Do LVLMs Understand Charts?
Analyzing and Correcting Factual Errors in Chart Captioning

Anonymous ACL submission

Abstract

Advances in large vision-language models (LVLMs) have led to significant progress in generating natural language descriptions for visual contents. These powerful models are known for producing texts that are factually inconsistent with the visual input. While some efforts mitigate such inconsistencies in natural image captioning, the factuality of generated captions for structured visuals, such as charts, has not received as much scrutiny. This work introduces a comprehensive typology of factual errors in generated chart captions. A large-scale human annotation effort provides insight into the error patterns in captions generated by various models, ultimately forming the foundation of a dataset, CHOCOLATE. Our analysis reveals that even advanced models like GPT-4V frequently produce captions laced with factual inaccuracies. To combat this, we establish the task of Chart Caption Factual Error Correction and introduce CHARTVE, a visual entailment model that outperforms current LVLMs in evaluating caption factuality. Furthermore, we propose C2TFEC, an interpretable two-stage framework that excels at correcting factual errors. This work inaugurates a new domain in factual error correction for chart captions, presenting a novel evaluation metric, and demonstrating an effective approach to ensuring the factuality of generated chart captions.

1 Introduction

Large vision-language models (LVLMs) have recently shown impressive capabilities in generating natural language descriptions of visual content like images, videos and charts (OpenAI, 2023b; Google, 2023a; Liu et al., 2023c; Wang et al., 2023). Chart captioning is particularly important for data analysts, business analysts, and journalists who rely on accurate chart interpretations for decision-making and reporting. However, no prior work has studied the *factuality*[1] of the generated captions. Given that factuality is vital for credibility in applications of chart captioning in news articles (Liu et al., 2021), educational resources (Fu et al., 2022), and social media (Monteiro et al., 2017), examining the truthfulness of generated captions is a critical concern.

To understand the factual errors in chart captioning models, we introduce a typology of factual errors for the chart domain. Using this scheme, we conduct a large-scale human annotation study to analyze the distributions of various error types, such as Value Error and Label Error, in captions from various models, from task-specific fine-tuned models to LVLMs (see Table 1). The annotated samples are then categorized into three splits, LVLM (Large-vision Language Models), LLM (Large Language Models), and FT (Fine-tuned Vision-language Models), based on the architecture and the scale of the underlying models, and form a dataset which we named CHOCOLATE. With this dataset collected, we aim to answer three main research questions. First, **are state-of-the-art chart captioning models able to produce factual captions? We find the answer is no** (§2). Specifically, 82.06% of the generated captions are non-factual (see Table 2). Even state-of-the-art LVLMs like GPT-4V (OpenAI, 2023b) produce a great portion of errors in its generated captions (see Figure 1).

The prevalence of factual inconsistencies observed in the generated captions by various models underscores the urgent need to mitigate the factual errors of such models. Hence, we introduce a new task, *Chart Caption Factual Error Correction* (§3), which presents a novel challenge of rectifying factual inaccuracies in chart captions generated by LVLMs. A pertinent question that arises from this task is: **how to automatically evaluate the factual consistency between charts and captions?** To

[1]Factuality is also known as the *faithfulness* or *factual consistency* between inputs and outputs

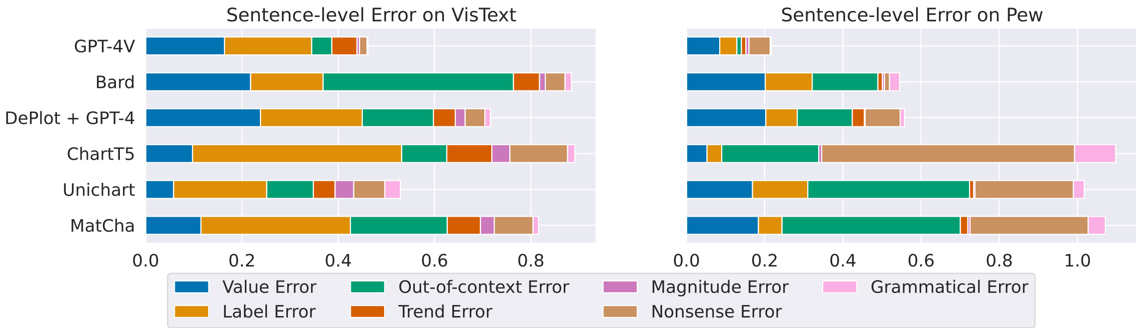


Figure 1: Error distribution for different models on VisText and Pew. The error rates are computed per sentence. An error rate of 0.4 indicates that 40% of the sentences in the generated captions contain such an error. Note that a single caption may contain multiple types of errors; hence, the maximum value for a stacked bar is greater than 1.0. We show that even the most advanced LVLM, GPT-4V, generates captions with a high rate of factual error.

tackle this question, we present CHARTVE, novel visual entailment approach to assess the factual consistency of chart captions. This model is trained by repurposing existing resources from chart summarization and chart question answering. Results show that CHARTVE performs competitively with proprietary LVLMs and outperforms the most advanced open-source LVLM, despite being 64 times less in size.

Now that we have set up the task, we turn to the challenge of **how to effectively correct factual errors in chart captions?** We propose C2TFEC (§4), an interpretable two-step framework that decomposes visual reasoning into image-to-structure rendering and text-based reasoning. C2TFEC first transforms the input chart into a structured data table representation. Grounded in this extracted tabular data, the second component then identifies and fixes any factual inconsistencies in the generated caption through an interpretable reasoning process. Our experiments demonstrate that this explicit decomposition enables more reliable factuality corrections compared to end-to-end approaches. The intermediate symbolic representation acts as an effective bridge between charts and captions, enabling C2TFEC to significantly outperform competitive baselines including GPT-4V (§6).

In summary, our contributions are as follows:

- We present the first analysis of factual errors in captions produced by models of various scales using a novel error typology, which results in the CHOCOLATE dataset.
- We introduce the Chart Caption Factual Error Correction task that challenges models to correct factual errors in generated chart captions.
- We present CHARTVE, a reference-free evaluation metric based on visual entailment that correlates better with human judges than LVLMs.
- We propose C2TFEC, an interpretable two-stage error correction framework that performs better than all existing LVLMs.

2 Analyzing Factual Errors

To understand the capabilities of existing models in summarizing key information from charts, we conduct a large-scale analysis on six most advanced chart captioning models on the VisText (Tang et al., 2023a) and Pew (Kantharaj et al., 2022) datasets. To facilitate this process, we introduce an error typology, as illustrated in §2.1. Upon gathering human annotations, we present a detailed analysis of different captioning models (§2.2) and discuss the quality of the collected data (§2.3).

2.1 Error Typology

To understand the frequency of various types of errors made by chart captioning systems, we define a typology of errors as detailed below and demonstrate examples in Table 1.

Value Error A quantitative data value from the chart is incorrectly stated in the caption. This includes numbers representing values on axes, percentages, or other numerical data points.

Label Error A non-numerical label, category, or text element from the chart is incorrectly referenced in the caption. This includes labels on axes, legend items, categorical variables, etc.

Trend Error The overall direction of change over time or comparison between groups is incorrectly described in the caption, such as stating an increasing trend when it is actually decreasing.

Magnitude Error The degree or amount of difference described for a trend is unfaithful to the chart, such as stating an increase "sharp" when the chart shows it is actually "smooth".

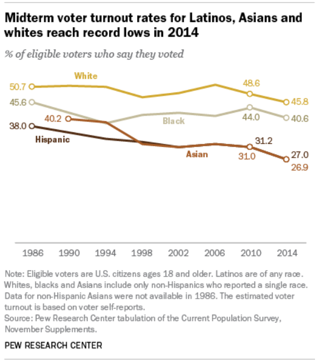


Chart	**Category**	**Example Caption**
(chart above)	Value Error	Asians have a turnout rate of *20.4%* in 1990.
	Label Error	*Asians* have the highest turnout rates across the years.
	Trend Error	From 1986-2014, the turnout rates are *increasing* overall.
	Magnitude Error	From 1986-2014, the turnout rates are *sharply* decreasing overall.
	Out-of-context Error	*Vietnamese* have the highest turnout rates among Asians.
	Nonsense Error	From 1986-2014, *#?sep #sep #sep #sep*.
	Grammatical Error	The turnout rates are *decrease* overall.

Table 1: Typology of errors illustrated with an example chart.

Out-of-context Error Concepts, variables, or any information introduced in the caption that does not exist at all in the content of the chart. The caption contains factual statements not grounded in the actual chart contents.

Nonsense Error The caption contains incomplete sentences, disconnected phrases that do not connect logically, or sequences of words that simply do not make coherent sense.

Grammatical Error There are grammatical mistakes in the structure or syntax of the caption.[2]

2.2 Captioning Model Analysis

We consider various types of models. First, ChartT5 (Zhou et al., 2023), MatCha (Liu et al., 2023b), and UniChart (Masry et al., 2023) are the most advanced task-specific models fine-tuned with in-domain data from the VisText and Pew datasets. Second, DePlot + GPT-4 (Liu et al., 2023a; OpenAI, 2023a) is a LLM-based pipeline approach. Finally, GPT-4V and Bard[3] are the strongest LVLMs. For each model and dataset, we randomly sample 100 chart figures and generate the corresponding captions. Invalid output sequences, such as empty strings, are filtered out.

We compute the percentage of sentences with factual errors for different models and datasets, with a breakdown of different error types. Error rates are computed at the sentence level instead of the caption level since different models generate captions of different lengths. A sentence-level evaluation helps mitigate this discrepancy and facilitates a fairer comparison.

From Figure 1, we made the following observations. First, **SOTA chart captioning models often fail to produce factual captions**. Additionally, as shown in Table 2, we calculated the percentage of non-factual captions, revealing that 82.06% of captions contain at least one factual error. More importantly, even models like GPT-4V and Bard, which have demonstrated proficiency in a variety of vision-language tasks, produce factually incorrect captions 81.27% of the time, as recorded in Table 7. These findings highlight the inherent difficulties of chart captioning tasks and the limitations of SOTA vision-language models.

Second, **task-specific chart captioning models and LVLMs show opposite trends on the two datasets**. Task-specific models, including ChartT5, MatCha, and UniChart, produce fewer errors on the VisText dataset. Conversely, LVLMs, including GPT-4V and Bard, generate significantly fewer errors on the Pew dataset. The key distinctions on these datasets are two: (1) the prevalent labeled values on charts from Pew and (2) the simpler structures in charts from VisText. We hypothesize that LVLMs may be better at utilizing the labeled numbers, while task-specific effectively interpret values via axis alignment. We show an example to validate this hypothesis in Figure 6.

Third, **LVLMs cannot consistently outperform task-specific fine-tuned models**. Despite their extensive training data and parameters, LVLMs may be surpassed by task-specific models with appropriate pre-training objectives and architectures. For example, on the VisText dataset, UniChart outperforms Bard and is comparable to GPT-4V in terms of producing more factual captions owing to UniChart's various pre-training objectives for chart comprehension, enabling better interpretation of the relationship between data points within charts.

The dataset resulting from the analysis is named CHOCOLATE (**C**aptions **H**ave **O**ften **Ch**O**sen **L**ies

[2] Note that we do not consider grammatical errors as factual inconsistency. They are analyzed for assessing fluency.

[3] We tested Bard before Gemini's release (Google, 2023b).

	# Factual	# Non-factual	# Total
Sentence	2,561	2,762	5,323
Caption	213	974	1,187

Table 2: Statistics of the captions we analyzed. A sentence is considered factual if and only if it does not contain any factual error. A caption is considered factual if all its sentences are factual.

About **T**he **E**vidence), where each instance consists of a chart, a generated chart caption, and error types labeled by human annotators. Drawing insights from Tang et al. (2023b) that factual errors produced by different kinds of models may be easier or more difficult to identify, we categorize CHOCOLATE into three splits: the **LVLM** split, with captions from GPT-4V and Bard; the **LLM** split, featuring DePlot + GPT-4 outputs; and the **FT** split, for ChartT5, UniChart, and MatCha captions. Split details are in Appendix C.

2.3 Dataset Quality

To evaluate the quality of CHOCOLATE, we measured inter-annotator agreement by calculating Fleiss' Kappa κ (Fleiss, 1971) and the majority vote agreement percentage p, in line with the metrics used by Pagnoni et al. (2021). We applied these metrics across all 5,323 sentences in CHOCOLATE. For determining factual consistency between chart sentences and their corresponding charts, we achieved a Fleiss' Kappa of $\kappa = 0.63$ and a majority vote agreement of $p = 91\%$. For context, Pagnoni et al. (2021) reported a Fleiss' Kappa of $\kappa = 0.58$ and a majority agreement level of $p = 91\%$. This suggests that CHOCOLATE exhibits a quality on par with well-established benchmarks in text-based factual inconsistency detection.

3 The Chart Caption Factual Error Correction Task

The dataset collected in §2 enables us to study the Chart Caption Factual Error Correction task. In this section, we first formally provide the definition of this task (§3.1) and propose an effective reference-free evaluation metric based on chart visual entailment (§3.2).

3.1 Task Definition

The input to our task is a chart $\mathcal{E}$ and chart caption $\mathcal{C}$ that may or may not be factually consistent with $\mathcal{E}$. The goal of chart caption factual error correction is to produce a corrected caption $\hat{\mathcal{C}}$ that fixes factual errors in $\mathcal{C}$ with the minimum amount of edits. If $\mathcal{C}$ is already faithful to $\mathcal{E}$, models should output the original caption (i.e. $\hat{\mathcal{C}} = \mathcal{C}$). Following prior work on text-based factual error correction (Thorne and Vlachos, 2021; Huang et al., 2023b; Gao et al., 2023), corrections should be made with as few substitution, insertion, and deletion operations as possible since one can trivially achieve 0% non-factual rate by deleting all words in a caption.

3.2 Reference-free Evaluation With Chart Visual Entailment

There was no established metric for evaluating the factual consistency between a chart and the corresponding chart caption. In addition, since our dataset does not contain annotated reference captions[4], text-based metrics cannot be adopted. As a solution, we propose CHARTVE, a reference-free evaluation metric based on chart visual entailment, as detailed in the following paragraphs.

CHARTVE Overview We formulate the inconsistency detection problem as a chart visual entailment task. Given a chart caption sentence c and a chart $\mathcal{E}$, the task is to predict whether the relationship from $\mathcal{E}$ to c as ENTAILMENT (factually consistent) or NOTENTAILMENT (factually inconsistent). The main challenge of learning a visual entailment model for this task is the lack of data. To overcome this challenge, we repurpose data from relevant tasks, such as chart QA, as positive samples. Then, we propose a table-guided negative data generation to produce negative samples.

Positive Data Creation We consider datasets from two tasks that are closely related to the chart visual entailment task: chart question answering and chart captioning. We utilize two datasets from chart question answering: ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020). Using a QA2Claim model (Huang et al., 2023b), we transform the question-answer pairs into declarative statements and pair them with the original charts to form positive instances (ENTAILMENT). For chart captioning, captions from VisText (Tang et al., 2023a) and Chart-to-Text (Kantharaj et al., 2022) are segmented into individual sentences. Each sentence is paired with the relevant chart to create a positive instance. These methods allow us to repurpose existing resources for training CHARTVE.

[4]Reference captions are not collected due to the challenges of curating high-quality references through crowd-sourcing.

Table-guided Negative Data Generation Generating negative training samples is achieved by perturbing the positive instances grounded in the underlying data tables of the charts. For a chart $\mathcal{E}_i$ and its underlying data table $\mathcal{A}_{\mathcal{E}_i}$, we locate values in $\mathcal{A}_{\mathcal{E}_i}$ that matches a substring within the positive caption c_i^+. When a match is found, the substring in the caption is substituted with a different value from the same column in $\mathcal{A}_{\mathcal{E}_i}$, yielding a value or label-error infused negative sentence c_i^-, maintaining relevance while ensuring inconsistency with $\mathcal{E}_i$. For trend-related errors, we replace trend-terms found in c_i^+ with their opposites, drawing on a specific lexicon of terms like "increase" and "decrease," thereby creating trend-contradictory statements. Furthermore, out-of-context errors are crafted by pairing $\mathcal{E}_i$ with a mismatched caption c_j^+ from another chart, where $i \neq j$. This simulates captions filled with unrelated data.

The above process is illustrated in Algorithm 1. We use the training, development, and test sets of the repurposed datasets for training, validating, and testing CHARTVE. This is vital for ensuring that CHARTVE is free from data contamination in downstream applications. In total, we collected over 595K instances partitioned into training, development, and test splits with a ratio of 522:36:37, respectively.

Learning CHARTVE We selected UniChart as our base model, given its superior performance amongst comparable-size models[5]. Recognizing that UniChart has been pre-trained on chart question answering tasks, we employ a tailored input template t as follows:

> *Does the image entail this statement: "SENTENCE"?*

In this template, *SENTENCE* replaces the chart caption sentence c. Taking in a chart $\mathcal{E}$ and template t as input, UniChart is fine-tuned to produce the token "yes" if the chart $\mathcal{E}$ entails the caption sentence c, and "no" otherwise using maximum likelihood estimate. During inference time, we use the same input format and probe the logits corresponding to the "yes" (l_{yes}) and "no" (l_{no}) decoder tokens. Following this, we apply the softmax function to convert these logits into an entailment score $s(\mathcal{E}, c)$ that ranges from 0 to 1:

$$s(\mathcal{E}, c) = \frac{e^{l_{\text{yes}}}}{e^{l_{\text{yes}}} + e^{l_{\text{no}}}}. \tag{1}$$

Here, e is the base of the natural logarithm. Finally, we compute the minimum of the entailment scores for all sentences within a caption, denoted by $S(\mathcal{E}, \mathcal{C})$, where $\mathcal{C}$ represents the set of all caption sentences for chart $\mathcal{E}$:

$$S(\mathcal{E}, \mathcal{C}) = \min_{c \in \mathcal{C}} s(\mathcal{E}, c). \tag{2}$$

Meta-evaluation of Different Evaluation Metrics To evaluate the effectiveness of different methods in assessing the factuality of generated captions on the CHOCOLATE dataset, we employ Kendall's Tau (Kendall, 1938) to compute the correlation between these methods and human judgments. Given the absence of prior work on factual inconsistency detection methods for chart captions, we compare our CHARTVE with zero-shot capable methods, including DePlot + GPT-4, Bard, GPT-4V, and the leading open-source LVLM, LLaVA-1.5-13B (Liu et al., 2023c). Text-based factuality metrics, SUMMAC (Laban et al., 2022) and QAFACTEVAL (Fabbri et al., 2022b), which compute the factual consistency between the reference caption and the generated caption, are also included. The prompts for these models are detailed in Appendix E.

Model	CHOCOLATE LVLM	LLM	FT
SUMMAC	-0.011	0.023	0.036
QAFACTEVAL	0.064	0.045	0.054
LLaVA-1.5-13B	0.002	0.057	0.214
Bard	-0.014	0.105	0.**291**
GPT-4V	0.157	**0.205**	0.215
DePlot + GPT-4	0.129	0.117	0.109
CHARTVE (Ours)	**0.178**	0.091	0.215

Table 3: Kendall's Tau correlation of different approaches on the CHOCOLATE dataset.

Meta-evaluation, summarized in Table 3, shows that, overall, **metrics exhibit the strongest correlation with human judgment on the FT split and the weakest on the LVLM split**. This pattern aligns with expectations: the FT captions are littered with more obvious mistakes, such as out-of-context and nonsense errors, while errors stemming from LVLMs are harder to detect since they often demand intricate inferences regarding the data points' positions relative to the axes, as detailed in Figure 1. Importantly, Our CHARTVE excels on the challenging LVLM split, but less so on the LLM split, likely due to shifts in token distribution, as DePlot + GPT-4 occasionally employs table-centric

[5] Our fine-tuning begins with this checkpoint: https://huggingface.co/ahmed-masry/unichart-base-960.

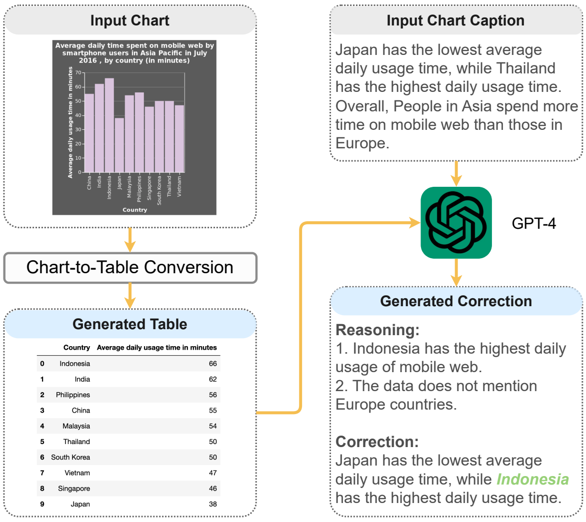


Figure 2: An overview of C2TFEC. Our approach decomposes visual reasoning into image-to-structure rendering and text-based reasoning, allowing for interpretability and better correction of chart captions.

terminology (e.g., "columns" and "entries") absent from CHARTVE's training data. Despite this, **CHARTVE compares favorably to proprietary LVLMs and outperforms LLaVA-1.5-13B, despite CHARTVE being *64 times smaller in scale*.**

Bard and GPT-4V lead on the LLM and FT splits, respectively. However, Bard shows a negative correlation on the LVLM split, hinting at LVLMs' limitations in assessing the factuality of chart captions. Thus, we advocate for using the best-performing metric for each split for evaluation.

4 Methodology

In correcting factual errors in generated captions, we propose C2TFEC, a two-step, interpretable framework, as shown in Figure 2. C2TFEC first transforms input charts into data tables (§4.1), then rectifies errors in the caption using the tabular data (4.2). This framework is motivated by our analysis on "DePlot + GPT-4", which shows that a notable proportion of errors in caption generation originated from the DePlot component. To mitigate this, we develop a stronger chart-to-table model based on UniChart, significantly improved with expansive fine-tuning datasets. The advantage of C2TFEC is its ability to harness the reasoning strengths of GPT-4 to faithfully correct errors, boosting caption factuality.[6]

[6] Here, we do not consider approaches based on LVLMs due to their tendency towards factual errors.

4.1 Chart-To-Table Conversion

The training data for our chart-to-table model is sourced from datasets including VisText, Chart-to-Text, ChartQA, and PlotQA, where we repurpose original charts and underlying data tables for our model's training. We collected a total of 65K instances with a train:dev:test split of 61:2:2. Similar to DePlot (Liu et al., 2023a), our model is also trained to generate chart titles, enhancing its ability to contextualize the data represented in table form. Let $\mathcal{M}$ denote our proposed model. For a given chart figure $\mathcal{E}$, the model autoregressively generates a chart title $\mathcal{T}$ and a corresponding table $\mathcal{A}$ (i.e. $\mathcal{T}, \mathcal{A} = \mathcal{M}(\mathcal{E})$).

4.2 Table-based Error Rectification

With the input chart now converted into structured tabular data, the second phase uses the reasoning capacity of LLMs to address the factual inconsistency between $\mathcal{C}$ and the generated table $\mathcal{A}$. Here, we use GPT-4 as the LLM. GPT-4 first provides an explanatory breakdown of detected factual errors in $\mathcal{C}$ based on the table contents. It then uses this explanation to produce a corrected caption $\hat{\mathcal{C}}$. This transparent process enables users to validate the reasoning behind each correction.

C2TFEC separates the factual verification from language generation, taking advantage of the complementary strengths of separate vision and language models tailored to their respective domains. The symbolic table representation acts as a bridge to enhance and validate factual consistency in chart captions.

5 Experimental Settings

To assess C2TFECs ability in factual error correction for chart captions, we experiment on the CHOCOLATE dataset.

Datasets Our CHOCOLATE dataset includes 1,187 chart-caption pairs with factually consistent and inconsistent captions, as detailed in §2. It is split into LVLM, LLM, and FT, reflecting the diversity of models that generated the captions.

Baselines Since CHOCOLATE does not comprise training data, we compare C2TFEC against zero-shot capable LVLMs and LLMs, including LVLMs, LLaVA-1.5-13B, GPT-4V, Bard, as well as DePlot + GPT-4. For a fairer comparison between our approach and DePlot, we continue fine-tuning DePlot

Dataset Split →	CHOCOLATE-LVLM		CHOCOLATE-LLM		CHOCOLATE-FT	
Evaluation Metric → **Correction Model ↓**	CHARTVE (%)	Levenshtein	GPT-4V (%)	Levenshtein	Bard (%)	Levenshtein
N/A	31.13	0.0	23.47	0.0	43.10	0.0
LLaVA	31.20	19.09	22.45	9.20	52.94	16.94
Bard	14.13	127.83	31.77	77.63	75.69	42.80
GPT-4V	33.30	31.26	**52.35**	50.57	76.55	30.92
DePlot + GPT-4	32.47	81.37	22.45	21.25	70.31	38.79
DePlot$_{\text{CFT}}$ + GPT-4	32.91	84.99	25.51	55.35	70.47	40.12
C2TFEC (Ours)	**34.34**	72.19	39.29	53.11	**81.14**	37.36

Table 4: Correction performance of different models on the CHOCOLATE dataset. CHARTVE measures factuality by computing the entailment probability from each chart to the corresponding caption sentences. GPT-4V and Bard, when used as evaluation metrics, rate each chart caption as factually consistent with the chart or not. Levenshtein computes the edit distance between the corrected caption and the original caption (denoted as "N/A"). Metric scores are shown separately for each of the three data splits based on captioning model source. The highest and second highest performing models per evaluation metric and split are highlighted in boldface and underlines respectively.

for an additional 5,000 steps on VisText, an approach which has been shown effective for adapting models to unseen domains (Huang et al., 2023b). We denote this model as DePlot$_{\text{CFT}}$. The prompts used for each model are described in Appendix E.

Evaluation Metrics We assess the factual consistency between corrected captions and input charts using CHARTVE, GPT-4V, and Bard, according to our recommendations in §3.2. In addition, since corrections should be made with as few edits as possible, we measure the number of edits using the Levenshtein distance (Levenshtein et al., 1966).

6 Results

6.1 Main Results

The results in Table 4 demonstrate that our C2TFEC achieves the best performance for factual consistency on the LVLM and FT splits, and takes the second place on the LLM split. This indicates that **the two-step process of first transforming charts into structured data tables and then rectifying factual inconsistencies using table-caption alignment is an effective strategy.**

Bard's underperformance on the LVLM split and its negative correlation with human judgments of factuality, as shown in Table 3, implies its unreliability in detecting errors in chart captions. Additionally, when used as an evaluator, GPT-4V tends to assign high factuality scores to its own corrected outputs on all three splits (see Table 8), while other metrics show GPT-4V lagging behind C2TFEC. This suggests GPT-4V may suffer from the *self-enhancement bias* (Zheng et al., 2023), overestimating its own performance when used for evaluation. We thus perform human evaluations in §6.2 to verify the effectiveness of our approach.

6.2 Human Evaluation

Our human assessments focus on comparing C2TFEC with GPT-4V by using the same annotation tasks detailed in §2 for factual error identification, with the same annotators evaluating. We sampled 30 charts from each split of LVLM, LLM, and FT. For each chart, human judges are presented with a caption generated by one of the models.

Figure 3 demonstrates C2TFECs superiority in multiple error categories, especially with a substantial decrease in Value Errors, over 20% better in the LVLM and LLM splits, and halving the overall error rate compared to GPT-4V. C2TFEC virtually eliminated Trend Errors, highlighting its strong error correction ability, particularly for axes-related errors like Label, Value, and Trend errors. A representative comparison is shown in Figure 4. GPT-4V's shortcomings seem to stem from its failure to accurately infer data point values from charts as evidenced in Figure 7.

In contrast, GPT-4V is better in addressing Out-of-context Errors, involving information out of the chart's scope. However, GPT-4V seemed challenged in rectifying errors within captions generated by itself, particularly within the LVLM split. This observation echoes recent findings on LLMs' inability to self-correct (Huang et al., 2023a; Valmeekam et al., 2023), we find that **LVLMs also cannot perform self-correction**. More importantly, our human evaluation results, combined with our findings in Table 4 and Table 8, reflect that GPT-4V is subject to serious self-enhancement bias. Consequently, **although GPT-4V's capabilities are formidable, we recommend not using them to assess their own outputs**.

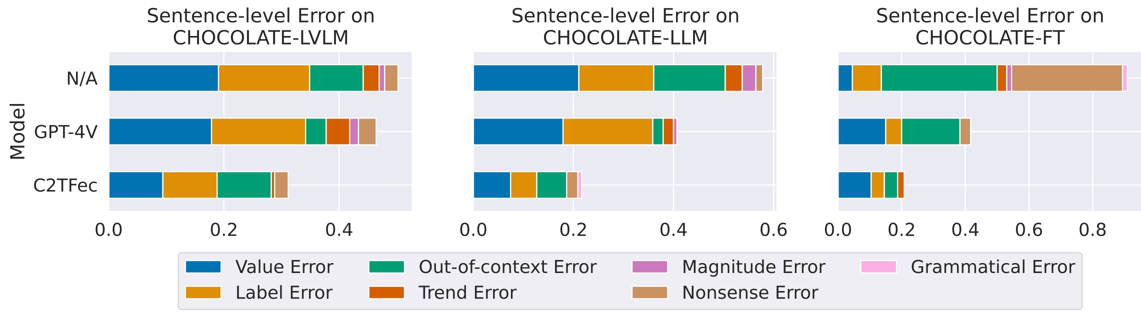


Figure 3: Human evaluation results on subsets of the CHOCOLATE dataset, comparing C2TFEC and GPT-4V. C2TFEC corrects significantly more errors compared to GPT-4V, especially Value, Label, and Trend Errors.

7 Related Work

7.1 Chart Captioning

Chart captioning is essential for accurately interpreting and communicating the information conveyed by chart images, particularly in news articles and social media, where factuality is imperative to prevent misinformation. While current datasets like FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018), PlotQA (Methani et al., 2020), VisText (Tang et al., 2023a), and Chart-to-Text (Kantharaj et al., 2022) offer chart image descriptions and question-answer pairs to train models, advancements in vision-language models like ChartT5 (Zhou et al., 2023), MatCha (Liu et al., 2023b), and UniChart (Masry et al., 2023) have largely prioritized relevance and fluency over factual accuracy. Our work provides a rigorous characterization of factual errors in chart captioning and comparisons of methods to address this gap. By focusing on faithfulness and correction, we complement the emphasis of prior work and aim to produce more trustworthy chart captions.

7.2 Factual Error Correction

Prior research in factual error correction has mainly targeted text summarization and fact-checking. Within summarization, the bulk of work has been carried out in the news domain and often involves methods that substitute inconsistent entities from the source text. Some studies have enhanced this approach through entity-replacement reranking techniques (Chen et al., 2021), autoregressive models for rewriting and perturbation filtering (Cao et al., 2020; Zhu et al., 2021; Adams et al., 2022), and editing strategies that focus on selective deletion (Wan and Bansal, 2022). In contrast, Fabbri et al. (2022a) employed sentence compression datasets to train their models. More recently, Gao et al. (2023) have expanded the focus of these studies to include dialogue summarization.

Moving to the domain of fact-checking, this area has experienced a flurry of activity, particularly with the increased attention on combating misinformation (Fung et al., 2021; Wu et al., 2022; Fung et al., 2022; Huang et al., 2023d,c; Qiu et al., 2023). Early approaches train a distantly supervised model that involves a masker and a corrector (Shah et al., 2020; Thorne and Vlachos, 2021). Thorne and Vlachos (2021) made significant strides by developing the first factual error correction dataset for fact-checking, thus enabling fully supervised training for error correctors. Recently, Huang et al. (2023b) propose an interpretable framework that breaks down the process of fact-checking into individual components. Our study builds on these insights and extends them to a multimodal context, which challenges models to understand the chart images and the consistency between different modalities.

8 Conclusion

Our study exposes the prevalent issue of factual errors in chart captions generated by various chart captioning models and introduces CHOCOLATE to scrutinize these errors. We establish the Chart Caption Factual Error Correction task to propel the creation of trustworthy captioning systems and present CHARTVE, an evaluation model surpassing LVLMs in mirroring human assessments of caption factuality. Our two-stage correction framework, C2TFEC, provides an interpretable means of improving caption factuality by transforming visual data into structured tables for more faithful error corrections. Our work marks an essential step in ensuring verifiable and trustworthy chart captions. Future directions include extending our approach to multimodal contexts beyond charts, developing more sophisticated error detection and correction algorithms, and creating datasets covering a broader range of visual content.

9 Ethical Considerations

Text generation models pre-trained on information from the Web are known to demonstrate various biases. Despite the primary focus on models and datasets that represent the English-speaking population's culture, manual examinations of the CHOCOLATE dataset reveal no evidence of biases related to gender, age, race, or other socioeconomic factors.

In §2 and §6.2, we recruited annotators to assess the factual consistency of chart captions. The annotators were fairly compensated for their efforts, as detailed in Appendix B. During the annotation process, we made provisions for open communication, allowing the annotators the flexibility to work at their preferred pace and the freedom to withdraw from the project at any point. Additionally, we took measures to protect the anonymity of the contributors by excluding any personally identifiable information from the dataset.

10 Limitations

We acknowledge that our study did not rigorously examine the sensitivity of different systems to the variations in the prompts used. The effectiveness of several natural language processing tasks is known to be influenced by the design of the input prompts. Our omission of a systematic sensitivity analysis means that there could be a range of responses to different prompts that we have not accounted for, which may affect the generalization of our results. However, we did not perform prompt tuning to craft prompts that benefit our proposed model. Therefore, the comparisons across all models are fair. Due to the scope of our study, we leave the prompt sensitivity experiments for future work.

In addition, charts in the datasets we used are mostly line plots and bar plots. Future efforts can extend our work with additional analyses for other types of charts, such as violin plots and distribution plots.

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Alex Fabbri, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu, and Caiming Xiong. 2022a. Improving factual consistency in summarization with compression-based post-editing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9149–9156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022b. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. 2022. Doc2ppt: automatic presentation slides generation from scientific documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 634–642.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Yi R. Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4790–4791, New York, NY, USA. Association for Computing Machinery.

Mingqi Gao, Xiaojun Wan, Jia Su, Zhefeng Wang, and Baoxing Huai. 2023. Reference matters: Benchmarking factual error correction for dialogue summarization with fine-grained evaluation framework.

In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13932–13959, Toronto, Canada. Association for Computational Linguistics.

Google. 2023a. Bard.

Google. 2023b. Gemini - google deepmind.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023b. Zero-shot faithful factual error correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5660–5676, Toronto, Canada. Association for Computational Linguistics.

Kung-Hsiang Huang, Hou Pong Chan, Kathleen McKeown, and Heng Ji. 2023c. Manitweet: A new benchmark for identifying manipulation of news on social media. *arXiv preprint arXiv:2305.14225*.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023d. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*.

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

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

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, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning.

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, Dublin, Ireland. 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.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

João Monteiro, Asanobu Kitamoto, and Bruno Martins. 2017. Situational awareness from social media photographs using automated image captioning. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 203–211. IEEE.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Gpt-4v(ision) system card.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. 2023. Amrfact: Enhancing summarization factuality evaluation with amr-driven training data generation. *arXiv preprint arXiv:2311.09521*.

Darsh Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8791–8798.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023a. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023b. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, Online. Association for Computational Linguistics.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. 2023. Paxion: Patching video-language foundation models with action knowledge. In *Proc. 2023 Conference on Neural Information Processing Systems (NeurIPS2023) [Spotlight Paper]*.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

A Further Discussions

Captioning Model Analysis In addition the findings we summarize in §2.2, we also found that **the error distribution for each model differs on different datasets**. Almost all models make significantly more Nonsense Errors on the Pew dataset. In addition, task-specific models observe a non-negligible increase in Out-of-context Errors on the Pew dataset. Both observations could be explained by the fact that these models are sometimes confused about the charts in Pew, which are often associated with more complicated structures.

Furthermore, in Figure 1, the error rates are computed as the number of such errors divided by the number of sentences. While this provides an overview of the *frequency* for each error, it does not indicate the likelihood of a value/label/trend/magnitude-related mention in the generated captions being factual. This limitation can result in an underrepresentation of certain error types – for instance, the infrequent occurrence of Magnitude Errors as shown in Figure 1 is more a consequence of the scarcity of magnitude-related mentions in the captions rather than an indication of the models' superior trend variance comprehension. To address this, we sample 30 generated captions for each model from each dataset and compute another error rate as the number of sentences containing such non-factual mentions over the number of sentences containing such mentions. The results are shown in Table 5. The outcomes corroborate the observations in §2.2, while Table 5 offers a supplementary perspective on model performance.

Meta-evaluation Results For the text-based metrics presented in Table 3, they both perform weakly in determining the factuality of the generated caption. This is largely because charts often contain much denser information compared to the corresponding reference. As a result, text-only factuality metrics are unsuitable for assessing factual consistency between charts and captions.

Main Results We see that C2TFEC outperforms the pipeline approaches of DePlot/DePlot$_{\text{CFT}}$ + GPT-4 across the board. While both methods utilize an intermediate tabular representation and leverage GPT-4 for language generation/correction, C2TFEC employs a superior chart-to-table conversion model with much more comprehensive training datasets. This results in extracted tables that more faithfully capture the underlying chart data, better facilitating the downstream factual error cor-

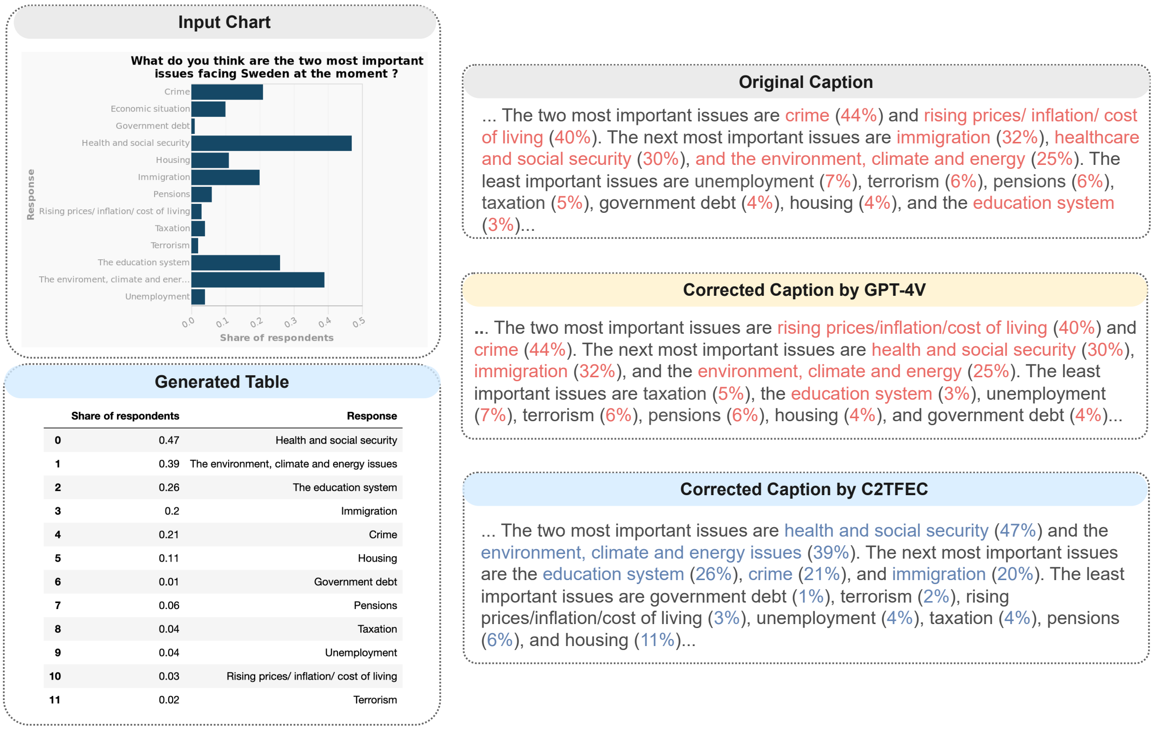


Figure 4: An example showing how decomposing the visual reasoning process into image-to-structure rendering and text-based reasoning allows C2TFEC to accurately rectify errors in chart captions. Texts marked in red indicate non-factual information units in the caption, whereas those marked in blue represent information units faithful to the chart. In this instance, C2TFEC successfully corrects all Value and Label Errors presented in the original caption. Conversely, GPT-4V fails to identify the factual inconsistencies and merely reorders the entities in the caption.

Dataset →	VisText				Pew			
Error Type → **Model ↓**	Value	Label	Trend	Magnitude	Value	Label	Trend	Magnitude
ChartT5	92.31 (12/13)	64.71 (33/51)	32.00 (8/25)	100.00 (3/3)	66.67 (2/3)	100.00 (2/2)	N/A (0/0)	N/A (0/0)
MatCha	71.43 (5/7)	50.00 (13/26)	23.33 (7/30)	50.00 (1/2)	100.00 (2/2)	66.67 (2/3)	N/A (0/0)	N/A (0/0)
UniChart	**33.33 (3/9)**	**29.41 (10/34)**	**0.00 (0/14)**	50.00 (2/4)	51.72 (15/29)	46.67 (14/30)	100.00 (1/1)	N/A (0/0)
DePlot + GPT-4	51.52 (34/66)	44.78 (30/67)	30.77 (8/26)	**0.00 (0/7)**	49.25 (33/67)	34.48 (10/29)	46.15 (6/13)	**0.00 (0/3)**
Bard	69.12 (47/69)	69.39 (34/49)	43.75 (14/32)	15.38 (2/13)	38.10 (40/105)	27.71 (23/83)	11.11 (2/18)	40.00 (2/5)
GPT-4V	40.48 (17/42)	33.33 (17/51)	20.75 (11/53)	23.53 (4/17)	**8.20 (10/122)**	**9.02 (11/122)**	**16.67 (2/12)**	33.33 (2/6)

Table 5: Error rates (%) are calculated by dividing the number of sentences containing such non-factual mentions (e.g. non-factual mentions of values) by the number of sentences containing such mentions (e.g. all mentions of values). The lower the error rate, the better the performance.

Dataset Split →	CHOCOLATE-LVLM		CHOCOLATE-LLM		CHOCOLATE-FT	
Evaluation Metric → **Correction Model ↓**	CHARTVE (%)	Levenshtein	GPT-4V (%)	Levenshtein	Bard (%)	Levenshtein
C2TFEC	29.29	62.85	40.63	35.63	49.49	23.48
C2TFEC (w/ GT Table)	**29.90**	52.82	**40.69**	32.59	**50.93**	23.47

Table 6: Correction performance of different models on the CHOCOLATE dataset. CHARTVE measures factuality by computing the entailment probability from each chart to the corresponding caption sentences. GPT-4V and Bard, when used as evaluation metrics, rate each chart caption as factually consistent with the chart or not. Levenshtein computes the edit distance between the corrected caption and the original caption. Metric scores are shown separately for each of the three data splits. Note that the Bard metric corresponds to Gemini Pro (Google, 2023b) since the experiments were conducted after its release.

rection. C2TFEC also requires a relatively small number of edits to captions according to Levenshtein distance, making focused changes to improve factuality while minimizing revisions. An example output from C2TFEC is shown in Figure 4. By comparison, the proprietary LVLM Bard produces corrected captions requiring 127.83 as many character-level edits on average. This signals excessive rewriting rather than targeted error correction. After manually inspecting Bard's outputs, we found the reason is that Bard oftentimes try to improve the fluency of the caption by paraphrasing. Hence, it makes more edits to the generated captions.

Understanding The Upper Bound We seek to understand the performance upper bound of our proposed two-stage framework by replacing generated tables with ground-truth data tables. Since the ground-truth data tables in Pew are not available, we experiment with only the instances from the VisText dataset. The results are demonstrated in Table 6.

B Annotation Details

In this section, we present the details of our human annotation conducted in §2.

B.1 Worker Qualification

We laid out specific preliminary criteria for the recruitment of MTurk workers with impressive performance records. These prerequisites comprise a HIT approval percentage of 99% or above, a minimum of 10,000 approved HITs, and the worker's location within the United Kingdom, Canada, or the United States.

Moreover, beyond these initial criteria, suitable workers have to successfully pass two staged qualification examinations focused on identifying factual errors in generated chart captions. To optimize the qualification procedure, the authors manually annotate two HITs, each consisting of one chart and one caption produced by one of our chart captioning models. In every qualification round, annotators are exposed to one of these annotated examples. Workers whose annotations fail to correspond closely with ours are eliminated from the selection procedure.

Finally, a group of 7 annotators who successfully navigated all three stages of qualification tests were chosen. Additionally, each HIT was meticulously crafted to ensure that annotators could achieve an equivalent hourly pay rate of $15 - $20, assuming they work without interruption.

B.2 Annotation Guidelines

In this task, you will evaluate the factual errors for a generated caption with regard to the reference chart. To correctly solve this task, follow these steps:

- Carefully read the generated caption and the reference chart.

- Compare the generated caption against the reference chart and decide whether the caption contains any factual error defined below.
- You should click/press the button if an error occurs. A blue button indicates the caption contains the corresponding factual error, while a white button means the caption does not contain such an error.

Warning: Annotations will be checked for quality against control labels, low-quality work will be rejected.

Error definition

- **Value error**: A quantitative data value is incorrect.
- **Label error**: A non-quantitative data value is incorrect.
- **Trend error**: The direction of a trend is wrong.
- **Magnitude error**: The magnitude or variance of a trend is wrong.
- **Out-of-context error**: The caption introduces concepts that are not present in the chart.
- **Grammatical error**: The grammar of the caption is wrong.
- **Nonsense error**: The caption is incomplete or does not make sense at all.

B.3 Annotation Interface

The interface for our human annotation is shown in Figure 5.

C Dataset Details

Table 7 presents the detailed statistics of each split in our dataset.

D Implementation Details

D.1 Details of the Chart-To-Table Model

Our chart-to-table model takes in as input a graphical chart and outputs a linearized data table format, using `\t` to delimit columns and `&&&` for row separation. The backbone of our approach is UniChart (Masry et al., 2023), due to its diverse chart-oriented pre-training objectives that have demonstrated strong performance on relevant tasks.

D.2 Table-guided Negative Data Generation

In Algorithm 1, we depict the details of how we generate negative data for our CHARTVE model.

D.3 Model Training

The Chart-To-Table model and CHARTVE are optimized using AdamW for a maximum of 20,000 and 50,000 steps, respectively. The learning rates for both models are set to 5e-5. During inference time, the Chart-To-Table model uses beam search with a beam width of 4.

E Prompts

The prompts for using LVLM and LLM as evaluation metrics are displayed in Figure 8 and Figure 9,

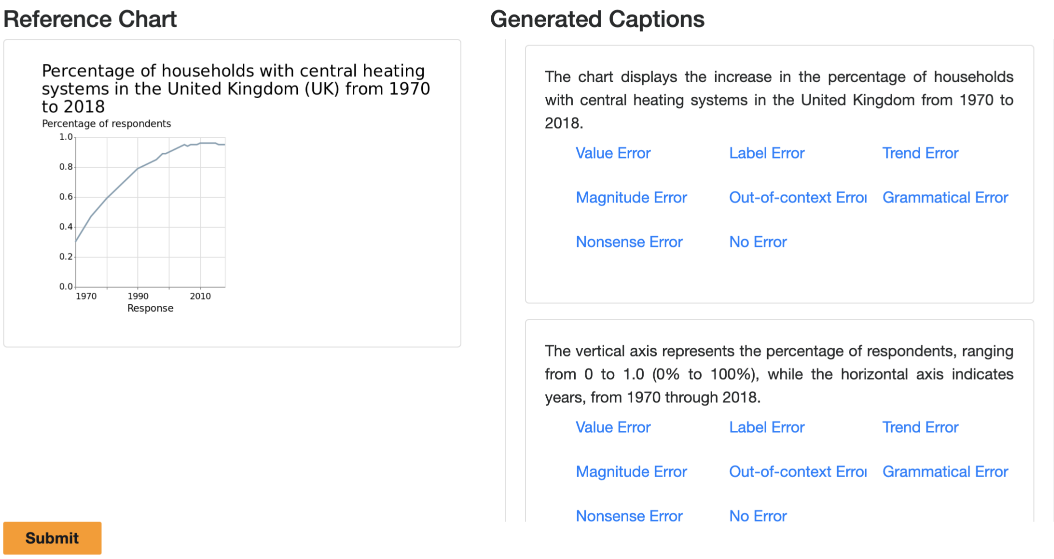


Figure 5: Human annotation interface for our data collection discussed in §2. Examples of each type of error from Table 1 are also displayed in the annotation interface. We were not able to show these examples in this figure due to space limits.

	Chocolate-LVLM		Chocolate-LLM		Chocolate-FT	
	# Factual	# Non-factual	# Factual	# Non-factual	# Factual	# Non-factual
Sentence	1,683	1,270	518	469	360	1,023
Caption	74	321	27	169	112	484

Table 7: Dataset statistics per split. A sentence is considered factual if and only if it does not contain any factual error. A caption is considered factual if all its sentences are factual.

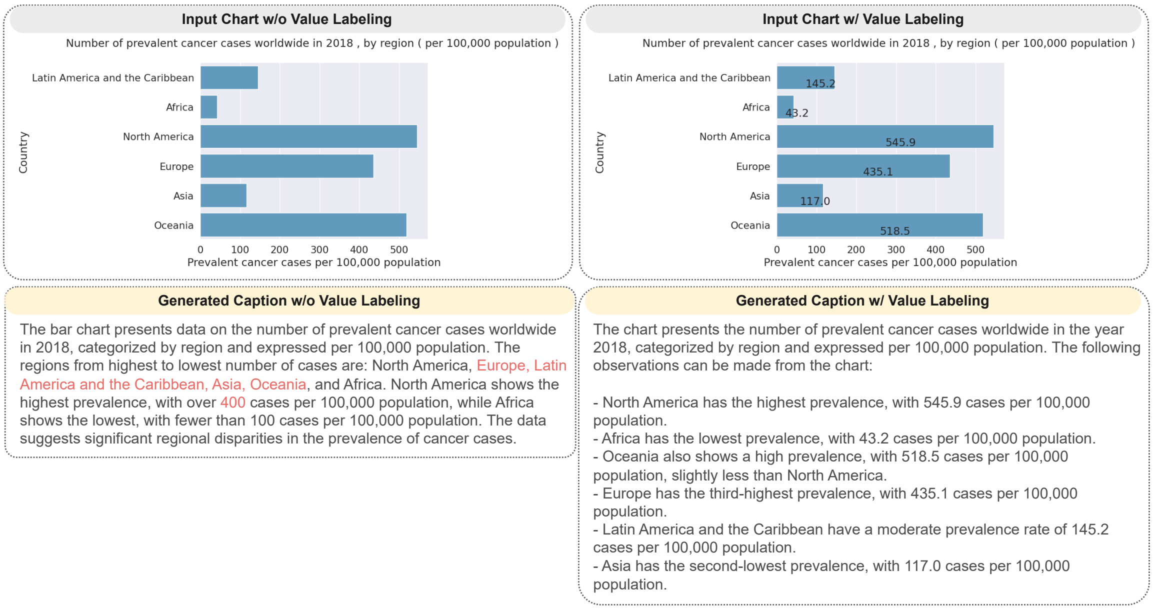


Figure 6: The impact of value labeling. We prompted GPT-4V to generate captions of two charts we created using the Seaborn library from an underlying table sampled from the Chart-to-Text dataset, with or without labeling the values of the bars on the chart. We see that when the labeled values are presented in the chart, GPT-4V is capable of producing more factual captions.

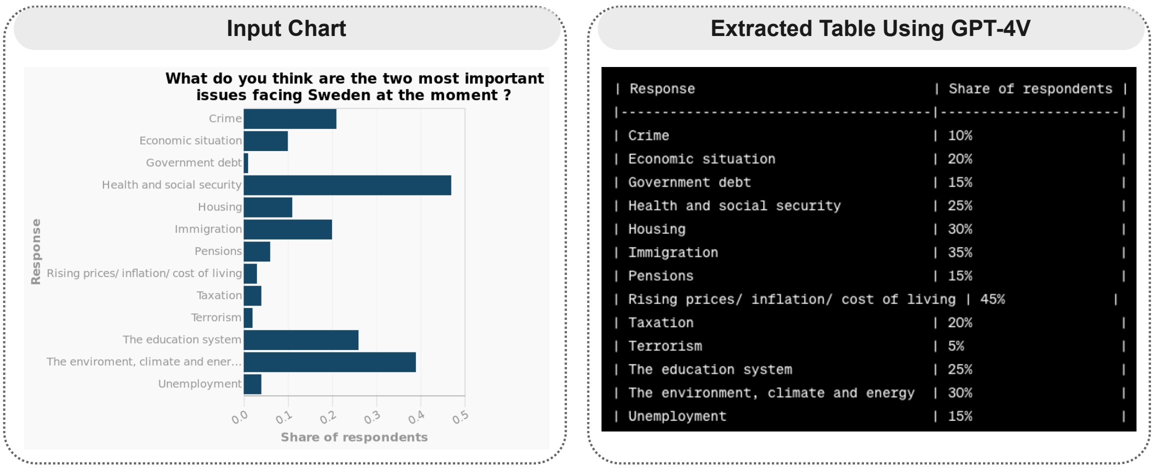
while the prompts for factual error correction are shown in Figure 10 and Figure 11.

Algorithm 1: Table-guided Negative Data Generation

Input: Data table $\mathcal{A}_{\mathcal{E}_i}$ for chart $\mathcal{E}_i$, Positive caption sentence c_i^+.

Output: Set of negative caption sentences $C_i^- = \{c_{i,\text{value}}^-, c_{i,\text{trend}}^-, c_{i,\text{context}}^-\}$.

Initialize C_i^- as an empty set;
Define a lexicon of trend terms T;
Define entailment threshold τ;
// Generate Value and Label Errors;
for *each cell value v in $\mathcal{A}_{\mathcal{E}_i}$* **do**
    if *v is a substring of c_i^+* **then**
        Randomly sample a new value v' from the same column in $\mathcal{A}_{\mathcal{E}_i}$;
        Replace v in c_i^+ with v' to get $c_{i,\text{value}}^-$;
        Add $c_{i,\text{value}}^-$ to C_i^-;
// Generate Trend Errors;
for *each trend term t in T* **do**
    if *t is found in c_i^+* **then**
        Replace t in c_i^+ with its antonym to get $c_{i,\text{trend}}^-$;
        Add $c_{i,\text{trend}}^-$ to C_i^-;
// Generate Out-of-Context Errors;
Randomly select a different chart $\mathcal{E}_j$ where $j \neq i$;
Pair $\mathcal{E}_i$ with unrelated caption sentence c_j^+ to get $c_{i,\text{context}}^-$;
Add $c_{i,\text{context}}^-$ to C_i^-;
return C_i^-;



Response	Share of respondents
Crime	10%
Economic situation	20%
Government debt	15%
Health and social security	25%
Housing	30%
Immigration	35%
Pensions	15%
Rising prices/ inflation/ cost of living	45%
Taxation	20%
Terrorism	5%
The education system	25%
The environment, climate and energy	30%
Unemployment	15%

Figure 7: An example showing GPT-4V cannot accurately extract tables from charts. This indicates its inability to infer the actual value of each data point within the chart.

Dataset Split → **Evaluation Metric →** **Correction Model ↓**	**CHOCOLATE-LVLM** GPT-4V	**CHOCOLATE-LLM** GPT-4V	**CHOCOLATE-FT** GPT-4V
N/A	<u>50.89</u>	23.47	24.83
LLaVA	29.87	22.45	39.45
Bard	37.37	31.77	44.86
GPT-4V	**61.34**	**52.35**	**74.79**
DePlot + GPT-4	23.79	22.45	40.63
C2TFEC (Ours)	35.96	<u>39.29</u>	<u>55.56</u>

Table 8: Correction performance on CHOCOLATE using GPT-4V as the evaluation metric. GPT-4V, when used as an evaluator, assigns significantly higher scores to its own generations. This suggests potential self-enhancement bias of GPT-4V. Note that GPT-4V also assign a high scores to the original captions (i.e. N/A) on the LVLM split. This is because half of these captions are directly generated from GPT-4V.

LVLM Evaluation Prompt

You are given a chart and a caption, you are tasked to detect whether the caption is factually consistent with the chart.
[Start of Caption]
{caption}
[End of Caption]
You should answer 'Answer: Yes' or 'Answer: No'. Do not provide explanation or other thing.

Figure 8: Prompts for using GPT-4V, Bard, and LLaVA-1.5 as a evaluator.

LLM Evaluation Prompt

You are given a table extracted from a chart and a caption. The table uses "<0x0A>" to delimit rows and "|" to delimit columns. The first row is the extracted chart title. You are tasked to detect whether the caption is factually consistent with the table.

[Start of Extracted Table]
{table}
[End of Extracted Table]
[Start of Caption]
{caption}
[End of Caption]

You should answer 'Answer: Yes' or 'Answer: No'. Do not provide explanation or other thing.

Figure 9: Prompts for using DePlot + GPT-4 as a evaluator.

LVLM Correction Prompt

You are given a chart and a chart caption. Your task is to correct errors in the caption based on the given chart. You should correct factual errors in the caption by as few substitution, insertion, and deletion operations as possible.

[Start of Caption]
{caption}
[End of Caption]
========

You must give your response in a structured JSON format that can be directly parsed with json.loads. Your response should contain two fields and two fields only:
"corrected_caption": the corrected caption based on the chart provided
"explanation": an explanation of your correction
Please follow the below rules:
1. Do not include "```json" in your response so that your output can be directly parsed with json.loads.
2. There are likely multiple errors in the caption. Please correct all factual errors. If there is no error, "corrected_caption" should be the same as input caption.

Figure 10: Prompts for using GPT-4V, Bard, and LLaVA as a factual error corrector.

LLM Correction Prompt

You are given a Markdown table, a chart title and a chart caption. The linearized table is assumed to faithfully represent the chart correpsonding to the caption. Your task is to correct errors in the caption based on the Markdown table and the chart title. You should correct factual errors in the caption by as few substitution, insertion, and deletion operations as possible.

[Start of Table]
{extracted_table}
[End of Table]

[Start of Chart Title]
{extracted_title}
[End of Chart Title]

[Start of Caption]
{caption}
[End of Caption]
========

You must give your in a structured JSON format that can be directly parsed with json.loads. Your response should contain two fields and two fields only:
"explanation"; an explanation of your correction
"corrected_caption": the corrected caption based on the table provided
Please follow the below rules:
1. Do not include "```json" in your response so that your output can be directly parsed with json.loads.
2. There are likely multiple errors in the caption. Please correct all factual errors. If there is no error, "corrected_caption" should be the same as input caption.

Figure 11: Prompts used for using DePlot + GPT-4 as a factual error corrector.