semantically correct but lexically mismatched outputs, thus potentially misrepresenting true model performance. We notice that sometimes the result is evaluated as wrong even if it is correct. A human evaluation could solve this issue. Furthermore, the work is limited in providing evaluations with other models or benchmarks.

Acknowledgments

This research was funded by the Bavarian State Ministry for Science and the Arts (StMWK: Bayerische Staatsministerium für Wissenschaft und Kunst - StMWK) as part of the Project "CHIASM" (Changenreiche industrielle Anwendungen für vortrainierte Sprachmodelle) and as part the High Tech Agenda of the Free State of Bavaria.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. 2025. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

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

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. LEAF-QA: Locate, Encode Attend for Figure Question Answering . In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3501–3510, Los Alamitos, CA, USA. IEEE Computer Society.

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

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. *Preprint*, arXiv:2311.16483.

Wei He, Zhiheng Xi, Wanxu Zhao, Xiaoran Fan, Yiwen Ding, Zifei Shan, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Distill visual chart reasoning ability from LLMs to MLLMs.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749, Bangkok, Thailand. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. 2024. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training. *Preprint*, arXiv:2404.14604.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding Data Visualizations via Question Answering . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, Los Alamitos, CA, USA. IEEE Computer Society.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An annotated figure dataset for visual reasoning.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. Acl-fig: A dataset for scientific figure classification. In *Proceedings of the 2023 Workshop on Scientific Document Understanding (SDU)*, volume 3656, pages 1–12. CEUR-WS. Presented at the 2023 Workshop on Scientific Document Understanding (SDU 2023).

Hajime Kato, Mitsuru Nakazawa, Hsuan-Kung Yang, Mark Chen, and Björn Stenger. 2022. Parsing line chart images using linear programming. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2553–2562.

Wonjoong Kim, Sangwu Park, Yeonjun In, Seokwon Han, and Chanyoung Park. 2025. SIMPLOT: Enhancing chart question answering by distilling essentials. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 573–593, Albuquerque, New Mexico. Association for Computational Linguistics.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *Preprint*, arXiv:2308.03349.

Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. 2024b. Synthesize Step-by-Step: Tools, Templates and LLMs as Data Generators for Reasoning-Based Chart VQA . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13613–13623, Los Alamitos, CA, USA. IEEE Computer Society.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023b. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310, Mexico City, Mexico. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. 2025. On the perception bottleneck of vlms for chart understanding. *Preprint*, arXiv:2503.18435.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1916–1924.

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 Xuan 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, Bangkok, Thailand. Association for Computational Linguistics.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Chinmayee Rane, Seshasayee Mahadevan Subramanya, Devi Sandeep Endluri, Jian Wu, and C. Lee Giles. 2021. Chartreader: Automatic parsing of bar-plots. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 318–325.

Yasaman Razeghi, Ishita Dasgupta, Fangyu Liu, Vinay Venkatesh Ramasesh, and Sameer Singh. 2024. Plot twist: Multimodal models don't comprehend simple chart details. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5922–5937, Miami, Florida, USA. Association for Computational Linguistics.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hrituraj Singh and Sumit Shekhar. 2020. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Peifang Wang, Olga Golovneva, Armen Aghajanyan, Xiang Ren, Muhao Chen, Asli Celikyilmaz, and Maryam Fazel-Zarandi. 2023. Domino: A dual-system for multi-step visual language reasoning. *Preprint*, arXiv:2310.02804.

Shulei Wang, Shuai Yang, Wang Lin, Zirun Guo, Sihang Cai, Hai Huang, Ye Wang, Jingyuan Chen, and Tao Jin. 2025. Omni-chart-600K: A comprehensive dataset of chart types for chart understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4051–4069, Albuquerque, New Mexico. Association for Computational Linguistics.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhengzhuo Xu, SiNan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2025a. Chartbench: A benchmark for complex visual reasoning in charts.

Zhengzhuo Xu, Bowen Qu, Yiyan Qi, SiNan Du, Chengjin Xu, Chun Yuan, and Jian Guo. 2025b. Chartmoe: Mixture of diversely aligned expert connector for chart understanding. In *The Thirteenth International Conference on Learning Representations*.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024a. TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1898, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. 2024b. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19228–19252, Miami, Florida, USA. Association for Computational Linguistics.

Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding via visual language pretraining on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1314–1326, Toronto, Canada. Association for Computational Linguistics.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2025. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11341–11359, Albuquerque, New Mexico. Association for Computational Linguistics.

A Appendix

A.1 Example of chart and corresponding prompts

We show here an example of a chart from the SciVQA dataset (Figure 1) and two different prompts (Figure 2), used as described in Section 3.2.

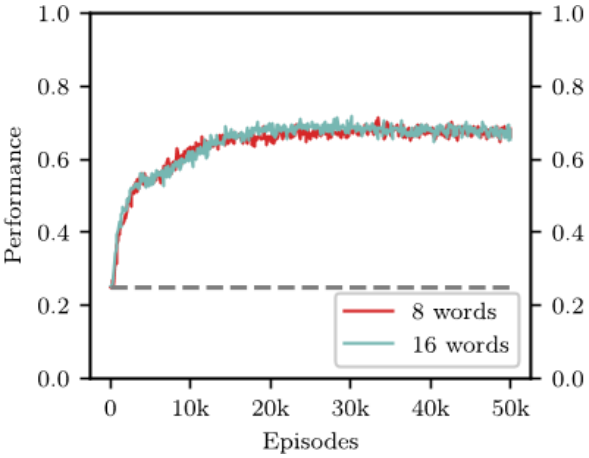

Figure 1: Chart paired with the prompts shown in Figure 2.

```
You are looking at a line chart.
The caption is: 'Figure 5: Performance of the model on images from the
CelebA dataset when the asking-agent has four images, two rounds of
question-answer are performed and with a vocabulary of eight and sixteen
words available. The dashed grey lines represents the baseline performance
where the asking-agent guesses randomly.'.
Question: What is the range of episodes?
[Data-only cue] Focus your response more on numeric or textual values.
Please also consider the caption of the figure to respond to the question.
Respond with a concise, one-word or very short phrase. No full sentences,
no explanations.
If the response is numeric, use digits only and include any units or
suffixes (e.g., %, kg, $).
If the answer cannot be inferred from the figure and caption, please reply
with the sentence: 'It is not possible to answer this question based only
on the provided data.'
---
<thinking> Reasoning (do NOT respond yet)
Step 1 Identify the figure type and its axes / legend.
Step 2 Locate the graphical elements relevant to the question.
Step 3 Extract the key-value information.
Step 4 Read the required values or qualitative trends.
Step 5 Form the short response requested above.
---
Final respond:
<answer>
```

```
Prompt: You are looking at one or more charts or graphs.
While inspecting the visual, pay attention to: color, position, shape,
size, height, direction, and any numeric values on axes, legends, or
labels.
Use the caption only if it clarifies the figure; otherwise rely on the
visual itself.

Answer format:
- Yes/No question -> reply 'Yes' or 'No' only.
- Multiple-choice question -> reply with the capital letter(s) of the
chosen option(s) (e.g. `A` or `A,B`, no spaces).
- Numeric answer -> digits only, include any units or symbols (e.g., %, kg,
$).
- If the answer cannot be inferred -> reply exactly: 'It is not possible to
answer this question based only on the provided data.'
- Please be concise and avoide explenations or reasoning in your final
answer.


Question: What is the range of episodes?

---
<thinking> Reasoning (do NOT respond yet)
1. Identify the chart type, axes, and legend.
2. Locate the graphical elements relevant to the question.
3. Extract the key values or qualitative trends.
4. Integrate helpful details from the caption (if any).
5. If multiple choice, match your finding to the option(s); if yes/no,
decide 'Yes' or 'No'.
6. Produce the concise answer following the formatting rules above.
---
Final respond:
<answer>
```

Figure 2: Examples of prompts for image-question pair. The image is shown in Figure 1.

A.2 Simple Question Prompt vs. Dynamic Prompt

The following example shows the answers generated with Qwen2.5-VL with Simple Question Prompt and with our Dynamic Prompt, as described in Section 3.2.1. The question is taken from the SciVQA dataset.

Question: "Does the accuracy of the red line decrease as the percentage of tokens selected increases?"

Simple Question Prompt Answer: “No, the accuracy of the red line does not decrease as the percentage of tokens selected increases. The red line remains relatively flat and stable across the range of percentages shown in the figure, indicating that the accuracy is consistent regardless of the number of tokens selected”. (Length: 47 words)

Dynamic Prompt Answer: "No".

Gold Answer: "No".