Faithful Chart Summarization with ChaTS-Pi

Syrine Krichene, Francesco Piccinno, Fangyu Liu, Julian Martin Eisenschlos
Google DeepMind, Zürich
{syrinekrichene,liufangyu,piccinno,eisenjulian}@google.com

Abstract

Chart-to-summary generation can help explore data, communicate insights, and help the visually impaired people. Multi-modal generative models have been used to produce fluent summaries, but they can suffer from factual and perceptual errors. In this work we present CHATS-CRITIC 🕵️, a reference-free chart summarization metric for scoring faithfulness. CHATS-CRITIC is composed of an image-to-text model to recover the table from a chart, and a tabular entailment model applied to score the summary sentence by sentence. We find that CHATS-CRITIC evaluates the summary quality according to human ratings better than reference-based metrics, either learned or n-gram based, and can be further used to fix candidate summaries by removing not supported sentences. We then introduce CHATS-PI 🍰, a chart-to-summary pipeline that leverages CHATS-CRITIC during inference to fix and rank sampled candidates from any chart-summarization model. We evaluate CHATS-PI and CHATS-CRITIC using human raters, establishing state-of-the-art results on two popular chart-to-summary datasets.[1]

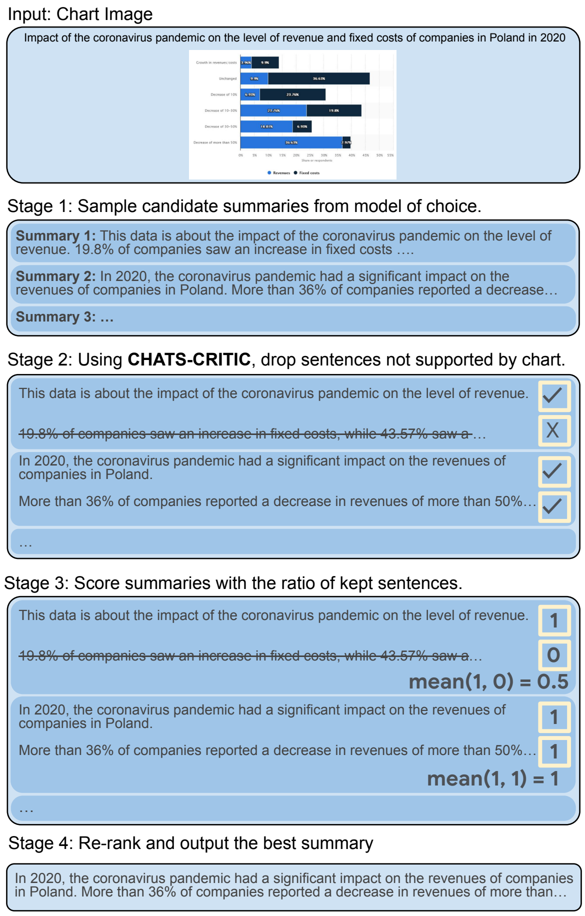


Figure 1: CHATS-PI generates multiple summaries given the chart using any summarization model. Each summary is then *repaired* by dropping refuted sentences according to the CHATS-CRITIC sentence scoring. Finally, we rank the summaries by computing the ratio of sentences that were kept.

1 Introduction

Chart summarization requires faithfully extracting quantitative data and describing them using natural language. Recent natural language generation (NLG) studies have explored different flavors of chart-to-summary generation tasks including caption generation for scientific figures (Hsu et al., 2021), chart summary generation (Kantharaj et al., 2022), or analytical textual descriptions for charts (Zhu et al., 2021). These tasks can be advantageous for the visually impaired (Benji Andrews, 2023) as well as for automating interpreting complex domains such as finance data-analysis, news reporting, and scientific domains (Siegel et al., 2016).

While a wide range of models and techniques have been applied for chart summarization, faithfulness remains a major challenge for the task. Specifically, the models often misread details in the charts (due to perceptual mistakes) or miscalculate the aggregations

[1]Code and demo at hf.co/spaces/chats-pi/chats-pi.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8705–8723
August 11-16, 2024 ©2024 Association for Computational Linguistics

(due to reasoning flaws). To overcome some of these limitations, *optical character recognition* (OCR) models and object detection systems are usually employed to extract meta-data such as axis, values, titles, legend (Luo et al., 2021; Masry et al., 2022). These data are then used as auxiliary inputs to finetune NLG models. Nonetheless, these modeling efforts are still limited by two fundamental issues (i) training & evaluation dataset quality and (ii) the reference-based metrics being used for evaluation. As examples, two widely used datasets, Chart-to-Text (Kantharaj et al., 2022) and SciCap (Hsu et al., 2021), are automatically extracted from web articles and academic journals. As a result, the summary references are prone to *hallucination*, i.e. the reference might contain context that cannot be entailed solely by the chart content. Training on this data can encourage the NLG models to improvise/hallucinate. Besides, the auto-extracted summaries sometimes emphasize only certain aspects of the chart, missing out critical insights.

On the other hand, n-gram based metrics such as BLEU (Papineni et al., 2002), or learned metrics such as BLEURT (Sellam et al., 2020) rely only on gold references. They are not capable of recognizing unreferenced but correct insights since they solely rely on the reference for scoring the summaries, as shown in Figure 3. This issue is especially pronounced when the gold references are noisy, which is the case for Chart-to-Text and SciCap. Last but not least, reference-based metrics also heavily penalize summary style mismatches, giving an artificial disadvantage to LLMs which are not tuned on the task data (Maynez et al., 2023).

This motivates building a reference-free critic CHATS-CRITIC (Figure 2) that can be used as a metric to score and re-rank summaries. We additionally introduce CHATS-PI (Figure 1) that leverage CHATS-CRITIC scores to generate a high quality summaries. We summarize our contributions as follows:

1. We present CHATS-CRITIC, a reference-free metric composed of a model that extracts the underlying table data from the chart and a table-entailment model acting on each sentence within a chart summary.
2. We design CHATS-PI, a pipeline that (i) generates multiple candidate summaries using a generative model, either fine-tuned or with in-context learning; (ii) then leverages CHATS-CRITIC to refine the summaries by dropping unsupported sentences; (iii) computes a summary score to rank the summaries by penalizing summaries with dropped sentences to increase the fluency, and (iv) outputs the best one.
3. To assess the efficacy of CHATS-CRITIC, we juxtapose human preferences against both CHATS-CRITIC and other prevailing metrics. Our results indicate that CHATS-CRITIC aligns more consistently with human evaluations. Furthermore, when contrasting CHATS-PI with other leading models that serve as baselines, CHATS-PI establishes state-of-the-art on two popular English benchmarks.

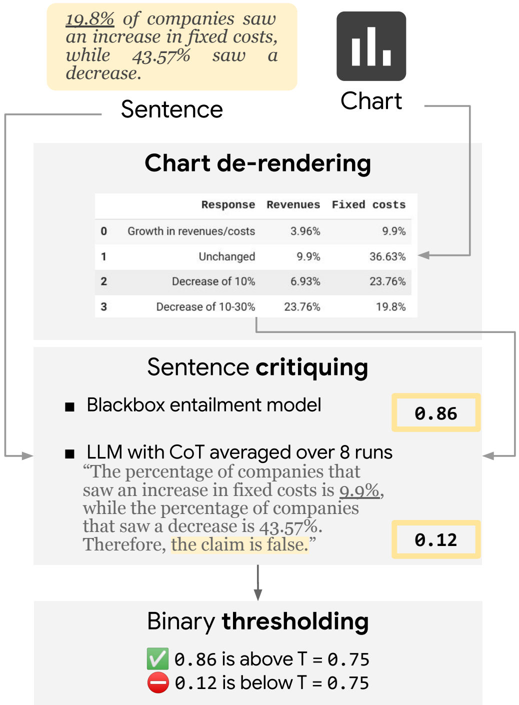
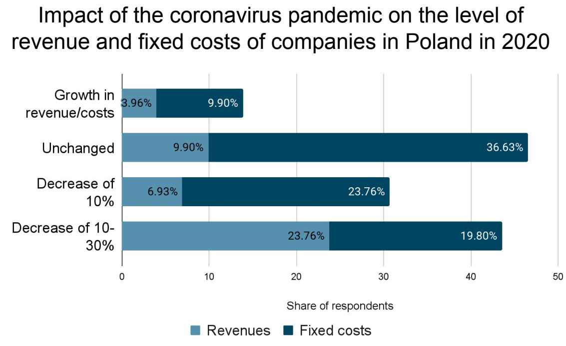


Figure 2: CHATS-CRITIC is composed of a de-rendering model to extract the table from the chart, and a table entailment model. The latter can be a blackbox table entailment model (e.g., TabFact as benchmarked in Table 3) or an LLM; in latter case, we use CoT prompt and average over 8 samples. In the figure, the threshold to reach a binary decision is set to $T = 0.75$. The chart icon refers to the same plot of Figure 3.



Metrics & Human Ratings for Entailed / Not Entailed Sentences

Reference
Eighty-six percent of managers expect their companies 'revenues from the outbreak of coronavirus in Poland to decrease within three months. Only 9.9 percent of them stated that the revenues would remain unchanged, and almost four percent expect growth . At the same time, 36.6 percent of respondents believe that fixed costs will not change, and 9.9 percent that they will go up. A drop of 10 percent is forecast by 23.8 percent of managers. For further information about the coronavirus (COVID-19) pandemic, please visit our dedicated Facts and Figures page.

BLEU: 100 **BLEURT-20**: 0.94 **PARENT**: 0.84
ChaTS-Critic (Ours): 0.4 = mean(0, 1, 1, 0, 0)

PALM-2
This data is about the impact of the coronavirus (COVID-19) pandemic on the level of revenue and fixed costs of companies in Poland in 2020. The data shows that more companies experienced a decrease in revenue than an increase in revenue. The data also shows that more companies experienced a decrease in fixed costs than an increase in fixed costs. The largest percentage of companies (36.63%) experienced a decrease of more than 50% in revenue. The smallest percentage of companies (2.97%) experienced a decrease of more than 50% in fixed costs.

BLEU: 9.3 **BLEURT-20**: 0.43 **PARENT**: 0.25
ChaTS-Critic (Ours): 1.0 = mean(1, 1, 1, 1, 1)

DePlot+Flan-T5
According to one of the scenarios, it is assumed that the outbreak of coronavirus (COVID-19) will cause a deeper recession in Poland. It is estimated that by the end of 2020, revenues will fall by more than 50 percent, with the number of companies with fixed costs estimated at 9.9 percent. For further information about the coronavirus (COVID-19) pandemic, please visit our dedicated Facts and Figures page.

BLEU: 27.55 **BLEURT-20**: 0.4 **PARENT**: 0.32
ChaTS-Critic (Ours): 0.0 = mean(0, 0, 0)

Figure 3: This example from Kantharaj et al. (2022) showcases the limits of reference-based metrics for summary evaluation: (1) the reference text often contains extra information that is not present in the chart which skews the evaluation, and (2) the reference-based metrics can fail at capturing unreferenced but correct sentences. In comparison, CHATS-CRITIC better reflects the human ratings for summary faithfulness.

2 The CHATS-CRITIC metric

As shown in Figure 2, CHATS-CRITIC is composed of a chart de-rendering model that generates the table content of the input chart image, and a table entailment model applied on a sentence level. This motivation stems from the observation that fine-grained evaluations are simpler than full-summary evaluations, mirroring the ease observed in human assessments (Krishna et al., 2023).

Chart de-rendering. To utilize the information in chart, previous works have incorporated a step to transcribe the image across modalities to a data table (Luo et al., 2021; Kantharaj et al., 2022; Liu et al., 2023a). This process of *de-rendering* enables leveraging downstream text model capabilities to process the information, rather than relying on an image model, which is typically only pre-trained on natural images. Similarly, in our work we start with a de-rendering step to extract the table t from an image of a chart C (Liu et al., 2023a).[2]

Sentence level faithfulness score $f(s)$ (interchangeably referred to as CHATS-CRITIC) is a sentence-level score defined as the probability of entailment of a sentence s conditioned on the de-rendered table t. This can be accomplished using a fine-tuned table-specialized model such as TAPEX (Liu et al., 2022) and TAPAS-CS (Eisenschlos et al., 2020), or by prompting an LLM such as PALM-2 (Anil et al., 2023). For the latter case, we can use few-shot examples with chain-of-thought as well as ensemble across K model runs by averaging the binary scores produced in each run, to improve the entailment accuracy, as shown in the example in Figure 2. The sampling produces a similar effect as Monte-Carlo dropout used by Steen et al. (2023).

3 The CHATS-PI pipeline

CHATS-PI, presented in in Figure 1, shorthand for Chart-To-Summary Pipeline , generates a set of candidate summaries in stage 1, then uses CHATS-CRITIC's per sentence scores to repair (stage 2) score the summaries (stage 3) and re-rank them (stage 4). This is done by removing sentences with low entailment scores and picking the candidate summary with the highest *Summary-level faithfulness score*.

Summary-level faithfulness score $F(S)$ is a per summary score defined as the ratio of kept sentences:

$$F(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} \mathbb{1}_{[T,1]} (f(s_i))$$

[2] We also tested end-to-end models using chart images as direct input, but the current de-rendering-based pipeline yielded the best performance.

where $\mathbb{1}_{[T,1]}(f(s_i))$ is the indicator function with T as the threshold, which is equal to 1 if $f(s_i) > T$ and 0 otherwise.

4 Experimental Setup

We assess our methods on diverse datasets to prove their broad applicability.

4.1 Datasets

Chart-to-Text (Kantharaj et al., 2022) is a large-scale benchmark for chart summarization including bar, line, area, scatter and pie charts, composed of two data sources: *Statista* (35k examples) and *Pew Research* (9k).[3]

SciCap (Hsu et al., 2021) is a large-scale benchmark for figure-captioning. It is extracted from science arXiv papers published between 2010 and 2020 and contains more than 2 million figures. We use 3 subsets of SciCap: *First Sentence collection* (133k), *Single-sentence Caption* (94k data points), and *Caption with No More than 100 Words* (131k).

TabFact (Chen et al., 2020) is a large-scale dataset for table-based fact verification. It contains 16k Wikipedia tables as evidence for 118k human-annotated statements. This dataset allows us to study fact verification with semi-structured inputs. We use it to evaluate the entailment accuracy of CHATS-CRITIC.

All our models are developed on the dev sets of the mentioned benchmarks and performances are reported on their test sets. We include more detailed descriptions and processing details of the benchmarks in Appendix A.1.

4.2 Setups for evaluation & comparison

Evaluating CHATS-CRITIC. We evaluate the quality of CHATS-CRITIC by comparing the model output entailment to human annotated examples randomly extracted from the Chart-To-Text (Statista). We also evaluate the metric's correlation with human judgments on summary level. We compare CHATS-CRITIC to reference-based metrics, including BLEU (Papineni et al., 2002), PARENT (Dhingra et al., 2019) that takes the table into account to compute n-gram similarity and as well as BLEURT-20 (Sellam et al., 2020; Pu et al., 2021), a learned metric.

Evaluating CHATS-PI. We report a wide range of metrics' scores across the three benchmarks. We compare CHATS-PI applied on different base models, as well as state-of-the-art baselines in the literature which do not rely on CHATS-PI where applicable. The SOTA baselines include PaLI (Chen et al., 2023) and MATCHA (Liu et al., 2023b) MATCHA on Chart-To-Text; M4C-Captioner (Horawalavithana et al., 2023) on SciCap. We additionally train and evaluate PaLI (Chen et al., 2023) ourselves to report more comprehensive results across different benchmarks and metrics. All metrics are reported on the sentence level except for the correlation study.

4.3 Our models

Plot-to-table model. As described, our approach relies on a plot-to-table translation model. For all our models, we make use of DePlot (Liu et al., 2023a), a state-of-the-art model for extracting table contents from chart images (i.e. chart de-rendering).[4] The de-rendered table is passed to a generative text-to-text model for further processing.

Generative models. We use two models for summary generation with the de-derendered table from last step as input. We adapt a FLAN-T5 (Suresh et al., 2023) base model with table embeddings to enhance table structure understanding, following the scheme of TabT5 (Andrejczuk et al., 2022). We fine-tune this model for each datasets for 220k training steps with a batch size of 128. We denote this setup as DePlot+FLAN-T5 (see Appendix A.4.1). The second approach is PALM-2 (L) (Anil et al., 2023) with in-context learning. The full prompt is described in Appendix A.4.3. We experiment with other models including end-to-end models in Appendix B.4.

[3] statista.com and pewresearch.org

[4] More details about DePlot can be found in Appendix A.2.

ChaTS-Critic is used for the ChaTS-Pi pipeline and as an additional metric in our experiments. We experiment with different model sizes and families for ChaTS-Critic's entailment component. When not specified, ChaTS-Critic uses DePlot and PaLM-2 (L) with Chain-of-thought (Wei et al., 2022) for the entailment model (shown in Figure 2). The full prompt is reported in Appendix A.4.4.

5 Results

5.1 Meta evaluation of ChaTS-Critic

ChaTS-Critic is evaluated by assessing its correlation with human ratings and the overall quality of the generated summaries. We randomly sampled 60 different charts from Chart-To-Text (Statista) test set and surveyed the *Entailment*, *Relevance*, and *Grammaticality* (see Appendix B.1) on the sentence and summary level when appropriate, making a multidimensional quality metric (Huang et al., 2020). The provenance of the summaries is hidden to prevent biasing the raters. The raters are 10 volunteering researchers from our institution (not including the authors). We refined the guidelines with a small sample of examples and raters before formally starting the survey. The Cohen's Kappa in this sample between pairs of raters is 0.61, which suggests substantial agreement (Landis and Koch, 1977). In the formal survey, the raters annotated the full set, one rater per example. As shown in Figure 7 in Appendix B.1, we display the chart alongside the title, then for each sentence we ask the rater if is (1) *entailed*, (2) *relevant*, and (3) *grammatically correct*. The full annotation guidelines are reported in Appendix B.2. We collect annotations for four data collections presented in table 1. In the two collections using ChaTS-Pi, the predictions are generated without dropping the unsupported sentences, to allow a thorough analysis of ChaTS-Critic quality.

Human annotation set (60 charts)	Avg # sentences	Avg # not entailed
Reference	2.5	1.0
PALM-2	4.3	1.1
🍰(PALM-2, 🕵(PALM-2))	5.4	1.0
🍰(PALM-2, 🕵(DePlot, PALM-2))	5.0	0.5

Table 1: Human annotation data collections sizes. 🍰: ChaTS-Pi generates a set of 10 candidates using PALM-2, and 🕵: ChaTS-Critic considers either the original tables or the generated ones using DePLot.

Entailment performance. We compare ChaTS-Critic against a no-op baseline $f(x) = 1$, where no sentences are filtered, as binary classifiers acting on each sentence. As such, we report Accuracy, F1 and AUC in Table 2. AUC measures the ability of a binary classifier to distinguish between the classes $\{0, 1\}$. When no sentences are dropped, no classifier is used, the AUC is equal to 50%. In this use case we can focus on F1 results to evaluate the quality of the input sentences. Precision and recall are reported to showcase the trade-off of ChaTS-Critic at a threshold of 0.75. We show that ChaTS-Critic significantly improves upon all the metrics reaching better Precision-Recall trade-off.

The reference summaries in SciCap are extracted automatically, implying that extra information might be present that cannot directly be deduced from the provided chart and metadata alone. As expected, the F1 score is low when considering all sentences entailed (i.e. baseline $f(x) = 1$). Our proposed metric improves F1 by 11 points and increases AUC by 31.5 points. For the three other datasets, the summaries' quality is already better than the reference. Thus, the gain is less significant: by 1 to 2 points for F1 and 20 to 22 for AUC.

We report the Pearson coefficient and the p-value in Table 2. For all the sets, the p-value is significantly small, indicating a high probability of observing a correlation to human ratings. The Pearson coefficient indicates that ChaTS-Critic has a human rating correlation from moderate (> 30) to strong (> 50).

Impact of critic model size. We compare in Table 3 different LLMs to implement the entailment component of ChaTS-Critic. We evaluate the performance of the models using the SciCap reference human annotation set and DePlot as a de-rendering model. We additionally study the entailment quality factoring out the de-rendering step by providing the original gold tables in SciCap and TabFact datasets.

Annotation set	Sentence Selection	Acc.	Balanced Acc.	Recall	Precision	F1	AUC	Pearson (p-value)
Reference	$f(x) = 1$	60.0	50.0	100.0	60.0	75.0	50.0	$--(--)$
	🕵	82.0	79.45	92.22	80.58	**86.01**	**81.56**	$62.2(1.9e-17)$
PALM-2	$f(x) = 1$	75.48	50.0	100.0	75.48	86.03	50.0	$--(--)$
	🕵	81.23	71.24	90.86	85.24	**87.96**	**70.54**	$45.15(1.6e-14)$
🍰(PALM-2, 🕵(PALM-2))	$f(x) = 1$	83.33	50.0	100.0	83.33	90.91	50.0	$--(--)$
	🕵	84.49	68.93	93.9	87.83	**91.81**	**70.6**	$43.6(1.7e-15)$
🍰(PALM-2, 🕵(DePlot, PALM-2))	$f(x) = 1$	89.4	50.0	100.0	89.4	94.41	50.0	$--(--)$
	🕵	92.38	66.76	99.26	92.73	**95.89**	**72.53**	$51.01(2.1e-21)$

Table 2: Evaluating CHATS-CRITIC (🕵) with a 0.75 threshold against human labels on Chart-To-Text, we contrast with a no-op baseline ($f(x) = 1$) and report sentence binary classifier metrics. For CHATS-PI (🍰), we generate 10 candidates at temperature $T = 0.7$, and CHATS-CRITIC (🕵) is computed with $T = 0.3$ over 8 samples.

Dataset	Sentence Selection metric	Accuracy	F1	AUC
Statista Reference	$f(x) = 1$	60.0	75.0	50.0
	🕵(DePlot, PALM-2(S))	70.67	76.6	71.72
	🕵(DePlot, PALM-2(L))	**82.0**	**86.01**	**81.56**
	🕵(PALM-2(S))	72.67	77.09	72.08
	🕵(PALM-2(L))	**81.33**	**86.0**	**81.67**
TabFact	$f(x) = 1$	50.32	66.95	50.0
	🕵(PALM-2(S))	81.37	79.45	81.93
	🕵(PALM-2(L))	**87.19**	**87.23**	**87.19**
	Eisenschlos et al. (2020)	81.0	-	-
	Liu et al. (2022)	84.2	-	-

Table 3: Comparing different critic models for CHATS-CRITIC accuracy. For the L model, we use a threshold of 0.75 for SciCap and 0.5 for TabFact. For the S model, we use a threshold of 0.5 for both sets. We also include two state-of-the-art finetuned models on the TabFact training set for reference. We experimented with applying these trained models to Statista as well but the accuracy was very low due to poor generalization.

As shown in the table, model size is a critical factor to improve CHATS-CRITIC overall quality. In SciCap using DePlot respectively gold tables, we see a 10.6 respectively 12.6 points increase on accuracy by using the small model compared to selecting all sentences (f(x)=1) and 11.3 respectively 8.6 increase when switching from small to large models. We observe the same behavior in TabFact with 4.2 increase from small to large.

5.2 Metrics correlation to human ratings

We investigate the correlation of the reference-based metrics to human ratings and compare it to CHATS-CRITIC. Since these metrics are applied on the summary level, we extract the human entailment rating per summary: if any sentence is not entailed, the entire summary is refuted. PARENT (Dhingra et al., 2019) uses also the table information on top of the summary. To thoroughly assess CHATS-CRITIC, we report the correlation on summary level.

Additionally, we study the p-value and Pearson coefficient in Table 4. To observe a possible correlation, reference-based metrics require optimizing for the entailment threshold (reported in the Appendix B.3.1). Even accounting for that aspect, most of the reference-based metrics fail at providing a p-value that is statistically significant to identify a correlation (less than 0.05). The majority of the metrics have a Pearson coefficient lower than 0.30, indicating a small correlation. However, these metrics are less reliable than CHATS-CRITIC, as these values are obtained by optimizing the threshold and the curve is not smooth; a deviation of 0.1 in the threshold reduces the Pearson coefficient dramatically and increases the p-value. The results reported in the table, further confirm that our metric is more reliable and has a higher correlation with respect to the reference-based metrics. We additionally report the precision and recall curves for all metrics in Appendix B.3.2.

5.3 Evaluation of the CHATS-PI pipeline

In the second experimental setup, we compare in Table 5 different models to solve the chart-to-summary task on three data collections. We show that adding CHATS-PI improves any of the presented generative models on CHATS-CRITIC. Additionally, it increases BLEURT-20 by around 1 point for all the data collections. The best generative model is PALM-2. CHATS-PI (PALM-2) consistently reaches between 93% and 96% of CHATS-CRITIC. For more details, models and metrics results see Appendix B.4.

Data collection	ChaTS-Critic$_{summary}$	BLEURT-20	BLEU	PARENT
Reference	$62.4(1.6e-07)$	$17.03(2.0e-1)$	*nan*	$24.2(6.7e-02)$
PALM-2	$40.56(1.5e-03)$	$22.42(9.0e-02)$	$14.92(2.6e-01)$	$-29.57(2.4e-02)$
🍰(PALM-2, 🕵(PALM-2))	$50.27(5.7e-05)$	$25.59(5.2e-02)$	$34.16(8.6e-03)$	$28.33(3.1e-02)$
🍰(PALM-2, 🕵(DePlot, PALM-2))	$51.97(6.9e-04)$	$45.68(3.4e-03)$	$26.92(9.7e-01)$	$23.62(1.4e-01)$

Table 4: Comparing models on ChaTS-Critic performance (i.e. 🕵 column), instantiated with PALM-2 and DePlot for chart de-rendering. ChaTS-Pi (i.e. rows with 🍰) uses ChaTS-Critic configured in the same way. We report SciCap (First sentence) split for brevity. Full results and additional evaluations are in Table 10.

Dataset	Model	🕵	BLEURT	BLEU
Statista	Chen et al. (2023) PaLI-17B (res. 588)	0.49	0.49	40.95
	DePlot+FLAN-T5	0.66	0.55	42.5
	PALM-2	0.89	0.44	14.8
	🍰 (DePlot+FLAN-T5)	0.76	**0.57**	**43.1**
	🍰 (PALM-2)	**0.96**	0.45	13.34
Pew	Liu et al. (2023b) MATCHA	–	–	12.2
	Chen et al. (2023) PaLI-17B (res. 588)	0.35	0.49	13.93
	DePlot+FLAN-T5	0.33	**0.5**	**15.33**
	PALM-2	0.87	0.47	8.83
	🍰 (DePlot+FLAN-T5)	0.41	**0.5**	15.09
	🍰 (PALM-2)	**0.95**	0.48	9.18
SciCap (First sentence)	Horawalavithana et al. (2023) M4C-Captioner	–	–	6.4
	Chen et al. (2023) PaLI-17B (res. 588)	0.41	0.3	11.05
	DePlot+FLAN-T5	0.34	0.29	15.12
	PALM-2	0.84	0.3	0.94
	🍰 (DePlot+FLAN-T5)	0.48	0.3	**15.53**
	🍰 (PALM-2)	**0.93**	**0.31**	0.76

Table 5: Comparing models on ChaTS-Critic performance (i.e. 🕵 column), instantiated with PALM-2 and DePlot for chart de-rendering. ChaTS-Pi (i.e. rows with 🍰) uses ChaTS-Critic configured in the same way. We report SciCap (First sentence) split for brevity. Full results and additional evaluations are in Table 10.

6 Analysis

Ablation study (ChaTS-Pi 4 stages) We report a performance study of the four stages of ChaTS-Pi, as depicted in Figure 1, in Table 6. Droppings sentences in Stage 2 increases F1 by 1.9 points compared to Stage 1. Ranking with ChaTS-Critic without repair shows 8.3 points compared to Stage 1 and 6.4 to Stage 2. Dropping the sentences of the top ranked summary increase F1 by 1.3 reaching 95.69% compared to using the top ranked summary.

Stage name		F1	AUC
S1	Summary generation	86.03	50.0
S2	Drop unentailed sentences	87.96	70.54
S3	Summary scoring	94.41	50.0
S4	↪ Filtering	**95.69**	**72.53**

Table 6: Binary classification F1 and AUC of ChaTS-Pi's different stages, on the PALM-2 annotation set. For an overview of the stages refer to Figure 1.

We ablated the impact of DePlot on ChaTS-Critic, using the original tables as a baseline. The findings are detailed in Table 7. Given that DePlot's extracted tables may include missing or inaccurate data, we anticipated a greater sentence drop in ChaTS-Critic with DePlot. Contrarily, the F1 remains consistent for the reference and even sees an increase in ChaTS-Pi sets. Upon examining specific instances, we discerned the primary reason as following: Some numbers in gold tables are "overly precise" (sometimes several digits after the decimal, making it hard for humans to distinguish). In contrast, DePlot always outputs a "rounded"/lossy value, which is preferred by human raters over those using the ultra-precise numbers from the gold table. Despite these observations, the overall difference remains marginal (less than 1 percentage point). This suggests that DePlot's performance is commendably accurate, even when juxtaposed with gold tables.

Annotation set	Table	F1	AUC
Reference	Gold	86.0	**81.67**
	DePlot	**86.01**	81.56
PALM-2	Gold	**88.19**	**71.97**
	DePlot	87.96	70.54
🍰(PALM-2, 🕵(PALM-2))	Gold	91.52	70.14
	DePlot	**91.81**	**70.6**
🍰(PALM-2, 🕵(DePlot, PALM-2))	Gold	95.19	**78.23**
	DePlot	**95.89**	72.53

Table 7: We compare performance of ChaTS-Critic when using a deplotter (DePlot) vs. using gold tables.

Grammaticality defined as the human ratings on grammatical errors (see Section 5.1) on non dropped sentences and summaries is reported in table 8. When applying ChaTS-Critic, we see a constant sentence-level *Grammaticality* for the ChaTS-Pi last stage –The quality is already at 98.6%, leaving little room for improvement– and a consistent improvement over all other sets. As for summary-level *Grammaticality (S)*, the story is more nuanced. On the Reference set (i.e. $\sim$ 3 sentences per summary), the impact on *Grammaticality (S)* is less prominent. On the PALM-2 annotation set, which features longer and more complex highlights (i.e. $\sim$ 5 sentences per summary), we can see a small drop of -1.97%. CHATS-PI last stage remains constant, showing the importance of ranking.

Relevance defined as the percentage of relevant sentences among the selected ones is reported in Table 8. We see a performance drop on this metric, mainly due to the design of CHATS-CRITIC. The relevant sentences usually feature a more complex structure. CHATS-CRITIC tends to prioritize less complex sentences during the entailment verification stage, thus producing an overall drop in *Relevance*. One such example occurs when multiple statistics and computations are included in a single sentence: *"The nonstore retailers increased by 22.8% in April 2020 compared to April 2019, whereas store retailers jumped 12% points in the same period."*. This is the case for the Reference and the CHATS-PI ranking stage.

Annotation Set	**Gram.**	**Gram. (S)**	**Relevance**
Reference	84.0	82.76	**43.33**
Drop unentailed sentences	**88.29**	**83.93**	41.44
PALM-2 Summary generation	88.12	87.93	68.97
Drop unentailed sentences	90.0	85.96	**69.09**
Summary scoring	**98.68**	**92.98**	58.94
↪ Filtering	98.61	**92.98**	57.14

Table 8: Human annotation rates for grammaticality and relevance computed on non dropped sentences. *Grammaticality* (Gram.) is the % of grammatically correct sentences; *Grammaticality (S)* the % of fully grammatically correct summaries; *Relevance* the percentage of relevant sentences. We report the reference, and different CHATS-PI stages using PALM-2. CHATS-CRITIC provides general improvements in sentence level grammaticality, whereas the performance on relevance and summary level grammaticality are mixed, due to CHATS-CRITIC design.

6.1 Multilingual generalization

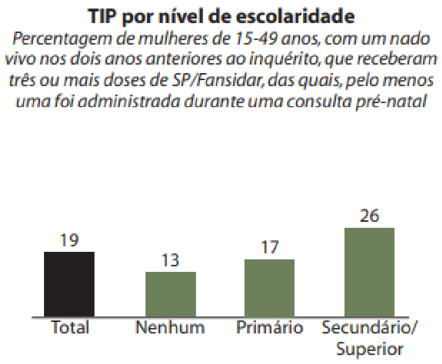
To evaluate the feasibility of our approach on *internationalization* (i18n) datasets, we investigated a controlled setting using the TATA dataset (Gehrmann et al., 2022), a multilingual table-to-text dataset. We selected a localized image per language from the test set, including Portuguese, Arabic, French, Russian, and Swahili. Regrettably, the test set did not include images in Hausa, Yoruba, or Igbo. Figures 4 and 9 include the images alongside the generated summaries.

Generally, the approach can provide an accurate summary if the deplotting component satisfies the following conditions: (a) no errors are introduced, or (b) a reasonable level of detail is provided. For Portuguese, the approach encounters failure case (b), as the deplotted table contains limited information (e.g., the subtext is not considered), resulting in a generic and factually deficient summary. Conversely, the error case (a) is observed for Arabic, where the information is extracted in an incorrect order, leading to factual inaccuracies.

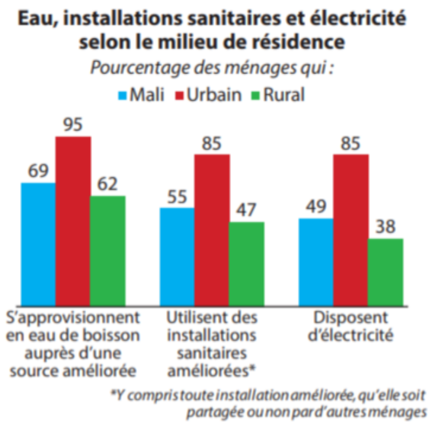
When the table extraction is executed without errors, the summary generated is also accurate and reliable.However, it is acknowledged that there are specific instances where potential stylistic errors may manifest within the summary. Such errors may include the reiteration of phrasal structures at the initiation of sentences or the potentially inappropriate integration of code-mixed information fragments.

7 Related work

Limits of reference-based metrics have been explored for text summarization tasks and different solutions have been proposed to address them. Scialom et al. (2019, 2021); Fabbri et al. (2022) use question generation pipelines to evaluate the faithfulness of summaries without relying on references. Anderson et al. (2016) study the use of intermediate dependency parse trees to rate natural image captions. For semi-structured data such as table-to-summary generation, PARENT (Dhingra et al., 2019) demonstrates the limitation of BLEU as it do not highlight the key knowledge from the table, whereas Opitz and Frank (2021) focuses on generating text from abstract meaning representations. Gehrmann et al. (2022) observed the poor correlation of BLEURT-20 to human ratings and proposed STATA, a learned



✗ SUMMARY: This statistic shows the percentage of women by education level . The percentage of women with a secondary or higher education is 26% . The percentage of women with no education is 13% . The percentage of women with primary education is 17% .



✓ SUMMARY: The statistics show the access to water, sanitation and electricity in Mali, by place of residence. In Mali, the access to improved drinking water sources is higher in urban areas (95%) than in rural areas (62%). The access to improved sanitation facilities is also higher in urban areas (85%) than in rural areas (47%). The access to electricity is higher in urban areas (85%) than in rural areas (38%). Overall, the access to basic services is lower in rural areas than in urban areas in Mali.

Figure 4: Examples from TATA with CHATS-PI summaries using the demo and publicly accessible models.

metric using human annotation.

In this work we explore building a reference-free metric for chart summary that does not require human-annotated references. We show that our metric CHATS-CRITIC has much higher correlation with human judgment than reference-based metrics such as BLEU.

Chart-to-summary generation has recently gained relevance within multimodal NLP. Obeid and Hoque (2020) created the Chart-to-Text dataset, using charts extracted from Statista. Kantharaj et al. (2022) extended it with more examples from Statista and Pew Research. Besides efforts on evaluation, multiple modeling methods have been proposed to reduce hallucinations. The approaches can be roughly divided into (1) pipeline-based methods which first extract chart components (e.g. data, title, axis, etc.) using OCR then leverage text-based models to further summarize the extracted information Kantharaj et al. (2022); Choi et al. (2019); (2) end-to-end models which directly input chart-attribute embeddings to Transformer-based models for enabling structured understanding of charts (Obeid and Hoque, 2020).

In this work we explored both (1) and (2). The best approach CHATS-PI generally follows the idea of (1). Instead of relying on OCR we use a de-rendering model for extracting structured information in charts and we explore a self-critiquing pipeline with LLMs for the best quality chart summarization.

8 Conclusion

In this paper, we tackle the chart-to-summary multimodal task, which has traditionally been challenging since it requires factual extraction and summarization of the insights presented in the image. To measure the quality of a summary (especially faithfulness which has been overlooked by previous metrics), we present a reference-free metric called CHATS-CRITIC for accurately and factually scoring chart-to-summary generation. CHATS-CRITIC obtains substantially higher correlations with human ratings compared to prior reference-based metrics. We additionally present CHATS-PI, a self-critiquing pipeline to improve chart-to-summary generation. CHATS-PI leverages CHATS-CRITIC scores to refine the output of any model by dropping unsupported sentences from the generated summaries and selecting the summary that maximizes fluency and CHATS-CRITIC'scores. Compared with state-of-the-art baselines, CHATS-PI demonstrates stronger summarization quality across the board, achieving better scores for both the CHATS-CRITIC which stresses faithfulness and also traditional metrics such as BLEURT.

Limitations

In the following, we outline the limitations of our work to ensure transparency and inspire future research. First, the chart domains we experimented with is limited to a few popular websites (e.g. Statista and Pew). This is due to the fact that existing academic chart-to-summary datasets only cover limited domains. However, to comprehensively evaluate the effectiveness of CHATS-CRITIC and CHATS-PI, it is desirable to also evaluate our approaches in other chart domains such as infographics and scientific/financial charts. Second, the CHATS-CRITIC depends on a deplotter (image-to-text) model, specifically DePlot (Liu et al., 2023a). DePlot has been trained on similar domains as the chart-to-summary datasets used in this work (e.g. Statista), and its performance may not generalize to other domains. The SciCap dataset evaluated in Table 5 provides some evidence of generalization. In future work, we plan to build out-of-domain evaluations to understand the impact of the deplotter's robustness better. Third, we focused mostly on English chart summary in this work. We plan to also expand multilingual chart summary in future works and expanding our evaluation on the TaTa dataset (Gehrmann et al., 2022) as a test bed.

Another potential limitation is the use of CHATS-CRITIC to evaluate CHATS-PI, although it is important to note that in order to mitigate the possibility for bias, we have cross checked with a variety of evaluation metrics and conducted ablation studies to analyze the individual components of our proposed solution. CHATS-PI has proven to also provide significant benefits when employing standard reference-based metrics. Comprehensive human studies have corroborated the usefulness of our results and we can confidently say that the improvement is consistent across the board.

We would also like to highlight the underlying risk of blindly trusting models to summarize content from an image accurately. Special care should be taken to verify outputs in accuracy-sensitive applications.

Despite its limitations, our work serves as an initial step in constructing reliable chart summarization evaluations and models. We hope future research can greatly benefit from this starting point.

Acknowledgments

We would like to thank Massimo Nicosia, Srini Narayanan, Yasemin Altun, Averi Nowak, Chenxi Pang, Jonas Pfeiffer, Mubashara Akhtar, Parag Jain, and the anonymous reviewers for their time, constructive feedback, useful comments and suggestions about this work.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

Ewa Andrejczuk, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yasemin Altun. 2022. Table-to-text generation and pre-training with TabT5. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6758–6766, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Ron Ellis Ryan Holbrook Benji Andrews, Hema Natarajan. 2023. Benetech - making graphs accessible.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. 2019. Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization. *Computer Graphics Forum*.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 143–152, New York, NY, USA. Association for Computing Machinery.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. 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.

Sebastian Gehrmann, Sebastian Ruder, Vitaly Nikolaev, Jan A Botha, Michael Chavinda, Ankur Parikh, and Clara Rivera. 2022. Tata: A multilingual table-to-text dataset for african languages. *arXiv preprint arXiv:2211.00142*.

Sameera Horawalavithana, Sai Munikoti, Ian Stewart, and Henry Kvinge. 2023. Scitune: Aligning large language models with scientific multimodal instructions. *arXiv preprint arXiv:2307.01139*.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

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.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. TAPEX: Table pre-training via learning a neural SQL executor. In *International Conference on Learning Representations*.

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

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.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9194–9213, Toronto, Canada. Association for Computational Linguistics.

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

Juri Opitz and Anette Frank. 2021. Towards a decomposable metric for explainable evaluation of text generation from AMR. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1504–1518, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. 2016. Figureseer: Parsing result-figures in research papers. In *Computer Vision – ECCV 2016*, pages 664–680. Springer.

Julius Steen, Juri Opitz, Anette Frank, and Katja Markert. 2023. With a little push, NLI models can robustly and efficiently predict faithfulness. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 914–924, Toronto, Canada. Association for Computational Linguistics.

Siddharth Suresh, Kushin Mukherjee, and Timothy T. Rogers. 2023. Semantic feature verification in FLAN-t5. *ICLR 2023 TinyPapers*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jiawen Zhu, Jinye Ran, Roy Ka-Wei Lee, Zhi Li, and Kenny Choo. 2021. AutoChart: A dataset for chart-to-text generation task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1636–1644, Held Online. INCOMA Ltd.

Appendix

A Experimental Setup

A.1 Datasets

We use two popular chat-to-summary datasets for our experiments. The first one is Chart-to-Text (Kantharaj et al., 2022), which can be found in https://github.com/JasonObeid/Chart2Text. The second one is SciCap (Hsu et al., 2021), which is available at https://github.com/tingyaohsu/SciCap. More details about the two datasets are introduced below.

Chart-To-Text has mainly two sources: (i) Statista and (ii) Pew Research. (i) Statista is automatically extracted from an online platform that publishes charts in different topics including economics, market, and opinion; it is composed of 34,811 table, charts and summary triplets. (ii) Pew is automatically extracted then manually annotated from data-driven articles about social issues, public opinion and demographic trends; it is composed of 9,285 chart summary pairs.

SciCap (Hsu et al., 2021) is a large-scale benchmark for figure-captioning. It is extracted from science arXiv papers published between 2010 and 2020 and contains more than 2 million figures. The figure-caption pairs are extracted using PDFFigures 2.0 (Clark and Divvala, 2016), then an automatic figure type classifier is used to select graph plots. To be comparable to the work of Hsu et al. (2021), we evaluate our model on the three subsets containing no sub-figures: *First Sentence collection* including $133,543$ figures, *Single-Sentence Caption* collection containing $94,110$ figures and *Caption with No More than* 100 *Words* composed of $131,319$ figures.

A.2 De-rendering

We use DePlot (Liu et al., 2023a) model in all our experiments. The model code and checkpoint are available at https://github.com/google-research/google-research/tree/master/deplot. We use the GCS path to the base model gs://deplot/models/base/deplot/v1 fine-tuned to solve the chart-to-table task. We do not perform any additional training, and use the model as a pre-processing step to extract the tables from the chart.

A.3 Baselines

We report the state-of-the-art models BLEU scores as presented in their papers. To be able to compare their models to ours and compute our new metric, we fine-tune a PaLI (Chen et al., 2023) model that gives a comparable results in BLEU as the other models. We select PaLI (Pathways Language and Image model) as our method of choice, because it takes the image as input directly, without the need for pre-processing or any OCR model to extract metadata, which can be difficult to reproduce. In our experiments, we use the larger $17B$ variant and fine-tune for $5k$ iterations with an image resolution of 588×588. The PaLI model is fine-tuned with 128 GCP-TPUv4. We use a batch size of 256 and max sequence length of 128.

A.4 Our models

A.4.1 DePlot+T5 and DePlot+Flan-T5

We adapt T5 (Raffel et al., 2020) and FLAN-T5 (Suresh et al., 2023) models: T5 is available at https://huggingface.co/t5-base and FLAN-T5 is available at https://huggingface.co/google/flan-t5-base. We adapt both base models to the chart-to-summary task. We add a de-rendering model to extract the table form the chart and use it as input of the models. Additionally, table embeddings are added to enhance table structure understanding. We fine-tune both models for $220k$ with 16 GCP-TPUv3 cores using a batch size of 128 and a max sequence length of 128.

A.4.2 MatCha-DePLot+FLAN-T5

We use in our experiments MatCha-DePlot+FLAN-T5, which is composed of a MatCha (Liu et al., 2023b) image understanding module coupled to a DePlot+FLAN-T5 model, both of which are base size. MatCha base is available at https://github.com/google-research/google-research/tree/master/deplot. This model takes in input both the a chart image and its table content (i.e. obtained by invoking DePlot). This setup should allow capturing visual aspects that DePlot ignores in its de-rendering process. MatCha-DePlot+FLAN-T5 is fine-tuned for $220k$ training steps with 32 GCP-TPUv3, 128 batch size, 1024 image length and a max sequence length of 128.

A.4.3 PALM-2

In our experiments for summary generation we use PALM-2(L) (Anil et al., 2023) with in-context learning. The prompt is displayed in Figure 5.

A.4.4 Critic model for CHATS-CRITIC

We use PALM-2 (Anil et al., 2023) as a critic model for CHATS-CRITIC. Prompting is crucial for the interpretability of the entailment results. PaLM-2 outputs a text to refute or entail the claim. Following Wei et al. (2022), we use Chain-of-thought prompting to emphasize the reasoning before making the decision on the claim. More precisely we use 2 shots prompting for the critic models as shown in Figure 6. We use the same prompting for the large and small PALM-2 models. The small model is available at https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text?hl=en.

B Results

B.1 Annotation framework

Figure 7 contains a screenshot of the annotation framework used to collect human ratings.

B.2 Annotation guidelines

We provided to the raters the following annotation guidelines:

1. **is_interesting** highlights an important insight from the chart such as min / max / avg value or comparison.

   The **title** copy is **not interesting**.

   If the sentence is **not entailed or grammatically not correct** but highlights important info please select **is_interesting**.

2. **cleaned_summary_is_grammatically_correct** = grammar and fluency. Here the critic **model drops some sentences**. Please focus on the **fluency of the paragraph**.

3. **Entailed** = you do not need additional info: **using the chart only**, be able to **extract the text**. (look at the chart the table can help you but not considered as ground truth.)

   If the prediction is equal to the title it is entailed you can consider it as not interesting.

   Please make sure that the meaning of the sentence does **not** add **additional info** about the chart.

   Examples:

   (a) The chart is about kids enrolled in kindergarten and nursery. The sentence contains: kids aged from 3 to 5. The title or the chart dose not refer to the age. **This adds a condition on the conducted study not referred in the title or the chart.** We considered it **not entailed**.

   (b) If the sentence contains a **general knowledge** such as definitions:
   - if **you know** that the definition **is correct** select **is_entailed**
   - if **you know** that it is **wrong** or **do not know** please select **not entailed**.

4. **Approximate numbers** is allowed up to **2 digits after the decimal.**

   Example: exact number in the chart between 2000 and 3000.

   - Text_1: "... around 2.51k" is **entailed**.

```
Use a neutral tone and decontextualize information when possible. Use markdown format and bullet points for clarity.

Timeseries or data frame to consider:
the title is L'Oréal S.A. - worldwide revenue by division from 2012 to 2019 ( in million euros )

data =
col : Year | Consumer products | Professional products | L'Oréal Luxe | Active Cosmetics | The Body Shop
row 1 : 2019 | 12748.2 | 3441.9 | 11019.8 | 2663.7 | -
row 2 : 2018 | 12032.2 | 3262.5 | 9367.2 | 2275.5 | -
row 3 : 2017 | 12118.7 | 3350.4 | 8471.7 | 2082.9 | -
row 4 : 2016 | 11993.4 | 3399.7 | 7662.4 | 1860.7 | 920.8
row 5 : 2015 | 11844.2 | 3399.7 | 7230.0 | 1816.3 | 967.2
row 6 : 2014 | 10767.5 | 3032.4 | 6197.9 | 1660.4 | 873.8
row 7 : 2013 | 10873.2 | 2973.8 | 5865.2 | 1576.3 | 835.8
row 8 : 2012 | 10713.2 | 3002.6 | 5568.1 | 1499.2 | 855.3


Describe the data frame or time series. Identify important values such as min, max, sum, peak, bottom, increase, drop and
      ↪ highlight important comparisons, trends and unexpected behaviour, using maximum 5 sentences *in total*. Only
      ↪ reference data points you see.

Answer: This statistic shows L'Oréal 's global revenue from 2012 to 2019 , by division .
In 2016 , the Body Shop division of L'Oréal generated approximately 920.8 million euros in revenue .
Between 2012 and 2019 , the consumer products , the Professional products and L'Oréal Luxe divisions reached the highest
      ↪ values in 2019 .
The reported sum of revenue in 2019 is approximately 30k including the consumer products , the Professional products and L'Oré
      ↪ al Luxe divisions .

------

Use a neutral tone and decontextualize information when possible. Use markdown format and bullet points for clarity.

Timeseries or data frame to consider:
the title is Quarterly average daily rate of hotels in Dallas in 2016 and 2017 ( in U.S. dollars )

data =
col : Quarter | 2016 | 2017
row 1 : Q4 | 164 | -
row 2 : Q3 | 163 | -
row 3 : Q2 | 167 | -
row 4 : Q1 | 169 | 170


Describe the data frame or time series. Identify important values such as min, max, sum, peak, bottom, increase, drop and
      ↪ highlight important comparisons, trends and unexpected behaviour, using maximum 5 sentences *in total*. Only
      ↪ reference data points you see.

Answer: This statistic shows the quarterly average daily rate of hotels in Dallas in 2016 and 2017 .
From Q1 2016 to Q3 2016 the average daily rate of hotels in Dallas in the United States decreased a minimum value of 163 .
The rate started to increase from Q3 2016 .
In the first quarter (Q1) of 2017 , the rate reached the highest value 170 U.S. dollars .

------

Use a neutral tone and decontextualize information when possible. Use markdown format and bullet points for clarity.

Timeseries or data frame to consider:
the title is Cities with the largest number of community gardens per 10,000 residents in the United States in 2019


data =
col : city | Community gardens per 10,000 residents | Affected Community gardens per 10,000 residents
row 1 : Portland, OR | 3.8 | 1.9
row 2 : Madison, WI | 3.1 | 1.3
row 3 : St. Paul, MN | 3.8 | 1.9
row 4 : Orlando, FL | 2.4 | 1.3
row 5 : Washington, DC | 3.2 | 1.3
row 6 : Seattle, WA | 2.1 | 1.2

Describe the data frame or time series. Identify important values such as min, max, sum, peak, bottom, increase, drop and
      ↪ highlight important comparisons, trends and unexpected behaviour, using maximum 5 sentences *in total*. Only
      ↪ reference data points you see.

Answer: The statistics display the Cities with the largest number of community gardens per 10,000 residents in the United
      ↪ States in 2019 .
The statistics show that some cities are home to more community gardens than others .
In 2019 , both Portland , OR and St. Paul , MN had the largest number with 3.8 community gardens per 10,000 residents .
In 2019 , Seattle, WA had the lowest number with 2.1 community gardens per 10,000 residents .
The average values of Community gardens per 10,000 residents is 3.06 .
```

Figure 5: PALM 3-shots prompting for summary generation

```
Read the table below regarding "Cities with the largest number of community gardens per 10,000 residents in the United States
      ↪ in 2019" to verify whether the provided claims are true or false.
col : city | Community gardens per 10,000 residents | Affected Community gardens per 10,000 residents
row 1 : Portland, OR | 3.8 | 1.9
row 2 : Madison, WI | 3.1 | 1.3
row 3 : St. Paul, MN | 3.8 | 1.9
row 4 : Orlando, FL | 2.4 | 1.3
row 5 : Washington, DC | 3.2 | 1.3
row 6 : Seattle, WA | 2.1 | 1.2

Claim: The statistics display the Cities with the largest number of community gardens per 10,000 residents in the United
      ↪ States in 2019 .
Answer: The title of the statistics is about the Cities with the largest number of community gardens per 10,000 residents in
      ↪ the United States in 2019 . Therefore, the claim is true .

Claim: The term 'community garden ' in the United States can mean a few different things .
Answer: The definition of community garden is not mentioned by the statistics . Therefore, the claim is false .

Claim:  For example , they can function as gathering places for the community and/or neighbors , however , they can also
      ↪ resemble the allotment gardens , often found in Europe , used by individuals and families .
Answer: The example is not mentioned by the statistics . Therefore, the claim is false .

Claim: Of course , some cities are home to more community gardens than others .
Answer: The statistics show that cities have different community gardens values . Therefore, the claim is true .

Claim: In 2019 , both Portland , OR and St. Paul , MN had the largest number with 3.8 community gardens per 10,000 residents .
Answer: Community gardens values are 3.8 > 3.2 > 3.1 > 2.4 > 2.1 . Portland , OR  Community gardens is 3.8 and St. Paul , MN
      ↪ Community gardens is 3.8 . Community gardens per 10,000 residents highest number is 3.8 . Therefore, the claim is
      ↪ true .

Claim: In 2019 , Portland , OR and St. Paul , MN had the largest number with 1.9 community gardens per 10,000 residents .
Answer: Portland Community gardens is 3.8, OR and St. Paul, MN  Community gardens is 3.8 . 3.8 is not 1.9 . Therefore, the
      ↪ claim is false .

Claim: In 2019 , Washington, DC had the largest number .
Answer: Community gardens values are 3.8 > 3.2 > 3.1 > 2.4 > 2.1 . Washington, DC Community gardens is 3.2 . Both Portland ,
      ↪ OR and St. Paul have the highest number 3.8 . Therefore, the claim is false .

Claim: In 2019  Portland, OR and Washington, DC had the largest number .
Answer: Community gardens values are 3.8 > 3.2 > 3.1 > 2.4 > 2.1 . Portland, OR Community gardens is 3.8 . St. Paul, MN is 3.8
      ↪  . Washington, DC Community gardens is 3.2 . Both Portland , OR and St. Paul have the highest number 3.8 followed by
      ↪ Washington, DC 3.2 . Therefore, the claim is false .

Claim: In 2019  Washington, DC,  Portland, OR and St. Paul, MN had the largest number .
Answer: Community gardens values are 3.8 > 3.2 > 3.1 > 2.4 > 2.1 . Washington, DC Community gardens is 3.2 . Portland, OR
      ↪ Community gardens is 3.8 . St. Paul, MN is 3.8 . Both Portland , OR and St. Paul have the highest number 3.8 followed
      ↪  by Washington, DC 3.2 . Therefore, the claim is true .

Claim: In 2019 , Portland, OR had the largest number .
Answer:  Portland, OR Community gardens is 3.8 . St. Paul, MN Community gardens is also 3.8 . Both Portland, OR and St. Paul,
      ↪ had the largest number . Therefore, the claim is false .

Claim: In 2019 , Portland , OR and St. Paul , MN had the lowes number of affected Community gardens per 10,000 residents .
Answer: Affected Community gardens per 10,000 residents values are 1.9 > 1.3 > 1.2 . Both Portland , OR and St. Paul have
      ↪ Affected Community gardens is 1.9 . Community gardens per 10,000 residents lowes number is 1.2 . Therefore, the claim
      ↪  is false .

Claim: the average values of Community gardens per 10,000 residents is 3.06 .
Answer: The sum of Community gardens is 3.8 + 3.1 + 3.8 + 2.4 + 3.2 + 2.1 = 18.400000000000002 . There are 6 cities. The
      ↪ average is 18.40 / 6 = 3.0666666666666664 . Therefore, the claim is true .

Claim: the median values of Community gardens per 10,000 residents is 3.15 .
Answer: The ordered values of Community gardens are 3.8 , 3.8 , 3.2, 3.1 , 2.4 , 2.1 . There are 6 cities. The median refers
      ↪ to the values at positions 2 and 3 (6 / 2) . The median is 3.1 + 3.2  / 2 = 3.15 . Therefore, the claim is true .

------

Read the table below regarding "Number of households in Denmark from 2018 to 2020 by type of households
" to verify whether the provided claims are true or false.

col : Year | Married couple | Single women | Single men | Other couples | Other households including more than 1 family |
      ↪ Children below 18 years not living with parents
row 1 : 2020 | 934630 | 691059 | 527750 | 341985 | 231680 | 1028
row 2 : 2019 | 932591 | 682152 | 518279 | 338268 | 233819 | 962
row 3 : 2018 | 932254 | 676468 | 512552 | 335335 | 230875 | 988

Claim: Between 2018 and 2020 , the amount of households increased from 40702 to roughly 2.7 million .
Answer: The number of Married couple households in 2018 was 932254 + 676468 + 512552 + 335335 + 230875 + 988 = 2688472 and not
      ↪  40702 . Therefore, the claim is false .

Claim: The number of households in Denmark increased by over 10 thousand in the period from 2019 to 2020 Answer .
Answer: The sum number of households in 2020 was 934630 + 691059 + 527750 + 341985 + 231680 + 1028 = 2728132 and in 2019 was
      ↪ 932591 + 682152 + 518279 + 338268 + 233819 + 962 = 2706071. The difference 2728132 - 2706071 = 22061 . 22061 is
      ↪ higher than 10000 . Therefore, the claim is true .

Claim: It reached its peak in 2020 .
Answer: The sum number of households in 2020 was 934630 + 691059 + 527750 + 341985 + 231680 + 1028 = 2728132. The sum number
      ↪ of households in 2019 was 932591 + 682152 + 518279 + 338268 + 233819 + 962 = 2706071. The sum number of households in
      ↪  2018 932254 + 676468 + 512552 + 335335 + 230875 + 988 = 2688472 . 2728132 is higher than 2706071 and 2688472. The
      ↪ highest sum is in 2020. Therefore, the claim is true .

Claim: In 2019 it reached roughly 2.7 million .
Answer: The sum in 2019 is 932591 + 682152 + 518279 + 338268 + 233819 + 962 = 2706071. 2706071 is roughly 2.7 million.
      ↪ Therefore, the claim is false .
```

Figure 6: PALM 2-shots prompting for CHATS-CRITIC.

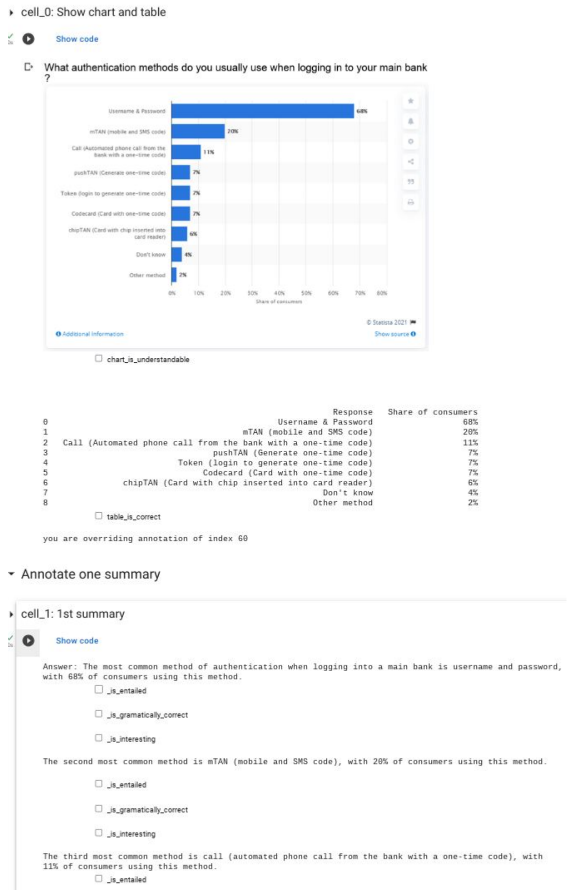


Figure 7: Annotation system example

- Text_2: "... around 2.5123k" is **not entailed**.

5. **grammatically_correct =** look at grammar errors / fluency / repetition. Punctuation only if **it changes the meaning of the sentence**. Small errors are acceptable.

   Example: forget a letter/ invert letters / forget punctuation.

B.3 Correlation to human ratings

B.3.1 Pearson's coefficient and p-value

We extract the summary level human annotation as following; when a summary contains at least one unsupported sentence it is considered as unfaithful. As a result, the human summary annotation is binary while all the summary metrics BLEU, BLEURT-20, ChaTS-Critic are continuous. Without the use of a threshold to binarize them, the results of the p-value and Pearson coefficients are extremely uninformative. We choose the best possible threshold for each of the metrics in order to compare them in a fair way. In other words, we evaluate the classification task by selecting a threshold. In the case of giving 2 binary vectors to compute the Pearson coefficient we have Pearson = Spearman = Phi (standardized Chi-square). Table 9 reports the different thresholds used to measure the p-value and Pearson's coefficient in Table 4.

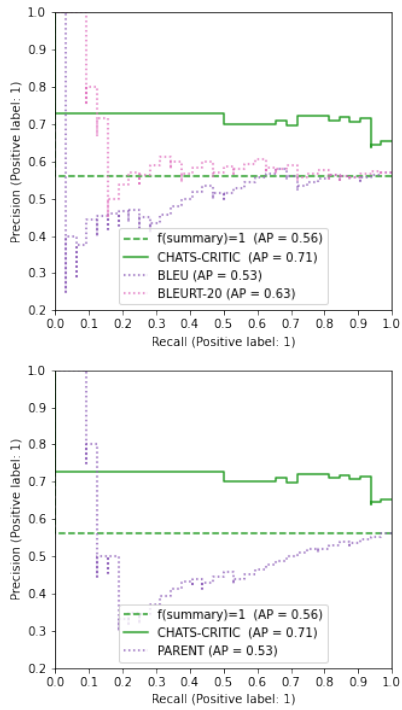


Figure 8: Human evaluation results for summary level to study the correlation of reference based metric to human ratings. Precision and Recall curve is displayed for the different metrics computed on Statista with summaries generated by PALM-2. The reported ChaTS-Critic Precision and Recall refers to the summary level.

B.3.2 Precision and Recall curves

Figure 8 shows the correlation of different metrics with human ratings by reporting Precision and Recall on the predicted summaries generated by PALM-2 compared to the original reference. A good correlation would display a continuously decreasing step function allowing to trade-off between Precision and Recall at a given threshold level. The ChaTS-Critic summary scores curve shows that it is a better classifier compared to all other metrics.

Data collection	CHATS-CRITIC$_{summary}$	BLEURT-20	BLEU	PARENT
Reference	0.9	0.9	nan	0.79
PALM-2	0.9	0.4	0.04	0.3
CHATS-PI (PALM-2, CHATS-CRITIC (PALM-2))	0.9	0.57	0.13	0.16
CHATS-PI (PALM-2, CHATS-CRITIC (DePlot, PALM-2))	0.9	0.37	0.16	0.16

Table 9: The thresholds used to report the values in Table 4 were selected as follows. For all metrics except for CHATS-CRITIC, we looked for the best threshold that maximized first the chance of observing a lower p-value and then a higher person coefficient. A constant threshold was considered for all sets when using CHATS-CRITIC.

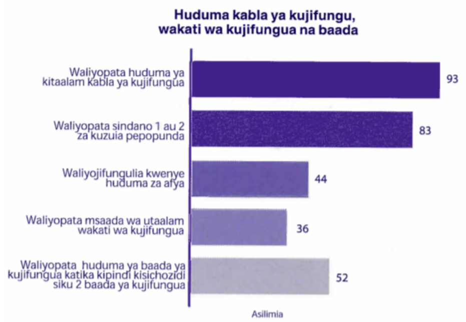
B.4 CHATS-PI pipeline evaluation

We report supplementary experiments and baselines in Table 10, alongside additional metrics. We report CHATS-CRITIC and CHATS-PI using DePlot as a de-rendering model and if the original table is provided we add an extra row to ablate the effect of DePlot. We use the following model checkpoint for BLEURT computation: https://storage.googleapis.com/bleurt-oss-21/BLEURT-20.zip.

B.5 Multilingual generalization

Figure 9 contains additional i18n examples from the TATA dataset (Gehrmann et al., 2022). Aside from a few stylistic quirks and redundancies, both summaries are accurate and faitfhful.

✓ **SUMMARY:** In the survey , 43% of women reported that they and their partner made decisions about money together . 29% of women reported that they made decisions about money on their own . 18% of women reported that their partner made decisions about money on their own . 11% of women reported that someone else made decisions about money .



✓ **SUMMARY:** The statistics display the Huduma kabla ya kujifungua, wakati wa kujifungua na baada . The statistics show that 93% of the people received professional care before giving birth . The statistics show that 83% of the people received 1 or 2 injections to prevent tetanus . The statistics show that 44% of the people gave birth in health services . The statistics show that 36% of the people received expert assistance during childbirth . The statistics show that 52% of the people received postpartum care within 2 days of giving birth .

Figure 9: Additional i18n examples from TATA alongside with the summaries produced by CHATS-PI instantiated as 🍰(Gemini, 🕵(Gemini)).

Dataset	Inputs	Model	ChaTS-Critic	BLEURT-20	BLEU	PARENT
Chart-To-Text (Statista)	original table title	Kantharaj et al. (2022) TAB-T5 + (pretrained-pew)	–	0.15	37.32	–
		TAB-T5	0.55	0.53	40.48	0.16
		TAB-FLAN-T5	0.67	0.56	41.48	0.32
		PALM-2	0.9	0.42	13.2	0.52
		🍰(TAB-T5)	0.67	0.54	41.45	0.16
		🍰(TAB-FLAN-T5)	0.76	**0.56**	**42.52**	0.32
		🍰(PALM-2)	**0.94**	0.45	12.43	**0.65**
	original table title + image	MATCHA-TAB-FLAN-T5	0.68	0.52	38.57	0.2
		🍰(MATCHA-TAB-FLAN-T5)	0.71	0.53	37.94	0.25
	OCR table	Kantharaj et al. (2022) OCR-T5	–	0.10	35.29	–
	image	Liu et al. (2023b) MATCHA	–	–	39.4	–
	image	Chen et al. (2023) PaLI-17B (res. 588)	0.49	0.49	40.95	–
	DePLot table title	DePLot+T5	0.54	0.54	41.83	0.15
		DePLot+FLAN-T5	0.66	0.55	42.5	0.19
		PALM-2	0.89	0.44	14.8	0.32
		🍰(DePLot+T5)	0.66	0.56	42.67	0.15
		🍰(DePLot+FLAN-T5)	0.76	**0.57**	**43.1**	0.15
		🍰(PALM-2)	**0.96**	0.45	13.34	**0.32**
	DePLot table title + image	MATCHA-DePLot+FLAN-T5	0.7	0.54	37.24	0.25
		🍰(MATCHA-DePLot+FLAN-T5)	0.79	0.55	39.24	0.25
Chart-To-Text (Pew)	OCR table	Kantharaj et al. (2022) OCR-T5	–	−0.35	10.49	–
	image	Liu et al. (2023b) MATCHA	–	–	12.2	–
		Chen et al. (2023) PaLI-17B (res. 588)	0.35	0.49	13.93	–
	DePLot table title	DePLot+T5	0.27	0.49	12.06	0.04
		DePLot+FLAN-T5	0.33	0.5	15.33	0.06
		PALM-2	0.87	0.47	8.83	**0.2**
		🍰(DePLot+T5)	0.34	0.5	14.9	0.04
		🍰(DePLot+FLAN-T5)	0.41	0.5	15.09	0.06
		🍰(PALM-2)	**0.95**	0.48	9.18	**0.2**
	DePLot table title	MATCHA-DePLot+FLAN-T5	0.36	**0.51**	**15.41**	0
		🍰(MATCHA-DePLot+FLAN-T5)	0.42	**0.51**	15.27	0
SciCap	SciTune info	Horawalavithana et al. (2023) LLaMA-SciTune (13B,CTOM)	–	–	5	–
		Horawalavithana et al. (2023) M4C-Captioner	–	–	6.4	–
SciCap (First Sentence)	image	Hsu et al. (2021) CNN+LSTM(vision only)	–	–	2.19	–
		Chen et al. (2023) PaLI-17B (res. 588)	0.41	0.3	11.05	–
	DePLot table	DePLot+T5	0.3	0.28	15.27	0
		DePLot+FLAN-T5	0.34	0.29	15.12	0.18
		PALM-2	0.84	0.3	0.94	0.36
		🍰(DePLot+T5)	0.44	0.29	15.2	0.22
		🍰(DePLot+FLAN-T5)	0.48	0.3	**15.53**	0.18
		🍰(PALM-2)	**0.93**	**0.31**	0.76	**0.36**
	DePLot-table image	MATCHA-DePLot+FLAN-T5	0.36	0.29	14.89	0.2
		🍰(MATCHA-DePLot+FLAN-T5)	0.49	0.3	14.19	0.49
SciCap (Single-Sent Caption)	Text	Hsu et al. (2021) CNN+LSTM(Text only)	–	–	2.12	–
	DePLot table	DePLot+T5	0.34	0.28	13.27	0
		DePLot+FLAN-T5	0.38	0.3	15.28	0.37
		PALM-2	0.84	0.31	0.69	0.35
		🍰(DePLot+T5)	0.52	0.3	15.72	0
		🍰(DePLot+FLAN-T5)	0.53	**0.32**	**18**	0.35
		🍰(PALM-2)	**0.93**	**0.32**	0.61	**0.42**
	DePLot-table image	MATCHA-DePLot+FLAN-T5	0.37	0.3	15.33	0.23
		🍰(MATCHA-DePLot+FLAN-T5)	0.51	**0.32**	16.96	0.23
SciCap (Caption w/ <=100 words)	Text	Hsu et al. (2021) CNN+LSTM(Text only)	–	–	1.72	–
	DePLot table	DePLot+T5	0.31	0.28	14.52	0.15
		DePLot+FLAN-T5	0.35	0.29	15.71	0.17
		PALM-2	0.82	0.3	0.81	0.41
		🍰(DePLot+T5)	0.45	0.29	14.2	0.15
		🍰(DePLot+FLAN-T5)	0.48	0.3	15.51	0.17
		🍰(PALM-2)	**0.93**	**0.32**	0.64	0.46
	DePLot-table image	MATCHA-DePLot+FLAN-T5	0.34	0.29	15.90	**0.48**
		🍰(MATCHA-DePLot+FLAN-T5)	0.46	0.30	**16.16**	**0.48**

Table 10: Comparing different models on ChaTS-Critic performance. ChaTS-Critic refers to ChaTS-Critic(PALM-2) using the original table if it is provided in the input, else it refers to ChaTS-Critic(DePLot, PALM-2). Additionally, ChaTS-Pi 🍰 uses ChaTS-Critic following the same logic. The reported numbers for Kantharaj et al. (2022) uses BLEURT-128 base. For the our experiment we use BLEURT-20