Improving Factual Accuracy of Neural Table-to-Text Output by Addressing Input Problems

Anonymous ACL submission

Abstract

Neural Table-to-Text models tend to hallucinate and produce text with factual errors. We worked with a mixture of models (different types and sizes) on the ToTTo dataset to study whether the output factual errors could be traced back to input problems. Outputs from each model were manually annotated (a total of 1,837 texts) for factual errors and the input problems responsible for many of them were identified and fixed. Addressing input problems reduced factual errors by between 52% and 76%, depending on the model. Overall, we observed that models struggled to work with input data that is structured in a non-standard way.

1 Introduction

Table-to-Text generation refers to the task of learning to generate natural language descriptions from tabular data (Parikh et al., 2020; Chen et al., 2020a,b) and is widely used in several application domains such as medical diagnosis (Pauws et al., 2018), financial (Zhu et al., 2023) and weather (Sripada et al., 2002; Gkatzia et al., 2017; Upadhyay and Massie, 2022) reports and sports summaries (Thomson et al., 2020). Neural language models are known to generate fluent texts but may generate outputs that are factually incorrect or unrelated to the provided input data. Such undesirable generation is called 'hallucination' (Wang and Sennrich, 2020; Raunak et al., 2021; Ji et al., 2023).

Previous studies on Table-to-Text tasks adopt traditional seq2seq methods to generate table descriptions (Wiseman et al., 2017; Puduppully et al., 2019; Rebuffel et al., 2019). Recently, transformer based models (Devlin et al., 2019; Raffel et al., 2020; OpenAI, 2023) have shown remarkable progress in textual content (Badaro et al., 2023), tabular data still needs more improvement to control hallucinations (Rebuffel et al., 2021). Neural models struggle with tabular data, especially when the inputs (i.e., cell values from rows and columns with respective headers) are not in a standard form. Some common tabular input problems are: when the cell values are 'non-atomic', for example, in Table 4a, while 'state' column is atomic, the 'candidates' column is non-atomic, where it contains name and % of votes of all candidates in one cell, missing important cell values (see Table 1) and use of hard-to-interpret column headers and row headers in tables[1].

Table 1 presents a sample[2] from ToTTo (Parikh et al., 2020). Only the highlighted cells from table 1a are passed to the model (as shown in Table 2). Passing Norton's % of votes and her party name (compare table 1b to table 1a) eliminates the hallucination (see output for table 1b); this correction is based on the data input problem described in Section 4.2.

In this work, the quality of output texts is scored by manually annotating output errors instead of using automated metric scores such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). These manual error annotations allowed us to identify patterns of errors which were then traced back to input problems. The main contributions of our work can be summarised as follows:

- We fine-tuned T5-base and T5-large models (Raffel et al., 2020), and conducted a pilot study, analysing 1,677 Politics domain texts from the ToTTo dataset (Parikh et al., 2020) through manual error analysis (Thomson and Reiter, 2020). This aimed to identify and understand errors in the generated text.

- We corrected the tabular input to adhere to a standard or canonical form to ensure the

[1] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Accessibility/Data_tables_tutorial#Column_headers:_bad_example

[2] In this paper, the term 'sample(s)' refers to tabular data (input) from the ToTTo dataset and/or its associated output text.

(a) Original cells highlighted in Yellow as Input

1996 United States House of Representatives election

Party	Candidate	Votes	%
Democratic	Eleanor Holmes Norton (inc.)	134,996	90.00
Republican	Sprague Simonds	11,306	7.54

LLAMA2-13B Output for tabular input a:
Eleanor Holmes Norton (inc.) won with **7.54%**[U] of the vote. Sprague Simonds was the Republican candidate and received **22.38%**[U] of the vote.

(b) Corrected Tabular data by including relevant cells

1996 United States House of Representatives election

Party	Candidate	Votes	%
Democratic	Eleanor Holmes Norton (inc.)	134,996	90.00
Republican	Sprague Simonds	11,306	7.54

LLAMA2-13B Output for tabular input b:
Democratic Party candidate Eleanor Holmes Norton won with 90% of the vote. Republican Party candidate Sprague Simonds received 7.54%.

Table 1: ToTTo example: Highlighted cells in yellow are passed as input to the model. Passing the appropriate cells (i.e., % votes and party name) as input, as shown in table 1b fixes the factual errors (compare output of table 1b to table 1a).

<page_title> 1996 United States House of Representatives election </page_title> <table> <cell> Eleanor Holmes Norton (inc.) <col_header> Candidate </col_header> </cell> <cell> Republican <col_header> Party </col_header> </cell> <cell> Sprague Simonds <col_header> Candidate </col_header> </cell> <cell> 7.54 <col_header> % </col_header> </cell> </table>

Table 2: Linearized Input from table 1a passed as Input to LLAMA2-13B model. Linearized Input for Table 1b is not shown here, it will include the related cells highlighted in Table 1b.

generation of factual texts from neural models. This correction was applied to a larger subset of 210 samples, resulting in a 62% decrease in factual errors for T5-base and a 57% decrease for T5-large in the generated text.

- We conducted experiments on the LLAMA2-7B and LLAMA2-13B models (Touvron et al., 2023) using 40 challenging samples selected from the previous 210. Tailoring zero-shot prompts for specific tabular input and manual error annotation on 160 texts showed that correcting input reduced factual errors by 52% in LLAMA2-7B and 76% in LLAMA2-13B.

- The inter-annotation agreement on the error annotation was good, with a Fleiss' Kappa of 0.622.

2 Pilot Study

2.1 Table to Text dataset, ToTTo

ToTTo is an open-domain English language dataset, where the input X is taken from Wikipedia table T, which includes the table's title (metadata) and specific highlighted cells, along with their headers. As shown in Table 1a, in many cases the Input X is observed to be problematic and fixing these problems is the main focus of this paper. This structured information is flattened to create a linearized text representation of the table, as mentioned in Table 2. The output Y, comprising the sequence of tokens $y_1, y_2, \ldots, y_n$ is a text description that coherently summarizes the content of the input table X. In this paper, we only look at politics domain.

2.2 T5 Baseline models

For this study, we fine-tuned both the T5-base (T5-b) and the T5-large (T5-l) models on the ToTTo dataset by following the baseline approach (Kale and Rastogi, 2020) and performed manual error annotation (Thomson and Reiter, 2020; Thomson et al., 2023) for our evaluation. We also added a new error category, ADDITION, which is used when the generated text have added words or phrases that diverge from the input. Definitions for all error categories adopted for this dataset can be found in Appendix B. Table 6 presents the eight categories of errors annotated for the two T5 models, and are detailed further in section 5.

Other hyperparameters and fine-tuning details for these two models are shown in Appendix A.

2.3 Insights from our Pilot Study

CATEGORY	T5-base (T5-b)		T5-large (T5-l)	
	Count	%	Count	%
NO ERROR	358	47	450	60
OMISSIONS	272	36	218	29
ERRORS	124	17	86	11
TOTAL COUNT	754		754	

Table 3: Pilot Study analysis: T5-b and T5-l Models in ToTTo Politics Domain. 'Errors' category is the main focus of our work; 'omissions' are excluded.

We focused on the Politics domain of the ToTTo

validation set. Our analysis (presented in Table 3) shows that:

No Error: 47% of the samples from T5-b and 60% of the samples from T5-l are error-free.

Omissions: Omissions occur when the generated text fails to mention some information from the input (González Corbelle et al., 2022) without making any factual errors. If a sample has both factual errors and omissions, we classify it as an error. The T5-b had 36% omissions and T5-l had 29%. The 'omissions' category is not our focus in this work and is excluded from our experiments.

Errors: Our analysis revealed that T5-b made factual errors in 17% and T5-l made factual errors in 11% of the total samples.

Hypothesis: Based on the insights from our error analysis, we hypothesised that when input data is structured in non-standard ways, models struggle to interpret these and therefore struggle to generate factually accurate output. We test this hypothesis by addressing input problems related to non-standard tabular structures. Our study focuses only on the 'errors' category (not omissions) of the ToTTo Politics domain. Further details are elaborated in section 3.

3 Input problems

The ToTTo dataset (Parikh et al., 2020) is created from Wikipedia tables, where a subset of table cells are paired with a Wikipedia description for approximately 121,000 samples. However, as illustrated in Table 1, data from the paired (highlighted) cells may not be structured in a standard and consistent way. We hypothesize that this leads to ambiguous interpretation of input by our models which in turn leads to output errors. Due to the practical challenges of correcting the entire dataset's tabular input, we focused on analyzing a subset of samples with errors (124 from T5-b and 86 from T5-l models, these were identified in our pilot study (shown in Table 3). Our aim is to segregate the errors that stem from non-standard or illogical table structures.

3.1 Generic input problems

Non-Atomic Tabular cell values: When a table column contains multiple atomic values in a cell (see Table 4). Examples of such non-atomic forms include multiple leaders' names, votes, their term dates, or election years in a single cell. We categorized these issues into 'single record lacking atomicity' and 'multiple records lacking atomicity' (shown in Figure 2 in Appendix E) to demonstrate how the models struggle when multiple records lack atomicity.

Complex table type: When tables contain election results in sentence form, models struggle to interpret and generate meaningful texts (see Table 11 in Appendix E).

Insufficient input: In some cases, the necessary cells are not highlighted in the tabular input, resulting in semantically incorrect outputs. Our analyses in Table 1 and Table 5 demonstrate that outputs become factually correct when relevant cells are included.

Longer table input: Models often struggle to generate accurate texts for lengthy table inputs, especially when the data is not in a standard form and lacks clear cell relationships.

3.2 Politics domain-specific input problems

Politics specific headers: In ToTTo, the use of symbols (e.g., '+' or '-' instead of clear semantic terms like *'swing percentage' or '% change compared to previous election'* as column headers in the tabular data caused 5% of errors. This lack of semantic guidance in the input made it difficult for models to accurately generate the correct output text (see Table 10 in Appendix E).

List of leader names in the input: In the politics domain, we observed a specific issue when input data contain list of leader names. Models tend to favor the leader whose name appears first in the list (for example, either from the table title or as the first leader in highlighted cells), even if they have lost the election (see Table 13, Table 14, Table 15 in Appendix E). This becomes even more complex when leader names are associated either with missing values (opponent leader's name or vote count) in other columns in the input or when the tabular input cell values are non-atomic.

4 Main Study

4.1 LLAMA2 models

In our main study, in addition to T5 models, we included LLAMA2-7B (L-7B) and LLAMA2-13B (L-13B) models to test our hypothesis on non-standard tabular input data, as discussed in section 2.3. We received official model weights from Meta, after filling this form[3] and followed the in-

[3] https://ai.meta.com/resources/models-and-libraries/llama-downloads/

(a) Input Table with Original Cells highlighted in Yellow

Page Title: 2014 United States Senate elections
Section Title: Elections leading to the next Congress

State	Candidates
Alaska	Dan Sullivan (Republican) 48.0% Mark Begich (Democratic) 45.8% Mark Fish (Libertarian) 3.7% Ted Gianoutsos (Independent) 2.0%
Virginia	Mark Warner (Democratic) 49.1% Ed Gillespie (Republican) 48.3% Robert Sarvis (Libertarian) 2.4%

(b) Input Table with Corrected Cells highlighted in Orange

State	Candidates	Party	% votes
Alaska	Dan Sullivan	Republican	48.0%
Alaska	Mark Begich	Democratic	45.8%
Virginia	Mark Warner	Democratic	49.1%
Virginia	Ed Gillespie	Republican	48.3%

(c) Outputs Before and After Input Data Correction

T5-base before fix: Alaska: Dan Sullivan (R) defeated Mark Begich (D) 45.8% to[W] 3.7%[U] and Ted Gianoutsos (Independent) 2.0%, and Virginia: Mark Warner won 49.1% to[W] 2.4%[U].
T5-base after fix: Alaska was won by Dan Sullivan (R) with 48.0% of the vote, and Virginia was won by Mark Warner (D) with 49.1% of the vote.

LLAMA2-7B before fix: 2014 United States Senate elections were held on November 4[A], 2014. The Republican Party gained control of the Senate for the first time since 2006. The Democrats lost their majority in the Senate and Republicans won a net gain of eight seats to take control of the chamber with a 54-46 majority. The Republicans also took control of the House of Representatives, giving them full control of Congress for the first time since 2010.[A]
LLAMA2-7B after fix: In the United States Senate elections of 2014, Dan Sullivan (Republican) won in Alaska with 48.0% of votes against Mark Begich (Democratic) who got 45.8%. In Virginia, Mark Warner (Democratic) won with 49.1% of votes against Ed Gillespie (Republican) who got 48.3%.

Table 4: Fixing Multiple leaders' votes that are not atomic, as detailed in Section 4.2.1. 'Before fix' shows the errors produced by models with table 4a as input and 'After fix' shows both models did not produce any factual error with table 4b as input. The error tokens in outputs are annotated with respective error code in superscript (W for WORD, U for NUMBER and A for ADDITION errors).

stallation instructions provided here[4] to quantize two of our large LLAMA2 models to 4-bit integer quantization. These quantized models were then used for local inference on a MacOS CPU. We set the temperature to 0.0 as it produced fewer errors. Further details on data corrections for these models are discussed in section 4.4.

4.2 Manual Fixes for Input Problems

4.2.1 Non-atomic cell values:

Correction: We corrected the input to store individual leader data, including votes, party, and election details, as a separate column variable. In Table 4a, when the candidates column is not atomic, all models made errors for this tabular data. The corrections involved separating each leader's party and votes into separate columns (see Table 4b). We passed the leading two leaders' data, excluding the remaining leaders, in separate rows for the respective states (Alaska and Virginia). After input correction, T5-b and L-7B outputs in Table 4 were error-free. While some models omitted other candidate and election details, *all models corrected the previous factual errors after the correction*.

Overall, T5-l and L-13B models corrected over 90% of the errors for this problem type, single record and multiple record lacking atomicity (see red and green bars in Figure 2, Appendix E). T5-b and L-7B models corrected over 60% of the errors but still struggled with a few samples even after correction.

[4] https://github.com/ggerganov/llama.cpp

4.2.2 Complex table type:

Correction: Complex tabular structure were generally difficult to correct because the election results are in sentence form e.g., see 'Results' column in Table 11 in Appendix E. We could only make corrections to 'non-atomic' cell values in 'Candidates' column. Three models generated 'the incumbent senator Coe I. Crawford lost renomination to Edwin S. Johnson, a Democrat candidate[C].' Losing renomination implies that the current senator lost the seat within their own party and not to the opponent party. This resulted in CONTEXT error, specifically *when the models made unsupported assumptions about the given input data*, as defined in Appendix B. All models except L-13B made this CONTEXT error, as shown in Table 11c.

Overall, T5-b could not fix this problem type. T5-l, L-7B, L-13B models partially fixed this input problem, as presented in Figure 2 in Appendix E.

4.2.3 Insufficient input:

Correction: This problem is straightforward to fix by including the missing cells (Table 1b) or headers from the tabular data. In Table 5, when the Minnesota Amendment's choices (i.e., 'Referendum failed No' and 'Yes' cells) were not included, all models incorrectly generated the favour of votes. As shown in Table 5, all models corrected the errors after including the choice of votes. In summary, this problem fixed all factual errors in T5-b, T5-l and L-13B after correction except for L-7B that added additional information for one sample.

(a) Input Table with Original Cells highlighted in Yellow

Page Title: Minnesota Amendment 2
Section Title: Results

Choice	Votes	% Votes
Referendum failed No	1,588,752	53.84
Yes	1,362,030	46.16

(b) Input Table with Corrected Cells highlighted in Orange

Choice	Votes	% Votes
Referendum failed No	1,588,752	53.84
Yes	1,362,030	46.16

(c) Outputs Before and After Input Data Correction

T5-base before fix The Minnesota Amendment 2 was approved[W] by 53.84 percent of voters, with a turnout of[C] 46.16 percent.

T5-base after fix: Minnesota Amendment 2 was defeated by 53.84 percent of voters, while 46.16 percent voted in favor of the measure.

LLAMA2-7B before fix: 53.84% of Minnesota voters voted for[W] Amendment 2.

LLAMA2-7B after fix: 53.84% of the voters voted no on Amendment 2, while 46.16% voted yes.%.

Table 5: Fixing Insufficient input problems, as detailed in Section 4.2.3. 'Before' shows WORD and CONTEXT errors due to missing choices of votes. 'After' shows these errors are resolved when these cell values are provided.

4.2.4 Longer table input:

Correction: When dealing with straightforward tabular data, it was easier to correct the input either by: i. *including the missing cells* and/or ii. *correcting the cell values to ensure atomicity*. However, complex tabular structures with hard-to-interpret column headers and row headers for over 100 rows pose difficulties to correct the input.

In general, the large models such as L-13B and T5-l corrected factual errors for straightforward input as shown in Figure 2 in Appendix E. T5-b and L-7B struggled to produce factual information. Example for this problem type is not presented, as this is a combination of Section 4.2.1 and Section 4.2.3.

4.2.5 ToTTo dataset: Politics specific

Correction: We included appropriate semantic cues in the headers to help the model differentiate between generic percentage of votes and swing percentage of votes. For other errors, we included 'party name' for write-in candidate and expanded the abbreviation of coalition party names to avoid errors. The Table 10 in Appendix E, the original input only had $\pm$ in the header and all models produced errors including L-13B. After explicitly including '$\pm$ seats compared to previous election', both LLAMA2 models corrected the previous error. However, both fine-tuned T5 models committed mistake even after this correction.

Specific semantic generation issue: For few samples, knowing the lifespan of ministers, including their birth and death years, is not enough to determine the right semantic whether the minister was the 'shortest-lived' or 'longest-lived'. It would require specific rules or instructions to produce accurate text. One such example is shown in Table 12 in Appendix E. All models with the original data produced incorrect text (WORD error). After we included 'Total time in office' cell and customised different prompts for LLAMA2, both LLAMA2 models corrected the WORD error. However, both T5 models did not show any improvement after including this detail. This could be due to the strong influence of the patterns learned from fine-tuning process for similar tabular structure.

Overall, L-13B fixed the factual mistakes for 5 out of 8 samples and L-7B fixed factual errors for 4 out of 8 samples. Both T5 models corrected 3 out of 8 samples (see Figure 2 in Appendix D).

4.2.6 List of Leader names:

Correction: We did not change the order of leader's name in the title, but we addressed missing vote counts and leader names to map the relations for each record as shown in Table 13, Table 14, and Table 15 in the Appendix E.

In Table 14, the title leader 'Chuck DeVore' lost the election among four other party nominees. Both T5 models made errors by focusing on *'Chuck DeVore' as the main candidate*, resulting in WORD and CONTEXT errors. NUMBER errors are present in all models due to missing votes. After including the votes and party details for all candidates, both LLAMA2 models corrected all errors. However, both T5 models still made errors stating the title leader either defeated or being defeated by all candidates, possibly due to *learned patterns during fine-tuning on the ToTTo dataset*.

In Table 15, the title leader, 'Joseph Haslet' ran as a candidate for Governor office in 1804 and 1807 but lost both elections. All models incorrectly generated Haslet won in both years. The correction is only limited to passing the party name and modifying the header from 'Subject' to 'Candidate'. Despite this change, all models continued to generate errors. Both T5 models struggled to generate the right winning candidate in 5 out of 11 samples (Figure 2a) for this input with same set of headers.

Overall, T5 models struggled the most to correct factual errors for this problem. L-7B model corrected three samples but produced factual errors for the remaining eight samples. After correction, L-13B made factual errors only for two samples (see Figure 2 in Appendix D).

4.3 T5 models on Corrected data

Based on the insights gathered from different input problems in section 3, we manually corrected the original tabular input to a possible standard form for 210 samples (taken from 124 and 86 'errors' category samples in Table 3). We ran both T5-b and T5-l models on the corrected data.

4.4 LLAMA2 models on Corrected data

From the corrected data, as described in section 4.3, we chose 40 challenging samples from the previous 210 as inputs to the LLAMA2-7B (L-7B) and LLAMA2-13B (L-13B) models. These samples were chosen to cover each of the input problem types described in section 3, guaranteeing a thorough analysis. Within these 40 samples, 21 were associated with general input problems and the remaining 19 are specific to the Politics domain, as shown in Figure 2c and Figure 2d (in Appendix E).

We studied the *zero-shot* capabilities of L-7B and L-13B models for the 40 challenging samples by employing different prompting techniques as proposed in Kojima et al. (2022). For each sample, we employed different prompts depending on the content of the tabular input. Few prompts are shown in Appendix C.

For each sample and for each of L-7B, L-13B, we examined the outputs before and after correcting the tabular input.

5 Results and discussion

The output texts from all the models working on corrected data have been manually annotated for errors following the same process as described in section 2. In this section, we summarize the results of input data corrections and discuss them. The previous section also presented and discussed error analysis locally while presenting individual corrections.

5.1 Error Reductions in T5 models

The input corrections significantly reduced factual errors, as evident in Table 6, which provides a comparison of error counts 'before data fix' and 'after data fix' for each error category. All errors in the generated text for 210 samples were manually annotated following the guidelines defined in Appendix B. It should be noted that *no prompt or instruction* was provided to these two fine-tuned T5 models.

CATEGORY	T5-base (T5-b)		T5-large (T5-l)	
	Before	After	Before	After
WORD	62	20	31	12
NAME	13	3	12	2
DATE_DIMENSION	7	1	6	1
NUMBER	12	1	5	0
OTHER	5	2	2	0
CONTEXT	13	3	16	4
ADDITION	2	1	2	1
NON-ENGLISH	20	20	21	21
TOTAL ERRORS	134	**51**	95	**41**
ERROR REDUCTION	**62%**		**57%**	

Table 6: Count of individual Error Annotations for 210 samples. The table compares the error counts between 'before' and 'after' applying tabular input correction. Each sample can contain multiple errors.

5.2 Error Reductions in LLAMA2 models

In Table 7, we present our error analysis of 40 difficult samples, validated using L-7B and L-13B models with tailored prompts (some shown in Appendix C). Outputs were manually error annotated, and the table compares the error counts before and after input correction. The *L-7B and L-13B* models showed **reductions of 52% and 76% in errors**, after addressing the tabular input to a standard form. The *T5-b and T5-l models* demonstrated **error reductions of 46% and 53%** respectively for the 40 difficult samples considered, which was shown for comparison purpose. Below is the summary of individual error categories after input correction:

- WORD errors, predominantly incorrect verbs and prepositions, were most common among all models. Post-correction, the L-13B model reduced WORD errors by 77%, while the T5-b, T5-l, and L-7B models achieved reductions of 33%, 42%, and 34%, respectively.

- Input correction led to a reduction in NAME and CONTEXT errors across all models.

- The original input data exhibited more NUMBER errors in LLAMA2 models compared to T5, which were significantly reduced after correction. Both LLAMA2 models completely resolved DATE_DIMENSION and OTHER errors.

CATEGORY	T5-base (T5-b) baseline		T5-large (T5-l) baseline		LLAMA2-7B (L-7B)		LLAMA2-13B (L-13B)	
	Before	After	Before	After	Before	After	Before	After
WORD	27	18	21	12	23	15	22	5
NAME	7	3	6	2	2	1	3	0
DATE_DIMENSION	2	1	1	1	2	0	1	0
CONTEXT	8	3	7	2	6	2	4	0
NUMBER	4	1	1	0	7	1	6	1
OTHER	0	0	0	0	5	0	1	0
ADDITION	0	0	0	0	5	5	7	5
TOTAL ERROR	48	**26**	36	**17**	50	**24**	45	**11**
ERROR REDUCTION(%)	**46%**		**53%**		**52%**		**76%**	

Table 7: Individual error count for the selected 40 challenging samples from the previous 210 from Table 6. Comparison of manual errors annotated for original data versus after providing data fix in T5 and LLAMA2 models.

CATEGORY	T5-base (T5-b) baseline		T5-large (T5-l) baseline		LLAMA2-7B (L-7B)		LLAMA2-13B (L-13B)	
	Before	After	Before	After	Before	After	Before	After
NO ERROR (Higher is better)	0	12	3	16	2	17	5	23
OMISSIONS	0	6	3	8	2	4	4	8

Table 8: Comparison of NO ERROR and OMISSIONS unique count for the same 40 samples. i. Increase in the number of 'NO ERROR' indicates that more output was found to be error-free after addressing input data issues. ii. Some outputs stopped making factual errors after correction but instead omitted part of the content from the input, resulting in a higher omission count post-fix.

- ADDITION errors are more common in LLAMA2 models than in T5. Despite including the prompt 'Use only the information mentioned in the input table data' and correcting the tabular data, both L-7B and L-13B models still produced five ADDITION errors each.

Table 8 shows two rows of data from our analysis of the 40 challenging samples. The first row shows that corrected data leads to an increase number of samples with 'no error' (shown in table 8, Higher is better). However, the second row shows that omissions have increased after input corrections. We observed that the models omitted part of the information either when the corrected tabular data had multiple column variables for more than two records or when the tabular structure is complex. In case of fine-tuned models, it learned to omit some information during the fine-tuning process. In our future work, we plan to study the reasons for why this is happening.

5.3 Model-specific errors for different input problem types

In Appendix D, Figure 2 presents how each of the four models is performing before and after data corrections for each of the input problem types, as described in section 3. For example, Figure 2b shows that T5-large was making more errors for the input problem type labelled 'Multiple records lacking atomicity' before (red bar) and made significant reductions after correction (green bar).

6 Inter-annotation experiment

One of the authors manually annotated errors in 1,508 outputs before input correction and 169 problematic outputs after correction (a total of 1,677 from both T5 models). Similarly, the annotation for both LLAMA2 models covered 160 outputs before and after input correction.

Two additional annotators annotated 60 outputs each, generated by four different models both before and after input correction. We provided them with a detailed document that included definitions of error categories, guidelines, and examples of error annotations. As illustrated in Table 1, we provided the tabular input along with the respective highlighted cells and output texts for error annotation. Annotators followed the guidelines and marked the errors. Each annotator spent approximately 3 hours to complete this experiment.

The annotated errors are shown in a confusion matrix in Table 9, where 'all agree' column shows all annotators agreed on the same error type and other columns show how often each annotator selected a different error type. Although the correct token was usually chosen, disagreements primarily occurred in choosing CONTEXT, ADDITION and WORD error types, as shown in red in Table 9. This might be due to the potential similarities in the definitions of CONTEXT errors and ADDITION errors (see Appendix B.3). While WORD error is comparatively straight-forward, one annotator chose 'CONTEXT' errors instead of WORD er-

ERROR	ALL AGREE	WORD	NAME	DATE_DIM	NUMBER	OTHER	CONTEXT	ADDITION	NO ERROR	TOTAL
WORD	17	0	0	0	2	0	9	3	0	31
NAME	3	0	0	0	0	0	2	0	0	5
DATE_DIM	1	0	0	0	0	0	2	0	0	3
NUMBER	2	2	0	0	0	0	0	0	0	4
OTHER	2	0	0	0	0	0	0	0	1	3
CONTEXT	7	9	2	2	0	0	0	6	0	26
ADDITION	12	1	0	0	0	0	5	0	0	18
NO ERROR	15	0	0	0	0	1	0	0	0	16

Fleiss' kappa overall agreement for three annotators, $k = (pa - pe)/(1 - pe)$ **= 0.622**

Table 9: Fleiss' Kappa coefficient: overall agreement on 60 samples among three annotators. 'All agree' column signifies unanimous agreement on error types, while other columns show unique error selections.

rors for few outputs. Cases where an annotator did not mark any errors were in the minority. The inter-annotation agreement on error category classification for 60 outputs, as shown by a Fleiss' kappa of 0.622, indicates substantial agreement between three annotators.

7 Related work

Table-to-Text. Other work on the ToTTo dataset include the LATTICE framework (Wang et al., 2022), preserving the structural information of the cell values for the same rows or columns and by removing the attention mechanism flow for other unrelated tokens; Hu et al. (2023) incorporated content planning and generation from Su et al. (2021) and synthetically added noisy cells in their fine-tuning regime. While these approaches are model agnostic and improved automated metric scores (BLEU, PARENT) by few points, the fundamental problem with the tabular input remains.

The Table Structure Understanding and Text Deliberation Approach (Chen et al., 2022), acknowledged that the target cells in the tabular input are not always highlighted. They used a template to capture all facts from raw table for 1.2K training samples (with rows and columns fewer than 8) and employed hierarchical multi-head attention to capture the structural information in their fine-tuning process. Though this approach is promising to retain the facts from raw table, it still has limitations for longer table, complex tabular structure and non-atomic cell values.

Error Analysis. While the metric scores (Papineni et al., 2002; Dhingra et al., 2019; Sellam et al., 2020) helps to evaluate the model's performance at a high level, relying only on automated metrics will not address specific weaknesses made by the neural models (van Miltenburg et al., 2021). Maynez et al. (2020) also emphasizes automated metrics are not sufficient and provided a detailed study on intrinsic and extrinsic hallucination in abstractive summarization. Our manual error analysis approach adopted from Thomson and Reiter (2020); Sundararajan et al. (2022); Thomson et al. (2023), helps to understand errors at a more granular level. Mapping our adopted methodology to Maynez et al. (2020)'s work, intrinsic refers to our main error categories (i.e., WORD, NAME, DATE, NUM, OTHER, CONTEXT) when the generated outputs are not faithful to the given input; extrinsic refers to the ADDITION category.

LLM prompts. Empirical evaluation of prompting strategies on the three large language models (LLMs) by Sivarajkumar et al. (2023) in clinical NLP found that tailoring task-specific prompt is crucial for achieving accuracy. In a study on Text-to-SQL, Chang and Fosler-Lussier (2023) investigated zero-shot prompting strategies, highlighting the significance of table representation. Their findings indicated that normalized database prompts outperformed unnormalized ones. In our work, we leveraged a recent LLM, llama2 (Touvron et al., 2023) and tailored our zero-shot prompt (Kojima et al., 2022) specific to the content of each table.

8 Conclusion

This paper presented a study that quantitatively demonstrates that fixing input problems such as insufficient data and data records containing non-atomic content improves the factual accuracy of output text by as high as 76% for one of the study models. Correcting inputs seems to also improve the number of output texts that are entirely error free. However, we still need to investigate why errors categorised as 'Omissions' increase after input corrections. In our future work, we aim to explore other tabular datasets for problems with input data, and study the generalization of the fixes explored in the current work.

Limitations

We acknowledge some limitations in this work. First, our corrections to tabular input are limited to errors identified within the Politics domain of ToTTo validation set. We excluded other domains from this dataset such as Sports, History, Arts, Entertainment, Demography, and more. We solely focus on the input problems within the ToTTo dataset, and our future work will involve exploring other datasets. Second, we did not extend the correction of tabular input for the Politics domain to the training split of ToTTo dataset due to the time-consuming process of handling different table headers and metadata.

In our future work, we aim to simplify the definitions of the 'context' and 'addition' error categories, as the annotation experiment revealed disagreement in choosing these error types for some samples, despite annotators marking the same error token.

Ethics Statement

This work seeks to address input problems in non-standard tabular structure and with the aim of reducing factual errors in the output text. We utilized the open-source dataset, ToTTo and maintained the ground-truth generation same as the original dataset. Our input correction did not introduce any further social bias to this dataset. We adopted error annotation methodology to annotate factual errors and one of the authors performed manual error analysis for this complete study. We sought consent from two additional annotators, the annotators volunteered to participate and annotated errors for 60 output texts each. They had the right to withdraw from the study at any point without facing any consequences. We provided necessary guidelines, instructions and examples for them to annotate errors.

References

Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, 11:227–249.

Shuaichen Chang and Eric Fosler-Lussier. 2023. How to prompt llms for text-to-sql: A study in zero-shot, single-domain, and cross-domain settings.

Miao Chen, Xinjiang Lu, Tong Xu, Yanyan Li, Zhou Jingbo, Dejing Dou, and Hui Xiong. 2022. Towards table-to-text generation with pretrained language model: A table structure understanding and text deliberating approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8199–8210, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. *CoRR*, abs/2004.10404.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020b. Logic2Text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2096–2111, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

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.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2017. Data-to-text generation improves decision-making under uncertainty. *IEEE Computational Intelligence Magazine*, 12(3):10–17.

Javier González Corbelle, Alberto Bugarín-Diz, Jose Alonso-Moral, and Juan Taboada. 2022. Dealing with hallucination and omission in neural natural language generation: A use case on meteorology. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 121–130, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Hanxu Hu, Yunqing Liu, Zhongyi Yu, and Laura Perez-Beltrachini. 2023. Improving user controlled table-to-text generation robustness. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2317–2324, Dubrovnik, Croatia. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Steffen C. Pauws, Albert Gatt, Emiel J. Krahmer, and Ehud Reiter. 2018. Making effective use of healthcare data using data-to-text technology. In *Data Science for Healthcare*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6908–6915.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2021. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 36:318 – 354.

Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2019. A hierarchical model for data-to-text generation. *Advances in Information Retrieval*, 12035:65 – 80.

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.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2023. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. *arXiv preprint arXiv:2309.08008*.

Somayajulu Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2002. Sumtime-meteo: Parallel corpus of naturally occurring forecast texts and weather data. *Computing Science Department, University of Aberdeen, Aberdeen, Scotland, Tech. Rep. AUCS/TR0201*.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barkavi Sundararajan, Somayajulu Sripada, and Ehud Reiter. 2022. Error analysis of ToTTo table-to-text neural NLG models. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 456–470, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Somayajulu Sripada. 2020. SportSett:basketball - a robust and maintainable data-set for natural language generation. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 32–40, Santiago de Compostela, Spain. Association for Computational Lingustics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech Language*, 80:101482.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Upadhyay and Stewart Massie. 2022. Content type profiling of data-to-text generation datasets. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5770–5782, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. 2022. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5037–5048, Seattle, United States. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Fengbin Zhu, Moxin Li, Junbin Xiao, Fuli Feng, Chao Wang, and Tat Seng Chua. 2023. Soargraph: Numerical reasoning over financial table-text data via semantic-oriented hierarchical graphs. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1236–1244, New York, NY, USA. Association for Computing Machinery.

Appendices

A Model fine-tuning specifications

The first model, T5-base (T5-b), was fine-tuned on the *full ToTTo training set of 120,761 samples* on a commodity server with a GeForce RTX 2080 GPU. The training took around seven days. The second model, T5-large (T5-l), was fine-tuned on *a subset of 50,000 ToTTo samples* on a secure cloud instance with an NVIDIA A100 GPU, completing in around 48 hours. Both models were fine-tuned using a constant learning rate of 0.0001, with the encoder's maximum length set to 512 tokens and the decoder's maximum length set to 128 tokens for ToTTo generations (Kale and Rastogi, 2020). Single-precision floating-point format (FP32) was employed for training on their respective GPU servers. The batch size, beam size and training steps for each model is shown in fig. 1.

Models	Batch size	Beam Size	Training steps
T5-base	2	10	180,000
T5-large	4	5	9,000

Figure 1: Model Specifications

B Inter-Annotation Procedure for Participants

B.1 Overview

The Input Data from ToTTo dataset, includes the Page title, Section Title and a Table with highlighted cells in yellow. These key parameters are conditionally used for training the neural models to summarize a meaningful and factual Text (as Output) focusing on: (i). highlighted cells in the table (ii). their corresponding header values (iii) The main Title and (iv) The Section Title.

For each of this tabular input data, you will see eight Outputs generated by different neural language models. Our goal is to evaluate whether the neural outputs remain faithful and produce factually accurate information based on the four parameters from the tabular input. The complete table, including the non-highlighted cells, is provided to offer you a clearer understanding of the error annotation task.

B.2 Domain

The inputs provided are specific to the domain of Politics, sourced from Wikipedia tables (as part of

the open-domain ToTTo dataset). The political data within these tables isn't limited to a single demography. Instead, it encompasses various details from the election processes across multiple countries, including:

- Election specifics such as Presidential, state, by-elections, council, district, Legislative Assembly, and other elections unique to particular countries.
- Information about Governors, Mayors, Ministers, and Ambassadors (pertaining to Foreign Affairs).
- Details regarding the Speaker of the Assembly.

The first item related to election details are predominantly used in this error annotation task.

B.3 Error Annotation guidelines

We are only interested if the highlighted cell values from the table were used to produce a factually correct sentence. The non-highlighted values in the table are not expected to be used for your annotation. Please read through the output texts and annotate cases for the error categories as mentioned below. Each error is denoted with a superscript for better readability.

NAME^{N}

- When names of the Party, Leader, place (Electorate), Ambassador etc., are wrong (mostly nouns).
- Annotation includes both single token or multiple tokens to include the complete names.
- Example Output text: **Urban Ahlin^{N}** is the Deputy Speaker of the Riksdag. Remarks: NAME error because the correct deputy speaker was Tobias Billström as per the tabular data.
- Example Output text: Kansas was won by Mitt Romney, Paul Ryan, **Barack Obama^{N}**, and **Joe Biden^{N}**, with Romney winning 59.66% of the popular vote, six electoral votes and 38.05 percent. In this example, Barack Obama and Joe Biden are two NAME errors because they did not win the election.
- In general, **Wednesday^{N}** instead of Tuesday is a NAME error but **May^{D}** instead of April is a DATE_DIMENSION error.

NUMBER^{U}

- When the number of seats and/or the number of votes and/or % of votes are incorrect. Single token is marked as error.
- When the A-party won with a majority of 5.5%. But the correct one is 4.4%. **5.5%^{U}** is a NUMBER error.
- Output: The voter turnout was 8,90%, with 10,052 votes. Remarks: The actual turnout was **81.90%^{U}**. Please note: the error here is NOT with the comma used as decimal (as it is an acceptable decimal operator for international use); Error because the number 81.90 was not correctly predicted.

DATE_DIMENSION^{D}

- When the Date and/or Month and/or Year are wrong in the generated text, it is annotated as one error.
- Example Output text: Cletus Avoka was the Minister for the Interior in the Mills government from 2009 to **2012^{D}**. Remarks: 2010 is the right end term of the year.
- As a general note, if the Output text did not capture Month and/or Date, but has the correct year, then this is NOT an ERROR. It could go to OMISSION with remarks Omission of Date and Month.

WORD^{W}

- When incorrect words such as verbs, prepositions, adjectives, adverbs and conjunctions are found in the output.
- Single token is marked as WORD error in most cases. Multiple tokens are annotated when the auxiliary verb (was), extension of prepositional phrase (along with) and others are incorrect.
- Example Output text: Carly Fiorina defeated Republican Tom Campbell with 56.4% of the vote to DeVore's 19.3%, **along with^{W}** Al Ramirez and Tim Kalemkarian. Remarks: Fiorina independently defeated all the leaders, so along with is wrong.
- Example Output text: Ling **won^{W}** the 2016 senate district against Democrat Josh Newman, with 49.6 percent of the vote. Remarks: Ling lost the election as per the tabular data.

- Some of the common WORD errors found in this data are won, defeated, lost, succeeded, adjectives such as current governor other prepositions (since, in and so on).

CONTEXT[C]

- When the model's output presents information that contradicts or make *unsupported assumptions about the given input data*. Group of tokens/span of text are annotated as CONTEXT error.

- It can sometimes be tricky to mark this error. Please follow the below sequence before marking this error.
  - In case of simple misrepresentation based on the information in the input data, it would be easier to mark the token as NAME, NUMBER, DATE_DIMENSION or WORD error.
  - In the case of a complex table structure, the outputs are likely to mess up completely with the overall information in the provided input data. In this case, it is hard to mark individual errors. Please go ahead and mark the group of tokens/ span of text and annotate it as a CONTEXT error.
  - For example, when the name of a person (Florence Pendleton) is contradicted as a jurisdiction ('Florence') in the Output text.

ADDITION[A]

- When the model's outputs have *added words, phrases, or details* that either diverge from the input's main topic or are unsupported by the given context.

- Single or group of tokens are annotated as ADDITION error.

- Example Output text: 2007 Algerian legislative election **was held on May 17, 2007**[A]. The results were as follows: 24 political parties won a total of 389 seats in the National Assembly. **Remarks:** The date and month is an additional information, that are not provided in the input. This is marked as an ADDITION error. Other parts of the output texts are correct.

OTHER[O]

- When the output repeats the same input multiple times producing garbage data. In some cases, when the table data has the Politics party name in the abbreviation, it produces garbage output. For example, when the tabular data has party name in abbreviation, it tries to produce a strange output. Output text: **GSSSDULSVDHSS**[O] gained 5.31% of the vote

- When the output is incomplete for longer table input or when when the output repeats the same input multiple times without producing a complete text.

- When punctuation symbols are placed at inappropriate places, for example, an apostrophe is missed for the Name of the Leader or Place.

NON-ENGLISH[NE]

- when the Unicode characters in non-English names are either replaced with special characters or when these Unicode characters are omitted.

- For example, **Pawe Gra**[NE] is a member of Sejm. **Remarks:** Paweł Graś is the correct name here.

Please identify the above-mentioned errors from the output texts and change the font colour of error tokens in red. If you're colour-blind, you can underline them instead. Please also add details of each error in the table below each example. The columns you need to fill are as follows:

- **Error:** The text you highlighted in red or underlined.

- **Category:** The type of error, choose from the eight categories mentioned above. The capitalization here does not matter, so you can write NAME, Name or even name.

- **Remarks:** Your additional remark if the error is not straight-forward. The intention here is to record your observations and not to correct the sentence.

- **Model :** This column will be already provided in the default template. You only need to include details in 'error', 'category' and 'remarks' columns to each of the model number.

If each model has multiple errors, please add rows, and update the Model accordingly to fill the other three columns.

C LLAMA2 prompts

Below prompt is for the corrected table data when the input table has party name, candidate and votes.

> 'Given the input table data, the task is to:
> (i). Identify the party name, candidate name, and the number of votes received by each candidate.
> (ii). Determine the winner based on the highest number of votes. Then, put together the gathered information from (i) and (ii) into a single coherent sentence. Input table data: <Linearized table data>'

Below is one of the prompt used for the original table data without mentioning any specific fields.

> 'The task is to summarize the information from the given input table data into a single coherent sentence. Use only the information mentioned in the input table data. Input table data is: <Linearized table data>'

D Input Problem Types for four models

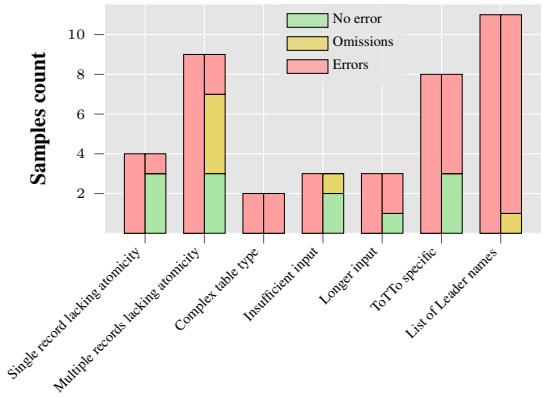
Figure 2 shows how each of the four models is performing before and after data corrections for each of the problem types. The left bars for each input problem type represent output scores before data correction, while the right bars represent scores after data correction. Color coding of output scores helps to identify and understand errors: green indicates no error, red indicates an error, and yellow indicates an omission.

The number of samples for each model differs, because we focus on outputs only when the model had 'errors' or 'omissions' before correction. This approach emphasizes the actual improvement in corrections. For example, T5-large had errors in 37 samples before correction, therefore the after correction also shows improvements for the same 37 samples.

E Input Data Corrections and Output Text

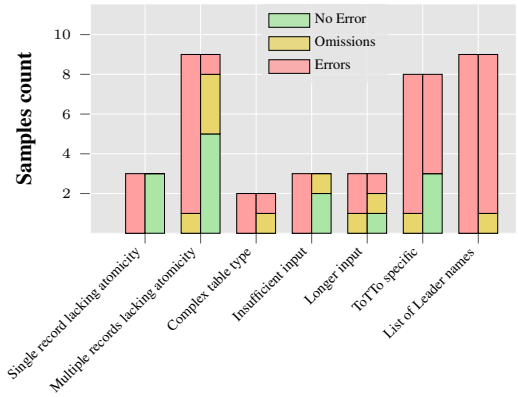
In this section, the input tabular data with outputs supplements the examples elaborated in Section 4.2 for each of the input problems as presented from Table 10 to Table 15.

(a) T5-base: Error Analysis of 40 samples.



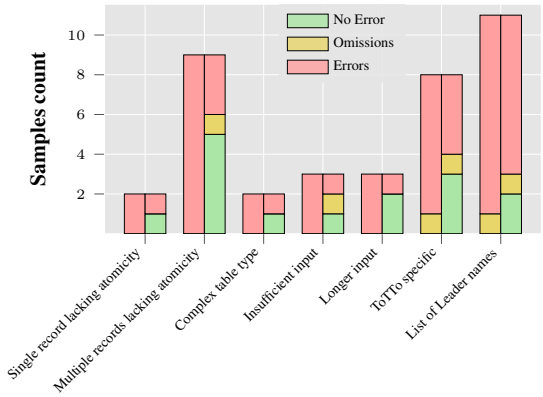
Input Problem Types: before data fix (left bar) versus after data fix (right bar)

(b) T5-large model: Error Analysis of 37 samples.



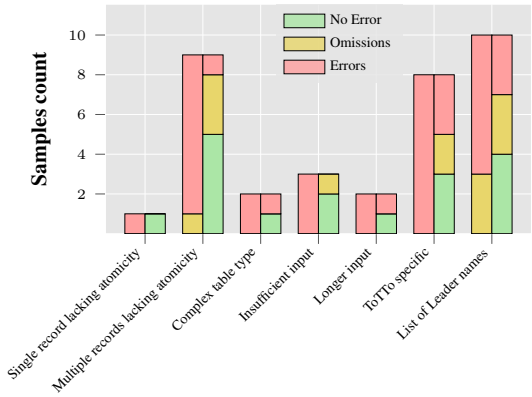
Input Problem Types: before data fix (left bar) versus after data fix (right bar)

(c) LLAMA2-7b: Error Analysis of 38 samples.



Input Problem Types: before data fix (left bar) versus after data fix (right bar)

(d) LLAMA2-13b: Error Analysis of 35 samples.



Input Problem Types: before data fix (left bar) versus after data fix (right bar)

Figure 2: Comparison of different input problem types before (original data) and after data correction across four models. It shows count of unique samples for three categories: (i) errors, (ii) omissions, and (iii) no errors.

(a) Input Table with Original Cells highlighted in Yellow

Page Title: 2012 Cardiff South and Penarth by-election
Section Title: By-election result

Party	%	±
Conservative	19.9	8.4
Liberal Democrat	10.8	-11.5

(b) Input Table with Corrected Cells highlighted in Orange

Page Title: 2012 Cardiff South and Penarth by-election
Section Title: By-election result

Party	%	± % difference with previous election
Conservative	19.9	8.4
Liberal Democrat	10.8	-11.5

(c) Outputs Before and After Input Data Correction

T5-large before fix: In the 2012 Cardiff South and Penarth by-election, the Liberal Democrats won[W] 11.5% of the vote.

T5-large after fix: In the 2012 Cardiff South and Penarth by-election, the Liberal Democrats lost by[W] 11.5%.

LLAMA2-13B before fix: 2012 Cardiff South and Penarth by-election was won[W] by the Liberal Democrats with a margin[W] of -11.5%.

LLAMA2-13B after fix: 2012 Cardiff South and Penarth by-election saw Liberal Democrat candidate lose by 11.5% of the vote compared to the previous election.

Table 10: ToTTo specific: Swing percentage difference with previous election, as detailed in Section 4.2.5. Before fix represents the output generated from input Table 10a and after fix represents the output generated after explicitly changing the header, as shown in input Table 10b.

(a) Input Table with Original Cells highlighted in Yellow

Page Title: 1914 United States Senate elections
Section Title: Races leading to the 64th Congress

State	Incumbent		Results	Candidates
	Senator	Party		
South Dakota	Coe I.Crawford	Republican	Incumbent lost renomination. New senator elected. Democratic gain.	Edwin S. Johnson (Democratic) 48.32% Charles H. Burke (Republican) 44.47%

(b) Outputs with Original Input before Input Data Correction

T5-base: In South Dakota, Republican Coe I. Crawford lost renomination **to Democrat Edwin S. Johnson**[C].

T5-large: South Dakota: Republican incumbent Coe I. Crawford lost his bid for the seat **to Democratic challenger Edwin S. Johnson**[C].

LLAMA2-7B: 1914 United States Senate elections in South Dakota were held on **November 3**[A], 1914. Incumbent Republican Senator Coe I. Crawford lost renomination **to Democratic candidate Edwin S. Johnson**[C]. Johnson defeated **Charles H. Dunnigan**[N] by **a margin of**[W] 48.32% to **47.56%**[U]. This was a Democratic gain from the previous election.

LLAMA2-13B: 1914 United States Senate elections were held on **November 3**[A], 1914 to elect members of the United States Senate in various states. The Republican Party lost one seat and the Democratic Party gained one seat.

(c) Input Table with Corrected Cells highlighted in Orange

Page Title: 1914 United States Senate elections
Section Title: Races leading to the 64th Congress

State	Incumbent Senator	Incumbent Party	Results	Candidate Senator	Candidate Party	Candidate votes %
South Dakota	Coe I.Crawford	Republican	Incumbent lost renomination. New senator elected. Democratic gain.	Edwin S. Johnson	Democratic	48.32%
South Dakota	Coe I.Crawford	Republican	Incumbent lost renomination. New senator elected. Democratic gain.	Charles H. Burke	Republican	44.47%

(d) Outputs After Input Data Correction

T5-base: In South Dakota, Republican Coe I. Crawford lost renomination **to Democrat Edwin S. Johnson**[C].

T5-large: South Dakota: Republican incumbent Coe I. Crawford lost his bid for the seat **to Democratic challenger Edwin S. Johnson**[C].

LLAMA2-7B: 1914 United States Senate elections were held on **November 3,**[A] 1914. In South Dakota, Coe I. Crawford, the Republican incumbent Senator lost renomination **to Edwin S. Johnson**[C], a Democratic candidate. Charles H., the Republican candidate also ran for the seat but he was defeated by Johnson. Therefore, it is a Democratic gain in this election.

LLAMA2-13B: 1914 United States Senate elections were held on **November 3**[A], 1914 to elect members of the United States Senate in various states. Races leading to the 64th Congress included South Dakota where Coe I. Crawford (Republican) lost renomination and Edwin S. Johnson (Democratic) was elected.

Table 11: Complex Table input problems, as detailed in Section 4.2.2. Difficult to fix this complex input type, yet LLAMA2-13B produced most of the facts right except for ADDITION error. Lost renomination 'to Democrat candidate' is a CONTEXT error, because the model made unsupported assumptions about the given input.

(a) Input Table with Original Cells highlighted in Yellow

Page Title: List of Japanese prime ministers by longevity
Section Title: Longevity of Japanese prime ministers

Prime Minister	Total time in Office	Date of Death	Lifespan
Sanetomi Sanjō	60 days	28 Feb 1891	53 years, 352 days

(b) Input Table with Corrected Cells highlighted in Orange

Page Title: List of Japanese prime ministers by longevity
Section Title: Longevity of Japanese prime ministers

Prime Minister	Total time in Office	Date of Death	Lifespan
Sanetomi Sanjō	60 days	28 Feb 1891	53 years, 352 days

(c) Outputs Before and After Input Data Correction

T5-base and T5-large before and after fix (no improvement): Sanj Sanetomi[NE] was the longest-lived[W] prime minister in Japan at 53 years, 352 days.

LLAMA2-7B before fix: The longest-serving[W] Japanese prime minister was Sanjō Sanetomi who served for 53 years and 352 days.
LLAMA2-7B after data fix: Sanjō Sanetomi was the Japanese prime minister and served for a total of 60 days in office. He died at age 53 years and 352 days.

LLAMA2-13B before data fix: The longest-lived[W] Japanese prime minister was Sanjō Sanetomi, who served for 53 years and 352 days.
LLAMA2-13B after data fix: Sanjō Sanetomi was the prime minister of Japan for a total of 60 days. He lived for 53 years and 352 days.

Table 12: ToTTo specific - WORD error made by all models before fix, as detailed in Section 4.2.5. After including 'Total time in Office', LLAMA2 models fixed the 'longest-lived' word error and introduced other addition errors. We revised the prompt and fixed the addition errors. *T5 models had no improvements after including information.*

(a) Input Table with Original Cells highlighted in Yellow

Page Title: Ling Ling Chang
Section Title: 2016 Senate run

California's 29th State Senate district election		
Party	**Candidate**	**% Votes**
Democratic	Josh Newman	50.4
Republican	Ling Ling Chang	49.6

(b) Input Table with Corrected Cells highlighted in Orange

Page Title: Ling Ling Chang
Section Title: 2016 Senate run

California's 29th State Senate district election		
Party	**Candidate**	**% Votes**
Democratic	Josh Newman	50.4
Republican	Ling Ling Chang	49.6

(c) Outputs Before and After Input Data Correction

T5-base and T5-large before fix: Chang defeated[W] Democrat Josh Newman with 49.6% of the vote.

T5-base and T5-large after fix: Ling Ling Chang was defeated by Josh Newman, a Democrat, with 50.4% of the vote to Chang's 49.6%.

LLAMA2-13B before fix: Ling Ling Chang won[W] the 2016 Senate run with 49.6% of the vote.

LLAMA2-13B after fix: Ling Ling Chang, who was running for the Senate seat in California, received 49.6% of votes while Josh Newman, her opponent, received 50.4%.

Table 13: Fixing ToTTo specific input problems: List of leader names for two leaders. 'Before' shows WORD errors due to missing vote percentages and party names for a leader. 'After' shows these errors are resolved by all models when the missing values are provided. Key point is T5 model did not produce error even though the title leader 'Ling Ling Chang' lost the election.

(a) Input Table with Original Cells highlighted in Yellow

Page Title: Chuck DeVore
Section Title: As U.S. Senate Candidate

Party	Candidate	Votes	%
Republican	Carly Fiorina	1,315,429	56.4
Republican	Tom Campbell	504,289	21.7
Republican	Chuck DeVore	452,577	19.3
Republican	Al Ramirez	42,149	1.8
Republican	Tim Kalemkarian	19,598	0.8

(b) Input Table with Corrected Cells highlighted in Orange

Page Title: Chuck DeVore
Section Title: As U.S. Senate Candidate

Party	Candidate	Votes	%
Republican	Carly Fiorina	1,315,429	56.4
Republican	Tom Campbell	504,289	21.7
Republican	Chuck DeVore	452,577	19.3
Republican	Al Ramirez	42,149	1.8
Republican	Tim Kalemkarian	19,598	0.8

(c) Outputs Before and After Input Data Correction

T5-base before fix: DeVore **defeated**[W] Republican Tom Campbell with **56.4%**[U] of the vote, **Al Ramirez with 19.3%, and Tim Kalemkarian with 19.3%**[C].
T5-base after fix: DeVore **defeated**[W] Republican nominees Carly Fiorina (56.4%), **Tom Campbell (21.7%), Al Ramirez (1.8%) and Tim Kalemkarian (0.8%)**[C].

T5-large before fix: DeVore was defeated by Republican Tom Campbell **(56.4%)**[U], **Al Ramirez (19.3%), and Tim Kalemkarian (21.7%)**[C].
T5-large after fix: DeVore was defeated in the U.S. Senate race by Republican nominees Carly Fiorina (56.4%), Tom Campbell (21.7%), **Al Ramirez (1.8%) and Tim Kalemkarian (0.8%).**[C]

LLAMA2-7B before fix: Chuck DeVore is a Republican candidate for the U.S. Senate in California. He received 19.3% of the vote in the primary election, which was held on June 8th, 2010. The other candidates were Carly Fiorina (56.4%), Tom Campbell **(24.1%)**[U] and Al Ramirez **(10.2%)**[U]. Tim Kalemkarian **did not run for this office**[C].
LLAMA2-7B after fix: The number of votes for each leader is as follows: Carly Fiorina - 56.4% Tom Campbell - 21.7% Chuck DeVore - 19.3% Al Ramirez - 1.8% Tim Kalemkarian - 0.8%. Carly Fiorina won the Republican primary for U.S. Senate in California with 56.4 percent of the vote, followed by Tom Campbell at 21.7 percent and Chuck DeVore at 19.3 percent.

LLAMA2-13B before fix: Carly Fiorina is the leading candidate for U.S. Senate with 56.4% of the vote, followed by Tom Campbell with **20.3%**[U], Chuck DeVore with 19.3%, and Al Ramirez with **4.0%**[U].
LLAMA2-13B after fix: The Republican Party candidate who won the most votes in the U.S. Senate election was Carly Fiorina with 56.4% of the total votes.

Table 14: ToTTo specific input problems: List of Leader names for five leaders, as detailed in Section 4.2.6. All models swapped the defeated leader names and/or % of votes, resulting in WORD, NUMBER and CONTEXT errors. After including the party names and % of votes for all leaders, LLAMA2 models corrected the errors. Fine-tuned T5 models still made errors. This could be because of the pattern learned from the fine-tuned training data of the ToTTo dataset.

(a) Input Table with Original Cells highlighted in Yellow

Page Title: Joseph Haslet

Section Title: Almanac

Year	Office	Subject	Party	% Votes	Opponent	Party	% Votes
1804	Governor	Joseph Haslet	Republican	48%	Nathaniel Mitchell	Federalist	52%
1807	Governor	Joseph Haslet	Republican	48%	George Truitt	Federalist	52%

(b) Input Table with Corrected Cells highlighted in Orange

Page Title: Joseph Haslet

Section Title: Almanac

Year	Office	Candidate	Party	% Votes	Opponent	Party	% Votes
1804	Governor	Joseph Haslet	Republican	48%	Nathaniel Mitchell	Federalist	52%
1807	Governor	Joseph Haslet	Republican	48%	George Truitt	Federalist	52%

(c) Outputs Before and After Input Data Correction

T5-base before fix: Joseph Haslet **was elected**[W] Governor in 1804, **defeating**[W] Federalist Party candidate **Nathaniel Mitchell**[N] in the election of 1807.

T5-base after fix: Joseph Haslet **was elected**[W] Governor **of**[W] the Republican Party in 1804 and 1807, **defeating**[W] the Federalist Party's Nathaniel Mitchell and George Truitt.

T5-large before fix: Joseph Haslet **was elected**[W] Governor in 1804, **defeating**[W] Federalist Party candidate Nathaniel Mitchell in the election of 1807.

T5-large after fix: In 1804 and 1807, Republican Joseph Haslet **defeated**[W] Federalist Nathaniel Mitchell and Federalist George Truitt.

LLAMA2-13B before fix: 1804, Joseph Haslet **was elected**[W] governor with 48% of the vote. In 1807, he **was re-elected**[W] with 48% of the vote.

LLAMA2-13B after fix: 1804: Joseph Haslet (Republican) **won**[W] with 48% of the votes. 1807: George Truitt (Federalist) won with 52% of the votes.

Table 15: ToTTo specific input problems: List of Leader names, as detailed in Section 4.2.6. Correction for this input is including the party name and modifying the header 'Subject' to 'Candidate'. All models generated Joseph Haslet to be the winning candidate including the LLAMA2 models that had specific instructions.