SciTables : A Dataset and Evaluation Framework for Complex Table-to-Text Generation

Mehrnoush Alizade, Tengrui Kong and Suman Kalyan Maity
Missouri S&T
Rolla, Missouri, USA
malizade@mst.edu,tkpqz@mst.edu,smaity@mst.edu

ABSTRACT

Generating coherent and factually grounded text from structured data is a core challenge in natural language generation, with applications in scientific communication, medical documentation, and automated reporting. Existing datasets primarily focus on open-domain or simplified table formats, limiting progress in more complex, high-stakes domains. We present **SciTables**, a new dataset and evaluation framework for scientific table-to-text generation, addressing the gap in existing resources that focus largely on open-domain or simplified tables. Our dataset is constructed from Computer Science papers on arXiv (2017–2023) and features complex tables rich in numeric, symbolic, and mathematical content paired with naturally occurring textual descriptions. We develop a scalable, semi-automated pipeline to extract, clean, and align tables with their associated text, preserving domain-specific language while minimizing annotation cost. The resulting benchmark poses realistic challenges for current models and supports evaluation beyond semantic similarity, including factual accuracy, relevance, and multiple forms of reasoning. We conduct extensive experiments with state-of-the-art generation models and show that while current models achieve strong semantic alignment with reference descriptions, they struggle with higher-order reasoning, aggregation, and factual grounding as table complexity increases. Our work provides a realistic and scalable benchmark for advancing faithful, informative, and reasoning-aware table-to-text generation in scientific domains.

PVLDB Reference Format:
Mehrnoush Alizade, Tengrui Kong and Suman Kalyan Maity. SciTables : A Dataset and Evaluation Framework for Complex Table-to-Text Generation. PVLDB, 19(11): 3400 - 3412, 2026.
doi:10.14778/3836663.3836697

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/NLP-Research-Insights/SciTables

1 INTRODUCTION

Generating coherent and factually grounded text from structured inputs is a core challenge in natural language generation (NLG) and a critical capability for many applications, including automated report generation, business intelligence, scientific publishing platforms, clinical documentation systems etc. In recent years, there has been notable progress in data-to-text generation, much of this progress has been driven by the availability of large-scale, high-quality datasets in domains such as sports [3, 30], restaurants [20], weather [16], biographies [2, 15], and open-domain settings [22, 28, 33]. However, in high-stakes domains like healthcare, pharmaceuticals, and scientific R&D, the challenge remains: structured data such as experimental tables, clinical trial results, or lab measurements must be translated into accurate and coherent natural language summaries. Yet, there is a striking lack of high-quality, large-scale datasets to support this task.

This gap limits the deployment of data-to-text systems in real-world setting, where outputs must not only be fluent, but also factually grounded, interpretable, and domain-specific. Creating such datasets at scale is expensive and often infeasible, requiring expert annotation to ensure that descriptions are faithful to underlying data. Manual annotation is costly, time-consuming, and often inconsistent, especially in domains that require specialized knowledge. These limitations can result in outputs that are not faithful to the source data—containing hallucinations [5], imprecise statements, or irrelevant information. These issues are especially problematic in high-stakes applications, where decisions may depend on the generated summaries.

To address these challenges, we introduce a new dataset and framework **SciTables** for scientific table-to-text generation. Our approach leverages a scalable, semi-automated pipeline that extracts and aligns complex tables from scientific research articles with high-quality textual descriptions. These scientific tables often contain rich, multi-dimensional data—posing significant challenges for current generation models and offering a rigorous testbed for benchmarking. By reducing annotation cost while maintaining factual consistency, our framework enables practical dataset construction at scale. In addition to dataset creation, we conduct a comprehensive evaluation of state-of-the-art generation models, using both LLMs-as-judges [18] and embedding-based semantic similarity metrics to assess the quality and faithfulness of generated outputs.

Our contributions are threefold: (1) we introduce a domain-specific dataset for complex table-to-text generation grounded in scientific and technical content; (2) we propose a scalable, pipeline for dataset creation with quality control; and (3) we benchmark models on this dataset, offering insights into the challenges of deploying faithful generation systems in real-world settings.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, Vol. 19, No. 11 ISSN 2150-8097.
doi:10.14778/3836663.3836697

Dataset	Pairs	Domain	Target Source	Annotation
WIKIBIO [15]	400K	Biographies	Wikipedia	Automated
ToTTo [22]	136K	Wikipedia (open-domain)	Wikipedia (Annotator Revised)	Human (Annotators iteratively revised the original text)
LOGICNLG [4]	37K	Wikipedia (open-domain)	Annotator Generator	Human/Automated (Based on TabFact; annotators wrote "refute", "entailment" statements)
SCIGEN (LARGE) [19]	53K	Scientific	arXiv (scientific articles)	Expert/Automated (Test data annotated by the article's authors)
NUMERICNLG [27]	1.3K	Scientific	ACL Anthology (scientific papers)	Expert/Automated (CS experts cleaned and annotated descriptions)
CTRLSCITAB [10]	9.0K	Scientific	arXiv (scientific articles)	Domain experts refined and annotated highlighted cells
ROTOWIRE [30]	11K	Basketball	RotoWire	Automated
WEBNLG [8]	2.53K	15 DBPedia categories	Annotator Generator	Crowd workers wrote natural text expressing triple sets
E2E [20]	51K	Restaurants	Annotator Generator	Crowd workers verbalized MR info, could skip unimportant ones
SCITABLES(OURS)	119K	Scientific	arXiv (scientific articles)	Automated

Table 1: Overview of Table-to-Text Datasets. "Pairs" indicates the number of structured data-description pairs (Table, Corresponding paragraph) in each dataset.

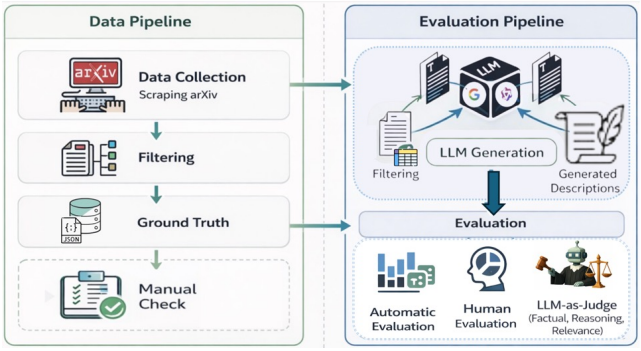

Figure 1: Overview of the SciTables semi-automated pipeline. The data pipeline begins with collecting LaTeX source files from arXiv, followed by table extraction, structural filtering, table-text alignment through explicit references, and manual quality verification. The resulting dataset is used to benchmark table-to-text generation models. The evaluation pipeline generates textual descriptions from tables and assesses them using automatic similarity metrics, human evaluation, and an LLM-as-a-judge framework that measures reasoning based metrics.

2 RELATED WORK

2.1 Domain Specific Datasets

Table-to-text generation has seen considerable progress with the release of large-scale datasets in open and semi-structured domains. Notably, ToTTo [22] and LogicNLG [4] have enabled model development at scale. These datasets, primarily based on Wikipedia, cover diverse topics and provide structured tables aligned with natural language descriptions. However, the tables tend to be relatively simple, with limited numeric or symbolic content, and are less suited to high-stakes domains requiring complex reasoning. Domain-specific datasets have also been introduced in narrower contexts. For instance, early work explored weather forecasts [16], restaurant recommendations [20], sports game summaries [3, 30], entity descriptions [24, 29] and biographies [2, 15]. These datasets typically involve controlled language and limited variability in table structures. Recently, Deng et al. in [6], introduced LiveSum, a new benchmark dataset composed of real-time commentary texts paired with summary tables to test models' ability to perform this task. [12] propose an LLM-driven system for text-to-table generation that aims to address the problem of unstable and error-prone table generation by LLMs.

Despite these advances, scientific and technical domains remain underrepresented. Scientific tables often involve dense numerical, symbolic, and domain-specific information, and their interpretation requires more than surface-level content alignment. This remains a notable gap in the literature, one that our work aims to address.

2.2 Scientific Table-to-Text Generation

Extending beyond existing domains, our work introduces a new dataset for table-to-text generation derived from scientific articles in the Computer Science section of arXiv. We collect tables and their corresponding natural language descriptions from a diverse set of papers published between 2017 and 2023. Table 3 provides an overview of the subdomain distribution within our dataset.

Unlike prior datasets, our tables often contain dense numeric and symbolic data, accompanied by highly technical descriptions. This results in a more realistic and challenging benchmark for testing the capabilities of data-to-text models, especially with respect to logical reasoning, factual consistency, and domain adaptation. As such, our dataset serves as a valuable resource for evaluating the performance of large language models on structured scientific content.

2.3 Annotation Strategies and Automation

Constructing large-scale, high-quality datasets through manual annotation is expensive, time-consuming, and difficult to scale - particularly in specialized domains where expert knowledge is required. To mitigate this, our approach shifts toward automation. In contrast to the ToTTo dataset [22], where annotators were instructed to rewrite contextually relevant sentences, we directly extract natural descriptive texts from scientific articles without manual rewriting. To ensure text quality, we apply a cleaning step that removes LaTeX notations, formatting commands, and other artifacts that can distort the semantic content or interfere with embedding-based processing. This process preserves the domain-specific terminology and natural linguistic structure, resulting in a more authentic and semantically aligned pairing between structured data and textual description.

Year	LaTeX Files	Original	Step 1	Step 2
2017	30,813	18,174	5,504	5,380
2018	41,950	27,848	8,627	8,448
2019	55,525	42,869	13,090	12,830
2020	71,425	32,286	9,866	9,655
2021	77,519	73,565	22,583	22,151
2022	82,109	84,176	26,056	25,537
2023	100,020	112,194	35,591	34,894
Total	**459,361**	**391,112**	**121,317**	**118,895**

Table 2: Statistics of our dataset construction. LaTeX Files is the number of files parsed; Original refers to table-reference pairs extracted. Step 1 removes entries with multiple paragraph references; Step 2 excludes descriptions outside the 100–500-token range.

Rank	CS Subfield	Abrv.	Percentage
1	Machine Learning	LG	27.1%
2	Computer Vision	CV	19.2%
3	Artificial Intelligence	AI	15.4%
4	Natural Language Processing	CL	10.3%
5	Information Theory	IT	6.4%
6	Robotics	RO	5.1%
7	Cryptography & Security	CR	4.9%
8	Systems & Control	SY	4.7%
9	Numerical Analysis	NA	3.6%
10	Data Structures & Algorithms	DS	3.3%

Table 3: Overview of the top 10 computer science subfields.

Prior work relying on manually constructed or crowdsourced annotations often suffers from limited domain expertise and systematic annotation artifacts, which may lead to reduced factual accuracy and alignment. Furthermore, the repeated use of similar phrasing or sentence structures by annotators can introduce stylistic bias and reduce linguistic diversity within the dataset[9, 11, 23]. Our automated pipeline addresses these limitations by leveraging naturally occurring, expert-authored scientific descriptions.

3 DATASET CONSTRUCTION

3.1 Semi-Automated Data Processing Pipeline

Figure 1 provides an overview of the SciTables framework, which consists of a data construction pipeline and an evaluation pipeline. The data construction pipeline transforms raw scientific articles into high-quality table-text pairs through four stages: data collection, table-text alignment, preprocessing and filtering, and quality verification. This semi-automated design combines deterministic extraction and filtering procedures with targeted manual inspection, enabling scalable dataset construction while maintaining data quality.

3.2 Source Collection and Table Extraction

We construct **SciTables** from Computer Science articles published on arXiv between 2017 and 2023. We focus on LaTeX source files because they preserve the structural information required to accurately identify table environments, captions, labels, and cross-references. As shown in Figure 1, the pipeline begins by downloading and parsing LaTeX sources from arXiv. From each article, we extract tables together with their captions and unique identifiers, yielding an initial collection of candidate table-text pairs. Table 3 shows the distribution of articles across the ten most frequent Computer Science subfields represented in our dataset. The resulting collection captures a diverse range of scientific tables containing dense numerical, symbolic, and categorical information accompanied by expert-authored descriptions.

3.3 Table-Text Alignment

A critical step in dataset construction is aligning each table with its corresponding textual description. Unlike prior datasets that rely on manual annotation or crowd-sourced summaries, SciTables leverages the explicit referencing structure present in scientific documents. For every extracted table, we first identify its LaTeX label (e.g., `\label{tab:results}`) and then locate all references to that label within the document body (e.g., `\ref{tab:results}`). Candidate paragraphs containing these references are extracted and linked to the corresponding table.

To ensure a clear and unambiguous correspondence between a table and its description, we retain only paragraphs that reference a single table and do not simultaneously mention other tables, figures, equations, or sections. When multiple qualifying paragraphs reference the same table, the example is discarded to avoid fragmented descriptions distributed across different parts of the paper. This procedure yields table-paragraph pairs that are tightly aligned and largely self-contained, facilitating reliable evaluation of table-to-text generation systems.

3.4 Preprocessing and Data Cleaning

Following alignment, we apply a series of preprocessing and quality-control steps to improve dataset consistency. First, we remove duplicate table-paragraph pairs that arise from repeated references within the same article. We then apply a length-based filter, retaining only paragraphs between 100 and 500 tokens. Shorter descriptions often lack sufficient contextual information, whereas longer passages frequently discuss multiple experimental findings and extend beyond the scope of a single table. Restricting the length range therefore improves both informativeness and alignment quality.

The extracted paragraphs are originally written in LaTeX and contain commands, citations, mathematical markup, and formatting constructs that are unsuitable for downstream text generation evaluation. To normalize these descriptions, we convert them into plain English using the prompt shown in Prompt 1. Specifically, we employ Llama-3.3-70B-Instruct to remove LaTeX-specific syntax while preserving the original semantics and wording. This normalization step produces a standardized textual representation that enables fair comparison between model-generated descriptions and reference texts. Table 2 provides a statistical summary of the dataset construction process, including the number of entries retained at each filtering stage by publication year.

3.5 Quality Verification

To validate the reliability of the semi-automated pipeline, we perform manual quality verification on a randomly sampled subset of the extracted examples. Annotators inspect the original table, the aligned paragraph, and the surrounding document context to verify extraction correctness, alignment accuracy, and semantic completeness. This final verification stage, illustrated in Figure 1, provides an additional quality assurance mechanism and helps ensure that the resulting dataset consists of faithful and informative table-text pairs suitable for benchmarking scientific table-to-text generation systems.

Prompt 1: LaTeX-to-Plain Text Conversion

Make the following paragraph LaTeX command free and convert it to plain English text. Keep everything else the same. Do not do paraphrasing. Add the $[END]$ token at the end of the cleaned paragraph.

Paragraph: {referencing_paragraphs}
Answer:

3.6 Dataset Characteristics and Complexity Analysis

Following construction and quality validation, we analyze the structural and semantic properties of the resulting dataset. These statistics provide quantitative evidence of the complexity and diversity of SciTables and help contextualize the challenges posed by scientific table-to-text generation. Table 4 summarizes key statistics related to table dimensions, content heterogeneity, numerical density, and information compression, providing quantitative evidence of the dataset's complexity and richness. The dataset contains 118,895 table-text pairs spanning a wide range of scientific domains and table layouts. On average, tables contain 6.7 rows and 4.9 columns, with over one-third (36.5%) exhibiting multi-level header structures that require hierarchical interpretation. The cell-type distribution further highlights the heterogeneous nature of scientific tables: while 43.4% of cells contain purely numerical values, a substantial proportion contain formulas (14.5%), mixed text-numeric content (13.4%), or numeric-symbolic expressions (5.0%), reflecting the prevalence of experimental results, statistical measures, and mathematical notation. Additionally, 46.1% of cells contain precision-sensitive numerical values, emphasizing the importance of numerical faithfulness during generation.

The dataset also exhibits a relatively high numeric density score (0.371), indicating that a large fraction of table content consists of quantitative information that must be accurately interpreted and verbalized. Finally, the average compression ratio of 5.609 suggests that the reference descriptions summarize information from tables that are substantially larger than the resulting text, requiring content selection, aggregation, and abstraction. Collectively, these characteristics demonstrate that SciTables presents significantly greater structural complexity and reasoning demands than existing open-domain table-to-text benchmarks.

Property	Value
Rows per Table (mean ± std)	6.7 ± 7.3
Columns per Table (mean ± std)	4.9 ± 3.6
Multi-level Headers	36.5%
Decimal / Precision-heavy cells	46.1%
Text-only cells	21.4%
Numeric-only cells	43.4%
Symbolic-only cells	2.3%
Numeric + Symbolic cells	5.0%
Text + Numeric cells	13.4%
Formula / Mathematical Expression cells	14.5%
Numeric Density Score	0.371
Compression Ratio	5.609

Table 4: Structural and content characteristics of SciTables. Numeric Density Score is computed as the ratio of numeric and symbolic cells to the total number of cells. Compression Ratio is defined as the ratio of table tokens to tokens in the corresponding reference paragraph.

3.7 Qualitative Examples and Generation Challenges

To further illustrate the complexity of SciTables, Figure 2 presents representative examples consisting of scientific tables, their corresponding author-written descriptions, and text generated by a state-of-the-art LLM. Unlike existing table-to-text benchmarks that primarily require localized cell verbalization, SciTables often demands higher-level reasoning, including comparative analysis, trend identification, aggregation across multiple rows and columns, and selective reporting of salient findings. As shown in the examples, LLM-generated descriptions frequently remain faithful to individual table entries but struggle to capture the deeper scientific conclusions expressed in the original text.

In the first example, the reference description synthesizes evidence across multiple benchmarks and highlights the key scientific claim that the proposed approach establishes a new state-of-the-art. In contrast, the generated text primarily enumerates benchmark results and dataset names while failing to articulate the broader conclusion and comparative significance. Similarly, in the second example, the reference text explains *why* performance improvements occur and provides domain-specific interpretation of the observed results. The generated description accurately reports numerical values but lacks the causal explanations, analytical insights, and content prioritization found in the author-written text. These examples demonstrate that SciTables requires substantially more than surface-level table verbalization and instead evaluates a model's ability to perform scientific reasoning, content selection, and evidence synthesis.

4 EXPERIMENTAL SETUP AND EVALUATION

First, we benchmark the performance of state-of-the-art large language models (LLMs) on our newly constructed **SciTables** dataset, which is designed to evaluate a model's ability to perform table-to-text generation in scientific domains. In this task, the input consists of a structured scientific table—often summarizing experimental results, measurements, or comparative analyses and the goal is to generate a coherent, factually accurate, and semantically faithful

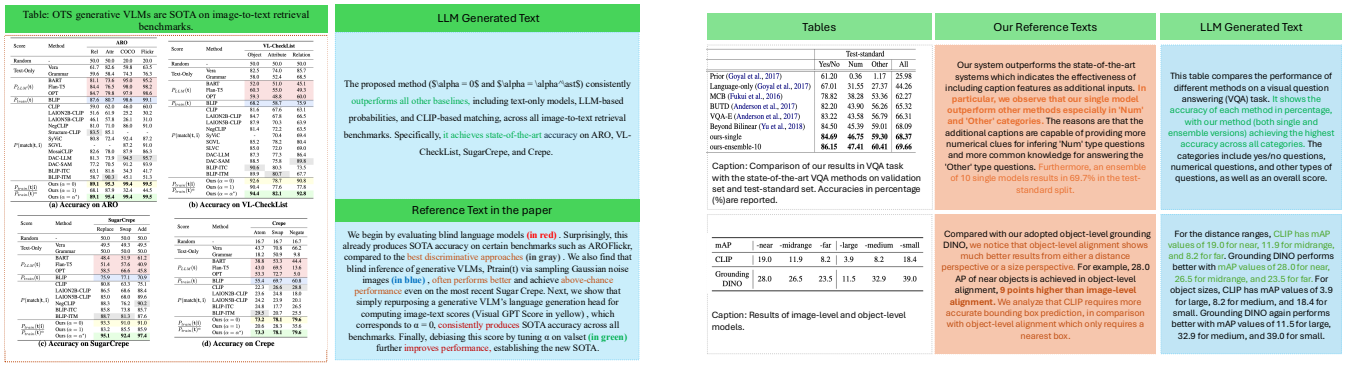

Figure 2: Representative examples from SciTables. Each example contains the extracted scientific table, the corresponding reference description from the original paper, and the output generated by an LLM.

paragraph that captures the key information and insights conveyed by the table. For benchmarking, we evaluate a representative subset of 8,104 examples sampled from the full dataset. This enables a computationally efficient yet representative evaluation, while the remaining examples are reserved for model training, and future large-scale benchmarking.

We evaluate five open-weight leading instruction-tuned LLMs: *Llama-3.3-70B-Instruct*, *Mistral-Small-24B-Instruct*, *Gemma-3-27B-it*, *Aya-expanse-32b* , *Phi-3-medium-4k-instruct*. To ensure a comprehensive assessment of model capabilities on tabular scientific data, we employ a combination of evaluation strategies: (i) automatic metrics that capture semantic similarity between the model-generated paragraph and the ground-truth reference, (ii) LLM-as-a-judge framework for scalable automated evaluation for reasoning capabilities with manual human evaluation.

Prompt 2: Table-Based Text Generation (With Caption)

Generate concise and informative sentences that summarize the key information from the given table and its caption. You may use the caption as additional context. Add the [*END*] token at the end.
Caption: {caption}
Table: {table}
Answer:

4.1 Automatic Metrics

To systematically assess the similarity between LLM-generated paragraphs and the corresponding reference texts, we employ a diverse set of automatic evaluation metrics. We use SBERT [25] and BERTScore [32] to measure semantic similarity, BLEU [21], ROUGE-L [17], and METEOR [1] to quantify lexical overlap, and KL_{LC} divergence [14] to assess differences between the lexical distributions of generated outputs and table captions. Together, these metrics provide complementary perspectives on semantic fidelity, content preservation, and linguistic variation.

Table 5 reports the performance of the LLMs when table captions are provided as part of the input. We analyze performance across three table-size categories (small, medium, and large) to understand how increasing table complexity affects generation quality. Across all models, performance degrades consistently as table size increases. For small tables, models achieve strong semantic similarity scores (SBERT: 0.73–0.80; BERTScore: 0.84–0.85), indicating that they can effectively leverage concise and focused table content. However, performance declines substantially for large tables, where SBERT drops to approximately 0.48–0.49 across all models. Similar trends are observed for BLEU, ROUGE-L, and METEOR, suggesting increasing difficulty in preserving both semantic content and lexical correspondence as input complexity grows. Among the evaluated models, Mistral achieves the highest overall SBERT score (0.753), Aya obtains the highest overall BERTScore F1 (0.846), Phi achieves the highest BLEU (0.026) and METEOR (0.221) scores, and Aya achieves the highest ROUGE-L score (0.185). Despite these differences, the performance gap between models is relatively small compared to the substantial degradation caused by increasing table size.

A similar trend is observed for lexical divergence. As table size increases, KL_{LC} rises sharply across all models, increasing from approximately 4.5–8.5 for small tables to more than 16 for large tables. This indicates that generated outputs become progressively less aligned with the lexical distribution of the original captions as the amount of information contained in the table grows. The simultaneous decrease in similarity metrics and increase in lexical divergence suggests that current LLMs struggle to maintain both semantic fidelity and content selection when generating descriptions for large, information-rich scientific tables.

4.1.1 Impact of providing captions in Table-to-Text Generation. To better understand the contribution of captions, we conduct an ablation study in which captions are masked from the input tables. Tables 5 and 6 compare model performance with and without caption information. Across all the models, providing captions consistently improves semantic similarity and lexical overlap metrics. Captions serve as a strong semantic anchor that guides content selection and helps models identify the primary findings represented in the table. When captions are removed, overall SBERT scores decrease for every model, while BLEU, ROUGE-L, and METEOR also decline. The effect becomes more pronounced for medium and

large tables, where the amount of information that must be summarized increases substantially. The absence of captions also leads to higher lexical divergence. For example, Llama's KL_{LC} increases from 9.07 to 12.89, Gemma's from 9.66 to 12.88, Phi's from 6.60 to 12.53, and Aya's from 8.51 to 12.86. These findings suggest that captions provide both semantic context and lexical grounding, enabling models to generate descriptions that more closely resemble human-authored scientific summaries.

4.1.2 Impact of Prompting Strategies. To investigate whether advanced prompting techniques can improve scientific table-to-text generation, we evaluate two representative models, Aya and Mistral, under three prompting settings: zero-shot prompting, few-shot prompting (3-shot and 5-shot), and Chain-of-Thought (CoT) prompting. All experiments are conducted in the masked-caption setting, which represents a more challenging scenario where models must rely solely on table content without additional contextual guidance. Table 7 summarizes the overall results across all table sizes.

The results indicate that **few-shot prompting provides modest but consistent improvements over standard zero-shot prompting**, particularly for semantic similarity metrics. For Aya, three-shot prompting improves SBERT from 0.666 to 0.711 and increases METEOR from 0.175 to 0.211, while simultaneously reducing lexical divergence (KL_{LC}) from 12.858 to 11.467. Similar trends are observed for Mistral, where three-shot prompting yields the highest SBERT score (0.755) among all Mistral variants while maintaining competitive performance across BLEU, ROUGE-L, and METEOR. These findings suggest that providing exemplar table-to-text pairs helps models better infer the desired output structure and content-selection strategy, leading to improved semantic alignment with the reference descriptions.Interestingly, increasing the number of demonstrations from three to five does not produce additional gains. For both Aya and Mistral, five-shot prompting either matches or slightly underperforms the corresponding three-shot configuration across most metrics. For example, Aya's SBERT score decreases from 0.711 to 0.695, while Mistral's ROUGE-L drops from 0.164 to 0.156. This suggests that the benefits of in-context examples quickly saturate for this task and that additional demonstrations may consume valuable context length without providing substantially new information.

In contrast, **Chain-of-Thought prompting does not consistently improve performance**. Although CoT achieves competitive BERTScore values (0.843 for Aya and 0.840 for Mistral), it yields lower SBERT, BLEU, ROUGE-L, and METEOR scores than both zero-shot and few-shot prompting. Moreover, CoT produces the highest KL divergence among all prompting strategies, increasing lexical drift from the reference descriptions. For example, KL_{LC} rises to 13.668 for Aya and 13.761 for Mistral. These results suggest that explicitly encouraging intermediate reasoning may lead models to generate longer and more explanatory outputs that deviate from the concise, results-oriented style typically found in scientific table descriptions.

Overall, the impact of prompting strategies is relatively modest compared to the differences observed across table complexity levels. While three-shot prompting provides measurable improvements in semantic similarity and lexical alignment, neither additional demonstrations nor Chain-of-Thought reasoning substantially alters overall generation quality. These findings suggest that the primary challenge in scientific table-to-text generation lies not in prompting alone, but in the model's ability to perform content selection, evidence aggregation, and faithful reasoning over complex structured scientific data.

4.1.3 Impact of Task-Specific Fine-Tuning. To investigate whether task-specific adaptation can improve scientific table-to-text generation, we fine-tuned Gemma using table–caption pairs from the SciTables training set. We then evaluated the fine-tuned model on the held-out test set and compared its performance against the base Gemma model using the automatic metrics described earlier.

Table 8 presents the overall results. Fine-tuning leads to notable improvements in several generation quality metrics. In particular, BERTScore F1 increases from 0.841 to 0.856, BLEU nearly doubles from 0.015 to 0.029, and METEOR improves substantially from 0.183 to 0.233. ROUGE-L also shows a modest increase (0.169 to 0.173), indicating better lexical and semantic alignment with the reference descriptions. These improvements suggest that task-specific fine-tuning helps the model better capture domain-specific terminology and generate descriptions that more closely resemble human-authored scientific narratives.

Interestingly, the fine-tuned model achieves slightly lower SBERT similarity (0.704 vs. 0.716) and a higher KL divergence (16.73 vs. 12.88). This suggests that fine-tuning encourages the model to produce richer and more diverse descriptions rather than closely mimicking the wording of the reference text. The simultaneous improvement in overlap-based metrics and lexical divergence indicates that the model is generating more detailed content while preserving semantic relevance.

Overall, the results demonstrate that task-specific fine-tuning provides meaningful gains over the base model, particularly in terms of lexical fidelity and content coverage. However, the improvements are not uniform across all metrics, suggesting that scientific table-to-text generation remains a challenging task even after adaptation and that further advances in structure-aware reasoning and content selection are needed.

4.1.4 Effect of Providing Complete Summaries as Ground Truth. One limitation of using author-written reference paragraphs as ground truth is that they might describe only the subset of findings discussed in the surrounding text rather than providing a comprehensive summary of the entire table. Prior work on table-to-text evaluation has similarly noted that reference texts may diverge from the information contained in the underlying table [7]. While these descriptions are highly reliable and grounded in the underlying data, they may omit relevant observations, secondary comparisons, or additional trends that are visible in the table. This raises an important question: *Do the extracted reference paragraphs adequately represent the information contained in the table, or would a more complete table-centric summary provide a better target for evaluation?*

To investigate this question, we constructed an alternative set of ground-truth descriptions. Starting from the original reference paragraphs, we prompted GPT-5.3 to generate **complete summaries (CS)** that comprehensively describe the information contained in the table while preserving factual consistency with both the table

Model	Table Size	SBERT	BERTScore F1	BLEU	ROUGE-L	METEOR	KL_{LC}
Llama	Small	0.779 ± 0.078	0.853 ± 0.019	0.021 ± 0.021	0.191 ± 0.057	0.204 ± 0.076	7.976 ± 4.384
	Medium	0.700 ± 0.156	0.841 ± 0.022	0.018 ± 0.020	0.167 ± 0.061	0.181 ± 0.090	10.686 ± 5.454
	Large	0.491 ± 0.060	0.817 ± 0.012	0.005 ± 0.007	0.113 ± 0.039	0.085 ± 0.045	16.444 ± 2.627
	Overall	0.746 ± 0.122	0.848 ± 0.021	0.020 ± 0.020	0.181 ± 0.060	0.193 ± 0.083	9.071 ± 5.032
Mistral	Small	0.796 ± 0.076	0.844 ± 0.020	0.024 ± 0.022	0.194 ± 0.059	0.224 ± 0.073	6.673 ± 4.282
	Medium	0.690 ± 0.156	0.832 ± 0.022	0.017 ± 0.020	0.164 ± 0.061	0.184 ± 0.094	10.306 ± 5.882
	Large	0.464 ± 0.067	0.810 ± 0.014	0.005 ± 0.007	0.115 ± 0.041	0.085 ± 0.044	16.417 ± 2.794
	Overall	0.753 ± 0.135	0.840 ± 0.022	0.021 ± 0.022	0.182 ± 0.062	0.207 ± 0.084	8.101 ± 5.292
Gemma	Small	0.787 ± 0.074	0.850 ± 0.019	0.020 ± 0.018	0.190 ± 0.054	0.211 ± 0.069	8.520 ± 4.160
	Medium	0.688 ± 0.161	0.837 ± 0.021	0.015 ± 0.017	0.161 ± 0.059	0.173 ± 0.087	11.415 ± 5.255
	Large	0.492 ± 0.064	0.817 ± 0.013	0.006 ± 0.008	0.113 ± 0.040	0.085 ± 0.044	16.429 ± 2.676
	Overall	0.747 ± 0.126	0.845 ± 0.020	0.018 ± 0.018	0.179 ± 0.058	0.195 ± 0.079	9.662 ± 4.838
Phi	Small	0.735 ± 0.145	0.844 ± 0.018	0.030 ± 0.033	0.194± 0.056	0.245 ± 0.069	4.492 ± 3.222
	Medium	0.641 ± 0.171	0.836 ± 0.019	0.021 ± 0.030	0.160 ± 0.063	0.185 ± 0.103	9.824 ± 6.749
	Large	0.487 ± 0.050	0.819 ± 0.012	0.003 ± 0.005	0.106 ± 0.040	0.072 ± 0.035	17.463 ± 2.000
	Overall	0.697 ± 0.163	0.841 ± 0.019	0.026 ± 0.032	0.181 ± 0.061	0.221 ± 0.089	6.603 ± 5.592
Aya	Small	0.730 ± 0.149	0.851 ± 0.019	0.028 ± 0.032	0.199 ± 0.057	0.218 ± 0.071	6.779 ± 3.901
	Medium	0.640 ± 0.171	0.839 ± 0.020	0.019 ± 0.028	0.162 ± 0.064	0.168 ± 0.092	11.144 ± 5.805
	Large	0.485 ± 0.048	0.820 ± 0.011	0.004 ± 0.004	0.105 ± 0.039	0.072 ± 0.034	17.400 ± 2.003
	Overall	0.694 ± 0.165	0.846 ± 0.020	0.024 ± 0.031	0.185 ± 0.063	0.198 ± 0.084	8.507 ± 5.229

Table 5: Caption Provided: Evaluation of LLM-generated text against reference paragraphs grouped by table size (Small: <500 tokens, Medium: 500–2000, Large: >2000, Overall). KL_{LC} is the KL divergence between the unigram distribution of the caption and LLM generated text.

and the original reference text. We perform this analysis on a representative sample of 1,000 tables. We then evaluate a representative model, Aya, using both the original reference paragraphs (RefP) and the GPT-5.3-generated complete summaries as evaluation targets.

Table 9 summarizes the results. When complete summaries are used as ground truth, performance improves substantially across all automatic metrics. In the captioned setting, SBERT increases from 0.744 to 0.822, BLEU from 0.024 to 0.039, and ROUGE-L from 0.182 to 0.218. Similar improvements are observed in the caption-masked setting, where SBERT rises from 0.715 to 0.785 and ROUGE-L from 0.168 to 0.197. These gains suggest that model-generated outputs align more closely with comprehensive table-level summaries than with the narrower author-written reference paragraphs. The improvements are particularly pronounced for semantic similarity metrics.

These findings do not imply that author-written descriptions are inadequate. Rather, they highlight two complementary evaluation perspectives. The original reference paragraphs capture the authors' selective interpretation of the most salient findings, whereas the complete summaries provide a more exhaustive representation of the table content. Together, they offer a richer framework for assessing scientific table-to-text generation. More broadly, the results suggest that future benchmarks may benefit from incorporating both human-authored descriptions and table-complete summaries, enabling evaluation of a model's ability to generate not only concise scientific narratives but also comprehensive table-grounded explanations.

4.2 Holistic Assessment Beyond Semantic Alignment

While semantic similarity metrics such as SBERT, BERTScore, BLEU, ROUGE-L, and METEOR provide useful indicators of content alignment between generated text and reference paragraphs, they do not capture deeper aspects of generation quality. A model may produce text that is semantically similar to the reference while still being factually inaccurate, logically shallow, numerically incorrect, or focused on less important aspects of the table. Scientific table-to-text generation requires capabilities that extend beyond lexical and semantic overlap, including comparative reasoning, trend identification, numerical faithfulness, content prioritization, and coherent synthesis of information across multiple rows and columns. To evaluate these dimensions, we employ **Qwen3.6-27B** as an LLM-as-a-Judge and assess generated outputs using the ten-dimensional rubric shown in Prompt 3.

Prompt 3: LLM-as-a-Judge Evaluation Prompt

You are an expert evaluator for scientific table-to-text generation. Your task is to assess the quality of the generated paragraph based on a given scientific table.
You must carefully compare the generated paragraph with the table and evaluate it along multiple dimensions. Focus strictly on the information present in the table. Do NOT use external knowledge.
1. Comparative Reasoning - ability to correctly compare entities (e.g., models, methods, conditions) based on one or more

Model	Table Size	SBERT	BERTScore F1	BLEU	ROUGE-L	METEOR	KL_{LC}
Llama	Small	0.711 ± 0.084	0.844 ± 0.017	0.017 ± 0.015	0.175 ± 0.057	0.197 ± 0.061	12.478 ± 3.850
	Medium	0.681 ± 0.126	0.838 ± 0.019	0.014 ± 0.014	0.158 ± 0.061	0.176 ± 0.078	13.371 ± 3.981
	Large	0.506 ± 0.068	0.818 ± 0.012	0.005 ± 0.007	0.108 ± 0.039	0.079 ± 0.044	17.344 ± 2.295
	Overall	0.696 ± 0.105	0.841 ± 0.018	0.016 ± 0.015	0.168 ± 0.052	0.187 ± 0.069	12.888 ± 3.948
Mistral	Small	0.778 ± 0.076	0.843 ± 0.020	0.023 ± 0.050	0.191 ± 0.059	0.221 ± 0.072	7.419 ± 4.698
	Medium	0.695 ± 0.156	0.834 ± 0.021	0.017 ± 0.054	0.163 ± 0.061	0.184 ± 0.094	10.704 ± 5.880
	Large	0.488 ± 0.067	0.815 ± 0.014	0.005 ± 0.007	0.111 ± 0.041	0.081 ± 0.049	17.252 ± 2.598
	Overall	0.744 ± 0.131	0.840 ± 0.021	0.021 ± 0.022	0.180 ± 0.062	0.206 ± 0.083	8.733 ± 5.471
Gemma	Small	0.739 ± 0.084	0.844 ± 0.019	0.017 ± 0.059	0.178 ± 0.051	0.195 ± 0.063	12.389 ± 3.831
	Medium	0.686 ± 0.137	0.835 ± 0.021	0.013 ± 0.061	0.155 ± 0.055	0.168 ± 0.078	13.517 ± 4.107
	Large	0.505 ± 0.067	0.818 ± 0.013	0.005 ± 0.007	0.109 ± 0.042	0.077 ± 0.038	17.401 ± 2.294
	Overall	0.716 ± 0.112	0.841 ± 0.019	0.015 ± 0.014	0.169 ± 0.054	0.183 ± 0.071	12.880 ± 3.994
Phi	Small	0.698 ± 0.129	0.836 ± 0.017	0.019 ± 0.024	0.172 ± 0.050	0.215 ± 0.060	11.797 ± 3.837
	Medium	0.629 ± 0.150	0.831 ± 0.017	0.014 ± 0.022	0.149 ± 0.053	0.172 ± 0.085	13.552 ± 4.299
	Large	0.487 ± 0.050	0.819 ± 0.012	0.003 ± 0.005	0.106 ± 0.040	0.072 ± 0.035	17.463 ± 2.000
	Overall	0.670 ± 0.142	0.834 ± 0.017	0.017 ± 0.023	0.163 ± 0.053	0.197 ± 0.074	12.525 ± 4.124
Aya	Small	0.692 ± 0.138	0.845 ± 0.019	0.019 ± 0.024	0.179 ± 0.055	0.189 ± 0.066	12.207 ± 3.851
	Medium	0.629 ± 0.157	0.836 ± 0.019	0.014 ± 0.022	0.151 ± 0.056	0.156 ± 0.078	13.766 ± 4.171
	Large	0.485 ± 0.048	0.820 ± 0.012	0.004 ± 0.004	0.105 ± 0.039	0.072 ± 0.034	17.400 ± 2.003
	Overall	0.666 ± 0.149	0.841 ± 0.020	0.017 ± 0.024	0.168 ± 0.057	0.175 ± 0.073	12.858 ± 4.060

Table 6: Masked caption: Evaluation of LLM-generated text against reference paragraphs grouped by table size (Small: <500 tokens, Medium: 500–2000, Large: >2000, Overall) with standard deviations. KL_{LC} is the KL divergence between caption and LLM output unigram distributions.

Model	SBERT	BERTScore F1	BLEU	ROUGE-L	METEOR	KL_{LC}
Aya-0shot	0.666 ± 0.149	0.841 ± 0.020	0.017 ± 0.024	0.168 ± 0.057	0.175 ± 0.073	12.858 ± 4.060
Aya-3shot	0.711 ± 0.141	0.834 ± 0.021	0.018 ± 0.022	0.163 ± 0.050	0.211 ± 0.058	11.467 ± 3.607
Aya-5shot	0.695 ± 0.143	0.824 ± 0.034	0.016 ± 0.022	0.148 ± 0.065	0.197 ± 0.069	11.508 ± 3.559
Aya-CoT	0.660 ± 0.144	0.843 ± 0.021	0.014 ± 0.020	0.163 ± 0.055	0.163 ± 0.064	13.668 ± 3.880
Mistral-0shot	0.744 ± 0.131	0.840 ± 0.021	0.021 ± 0.022	0.180 ± 0.062	0.206 ± 0.083	8.733 ± 5.471
Mistral-3shot	0.755 ± 0.081	0.834 ± 0.021	0.018 ± 0.023	0.164 ± 0.051	0.208 ± 0.058	11.906 ± 3.643
Mistral-5shot	0.745 ± 0.089	0.829 ± 0.029	0.018 ± 0.023	0.156 ± 0.059	0.201 ± 0.065	11.890 ± 3.605
Mistral-CoT	0.706 ± 0.110	0.840 ± 0.021	0.016 ± 0.022	0.167 ± 0.055	0.171 ± 0.061	13.761 ± 3.803

Table 7: Masked-caption setting: Overall performance of various prompting strategies. Results are averaged across all table sizes. KL_{LC} denotes the KL divergence between the unigram distributions of the generated text and the reference captions.

Model	SBERT	BERTScore F1	BLEU	ROUGE-L	METEOR	KL_{LC}
Gemma	0.716 ± 0.112	0.841 ± 0.019	0.015 ± 0.014	0.169 ± 0.054	0.183 ± 0.071	12.880 ± 3.994
Gemma-FT-caption	0.704 ± 0.089	0.856 ± 0.025	0.029 ± 0.047	0.173 ± 0.099	0.233 ± 0.132	16.744 ± 5.485

Table 8: Overall performance comparison between the base Gemma model and the caption-fine-tuned Gemma model.

metrics.
3 (Correct): All comparisons are accurate and supported by table values
2 (Partially correct): Comparison attempted but contains minor error (e.g., wrong margin, missing qualifier)
1 (Incorrect): hallucinated comparison

0: No comparative Reasoning

2. Trend Analysis - ability to describe patterns across rows/-columns (e.g., increase, decrease, stability).
3: Trend correctly identified and described with proper context
2: Trend mentioned but incomplete or slightly inaccurate

Setting	SBERT	BERTScore F1	BLEU	ROUGE-L	METEOR	KL_{LC}
Aya-WithCap-CS	0.822 ± 0.164	0.857 ± 0.032	0.039 ± 0.036	0.218 ± 0.090	0.173 ± 0.083	8.407 ± 5.302
AYA-WOCap-CS	0.785 ± 0.150	0.852 ± 0.030	0.026 ± 0.027	0.197 ± 0.079	0.154 ± 0.072	12.600 ± 4.097
AYA-WithCap-RefP	0.744 ± 0.139	0.845 ± 0.022	0.024 ± 0.031	0.182 ± 0.064	0.207 ± 0.089	8.401 ± 5.357
AYA-WOCap-RefP	0.715 ± 0.128	0.840 ± 0.021	0.018 ± 0.025	0.168 ± 0.059	0.185 ± 0.079	12.631 ± 4.122

Table 9: Effect of providing complete summary vs. Reference Paragraph

1: Incorrect trend described
0: No trend described

3. Aggregation / Summarization - ability to synthesize information across multiple cells (e.g., averages, ranges, general summaries).
3: Accurate and complete aggregation or summary
2: Partial aggregation or minor omission
1: Incorrect or misleading aggregation
0: No aggregation or summarization
4. Multi-Row / Multi-Column Reasoning - ability to combine information from multiple parts of the table to produce a coherent statement.
3: Correct multi-source synthesis
2: Partial integration but missing linkage
1: Fails to combine information or produces inconsistent statements
0: No multi-row / multi-column reasoning

5. Numerical Precision and Faithfulness - accuracy in reporting numeric values and relationships.
3: All numerical references are correct
2: Minor numerical deviation (e.g., rounding error)
1: Incorrect, fabricated, or inconsistent values
0: No numerical values and relationships reported

6. Uncertainty / Variance Interpretation - ability to correctly interpret variability (e.g., ± standard deviation).
3: Correct and meaningful interpretation of uncertainty
2: Mentions uncertainty but lacks clarity or precision
1: Misinterprets or ignores uncertainty
0: No uncertainty / variance interpretation reported

7. Content selection (Salience) - ability to identify and describe the most important information from the table.
3: Highly relevant and focused content selection
2: Some important points missing or minor irrelevant details included
1: Misses key insights or focuses on trivial content
0: No important information is provided

8. Logical Consistency and Coherence - internal consistency and logical flow of the generated paragraph.
3: Fully coherent and logically consistent
2: Mostly coherent and logically consistent
1: Minor inconsistencies or awkward flow
0: Contradictory or incoherent

9. Factual Accuracy - extent to which the generated paragraph is faithfully grounded in the table, with correct numerical values, relationships, and claims, and without hallucinations.

Dimension	Gemma	AYA	Phi	Mistral	Llam
Comparative Reasoning	1.256	0.988	1.244	1.482	1.021
Trend Analysis	0.987	0.744	0.984	1.248	0.651
Aggregation	1.785	1.347	1.637	2.082	1.519
Multi-row Reasoning	1.315	1.087	1.420	1.658	0.998
Numerical Precision	2.393	1.929	1.990	2.519	2.286
Uncertainty Interpretation	1.081	0.668	1.013	1.326	0.667
Content Selection (Salience)	1.933	1.529	1.764	2.197	1.669
Logical Consistency	2.481	1.581	1.845	2.553	2.522
Factual Accuracy	2.338	1.836	1.835	2.520	2.428
Relevance	2.413	1.888	2.179	2.555	2.320

Table 10: Average scores across 10 evaluation dimensions using Qwen3.6-27B as the LLM-as-Judge.

3 (Fully accurate): All statements are correct and fully supported by the table. Numerical values, comparisons, and interpretations are precise. No hallucinations or contradictions.
2 (Mostly accurate): Minor inaccuracies (e.g., small rounding issues, slight mis phrasing) but overall faithful to the table. No major factual errors.
1 (Partially accurate): Contains noticeable errors (e.g., incorrect values, wrong comparisons, misinterpretation of metrics), but some parts remain correct.
0 (Inaccurate): Major factual errors, hallucinated content, or statements not supported by the table. Output is unreliable.

10. Relevance - extent to which the generated paragraph focuses on salient and meaningful information from the table, avoiding irrelevant, redundant, or trivial details
3 (Highly relevant): Focuses on the most important insights from the table. Concise, informative, and free of irrelevant or redundant content.
2 (Mostly relevant): Covers key information but includes minor irrelevant details or misses a small important aspect.
1 (Partially relevant): Mix of relevant and irrelevant content; may overlook major insights or include unnecessary details.
0 (Irrelevant): Fails to capture key information or is dominated by irrelevant, trivial, or off-topic content.

Table 10 reports the average scores across the ten evaluation dimensions. Overall, **Mistral consistently achieves the strongest performance**, obtaining the highest score in every evaluation category. It achieves the best results in comparative reasoning (1.482), trend analysis (1.248), aggregation (2.082), multi-row reasoning (1.658), numerical precision (2.519), uncertainty interpretation (1.326), content selection (2.197), logical consistency (2.553), factual accuracy (2.520), and relevance (2.555). These results indicate that Mistral is not only effective at producing semantically similar descriptions, but also better at identifying salient findings,

synthesizing evidence across table regions, and generating factually grounded scientific narratives.

A clear distinction emerges between dimensions related to **reasoning** and those related to **factual grounding**. Across all models, numerical precision, factual accuracy, logical consistency, and relevance receive substantially higher scores than comparative reasoning, trend analysis, uncertainty interpretation, and multi-row reasoning. For example, Mistral achieves a score of 2.519 for numerical precision and 2.520 for factual accuracy, but only 1.248 for trend analysis and 1.326 for uncertainty interpretation. Similar patterns are observed for Gemma and Llama. This suggests that current LLMs are generally capable of accurately reporting values and producing coherent descriptions, yet they struggle to extract higher-level analytical insights from scientific tables.

The results also reveal that **reasoning over distributed evidence remains a major challenge**. Multi-row reasoning scores remain relatively low across all models, ranging from 0.998 (Llama) to 1.658 (Mistral), indicating difficulties in combining information across multiple rows and columns to form integrated conclusions. Likewise, trend analysis scores remain below 1.3 for all models, suggesting that identifying performance trajectories, trade-offs, or broader patterns is considerably more difficult than simply reporting individual results. Uncertainty interpretation is particularly challenging, with all models scoring below 1.35, highlighting a limited ability to reason about variability, confidence intervals, or standard deviations commonly found in scientific tables.

Among the remaining models, Llama and Gemma demonstrate competitive performance, particularly on factual accuracy, numerical precision, and logical consistency. Llama achieves the second-highest scores in factual accuracy (2.428) and logical consistency (2.522), while Gemma performs similarly well on numerical precision (2.393) and relevance (2.413). In contrast, Aya and Phi exhibit noticeably weaker performance across most reasoning dimensions. Although both models maintain reasonable factual grounding, their lower scores on comparative reasoning, trend analysis, aggregation, and content selection suggest greater difficulty in identifying and communicating the key scientific insights contained within complex tables.

Taken together, these findings complement the automatic evaluation results. While semantic similarity metrics indicate that many models generate text broadly aligned with the reference descriptions, this evaluation reveals substantial differences in reasoning quality, factual grounding, and scientific interpretation. The relatively low scores on comparative reasoning, trend analysis, multi-row reasoning, and uncertainty interpretation further demonstrate that SciTables poses challenges that extend beyond surface-level text generation, requiring models to perform deeper analytical reasoning over structured scientific data.

4.2.1 Human–LLM Judge Alignment. Although LLM-as-a-Judge evaluation offers a scalable alternative to human assessment, its reliability depends on the degree to which its judgments align with expert human evaluations. To quantify this alignment, we randomly sampled 200 generated table-to-text descriptions and obtained human ratings using the ten-dimensional evaluation rubric [13] introduced in section4.2. We then compared these scores against those produced by Qwen3.6-27B [31] operating as an automatic judge.

Dimension	Base-ρ	FT-ρ	Base-τ	FT-τ
Factual-Accuracy	0.9120	0.9628	0.8537	0.9384
Comparative-Reasoning	0.7765	0.9230	0.7339	0.8844
Multi-Row-Column-Reasoning	0.8957	0.9446	0.8485	0.9167
Aggregation-Summarization	0.8712	0.9507	0.7989	0.9278
Numerical-Precision-Faithfulness	0.9435	0.9562	0.9376	0.9595
Trend-Analysis	0.8786	0.9904	0.8409	0.9837
Uncertainty-Variance-Interpretation	0.8788	0.9712	0.8465	0.9530
Content-Selection-Salience	0.9016	0.9425	0.8393	0.9046
Relevance	0.9449	0.9906	0.9093	0.9843
Logical-Consistency-Coherence	0.8820	0.9809	0.7981	0.9682

Table 11: Agreement between human evaluators and Qwen3.6-27B judges before (Base) and after fine-tuning (FT). ρ denotes Spearman's ρ and τ for Kendall's τ.

While the base Qwen3.6-27B judge exhibited strong agreement with human evaluators across most dimensions, we hypothesized that its performance could be further improved through task-specific adaptation. A key challenge is that naturally occurring examples often exhibit limited score diversity, with many outputs clustered around similar quality levels. As a result, the judge may not be sufficiently exposed to the full spectrum of errors and reasoning behaviors required for robust evaluation. To address this limitation, we constructed a synthetic training dataset using GPT-5.3. The synthetic examples were designed to cover diverse quality levels across all ten evaluation dimensions, including both high-quality and intentionally flawed outputs exhibiting factual errors, weak reasoning, poor content selection, numerical inaccuracies, and coherence issues. All synthetic annotations were subsequently reviewed and verified by human annotators to ensure scoring consistency and quality. We then fine-tuned Qwen3.6-27B on this Human-GPT verified dataset to better calibrate its judgments.

Table 11 reports the agreement between human judgments and both the base and fine-tuned judges using Spearman's ρ and Kendall's τ. The results show consistent improvements across all ten evaluation dimensions. The largest gains are observed for reasoning-oriented dimensions, including comparative reasoning (Spearman: 0.7765 $\rightarrow$ 0.9230), trend analysis (0.8786 $\rightarrow$ 0.9904), uncertainty interpretation (0.8788 $\rightarrow$ 0.9712), and logical consistency (0.8820 $\rightarrow$ 0.9809). Similar improvements are observed in Kendall's τ, indicating stronger ranking consistency with human evaluators after fine-tuning. Further, dimensions related to factual grounding already exhibit high agreement in the base model, yet still benefit from adaptation. For example, factual accuracy improves from 0.9120 to 0.9628 (Spearman's ρ), while numerical precision increases from 0.9435 to 0.9562. The strongest overall agreements are achieved for relevance (0.9906), trend analysis (0.9904), and logical consistency (0.9809), suggesting that the fine-tuned judge is highly effective at identifying salient content and evaluating the quality of scientific narratives.

Overall, these results demonstrate that targeted fine-tuning substantially improves human–LLM judge alignment. The consistent gains across all ten dimensions indicate that exposing the judge to a diverse set of human-verified scoring scenarios leads to more reliable and human-like evaluations.

4.3 Impact of Source Table Characteristics on Generation Quality

Scientific tables vary considerably in their structural organization, content composition, and information density. To quantify these differences, we introduce a **Table Complexity Score (TCS)** that combines three complementary aspects of complexity: cell-type diversity, numerical density, and information compression. Let H_{cell} denote the cell-type entropy, NDS the Numeric Density Score, and CR the Compression Ratio. Let a_i denote the proportion of cells belonging to cell type entropy, where the six cell types correspond to text-only, numeric-only, symbolic-only, numeric-symbolic, text-numeric, and formula/ mathematical-expression cells. We first compute the cell-type entropy following standard information-theoretic entropy [26]:

$$H_{cell} = -\sum_{i=1}^{6} a_i \log a_i \tag{1}$$

which captures the heterogeneity of content within a table. We then compute the Numeric Density Score (NDS) and Compression Ratio (CR) as:

$$NDS = \frac{\#\text{numeric cells} + \#\text{symbolic cells}}{\#\text{total cells}} \tag{2}$$

$$CR = \frac{\#\text{tokens in table}}{\#\text{tokens in reference paragraph}} \tag{3}$$

Finally, after min-max normalization of the three components, we define the Table Complexity Score (TCS) as:

$$TCS = \frac{\hat{H}_{\text{cell}} + \hat{NDS} + \hat{CR}}{3} \tag{4}$$

where $\hat{H}_{\text{cell}}$, $\hat{NDS}$, and $\hat{CR}$ denote the normalized entropy, numeric density, and compression ratio, respectively. Higher TCS values correspond to tables that are more heterogeneous, numerically dense, and information-rich, thereby requiring greater reasoning and compression during table-to-text generation.

Table 12 reports the correlation between TCS and the Qwen3.6-27B LLM-as-a-Judge scores for each dimensions. Across all the models, increasing table complexity is consistently associated with lower generation quality. The overall ten-dimensional score exhibits significant negative correlations ranging from $r = -0.101$ (Phi) to $r = -0.175$ (Aya), indicating that more complex tables systematically reduce generation performance irrespective of model architecture. The strongest effects are observed for **aggregation/ summarization** ($r = -0.163$ to -0.256) and **content selection** ($r = -0.123$ to -0.255). This suggests that complex tables primarily challenge a model's ability to identify salient information and synthesize evidence across multiple cells into concise scientific narratives. Multi-row and multi-column reasoning also degrades consistently with increasing complexity ($r = -0.138$ to -0.196), indicating difficulties in integrating information distributed across different parts of the table.

Complexity additionally affects **faithfulness**. Both factual accuracy and numerical precision exhibit significant negative correlations across all models, suggesting that numerically dense and heterogeneous tables increase the likelihood of factual and numerical errors. In contrast, comparative reasoning, trend analysis, and uncertainty interpretation show relatively weak correlations with complexity, likely because these dimensions remain challenging even for simpler tables.

Overall, the results demonstrate that table complexity is an important predictor of generation difficulty. As tables become more heterogeneous, information-dense, and compression-intensive, models increasingly struggle with content selection, aggregation, reasoning, and factual grounding, highlighting the need for future systems that are robust to complex scientific table structures.

5 DISCUSSIONS

5.1 Potential Downstream Applications

While SciTables is introduced as a benchmark for scientific table-to-text generation, its rich table–text alignments support a broader range of downstream NLP tasks. First, the dataset can be used to evaluate **table understanding and reasoning**, including comparative reasoning, trend analysis, aggregation, multi-row reasoning, and numerical interpretation. These capabilities are critical for scientific AI systems that must analyze experimental results and communicate key findings. Second, SciTables provides a valuable resource for **scientific summarization** and **evidence-grounded generation**, where models must generate concise and faithful descriptions from structured data. The paired table and textual descriptions also enable **fact verification** tasks, allowing researchers to assess whether claims made in scientific text are supported by the underlying tabular evidence. Finally, the benchmark can support the development of **retrieval-augmented scientific assistants**, automated research analysis tools, and **instruction-tuned models** specialized for scientific communication. By combining complex tables, human-authored descriptions, and multi-dimensional evaluation frameworks, SciTables serves as a useful testbed for advancing trustworthy AI systems capable of reasoning over structured scientific data.

5.2 Limitations

Despite its scale and realism, SciTables has some limitations. First, the dataset is derived exclusively from Computer Science papers on arXiv, which may limit its generalizability to other scientific domains with different reporting styles and table structures. Second, although our semi-automated pipeline employs multiple cleaning and filtering stages, table–text alignment is inferred from document references and may occasionally miss relevant contextual information. Third, the reference descriptions reflect the authors' selective interpretation of table contents rather than exhaustive summaries, meaning that some valid table-grounded observations may not appear in the ground truth. Finally, while our LLM-as-a-Judge framework demonstrates strong agreement with human evaluators, automated judges may still exhibit biases and cannot fully replace expert human assessment. Future work will expand the benchmark to additional scientific disciplines, incorporate richer document context, and explore more comprehensive evaluation frameworks.

6 CONCLUSIONS

We introduce **SciTables**, a large benchmark dataset for scientific table-to-text generation, featuring complex numeric and symbolic

Model	Dimension	Pearson r	Spearman ρ	R^2	Model	Dimension	Pearson r	Spearman ρ	R^2
Aya	Overall 10Dim Score Mean	-0.175**	-0.245**	0.031	Mistral	Overall 10Dim Score Mean	-0.123**	-0.173**	0.015
	Reasoning Score Mean	-0.102**	-0.205**	0.010		Reasoning Score Mean	-0.112**	-0.170**	0.012
	Faithfulness Score Mean	-0.219**	-0.267**	0.048		Faithfulness Score Mean	-0.123**	-0.168**	0.015
	Comparative Reasoning	0.011	-0.054*	0.000		Comparative Reasoning	-0.015	-0.060*	0.000
	Trend Analysis	0.054*	0.013	0.003		Trend Analysis	-0.001	-0.046*	0.000
	Aggregation Summarization	-0.255**	-0.295**	0.065		Aggregation Summarization	-0.215**	-0.254**	0.046
	Multi Row Multi Column Reasoning	-0.171**	-0.214**	0.029		Multi Row Multi Column Reasoning	-0.196**	-0.216**	0.038
	Numerical Precision Faithfulness	-0.180**	-0.200**	0.032		Numerical Precision Faithfulness	-0.108**	-0.149**	0.012
	Uncertainty Variance Interpretation	0.006	-0.059*	0.000		Uncertainty Variance Interpretation	0.008	-0.032*	0.000
	Content Selection Salience	-0.255**	-0.305**	0.065		Content Selection Salience	-0.199**	-0.255**	0.040
	Logical Consistency Coherence	-0.204**	-0.280**	0.042		Logical Consistency Coherence	-0.120**	-0.169**	0.015
	Factual Accuracy	-0.184**	-0.203**	0.034		Factual Accuracy	-0.114**	-0.154**	0.013
	Relevance	-0.233**	-0.289**	0.054		Relevance	-0.130**	-0.193**	0.017
Gemma	Overall 10Dim Score Mean	-0.133**	-0.181**	0.018	Phi	Overall 10Dim Score Mean	-0.101**	-0.190**	0.010
	Reasoning Score Mean	-0.100**	-0.174**	0.010		Reasoning Score Mean	-0.056*	-0.144**	0.003
	Faithfulness Score Mean	-0.147**	-0.181**	0.022		Faithfulness Score Mean	-0.147**	-0.224**	0.021
	Comparative Reasoning	0.006	-0.047*	0.000		Comparative Reasoning	0.033*	-0.041*	0.001
	Trend Analysis	0.016	-0.037*	0.000		Trend Analysis	0.055*	-0.011	0.003
	Aggregation Summarization	-0.243**	-0.263**	0.059		Aggregation Summarization	-0.163**	-0.214**	0.026
	Multi Row Multi Column Reasoning	-0.157**	-0.183**	0.025		Multi Row Multi Column Reasoning	-0.138**	-0.197**	0.019
	Numerical Precision Faithfulness	-0.143**	-0.169**	0.020		Numerical Precision Faithfulness	-0.134**	-0.204**	0.018
	Uncertainty Variance Interpretation	0.023	-0.032*	0.001		Uncertainty Variance Interpretation	0.033*	-0.045*	0.001
	Content Selection Salience	-0.232**	-0.272**	0.054		Content Selection Salience	-0.155**	-0.213**	0.024
	Logical Consistency Coherence	-0.130**	-0.175**	0.017		Logical Consistency Coherence	-0.131**	-0.199**	0.017
	Factual Accuracy	-0.131**	-0.163**	0.017		Factual Accuracy	-0.147**	-0.212**	0.022
	Relevance	-0.161**	-0.214**	0.026		Relevance	-0.123**	-0.174**	0.015
Llama	Overall 10Dim Score Mean	-0.132**	-0.206**	0.017					
	Reasoning Score Mean	-0.111**	-0.189**	0.012					
	Faithfulness Score Mean	-0.108**	-0.206**	0.012					
	Comparative Reasoning	0.001	-0.015	0.000					
	Trend Analysis	0.010	-0.009	0.000					
	Aggregation Summarization	-0.256**	-0.248**	0.065					
	Multi Row Multi Column Reasoning	-0.151**	-0.184**	0.023					
	Numerical Precision Faithfulness	-0.104**	-0.139**	0.011					
	Uncertainty Variance Interpretation	-0.003	-0.016	0.000					
	Content Selection Salience	-0.250**	-0.266**	0.062					
	Logical Consistency Coherence	-0.088*	-0.146**	0.008					
	Factual Accuracy	-0.083*	-0.143**	0.007					
	Relevance	-0.130**	-0.213**	0.017					

Table 12: Correlation between structural complexity score and Qwen-judged scores across models and evaluation dimensions. Significance markers: $^{**}p < 0.01$, $^{*}p < 0.5$.

tables paired with naturally occurring textual descriptions from arXiv CS papers. We evaluated several state-of-the-art LLMs across semantic similarity, factual accuracy, reasoning quality, relevance, and content selection dimensions. While current models achieve relatively strong semantic alignment with reference descriptions, they consistently struggle with higher-order reasoning tasks such as aggregation, comparative reasoning, trend analysis, uncertainty interpretation, and multi-row reasoning. Our analyses further demonstrate that generation quality degrades as table complexity increases. We also show that captions provide important semantic grounding, while complete table summaries offer a complementary evaluation perspective to author-written descriptions. Collectively, these findings highlight the substantial challenges posed by scientific table-to-text generation and establish **SciTables** as a realistic and scalable benchmark for advancing faithful, coherent, and reasoning-aware generation systems for scientific data.

A APPENDIX

A.1 Annotator Demographics and Evaluation Procedure:

Three authors were chosen as annotators for human-based evaluations. The annotators were provided with the same template with rubrics that was used for the LLM-based evaluation.

A.2 AI assistance in Writing

AI Assistants were used solely to assist in improving the writing clarity and language of this paper. Specifically, AI-assisted refinements were applied to enhance readability, coherence, and grammatical accuracy. No AI-generated content was used to replace critical thinking or fabricate results. Ideas, methodology, experimental design, analysis, and conclusions were entirely conceived, developed, and executed by the authors.

REFERENCES

[1] 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*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909/

[2] Eva Banik, Claire Gardent, and Eric Kow. 2013. The kbgen challenge. In *the 14th European Workshop on Natural Language Generation (ENLG)*. 94–97.

[3] David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*. 128–135.

[4] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. *arXiv preprint arXiv:2004.10404* (2020).

[5] Wenqing Chen, Jidong Tian, Yitian Li, Hao He, and Yaohui Jin. 2021. Deconfounded variational encoder-decoder for logical table-to-text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5532–5542.

[6] Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. *arXiv preprint arXiv:2404.14215* (2024).

[7] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081* (2019).

[8] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *10th International Conference on Natural Language Generation*. ACL Anthology, 124–133.

[9] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1161–1166.

[10] Zhixin Guo, Jianping Zhou, Jiexing Qi, Mingxuan Yan, Ziwei He, Guanjie Zheng, Zhouhan Lin, Xinbing Wang, and Chenghu Zhou. 2024. Towards controlled table-to-text generation with scientific reasoning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9951–9955.

[11] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 107–112.

[12] Sarthak Harne, Arvind Agarwal, et al. 2025. Llm driven text-to-table generation through sub-tasks guidance and iterative refinement. *arXiv preprint arXiv:2508.08653* (2025).

[13] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=8euJaTveKw

[14] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79–86. https://doi.org/10.1214/aoms/1177729694

[15] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771* (2016).

[16] Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 91–99.

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[18] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634* (2023).

[19] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. SciGen: A Dataset for Reasoning-Aware Text Generation from Scientific Tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[20] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254* (2017).

[21] 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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[22] Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373* (2020).

[23] Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 1779–1789.

[24] Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. 2018. Generation of company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In *11th International Conference on Natural Language Generation*.

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[26] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[27] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards Table-to-Text Generation with Numerical Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1451–1465.

[28] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 645–654.

[29] Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. 2018. Describing a knowledge base. *arXiv preprint arXiv:1809.01797* (2018).

[30] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052* (2017).

[31] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mehran Xue, Mei Li, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025). https://arxiv.org/abs/2505.09388

[32] 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*. https://openreview.net/forum?id=SkeHuCVFDr

[33] Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024. Tablellama: Towards open large generalist models for tables. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6024–6044.