TabXEval: Why this is a Bad Table? An eXhaustive Rubric for Table Evaluation

[*]Vihang Pancholi [*]Jainit Bafna [*]Tejas Anvekar
Manish Shrivastava [†]Vivek Gupta

Arizona State University IIIT Hyderabad
{vpancho1,tanvekar,vgupta140}@asu.edu
{jainit.bafna,m.shrivastava}@research.iiit.ac.in

Abstract

Evaluating tables qualitatively and quantitatively poses a significant challenge, as standard metrics often overlook subtle structural and content-level discrepancies. To address this, we propose a rubric-based evaluation framework that integrates multi-level structural descriptors with fine-grained contextual signals, enabling more precise and consistent table comparison. Building on this, we introduce **TabXEval**, an eXhaustive and eXplainable two-phase evaluation framework. TabXEval first aligns reference and predicted tables structurally via TabAlign, then performs semantic and syntactic comparison using TabCompare, offering interpretable and granular feedback. We evaluate TabXEval on **TabXBench**, a diverse, multi-domain benchmark featuring realistic table perturbations and human annotations. A sensitivity-specificity analysis further demonstrates the robustness and explainability of TabXEval across varied table tasks. Code and data are available at https://coral-lab-asu.github.io/tabxeval/.

1 Introduction

Tables are a ubiquitous data format across critical workflows: budget forecasts, patient dashboards, and experimental logs alike where even a one-cell error can trigger costly re-statements or clinical misinterpretations. As large language models (LLMs) and other neural systems are increasingly tasked with *generating* or *transforming* such tables, reliable automatic *evaluation* becomes a bottleneck.

Despite the structured nature of tables, most evaluation metrics treat them as plain text. Metrics like BLEU, ROUGE, METEOR, and chrF rely on n-gram overlap, ignoring row–column alignment and unit consistency (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Popović, 2015).

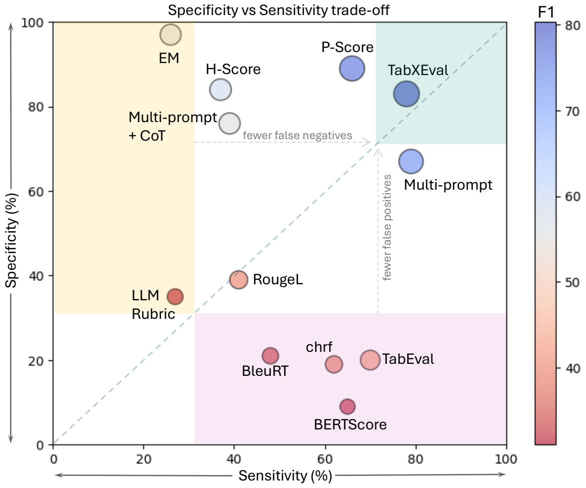


Figure 1: Sensitivity–Specificity trade-off across metrics. Bubble size reflects the harmonic mean of accuracy, sensitivity, and specificity; color shows F1-score. The dashed line marks ideal balance top-right methods perform best. Green denotes optimal (**Goldilocks**) zone; pink favors sensitivity, yellow favors specificity.

Embedding-based scores such as BERTScore improve semantic sensitivity but overlook structural errors like column swaps (Zhang* et al., 2020). Token-level metrics, including Exact Match and PARENT, address factual grounding but fail under reordered or merged schemas (Dhingra et al., 2019). Structural benchmarks highlight these issues: StructBench exposes failures on partial cell mismatches (Gu et al., 2024), TanQ reveals brittleness under unit conversions (Akhtar et al., 2025), and Data-QuestEval sacrifices structure for corpus-level QA-based comparisons (Rebuffel et al., 2021). Atomic decomposition methods like "Is this a bad table?" improve detection but add opacity and computational cost (Ramu et al., 2024). In contrast, work like THumB shows the benefit of rubric-based human ratings in improving evaluation transparency (Kasai et al., 2022).

Taken together, existing metrics tend to emphasize either semantics or structure, but rarely both. They offer limited diagnostic insight, often masking

[*]These authors contributed equally to this work.
[†]Primary supervisor and corresponding author of this work.

Findings of the Association for Computational Linguistics: ACL 2025, pages 22913–22934
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

specific error types and failing to provide actionable feedback for model improvement. As shown in image-captioning work like THumB, coupling automatic scores with rubric-based human ratings yields more interpretable and reliable evaluations (Kasai et al., 2022). These findings underscore the need for a rubric-based evaluation framework that explicitly assesses both structural alignment and semantic fidelity. In contrast to single-score metrics that collapse diverse errors—such as schema mismatches, contextual omissions, or subtle content shifts—into a single value, a rubric-based approach offers fine-grained, interpretable feedback. Such granularity is essential for complex or high-stakes tasks, where even minor discrepancies can significantly affect downstream performance.

To overcome these challenges, we introduce **TabXEval**, a novel evaluation framework built on a structured, multi-level rubric that combines high-level structural descriptors with fine-grained contextual signals. TabXEval operates in two phases: *TabAlign* first performs precise alignment of table elements using both rule-based and LLM-assisted strategies, followed by *TabCompare*, which conducts detailed semantic and syntactic analysis over the aligned cells. This design allows **TabXEval** to capture both table-level and cell-level discrepancies that prior metrics often overlook.

To rigorously test our rubric and framework, we construct **TabXBench**, a diverse, synthetic benchmark that emulates realistic table perturbations across multiple domains. TabXBench includes human-annotated ratings grounded in our rubric, serving as a gold standard for evaluating metric sensitivity, specificity, and alignment with human judgment. By providing controlled, interpretable scenarios, TabXBench fills a critical gap in current evaluation practice, enabling robust and explainable assessment of structured table outputs. Unlike prior metrics, **TabXEval**, as shown in Figure 1, excels at detecting subtle discrepancies i.e. sensitive and accurately localizing errors between tables i.e. specific enough. We summarize our main contributions below:

- We introduce the first rubric integrating multi-level structural descriptors and fine-grained contextual quantification for robust table comparisons.

- We propose **TabXEval**, a two-phase LLM-based table evaluation method that aligns reference tables structurally and compares them semantically and syntactically via our rubric.

- We construct **TabXBench**, a diverse benchmark derived from multi-domain datasets, validating evaluation metrics through structured perturbations and human assessments.

- We analyze the strengths and weaknesses of existing evaluation methods via Sensitivity-Specificity Trade-off.

- We present TabXEval's qualitative and quantitative effectiveness in table generation task, enabling explainable automatic evaluation.

2 TabXEval

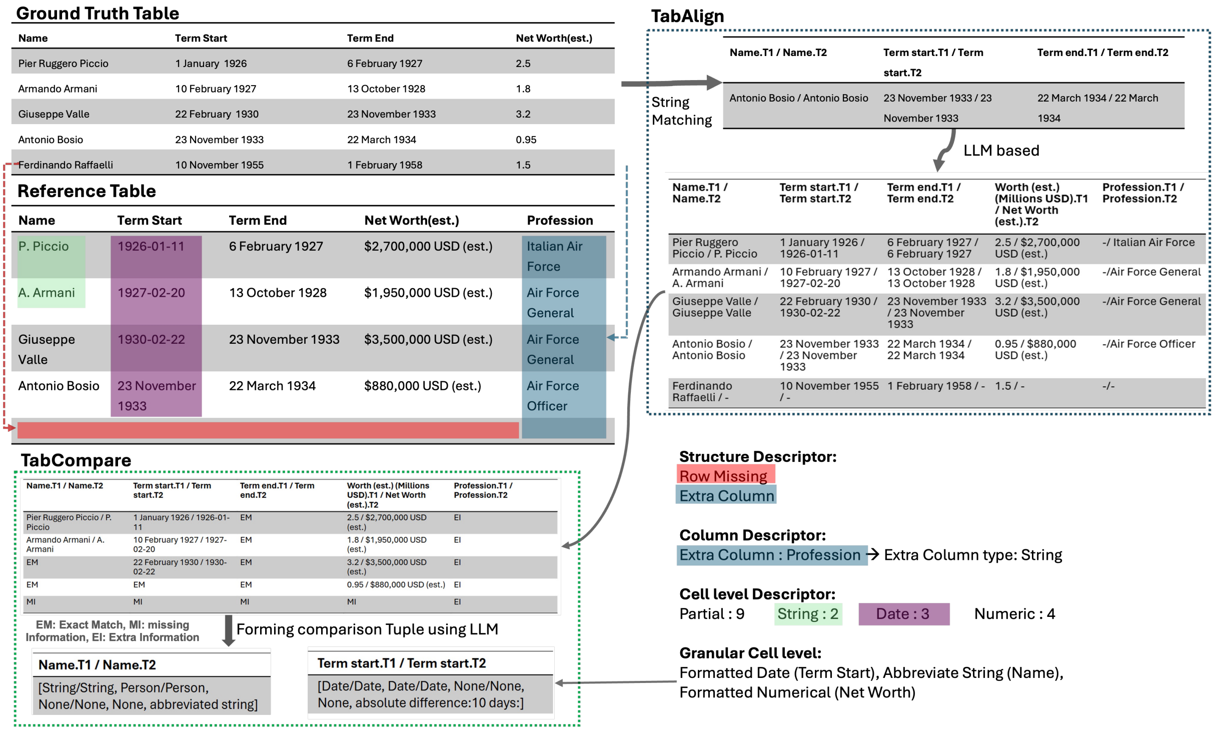
We establish a transparent and systematic protocol for table evaluation by comparing a reference table—the candidate table produced by a human or an LLM with a ground-truth table. These tables may differ in formatting, interpretation, or unit representation, necessitating a rigorous evaluation framework. For instance, an LLM-generated table might omit entire rows or abbreviate numeric values (e.g., “100k” vs. “100,000”), while a human-curated table may specify units only in the header (e.g., “velocity” vs. “velocity (m/s)”). To address such discrepancies, we design a set of rules based rubrics TabXEval, which aim to improve the reliability and interpretability of table evaluation.

2.1 Evaluation Rubric

Towards ensuring consistency and fairness in structure evaluation we provide an exhaustive rubric that enhances clarity and transparency in reference-based table evaluation. To quantify the degree of correctness and coverage of information at the most granular level, we advocate 4 categories of evaluation protocols:

Structure Descriptor coarsely assesses the overall structure against ground truth. This component compares information at the table level (missing information, extra information and exact matches), giving a high level description of information integrity.

Column Descriptor identifies data types of each column based on missing and extra information. This allows for a fine-grained strategy for evaluating cell values, as tables with heterogeneous columns like dates, numbers, strings etc should not be penalized on the same scale.



Ground Truth Table

Name	Term Start	Term End	Net Worth(est.)
Pier Ruggero Piccio	1 January 1926	6 February 1927	2.5
Armando Armani	10 February 1927	13 October 1928	1.8
Giuseppe Valle	22 February 1930	23 November 1933	3.2
Antonio Bosio	23 November 1933	22 March 1934	0.95
Ferdinando Raffaelli	10 November 1955	1 February 1958	1.5

Reference Table

Name	Term Start	Term End	Net Worth(est.)	Profession
P. Piccio	1926-01-11	6 February 1927	$2,700,000 USD (est.)	Italian Air Force
A. Armani	1927-02-20	13 October 1928	$1,950,000 USD (est.)	Air Force General
Giuseppe Valle	1930-02-22	23 November 1933	$3,500,000 USD (est.)	Air Force General
Antonio Bosio	23 November 1933	22 March 1934	$880,000 USD (est.)	Air Force Officer

TabAlign

Name.T1 / Name.T2	Term start.T1 / Term start.T2	Term end.T1 / Term end.T2
Antonio Bosio / Antonio Bosio	23 November 1933 / 23 November 1933	22 March 1934 / 22 March 1934

Name.T1 / Name.T2	Term start.T1 / Term start.T2	Term end.T1 / Term end.T2	Worth (est.) (Millions USD).T1 / Net Worth (est.).T2	Profession.T1 / Profession.T2
Pier Ruggero Piccio / P. Piccio	1 January 1926 / 1926-01-11	6 February 1927 / 6 February 1927	2.5 / $2,700,000 USD (est.)	-/ Italian Air Force
Armando Armani / A. Armani	10 February 1927 / 1927-02-20	13 October 1928 / 13 October 1928	1.8 / $1,950,000 USD (est.)	-/Air Force General
Giuseppe Valle / Giuseppe Valle	22 February 1930 / 1930-02-22	23 November 1933 / 23 November 1933	3.2 / $3,500,000 USD (est.)	-/Air Force General
Antonio Bosio / Antonio Bosio	23 November 1933 / 23 November 1933	22 March 1934 / 22 March 1934	0.95 / $880,000 USD (est.)	-/Air Force Officer
Ferdinando Raffaelli / -	10 November 1955 / -	1 February 1958 / -	1.5 / -	-/-

TabCompare

Name.T1 / Name.T2	Term start.T1 / Term start.T2	Term end.T1 / Term end.T2	Worth (est.) (Millions USD).T1 / Net Worth (est.).T2	Profession.T1 / Profession.T2
Pier Ruggero Piccio / P. Piccio	1 January 1926 / 1926-01-11	EM	2.5 / $2,700,000 USD (est.)	EI
Armando Armani / A. Armani	10 February 1927 / 1927-02-20	EM	1.8 / $1,950,000 USD (est.)	EI
EM	22 February 1930 / 1930-02-22	EM	3.2 / $3,500,000 USD (est.)	EI
EM	EM	EM	0.95 / $880,000 USD (est.)	EI
MI	MI	MI	MI	EI

EM: Exact Match, MI: missing Information, EI: Extra Information

Name.T1 / Name.T2
[String/String, Person/Person, None/None, None, abbreviated string]

Term start.T1 / Term start.T2
[Date/Date, Date/Date, None/None, None, absolute difference:10 days:]

Structure Descriptor:
Row Missing
Extra Column

Column Descriptor:
Extra Column : Profession → Extra Column type: String

Cell level Descriptor:
Partial : 9 String : 2 Date : 3 Numeric : 4

Granular Cell level:
Formatted Date (Term Start), Abbreviate String (Name), Formatted Numerical (Net Worth)

Figure 2: End-to-end schematic of **TabXEval**. (1) *TabAlign* aligns rows, columns, and cells using deterministic rules plus an LLM refinement loop. (2) *TabCompare* classifies each aligned cell as extra, missing, or partial and combines the counts with rubric weights (α, β, γ). This workflow populated the rubrics and outputs a table-level score and cell-level error trace, enabling fine-grained analysis.

Cell Level Descriptor looks at both semantic and syntactical representation of cell values. Other methods such as (Zhang* et al., 2020; Ramu et al., 2024; Gu et al., 2024) fail in recognizing different representations of the same data and hence can wrongly penalize correct information. This component of the rubric is necessary to craft an ideal table evaluation framework.

Granular Cell Level Difference that determines the magnitude of discrepancies between reference and ground truth table necessary to quantify instances w.r.t. cell level descriptions. It also captures variations by changing the format(e.g. m to cm, years to date) to report the absolute differences.

2.2 TabXEval Rubric

We propose **TabXEval**, a two-phase framework that combines deterministic rules and LLM-based analysis for robust and interpretable table evaluation. This design strikes a balance between precision (capturing exact matches) and flexibility (handling semantic or structural variations).

Phase 1: *TabAlign* matches columns, rows, and cells between the reference and candidate tables. We begin with exact string matching to establish a precise baseline alignment. Next, we refine this alignment using an LLM to account for abbreviations, synonyms, and structural transformations (e.g., merged columns, row/column transpositions). Purely exact matching can be overly strict, missing semantically equivalent but syntactically different cells. The LLM-driven refinement ensures a more comprehensive alignment while preserving high precision. Finally, we get an output table which has a combination of strict and relaxed mapping as shown in Figure 2.

Phase 2: *TabCompare* performs a fine-grained evaluation of the aligned tables. From the refined alignment, we extract table-level statistics (e.g., missing/extra rows or columns) and focus on partially matched cells. These cells are compared in detail using LLM-generated "comparison tuples" as shown in Figure 2 which capture numeric, string, date/time, and unit mismatches. We also compute magnitudes of differences (e.g., converting months to days) for precise reporting of discrepancies. Table-level summaries alone cannot uncover subtle cell-level errors, such as unit mismatches or minor numeric discrepancies. By combining table-level statistics with granular cell comparisons, TabXEval yields a more reliable and transparent assessment of content fidelity.

Score Our scoring function for TabXEval is defined as follows:

$$\text{TabXEval} = \sum_{I \in \{\text{Missing, Extra, Partial}\}} \beta_I \times \left(\sum_{E \in \{\text{row, column, cell}\}} \alpha_E \frac{f_E}{N_E} \right) \gamma_p \, .$$

where β_I is the weight assigned to each type of information error (I) such as Missing, Extra, and Partial; α_E is the weight for each entity type (E) including rows, columns, and cells; f_E represents the number of correctly matched entities; and N_E is the total number of entities in the ground truth.

For partial matches at the cell level, the modifier γ_p is defined as:

$$\gamma_p = \begin{cases} 1, & \text{if no partial cell,} \\ \omega_p \left| \frac{GT - Ref}{Ref} \right|, & \text{if partial cell detected.} \end{cases}$$

This formulation captures the multi-level nature of table evaluation insipred by proposed rubric. First aggregating errors across different information types (Missing, Extra, Partial) via the outer summation, and then evaluating the correctness at various entity levels (row, column, cell) using the inner summation. The term γ_p further refines the score by quantifying discrepancies in partially matched cells through a normalized absolute difference between the ground truth (GT) and the reference (Ref), ensuring that both coarse structural errors and fine-grained content differences are robustly accounted for in a single, interpretable metric. An illustrative example demonstrating the application of the above equations is provided in Appendix B.

Overall, TabXEval's two-phase structure ensures that both coarse (table-level) and fine (cell-level) differences are captured, providing an adaptable and explainable approach to table evaluation.

2.3 TabXBench Benchmark

Evaluating table metrics across diverse domains and error types remains a significant challenge due to the limited scope of existing datasets, which often focus on a single domain or contain only select data types. To bridge this gap, we introduce **TabXBench**, a controlled multi-domain test bed that comprehensively captures real-world nuances of tabular data generation and evaluation.

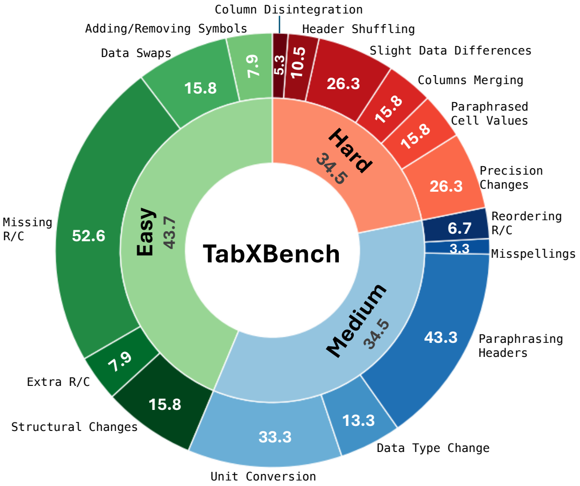


Figure 3: **Perturbation spectrum in TabXBench.** The outer ring enumerates the frequency (numeric labels) of the 16 fine-grained perturbation types applied to reference tables. The inner ring groups these edits into three difficulty bands *Easy* (light green, ≈44%), *Medium* (blue, ≈34%), and *Hard* (red, ≈35%).

TabXBench is designed to rigorously assess the sensitivity and specificity of reference-based table evaluation methods. Unlike prior datasets that narrowly target specific tasks (e.g., finance or sports), our benchmark spans multiple domains (finance, sports, knowledge bases, and more) while systematically incorporating an extensive range of potential table perturbations. This diversity ensures that TabXBench captures common pitfalls such as missing rows/columns, reordered headers, unit mismatches, numeric discrepancies, and complex structural variations (e.g., row/column transposition).

Dataset	# of Tables	# Perturb/Table	Average		
			Rows	Cols	Cells
FINQA	13	5	13.08	5.54	71.00
TANQ	8	5	8.38	5.38	44.25
ROTOWIRE	12	5	12.08	14.08	165.67
FETAQA	7	5	15.86	5.14	88.14
WIKITABLES	5	5	16.40	3.60	53.60
WIKISQL	6	5	8.33	5.17	42.17

Table 1: Composition of the TabXBench corpus across six source datasets.

TabXBench consists of 50 handpicked "clean" (reference) tables from six popular text-to-table and table QA datasets: RotoWire (Wiseman et al., 2017), TANQ (Akhtar et al., 2025), FetaQA (Nan et al., 2022), FinQA (Chen et al., 2021), WikiTable (Pasupat and Liang, 2015), and WikiSQL (Zhong et al., 2017) and the statistics can be found in Table 1. Each table was augmented with five distinct perturbations spanning over 16 error types, carefully curated to reflect common generation mistakes. The perturbations were first drafted with LLM assistance (e.g., reformatting numeric values, altering units, swapping headers) and subsequently validated by human experts for correctness and variety. We categorize the resulting perturbed tables into three difficulty levels (Easy, Medium, Hard) as demonstrated in Figure 3, ensuring coverage of both straightforward errors (e.g., minor typos) and complex structural manipulations (e.g., merged cells or shifted rows).

By offering a controlled yet diverse environment, TabXBench enables: **Fine-Grained Analysis**: Researchers can systematically evaluate how well metrics detect specific error types (e.g., unit mismatches vs. missing rows). **Sensitivity-Specificity Trade-offs**: The benchmark's difficulty levels and variety of perturbations allow detailed insights into a method's robustness and tolerance for minor vs. major errors. **Realistic Scoring Correlations**: It includes human annotations (aligned with our rubric in Section 2.1), enabling correlation studies that compare machine-generated scores to human judgments.

Finally, to illustrate TabXEval pragmatic value, we apply it to evaluate table outputs from three LLMs across four standard tasks - RotoWire, TANQ, WikiBio, and WikiTable highlighting how each evaluation method fares in a realistic setting. As a pioneering multi-domain resource, TabXBench thus provides a solid foundation for advancing research in reliable, explainable table evaluation.

3 Experiments

To validate efficacy of TabXEval, we conduct experiments using our synthetic dataset TabXBench. We report GPT-4o (OpenAI et al., 2024) and LLaMA-3.3-INSTRUCT results for our framework TabXEval for both components *TabAlign* and *TabCompare*.

Baselines. Our evaluation compares TabXEval against a broad range of baselines, which we classify into deterministic and non-deterministic approaches. Deterministic metrics (e.g., Exact Match (EM), chrF, ROUGE-L) yield fixed outputs based on string- or character-level comparisons, ensuring reproducibility. In contrast, non-deterministic metrics (e.g., BERTScore, BLUERT, H-Score) leverage contextualized neural representations to capture subtle linguistic nuances albeit with potential variability. We also include two recent methods: P-Score, an LLM-based metric outputting scores on a 0–10 scale, and TabEval, an embedding-based method that unrolls tables using an LLM and computes entailment via RoBERTa-MNLI. For a fair comparison, we further propose a Direct-LLM baseline that is prompted with the same evaluation rubric detailed in Section 2.1.

LLMs. Throughout our experiments, we used GPT-4o, Gemini 2.0-flash, and LLaMA-3.3-Instruct 70B. All models were executed with identical sampling settings (default temperature, top-k, and top-p) unless specified otherwise. All our prompts for TabAlign, TabCompare and Direct-LLM basline are given in Appendix C.

3.1 Human correlation

For each sample in the ground truth set of TabXBench, two human evaluators evaluated the tables using our proposed rubrics and guidelines. For both Human and TabXEval, we present ground-truth and a randomly selected perturbation (out of 5) to fill the rubrics. The detailed human annotation protocol is described in Appendix F. Finally these rubrics are compared using both Pearson (Sedgwick, 2012) and Kendall's Tau (Sen, 1968) correlation coefficients to quantify the degree of alignment between our method and human ratings. We observe very high correlation, i.e, $\mathbf{99.7}\%$ and $\mathbf{95.1}\%$ Pearson's ρ correlation for the Rubric Structure Descriptor and Cell Level Descriptor respectively.

In contrast, Direct LLM based baseline correlates, only $\mathbf{30.6}\%$ and $\mathbf{40.6}\%$ Pearson's ρ correlation for the Rubric Structure Descriptor and Cell Level Descriptor respectively. Revealing that it fails to understand the rubric, and quantify the structural and contextual challenges in table evaluation, and hence is unable to correctly align with humans correlation. Similarly we report $\mathbf{99.1}\%$ and $\mathbf{92.8}\%$ human correlation using Kendall's τ. While the baseline is $\mathbf{30.7}\%$ and $\mathbf{55}\%$ on Structure Descriptor and Cell Level Descriptor rubric respectively. These results demonstrate that disentangling the alignment and comparison phases is critical for robust table evaluation as Direct method falls short even hen presented with evaluation rubrics. TabXEval's correlation with human are consistent in capturing multi-level nuances for real-world challenges in table evaluation resembling humans.

3.2 Human Ranking Correlation Study

To further validate the robustness of our evaluation framework, we conducted a human ranking correlation study using outputs from TabXBench.

In this study, expert annotators ranked the quality of a ground-truth table and its perturbed variants each reflecting real-world errors such as structural, semantic, and formatting issues. These human rankings serve as our gold standard for assessing table quality.

For each table, we computed an aggregated score (based on cell-level f_1 measures) using various evaluation metrics, including deterministic baselines (e.g., Exact Match, chrf, ROUGE-L) and non-deterministic methods (e.g., BERTScore, H-Score, P-Score, BLUERT, TabEval), alongside our proposed TabXEval. Further, for baselines implementation we run both a single-step and multi-step LLM baseline to populate our proposed rubric tables that mirrors our two-stage TabXEval pipeline more closely along with LLM based ranking and multi-step LLM baseline with Chain-of-thoughts. We then measured the correlation between the automatic rankings and the human judgments using multiple metrics: Spearman's ρ, Kendall's τ, Weighted Kendall's $\tau^\dagger$, Rank-Biased Overlap (RBO), and Spearman's Footrule. These measures collectively assess both the overall ranking order and positional differences.

As shown in Table 2, TabXEval achieves the strongest correlation with human rankings across all metrics. Specifically,

Overall Ranking Order: TabXEval attains a Spearman's ρ of 0.44 a relative improvement of nearly 47% over the next-best method (P-Score at 0.30). Its Kendall's τ of 0.40 and Weighted Kendall's $\tau^\dagger$ of 0.38 further indicate strong monotonic agreement with human assessments.

Top-Weighted Agreement: With an RBO of 0.34, TabXEval demonstrates superior alignment in the higher-ranked items, compared to values ranging from 0.23 to 0.31 for other methods.

Positional Accuracy: TabXEval records the lowest Spearman's Footrule distance (0.29), reflecting minimal positional discrepancy relative to the human gold standard.

Notably, the TabEval method not only fails to capture these nuances as it negatively correlates, highlighting its inability to account for the multi-faceted nature of table quality. This analysis reinforces the importance of our two-phase approach, disentangling structural alignment (TabAlign) from detailed cell-level comparison (TabCompare) to effectively mirror human judgment. Finally, the human ranking correlation study clearly demonstrates that TabXEval provides a more robust, interpretable, and human-aligned evaluation of table outputs, capturing both coarse and fine-grained discrepancies that are critical in real-world scenarios.

3.3 What Sets TabXEval Apart?

A key criteria of any evaluation metric is to achieve a balance between *specificity* (i.e., avoiding false positives) and *sensitivity* (i.e., avoiding false negatives). In Figure 1, we visualize this trade-off by plotting each metric's specificity (y-axis) against its sensitivity (x-axis). The background colormap in the figure corresponds to the F1 score. Finally, Bubble Size represented by harmonic mean of specificity, sensitivity and accuracy, providing a quick visual cue for overall performance.

Goldilocks Zone for Ideal Metrics. Metrics positioned in the top-right portion of the chart (the green-shaded "Goldilocks zone") demonstrate the desired trait of consistently identifying correct table content (*high sensitivity*) while minimizing the likelihood of falsely flagging errors (*high specificity*). TabXEval resides firmly in this zone, illustrating its balanced performance across diverse table perturbations.

Comparisons with Other Metrics. We compare **TabXEval** against several widely used metrics: (1) P-Score performs well at a high level and sits near the Goldilocks zone in our evaluations, reflecting strong table-level correctness. However, it lacks *explainability* and *granular insights*, offering only a single 0–10 score that limits interpretability and error traceability. (2) H-Score and BERTScore better capture semantics than string-based metrics such as EM, ROUGE-L, and chrF, but often overlook structural errors like swapped columns or missing rows—resulting in moderate sensitivity but poor specificity. (3) TabEval uses entailment over LLM-generated atomic statements, but frequently misses fine-grained numeric or unit mismatches. In our experiments, it produced false positives on tables that were re-formatted yet semantically equivalent.

Why TabXBench Matters. TabXBench introduces diverse table perturbations (e.g., missing rows/columns, numeric/unit mismatches) to stress-test metrics across real-world errors (Figure 3). Simple metrics like *Exact Match* fail under reordering

Metrics	Spearman's $\rho \uparrow$	Kendall's $\tau \uparrow$	W-Kendall's $\tau^{\dagger} \uparrow$	RBO $\uparrow$	Spearman's Footrule $\downarrow$
EM	0.18	0.16	0.16	0.26	0.57
chrF	0.12	0.11	0.08	0.25	0.59
H-Score	0.14	0.11	0.09	0.28	0.51
BERTScore	0.19	0.15	0.13	0.25	0.57
ROUGE-L	0.21	0.18	**0.40**	0.27	0.53
BLEURT	0.29	0.25	0.25	0.27	0.51
TabEval	-0.04	-0.04	-0.03	0.23	0.63
P-Score	0.30	0.27	0.24	0.31	0.39
LLM rubric	0.23	0.16	0.17	0.28	0.47
LLM ranking	0.29	0.24	0.23	0.30	0.41
Multi-prompt	0.29	0.24	0.23	0.30	0.42
Multi-prompt + CoT	0.30	0.25	0.24	0.29	0.45
TabXEval	**0.44**	**0.40**	0.38	**0.34**	**0.29**

Table 2: Correlation between automatic rankings and human judgments. Higher Spearman's ρ, Kendall's τ, Weighted Kendall's $\tau^{\dagger}$, and RBO values indicate better agreement, while lower Spearman's Footrule values are preferable.

or semantic shifts, while LLM/embedding-based methods miss unit mismatches or partial errors. TabXEval's *two-phase* approach first aligning structure (TabAlign), then systematically comparing content (TabCompare) ensures precise discrepancy detection aligned with human judgment.

Significance of TabXEval. By balancing sensitivity, specificity, and F1 scores, TabXEval not only outperforms across evaluation dimensions but also *explains* its judgments via structured rubrics. This transparency is crucial for financial reporting, scientific validation, and knowledge curation, where subtle errors can be costly. Identifying *what* went wrong and *where*, TabXEval provides both quantitative and qualitative insights for improving table generation.

In summary, TabXEval's interpretability, robust performance, and alignment with TabXBench's challenging setup establish it as the new standard for explainable, human-aligned table evaluation.

3.4 Performance Analysis

The results from our Endurance Test on Text-to-Table Generation as depicted in Table 3 and Table 4 clearly demonstrate how TabXEval's two-phase evaluation framework enables us to drill down from overall table structure to fine-grained cell details.

Table-Level Performance Across datasets, `GPT-4o` frequently exhibits higher exact match scores (EM) for both rows and columns compared to `LLaMA-3.3` and `Gemini-2.0-flash`. For instance, on WikiTables, `GPT-4o` achieves a row EM score of 27.11, surpassing the performance of the other models. Additionally, `GPT-4o` consistently maintains low Extra Information (EI) values at the column level (e.g., only 0.07 EI on WikiTables), indicating that it preserves the intended table structure with minimal unintended additions. Such metrics underscore the model's ability to capture overall table integrity across diverse datasets, including WikiBio, TANQ, and RotoWire.

Cell-Level Performance A closer examination at the cell level reveals further nuances. In datasets such as WikiTables and WikiBio, GPT-4o records significantly fewer errors in string cells; for example, its string EI on WikiTables is only 1.17 compared to 4.33 for `LLaMA-3.3`. On TANQ, both `GPT-4o` and `Gemini-2.0-flash` show lower partial errors in numerical and string cell types relative to `LLaMA-3.3`, suggesting more robust semantic and syntactic matching. Notably, on RotoWire, `GPT-4o` also demonstrates lower partial error counts in both numerical and string cells when compared with `Gemini-2.0-flash`. These detailed cell-level insights are crucial as they highlight the models' abilities to handle fine-grained discrepancies such as unit mismatches or subtle formatting errors. This course-to-fine grain evaluation not only facilitates the identification of specific error types but also offers interpretable insights into the strengths and weaknesses of each model.

Adaptability and Robustness Evaluation Figure 4 illustrates that TabXEval consistently outperforms existing metrics in human-correlation across a broad range of weighting schemes (Appendix A Figure 5). Each box plot corresponds to 2^6 permutations of weights for key dimensions (e.g., missing, extra, row, column, cell, partial); these weights contribute to Equations 2.2 and 2.2, set either as perm(0)=$\{0, 1\}$ or perm(0.25)=$\{0.25, 1\}$. This flexibility allows TabXEval to adapt to dataset-specific priorities (e.g., penalizing missing rows more heavily) while maintaining robust, domain-agnostic performance. Notably, our best-

	LLaMA-3.3 70B							GPT-4o							Gemini-2.0-flash						
Stat	Num	String	Bool	Date	List	Time	Others	Num	String	Bool	Date	List	Time	Others	Num	String	Bool	Date	List	Time	Others
										WikiTables											
EI	0.05	4.33	0.00	0.17	0.00	0.00	0.13	0.03	**1.17**	0.00	0.02	0.00	0.00	0.13	0.02	2.33	0.00	0.11	0.00	0.00	0.07
MI	0.01	0.80	0.00	0.03	0.00	0.00	0.00	0.01	0.93	0.00	0.00	0.00	0.00	0.01	0.00	0.73	0.00	0.00	0.00	0.00	0.00
Partial	0.22	25.00	0.00	0.35	0.00	0.01	0.09	**0.32**	20.34	0.00	0.55	0.00	0.02	0.10	0.30	22.50	0.00	0.48	0.00	0.02	0.07
										WikiBio											
EI	0.04	2.84	0.00	0.09	0.00	0.00	0.12	0.04	2.02	0.00	0.03	0.00	0.00	0.09	0.04	2.30	0.00	0.07	0.00	0.00	0.06
MI	0.02	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.01	0.00	0.00	0.00	0.00	0.37	0.00	0.01	0.00	0.00	0.00
Partial	0.16	14.38	0.00	2.60	0.00	0.00	0.03	0.16	15.80	0.00	0.86	0.00	0.00	0.03	0.15	13.59	0.00	2.97	0.00	0.00	0.04
										TANQ											
EI	0.05	0.84	0.00	0.16	0.09	0.00	0.00	0.02	0.18	0.00	0.06	0.03	0.01	0.00	0.00	0.29	0.00	0.07	0.02	0.00	0.00
MI	0.01	0.24	0.00	0.11	0.01	0.04	0.00	0.01	0.08	0.00	0.05	0.02	0.01	0.00	0.02	0.21	0.00	0.05	0.00	0.02	0.00
Partial	2.73	20.50	0.00	4.72	4.82	2.19	0.07	**1.28**	**11.92**	0.00	3.48	3.46	1.32	0.01	**1.22**	**9.35**	0.00	2.02	2.64	1.12	0.02
										RotoWire											
EI	0.84	0.50	0.00	0.00	0.00	0.00	0.01	0.68	0.23	0.00	0.00	0.00	0.00	0.00	1.31	0.54	0.00	0.00	0.00	0.00	0.09
MI	0.97	0.32	0.00	0.00	0.00	0.00	0.04	0.56	0.08	0.00	0.00	0.00	0.00	0.00	1.06	0.28	0.00	0.00	0.00	0.00	0.02
Partial	**0.66**	**0.92**	0.00	0.00	0.00	0.00	0.00	**0.31**	**0.69**	0.00	0.00	0.00	0.00	0.00	**2.72**	**3.87**	0.00	0.00	0.00	0.00	0.03

Table 3: **Cell-Level Performance Analysis** of Extra (EI), Missing (MI), and Partial mismatches across data types numerical, string, boolean, date, list, time, and other for WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o shows fewer string EI in WikiTables and lower partial errors in numerical and string cells in TANQ and RotoWire.

	LLaMA-3.3 70B			**GPT-4o**			**Gemini-2.0-flash**		
	MI	**EI**	**EM**	**MI**	**EI**	**EM**	**MI**	**EI**	**EM**
					WikiTable				
Row	8.69	15.82	25.59	23.44	11.51	**27.11**	20.81	10.67	26.03
Col	0.37	0.92	1.67	**4.47**	**0.07**	1.02	2.97	0.37	1.55
					WikiBio				
Row	25.16	29.33	16.17	26.09	27.63	19.39	30.08	24.38	16.89
Col	0.10	0.0	0.05	0.05	0.025	0.0	0.12	0.0	0.0
					TANQ				
Row	7.6	5.83	10.97	8.27	2.80	13.00	8.51	4.01	13.01
Col	2.69	1.82	23.89	2.24	0.19	22.42	2.82	0.63	21.78
					RotoWire				
Row	3.48	39.22	17.62	1.32	28.65	38.41	3.10	22.82	13.80
Col	10.87	16.21	52.70	17.57	5.71	60.24	16.35	10.52	48.51

Table 4: **Table-Level Performance Analysis**: Row/Column MI, EI, and EM rates on WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o leads with highest Row EM (27.11) and lowest Col EI (0.07) on WikiTables.

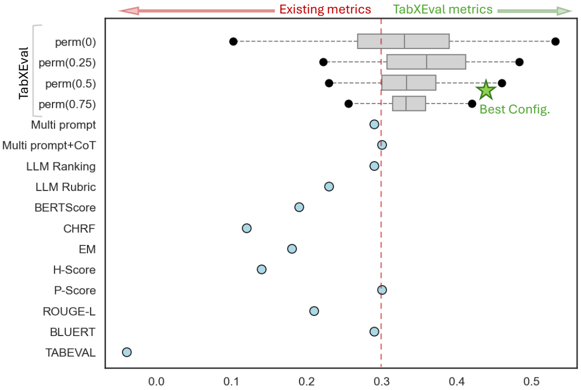


Figure 4: Human Ranking Correlation. This plot compares existing metrics (with the best performance indicated by the red dashed line) against various configurations of our TabXEval metric. The green star highlights the best performing TabXEval configuration.

performing configuration (marked by the star) demonstrates both high average correlation and low variance, underscoring TabXEval's capacity to balance fine-grained descriptors with overall structural fidelity.

Complementing this, Table 5 demonstrates TabXEval's backbone-agnostic robustness: it achieves strong alignment with human rankings across diverse LLMs, including GPT-4o, LLaMA, Qwen, and Gemini. Even under varying model architectures, TabXEval maintains a correlation of ≥ 0.30 across all metrics, reinforcing its reliability beyond just weighting flexibility. A detailed qualitative analysis, provided in Appendix D, further demonstrates TabXEval's effectiveness across both domain-agnostic and specific weighting configurations.

4 Comparison with Related Work

Text-to-Table Generation. Early text–to–table research exploited single–domain corpora such as RotoWire for basketball summaries (Wiseman et al., 2017), the E2E restaurant set (Novikova et al., 2017), WikiBio infobox–biography pairs (Lebret et al., 2016), and WikiTableText (Pasupat and Liang, 2015). While pioneering, these resources offer limited schema variety and often encourage hallucinated or under-structured outputs. Recent collections address these gaps: StructBench permutes headers, merges columns, and shuffles schemas to test structural generalisation (Gu et al.,

Metrics	Spearman's $\rho \uparrow$	Kendall's $\tau \uparrow$	W-Kendall's $\tau^{\dagger} \uparrow$	RBO $\uparrow$	Spearman's Footrule $\downarrow$
P-Score (GPT-4o)	0.30	0.27	0.24	0.31	0.39
BleuRT	0.29	0.25	0.25	0.27	0.51
TabXEval (GPT-4o)	**0.44**	**0.40**	**0.38**	**0.34**	**0.29**
TabXEval (LLaMA)	0.37	0.30	0.30	0.29	0.44
TabXEval (Qwen)	0.33	0.27	0.28	0.30	0.38
TabXEval (Gemini)	0.30	0.23	0.21	0.30	0.40

Table 5: **Robustness of TabXEval across LLM back-bones.** Human-ranking correlations for TabXEval run with `GPT-4o`, `LLaMA-3.3 70B-Instruct`, `Qwen-72B-Instruct` and `Gemini-2.0-pro`, compared to BLEURT and the P-Score baseline. TabXEval + `GPT-4o` attains the strongest alignment (Spearman's $\rho = 0.44$), but all back-bones retain a ≥ 0.30 correlation, indicating backbone-agnostic reliability.

2024), whereas TanQ requires multi-hop, multi-source reasoning to generate answer tables (Akhtar et al., 2025). Such challenging benchmarks expose systematic weaknesses in both generation models and legacy evaluation metrics, motivating the fine-grained rubric employed by TabX.

Other Evaluation Metrics *1. Surface and embedding overlap.* Classic n-gram scores BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015) treat a table as flat text, ignoring header alignment or cell hierarchy. Embedding-based BERTScore (Zhang* et al., 2020) improves semantic sensitivity, yet still overlooks structural fidelity. Token-level Exact Match and PARENT (Dhingra et al., 2019) partially reward factual grounding, but cannot detect column swaps or unit shifts.

2. Structure-aware and reference-less scores. StructBench introduces H-Score and P-Score, targeting hierarchical integrity and holistic quality, respectively (Gu et al., 2024). TabEval ("Is this a bad table?") decomposes each table into atomic statements and uses textual entailment to capture fine-grained errors (Ramu et al., 2024). Complementarily, Data-QuestEval dispenses with references altogether by generating and answering questions directly over the source data, achieving strong human correlation in data-to-text tasks (Rebuffel et al., 2021). We further show that our approach performs robustly across a wide range of table structures and can be extended to handle hierarchical formats, as discussed in Section E.

Despite these advances, existing metrics still emphases either semantics or structure and provide a single numeric values with limited error traceability or explainability. Our two-phase TabX closes this gap by disentangling alignment (*TabAlign*) from cell-level comparison (*TabCompare*), producing an interpretable score that balances sensitivity and specificity.

5 Conclusion and Future Work

In this work, we have introduced TabXEval an eXhaustive and eXplainable, two-phase framework that transforms table evaluation by disentangling structural alignment from detailed cell-level comparison. Our method leverages a comprehensive rubric to quantify both coarse and fine-grained errors, yielding results that strongly correlate with human assessments. By developing TabXBench, a challenging multi-domain benchmark with diverse perturbations, we have demonstrated the robustness, explainability, and human-alignment of our approach. While limitations such as computational overhead and handling of hierarchical tables remain, the promising performance of TabXEval opens avenues for further research and refinement in automatic table evaluation.

We found two key directions for future work. Firstly, while our current approach leverages large language models to ensure robustness and generalization to unseen table structures, their computational overhead can hinder practical deployment. Developing more compact alternatives using smaller models (e.g., BART, T5) would require large-scale, heterogeneous fine-tuning data to preserve performance on out-of-distribution tables—a resource that is currently limited. Future efforts could explore model distillation or semi-supervised training to create lightweight yet reliable variants of TabXEval. Secondly, although TabXEval effectively handles structural variations—such as merged or decomposed cells—through LLM-guided alignment, directly supporting complex hierarchical tables (e.g., nested cells or multi-level headers) remains a challenge. A promising direction is to incorporate a preprocessing step that flattens hierarchical structures into standardized key paths (e.g., transforming "Release Date" with sub-headers "Month" and "Year" into "Release Date.Month" and "Release Date.Year"), enabling compatibility with our approach while preserving semantic structure.

Limitations

While TabXEval demonstrates strong performance, it is not without its limitations. Its reliance on large-scale language models comes at the cost of increased computational overhead, which may impact scalability. Moreover, the method faces challenges when dealing with hierarchical tables, where nested headers and multi-level groupings make alignment significantly more complex. Lastly, as a reference-based evaluation approach, TabXEval necessitates access to ground-truth tables, leaving the question of referenceless evaluation an open and compelling challenge for future research.

Ethics Statement

The authors affirm that this work adheres to the highest ethical standards in research and publication. Ethical considerations have been meticulously addressed to ensure responsible conduct and the fair application of computational linguistics methodologies. Our findings are aligned with experimental data, and while some degree of stochasticity is inherent in black-box Large Language Models (LLMs), we mitigate this variability by maintaining fixed parameters such as temperature, top_p, and top_k. Furthermore, our use of LLMs, including `GPT-4o`, `Gemini`, and `LLaMA`, complies with their respective usage policies. To refine the clarity and grammatical accuracy of the text, AI based tools such as Grammarly and ChatGPT were employed. Additionally, human annotators who are also among the authors actively contributed to data labeling and verification, ensuring high-quality annotations. To the best of our knowledge, this study introduces no additional ethical risks.

Acknowledgements

We thank the Complex Data Analysis and Reasoning Lab at Arizona State University for computational support, and Krishna Singh Rajput for his support during the submission process. Lastly, we thank our lab cat, Coco, for her unwavering commitment to keeping our professor creatively unhinged during deadline season.

References

Mubashara Akhtar, Chenxi Pang, Andreea Marzoca, Yasemin Altun, and Julian Martin Eisenschlos. 2025. TANQ: An open domain dataset of table answered questions. *Preprint*, arXiv:2405.07765.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. 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.

Zhouhong Gu, Haoning Ye, Xingzhou Chen, Zeyang Zhou, Hongwei Feng, and Yanghua Xiao. 2024. StrucText-Eval: Evaluating Large Language Model's Reasoning Ability in Structure-Rich Text. *Preprint*, arXiv:2406.10621.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent Human Evaluation for Image Captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478, Seattle, United States. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

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.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E Dataset: New Challenges For End-to-End Generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

OpenAI, :, and Aaron Hurst et. al. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

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.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Pritika Ramu, Aparna Garimella, and Sambaran Bandyopadhyay. 2024. Is this a bad table? a closer look at the evaluation of table generation from text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22206–22216, Miami, Florida, USA. Association for Computational Linguistics.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A Referenceless Metric for Data-to-Text Semantic Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philip Sedgwick. 2012. Pearson's correlation coefficient. *Bmj*, 345.

Pranab Kumar Sen. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American statistical association*, 63(324):1379–1389.

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.

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

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Appendix

A Performance Analysis

Figure 5 represent human ranking correlation of TabXEval across various confirgrations of parameters.

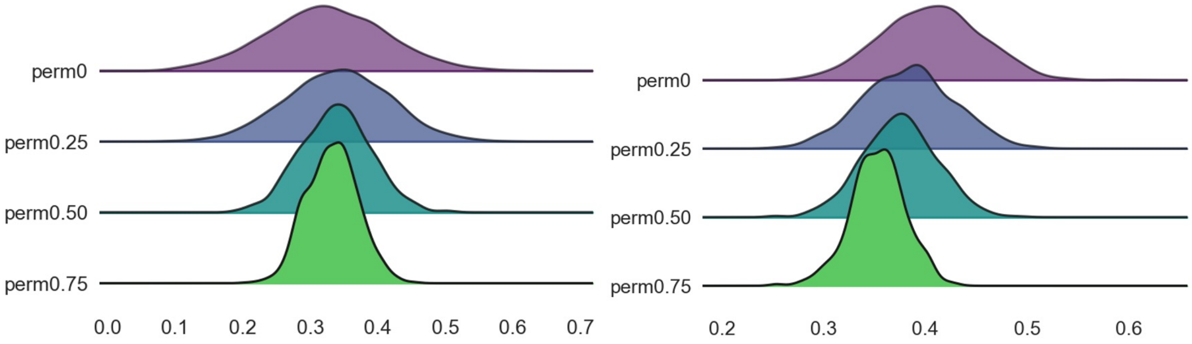


Figure 5: Human Ranking Correlation, ours all configurations.

B Illustration Example

Below is a detailed breakdown of our scoring method. For example we are evaluating two movie examples where the Ground Truth Table (GT): 5x5 table and a Reference Table (Ref): 4x6 table.

$$\textbf{TabXEval} = \sum_{I \in \{\text{Missing, Extra, Partial}\}} \beta_I \left(\sum_{E \in \{\text{row, column, cell}\}} \alpha_E \frac{f_E}{N_E} \right) \times \gamma_p \tag{1}$$

Partial Score Weight Definition

$$\gamma_p = \begin{cases} 1, & \text{if no partial cell,} \\ \omega_p \times \left| \dfrac{GT - Ref}{Ref} \right|, & \text{if partial detected.} \end{cases}$$

Observed Errors

- **Missing Row:** The reference table is missing one row compared to the ground truth.
- **Extra Column:** The reference table has an additional column about directors.
- **Partial Match in a Cell:** One cell in the reference table (release date) has a typo and increases the date by 2 days.

Weights

Entity Weights

- Column: $\alpha_{\text{column}} = 1$
- Row: $\alpha_{\text{row}} = 0.9$
- Cell: $\alpha_{\text{cell}} = 0.8$

Information Error Weights

- Missing: $\beta_{\text{Missing}} = 1$
- Extra: $\beta_{\text{Extra}} = 0.9$
- Partial: $\beta_{\text{Partial}} = 0.8$

Gamma Modifier for Partial Matches

For numerical data, we use a weight $\omega_p = 0.9$ and compute a normalized difference:

$$\left|\frac{GT - Ref}{Ref}\right| \;=\; 0.4.$$

Thus,

$$\gamma_p = 0.9 \times 0.4 = 0.36.$$

Our proposed rubric is domain agnostic, the beauty of our work is that our weighting scheme is flexible which makes the evaluation metric domain specific based on domain knowledge, for instance one can set higher weights to cell values with numeric types for financial data, in contrast the same rubric can be tuned for sports data where structural nuances should be penalized. Moreover, evidenced by our findings (Figure 4), highlights the strength and generalizability of the underlying scoring mechanism.

Scoring:

Missing Row:

- Weight for rows: $\alpha_E = 0.9$
- Total rows in GT: 5
- Weight for missing: $\beta_I = 1$

$$\text{Contribution to error} = \alpha_E \times \beta_I \times \frac{\text{Number of missing rows}}{\text{Total rows}} = 0.9 \times 1 \times \frac{1}{5} = 0.18$$

Extra Column:

- Weight for columns: $\alpha_E = 1$
- Weight for extra: $\beta_I = 0.9$
- Total columns in GT: 5

$$\text{Contribution to error} = \alpha_E \times \beta_I \times \frac{1}{5} = 1 \times 0.9 \times \frac{1}{5} = 0.18$$

Partial Match (Cell):

- Weight for cells: $\alpha_E = 0.8$
- Weight for partial: $\beta_I = 0.8$
- Total cells in GT: $5 \times 5 = 25$
- $\gamma_p = 0.36$

$$\text{Contribution to error} = \alpha_E \times \beta_I \times \frac{1}{25} \times \gamma_p = 0.8 \times 0.8 \times \frac{1}{25} \times 0.36 = 0.0088$$

Total Error Score:

$$0.18 + 0.18 + 0.0088 = \boxed{0.368}$$

By adjusting the weights assigned to entity types (rows, columns, cells), information errors (missing, extra, partial), and partial matches, **the framework can be fine-tuned to emphasize aspects critical to particular applications.**

C Prompts

We provide detailed prompts for TabAlign in Figure 6, TabCompare in Figure 7 and Direct-LLM Baseline in Figure 8.

D Qualitative Example

Figure 9 presents sample perturbed tables from the TabXBench benchmark, illustrating domain-specific corruptions across varying difficulty levels. It includes qualitative examples and a performance comparison of **TabXEval** against prior metrics.

E Hierarchial Tables

Though direct processing of complex hierarchical tables as shown in table 6 (e.g., nested merged cells, multi-level headers) is currently a limitation, we've ideated a viable workaround. We can employ a preliminary structure decoding step, as illustrated in table 7, which effectively unrolls hierarchical headers into a flattened format. For example, parent header 'Release Date' and sub-headers 'Month'/'Year' transform into 'Release Date.Month' and 'Release Date.Year.' This standardized table representation is then fully compatible with our evaluation framework, accommodating both strict and relaxed mapping criteria

Table 6: Hierarchical table

Movies	Parts	Release Date		Ratings
		Month	Year	
When we die	Part 1	May	2010	5
	Part 2	December	2022	3

Table 7: Flattened formatted table

Movies $\vee$ Parts	Release Date $\wedge$ Month	Release Date $\wedge$ Year	Ratings
When we die $\vee$ Part 1	May	2010	5
When we die $\vee$ Part 2	December	2022	3

TabAlign Prompt

INSTRUCTIONS:

Given **three tables** — **Table 1**, **Table 2**, and a **Partially Aligned Table** — your goal is to **align Table 1 and Table 2** using the Partially Aligned Table and output a **Final Aligned Table** in **Markdown**.

Follow this general **algorithm**:

1. **Compare** Table 1 and Table 2, keep differences in mind.
2. If one table appears to be a **transpose** of the other, take the transpose to match structures.
3. Use the **Partially Aligned Table** to align remaining rows and columns.
4. If a row or column **cannot be matched**, keep it as **extra** and fill with **-** (dash).
5. Handle **multiple possible mappings** carefully (multi-mapping).
6. Ensure the **Partially Aligned Table** is part of the final output.
7. Place **unmatched** rows/columns at the **end** of the table.
8. Recheck for correct alignment of **columns, rows, and cells**.
9. Include **all** cells from both Table 1 and Table 2 in the Final Aligned Table.
10. **Do not omit** any columns from either table.

If the Partially Aligned Table is **None**, simply perform alignment without it. Output **only** the final aligned table in Markdown (no extra text).

Format for the Final Aligned Table:

- Each cell is written as **`cell1/cell2`** (where `cell1` is from Table 1 and `cell2` from Table 2).
- If a value is **missing** in one table, use a dash: **`-/cell2`** or **`cell1/-`**.
- Each row must have the **same number of columns**.
- Do **not add** columns that do not exist in either table.

Examples:

Table 1:
| Year | Competition | Venue | ...
| ... | ... | ... | ...

Table 2:
| Year | Place | Country | ...
| ... | ... | ... | ...

Partially Aligned Table:
| Year.T1/Year.T2 | Position.T1/Position.T2 | ...
| 2011/2011 | 2nd/2nd | ...
| ... | ... | ...

Output (shortened):
| Year.T1/Year.T2 | Competition.T1/- | Venue.T1/Place.T2 | ... |
| ... | ... | ... | ... |

The final alignment shows all **matched** and **unmatched** columns/rows, with **slashes** and **dashes** where appropriate.

...

Figure 6: Prompt for tabular alignment, leveraging Partially Aligned Table and Reference Tables to generate a final structured table while preserving unmatched elements.

TabCompare Prompt

INSTRUCTIONS:
You are given a table where each cell is "`value1/value2`". Determine differences between value1 (from Table 1) and value2 (from Table 2). If a cell is "`-`", skip it; if both parts are empty, output "`[-]`". Use the column headers (also "`header1/header2`") as context for:

1. **Data Type** (Numerical, String, List, Date, Time, Boolean, Others, Empty)
2. **Entity** (Person, Organization, Location, Date, Time, Money, Percent, Facility, Event, Product, Work of Art, Language, Nationality, Ordinal, Cardinal, Others)
3. **Unit** (determine from context or values; if none, use "`None`")
4. **Missing/Extra Info** (e.g., if something appears only in one part)
5. **Difference** (format depends on Data Type: numerical → absolute difference, date → difference in days, time → difference in seconds, etc.)

For each cell, output a **5-element tuple**:

```
[DataType1/DataType2, Entity1/Entity2, Unit1/Unit2, Missing/Extra Info, Difference]
```

Output: **Only** the final table (Markdown) with these tuples, keeping the same structure as the input table. No extra explanations.

EXAMPLES:

Example 1
Input Table:

```
| Director.T1/Director.T2 | Writer.T1/Writer.T2 | Original Air Date/Air Date | Production
Code/Prod. Code |
|:------------------------|:--------------------|:---------------------------|:-------------
--------------|
| Richard Dale/R. Dale    | Tim Loane/T. Loane  | 21 March 2001/21/03/2001   | 101/106
            |
| ...                     | ...                 | ...                        | ...
            |
```

Output Table:

```
| Director.T1/Director.T2                                              | Writer.T1/Writer.T2
                                        | Original Air Date/Air Date
              | Production Code/Prod. Code                                     |
|:--------------------------------------------------------------------|:--------------------
---------------------------------------|:---------------------------------------------------
--------------|:-------------------------------------------------------------|
| [String/String, Person/Person, None/None, None, abbreviated string:Richard Dale -> R. Dale]
| [String/String, Person/Person, None/None, None, abbreviated string:Tim Loane -> T. Loane] |
[Date/Date, Date/Date, None/None, None, absolute difference:0:days:] | [Numerical/Numerical,
Cardinal/Cardinal, None/None, None, ab                                                    |
...                                                                 | ...
                                    |
```

Figure 7: Prompt for identifying data type, entity, and unit differences between two tables, outputting structured tuples to capture variations in numerical, string, date, and categorical values.

Baseline comparison prompt

You are an advanced data comparison and analysis engine. Your task is to receive two tables (GT and Generated) and perform a thorough comparison to extract difference metrics. Output four statistics:

Row and Column Stats
For each, report EM, MI, EI ...
Format:

```
| Type   | MI  | EI  | EM  |
|--------|-----|-----|-----|
| Row    | ... | ... | ... |
| Column | ... | ... | ... |
```

1.

Detailed Column Stats
Analyze each column by data type (Numerical, String, Bool, Date, List, Time, Others) ...
Format:

```
|           | Numerical | String | Bool | Date | List | Time | Others|
|-----------|-----------|--------|------|------|------|------|-------|| EI        | ...
  | ...     | ...  | ...  | ...  | ...  | ...    |
| MI        | ...        | ...    | ...  | ...  | ...  | ...  | ...   |
```

2.

Detailed Cell Stats
Exclude missing/extra rows/columns, then classify cell differences (MI, EI, Partial) ...
Format:

```
| Category | Numerical | String | Bool | Date | List | Time | Others |
|----------|-----------|--------|------|------|------|------|--------|
| MI       | ...       | ...    | ...  | ...  | ...  | ...  | ...    |
| EI       | ...       | ...    | ...  | ...  | ...  | ...  | ...    |
| Partial  | ...       | ...    | ...  | ...  | ...  | ...  | ...    |
```

3.

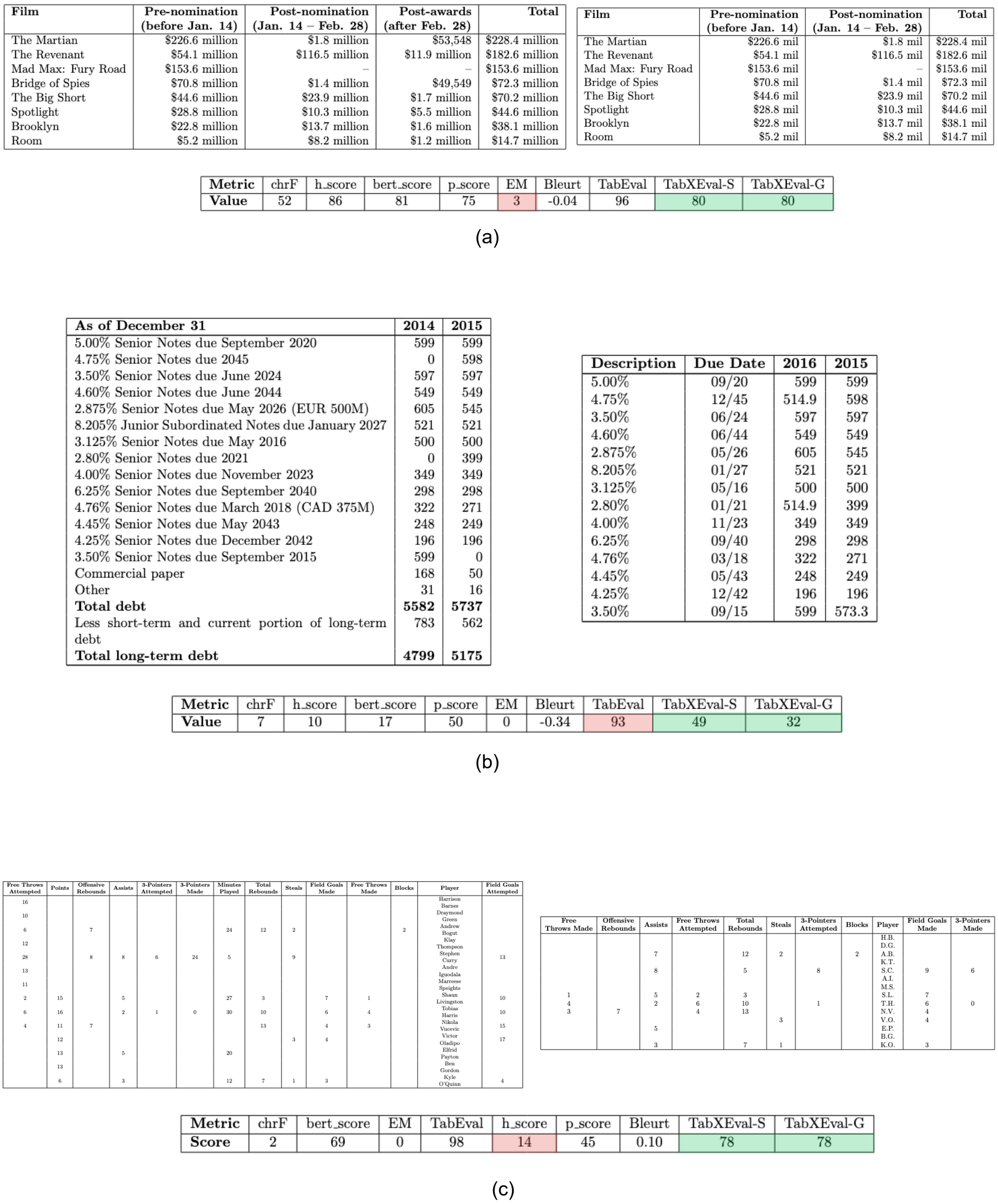
Cell Level Difference with Magnitude
For nuanced differences (unit, format, etc.), output JSON with ...
Example Structure:

```
{
  "bool": { "same": ..., "different": ... },
  "Date": { "date": ..., "time": ... },
  "List": { "MI": ..., "EI": ..., "EM": ... },
  "Distance (yards).T1/Distance (meters).T2": {
    "Numerical": { "unit_mismatch": ..., "ner_mismatch": ..., "delta": ..., "MI": ..., "EI":
... },
    "String": { "ner_mismatch": ..., "spell_errors": ..., "abbreviated_string": ...,
"semantically": { "same": ..., "different": ... }, "other": ..., "MI": ..., "EI": ... }
  }
}
```

4. **Additional Notes:**
- MI: ...
- EI: ...
- EM: ...
- Partial matches: ...

OUTPUT:
Only output the 4 tables (Row and Column Stats, Detailed Column Stats, Detailed Cell Stats, and Cell Level Difference with Magnitude) in the specified formats with no extra text.

Figure 8: Baseline comparison prompt for evaluating differences between ground truth (GT) and generated data, providing structured metrics for row, column, and cell-level analysis.



Film	Pre-nomination (before Jan. 14)	Post-nomination (Jan. 14 – Feb. 28)	Post-awards (after Feb. 28)	Total
The Martian	$226.6 million	$1.8 million	$53,548	$228.4 million
The Revenant	$54.1 million	$116.5 million	$11.9 million	$182.6 million
Mad Max: Fury Road	$153.6 million	–	–	$153.6 million
Bridge of Spies	$70.8 million	$1.4 million	$49,549	$72.3 million
The Big Short	$44.6 million	$23.9 million	$1.7 million	$70.2 million
Spotlight	$28.8 million	$10.3 million	$5.5 million	$44.6 million
Brooklyn	$22.8 million	$13.7 million	$1.6 million	$38.1 million
Room	$5.2 million	$8.2 million	$1.2 million	$14.7 million

Film	Pre-nomination (before Jan. 14)	Post-nomination (Jan. 14 – Feb. 28)	Total
The Martian	$226.6 mil	$1.8 mil	$228.4 mil
The Revenant	$54.1 mil	$116.5 mil	$182.6 mil
Mad Max: Fury Road	$153.6 mil	–	$153.6 mil
Bridge of Spies	$70.8 mil	$1.4 mil	$72.3 mil
The Big Short	$44.6 mil	$23.9 mil	$70.2 mil
Spotlight	$28.8 mil	$10.3 mil	$44.6 mil
Brooklyn	$22.8 mil	$13.7 mil	$38.1 mil
Room	$5.2 mil	$8.2 mil	$14.7 mil

Metric	chrF	h_score	bert_score	p_score	EM	Bleurt	TabEval	TabXEval-S	TabXEval-G
Value	52	86	81	75	3	-0.04	96	80	80

(a)

As of December 31	**2014**	**2015**
5.00% Senior Notes due September 2020	599	599
4.75% Senior Notes due 2045	0	598
3.50% Senior Notes due June 2024	597	597
4.60% Senior Notes due June 2044	549	549
2.875% Senior Notes due May 2026 (EUR 500M)	605	545
8.205% Junior Subordinated Notes due January 2027	521	521
3.125% Senior Notes due May 2016	500	500
2.80% Senior Notes due 2021	0	399
4.00% Senior Notes due November 2023	349	349
6.25% Senior Notes due September 2040	298	298
4.76% Senior Notes due March 2018 (CAD 375M)	322	271
4.45% Senior Notes due May 2043	248	249
4.25% Senior Notes due December 2042	196	196
3.50% Senior Notes due September 2015	599	0
Commercial paper	168	50
Other	31	16
Total debt	**5582**	**5737**
Less short-term and current portion of long-term debt	783	562
Total long-term debt	**4799**	**5175**

Description	**Due Date**	**2016**	**2015**
5.00%	09/20	599	599
4.75%	12/45	514.9	598
3.50%	06/24	597	597
4.60%	06/44	549	549
2.875%	05/26	605	545
8.205%	01/27	521	521
3.125%	05/16	500	500
2.80%	01/21	514.9	399
4.00%	11/23	349	349
6.25%	09/40	298	298
4.76%	03/18	322	271
4.45%	05/43	248	249
4.25%	12/42	196	196
3.50%	09/15	599	573.3

Metric	chrF	h_score	bert_score	p_score	EM	Bleurt	TabEval	TabXEval-S	TabXEval-G
Value	7	10	17	50	0	-0.34	93	49	32

(b)

Free Throws Attempted	Points	Offensive Rebounds	Assists	3-Pointers Attempted	3-Pointers Made	Minutes Played	Total Rebounds	Steals	Field Goals Made	Free Throws Made	Blocks	Player	Field Goals Attempted
16												Harrison Barnes	
10												Draymond Green	
6		7				24	12	2			2	Andrew Bogut	
12												Klay Thompson	
28		8	8	6	24	5		9				Stephen Curry	13
13												Andre Iguodala	
11												Marreese Speights	
2	15		5			27	3		7	1		Shaun Livingston	10
6	16		2	1	0	30	10		6	4		Tobias Harris	10
4	11	7					13		4	3		Nikola Vucevic	15
	12							3	4			Victor Oladipo	17
	13		5			20						Elfrid Payton	
	13											Ben Gordon	
	6		3			12	7	1	3			Kyle O'Quinn	4

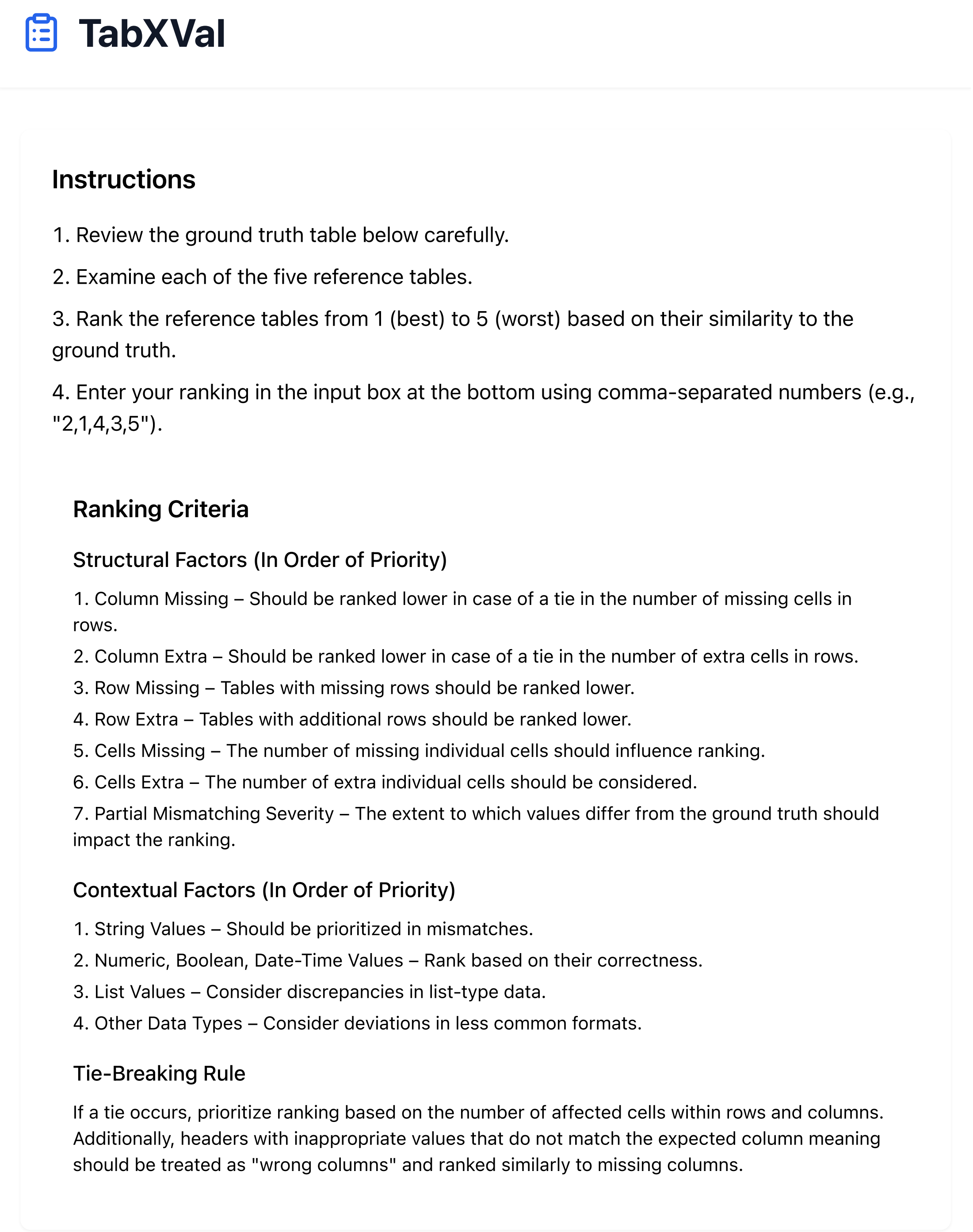
Free Throws Made	Offensive Rebounds	Assists	Free Throws Attempted	Total Rebounds	Steals	3-Pointers Attempted	Blocks	Player	Field Goals Made	3-Pointers Made
								H.B.		
								D.G.		
		7		12	2		2	A.B.		
								K.T.		
		8		5		8		S.C.	9	6
								A.I.		
								M.S.		
1		5	2	3				S.L.	7	
4		2	6	10		1		T.H.	6	0
3	7		4	13				N.V.	4	
					3			V.O.	4	
		5						E.P.		
								B.G.		
		3		7	1			K.O.	3	

Metric	chrF	bert_score	EM	TabEval	h_score	p_score	Bleurt	TabXEval-S	TabXEval-G
Score	2	69	0	98	14	45	0.10	78	78

(c)

Figure 9: Sample perturbed tables from the TabXBench benchmark, illustrating domain-specific corruptions at different difficulty levels. **(a) Movie domain with Easy:** "Easy" perturbations applied to a clean movie-metadata table, including minor spelling errors in film titles, superficial header rephrasing, simple date-format conversions (e.g., "March 3, 2020" ↔ "03/03/2020"), trivial numeric formatting changes (addition/removal of thousands separators), and basic unit shifts (e.g., runtime in minutes vs. hours). **(b) Finance domain with (Easy + Hard):** A financial report table subjected to both "Easy" (currency-symbol normalization, decimal rounding) and "Hard" modifications, such as inconsistent metric abbreviations (e.g., "Rev." vs. "Revenue"), merged indicator columns, omitted quarterly rows, and large-scale unit mismatches (millions vs. billions). **(c) Sports domain with Medium :** A sports-stats table with "Medium" perturbations, featuring moderate header reordering (e.g., swapping "Team" and "Position"), slight numeric shifts in game statistics (win/loss counts adjusted by one or two), merged athlete performance rows, and partial row/column transpositions to emulate realistic table-generation errors.

F Human Evaluation Setup



TabXVal

Instructions

1. Review the ground truth table below carefully.

2. Examine each of the five reference tables.

3. Rank the reference tables from 1 (best) to 5 (worst) based on their similarity to the ground truth.

4. Enter your ranking in the input box at the bottom using comma-separated numbers (e.g., "2,1,4,3,5").

Ranking Criteria

Structural Factors (In Order of Priority)

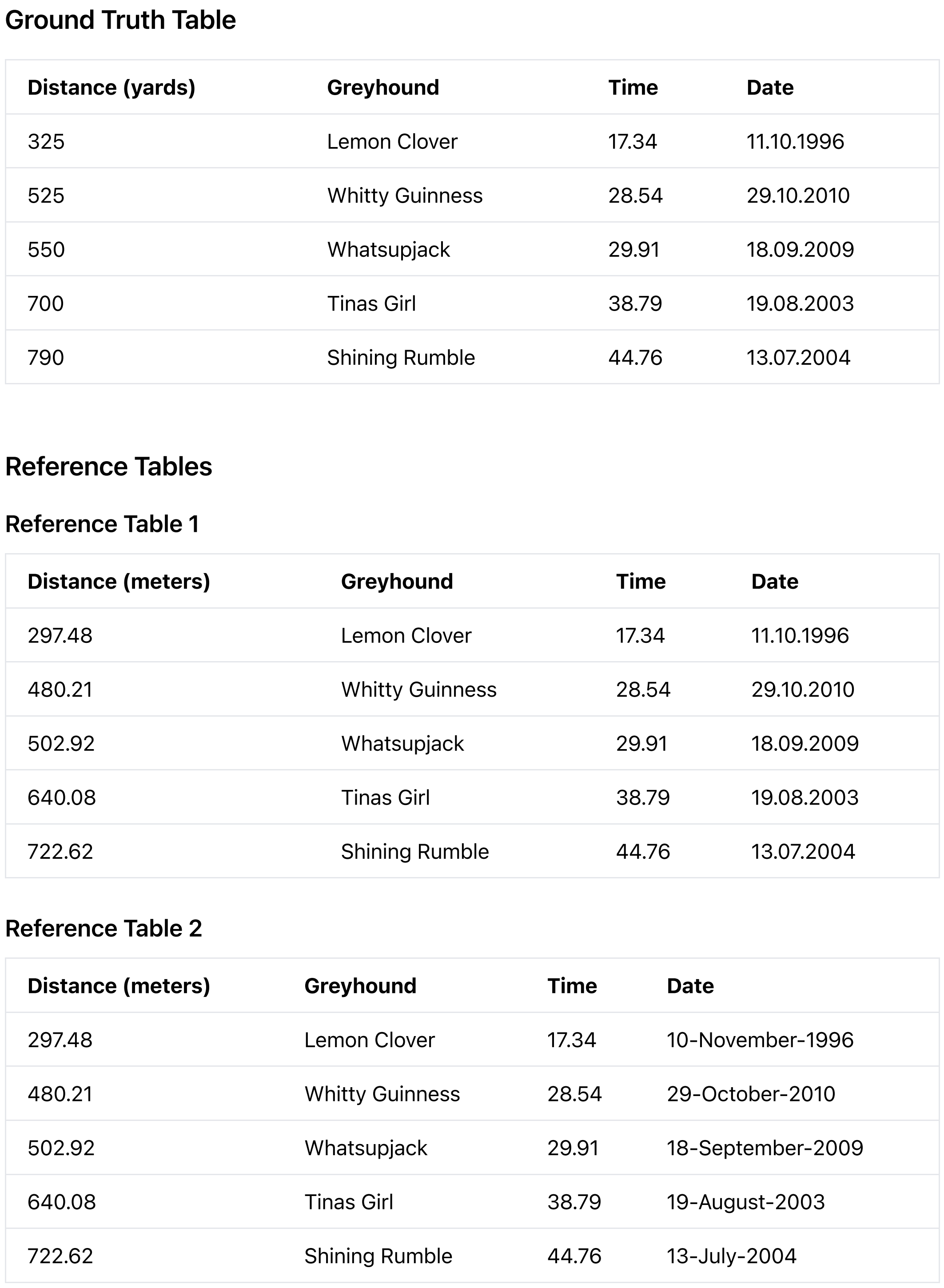
1. Column Missing – Should be ranked lower in case of a tie in the number of missing cells in rows.
2. Column Extra – Should be ranked lower in case of a tie in the number of extra cells in rows.
3. Row Missing – Tables with missing rows should be ranked lower.
4. Row Extra – Tables with additional rows should be ranked lower.
5. Cells Missing – The number of missing individual cells should influence ranking.
6. Cells Extra – The number of extra individual cells should be considered.
7. Partial Mismatching Severity – The extent to which values differ from the ground truth should impact the ranking.

Contextual Factors (In Order of Priority)

1. String Values – Should be prioritized in mismatches.
2. Numeric, Boolean, Date-Time Values – Rank based on their correctness.
3. List Values – Consider discrepancies in list-type data.
4. Other Data Types – Consider deviations in less common formats.

Tie-Breaking Rule

If a tie occurs, prioritize ranking based on the number of affected cells within rows and columns. Additionally, headers with inappropriate values that do not match the expected column meaning should be treated as "wrong columns" and ranked similarly to missing columns.



Ground Truth Table

Distance (yards)	Greyhound	Time	Date
325	Lemon Clover	17.34	11.10.1996
525	Whitty Guinness	28.54	29.10.2010
550	Whatsupjack	29.91	18.09.2009
700	Tinas Girl	38.79	19.08.2003
790	Shining Rumble	44.76	13.07.2004

Reference Tables

Reference Table 1

Distance (meters)	Greyhound	Time	Date
297.48	Lemon Clover	17.34	11.10.1996
480.21	Whitty Guinness	28.54	29.10.2010
502.92	Whatsupjack	29.91	18.09.2009
640.08	Tinas Girl	38.79	19.08.2003
722.62	Shining Rumble	44.76	13.07.2004

Reference Table 2

Distance (meters)	Greyhound	Time	Date
297.48	Lemon Clover	17.34	10-November-1996
480.21	Whitty Guinness	28.54	29-October-2010
502.92	Whatsupjack	29.91	18-September-2009
640.08	Tinas Girl	38.79	19-August-2003
722.62	Shining Rumble	44.76	13-July-2004

Reference Table 3

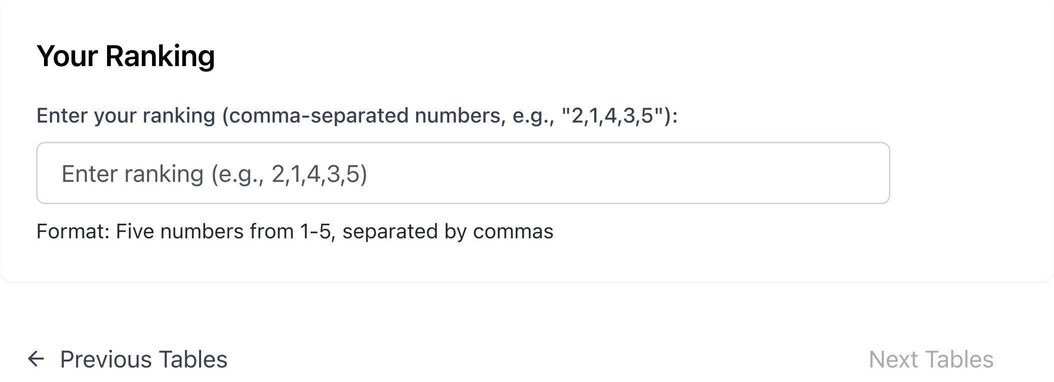
Distance (meters)	Greyhound	Time	Date	Date
297.48	Lemon Clover	17.34	10-November-1996	18.09.2009
480.21	Whitty Guinness	28.54	29-October-2010	13.07.2004
502.92	Whatsupjack	29.91	18-September-2009	19.08.2003
640.08	Tinas Girl	38.79	19-August-2003	10.11.1996
722.62	Shining Rumble	44.76	13-July-2004	29.10.2010

Reference Table 4

Distance (yards)	Greyhound	Duration	Date
525		29.91	29-October-2010
550	Whitty Guinness		11-October-1996
700	Shining Rumble		
525		29.91	29-October-2010
550	Whitty Guinness		11-October-1996
700	Shining Rumble		

Reference Table 5

Distance (yards)	Greyhound	Time	Date
325	Nova Eclipse	28.54	29.10.2010
525	Galaxy Flame	17.34	19.08.2003
550	Sonic Dash	44.76	13.07.2004
700	Midnight Star	29.91	18.09.2009
790	Thunder Strike	38.79	11.10.1996



Your Ranking

Enter your ranking (comma-separated numbers, e.g., "2,1,4,3,5"):

Format: Five numbers from 1-5, separated by commas