

# TableEval: A Real-World Benchmark for Complex, Multilingual, and Multi-Structured Table Question Answering

Junnan Zhu<sup>2\*</sup>, Jingyi Wang<sup>1,2\*</sup>, Bohan Yu<sup>1,3</sup>,  
Xiaoyu Wu<sup>1</sup>, Junbo Li<sup>1</sup>, Lei Wang<sup>1,2</sup>, Nan Xu<sup>1,2†</sup>

<sup>1</sup> Beijing Wenge Technology Co., Ltd.

<sup>2</sup> State Key Laboratory of Multimodal Artificial Intelligence Systems,  
Institute of Automation, CAS, Beijing, China

<sup>3</sup> School of Advanced Interdisciplinary Sciences, UCAS, China  
junnan.zhu@nlpr.ia.ac.cn, nan.xu@wenge.com

## Abstract

LLMs have shown impressive progress in natural language processing. However, they still face significant challenges in TableQA, where real-world complexities such as diverse table structures, multilingual data, and domain-specific reasoning are crucial. Existing TableQA benchmarks are often limited by their focus on simple flat tables and suffer from data leakage. Furthermore, most benchmarks are monolingual and fail to capture the cross-lingual and cross-domain variability in practical applications. To address these limitations, we introduce TableEval, a new benchmark designed to evaluate LLMs on realistic TableQA tasks. Specifically, TableEval includes tables with various structures (such as concise, hierarchical, and nested tables) collected from four domains (including government, finance, academia, and industry reports). Additionally, TableEval features cross-lingual scenarios with tables in Simplified Chinese, Traditional Chinese, and English. To reduce potential data leakage, we curate data from recent real-world documents. Considering that existing TableQA metrics fail to capture semantic accuracy, we further propose SEAT, a new evaluation framework that assesses the alignment between model responses and reference answers at the sub-question level. Experimental results have shown that SEAT achieves high agreement with human judgment. Extensive experiments on TableEval reveal critical gaps in the ability of state-of-the-art LLMs to handle these complex, real-world TableQA tasks, offering insights for future improvements. We make our dataset available here: <https://github.com/wenge-research/TableEval>.

## 1 Introduction

In recent years, Large Language Models (LLMs) have made remarkable progress in natural language

processing tasks such as text generation, reasoning, and question answering. However, the complexity of table-based Question Answering (TableQA) poses distinctive challenges. Unlike texts, tables often exhibit nested structures, multi-row/column spans, and cross-references that demand domain-specific and structure-aware reasoning.

Despite significant advances in TableQA, current benchmarks fail to capture the key challenges encountered in industrial and real-world applications. First, most (Chen et al., 2020; Parikh et al., 2020; Chen et al., 2021a; Nan et al., 2022; Wu et al., 2024) focus on simple flat tables, ignoring the diverse and complex structures commonly encountered in real-world data, such as hierarchical, concise, and nested structures. Second, existing benchmarks suffer from data leakage, where test data overlaps with the pretraining corpora of LLMs. This issue leads to a bias in performance evaluation, producing misleading results that fail to reflect actual generalization ability. Finally, the predominantly monolingual nature of most benchmarks fails to address the cross-lingual and cross-domain challenges critical for real-world applications. Practical use cases often require the ability to understand and reason over data presented in multiple languages and originating from diverse domains.

To address these limitations, we introduce TableEval, a new benchmark to evaluate LLMs on broader and more representative tasks. TableEval contains diverse tables covering various structures, such as concise<sup>1</sup>, hierarchical, and nested tables. We minimize the risks of data leakage by ensuring that all data is sourced from recent documents (published in 2024) that have not been included in the pretraining data of most existing LLMs. Besides, we highlight the challenges of both cross-lingual and cross-domain settings by including tables from

\*Equal Contribution.

†Corresponding Author.

<sup>1</sup>We refer to concise tables as defined in Lautert et al. (2013), where merged cells are used to enhance compactness.

various domains (such as administrative records, financial disclosures, academic papers, and industry reports) in multiple languages (Simplified and Traditional Chinese, English), with all QA pairs presented in Simplified Chinese.

TableEval covers six high-level tasks and 16 fine-grained question types, ranging from simple lookup queries to numerical calculations, comparative analysis, and multi-hop questions. The benchmark also includes multi-turn conversation tasks, where models are required to perform dynamic reasoning across multiple steps. Critically, all questions have unambiguous, objectively verifiable answers to ensure consistent and reliable evaluation.

Evaluating LLMs in TableQA tasks is challenging, primarily because LLMs often generate natural language responses that vary significantly in phrasing and structure. Even when specific output formats are enforced via prompting, LLMs tend to deviate, making it difficult to evaluate the semantic accuracy of responses using traditional metrics like exact match (Zhu et al., 2021), F1 score (Chen et al., 2021b), or n-gram matching (Parikh et al., 2020; Nan et al., 2022). To address this limitation, we propose SEAT (Structured Evaluation for Answers in TableQA), a novel TableQA evaluation framework that leverages an LLM with crafted prompts to compare generated responses against structured reference answers. SEAT evaluates the semantic correctness of responses by following a two-step process: (1) Extract key answers to each sub-question from the model’s response and compare each with the reference to identify matching or divergent elements. If no multiple sub-questions are provided, the original question is treated as a single sub-question. (2) Present the evaluation results in a structured JSON format and aggregate the final scores, facilitating easy tracking and verification. This approach provides a reliable, scalable evaluation method that goes beyond surface-level matching, offering a more accurate assessment.

We systematically assess LLMs on TableEval. Our extensive experiments reveal several notable observations: (1) Closed-source models (e.g., o1-preview, Claude 3.5 Sonnet) consistently lead in performance, but large open-source models (e.g., DeepSeek-R1, QwQ-32B-Preview) show significant promise with proper scaling and enhanced reasoning capabilities. (2) Understanding complex table structures remains a fundamental challenge, with performance drops of 10–15% on nested/hierarchical tables compared to flat tables. (3) Domain-

specific and language-specific gaps persist, as models often struggle more in Chinese or specialized domains than in English or general settings.

Our main contributions are as follows:

- We introduce TableEval, a comprehensive real-world benchmark with diverse table structures and cross-lingual/cross-domain challenges using recent data to prevent leakage and realistically evaluate LLMs’ generalization capabilities.
- We propose SEAT, a novel evaluation framework that assesses model responses at the sub-question level, ensuring semantic alignment with reference answers. Through extensive experiments on multiple LLM backbones, SEAT demonstrates consistently higher correlations with human judgments than existing metrics.
- Our experiments reveal critical performance gaps in real-world TableQA: closed-source models lead overall but open-source alternatives become competitive with sufficient scaling and enhanced reasoning; complex table structures significantly degrade performance; and domain/language-specific weaknesses indicate the need for more effective strategies.

## 2 Related Work

**TableQA Methods.** Question answering over tabular data has attracted increasing interest due to its potential to make complex data more accessible to non-experts. Early studies on TableQA primarily explored semantic parsing methods, treating the task as Text-to-SQL, where natural language queries and table schemas are directly converted into executable SQL (Pasupat and Liang, 2015; Wang et al., 2020; Shi et al., 2020; Guo et al., 2021). While these methods show great potential, they require a large amount of SQL-labeled training data and often assume that column headers are enough for reasoning, ignoring the valuable information in table cells. To overcome these problems, recent studies have shifted towards the retrieve-and-reasoning paradigm (Eisenschlos et al., 2021; Zhong et al., 2022; Kumar et al., 2023; Lei et al., 2023) instead of directly generating SQL. Large-scale table pre-training (Yin et al., 2020; Herzog et al., 2020; Liu et al., 2022; Jiang et al., 2022; Zhao et al., 2022) has further enhanced this shift, which has demonstrated strong performance on

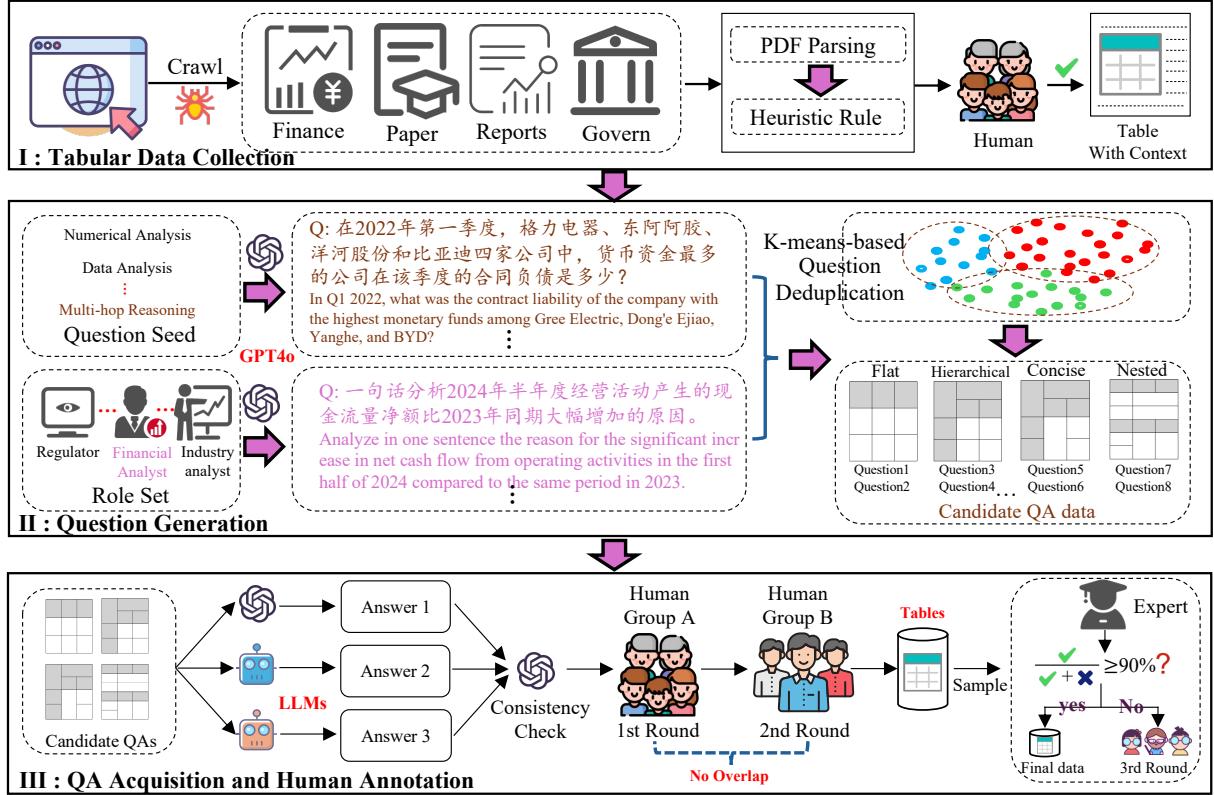


Figure 1: Overview of data collection. (1) Tabular Data Collection, collecting tables from financial reports, industry research, academic papers, and governmental data; (2) Question Generation, using Template-Prompted and Role-Prompted strategies to generate TableQA questions, filtered through clustering and deduplication; (3) QA Acquisition and Human Annotation—iteratively refining answers through LLM consistency checks, human reviews, and structured answer extraction, ensuring accuracy, completeness, and alignment with the original tabular data.

standard benchmarks. Furthermore, the advent of LLMs enhanced with few-shot prompting (Zhang et al., 2024a,b; Wang et al., 2024; Sui et al., 2024; Li et al., 2024) has significantly extended these capabilities, enabling more flexible, scalable, and generalizable reasoning over tabular data.

**TableQA Benchmarks.** While TableQA solutions continue to advance, existing benchmarks have yet to fully catch up with this progress. Early benchmarks like WQA (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), and TabFact (Chen et al., 2020) use HTML tables from Wikipedia, focusing on cell retrieval and extraction tasks with relatively simple table structures. To address more complex reasoning, later benchmarks introduce greater diversity and difficulty. Datasets such as ToTTo (Parikh et al., 2020), OTTQA (Chen et al., 2021a), and FeTaQA (Nan et al., 2022) require generating answers beyond the table content, while numeric-focused datasets like FinQA (Chen et al., 2021b) and AIT-QA (Katsis et al., 2022) emphasize computational reasoning in financial contexts. Additionally, logical expression-based benchmarks like Spider (Yu et al., 2018) and Bird (Li et al.,

2023) incorporate structured logical reasoning for handling intricate, logic-driven queries. However, existing benchmarks have the following limitations. Most focus on simple flat table structures, ignoring the hierarchical and nested structures encountered in the real world. Another challenge is data leakage, where overlaps between test data and pretraining corpora lead to biased evaluation. Additionally, most benchmarks are monolingual, lacking the cross-lingual and cross-domain diversity required for real-world applicability.

Therefore, we introduce a comprehensive benchmark that covers various table structures, supports multiple languages, and spans diverse real-world tasks and question types. It offers a more accurate evaluation of TableQA systems. We compare TableEval with existing benchmarks in Table 1.

### 3 Construction of TableEval

#### 3.1 Task Formulation

Given a table  $T$ , which may be presented in complex structures such as hierarchical, nested, or others, and a corresponding natural language question

Dataset	Flat	Hierarchical Bodies	Hierarchical Headers	Nested	Multi-Source	Table Language	QA Language	Information Retrieval	Numerical Analysis	Reasoning	Data Analysis	Multi-turn Conversation	Table Structure Understanding
WTQ (Pasupat and Liang, 2015)	✓	✗	✗	✗	✗	EN	EN	✓	✓	✗	✗	✗	✗
SQA (Iyyer et al., 2017)	✓	✗	✗	✗	✗	EN	EN	✗	✗	✗	✓	✗	✗
TabFact (Chen et al., 2020)	✓	✗	✗	✗	✗	EN	EN	✗	✗	✗	✓	✗	✗
FinQA (Chen et al., 2021b)	✓	✗	✗	✗	✗	EN	EN	✗	✓	✗	✗	✗	✗
TAT-QA (Zhu et al., 2021)	✓	✗	✗	✗	✗	EN	EN	✓	✓	✗	✗	✗	✗
OTT-QA (Chen et al., 2021a)	✓	✗	✗	✗	✗	EN	EN	-	-	✓	-	-	-
FéTaQA (Nan et al., 2022)	✓	✗	✗	✗	✗	EN	EN	✓	-	-	-	-	-
AIT-QA (Katsis et al., 2022)	✓	✗	✓	✗	✗	EN	EN	✓	✓	✗	✗	✗	✗
IM-TQA (Zheng et al., 2023b)	✓	✓	✓	✗	✓	ZH	ZH	✓	✗	✗	✗	✗	✗
TableBench (Wu et al., 2024)	✓	✗	✗	✗	✓	EN	EN	✓	✓	✓	✓	✗	✗
TableEval (Ours)	✓	✓	✓	✓	✓	EN/ZH/ZH-HK	ZH	✓	✓	✓	✓	✓	✓

Table 1: Comparison of TableEval with other TableQA datasets.

$Q$  related to  $T$ , the objective is to generate an accurate natural language answer  $A$ . Unlike existing TableQA tasks that require varied output formats based on the task type, our benchmark standardizes the output to natural language responses to precisely assess the comprehensive TableQA capabilities of LLMs. This approach accommodates various task types within diverse real-world industrial scenarios, including information retrieval, numerical analysis, table reasoning, data analysis, multi-turn dialogue, and understanding of table structures. By supporting complex table structures and diverse data sources, our benchmark ensures that LLMs are evaluated on their ability to understand, interpret, and interact with intricate table data consistently and effectively using natural language.

### 3.2 Dataset Construction

**Tabular Data Collection.** To reduce the risk of data leakage and ensure our dataset accurately reflects current TableQA challenges, we exclusively collect documents published in 2024. We select four document types to capture a broad coverage of industrial and academic contexts: (i) financial reports and company announcements, (ii) industry/stock research reports, (iii) academic papers, and (iv) governmental data. These materials are sourced from publicly available authoritative channels to ensure completeness and legal compliance. Specifically, we collect financial documents and company announcements from the Shenzhen Stock Exchange<sup>2</sup>, industry/stock research reports from Eastmoney Securities<sup>3</sup>, academic papers from arXiv<sup>4</sup> and the China National Knowledge Infrastructure<sup>5</sup>, and governmental data from the National Bureau of Statistics<sup>6</sup>. During the selection, we maintain a balanced distribution of document types to avoid the dataset being biased towards any single source.

After collecting PDFs (and a limited number

of HTML pages), we use parsing tools designed explicitly for PDF and HTML formats to extract tables into Excel worksheets, yielding 29,241 tables. To keep critical contextual information, we apply heuristic rules (see Figure 14 and Algorithm 1 in Appendix for details) to capture any adjacent text snippets near the tables, such as table captions or explanatory notes, so that the semantic context essential for question answering remains. Subsequently, three graduate students with finance and statistics backgrounds review the Excel files to correct parsing errors and ensure the table structures align accurately with the original PDFs. Each table is converted to Markdown format after reviewing. We provide an experiment on the impact of table format in Appendix B.3.

Next, annotators assign each table to one of seven categories: vertical, horizontal, matrix, concise, hierarchical, nested, and others. Finally, we sample 617 tables by considering factors such as table size, source, content length, and structural type. This sampling approach ensures a diverse yet balanced tabular dataset for benchmarking TableQA.

**Question Generation.** Based on the verified tables, we generate questions using two strategies, each designed to produce diverse questions.

**Template-Prompted Strategy.** We design prompt templates to guide GPT-4o in creating questions for various TableQA tasks, including simple lookups, numerical analysis, table size detection, and multi-hop inferences. The template structure and a curated seed pool of examples help guide GPT-4o to generate high-quality questions specific to different task types. However, these generated questions sometimes mirror the examples in the seed pool, limiting the question diversity.

**Role-Prompted Strategy.** To address the lack of diversity in questions, we simulate role-based scenarios that focus on different aspects of tabular data, such as those of investors, market analysts, and domain experts. Investors might inquire about liquidity or financial stability, while analysts could be more concerned with product strategies

<sup>2</sup><http://www.szse.cn/disclosure/listed/notice/index.html>

<sup>3</sup><https://data.eastmoney.com/report/>

<sup>4</sup><https://arxiv.org>

<sup>5</sup><https://www.cnki.net>

<sup>6</sup><https://www.stats.gov.cn/sj/zxfb/>

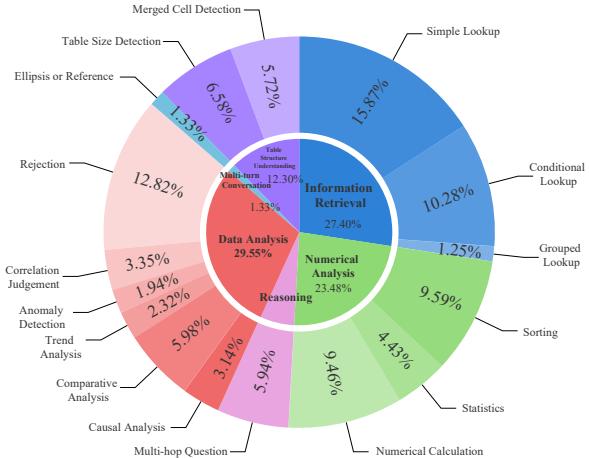


Figure 2: Task distribution of our TableEval.

and market trends. To ensure consistency with the task types in the Template-Prompted Strategy, we prompt GPT-4o to categorize each generated question into a predefined task type (e.g., simple lookups, numerical analysis). This approach better reflects actual usage contexts yet finds it challenging to produce highly complex or long-tail queries beyond typical stakeholder interests.

Combining both strategies, we obtain an extensive collection of QA pairs associated with multiple user perspectives and table formats. However, the automatically generated answers are often concise and lack detailed reasoning. Therefore, we only keep the generated questions. Additionally, questions often show an uneven distribution of task types due to two primary factors. First, the structure of the tables imposes inherent limitations on the complexity of the questions that can be generated. Second, while the role-prompted strategy uses role-driven scenarios to create questions, predicting the types of questions in advance remains challenging. For example, flat, horizontally oriented tables typically generate simple fact-based questions, such as “*What is the price of X?*”. In contrast, hierarchical tables, which can represent more complex relationships, are better suited for supporting multi-step inferential or aggregative questions.

We use K-means clustering within each task for sampling and deduplication to enhance question diversity and ensure a balanced task distribution. For each task, we set the expected number of QA pairs as the number of clusters. Specifically, we use BGE-M3 (Chen et al., 2024) to obtain question embeddings and cluster them by similarity, retaining only one representative query per cluster. This process yields 34,161 distinct QA pairs. Among these, 5,422 pairs (spanning 16 tasks) are selected

Type	Flat			Hier.	Conc.	Nested	Oth.
	V	H	M				
#Table	166	6	69	243	99	13	21

Table 2: Table Structure-wise Statistics. V, H, M, Hier, Conc, and Nested represent vertical, horizontal, matrix, hierarchical, concise, and nested table types.

as candidates for the final test set, with the remainder serving as the candidate training set.

**QA Acquisition and Human Annotation.** Human annotation is essential for ensuring accuracy and consistency in our final question-answer pairs. Initially, we employ GPT-4o to generate free-text answers, supplemented by outputs from multiple additional LLMs. Our annotation team consists of six graduate students with finance and statistics backgrounds, trained on comprehensive guidelines covering task categorization, question decomposition, and structured answer formats.

**First-stage annotation** focuses on quality control and accuracy verification. Annotators review each QA pair according to four principles: eliminating ambiguous or impractical questions, verifying relevance to the underlying table content, correcting inaccuracies in LLM-generated answers, and filtering out overly simple questions (those correctly answered by all LLMs with brief responses). Two independent annotators review each pair, with team discussions to resolve any inconsistencies, followed by expert sampling reviews. If acceptance rates drop below 90% during sampling review, we initiate additional human verification.

**Second-stage annotation** employs a structured answer extraction process that determines whether original questions require decomposition into sub-questions and extracts JSON-formatted answers from annotated responses. This stage filters redundant and low-relevance information from natural language responses, refining answer content to improve comparison accuracy. The JSON-formatted answers undergo two rounds of human verification and one round of expert review to ensure accuracy and consistency with the original table data and question context. The comprehensive annotation process averaged approximately \$2.5 per sample.

**Second-stage annotation** employs a structured answer extraction process. We use LLMs to determine if the original question needs to be decomposed into sub-questions and to extract JSON-formatted answers from the annotated responses. This process filters redundant and low-relevance

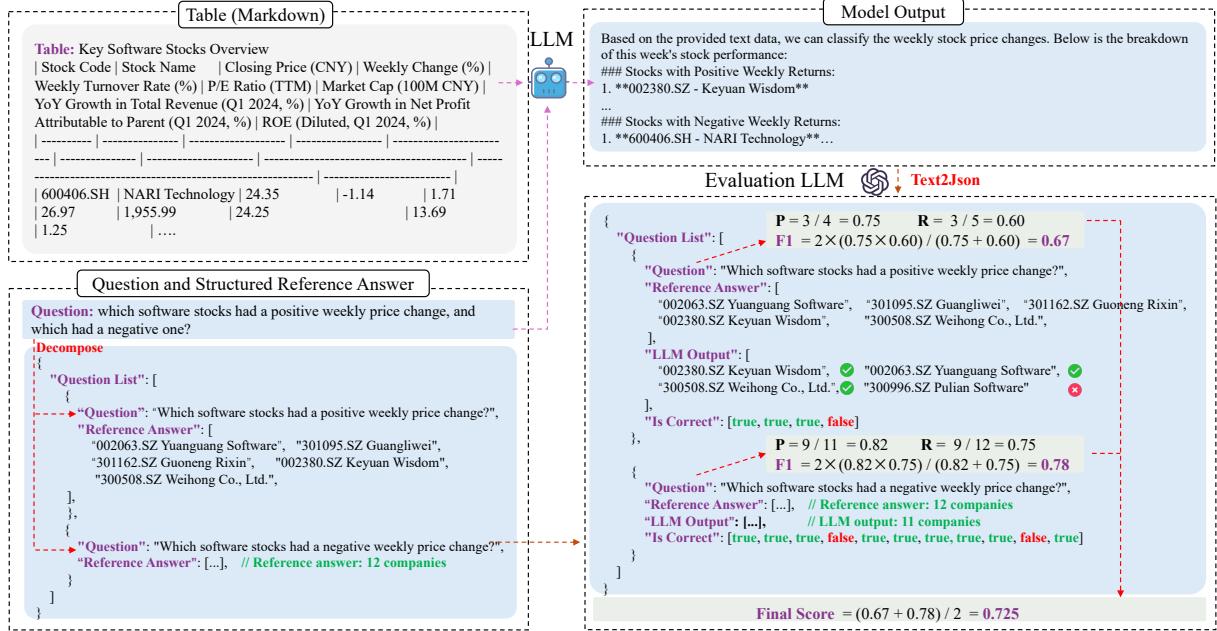


Figure 3: Overview of Our SEAT evaluation method.

information from natural language responses, refining the answer content to improve comparison accuracy. The resulting JSON-formatted answers then undergo two rounds of human verification and one round of expert review to ensure accuracy and consistency with the original table data and question context. The comprehensive annotation process averaged approximately \$2.5 per sample.

### 3.3 Dataset Statistics

We present the task and table structure distributions of our dataset in Figure 2 and Table 2, respectively. Our dataset comprises six main task categories, further divided into 16 subcategories, covering real-world scenarios such as simple and conversational retrieval, reasoning, numerical and data analysis, and table structure understanding. It includes 2,325 QA pairs, spanning Simplified Chinese, Traditional Chinese (Hong Kong), and English tables. See more details in Table 7 and Table 8.

Our dataset provides comprehensive coverage of various domain-specific subtasks, supporting a more fine-grained evaluation of cross-domain capabilities. For financial reports, we include subtasks such as financial metric retrieval, ratio and indicator trend analysis, and cross-period comparisons. Industry research reports encompass subtasks like industry performance comparison and financial indicator ranking. Academic papers contain experimental result queries and definition lookups. Government data includes price change calculations

and industry growth rate queries.

## 4 Evaluation Method for TableQA

Evaluating LLMs in TableQA is challenging because natural language responses can vary widely in phrasing and structure. Existing methods often use exact match (Zhu et al., 2021), F1-score (Chen et al., 2021b), or n-gram-matching (Parikh et al., 2020; Nan et al., 2022) metrics. However, these approaches often fail to assess the accuracy of LLM-generated responses effectively. To solve this problem, we propose a simple and effective evaluation framework, SEAT (Structured Evaluation for Answers in TableQA), that leverages an LLM with crafted prompts to compare generated responses with structured reference answers.

As shown in Figure 3, SEAT follows a structured process to evaluate the accuracy and relevance of model-generated responses, guided by the original question and a corresponding JSON-formatted reference answer. SEAT consists of two steps: (1) Extract key answers to each sub-question from the model's final response, compare each with the reference answers to identify matching or divergent elements, and label each extracted answer as "true" or "false" based on its alignment with the reference. (2) Present the evaluation results in a JSON format, including the original question, the model's response, and the evaluation record for each answer to facilitate tracking and verification. Compute the overall F1 for each QA pair by averaging the F1

scores across all sub-questions. For multi-turn conversational QA, first compute the average F1 for each turn, then average across all turns to obtain the final F1 score for the QA session.

Specifically, given a model-generated answer  $A$  and a reference answer  $R$ , both in natural language, SEAT converts them into structured forms and evaluates their alignment in a single step. It uses an LLM guided by a specific prompt (see Figure 12), which can be formalized as:

$$J_{\text{eval}} = \text{Text2Json}(A, R) \quad (1)$$

where Text2Json instructs the LLM to parse both  $A$  and  $R$ . The LLM is prompted to first extract key answer components from  $A$  corresponding to the question, then compare these components against the information in  $R$  to determine their correctness (labeling each as "true" or "false"). The final output,  $J_{\text{eval}}$ , is a single JSON object that contains the structured representations of both answers alongside the correctness judgments for each sub-question. Precision, recall, and the F1 score are then calculated from this structured output. This integrated approach ensures that the semantic comparison is directly tied to the structuring process, providing a robust and consistent evaluation.

## 5 Experiment

### 5.1 Experimental Setup

We evaluate 19 models, ranging from 7B to 671B parameters, to assess their performance on TableEval. The evaluated models include open-source and closed-source models to comprehensively compare different model architectures and capabilities. For open-source models, we evaluate Qwen2.5-Instruct series (Qwen et al., 2025): Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, Qwen2.5-72B-Instruct; Qwen2.5-Coder-Instruct-32B (Hui et al., 2024); QwQ-32B-Preview; DeepSeek series (DeepSeek-AI et al., 2024): DeepSeek-V2-Lite-Chat (16B), DeepSeek-Coder-V2-Lite-Instruct (16B), DeepSeek-V2.5-1210 (236B), DeepSeek-V3 (671B), DeepSeek-R1 (671B) (DeepSeek-AI et al., 2025); glm-4-9b-Chat (9B) (GLM et al., 2024); Llama-3 series (Grattafiori et al., 2024): Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct. For closed-source models, we evaluate qwen-max-2024-09-19, Claude 3.5 sonnet-20241022, GPT-4o-mini-2024-07-18, GPT-4o-2024-11-20 (OpenAI et al., 2024), and o1-preview. We set the temperature to 0 and

Metric		$r$	$\rho$	$\tau$
F1		.1776	.1222	.1014
ROUGE-L		.6132	.6030	.4988
LLM-as-a-judge	GPT4o	.8673	.8792	.7733
	Qwen-2.5-72B	.7080	.7333	.6592
	Llama-3.3-70B	.8052	.8311	.7214
SEAT	GPT4o	<b>.9373</b>	<b>.9346</b>	<b>.9062</b>
	Qwen-2.5-72B	.8257	.8137	.7726
	Llama-3.3-70B	.8410	.8359	.8214

Table 3: Correlation with human scores, measured with Pearson  $r$ , Spearman  $\rho$ , and Kendall  $\tau$  coefficients. We evaluate SEAT with different LLM backbones. LLM-as-a-judge (Zheng et al., 2023a) prompts GPT-4o to provide a 1-10 rating for evaluation.

use a fixed seed (42) to make the model’s output more deterministic. For models that do not allow the configuration of parameters (e.g., temperature or seed), we use their default settings.

To obtain detailed, interpretable, and structured responses for each question in the TableQA task, we conducted comparative experiments based on Chain-of-Thought (CoT) prompting (Wei et al., 2022). The specific CoT prompt used in our experiments are shown in the Figure 13, which align with those in previous studies (Zheng et al., 2024; Deng et al., 2024; Zhang et al., 2024c).

### 5.2 Effectiveness of SEAT

To assess the effectiveness of SEAT, we analyze the correlation between SEAT and human scores. Annotators rate model responses from 1 to 10 based on reference quality. We collect ratings for 5% randomly selected samples and evaluate the correlation using three metrics: 1) Pearson correlation coefficient ( $r$ ), 2) Spearman rank coefficient ( $\rho$ ), and 3) Kendall rank coefficient ( $\tau$ ). As shown in Table 3, SEAT consistently achieves higher correlations with human judgments than existing metrics (F1 and ROUGE-L) and the LLM-as-a-judge. When using GPT-4o as the backbone, SEAT attains the highest agreement with human judgments. This improvement over the direct rating approach (LLM-as-a-judge) remains significant across all tested backbones, including Qwen-2.5-72B and Llama-3.3-70B, demonstrating SEAT’s robustness.

### 5.3 Main Results

Table 4 reports how well each LLM handles six main tasks in TableEval: *Information Retrieval*, *Numerical Analysis*, *Reasoning*, *Data Analysis*, *Multi-turn Conversation*, and *Table Structure Under-*

Model	Size (B)	Avg	Information Retrieval	Numerical Analysis	Reasoning	Data Analysis	Multi-turn Conversation	Table Structure Understanding
Qwen2.5-7B-Instruct	7	59.60	69.23	64.29	59.38	69.71	68.67	26.35
Llama-3.1-8B-Instruct	8	49.26	67.40	53.35	48.82	57.06	53.15	15.76
glm-4-9b-chat	9	53.61	66.19	51.09	55.09	62.47	64.36	22.44
Qwen2.5-14B-Instruct	14	70.02	84.72	78.93	68.65	75.06	75.05	37.72
DeepSeek-V2-Lite-Chat	16	36.75	48.52	35.43	35.97	51.80	41.61	7.15
DeepSeek-Coder-V2-Lite-Instruct	16	48.30	60.40	56.39	50.03	51.51	50.62	20.83
QwQ-32B-Preview	32	78.14	89.33	85.75	81.37	71.69	82.15	58.53
Qwen2.5-32B-Instruct	32	75.50	86.32	84.10	76.09	77.60	82.25	46.61
Qwen2.5-Coder-32B-Instruct	32	70.75	79.82	77.00	73.03	76.33	74.89	43.44
Llama-3.3-70B-Instruct	70	72.94	87.42	76.70	73.38	<b>81.27</b>	80.62	38.24
Qwen2.5-72B-Instruct	72	74.23	82.68	81.53	74.85	78.94	81.90	45.50
DeepSeek-V2.5-1210	236	73.27	87.41	79.10	71.49	77.97	78.72	44.94
DeepSeek-V3	671	77.95	<b>91.20</b>	82.61	81.72	77.45	<b>85.83</b>	48.89
DeepSeek-R1	671	<b>82.46</b>	90.15	<b>88.56</b>	<b>87.91</b>	77.79	78.29	<b>72.05</b>
qwen-max-2024-09-19	N/A	73.34	84.42	81.35	72.64	78.09	80.18	43.35
Claude-3.5-sonnet-20241022	N/A	83.32	<b>89.62</b>	<b>91.06</b>	<b>85.76</b>	<b>84.01</b>	<b>87.94</b>	61.51
gpt-4o-mini-2024-07-18	N/A	68.47	82.64	76.15	73.13	70.70	73.66	34.56
gpt-4o-2024-11-20	N/A	78.79	88.24	86.00	83.05	81.47	83.20	50.79
o1-preview	N/A	<b>83.43</b>	88.30	87.08	82.88	77.89	83.38	<b>81.03</b>

Table 4: Performance of LLMs on TableEval. We report the F1 score calculated by our SEAT. Bold values indicate the best result within each category. Avg denotes the overall score.

standing, along with the average score. Larger models generally perform better, with closed-source models like Claude 3.5 Sonnet, GPT-4o, and o1-preview often leading the rankings. o1-preview achieves the highest average score (83.43%), while Claude 3.5 Sonnet performs comparably (83.32%). Both perform well in complex tasks such as reasoning and data analysis, which require deep reasoning and domain knowledge. Among open-source models, DeepSeek-R1 shows strong, complex reasoning abilities and performs best in reasoning, numerical analysis, and table structure understanding. Qwen2.5-14B-Instruct significantly improves over its 7B version, demonstrating the benefits of increased parameters. Scaling up to 32B further enhances overall performance, with models like QwQ-32B-Preview and Qwen2.5-32B-Instruct improving across most metrics. However, even these large models still struggle with table structure understanding, suggesting that specialized training is necessary for tasks requiring layout comprehension. The significant gap between top closed-source LLMs and mid-sized open-source ones underscores the advantages that closed-source models may have, potentially through specialized training strategies, proprietary datasets, or more computational resources.

**Structure-aspect Performance.** Figure 4 (See Table 12 for details) depicts the performance of different LLMs across different table structures, including *vertical*, *horizontal*, *matrix*, and more complex *concise*, *hierarchical*, and *nested* structures. Closed-source models generally demonstrate more balanced and superior performance. GPT-4o,

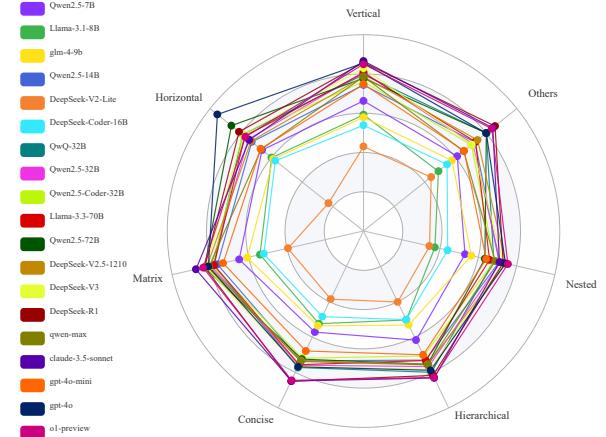


Figure 4: Performance of LLMs across table structures.

Claude, and o1-preview outperform across different table structures, demonstrating their strong capabilities. However, specific open-source models can be highly competitive in specific structures. For example, DeepSeek R1 performs particularly well on *hierarchical* and *others* table structures. Besides, all models, whether open or closed, face more significant challenges with *nested* and *hierarchical* tables compared to simpler formats, underscoring the ongoing difficulty in structure-aware comprehension and structured reasoning.

**Language-aspect Performance.** Table 5 depicts the performance of different LLMs across tables in Simplified Chinese, Traditional Chinese, and English. Closed-source models demonstrate strong multilingual capabilities, maintaining high and consistent accuracy across these languages. Claude 3.5 Sonnet and o1-preview consistently achieve the top performance. A notable trend is a model's develop-

Model	Size	ZH	ZH-HK	EN
Qwen2.5-7B-Instruct	7	61.98	61.62	63.67
Llama-3.1-8B-Instruct	8	53.24	52.48	54.66
glm-4-9b-chat	9	55.77	55.19	55.30
Qwen2.5-14B-Instruct	14	72.00	74.49	76.96
DeepSeek-V2-Lite-Chat	16	40.56	43.97	38.09
DeepSeek-Coder-V2-Lite-Instruct	16	51.32	51.66	50.66
QwQ-32B-Preview	32	78.55	76.40	81.62
Qwen2.5-32B-Instruct	32	76.99	78.43	78.84
Qwen2.5-Coder-32B-Instruct	32	73.03	70.12	75.70
Llama-3.3-70B-Instruct	70	75.52	77.13	76.79
Qwen2.5-72B-Instruct	72	76.16	74.51	77.68
DeepSeek-V2.5-1210	236	75.51	78.15	77.31
DeepSeek-V3	671	78.20	79.69	81.60
DeepSeek-R1	671	<b>83.46</b>	<b>82.27</b>	<b>84.86</b>
qwen-max-2024-09-19	N/A	74.94	78.36	76.88
Claude-3.5-sonnet-20241022	N/A	<b>84.50</b>	<b>83.84</b>	85.34
gpt-4o-mini-2024-07-18	N/A	69.64	73.80	72.17
gpt-4o-2024-11-20	N/A	80.14	79.73	82.75
o1-preview	N/A	82.69	82.85	<b>86.41</b>

Table 5: Language-aspect performance.

ment background’s influence on language strengths. Models trained in large Chinese corpora, such as the Qwen series and DeepSeek-R1, tend to perform best in English, followed by Simplified Chinese, with Traditional Chinese ranking lowest. In contrast, models primarily trained in English often achieve their highest accuracy in English or Traditional Chinese, while Simplified Chinese tends to perform worse.

**Domain-aspect Performance.** Table 6 shows the performance of different LLMs across tables in finance, government, paper, and report domains. Larger models consistently outperform smaller ones across all four domains. Smaller models show more significant performance variation, struggling in specific areas, whereas larger models remain more stable, indicating more substantial generalization across diverse tabular data. The results show that Finance and Government data are more challenging, as smaller models tend to have greater performance fluctuations. On the other hand, Paper and Report tasks are handled more steadily, suggesting structured data complexity varies by field.

**Key Insights and Observations.** (1) Closed-Source Models Lead, but Open-Source Models Show Promise. Closed-source models (e.g., Claude 3.5 Sonnet, GPT-4o, o1-preview) consistently achieve top-tier performance across tasks, table structures, languages, and domains, benefiting from specialized training resources. However, large-scale open-source models (e.g., DeepSeek-R1, QwQ-32B-Preview) can be highly competitive with test-time scaling, indicating that open models have the potential to reduce the gap on complex TableQA with continued improvements. (2) Table Structure Understanding Remains a Challenge. Both closed- and open-source models strug-

Model	Size	Finance	Government	Paper	Report
Qwen2.5-7B-Instruct	7	62.69	60.44	63.60	60.94
Llama-3.1-8B-Instruct	8	53.47	49.04	55.87	52.63
glm-4-9b-chat	9	55.49	54.42	54.94	57.10
Qwen2.5-14B-Instruct	14	73.11	73.32	74.90	73.55
DeepSeek-V2-Lite-Chat	16	43.02	34.84	38.71	40.57
DeepSeek-Coder-V2-Lite-Instruct	16	50.88	46.14	50.76	55.23
QwQ-32B-Preview	32	77.71	79.24	80.09	79.97
Qwen2.5-32B-Instruct	32	77.30	74.56	79.68	77.64
Qwen2.5-Coder-32B-Instruct	32	71.65	72.28	76.19	73.06
Llama-3.3-70B-Instruct	70	76.43	70.77	76.93	77.23
Qwen2.5-72B-Instruct	72	73.71	76.20	78.90	78.32
DeepSeek-V2.5-1210	236	77.31	74.85	75.30	76.86
DeepSeek-V3	671	79.03	79.73	79.45	79.57
DeepSeek-R1	671	81.30	90.13	84.86	83.48
qwen-max-2024-09-19	N/A	76.19	75.57	75.78	76.27
Claude-3.5-sonnet-20241022	N/A	82.75	86.18	86.61	85.08
gpt-4o-mini-2024-07-18	N/A	71.34	68.37	70.79	71.96
gpt-4o-2024-11-20	N/A	79.63	81.29	83.18	79.57
o1-preview	N/A	82.85	86.36	84.34	83.06

Table 6: Domain-aspect performance.

gle with complex table structures, particularly nested and hierarchical structures. Performance drops of 10–15% are observed compared to simple structures (flat tables), highlighting the importance of future research on explicit table structures or structure-aware encoding methods. (3) Domain and Language-Specific Performance Gaps. Open-source models like DeepSeek-R1 perform strongly in specific domains, such as government data, while closed-source models demonstrate strong cross-domain adaptability. Similarly, language-aspect evaluation shows that while models often perform well in English, their performance declines in Simplified or Traditional Chinese. These findings suggest that domain-specific or language-specific training strategies should be explored.

## 6 Conclusion

We introduce TableEval, a real-world benchmark that fills a critical gap in evaluating LLMs on TableQA by featuring various table structures, languages, and domains. By curating recently published tables, TableEval minimizes data leakage and provides a more accurate evaluation of LLMs. Experimental results show that while closed-source models generally perform better, well-scaled and fine-tuned open-source models remain competitive. However, both struggle with complex table structures like nested or hierarchical ones, leading to a 10–15% performance drop. Additionally, models still face challenges with domain-specific and cross-lingual understanding. Besides, our SEAT evaluation method assesses answers structurally, ensuring semantic accuracy where traditional exact-match and n-gram matching metrics fall short and achieving high consistency with human judgments. We believe TableEval and SEAT will advance research in structure-aware approaches, multilingual representation, and domain adaptation for TableQA.

## 7 Limitations

We conclude the limitations of our study as follows: (1) Despite efforts to source heterogeneous tables (including concise, hierarchical, and nested structures), real-world data includes many other layouts (such as highly irregular formats or image-based tables), that are not fully captured in our benchmark. (2) Our dataset primarily presents questions in Simplified Chinese, while we include table content in Simplified Chinese, Traditional Chinese, and English. Deeper investigations into fully multilingual TableQA scenarios are still needed.

## Acknowledgements

This work is supported by Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park No.Z231100007423016, the National Natural Science Foundation of China under Grants No.62206287. We sincerely thank the anonymous reviewers for their insightful comments and constructive suggestions.

## References

Jianlyu Chen, Shitao Xiao, Peitian Zhang, et al. 2024. **M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021a. **Open question answering over tables and text**. In *International Conference on Learning Representations (ICLR)*, pages 1–18.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. **Tabfact: A large-scale dataset for table-based fact verification**. In *International Conference on Learning Representations (ICLR)*, pages 1–25.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, et al. 2021b. **FinQA: A dataset of numerical reasoning over financial data**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3697–3711. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. **Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning**. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, et al. 2024. **Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model**. *Preprint*, arXiv:2405.04434.

Naihao Deng, Zhenjie Sun, Ruiqi He, et al. 2024. **Tables as texts or images: Evaluating the table reasoning ability of LLMs and MLLMs**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 407–426. Association for Computational Linguistics.

Julian Eisenschlos, Maharshi Gor, Thomas Müller, and William Cohen. 2021. **MATE: Multi-view attention for table transformer efficiency**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7606–7619. Association for Computational Linguistics.

Team GLM, :, Aohan Zeng, Bin Xu, et al. 2024. **Chatglm: A family of large language models from glm-130b to glm-4 all tools**. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. **The llama 3 herd of models**. *Preprint*, arXiv:2407.21783.

Jiaqi Guo, Ziliang Si, Yu Wang, Qian Liu, Ming Fan, Jian-Guang Lou, Zijiang Yang, and Ting Liu. 2021. **Chase: A large-scale and pragmatic Chinese dataset for cross-database context-dependent text-to-SQL**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 2316–2331. Association for Computational Linguistics.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, et al. 2020. **TaPas: Weakly supervised table parsing via pre-training**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4320–4333. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, et al. 2024. **Qwen2.5-coder technical report**. *Preprint*, arXiv:2409.12186.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. **Search-based neural structured learning for sequential question answering**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1821–1831. Association for Computational Linguistics.

Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. **OmniTab: Pretraining with natural and synthetic data for few-shot table-based question answering**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 932–942. Association for Computational Linguistics.

Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, et al. 2022. **AIT-QA: Question answering dataset over complex tables in the airline industry**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 305–314. Association for Computational Linguistics.

Vishwajeet Kumar, Yash Gupta, Saneem Chemmengath, Jaydeep Sen, Soumen Chakrabarti, Samarth Bharadwaj, and Feifei Pan. 2023. **Multi-row, multi-span distant supervision for Table+Text question answering**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8080–8094. Association for Computational Linguistics.

Larissa R. Lautert, Marcelo M. Scheidt, and Carina F. Dorneles. 2013. **Web table taxonomy and formalization**. *ACM SIGMOD Record*, 42(3):28–33.

Fangyu Lei, Xiang Li, Yifan Wei, Shizhu He, Yiming Huang, Jun Zhao, and Kang Liu. 2023. **S3HQA: A three-stage approach for multi-hop text-table hybrid question answering**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1731–1740. Association for Computational Linguistics.

Jinyang Li, Binyuan Hui, Jiaxi Yang, Ge Qu, and others. 2023. **Can LLM already serve as a database interface? a big bench for large-scale database grounded text-to-SQLs**. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, pages 1–28. Curran Associates Inc.

Peng Li, Yeye He, and others. 2024. **Table-gpt: Table fine-tuned gpt for diverse table tasks**. In *Proceedings of the ACM on Management of Data*, pages 1–28. Association for Computing Machinery.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. **TAPEX: Table pre-training via learning a neural SQL executor**. In *International Conference on Learning Representations (ICLR)*, pages 1–19.

Linyong Nan, Chiachun Hsieh, Ziming Mao, et al. 2022. **FeTaQA: Free-form table question answering**. *Transactions of the Association for Computational Linguistics*, 10:35–49.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. **Gpt-4 technical report**. *Preprint*, arXiv:2303.08774.

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

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

Qwen, :, An Yang, Baosong Yang, et al. 2025. **Qwen2.5 technical report**. *Preprint*, arXiv:2412.15115.

Tianze Shi, Chen Zhao, Jordan Boyd-Graber, et al. 2020. **On the potential of lexico-logical alignments for semantic parsing to SQL queries**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1849–1864. Association for Computational Linguistics.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. **Table meets Ilm: 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 (WSDM)*, pages 645–654. Association for Computing Machinery.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. **RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7567–7578. Association for Computational Linguistics.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Minculich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. **Chain-of-table: Evolving tables in the reasoning chain for table understanding**. In *The Twelfth International Conference on Learning Representations (ICLR)*, pages 1–24.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. **Chain of thought prompting elicits reasoning in large language models**. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–14.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. 2024. **Tablebench: A comprehensive and complex benchmark for table question answering**. *arXiv preprint arXiv:2408.09174*.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. **TaBERT: Pretraining for joint understanding of textual and tabular data**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8413–8426. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, et al. 2018. **Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 3911–3921. Association for Computational Linguistics.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024a. **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 (NAACL)*, pages 6024–6044. Association for Computational Linguistics.

Yunjia Zhang, Jordan Henkel, Avrilou Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M. Patel. 2024b. **Reactable: Enhancing react for table question answering**. In *Proceedings of the VLDB Endowment (VLDB)*, pages 1981–1994. VLDB Endowment.

Zhehao Zhang, Yan Gao, and Jian-Guang Lou. 2024c.  **$e^5$ : Zero-shot hierarchical table analysis using augmented LLMs via explain, extract, execute, exhibit and extrapolate**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1244–1258. Association for Computational Linguistics.

Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir Radev. 2022. **ReasTAP: Injecting table reasoning skills during pre-training via synthetic reasoning examples**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9018. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023a. **Judging LLM-as-a-judge with MT-bench and chatbot arena**. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, pages 1–29.

Mingyu Zheng, Xinwei Feng, Qingyi Si, et al. 2024. **Multimodal table understanding**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9102–9124. Association for Computational Linguistics.

Mingyu Zheng, Yang Hao, Wenbin Jiang, et al. 2023b. **IM-TQA: A Chinese table question answering dataset with implicit and multi-type table structures**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5074–5094. Association for Computational Linguistics.

Wanjun Zhong, Junjie Huang, Qian Liu, et al. 2022. **Reasoning over hybrid chain for table-and-text open domain question answering**. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4531–4537.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, et al. 2021. **TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 3277–3287. Association for Computational Linguistics.

## A Dataset Construction Details

Task Type	Question Category	Count
Information Retrieval	Simple Lookup	369
	Conditional Lookup	239
	Grouped Lookup	29
Numerical Analysis	Sorting	223
	Statistics	103
	Numerical Calculation	220
Reasoning	Multi-hop Question	138
	Causal Analysis	73
	Comparative Analysis	139
	Trend Analysis	54
	Anomaly Detection	45
	Correlation Judgement	78
	Rejection	298
	Ellipsis or Reference	31
	Total	2,325
Table Structure Understanding	Table Size Detection	153
	Merged Cell Detection	133

Table 7: Task Types and Question Categories

Language	#Table	#QA Pairs
ZH-CN	364	1,331
ZH-HK	114	406
EN	139	588
Total	617	2,325

Table 8: TableEval Language-wise Statistics.

### A.1 Multi-answer Consistency Checking

To enhance the quality of data annotation, we leverage GPT-4o, DeepSeek 2.5, and Qwen 2.5-72B-Instruct to generate annotated answers for all questions. A consistency check is then performed on these annotations, with the results serving as guidance for subsequent human check. To this end, we propose an LLM-based consistency-checking method that rigorously evaluates the alignment of answer content (prompt can be found in Figure 7). This method assesses semantic matching, expression style, core information, and the complete consistency of table content. Any semantic conflicts, redundant details, or incomplete matches identified in the answers are considered inconsistencies during the evaluation process.

Task	$r$	$\rho$	$\tau$
Information Retrieval	0.9687	0.9141	0.9046
Numerical Analysis	0.8509	0.8330	0.8163
Reasoning	1.0000	1.0000	1.0000
Data Analysis	0.8972	0.8871	0.8361
Multi-turn Conversation	1.0000	1.0000	1.0000
Table Structure Understanding	0.9957	0.9957	0.9951

Table 9: Task-specific correlation of SEAT with human scores, measured with Pearson  $r$ , Spearman  $\rho$ , and Kendall  $\tau$  coefficients, using GPT-4o as backbone.

**Semantic Analysis.** The core information in all three answers is compared individually to ensure there are no semantic differences in their expressions. Variations in phrasing, wording, or inconsistencies in described details will make the answers considered inconsistent.

**Redundancy Check.** If any answer contains unnecessary explanations or repeated content that unnecessarily increases the length, it will also be considered inconsistent.

## A.2 Structured Answer Extraction

We use GPT-4o to extract structured answers from LLM-generated responses. This method aims to precisely extract the content corresponding to the question from the answer text, which is then used for subsequent model evaluation. It follows the following rules to ensure the accuracy of the extracted answers in terms of values, format, and content. We show the prompt in Figure 10.

**Multi-part Question Processing.** The original question should be decomposed into a list of sub-questions when it contains multiple sub-questions.

**Single Question Processing.** For a single question, keep the answer to that question without unnecessary decomposition, ensuring a concise and clear response.

**Answer Extraction.** Our goal is to extract concise final answers for each sub-question to ensure an accurate evaluation of all models’ natural language responses. Additionally, we have two specific requirements: (a) During the answer extraction process, the original format of numerical values, including percentage signs, units, commas, and scientific notation, must be strictly kept, ensuring the precision of the extracted answers. (b) If the answer text contains ambiguity, errors, or incomplete information, the answer will be labeled as “Reference Answer Incorrect”, helping the user understand the potential uncertainty of the answer.

Domain	$r$	$\rho$	$\tau$
financial reports and company announcements	0.9575	0.9295	0.8967
industry/stock research reports	0.9605	0.9642	0.9516
academic papers	0.8794	0.8777	0.8470
governmental data	0.9973	0.9982	0.9962

Table 10: Domain-specific correlation of SEAT with human scores, measured with Pearson  $r$ , Spearman  $\rho$ , and Kendall  $\tau$  coefficients, using GPT-4o as backbone.

## B Additional Experimental Analysis

### B.1 Robustness of SEAT

To provide a more detailed analysis of SEAT’s effectiveness across different contexts, we further categorize our main experimental results (Table 3) by task type and domain. This analysis aims to determine whether SEAT maintains consistent correlation with human judgments regardless of the specific nature of the TableQA task or the domain from which the tables originate. Rather than conducting separate experiments, we classify our existing evaluation data according to six distinct task types and four domains to examine SEAT’s robustness across these dimensions.

**Task-Specific Robustness.** Table 9 presents the correlation between SEAT and human judgments across six task types in TableEval, using GPT-4o as the evaluation backbone. The results demonstrate SEAT’s high consistency across varied task complexities. For reasoning and multi-turn conversation tasks, SEAT achieves perfect correlation with human judgments, indicating that the structured evaluation approach is particularly effective for complex, multi-step reasoning processes. Information retrieval and table structure understanding tasks also show very high correlations (Pearson values of 0.9687 and 0.9957 respectively), suggesting SEAT accurately captures the quality of responses for these fundamental TableQA capabilities.

Notably, while numerical analysis and data analysis tasks show slightly lower correlation coefficients (Pearson values of 0.8509 and 0.8972 respectively), these values remain substantially higher than those achieved by traditional metrics as shown in our main results. This slight decrease in correlation for quantitative tasks may reflect the inherent complexity of evaluating numerical reasoning, where small computational errors can lead to significant discrepancies in final answers. Nevertheless, the consistently high correlation values across all task types demonstrate SEAT’s robustness as an evaluation framework regardless of the specific

Model	Size	Avg	Information Retrieval			Numerical Analysis			Reasoning			Data Analysis						Conversation		Table Structure Understanding	
			Simple	Condition	Grouped	Sorting	Statistics	Calculation	Multi-hop	Causal	Comparative	Trend	Anomaly	Correlation	Judgement	Rejection	Ellipsis/ Reference	Table Size Detection	Merged Cell Detection		
Qwen2.5-7B-Instruct	7	59.94	69.79	68.88	65.03	74.57	57.28	57.14	59.38	54.25	74.03	73.93	44.84	65.14	75.66	68.67	41.18	9.30			
Llama-3.1-8B-Instruct	8	51.76	70.31	62.58	70.18	53.55	50.49	54.49	48.82	57.21	61.85	63.91	39.17	55.24	56.72	53.15	21.57	8.86			
glm-4-9b-chat	9	53.83	68.78	63.25	57.43	59.84	47.96	43.69	55.09	45.23	57.51	72.70	51.63	62.45	68.79	64.36	34.64	7.86			
Qwen2.5-14B-Instruct	14	70.54	85.20	84.40	81.39	85.98	68.94	76.46	68.65	70.04	80.19	72.24	60.24	69.08	78.06	75.05	56.54	15.57			
DeepSeek-V2-Lite-Chat	16	39.45	55.57	52.93	38.84	27.08	34.02	39.29	62.09	63.99	39.29	55.54	41.61	9.48	42						
DeepSeek-Coder-V2-Lite-Instruct	16	49.40	66.51	50.46	64.25	51.74	55.02	61.75	50.03	49.15	60.18	58.85	37.51	42.53	51.17	50.62	27.45	13.20			
QwQ-32B-Preview	32	77.17	88.07	91.22	89.75	<b>92.35</b>	79.94	81.76	81.37	67.32	<b>87.95</b>	75.34	<b>37.12</b>	64.64	66.15	82.15	77.12	36.47			
Qwen2.5-32B-Instruct	32	75.18	86.41	86.84	80.90	79.43	79.82	76.09	69.84	79.73	77.70	66.44	<b>76.73</b>	80.42	82.25	67.32	22.43				
Qwen2.5-Coder-32B-Instruct	32	71.34	78.72	80.48	88.62	86.20	69.90	71.01	73.03	60.78	79.25	76.19	67.17	69.29	82.03	74.89	62.75	21.06			
Llama-3.0-70B-Instruct	70	73.67	89.39	84.86	83.39	80.08	75.90	73.65	73.38	<b>73.97</b>	81.91	84.02	64.33	73.36	86.90	80.62	60.13	12.87			
Qwen2.5-72B-Instruct	72	74.80	80.31	86.35	82.70	88.63	67.38	80.97	74.85	68.95	85.21	85.83	73.10	79.76	81.90	60.78	27.92				
DeepSeek-V2.5-1210	236	73.60	91.32	81.08	89.81	77.91	73.71	82.83	71.49	63.67	80.86	<b>86.81</b>	62.88	66.00	83.94	78.72	63.73	22.84			
DeepSeek-V3	671	77.03	<b>92.84</b>	88.24	90.74	82.49	78.23	84.73	81.72	65.00	86.63	85.75	67.67	69.91	78.24	<b>85.83</b>	67.97	26.44			
DeepSeek-R1	671	<b>80.14</b>	87.26	<b>93.85</b>	90.49	<b>96.37</b>	<b>87.59</b>	85.69	77.44	77.50	59.77	65.41	<b>87.39</b>	78.29	<b>86.60</b>		<b>55.30</b>				
qwen-max-2024-09-19	N/A	73.30	84.72	84.76	77.60	84.94	75.41	80.50	72.64	69.96	82.17	74.66	65.88	78.36	80.58	80.18	63.40	11.12			
Claude-3.5-sonnet-20241022	N/A	81.49	89.83	<b>90.33</b>	<b>92.35</b>	<b>95.79</b>	<b>87.46</b>	<b>87.76</b>	69.55	87.56	75.94	<b>87.94</b>	74.81	82.45	<b>87.94</b>	78.40	42.43				
gpt-4o-mini-2024-07-18	N/A	68.69	84.82	79.99	76.62	81.13	78.83	69.85	73.13	81.16	80.45	56.71	65.89	86.50	73.66	47.39	19.70				
gpt-4o-2024-11-20	N/A	78.21	<b>89.59</b>	86.26	87.42	88.10	86.74	83.53	83.05	<b>72.20</b>	<b>88.26</b>	75.07	67.32	75.96	85.31	83.20	65.34	34.06			
o1-preview	N/A	<b>82.03</b>	87.98	88.52	90.57	92.93	84.80	82.21	82.88	71.45	85.64	<b>85.79</b>	60.46	<b>78.40</b>	76.93	83.38	<b>90.52</b>	<b>70.03</b>			

Table 11: Performance of LLMs on 16 TableQA sub-tasks.

Model	Size	Vertical	Horizontal	Matrix	Concise	Hierarchical	Nested	Others
Qwen2.5-7B-Instruct	7	<b>66.26</b>	66.10	64.97	57.12	61.65	53.04	61.07
Llama-3.1-8B-Instruct	8	<b>59.24</b>	59.95	54.09	52.34	50.23	37.30	48.83
glm-4-9b-chat	9	<b>57.95</b>	58.86	60.76	53.53	53.03	56.36	57.64
Qwen2.5-14B-Instruct	14	74.38	72.86	76.36	73.34	72.93	70.91	72.70
DeepSeek-V2-Lite-Chat	16	43.05	22.81	39.48	38.49	40.08	34.46	44.10
DeepSeek-Coder-V2-Lite-Instruct	16	53.90	57.48	52.11	48.35	50.08	43.95	54.43
QwQ-32B-Preview	32	79.67	78.86	78.35	77.11	79.81	68.05	80.07
Qwen2.5-32B-Instruct	32	80.99	75.05	79.60	75.81	76.59	71.53	71.26
Qwen2.5-Coder-32B-Instruct	32	76.10	78.19	79.10	71.78	70.55	68.47	70.33
Llama-3.3-70B-Instruct	70	80.54	77.43	78.15	75.53	73.11	65.83	72.94
Qwen2.5-72B-Instruct	72	77.77	<b>86.10</b>	<b>83.13</b>	72.42	75.27	63.43	80.03
DeepSeek-V2.5-1210	236	78.69	73.43	81.94	73.21	75.31	<b>74.28</b>	74.34
DeepSeek-V3	671	82.96	78.76	82.26	76.68	78.24	73.13	70.57
DeepSeek-R1	671	<b>85.83</b>	81.10	83.00	<b>84.86</b>	<b>81.59</b>	72.96	<b>85.69</b>
qwen-max-2024-09-19	N/A	78.68	66.76	80.74	73.75	75.66	68.86	65.29
Claude-3.5-sonnet-20241022	N/A	<b>86.61</b>	74.43	<b>87.59</b>	<b>84.77</b>	<b>83.03</b>	71.05	<b>84.27</b>
gpt-4o-mini-2024-07-18	N/A	74.82	67.10	73.38	67.80	69.97	64.11	65.55
gpt-4o-2024-11-20	N/A	85.59	<b>95.24</b>	81.13	76.76	78.70	74.17	79.97
o1-preview	N/A	84.96	77.00	83.81	84.40	82.71	<b>75.46</b>	83.60

Table 12: Structure-aspect performance.

TableQA function being assessed.

**Domain-Specific Robustness.** Table 10 extends our analysis to examine SEAT’s performance across four domains from which tabular data originate. The results reveal SEAT’s consistent effectiveness across various subject matters. Governmental data shows the highest correlation, likely due to the structured and standardized nature of such information. Both financial reports/company announcements and industry/stock research reports domains also demonstrate very high correlations (Pearson values of 0.9575 and 0.9605 respectively), indicating SEAT effectively evaluates responses in domains with specific terminology and numerical components.

Academic papers represent the most challenging domain, though the correlation values remain strong. This slight decrease may be attributed to the complexity and diversity of academic content, which often contains specialized vocabulary, complex methodology descriptions, and multi-faceted findings that can be challenging to evaluate in a structured format. Despite this relative difference, the consistently high correlation across all domains

confirms that SEAT maintains its effectiveness regardless of the subject matter or data source.

These comprehensive experiments across both task types and domains provide strong evidence for SEAT’s robustness as an evaluation framework.

## B.2 Performance on Sub-tasks

Table 11 provides a more granular evaluation of each model’s performance across 16 sub-tasks, offering more profound insight into LLMs’ specific strengths and weaknesses. These sub-tasks include broader categories of *information retrieval*, *numerical analysis*, *reasoning*, *data analysis*, *multi-turn conversation*, and *table structure understanding*, each targeting a unique aspect of TableQA.

It can be found in the *information retrieval* sub-tasks (simple, condition, and grouped lookups) that larger or more specialized models (particularly DeepSeek-R1, Claude 3.5 sonnet, GPT-4o, and QwQ-32B-Preview) consistently achieve the highest performance. QwQ-32B-Preview and DeepSeek-R1 better handle more complex filtering (condition, grouped), suggesting that the enhanced reasoning capabilities help manage varied

Model	Avg	Information Retrieval	Numerical Analysis	Reasoning	Data Analysis	Multi-turn Conversation	Table Structure Understanding
<b>HTML</b>							
gpt-4o-2024-11-20	81.09	89.98	87.26	80.92	79.80	75.65	72.95
Qwen2.5-14B-Instruct	72.51	80.93	79.85	73.21	73.17	75.37	52.56
Llama-3.1-8B-Instruct	53.19	70.28	53.32	50.52	56.86	56.87	31.32
<b>Latex</b>							
gpt-4o-2024-11-20	81.04	88.00	86.19	83.04	80.66	82.95	65.38
Qwen2.5-14B-Instruct	72.67	82.66	81.10	72.56	73.19	76.83	49.69
Llama-3.1-8B-Instruct	54.16	69.51	56.90	52.77	56.85	59.10	29.84
<b>Markdown</b>							
gpt-4o-2024-11-20	78.79	88.24	86.00	83.05	81.47	83.20	50.79
Qwen2.5-14B-Instruct	70.02	84.72	78.93	68.65	75.06	75.05	37.72
Llama-3.1-8B-Instruct	49.26	67.40	53.35	48.82	57.06	53.15	15.76

Table 13: Performance of LLMs on TableEval with different table formats.

data partitions. A similar pattern is shown in *numerical analysis* (sorting, statistics, calculation), with Claude 3.5 sonnet exceeding 95% on sorting and DeepSeek-R1 exceeding 86% on statistics and calculation. This performance indicates that large model sizes and specialized data or method-specific fine-tuning can improve arithmetic-driven tasks.

For *reasoning*, we focus on multi-hop questions requiring chained inference across table cells. DeepSeek-R1 achieves a top score of 87.91%, followed by closed-source models such as Claude 3.5 sonnet and GPT-4o, highlighting the benefits of size and specialized training. Smaller open-source models (e.g., glm-4-9b-chat, DeepSeek-V2-Lite-Chat) fall behind, reflecting the complexity of multi-hop inference towards small-scale models.

*Data analysis* includes six sub-tasks (causal analysis, comparative analysis, trend analysis, anomaly detection, correlation judgment, and rejection). These tasks require higher-level inference and domain knowledge, and the results reveal significant performance spread across models. For example, Llama-3.3-70B-Instruct achieves high scores in causal and trend analysis, while QwQ-32B-Preview achieves one of the highest scores on anomaly detection. Conversely, o1-preview demonstrates strong correlation judgment (78.40%), and Claude 3.5 sonnet outperforms all other models on rejection (92.05%), reflecting its robustness when queries demand domain knowledge or interpretation of numeric relationships. The difference in these data analysis tasks suggests that even leading LLMs specialize differently, implying opportunities for method-level customizations to boost performance on specific analytical tasks.

In the *ellipsis/reference* sub-task, Claude 3.5 sonnet achieves the best (87.94%), followed closely

by DeepSeek-V3. Large-scale open-source models (e.g., Llama-3.3-70B-Instruct, Qwen2.5-72B-Instruct) also obtain high scores, yet they still fall behind top closed-source systems, demonstrating the benefit of extensive multi-turn dialogue training data in proprietary setups.

Finally, *table structure understanding* is assessed via table size detection and merged cell detection, which reflect the ability to parse and interpret structural details. While some models perform well on table size detection (led by o1-preview at 90.52%), performance on merged cell detection drops significantly; no model exceeds 70.03%, also achieved by o1-preview. This large gap highlights the challenge of complex structural tasks, where even advanced LLMs struggle to accurately capture and interpret non-trivial table structures (layouts) or spanning cell configurations.

Overall, the fine-grained results demonstrate the multi-dimensional capability of LLMs in tabular data. Although large models dominate most categories, tasks such as merged cell detection or advanced data analysis remain challenging. These observations highlight the importance of structure-aware designs, targeted fine-tuning for analytical tasks, and more elaborate training on real-world tables (particularly those with nested or merged cells). By bridging these gaps, LLMs will be better equipped to handle the broad tasks of TableQA.

### B.3 Impact of Table Format

We conduct an experiment to explore the impact of table format on LLM performance by converting tables to LaTeX, HTML, and Markdown formats. We report the main results and the fine-grained results of 3 LLMs on Table 13 and Table 14.

We find that all three models achieve approx-

Model	Avg	Information Retrieval			Numerical Analysis			Reasoning			Data Analysis						Conversation		Table Structure Understanding	
		Simple	Condition	Grouped	Sorting	Statistics	Calculation	Multi-hop	Causal Analysis	Comparative Analysis	Trend Analysis	Anomaly Detection	Correlation Judgement	Rejection	Ellipsis/ Reference	Table Size Detection	Merged Cell Detection			
<b>HTML</b>																				
gpt-4o-2024-11-20	80.29	90.10	89.20	94.94	91.86	89.32	81.63	80.92	63.84	81.52	76.94	63.36	70.96	88.23	75.65	71.35	74.78			
Qwen2.5-14B-Instruct	71.87	79.60	82.73	83.07	85.37	72.84	77.44	73.21	62.05	76.07	74.94	65.09	60.97	78.62	75.37	69.61	32.94			
Llama-3.1-8B-Instruct	54.98	71.30	68.66	70.75	54.83	48.72	53.93	50.52	63.16	61.85	71.44	39.67	53.05	53.94	56.87	41.57	19.34			
<b>Latex</b>																				
gpt-4o-2024-11-20	79.51	87.66	88.54	87.82	92.76	84.80	80.15	83.04	70.04	80.69	79.03	63.94	72.51	88.17	82.95	70.05	59.93			
Qwen2.5-14B-Instruct	72.42	82.88	82.03	85.19	85.17	76.77	78.93	72.56	58.25	79.61	80.36	64.61	61.67	76.86	76.83	64.38	32.67			
Llama-3.1-8B-Instruct	54.98	71.18	67.32	66.46	59.40	54.55	55.48	52.77	55.88	62.21	69.56	42.63	49.84	56.28	59.10	46.73	10.27			
<b>Markdown</b>																				
gpt-4o-2024-11-20	78.21	89.59	86.26	87.42	88.10	86.74	83.53	83.05	72.20	88.26	75.07	67.32	75.96	85.31	83.20	65.34	34.06			
Qwen2.5-14B-Instruct	70.54	85.20	84.40	81.39	85.98	68.94	76.46	68.65	70.04	80.19	72.24	60.24	69.68	78.06	75.05	56.54	15.57			
Llama-3.1-8B-Instruct	51.76	70.31	62.58	70.18	53.55	50.49	54.49	48.82	57.21	61.85	63.91	39.17	55.24	56.72	53.15	21.57	8.86			

Table 14: Performance of LLMs on 16 sub-tasks with different table formats.

Model	Avg	Information Retrieval	Numerical Analysis	Reasoning	Data Analysis	Multi-turn Conversation	Table Structure Understanding
gpt-4o-mini-2024-07-18	68.47	82.64	76.15	73.13	70.70	73.66	34.56
Qwen2.5-7B-Instruct (fine-tuned)	66.19	80.06	61.71	58.73	70.56	70.92	55.16
glm-4-9b-chat (fine-tuned)	57.62	76.32	48.57	45.28	66.67	62.60	46.30
Llama-3.1-8B-Instruct (fine-tuned)	63.83	79.88	60.09	53.31	68.84	71.71	49.18
Qwen2.5-7B-Instruct	59.60	69.23	64.29	59.38	69.71	68.67	26.35
glm-4-9b-chat	53.61	66.19	51.09	55.09	62.47	64.36	22.44
Llama-3.1-8B-Instruct	49.26	67.40	53.35	48.82	57.06	53.15	15.76

Table 15: Performance of fine-tuned LLMs on TableEval.

imately 2 percentage points higher F1 scores on HTML and Latex formats than Markdown. We further look into Table 14: the main improvement comes from the table structure understanding task, where all models significantly improve F1 scores. Specifically, GPT-4o shows a 22.16% and 14.59% increase in F1 scores on HTML and LaTeX tables, respectively, compared to Markdown. The improvement is even more significant for Llama-3.1-8B-Instruct, with nearly double the performance gain.

In sub-task evaluations, all models exhibit around a 2× improvement in merged cell detection on HTML tables. Additionally, in table size detection, all models show varying degrees of progress. These results indicate that structured table formats (HTML and LaTeX) significantly enhance the models' ability to understand table structures, leading to improved TableQA performance.

#### B.4 Impact of Fine-tuning

We initially obtained 34,161 distinct questions, from which 2,325 were selected as the final dataset (§3.2). Based on the remaining questions, we applied the QA acquisition process described in §3.2 to generate QA pairs. These pairs undergo a consistency check but do not undergo a human review. Finally, we select 16,844 QA pairs for LLM fine-tuning. The results are presented in Table 15.

Our experimental results demonstrate that fine-tuning significantly enhances the TableQA capabilities of LLMs. Even small-scale LLMs can achieve comparable performance with GPT-4o-mini after fine-tuning. Specifically, the fine-tuned Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct

achieved notable improvements, particularly in table structure understanding, where they even outperform GPT-4o-mini. These results suggest that fine-tuning with task-specific data can help models develop a more structured understanding of tables.

Despite these advancements, challenges remain, particularly in numerical reasoning and complex analytical tasks, where even the fine-tuned models still fall behind. While information retrieval and data analysis have improved, the ability to perform deep numerical computations and multi-step reasoning over structured tables continues to be a key limitation. These findings indicate that further research is needed to enhance models' structured reasoning abilities, potentially through structure-aware training, hybrid retrieval-based approaches, or integration with external knowledge sources.

Overall, our study highlights the importance of fine-tuning for improving TableQA performance, demonstrating that with sufficient task-specific adaptation, small-scale open-source LLMs can achieve competitive results against stronger proprietary baselines. In the future work, we can focus on refining structured comprehension and enhancing numerical reasoning.

#### C Prompts

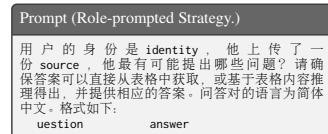


Figure 5: Prompt in role-prompted strategy.

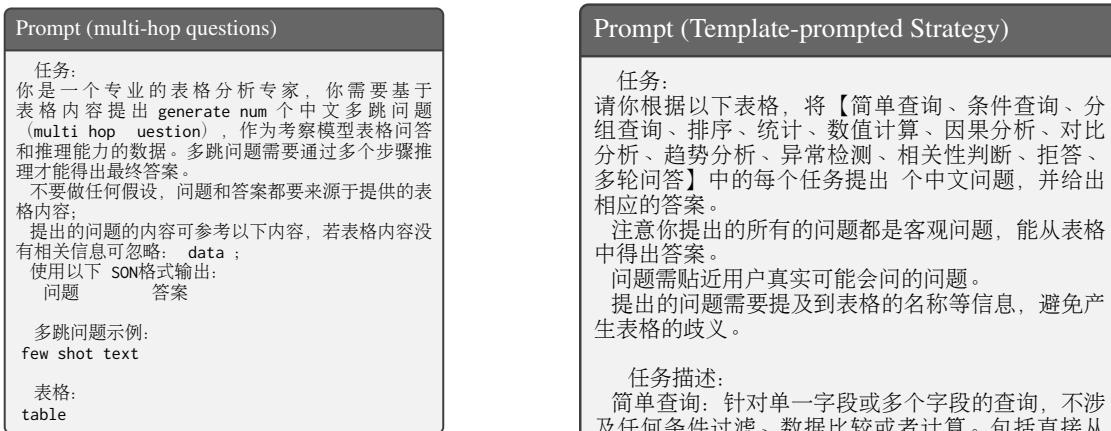


Figure 6: Prompt for multi-hop questions in template-prompted strategy.

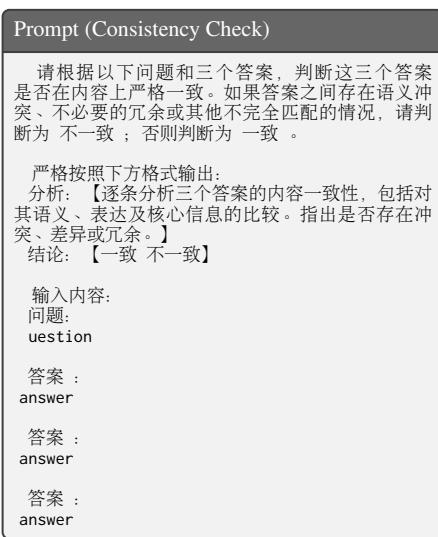


Figure 7: Consistency Check prompt.

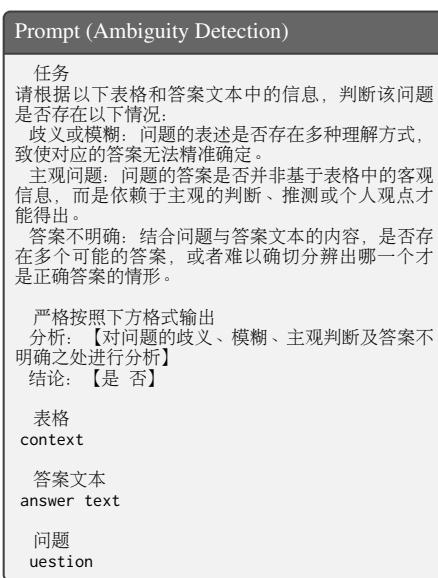


Figure 8: Ambiguity Detection prompt.

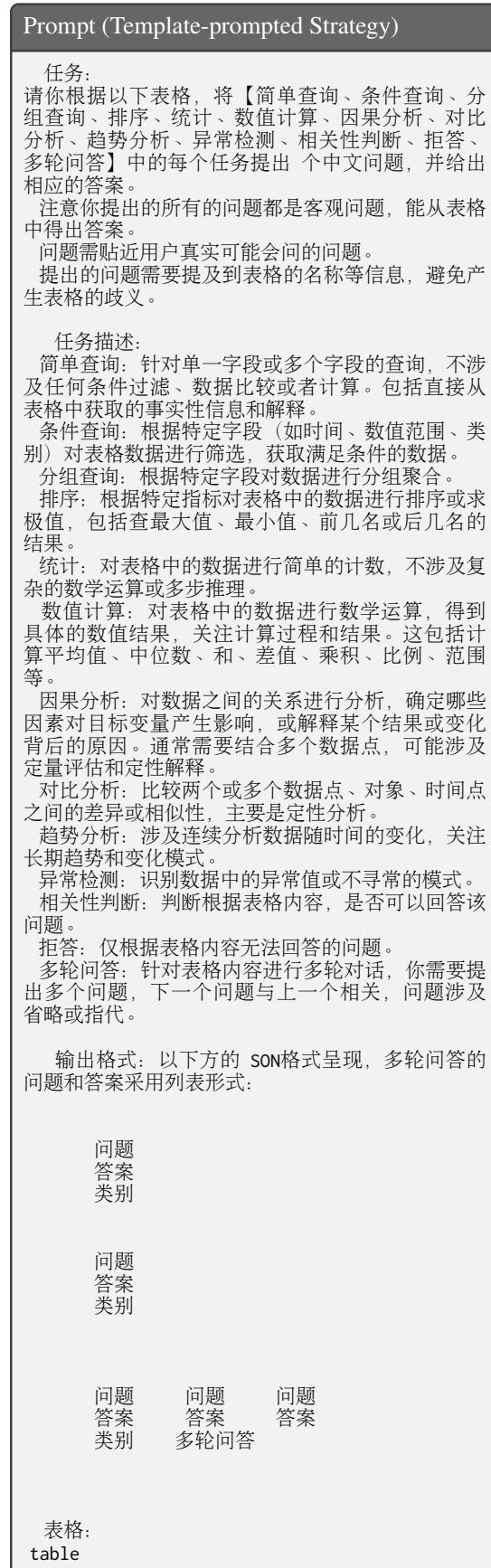


Figure 9: QA generation prompt in template-prompted strategy.

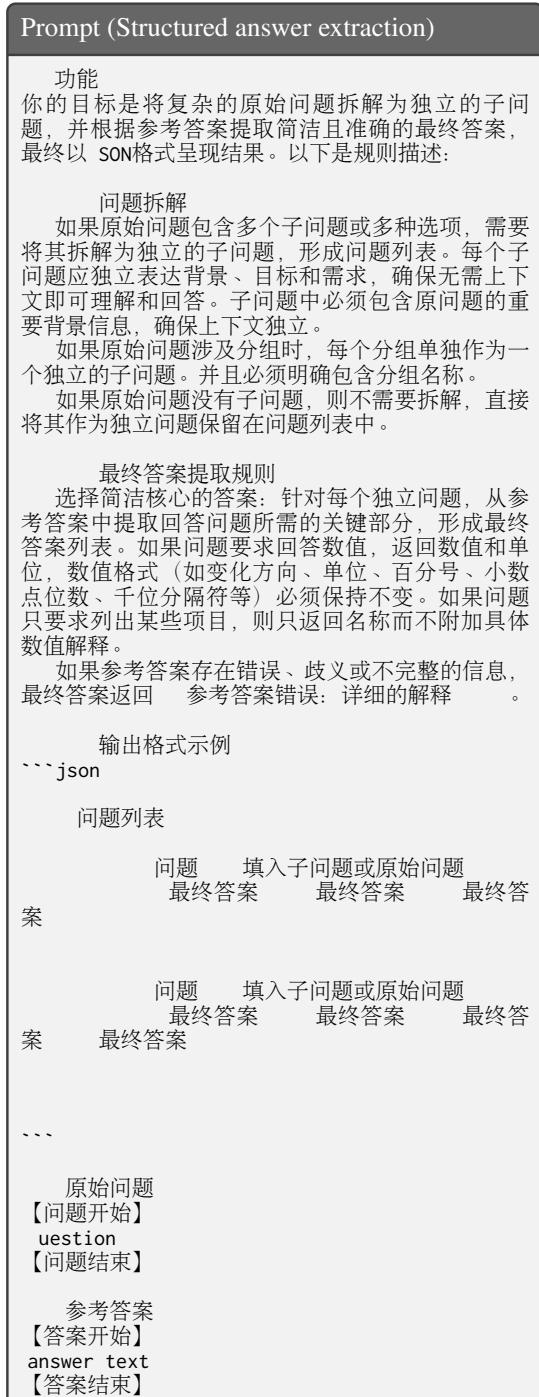


Figure 10: Structured answer extraction prompt.

Prompt (QA type classification)		
仔细阅读表格、问题和答案，将问题分类为以下类别之一。如果你不确定问题的类别，请输出 不确定。		
<b>类别：</b>		
简单查询：针对单一字段或多个字段的查询，不涉及任何条件过滤。包括直接从单元格中获取的事实性信息和解释。示例：哈一药业股份有限公司的专利有哪些？示例：稳健 号多少钱起投？收益率是多少？示例：为什么水通道蛋白膜的过滤性能能够持久如一？（水通道蛋白膜特点 中有介绍）		
条件查询：根据特定字段（如时间、数值范围、类别）对表格数据进行筛选，获取满足条件的数据。示例：光云科技 年的财务费用是多少元？示例：列出所有销售额超过 万元且属于电子行业的公司及其法定代表人。		
分组查询：根据特定字段对数据进行分组聚合。示例： 年的已授权的和未授权的专利分别有哪些？示例：请将下列公司按净利润在 万以内、 万、 万以上进行划分。		
排序：根据特定指标对表格中的数据进行排序，包括查最大值、最小值、前几名或后几名的结果。示例： 年总资产最高的前三家上市公司是哪些家？示例：贵阳银行在 年和 年的各个业务部门按收入贡献排序如何？		
统计：对表格中的数据进行简单的计数，不涉及复杂的数学运算或多步推理。示例：公司博士学历的人数有多少？示例：百度公司现有多少名独立董事？		
数值计算：对表格中的数据进行数学运算，得到具体的数值结果，关注计算过程和结果。这包括计算平均值、中位数、和、差值、乘积、比例、范围等。示例：营业收入占收入的比例是多少？（计算百分比）示例：百亚股份本期公允价值变动收益较上一报告期变化是多少？（计算差值）		
多跳问题：将复杂问题分解为多个相关联的子问题，逐步解决每个子问题，最终得到完整的推理过程。每个子问题可能涉及不同的操作，如查询、计算、排序等。关键在于需要通过一系列推理步骤，每一步的结果为下一步提供必要的信息或基础。示例：表格中基金经理人最多管理了几只基金？（首先统计，然后排序）示例：万方药业公司在本次上市前的股东结构中，持股比例最高的股东是谁？持股比例是多少？（需要先排序，然后查询）		
因果分析：对数据之间的关系进行分析，确定哪些因素对目标变量产生影响，或解释某个结果或变化背后的原因。通常需要结合多个数据点，可能涉及定量评估和定性解释。示例：公司的净利润同比增长的主要驱动因素是什么？（寻找增长的内部原因）示例：哪些费用项目的变化对净利润影响最大？（评估影响程度并解释原因）		
对比分析：比较两个或多个数据点、对象、时间点之间的差异或相似性，主要是定性分析。示例：上周债基指数成分内部离差比本周有何变化？示例：永顺生物 年比 年总的营收多了还是少了？		
趋势分析：涉及连续分析数据随时间的变化，关注长期趋势和变化模式。示例：公司的毛利率趋势如何，是否有改善？示例：股基指数本周和上周的内部离差变化趋势是什么？		
异常检测：识别数据中的异常值或不寻常的模式。例如：哪些项目的变动幅度超过了历史平均水平？		
归纳总结：提取关键信息归纳出最终的结论，表格中的信息无法直接回答问题。		

Figure 11: QA type classification prompt used in role-prompted strategy.

Prompt (Evaluation)																																	
<p><b>目标</b> 请将大模型的回答与用户提供的参考答案进行对比，步骤如下：</p>																																	
<p><b>提取关键答案</b> 定位大模型回答的 最终总结，逐个对照参考答案中的子问题，从大模型的 最终总结 中提取每个问题对应所有关键回答。关键回答应仅包含核心的、直接回答问题的内容。</p>																																	
<p>对已识别出的某个关键回答进行补充说明的内容，应与该关键回答合并为一个整体，不要拆分成新的答案要素。只有在内容明显独立、可与参考答案中不同要素相对应时，才视为新答案。</p>																																	
<p><b>对比并标注：</b>将提取出的回答与参考答案逐一对比，按以下标准进行标注：</p>																																	
<p><b>错误答案 (false) :</b> 如果大模型多输出了一些要素，并且这些要素与参考答案无法对应或仅是多余的补充信息（不是在同一个要素中补充，而是产生了多余答案要素），则判定为错误。</p>																																	
<p><b>正确答案 (true) :</b> 如果该条回答与参考答案某一要素含义一致或高度吻合，视为正确。</p>																																	
<p>注意：每个从大模型回答中提取出的答案要素，都要有相应的 true 或 false 标签，确保每个回答要素都被检查。</p>																																	
<p><b>输出格式</b></p>																																	
<p>```</p>																																	
<p><b>问题列表</b></p>																																	
<table border="1"> <thead> <tr> <th>问题</th><th>子问题</th><th>参考答案</th><th>答案</th><th>答案</th><th>关键回答</th><th>关键回答</th></tr> </thead> <tbody> <tr> <td>大模型的回答</td><td></td><td>大模型的回答</td><td></td><td>关键回答</td><td></td><td>关键回答</td></tr> <tr> <td>是否正确</td><td></td><td>true</td><td>false</td><td></td><td></td><td></td></tr> </tbody> </table>						问题	子问题	参考答案	答案	答案	关键回答	关键回答	大模型的回答		大模型的回答		关键回答		关键回答	是否正确		true	false										
问题	子问题	参考答案	答案	答案	关键回答	关键回答																											
大模型的回答		大模型的回答		关键回答		关键回答																											
是否正确		true	false																														
<table border="1"> <thead> <tr> <th>问题</th><th>子问题</th><th>参考答案</th><th>答案</th><th>答案</th><th>关键回答</th><th>关键回答</th></tr> </thead> <tbody> <tr> <td>大模型的回答</td><td></td><td>大模型的回答</td><td></td><td>关键回答</td><td></td><td>关键回答</td></tr> <tr> <td>关键回答</td><td></td><td></td><td>关键回答</td><td></td><td></td><td></td></tr> <tr> <td>是否正确</td><td></td><td>false</td><td>true</td><td>false</td><td>false</td><td></td></tr> </tbody> </table>						问题	子问题	参考答案	答案	答案	关键回答	关键回答	大模型的回答		大模型的回答		关键回答		关键回答	关键回答			关键回答				是否正确		false	true	false	false	
问题	子问题	参考答案	答案	答案	关键回答	关键回答																											
大模型的回答		大模型的回答		关键回答		关键回答																											
关键回答			关键回答																														
是否正确		false	true	false	false																												
<p>```</p>																																	
<p><b>参考答案</b></p>																																	
<p>reference answer</p>																																	
<p>大模型的回答</p>																																	
<p>predicted answer</p>																																	

Figure 12: Evaluation prompt in SEAT.

Prompt (CoT inference)	
You are a helpful assistant. You should think step-by-step and provide a detailed and helpful response.	
文本: {context}	
问题: {question}	
<p><b>Infomation Retrieval, Sorting, Statistics, Multi-hop , Data Analysis (Except Rejection), Conversation,</b></p>	
文本: {context}	<b>Calculation</b>
问题: {question}	请对计算结果进行处理，若为整数，保持不变；若为小数，保留两位小数，遵循四舍五入规则，特别注意当四舍五入结果为0时，保留原始计算结果。
文本: {context}	<b>Rejection</b>
问题: {question}	如果文本内容无法回答该问题，请明确指出。
文本: {context}	<b>Table Size Detection</b>
问题: {question}	统计行列数量时需包含表头，使用以下JSON格式输出Table Size Detection: {"行数": 行数, "列数": 列数}。
文本: {context}	<b>Merged Cell Detection</b>
<p>问题: {question} 寻找合并单元格时需包含表头，使用以下JSON格式输出，left_top表示合并单元格的左上角索引位置，right_bottom表示合并单元格的右下角索引位置，cell_value是合并单元格的内容，所有的索引从1开始：</p> <pre>[   {     "left_top": [],     "right_bottom": [],     "cell_value": ""   } ]</pre> <p>若没有合并单元格则直接输出“未找到合并单元格”。</p>	

Figure 13: CoT prompt in inference. The prompt varies depending on the tasks, with the text in red representing the tasks corresponding to each

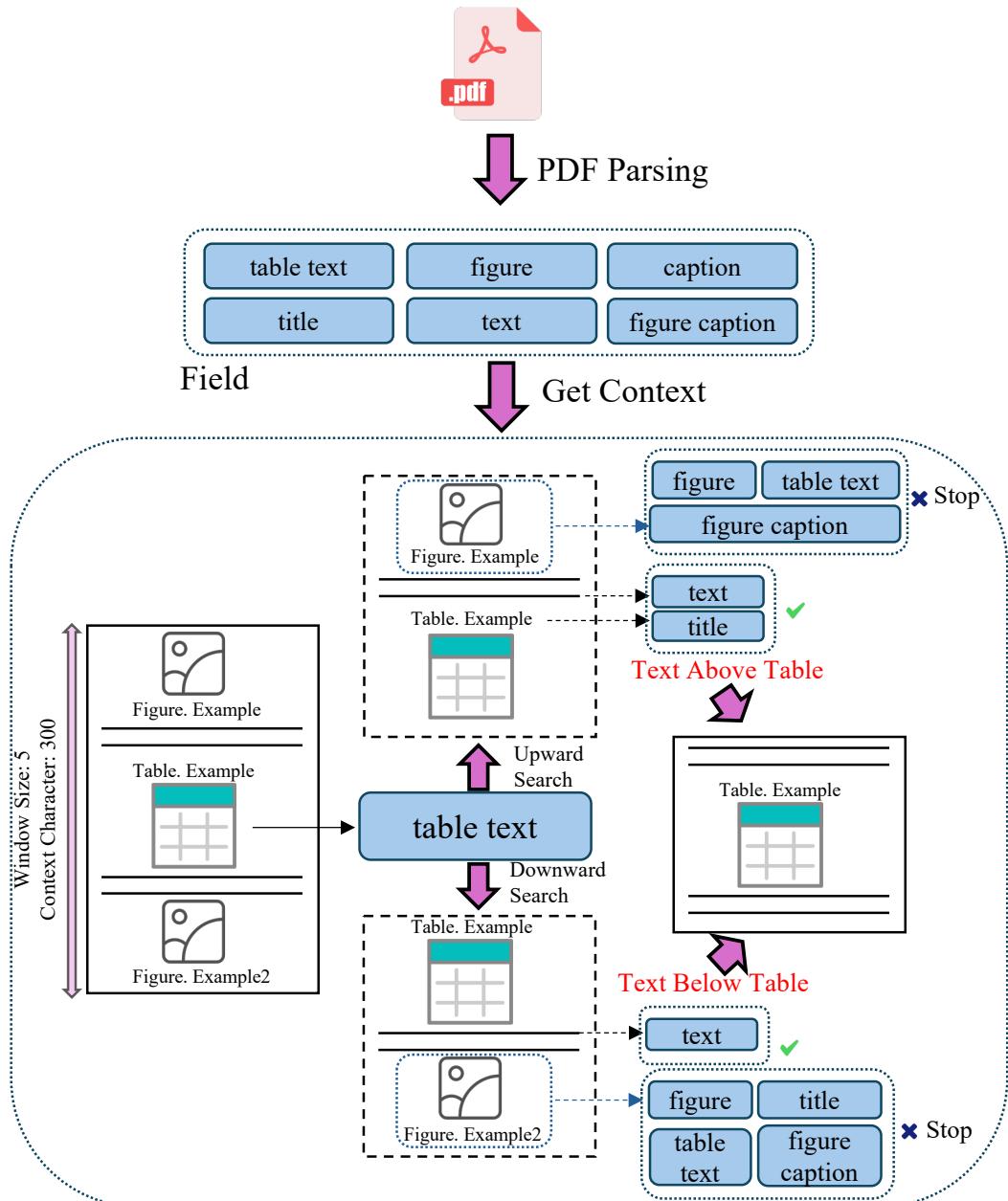


Figure 14: Overview of Heuristic Rules for Table Context Extraction.

---

**Algorithm 1** Heuristic Rules for Table Context Extraction

---

**Input:**

- $\mathcal{L}$ : PDF layout analysis results, a list of elements, including layout type and textual content.
- $\omega$ : Window range for context extraction, expressed as the number of paragraphs or layout blocks.
- $\kappa$ : Character count limit for the extracted context.

**Output:**

- $\mathcal{T}$ : A set of tables, each table augmented with its extracted upward and downward context.

```
1: procedure EXTRACTTABLECONTEXT( $\mathcal{L}, \omega, \kappa$ )
2:    $\mathcal{T} \leftarrow \text{LocateTables}(\mathcal{L})$                                  $\triangleright$  Identify all tables in the layout
3:   for each Table  $T \in \mathcal{T}$  do
4:      $C_{\text{up}} \leftarrow \emptyset$                                           $\triangleright$  Initialize upward context
5:      $C_{\text{down}} \leftarrow \emptyset$                                           $\triangleright$  Initialize downward context
6:     for each Window  $w$  in  $\text{TraverseUpward}(T, \omega)$  do           $\triangleright$  Upward Context Search
7:       if  $\text{IsHeader}(w)$  then
8:          $C_{\text{up}} \leftarrow \text{ExtractText}(w, T)$ 
9:         break
10:      else if  $\text{IsNonTextContent}(w)$  then
11:         $C_{\text{up}} \leftarrow \text{ExtractTextBelow}(w)$ 
12:        break
13:      else if  $\text{CharCountExceeded}(C_{\text{up}}, \kappa)$  then
14:        break
15:      else if  $\text{WindowRangeExceeded}(w)$  then
16:        break
17:      end if
18:       $\text{SkipIrrelevantContent}(w)$ 
19:   end for                                                  $\triangleright$  Downward Context Search
20:   for each Window  $w$  in  $\text{TraverseDownward}(T, \omega)$  do
21:     if  $\text{IsHeader}(w)$  then
22:        $C_{\text{down}} \leftarrow \text{ExtractText}(T, w)$ 
23:       break
24:     else if  $\text{IsNonTextContent}(w)$  then
25:        $C_{\text{down}} \leftarrow \text{ExtractTextAbove}(w)$ 
26:       break
27:     else if  $\text{CharCountExceeded}(C_{\text{down}}, \kappa)$  then
28:       break
29:     else if  $\text{WindowRangeExceeded}(w)$  then
30:       break
31:     end if
32:      $\text{SkipIrrelevantContent}(w)$ 
33:   end for
34:    $\text{CombineContext}(T, C_{\text{up}}, C_{\text{down}})$ 
35:   end for                                                  $\triangleright$  Return tables with augmented context
36:   return  $\mathcal{T}$ 
37: end procedure
38: function LOCATETABLES( $\mathcal{L}$ )
39:   return  $\text{ElementsOfType}(\mathcal{L}, \text{"Table"})$ 
40: end function
```

---