

# CHARTCAP: Mitigating Hallucination of Dense Chart Captioning

## Supplementary Material

Chart Type	Title	Axes	Categories	Bubble	Legends	Labels
Line	✓	✓			✓	✓
Bar	✓	✓	✓		✓	✓
Pie	✓		✓		✓	✓
Histogram	✓	✓			✓	✓
Scatter	✓	✓			✓	✓
Area	✓	✓			✓	✓
Bubble	✓	✓	✓	✓	✓	✓

Table 6. Caption schema specifying the structural elements required for each chart type. For readability, choropleth maps and treemaps are excluded due to their distinct characteristics. Choropleth maps include title, base map, color scale, geographic labels, data classes, and north arrow. Treemaps include title, tiles, hierarchy levels, and color coding.

### A. Type-specific Caption Schema

#### A.1. Caption Schema for Structural Description

The caption schema that defines the structural elements included in each chart type is shown in Table 6.

#### A.2. Caption Schema for Key Insights

The caption schema that specifies the key insights to be included for each chart type can be found in Table 7. In the case of “Retrieve Value”, the task involves reading data points to answer a question, making it inapplicable to the captioning task. Therefore, if the data points had labels, those were extracted; otherwise, the initial, middle, and final data values were extracted in the caption.

### B. Prompt Demonstrations

We present the prompts used in the dataset-generation pipeline and in chart regeneration for the cycle-consistency-based human-verification process and the Visual Consistency Score.

- **Filtering Non-Chart Images:** See Table 8
- **Type Classification and Title Extraction:** See Table 9
- **Retrieving Type-Specific Information:** See Tables 10 and 11
- **Finalizing the Caption:** See Table 12
- **Chart Regeneration:** See Table 13
- **Code Debugging:** See Table 14

### C. Task Allocation Experiment

#### C.1. Experiment Setup

This experiment was designed to optimize the effective utilization of GPT-4o and Claude 3.5 Sonnet in extracting

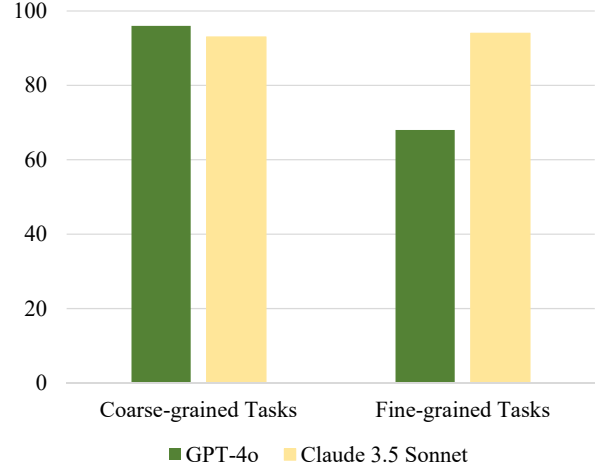


Figure 8. Accuracy of GPT-4o and Claude 3.5 Sonnet on coarse-grained tasks and fine-grained tasks.

accurate information based on the caption schema. The tasks defined in Tables 6 and 7 were categorized into fine-grained and coarse-grained tasks, and the performance of each model was evaluated accordingly.

Coarse-grained tasks require broad attention across an image, such as understanding of overall data trends or comparisons between data series: “Type Classification”, “Title”, “Category”, “Bubble”, “Legend”, “Label”, “Make Comparisons”, “Find Correlations / Trends”, and “Characterize Distribution”. In contrast, fine-grained tasks require more localized attention, such as extracting precise numerical values or reading specific data points: “Axes”, “Retrieve Value”, “Find Extremum”, “Determine Range”, “Find Clusters”, and “Find Anomalies”. The experiment was conducted on 100 randomly sampled data, and the performance of both models was manually evaluated for each task category. The prompts used in this experiment can be found in Tables 10 and 11.

#### C.2. Results

The experimental results are shown in Figure 8. For coarse-grained tasks, GPT-4o achieved an accuracy of 96%, while Claude 3.5 Sonnet achieved 93%, with GPT-4o demonstrating a slight advantage. The primary sources of error for Claude 3.5 Sonnet were in type classification and title generation, where the model tended to introduce hallucinated information by attempting to predict the overall theme of a chart. Based on these findings, GPT-4o was selected for coarse-grained tasks in our pipeline.

Chart Type	Retrieve Value	Find Extremum	Make Comparison	Determine Range	Find Correlations / Trend	Characterize Distribution	Find Clusters	Find Anomalies
Line	✓	✓	✓	✓	✓			
Bar	✓	✓	✓	✓				
Pie	✓	✓	✓					
Histogram	✓	✓	✓			✓		
Scatter	✓	✓	✓	✓	✓	✓	✓	✓
Area	✓	✓	✓	✓	✓			
Bubble	✓	✓	✓	✓	✓	✓	✓	✓
Choropleth Map	✓	✓	✓					
Treemap	✓	✓	✓					

Table 7. Caption schema specifying the key insights required for each chart type.

#### Filtering non-chart images.

Please determine whether the image contains a single, data-driven chart only. A data-driven chart is a visual representation directly based on numerical data. Note that an inset chart (a smaller chart embedded within a larger chart) is not considered a multi-chart.

- If the image consists exclusively of a single data-driven chart (with no additional visuals, such as natural images, illustrations, conceptual diagrams, or schematics) and does not contain multiple subplots, respond with:

Single-Chart: yes

- If the image contains any non-data-driven elements (e.g., natural images, illustrations, conceptual diagrams, schematics) or features multiple charts/subplots, respond with:

Single-Chart: no

Follow the exact response format:

Single-Chart:

Table 8. Prompt used for filtering non-chart images.

For fine-grained tasks, GPT-4o achieved an accuracy of 68%, whereas Claude 3.5 Sonnet significantly outperformed it with an accuracy of 94%. The primary weakness of GPT-4o was its difficulty in accurately reading data coordinates, a critical skill for tasks such as “Retrieve Value” and “Find Extremum”, resulting in incorrect values for maxima, minima, and other key numerical indicators. Because of this fundamental limitation, Claude 3.5 Sonnet was selected for fine-grained tasks in our pipeline.

## D. Validation of Cycle Consistency-based Human Verification Process

### D.1. Quantitative Evaluation

To quantify the effectiveness of our verification process, we benchmark it against direct chart–caption comparison in terms of both accuracy and efficiency.

- **Accuracy:** On 100 randomly sampled pairs from CHARTCAP, our process achieved an F1 score of 94.7%, with a recall of 90.0% and precision of 100.0%, ensuring that no incorrect captions were falsely validated.
- **Efficiency:** Direct chart-caption comparison required approximately 145 seconds per sample, whereas our process took only 6 seconds per sample, achieving a 24× speedup.

### D.2. Qualitative Analysis

To qualitatively assess the logical validity of our method, we define one axiom and one premise:

- **Axiom:** A correct verification process should not classify incorrect data as correct.
- **Premise:** Human inspectors make no mistakes during verification.

Figure 9 illustrates the four main scenarios that arise when regenerating charts from captions:

1. **Scenario A.** If the caption is incorrect, it produces faulty code leading to a mismatched image, which is identified and removed.
2. **Scenario B.** If the caption lacks sufficient detail, an over-simplified chart is generated and subsequently filtered out.
3. **Scenario C.** Even if the caption is accurate, errors in code generation or execution can result in a failed reconstruction, leading the sample to be excluded.
4. **Scenario D.** Only when the caption is both accurate and informative, and the chart regenerates without errors, does the sample pass verification.

This process ensures that only captions containing both correct and adequately detailed information are retained.

In summary, our process is designed to guarantee both the correctness and depth of information in CHARTCAP,

---

**Type Classification and Title Extraction.**

---

**[System]**

You are an expert in data visualization and chart interpretation. Your task is to provide accurate analysis of charts such as identifying and classifying the chart.

**[User]**

Please analyze the image to classify the chart type(s) and extract the main title according to the instructions below.

- Identify the chart type(s) from the following list: [line, bar, pie, histogram, scatter, area, bubble, choropleth map, treemap].

- If it belongs to multiple chart types, list them separated by commas (e.g., "bar", "line"). If it does not match any listed chart types or is a 3D visualization, respond with:

Type: none

Title: not specified

- If the image contains one or more valid chart types, extract the main title of the chart. If the title is not visible or unclear, respond with 'not specified'.

Follow the exact response format:

Type: <list of chart\_type(s) or 'none'>

Title: <chart\_title or 'not specified'>

---

Table 9. Prompt used for type classification and title extraction.

---

**Extracting Type-specific Information (Coarse-grained)**

---

**[System]**

You are an expert in data visualization and chart interpretation. Your task is to provide accurate analysis of charts such as identifying the components, key trends, and insights, without making any guesses.

**[User]**

Identify and describe the components of the given line chart. Only explain the components if they exist; otherwise, respond with not specified. Do not guess or include information not visible in the image, except for approximations in axes ranges, retrieving value of data points, and determining data point ranges.

If the chart is multi-series, grouped, or includes an inset chart, compute and report information for each data series or category separately.

\* Type: Provide the type or types of the chart from line chart, bar chart, pie chart, histogram, scatter plot, area chart, bubble chart, choropleth map, and treemap.

\* Legends: Identify any legends or keys that globally explain symbols, colors, or data series.

\* Labels: Identify specific labels that annotate or describe individual elements, such as data points, bars, or segments of a chart. Exclude axis labels and legends.

\* Data Comparison: Highlight specific similarities or differences between data points or categories in the chart. Focus on relative comparisons rather than extracting or explaining precise values. Avoid analyzing overall trends.

\* Data Correlations/Trends: Analyze patterns or relationships between variables, noting any trends.

Only respond with the analyzed results, avoiding any additional statements or extraneous text.

Follow the exact response format:

<attribute 1>: <analysis result>

<attribute 2>: <analysis result>

...

---

Table 10. Prompt used for extracting coarse-grained, type-specific information from line charts.

while substantially boosting the efficiency of large-scale verification. By combining logical verification with cycle consistency-based human verification process, we enable efficient quality control of CHARTCAP and mitigate the burden of manual inspection.

## E. Validation of VCS with Human Evaluation

To validate the effectiveness of the Visual Consistency Score (VCS), we performed head-to-head human evaluations. For every comparison between two baselines, we randomly sampled 100 chart-caption pairs from the three test sets-CHARTCAP, VisText, and Chart-to-Text. Following the protocol in Appendix J, human annotators compared caption pairs and selected the better one with respect to informativeness, accuracy, and fewer hallucinations. We then computed the agreement rate as the proportion of comparisons in which the caption preferred by human annotators also received a higher metric score.

As shown in Table 15, VCS achieved the highest agree-

ment rates across all three criteria, followed by OCRScore. These results indicate that both metrics reliably capture key aspects of caption quality as perceived by humans, validating their effectiveness as automatic metrics.

## F. LLM Fidelity in Caption-to-Code Translation

We observe that caption distortions during the first phase of VCS evaluation-LLM caption-to-code translation-are rare in practice. Interestingly, the distortion rate increases slightly when a caption is less informative. To investigate this, we analyzed 100 caption-code pairs each from Phi3.5-Vision-4B<sub>CHARTCAP</sub> (with the highest VCS) and Phi3.5-Vision-4B<sub>Original</sub> (with the lowest VCS). We examined (1) whether the elements defined in the caption were correctly preserved, and (2) whether any content not described in the caption appeared in the code.

As a result, Phi3.5-Vision-4B<sub>CHARTCAP</sub> achieved a caption-to-code accuracy of 99%, with the remaining 1%

---

**Extracting Type-specific Information (Fine-grained).**

---

**[System]**

You are an expert in data visualization and chart interpretation. Your task is to provide accurate analysis of charts such as identifying the components, retrieving data points, statistics, key trends, insights, without any guesses.

**[User]**

Identify and describe the components of the given line chart. Only explain the components if they exist; otherwise, respond with not specified. Do not guess or include information not visible in the image, except for approximations in axes ranges, retrieving value of data points, and determining data point ranges.

If the chart is multi-series, grouped, or includes an inset chart, compute and report information for each data series or category separately.

\* Axes: Describe the axes, including titles, units, scales, and ranges. If categories are involved in the axes, list their names as well.

\* Retrieve Value: Retrieve the coordinates of the initial, middle, and end data points. Additionally, if specific numbers or values are labeled for any data points, also provide the coordinates of those points as well.

\* Find Extremum: Find the coordinate of the minimum and maximum data points for each data series.

\* Determine Range: Specify the range (span) of the dependent (response) variable's values from the data points, not the range of the axis.

Only respond with the analyzed results, avoiding any additional statements or extraneous text.

Follow the exact response format:

<attribute 1>: <analysis result>

<attribute 2>: <analysis result>

...

---

Table 11. Prompt used for extracting fine-grained, type-specific information from line charts.

---

**Finalizing the Caption.**

---

**[System]**

You are an expert in converting provided bullet points into continuous sentences without omitting or adding any information.

**[User]**

Generate a natural language caption for the chart based strictly on the provided information. Ensure the caption includes all the details given in the input without omitting anything or adding new information beyond what is explicitly stated.

Explicitly mention that information is not provided if it is stated as not provided in the chart information.

[Chart Information]

{chart\_info}

Respond only with the generated caption, including all the information provided.

Caption:

---

Table 12. Prompt used for finalizing the caption.

due to the omission of the axis title. For Phi3.5-Vision-4B<sub>Original</sub>, the accuracy was 96%, and in the remaining 4% of cases, the LLM hallucinated placeholder or arbitrary data values to fill in the missing details of oversimplified captions. These results show that (1) translation errors are infrequent, and (2) lower information density in captions tends to increase the likelihood of LLM’s caption-to-code distortions, ultimately resulting in lower VCS.

## G. Sensitivity of VCS to Structural Errors

Although SigLIP’s attention mechanism on charts is not fully interpretable, we find that the model is reasonably sensitive to structural elements. We analyze 100 captions collected from baseline models and found three major error types: (1) misidentification of maxima/minima (20%), (2) axis hallucinations (13%), and (3) omission of data series (7%). After manually correcting these errors, VCS increased by 1.3%, 6.1%, and 4.7%, respectively, indicating that VCS is capable of detecting such structural issues.

## H. Additional Baselines

We additionally fine-tuned Qwen2.5-VL-7B on CHARTCAP and evaluated it on the VisText and Chart-to-Text benchmarks. We also evaluated Phi3.5-Vision-4B<sub>ChartSumm</sub> on the same benchmarks.

As shown in Table 16, Qwen2.5-VL-7B<sub>CHARTCAP</sub> consistently outperforms its base model, whereas Phi3.5-Vision-4B<sub>ChartSumm</sub> underperforms Phi3.5-Vision-4B<sub>CHARTCAP</sub> and even degrades performance relative to its own base model.

We conduct additional human evaluation under the protocol in Appendix J directly comparing Phi3.5-Vision-4B<sub>CHARTCAP</sub> and Phi3.5-Vision-4B<sub>ChartSumm</sub> on VisText test set. As shown in Figure 10, Phi3.5-Vision-4B<sub>ChartSumm</sub> received fewer preferences than Phi3.5-Vision-4B<sub>CHARTCAP</sub> across all three evaluation aspects. In summary, both automatic and human evaluations indicate that (1) CHARTCAP consistently improves the captioning performance of state-of-the-art models, and (2) CHARTCAP is a more effective training dataset than ChartSumm.

---

**Regenerating a Chart from a Caption.**

---

[System]

You are an expert in Python and the Matplotlib library. Your task is to generate a complete Python script that precisely reflects every detail in the given chart description, without making any guesses.

[User]

Generate accurate Python code using Matplotlib library strictly based on the given description about a chart.

If the description lacks details about required chart components or data points, omit them from the code instead of making assumptions, but ensure that every detail in the description is included.

Instead of using numpy’s sin, cos, or exp function, manually define data points to represent the chart if needed.

Labels are elements that display and specify data points in the chart. They are different from axis labels (titles).

[Description]

“{caption}”

Respond only the generated code.

Code:

---

Table 13. Prompt used for regenerating a chart from a caption.

---

**Debugging Erroneous Code.**

---

[System]

You are an expert in Python and the Matplotlib library. Your task is to fix the code based on the provided error message.

[User]

[Erroneous Code]

“{code}”

[Error Message]

“{error\_message}”

Analyze the provided error message and fix the code accordingly. Make only the necessary changes to resolve the error while keeping all correctly functioning attributes unchanged. Return only the corrected code without any explanations or additional output.

Corrected Code:

---

Table 14. Prompt used for debugging erroneous code.

Metric	Informativeness	Accuracy	Fewer hallucinations
SacreBLEU	60.50	59.50	59.83
ROUGE	55.34	55.67	57.00
METEOR	70.34	68.34	69.34
BERTScore	68.67	67.00	68.34
VCS (so400m)	<b>79.33</b>	<b>77.00</b>	<b>77.33</b>
OCRScore	<u>76.00</u>	<u>75.00</u>	<u>74.00</u>

Table 15. Agreement between automated metrics and human judgments (%). Higher is better.

Model	VisText		Chart-to-Text	
	VCS	OCRScore	VCS	OCRScore
Qwen2.5-VL-7B	0.9044	0.3197	0.7739	0.1622
Qwen2.5-VL-7B <sub>CHARTCAP</sub>	0.9328	0.3436	<b>0.8084</b>	<b>0.1817</b>
Phi3.5-Vision-4B	0.8814	0.3414	0.7490	0.1786
Phi3.5-Vision-4B <sub>CHARTCAP</sub>	<b>0.9382</b>	<b>0.3826</b>	0.8075	0.1789
Phi3.5-Vision-4B <sub>ChartSumm</sub>	0.8677	0.1414	0.7281	0.0789

Table 16. Results of VCS and OCRScore on VisText and Chart-to-Text. VCS is computed using SigLIP2-So400M-512.

## I. Training Hyperparameters

Training was conducted using 6 RTX A6000 GPUs. The total training time was approximately 60 hours for InternVL2.5-8B<sub>CHARTCAP</sub>, 30 hours for Phi3.5-Vision-

4B<sub>CHARTCAP</sub>, 12 hours for Phi3.5-Vision-4B<sub>Original</sub>, 2 hours for Phi3.5-Vision-4B<sub>ChartSumm</sub>, and 50 hours for Qwen2.5-VL-7B<sub>CHARTCAP</sub>. The training hyperparameters used for these models are summarized in Table 17.

## J. Details of Human Evaluation

**Sampling and Setup.** For each comparison, we randomly sampled 100 chart–caption pairs from two competing baselines. Crowd workers were shown the two pairs in a random left-right order and asked to choose the better pair under three criteria:

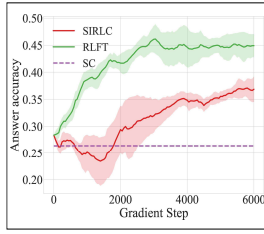
1. **Informativeness** – Does the caption adequately describe the chart’s structure and key insights (highlighted in green and blue in Fig 1)?
2. **Accuracy** – How faithfully does the caption reflect the chart’s structure and key insights?
3. **Fewer Hallucinations** – Does the caption avoid information that cannot be inferred from the chart (highlighted in red in Fig 1)?

For the dataset-level study in § 3.5 (CHARTCAP vs ChartSumm), we added a fourth question—overall preference—to capture holistic quality while accounting for chart complexity.

**User Interfaces.** Figure 11 illustrates the user interface

## Scenario A.

### Chart Image



### Caption

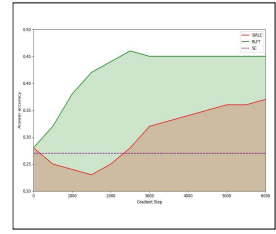
The line and **area chart** contains three legends: "SIRC" represented by a red line and red shaded area, "RLFT" shown in green line and green shaded area, and "SC" displayed as a purple dashed line. No specific labels are visible on individual data points or areas. The data comparison illustrates that "RLFT" consistently achieves higher answer accuracy than both "SIRC" and "SC" after the initial...

### Python Code

```
-- Create figure and axis
fig, ax = plt.subplots(figsize=(10, 6))

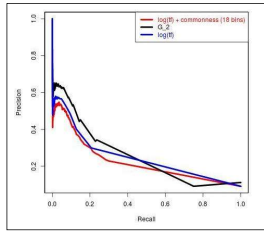
# Plot lines and areas
ax.fill_between(x, sirc_y, alpha=0.2, color='red')
ax.plot(x, sirc_y, color='red', label='SIRC')
ax.fill_between(x, rlft_y, alpha=0.2, color='green')
ax.plot(x, rlft_y, color='green', label='RLFT')
ax.plot(x, sc_y, color='purple' ...
```

### Reconstructed Chart Image



## Scenario B.

### Chart Image



### Caption

The line chart presents three series labeled as "log(tf) + commonness (18 bins)" in red, "G\_2" in black, and "log(tf)" in blue, illustrating their respective precision values across a recall range from 0.0 to 1.0. The data reveals that "G\_2" generally ... All three series show a **trend of decreasing precision as recall increases**, albeit at different rates...

### Python Code

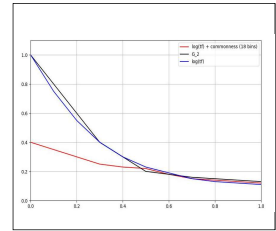
```
import matplotlib.pyplot as plt

# Define data points
recall = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]

# Precision values for each series
log_tf_commonness = [0.4, 0.35, 0.3, 0.25, 0.23, 0.22, 0.18, 0.15, 0.14, 0.13, 0.12]
g2 = [1.0, 0.8, 0.6, 0.4, 0.3, 0.2, 0.18, 0.16, 0.15, 0.14, 0.13]
log_tf = [1.0, 0.75, 0.55, 0.4, 0.3, 0.23, 0.19, 0.15, 0.13, 0.12, 0.11]

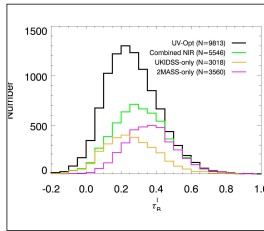
# Create the line plot ...
```

### Reconstructed Chart Image



## Scenario C.

### Chart Image



### Caption

The chart, which is a histogram, displays the distributions of four data series represented by black for UV-Opt (N=9813), green for Combined NIR (N=5546), orange for UKIDSS-only (N=3018), and magenta for 2MASS-only (N=3560). Notably, the black distribution (UV-Opt) generally has higher counts for most bins compared to the others, with the green distribution (Combined NIR) ...

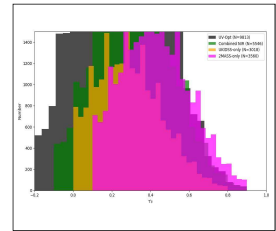
### Python Code

```
def normal_like(x, mu, sigma, amplitude):
    return amplitude * np.exp(-(x - mu)**2 / (2 * sigma**2))

y_uv = normal_like(x_uv, 0.25, 0.2, 1350)
y_nir = normal_like(x_nir, 0.3, 0.2, 700)
y_ukidss = normal_like(x_ukidss, 0.25, 0.2, 400)
y_2mass = normal_like(x_2mass, 0.4, 0.2, 500)

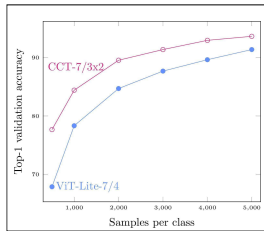
plt.figure(figsize=(10, 6))
plt.hist(x_uv, weights=y_uv, bins=30, color='black', alpha=0.7, label=f'UV-Opt ...
```

### Reconstructed Chart Image



## Scenario D.

### Chart Image



### Caption

The line chart compares the top-1 validation accuracy of two models, with the pink line labeled "CCT-7/3x2" and the blue line labeled "ViT-Lite-7/4." The pink line consistently achieves higher top-1 validation accuracy values than the blue line across all sample counts. Both lines indicate a positive correlation between the number of samples per class, ranging from ...

### Python Code

```
import matplotlib.pyplot as plt

# Data points
samples = [500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000]
cct_accuracy = [77.5, 85.0, 88.5, 90.0, 91.0, 92.0, 92.5, 93.0, 93.2, 93.5]
vit_accuracy = [67.5, 77.0, 82.0, 84.5, 86.0, 87.0, 88.5, 89.5, 90.5, 91.5]

# Create the plot
plt.figure(figsize=(10, 6))

# Plot lines
plt.plot(samples, cct_accuracy, color='pink', label='CCT-7/3x2', linewidth=2) ...
```

### Reconstructed Chart Image

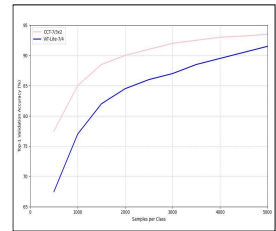


Figure 9. Examples of the four main scenarios that arise during the cycle consistency-based human verification process. In Scenario A, the caption incorrectly describes a *line chart with a shaded area* as an *area chart*. Consequently, the generated code reflects this incorrect information, leading to the reconstruction of a chart that does not match the original. In Scenario B, the caption oversimplifies the data trend by merely describing it as *decreasing*. As a result, the reconstructed chart follows a simple downward trend, failing to capture the original complexity. In Scenario C, an error occurs during the code generation process, leading to the creation of an incorrect chart. Such coding errors were primarily observed when using NumPy's nonlinear functions. Scenario D shows that both the caption and the generated code must be accurate for the reconstructed chart to correctly match the original chart, demonstrating the necessity of precise and informative captions.

Hyperparameter	InternVL2.5-8B <sub>CHARTCAP</sub>	Phi3.5-vision-4B <sub>CHARTCAP</sub>	Phi3.5-vision-4B <sub>Original</sub>	Phi3.5-vision-4B <sub>ChartSumm</sub>	Qwen2.5-VL-7B <sub>CHARTCAP</sub>
Epochs	2	1	1	1	2
Batch size	12	192	192	192	12
Learning rate	2e-5	2e-5	2e-5	2e-5	2e-5
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Warmup ratio	0.05	0.05	0.1	0.05	0.1
Scheduler	cosine	constant	cosine	constant	cosine
LoRA rank	32	32	32	32	32
LoRA alpha	64	32	32	32	64
LoRA dropout	0.0	0.05	0.05	0.05	0.05

Table 17. Training hyperparameters for InternVL2.5-8B, Phi3.5-vision-4B, and Qwen2.5-VL-7B.

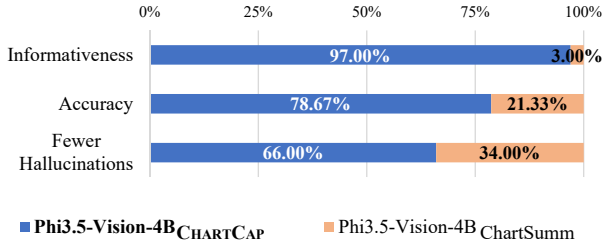


Figure 10. Human evaluation results comparing Phi3.5-Vision-4B<sub>CHARTCAP</sub> and Phi3.5-Vision-4B<sub>ChartSumm</sub> on the VisText test set.

used for model comparisons, while Fig 12 shows the interface used for dataset comparisons.

**Worker Qualification and Quality Control.** To ensure reliable judgments, we administered a qualification test to assess workers’ understanding of the task. Only those who passed were allowed to participate in the main evaluation. During the evaluation, workers were also required to provide brief justification for their choices, discouraging random or inattentive responses.

**Platform and Demographics.** Evaluations were conducted on Amazon Mechanical Turk, with participation restricted to English-speaking countries (Australia, Canada, New Zealand, the United States, and the United Kingdom).

**Inter-annotator agreement.** We measured the inter-annotator agreement using Gwet’s AC1.

- § 3.5: 0.84 (informativeness), 0.80 (accuracy), 0.23 (fewer hallucinations), 0.80 (overall preference).
- § 4.1: 0.47, 0.27, 0.27.
- § 4.2: 0.70, 0.52, 0.22.
- § E: 0.84, 0.71, 0.71.
- § H: 0.91, 0.37, 0.24

**Compensation.** Workers were compensated \$0.50 per HIT, corresponding to approximately \$15 per hour, which exceeds the U.S. federal minimum hourly wage.



## Instructions

Welcome! In this main HIT, you will evaluate the informativeness and accuracy of chart captions based on a given chart image. Please read the instructions carefully before submitting your responses.

### Steps:

1. Examine the provided chart image.
2. Read the given chart captions carefully.
3. Answer the questions comparing the captions.

## Your Task



**Previewing Answers Submitted by Workers**  
This message is only visible to you and will not be shown to Workers.  
You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

**Compare the Captions:**

**Caption #1**

This chart shows the operating revenue of NextEra Energy from fiscal year 2006 to fiscal year 2019, measured in million U.S. dollars. The data is presented as an area chart with a blue fill. The y-axis ranges from 0 to approximately 20,000 million (or \$20 billion) dollars, while the x-axis spans 13 years. The revenue trend shows some fluctuations over the years but generally maintains a level between \$15-20 billion, with a notable increase toward the end of the period in 2019. The visualization suggests NextEra Energy has maintained relatively stable revenue over most of this period, with recent growth in the later years of the decade shown.

**Caption #2**

The area chart titled "Operating revenue of NextEra Energy from FY 2006 of FY 2019 (in million U.S. dollars)" illustrates the operating revenue of NextEra Energy over the years 2006 to 2018. The chart shows that the operating revenue was approximately 16,000 million U.S. dollars in 2006, decreased to around 14,000 million U.S. dollars in 2012, and then increased to about 19,000 million U.S. dollars by 2018. The minimum operating revenue recorded was approximately 14,000 million U.S. dollars in 2012, while the maximum was around 19,000 million U.S. dollars in 2018, resulting in a range of approximately 5,000 million U.S. dollars. The X-axis represents the years from 2006 to 2018, and the Y-axis indicates the operating revenue in million U.S. dollars, ranging from 0 to 20,000. Legends and labels are not specified in the chart information.

**Question #1. Which caption has more information?**

☐ Definitely Caption #1 ☐ Slightly Caption #1 ☐ Slightly Caption #2 ☐ Definitely Caption #2

**Question #2. Which caption is more accurate?**

☐ Definitely Caption #1 ☐ Slightly Caption #1 ☐ Slightly Caption #2 ☐ Definitely Caption #2

**Question #3. Which caption contains fewer hallucinations?**

☐ Definitely Caption #1 ☐ Slightly Caption #1 ☐ Slightly Caption #2 ☐ Definitely Caption #2

**Explain your choices (required):**

---

(Optional) Any feedback or issues?

---

**Submit**

Figure 11. User interface for human evaluation comparing captions from different models on informativeness, accuracy, and fewer hallucinations.



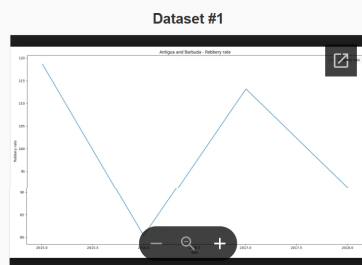
### Instructions

Welcome! In this task, you will compare two different datasets, each containing an chart image and its caption. Please read the instructions carefully before submitting your responses.

**Steps:**

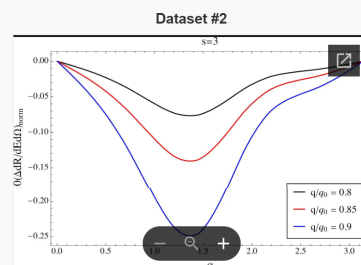
1. Examine both chart-caption pairs carefully.
2. Compare the quality, informativeness, and accuracy of each dataset.
3. Answer the questions comparing the two datasets.

## Your Task



**Caption #1**

In 2018 , robbery rate for Antigua and Barbuda was 91.4 cases per 100,000 population . Though Antigua and Barbuda robbery rate fluctuated substantially in recent years , it tended to decrease through 2015 - 2018 period ending at 91.4 cases per 100,000 population in 2018 .



**Caption #2**

The chart, which is a line chart with legends explaining the colors of the lines indicating three data series (black for  $q/q_0 = 0.8$ , red for  $q/q_0 = 0.85$ , and blue for  $q/q_0 = 0.9$ ), shows that for the same value of alpha, the black line consistently has the highest values, followed by the red line, while the blue line has the lowest. All three lines reveal a similar trend, exhibiting a downward curve that reaches a minimum near alpha = -1.5, followed by an upward rise, with consistent spacing between the lines across alpha values. The x-axis (alpha) ranges from 0.0 to 3.0, and the y-axis ( $10(\Delta\text{d}/d \cdot \text{Rf}/d \cdot \text{Omega}/n)\%$ ) ranges approximately from -0.25 to 0.00, with the title Omega=3. The data points for series  $q/q_0 = 0.8$  include an initial value at (0.0, 0.0), a middle value at (1.5, -0.075), and an endpoint at (3.0, 0.0), with a minimum at (1.5, -0.075) and maximum values at (0.0, 0.0) and (3.0, 0.0), resulting in a range of 0.075. For series  $q/q_0 = 0.85$ , the initial value is (0.0, 0.0), the middle value is (1.5, -0.14), and the endpoint is (3.0, 0.0), with a minimum at (1.5, -0.14) and maximum values again at (0.0, 0.0) and (3.0, 0.0), leading to a range of 0.14. Lastly, for series  $q/q_0 = 0.9$ , the initial value is (0.0, 0.0), the middle value is (1.5, -0.25), and the endpoint is (3.0, 0.0), with a minimum at (1.5, -0.25) and maximum values at (0.0, 0.0) and (3.0, 0.0), resulting in a range of 0.25. The chart does not specify a title or labels.

### Compare the Datasets:

Question #1. Which dataset's caption has more information?

- Definitely Dataset #1   ○ Slightly Dataset #1   ○ Slightly Dataset #2   ○ Definitely Dataset #2

Question #2. Which dataset's caption is more accurate relative to its image?

- Definitely Dataset #1   ○ Slightly Dataset #1   ○ Slightly Dataset #2   ○ Definitely Dataset #2

Question #3. Which dataset's caption contains fewer hallucinations?

- Definitely Dataset #1   ○ Slightly Dataset #1   ○ Slightly Dataset #2   ○ Definitely Dataset #2

Question #4. Overall, which dataset provides a better image-caption pair?

- Definitely Dataset #1 ○ Slightly Dataset #1 ○ Slightly Dataset #2 ○ Definitely Dataset #2

**Explain your choices (required):**

(Optional) Any feedback or issues?

Submit

Figure 12. User interface for human evaluation comparing datasets (CHARTCAP vs. ChartSumm) on informativeness, accuracy, fewer hallucinations, and overall preference.