SciTableQA: A Question-Answering Benchmark for Complex Scientific Tables

Kehinde Ajayi[1], Yi He[2], Matthew Maisonave[1], Kris SeekFord[1], and Jian Wu[1]

[1] Old Dominion University, Norfolk, VA, USA
{kajay001,j1wu}@odu.edu
[2] William & Mary, Williamsburg, VA, USA yihe@wm.edu

Abstract. Reasoning over complex scientific tables is an essential yet largely underexplored topic. Existing benchmarks often fail to capture the structural and content challenges of complex scientific tables, such as irregular layouts, multi-row headers, and symbolic notations. To address these challenges, we introduce SciTableQA, a benchmark comprising 320 scientific tables in 5 scientific domains, and 8,730 human-verified QA pairs categorized into two types of reasoning tasks. Our tables cover 8 structural features and 4 content features. The questions include 4 interrogative and 3 imperative question styles. To answer the questions, the model needs to reason across up to 14 rows and 13 columns of a table. We evaluate the performance of three widely used Large Language Models (LLMs)—GPT-3.5-turbo, Llama 3-8B, and Mistral-7B against the benchmark dataset. Our evaluation focuses on reasoning accuracy, cross-LLM generalization, and explanation validity. Our findings reveal that arithmetic tasks are more challenging than cell selection tasks. LLMs' performances on both types of tasks vary across domains. Compared with existing TableQA benchmarks, we believe SciTableQA provides a more challenging dataset for evaluating machine learning models' reasoning capability on complex scientific tables. Our dataset is available at https://huggingface.co/datasets/Kehindeajayi01/SciTableQA.

Problem

1 Introduction

Scientific tables are a primary vehicle for summarizing and communicating dense, structured data in a compact format in scientific publications. These tables often present experimental results, statistical comparisons, and observational findings that are critical to support scientific discoveries. Automatic interpretation and reasoning over such data are key to accelerating advanced research applications, such as hypothesis generation [11], claim verification [15], and building knowledge bases [9].

Advancements in large language models (LLMs) have brought remarkable improvements in natural language understanding and reasoning, including table-based question answering (TableQA). However, existing benchmark datasets for TableQA, such as WikiTableQuestions [14] and Spider [18], are designed for general-purpose tables, often derived from Wikipedia or financial datasets.

Problem - Existing TableQA benchmarks and datasets do not adequately capture the structural complexity and reasoning challenges of standalone scientific tables, making it difficult to properly evaluate LLM reasoning over complex scientific tables.

General-purpose tables usually contain canonical structures and monolithic content types (e.g., numbers, text). A portion of tables in scientific papers also resemble general-purpose tables, but a significant fraction are complex tables, containing much more complicated structure and content types, such as multi-row headers, irregular layouts, and symbolic notations [12,10]. Additionally, many TableQA benchmark datasets focus on tasks like SQL generation [18] or simple cell look-ups [21], with limited focus on complex numerical reasoning and advanced inference.

Recent TableQA datasets, such as SCITAB [10] and SCITAT [19], have attempted to close this gap by incorporating datasets tailored to scientific domains. SCITAB, for example, focuses on compositional reasoning and fact verification for scientific tables, while SCITAT emphasizes reasoning that integrates both tables and text. However, SCITAB exhibits a relatively narrow range of reasoning types, focusing on arithmetic reasoning and cell selection tasks. SCITAT, although containing more diverse reasoning types, requires combining text and tables and, therefore, is not appropriate for testing the models' reasoning capabilities for standalone tables.

In this work, we introduce SciTableQA, a benchmark dataset designed to evaluate the reasoning capabilities of LLMs on standalone scientific tables. Our approach focuses on tables and the diverse reasoning challenges they pose, without requiring additional context from textual narratives. To this end, we curate 320 scientific tables obtained from PDFs in various domains, including Materials Science (MatSci), Biology, Computer Science (CompSci), scientific reports (Reports), and ICDAR-2013, and employ LLMs to automatically generate 8,730 high-quality question-answer (QA) pairs. These pairs are manually verified and categorized into two distinct reasoning types, namely cell selection and arithmetic reasoning, with explanations for each answer for transparency and interpretability. This approach of utilizing LLMs to generate the QA pairs mitigates the need for extensive human labor and domain expertise, thereby overcoming the limitation typically associated with the high cost of developing high-quality Scientific TableQA benchmarks.

This work explores four research questions aligned with the broader challenges of reasoning over complex scientific tables.

RQs

1. How well do LLMs generalize across different domains when reasoning over complex scientific tables?
2. How are the performances of LLMs affected by structural and content complexities?
3. How consistent and coherent are LLMs in explaining the answers to table-related questions?
4. How effectively can LLMs answer questions generated by other LLMs, and what types of questions do they work well on or fail at?

By addressing these questions, SciTableQA offers a benchmark that allows comprehensive and robust evaluations of TableQA models, drawing insights into specific table features that challenge the models. Our contributions are as follows:

Contribution - Introduces SciTableQA, a benchmark dataset and evaluation framework for assessing LLM reasoning on standalone complex scientific tables and identifies key challenges in structural and content-based table reasoning.

contributions

– We introduce SciTableQA, a benchmark dataset specifically designed to evaluate the capabilities of LLMs on answering questions related to standalone scientific tables.
– We propose a novel framework that evaluates the capabilities of LLMs to answer questions generated by other LLMs.
– By analyzing the LLMs' performance across structural and content-based features, we identified key challenges in complex scientific table reasoning.

2 Related Work

Research on QA over tabular data has witnessed significant advancements, with benchmarks evolving from general tabular datasets to more domain-specific and reasoning-focused benchmarks. Early efforts, such as WikiTableQuestions [14,13], introduced simple retrieval, arithmetic reasoning, and free-form answer generation of questions related to Wikipedia tables. Spider [18] studies the task of complex SQL generation for multi-table databases. Although these datasets helped advance parsing and table comprehension tasks, they largely focused on general-purpose tables, e.g., tables with clear boundaries and values in most cells, limiting their applicability to reasoning in complex scientific tables.

Incorporating numerical reasoning into TableQA tasks marked a milestone. HybridQA [2,20] combined tabular and textual data, requiring multi-hop reasoning across both modalities. Financial datasets such as TAT-QA [21] and FinQA [3] further emphasized numerical reasoning, involving operations such as aggregation, comparison, and trend analysis. These datasets are instrumental in evaluating LLMs' ability to handle numerical operations but fall short in addressing the structural complexities and domain-specific features in scientific tables.

For scientific domains, benchmarks such as SciGen [12], which focused on reasoning-aware text generation from scientific tables, and SCITAB [10], which introduced claim verification tasks, began addressing domain-specific requirements. SCITAB incorporates compositional reasoning tasks, such as verifying claims based on tabular data. However, its emphasis on claim verification limits its scope for evaluating diverse question types directly on tables.

Table 1. Comparison of SciTableQA with existing TableQA datasets. SciTableQA uniquely combines scientific domain diversity, structural complexity, and verified QA quality in a table-only benchmark.

Dataset	QA Pairs	#Tables	Domains	Reasoning Types	Text Context	Annotation
SciTableQA	**8,700**	**320**	**MatSci, Biology, CompSci, ICDAR, Reports**	**Cell selection + Arithmetic**	**Table only**	**LLM-generated & human verification**
SCITAT	953	871	CS	Cell selection + Arithmetic	**Table + text**	LLM-generated & human verification
SCITAB	1,225	872	CS	Fact verification	Table only	Expert-verified
SciGen	220	220	CS	Data-to-text	Generates text	Expert & auto-scaling
FeTaQA	10,000	∼10K	Wikipedia	Free-form answers	Table only	Human-generated answers

Table 2. Distribution of tables across structural and content-based categories.

Categories	Table Features	CompSci	MatSci	Report	ICDAR	Biology	Total
Structural	Compact	8	3	5	10	-	26
	Partially-bordered	10	10	5	10	5	40
	Multi-row	10	5	6	10	6	37
	Borderless	10	5	5	10	5	35
	Simple	10	10	5	10	5	40
	LongContentRow	11	-	5	-	6	22
	Sparse	-	9	6	9	4	28
	Multi-column	11	6	5	10	4	36
Content-based	Subscript	-	7	-	-	3	10
	Superscript	-	5	-	-	7	12
	Notations	4	5	-	-	5	14
	Equations	5	10	-	-	5	20
	Total Tables	79	75	42	69	55	320

Recent benchmarks have expanded the evaluation landscape for TableQA. For instance, TableBench [17] provides a broad evaluation across 18 reasoning dimensions and diverse industrial scenarios. However, its multi-modal design and emphasis on real-world data heterogeneity make it less effective for standalone tables. Domain-specific datasets such as AIT-QA [7], which targets the airline industry, and KET-QA [5], which integrates external knowledge for enhanced reasoning, are constrained by their reliance on information beyond the tables.

More recently, the SCITAT benchmark [19], which consists of several reasoning types, including look-up, numerical reasoning, data analysis, and tabulation, was proposed. SCITAT requires reasoning over both tables and the associated textual context. However, its reliance on textual information makes it unsuitable for evaluating standalone tables. Additionally, although SCITAT covers a broad range of reasoning subtypes, its coverage of domain-specific scientific calculations remains limited. Table 1 shows a comparison of the SciTableQA dataset with existing TableQA datasets.

Our work, SciTableQA, addresses these limitations by focusing exclusively on questions generated from standalone complex scientific tables. SciTableQA emphasizes high-level numerical reasoning, structural interpretation, and data interdependencies without requiring external textual references. By curating a diverse set of complex scientific tables spanning disciplines including MatSci, Biology, CompSci, ICDAR-2013, and Reports, and generating QA pairs across reasoning types, SciTableQA provides a challenging benchmark for evaluating LLMs on complex scientific tables.

3 SciTableQA Dataset

Our workflow (Fig. 1) combines automatic QA generation with manual human verification to efficiently create the dataset without sacrificing the quality.

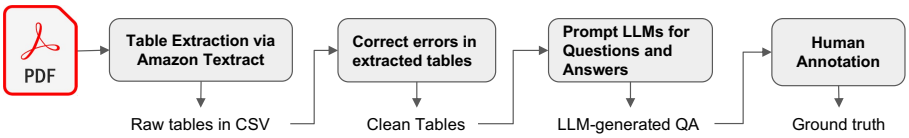


Fig. 1. The workflow to generate SciTableQA.

3.1 Data Collection and Pre-processing

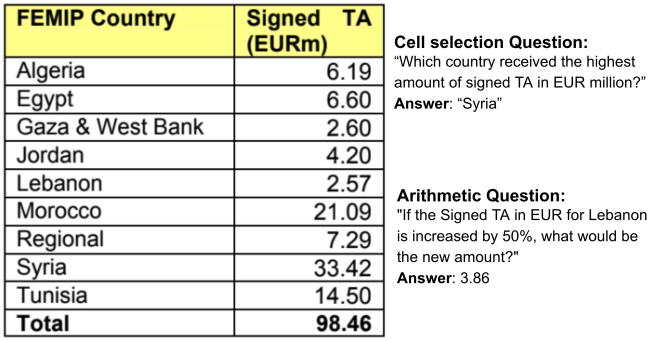
Data Source We construct SciTableQA by sourcing tables from publicly available scientific documents across multiple repositories, including arXiv, PubMed, IEEE Xplore, epa.gov, and ICDAR-2013. We surveyed 300 PDFs and manually curated tables that exhibit distinct or combined structural and content-based features to ensure a diverse coverage. To maintain relevance with modern table presentation styles, we limited the paper selection to publications from 2013 onward. The ICDAR-2013 dataset provides tables used for document recognition tasks, featuring structures such as borderless tables and sparse layouts. Tables from MatSci and Biology include experimental data and computational results, often with nested headers and domain-specific symbols. Scientific reports include tables from diverse scientific fields, such as Physics and Environmental Science, with diverse formats, and CompSci papers contribute data from algorithm analyses, benchmarks, and performance evaluations.

Table Extraction From PDF Papers We extract tables in text form from the obtained scientific papers in PDF format using Amazon Textract, a commercial optical character recognition (OCR) tool designed for high-accuracy text and table extraction. Amazon Textract has its ability to extract tables containing complex structures in PDFs. Textract also preserves the row-column relationships within tables, facilitating structured data conversion.

After extraction, we export each table into a CSV file. Saving the table as CSV is necessary because it enables precise programmatic access to each cell's content and location, which is essential for generating accurate QA pairs. However, the output of OCR engines may contain errors, especially for scientific tables with intricate layouts. Errors such as misaligned columns, missing cell values, incorrect merging of multi-row headers, and misinterpretation of symbols or units were common. If left uncorrected, they could propagate and introduce downstream errors, leading to unreliable QA generation and incorrect ground-truth answers. Therefore, we manually verify and revise extracted tables to match their original structure in the source papers.

3.2 Question-Answer Generation

To generate the SciTableQA dataset, we first prompt three representative LLMs: GPT-3.5-turbo [1], Llama 3-8B Instruct [4], and Mistral 7B [6] to generate QA pairs. In this paper, we will refer to these models as GPT-3.5, Llama-3, and Mistral for brevity. The use of multiple LLMs was motivated by the need to evaluate model performance across diverse scientific domains and to ensure variability in



FEMIP Country	Signed TA (EURm)
Algeria	6.19
Egypt	6.60
Gaza & West Bank	2.60
Jordan	4.20
Lebanon	2.57
Morocco	21.09
Regional	7.29
Syria	33.42
Tunisia	14.50
Total	**98.46**

Fig. 2. Examples of cell selection and arithmetic questions on a bordered table.

question formulation and reasoning complexity. By leveraging different models, we introduce diversity in linguistic structure, reasoning styles, and potential error patterns. Although newer LLMs have been developed, as we will show, these LLMs are sufficient to generate diverse questions. The major goal here is to develop the dataset and the evaluation framework, which can be applied to any LLMs.

We prompt LLMs to generate two types of reasoning questions: **cell selection** and **arithmetic** reasoning. Cell selection questions require retrieving a specific value from a single table cell or a set of related cells based on one or multiple column and row headers. An example of cell selection and arithmetic questions is shown in Fig. 2. These questions test the ability of LLMs to accurately extract information from the given table. Arithmetic questions involve performing numerical operations such as summation, averaging, percentage difference, and comparison across a minimum of three cells. These require the LLMs to correctly interpret numerical values and perform precise mathematical operations. Given that scientific tables often require complex reasoning over quantitative data, incorporating both types of questions is essential for a well-rounded evaluation.

3.3 Prompting Strategy

To generate high-quality QA pairs, we employ a zero-shot chain-of-thought (CoT) [8,16] prompting strategy, which instructs LLMs to ask questions on a given table, provide answers to the questions asked, and supply an explanation on how they arrive at the answers.

For consistency, all LLMs use the same prompting templates for asking and answering questions – one for cell selection and one for arithmetic reasoning. The templates are carefully crafted to elicit diverse and logically structured questions while maintaining uniformity in reasoning complexity. Each table is associated with 10 unique QA pairs, with 5 cell selection questions and 5 arithmetic reasoning questions.

Table 3. Distribution of QA pairs by Question-Types generated by each LLM.

LLMs	#Cell Selection	#Arithmetic	Total
Llama 3-8B	1,453	1,452	2,905
GPT 3.5-Turbo	1,471	1,425	2,896
Mistral-7B	1,471	1,458	2,929
Total	4,395	4,335	8,730

3.4 Human Verification

Because LLMs can hallucinate, generate duplicate questions, and produce incorrect answers, human verification is essential to ensure the quality and reliability of the final data. Unlike traditional annotation workflows, where humans correct errors in model-generated questions and answers, we recruited and trained 5 graduate students in computer science to answer questions based on the table data **without** seeing the answers generated by LLMs, to mitigate the potential prejudgment bias. The LLM-generated questions are not changed.

To maintain verification quality, we employ two annotators to answer each question and resolve any discrepancies between them through consensus discussions; any disagreements are escalated to a senior reviewer. We achieved an inter-annotator agreement (Kappa) score of 0.8, reflecting strong consistency in the annotation process. The final dataset contains all verified QA pairs and LLM explanations along with their ground-truth answers.

By structuring the human annotation process in this manner, SciTableQA enables a direct comparison between the answers generated by LLMs and the ground-truth answers provided by human annotators. Therefore, the performance metrics directly measure the models' reasoning capabilities, without being affected by post hoc adjustments or corrections.

3.5 Data Diversity

This section provides a detailed analysis of the dataset's diversity, covering the distribution of tables across domains and structural categories, as well as the breakdown of QA pairs generated by each LLM.

Distribution of Tables Across Scientific Domains SciTableQA includes tables from five scientific domains. As shown in Table 2, CompSci and MatSci have the most number of tables (79 and 75, respectively). Report and Biology have the least number of tables (42 and 55, respectively).

Diversity of Table Features The tables in SciTableQA incorporate diverse structural and content-based features. Table 2 shows 8 structural features and 4 content features, at least 1 structural feature (long content row) and 4 content features (subscript, superscript, notation, equation) being new compared to the existing (i.e., ICDAR-2013) datasets. Notations and equations introduce domain-specific complexity, requiring advanced reasoning. Additionally, sparse and long-content row tables vary in data density, testing LLMs' ability to extract relevant

information from scattered or text-heavy rows. The combined feature analysis (Fig. 3) further shows that non-ICDAR tables exhibit greater diversity across both structural and content-based categories, highlighting the broader challenge posed by SciTableQA compared to prior benchmarks.

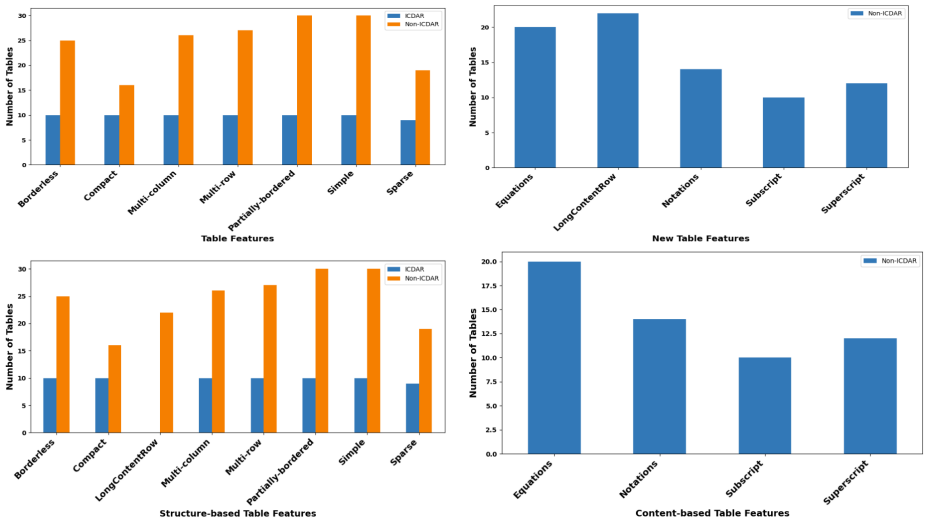


Fig. 3. The number of tables with different structural and content-based features in SciTableQA for ICDAR-2013 and non-ICDAR-2013 tables. Upper left: table features in ICDAR and other domains; Upper right: table features that ICDAR-2013 do not contain; Lower left: table features describing structural complexities of the tables; Lower right: table features describing content complexities of the tables.

Diversity of Question Types Our dataset contains 4 interrogative types of questions and 3 imperative types of questions. The interrogative type includes "What"-type questions, which constitute approximately 70% of all questions. Other notable interrogative types include "Which" (16%), "How many" (4%), and "How much" (1.5%). The remaining 5% are imperative questions that start with "Find", "Calculate", and "Determine", signaling fact-based retrieval tasks.

Multirow and Multicolumn Reasoning One characteristic of our questions is that the reasoning may need multiple rows and/or columns. In Fig. 4, we examine the joint distribution of rows and columns involved in answering each question. Although most questions involve a single row and/or a single column (41%), the questions also includes a wide range of header pairings, such as two rows and 1 column (2R-1C), 1R-2C, and more complex combinations such as 3R-8C and 3R-3C demonstrating that SciTableQA is designed to challenge models with multi-dimensional data extraction and aggregation tasks. This variety ensures that the benchmark evaluates LLMs' ability to reason over both simple and complex table structures.

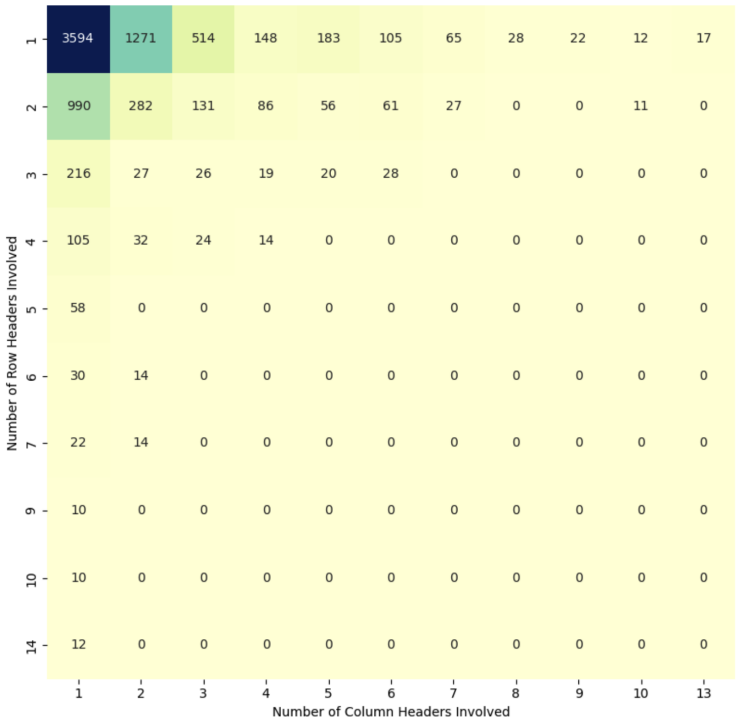


Fig. 4. Joint distribution of rows and columns involved in questions generated by LLMs, highlighting the diverse range of information required to answer questions .

4 Evaluation

4.1 Accuracy and Reasoning on Self-Generated QA

We first evaluate the accuracy of GPT-3.5, Llama-3, and Mistral on their self-generated QA pairs. For each LLM-generated QA pair, we compare the LLM's answer to the ground truth answer provided by the human answerer. In addition, we assess the reasoning process by analyzing the explanations provided by the LLMs for each answer. We analyze the model performance in three dimensions: question types, domains, and table features.

Model Performance on Arithmetic Reasoning As shown in Fig. 5, model performance varies significantly between cell selection and arithmetic reasoning questions, with cell selection yielding higher accuracy (Acc) across all models. Accuracy in this context is computed based on the exact match between the LLM-provided answer and the ground truth answer. Specifically, GPT-3.5 achieves the highest accuracy scores, with (Acc_CS, Acc_Arith) = (0.790, 0.490) for cell selection and arithmetic types of questions, respectively, followed by Mistral (0.647, 0.410) and Llama-3 (0.585, 0.278).

Approach - The paper constructs the SciTableQA benchmark by collecting complex scientific tables from multiple domains, generating question–answer pairs using LLMs, verifying them through human annotation, and evaluating LLM performance across structural, content, and domain-specific reasoning tasks.

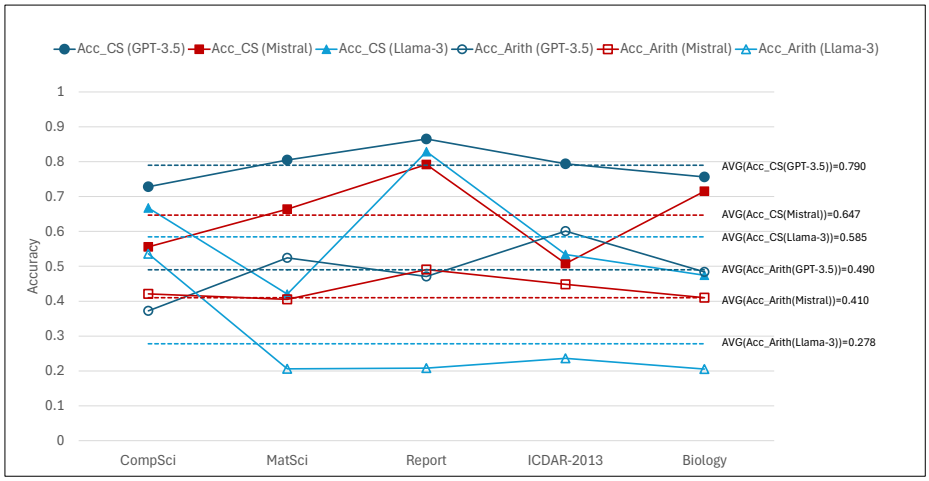


Fig. 5. Accuracy scores (Acc) on cell selection (CS) and arithmetic (Arith) reasoning tasks for self-generated QA pairs.

For Acc_CS, all LLMs exhibit domain dependencies. Mistral and Llama-3 exhibit larger standard deviation (STD_S) (0.116 and 0.165, respectively) than GPT (0.052). The LLMs ranked by the average Acc_CS from the highest to the lowest are GPT-3.5, Mistral, and Llama-3.

All LLMs achieve the highest Acc_CS for the Report tables. The domains they perform the worst in are CompSci, ICDAR-2023, and MatSci for GPT-3.5, Mistral, and Llama-3, respectively.

For Acc_Arith, similar to Acc_CS, all LLMs exhibit domain dependencies, but this time, GPT-3.5 and Llama-3 exhibit larger STD_S (0.083 and 0.145, respectively). The LLMs ranked by the average Acc_Arith from the highest to the lowest are GPT-3.5, Mistral, and Llama-3.

4.2 Reasoning Evaluation on Arithmetic Questions

We assess the quality of reasoning processes used by LLMs when answering scientific table-based questions based on the correctness of answers to the arithmetic questions and how accurately they explain the answers. This measures how well the explanations provided by LLMs align with the ground truth explanations. We manually score each explanation on a 0–3 scale, called reasoning validity (RV) scores, defined as follows:

- **3**: The explanation is aligned with the human evaluation and leads to the correct answer (matching the ground truth).
- **2**: The explanation is mostly aligned with human evaluations but contains minor wrong steps that lead to the wrong answer.
- **1**: The explanation is partially aligned with human evaluations or contains major flaws leading to a wrong answer.
- **0**: No explanation is provided, or the reasoning is completely incorrect.

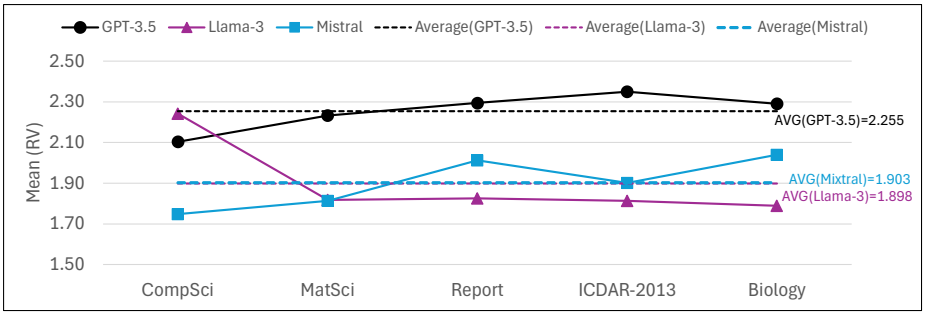


Fig. 6. The average reasoning validity (RV) scores on arithmetic tasks for GPT-3.5, Llama-3, and Mistral across the five domains.

The evaluation results (Fig. 6) reveal key insights into the models' reasoning capability across domains. GPT-3.5 consistently achieves the higher reasoning validity scores, demonstrating superior logical coherence and step-by-step explanations across all domains except computer science. All three models show unequal performance across all domains with Llama-3 exhibiting significant drops in Biology and MatSci. Llama-3's reasoning validity is notably higher in CompSci compared to GPT-3.5 and Mistral.

4.3 Cross-LLM QA

To further assess the robustness of the LLMs in reasoning over complex scientific tables, we conduct a cross-LLM evaluation on arithmetic tasks using the accuracy metric, where each model answers questions generated by other models. We focus on arithmetic tasks because arithmetic reasoning is substantially more challenging for the models than cell selection, making it a more meaningful target for evaluating cross-model generalization. The cross-LLM evaluation on arithmetic tasks highlights significant variations in model performance across table features and scientific domains. We organize table features into a structural category and a content-based category, shown in Table 2. Structural features delineates complex layouts, while content-based features include domain-specific symbols, notations, and mathematical content.

Structural Category Fig. 7 shows that GPT-3.5 consistently outperforms Mistral, when answering questions generated by Llama-3. Specifically, GPT-3.5 achieves accuracy scores above 0.5 on challenging features such as notations and long content-row tables. In contrast, Mistral struggles across most table features (Fig. 7). Although Llama-3 performs relatively well in table with features such as superscripts and subscripts, its accuracy decreases for tables with more intricate structures, such as borderless and partially-bordered, highlighting its limitations in managing structural complexity.

Content-Based Categories Content-based categories, such as notations, equations, superscripts, and subscripts, test the models' ability to process domain-specific and semantic content. As shown in Fig. 8, Llama-3 outperforms GPT-

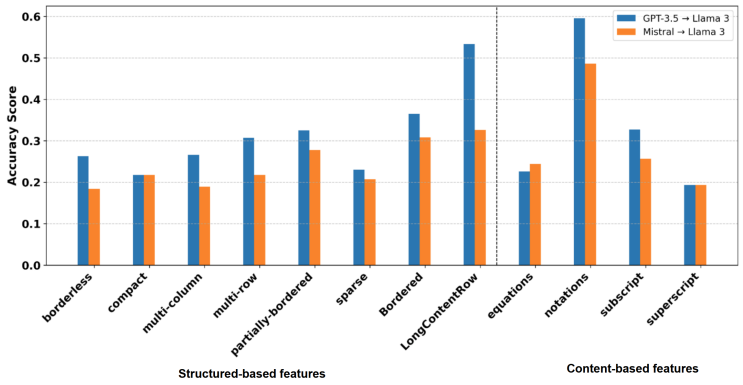


Fig. 7. Comparison of GPT-3.5 and Mistral on Llama-3 questions across two categories of table features.

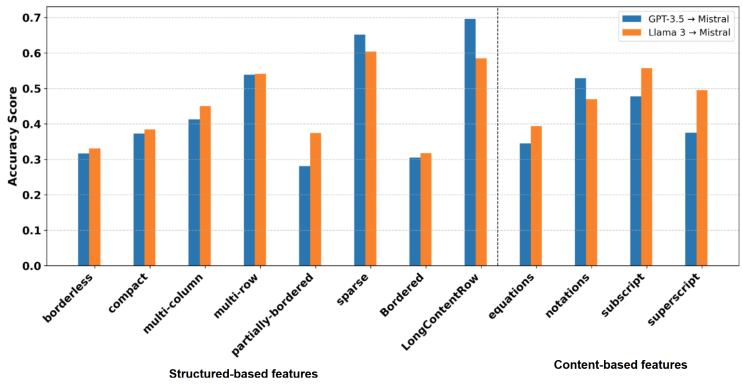


Fig. 8. Comparison of GPT-3.5 and Llama-3 on Mistral questions about tables with two categories of features.

3.5 for tables with 3 out of 4 content-based features, demonstrating greater consistency in handling content-dense tables. For instance, Llama-3 achieves higher accuracy scores than GPT-3.5 for features such as subscripts and superscripts. Mistral, whose performance is relatively even across both structural and content-based groups, still exhibits lower performance, particularly in equations, where domain-specific reasoning is critical. These findings highlight the challenges posed by content-heavy categories and reveal Llama-3's comparative advantage in scenarios requiring semantic understanding, particularly when contrasted with GPT-3.5.

Performance by Domain At the domain level, the evaluation reveals distinct patterns in model performance influenced by the structure and content of the tables in each domain. As shown in Fig. 9 and Fig. 11, domains such as CompSci and ICDAR-2013 demonstrate relatively better accuracy scores especially by

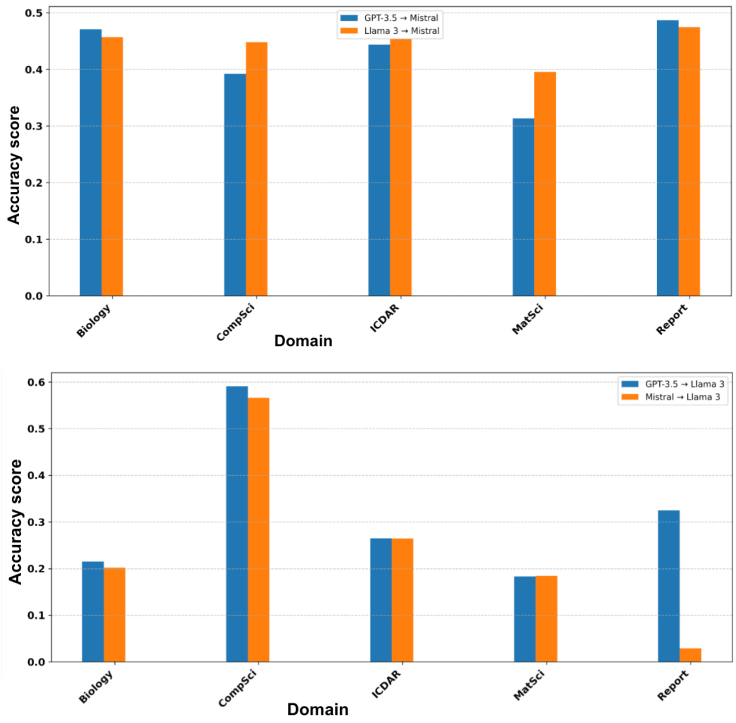


Fig. 9. Comparison of GPT-3.5 and Llama-3 on Mistral questions (upper), and GPT-3.5 and Mistral on Llama-3 questions (lower) across five domains.

GPT-3.5 and Llama-3. GPT-3.5, in particular, exhibits a strong performance advantage in the CompSci domain, achieving higher accuracy scores than Mistral when answering questions generated by Llama-3 (Fig. 9).

Conversely, all models face significant challenges for tables in Biology and Report domains. For example, when answering questions generated by Llama-3 in the Report domain as shown in Fig. 9 (lower panel), GPT-3.5's performance shows a noticeable decline. This indicates that domain-specific complexities, such as dense semantic content or specialized notations, remain a bottleneck for effective reasoning, specifically for GPT-3.5.

Interestingly, while GPT-3.5 generally outperforms Llama-3 across most domains, its advantage diminishes in MatSci, where Llama-3 achieves competitive scores, particularly when answering Mistral-generated questions (Fig. 9 upper panel).

4.4 Reasoning Scoring

We adopted a manual scoring and an automatic scoring method to assess the explanations generated by LLMs. This alignment between the two scores is critical for ensuring that LLMs provide semantically and logically consistent reasoning.

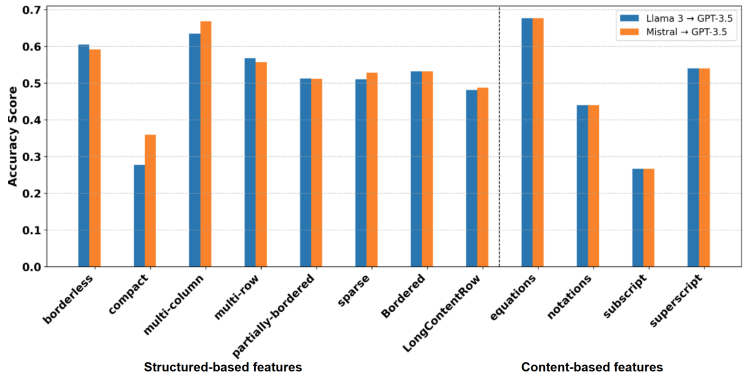


Fig. 10. Comparison of Llama-3 and Mistral on GPT-3.5 questions for tables with two categories of features.

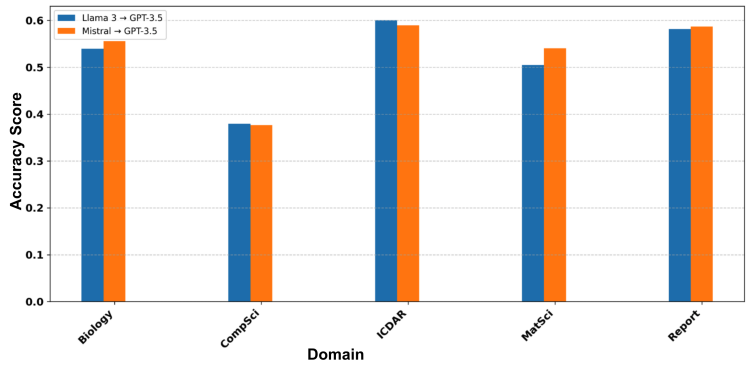


Fig. 11. Comparison of Llama-3 and Mistral on GPT-3.5 questions across five domains.

Alignment for Manual & Automatic Scoring For manual scoring, we randomly select 50 QA pairs in which one LLM generates the question, and a different LLM provides the answer and explanations (e.g., GPT-3.5 answering Llama-3's questions), resulting in 150 QA pairs. Each pair is evaluated based on whether the explanations of the answers to the same question provided by both LLMs are semantically and logically equivalent. If they are, the pair receives a score of 1; otherwise, it receives a score of 0. For example, suppose Llama-3 generates the question, "What is the sum of the values in the third column?" GPT-3.5 answers with "42," explaining that the values in the third column are 10, 15, and 17, and their sum is 42. Llama-3, when answering the same question, also provides "42" as the answer, explaining that adding 10, 15, and 17 results in 42. Since both explanations align in their reasoning and conclusion, this QA pair is assigned a score of 1.

Table 4. Comparison of Automatic vs. Manual Scoring Differences Across Evaluations. → indicates left model responding to questions asked by the right model.

Model Pair	Count		% Agreement
	Agree	Disagree	
GPT → Llama	35	15	70%
GPT → Mistral	35	15	70%
Llama → GPT	40	10	80%
Llama → Mistral	35	15	70%
Mistral → GPT	40	10	80%
Mistral → Llama	40	10	80%

For automatic scoring, we employ the Sentence Transformer model (all-MiniLM-L6-v2[3]) to compute the cosine similarity scores between embeddings of the two explanations. The equivalence label is determined by whether the similarity score is greater (equivalent) or smaller (non-equivalent) than a threshold. We tested 20 threshold values in a range between 0.5 and 1.0 for each of the 50 QA sets answered by two LLMs to determine the threshold that results in the highest value of F1 scores grounded on the human's scores. We found 0.7, 0.72, and 0.75 as best thresholds for three sets, which are very close.

To validate the automatic scoring method, we calculate the difference between the manual and automatic scores for each QA pair. The higher proportion of equivalent pairs of explanations, as shown in Table 4, indicates a better alignment between the two methods, suggesting that the automatic scoring method is a reliable measure.

5 Conclusion

This paper introduces SciTableQA, a benchmark dataset designed to evaluate LLMs on the QA ability on standalone complex scientific tables. We evaluated three representative LLMs in three families, namely GPT-3.5, Llama-3, and Mistral. Our findings show that these models' TableQA capability varies depending on tables' structure, content types, and scientific fields. In general, GPT-3.5 excels in structured-based features, while Llama-3 performs well in content-based features, and Mistral performs evenly across both table features. We also demonstrate that semantic similarity-based automatic scoring has a high alignment with human's scoring and thus may be used for the consistency of explanations provided by two LLMs for the same QA pair. These results highlight the need for future LLMs to improve in numerical reasoning, structural understanding, and semantic consistency. SciTableQA sheds the light for more rigorous testing of LLMs on robust TableQA tasks. It supports the development of stronger AI systems for scientific research and beyond.

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

References

1. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
2. Chen, W., Zha, H., Chen, Z., Xiong, W., Wang, H., Wang, W.Y.: HybridQA: A dataset of multi-hop question answering over tabular and textual data. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1026–1036. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.91, https://aclanthology.org/2020.findings-emnlp.91/
3. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.H., Routledge, B., Wang, W.Y.: FinQA: A dataset of numerical reasoning over financial data. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3697–3711. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.300, https://aclanthology.org/2021.emnlp-main.300/
4. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
5. Hu, M., Dong, H., Luo, P., Han, S., Zhang, D.: KET-QA: A dataset for knowledge enhanced table question answering. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 9705–9719. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.848/
6. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)
7. Katsis, Y., Chemmengath, S., Kumar, V., Bharadwaj, S., Canim, M., Glass, M., Gliozzo, A., Pan, F., Sen, J., Sankaranarayanan, K., Chakrabarti, S.: AIT-QA: Question answering dataset over complex tables in the airline industry. In: Loukina, A., Gangadharaiah, R., Min, B. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track. pp. 305–314. Association for Computational Linguistics, Hybrid: Seattle, Washington + Online (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-industry.34, https://aclanthology.org/2022.naacl-industry.34/
8. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
9. Kruit, B., He, H., Urbani, J.: Tab2know: Building a knowledge base from tables in scientific papers. In: International Semantic Web Conference. pp. 349–365. Springer (2020)

10. Lu, X., Pan, L., Liu, Q., Nakov, P., Kan, M.Y.: SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 7787–7813. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.483, https://aclanthology.org/2023.emnlp-main.483/
11. Majumder, B.P., Surana, H., Agarwal, D., Mishra, B.D., Meena, A., Prakhar, A., Vora, T., Khot, T., Sabharwal, A., Clark, P.: Discoverybench: Towards data-driven discovery with large language models. arXiv preprint arXiv:2407.01725 (2024)
12. Moosavi, N.S., Rücklé, A., Roth, D., Gurevych, I.: 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) (2021)
13. Nan, L., Hsieh, C., Mao, Z., Lin, X.V., Verma, N., Zhang, R., Kryściński, W., Schoelkopf, H., Kong, R., Tang, X., et al.: FeTaQA: Free-form table question answering. Transactions of the Association for Computational Linguistics **10**, 35–49 (2022)
14. Pasupat, P., Liang, P.: Compositional semantic parsing on semi-structured tables. In: Zong, C., Strube, M. (eds.) 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). pp. 1470–1480. Association for Computational Linguistics, Beijing, China (Jul 2015). https://doi.org/10.3115/v1/P15-1142, https://aclanthology.org/P15-1142/
15. Wang, N.X., Mahajan, D., Danilevsky, M., Rosenthal, S.: Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts). arXiv preprint arXiv:2105.13995 (2021)
16. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
17. Wu, X., Yang, J., Chai, L., Zhang, G., Liu, J., Du, X., Liang, D., Shu, D., Cheng, X., Sun, T., et al.: TableBench: A Comprehensive and Complex Benchmark for Table Question Answering. arXiv preprint arXiv:2408.09174 (2024)
18. Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., Radev, D.: Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3911–3921. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1425, https://aclanthology.org/D18-1425/
19. Zhang, X., Wang, D., Wang, B., Dou, L., Lu, X., Xu, K., Wu, D., Zhu, Q., Che, W.: SCITAT: A Question Answering Benchmark for Scientific Tables and Text Covering Diverse Reasoning Types. arXiv preprint arXiv:2412.11757 (2024)
20. Zhong, W., Huang, J., Liu, Q., Zhou, M., Wang, J., Yin, J., Duan, N.: Reasoning over hybrid chain for table-and-text open domain question answering. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022. pp. 4531–4537. ijcai.org (2022). https://doi.org/10.24963/IJCAI.2022/629, https://doi.org/10.24963/ijcai.2022/629
21. Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., Chua, T.S.: TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceed-

ings 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). pp. 3277–3287. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.254, https://aclanthology.org/2021.acl-long.254/