T2R-bench: A Benchmark for Generating Article-Level Reports from Real World Industrial Tables

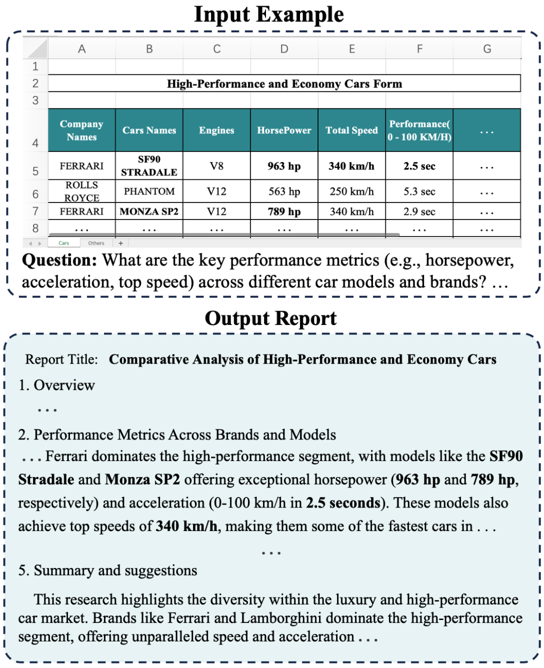
**Jie Zhang[1,*], Changzai Pan[1,*], Sishi Xiong[1,*], Kaiwen Wei[2,*],
Yu Zhao[1], Xiangyu Li[1], Jiaxin Peng[1], Xiaoyan Gu [1],
Jian Yang[3], Wenhan Chang[1], Zhenhe Wu[3], Jiang Zhong[2],
Shuangyong Song[1], Xuelong Li[1,†]**

[1] Institute of Artificial Intelligence (TeleAI), China Telecom,
[2] Chongqing University, [3] Beihang University

Abstract

Extensive research has been conducted to explore the capabilities of large language models (LLMs) in table reasoning. However, the essential task of transforming tables information into reports remains a significant challenge for industrial applications. This task is plagued by two critical issues: 1) the complexity and diversity of tables lead to suboptimal reasoning outcomes; and 2) existing table benchmarks lack the capacity to adequately assess the practical application of this task. To fill this gap, we propose the **table-to-report** task and construct a bilingual benchmark named **T2R-bench**, where the key information flow from the tables to the reports for this task. The benchmark comprises 457 industrial tables, all derived from real-world scenarios and encompassing 19 industry domains as well as 4 types of industrial tables. Furthermore, we propose an evaluation criteria to fairly measure the quality of report generation. The experiments on 25 widely-used LLMs reveal that even state-of-the-art models like Deepseek-R1 only achieves performance with 62.71% overall score, indicating that LLMs still have room for improvement on T2R-bench.

Data:
https://huggingface.co/datasets/Tele-AI/TeleTableBench
https://github.com/Tele-AI/TeleTableBench



Input Example

	A	B	C	D	E	F	G
1							
2	**High-Performance and Economy Cars Form**						
3							
4	**Company Names**	**Cars Names**	**Engines**	**HorsePower**	**Total Speed**	**Performance(0 - 100 KM/H)**	. . .
5	FERRARI	**SF90 STRADALE**	V8	**963 hp**	**340 km/h**	**2.5 sec**	. . .
6	ROLLS ROYCE	PHANTOM	V12	563 hp	250 km/h	5.3 sec	. . .
7	FERRARI	**MONZA SP2**	V12	**789 hp**	340 km/h	2.9 sec	. . .
8	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Question: What are the key performance metrics (e.g., horsepower, acceleration, top speed) across different car models and brands? …

Output Report

Report Title: **Comparative Analysis of High-Performance and Economy Cars**

1. Overview

. . .

2. Performance Metrics Across Brands and Models

. . . Ferrari dominates the high-performance segment, with models like the **SF90 Stradale** and **Monza SP2** offering exceptional horsepower (**963 hp** and **789 hp**, respectively) and acceleration (0-100 km/h in **2.5 seconds**). These models also achieve top speeds of **340 km/h**, making them some of the fastest cars in . . .

. . .

5. Summary and suggestions

This research highlights the diversity within the luxury and high-performance car market. Brands like Ferrari and Lamborghini dominate the high-performance segment, offering unparalleled speed and acceleration . . .

Figure 1: The illustration of table-to-report task. The goal of this task is to analyze numerical data from table to generate comprehensive, coherent and accurate report, including descriptions, analysis and conclusions.

1 Introduction

The rapid development of large language models (LLMs) has significantly advanced research progress in table reasoning (Lu et al., 2024). Traditional research has primarily focused on tasks such as table-to-text generation (Parikh et al., 2020), table question answering (Pasupat and Liang, 2015a; Hu et al., 2024b; Xiong et al., 2025a,b,c), fact verification (Chen et al., 2019), text2sql (Li et al., 2024c; Wu et al., 2025b,a) and table data analysis (Chen, 2023).

However, the automation of report generation from tables is far more widely adopted in industrial applications[1], such as industrial data analysis systems (Ma et al., 2023), business intelligence (BI), table analysis tools and enterprise reporting tools. Notably, systematic research in this field remains largely unexplored and urgently requires further in-depth investigation. In addition, industrial tables commonly exhibit high complexity and diversity, creating a significant gap between existing academic benchmarks and industrial demands,

* These authors contributed equally to this work.
† Corresponding author: xuelong_li@ieee.org

[1] Typical industrial table applications include Microsoft Power BI, SAP BusinessObjects, IBM Cognos, MicroStrategy, Smartbi, etc.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22427–22451
November 4-9, 2025 ©2025 Association for Computational Linguistics

which poses new challenges for related research.

In light of the above considerations, we propose **table-to-report**, a novel task designed to convert structured tabular data into natural language reports, aiming to present data, trends, and insights to enhance table information flow efficiency (Shao and Li, 2025) as illustrated in Figure 1.

As an emerging task, it still faces several key challenges: (1) **Lack of high-quality benchmark**: current table benchmarks, such as Text2Analysis (He et al., 2024), TableBench (Wu et al., 2024) and MiMoTable (Li et al., 2024b) primarily evaluate LLMs on table question answering tasks, each with a distinct focus. However, these benchmarks are not designed to assess table-to-report task. Besides, the table datasets used in most benchmarks predominantly consist of open-source academic data, failing to fully capture the main features and types of industrial tabular data, such as multiple tables, complex structure, and extremely large-size tables. The data volume and scale remain significantly constrained for extremely large-size tables (Sui et al., 2024). (2) **Lack of targeted evaluation criteria**: existing criteria like BLEU (Papineni et al., 2002a) and ROUGE (Lin, 2004) designed for summarization tasks, are unsuitable for table-to-report task due to non-unique gold standards. While general LLM-as-a-judge (Li et al., 2024a) method performs excel in text quality assessment,it neglects to evaluate numerical accuracy and table topic coverage, therefore limiting its applicability.

To address the aforementioned issues, we introduce **T2R-bench**, a high-quality benchmark designed to evaluate the reasoning capabilities of LLMs in the table-to-report task. T2R-bench encompasses Chinese and English tables from real-world industrial scenarios, covering 6 domains and 19 secondary industry categories. Compared to existing table benchmarks, as shown in Table 1, our benchmark features a comprehensive and diverse collection of single tables, multiple tables, complex structured tables, and extremely large-size tables, enhancing the benchmark's practicality and challenge. We also craft the designed approaches for table question annotation and report reference annotation. Furthermore, we develop an evaluation system that incorporates three criteria of numerical accuracy, information coverage, and general quality to comprehensively assess report quality. In the experiment, we select 25 widely-used methods for evaluation, the results demonstrate that strongest models struggle to achieve satisfactory performance on the table-to-report task.

Our contributions are summarized as follows:

Task and Benchmark	Multiple Table	Complex Structure Table	Extremely Large-Size Table	Answer Lengths
TableQA				
WikiSQL (Zhong et al., 2017)	×	×	×	1.9
WTQ (Pasupat and Liang, 2015b)	×	×	×	10.39
TAT-QA (Zhu et al., 2021)	×	×	×	20.3
FeTaQA (Nan et al., 2021)	×	×	×	18.9
AIT (Katsis et al., 2021)	×	×	×	1.1
TabFact (Chen et al., 2020)	×	×	×	18.3
TableBench (Wu et al., 2024)	×	×	×	8.5
HiTab (Cheng et al., 2022)	×	✓	×	12.9
DataBench (Grijalba et al., 2024)	×	×	✓	3.2
Mimo (Li et al., 2024b)	✓	✓	×	44.2
Spider (Yu et al., 2018)	✓	×	✓	35.5
Table2Text				
ToTTo (Parikh et al., 2020)	×	×	✓	17.4
DAE-val (Hu et al., 2024a)	×	×	✓	3.6
DataTales (Yang et al., 2024b)	×	×	✓	108.0
Text2Analysis (He et al., 2024)	×	×	×	/
Table2Report				
T2R-Bench (ours)	✓	✓	✓	950.2

Table 1: Comparison with existing datasets on table types and answer lengths. Since Text2Analysis benchmark dose not provide the publicly accessible download links, the average length could not be calculated. A detailed comparison is provided in Appendix B.

- We introduce **T2R-bench**, the first real world industrial benchmark for the table-to-report task. It encompasses 457 real-world tables across 19 domains, covering 4 industrially relevant types, including single tables, multiple tables, complex structured tables, and extremely large-size tables.
- We propose an evaluation system for table-to-report generation, incorporating 3 carefully designed criteria to assess report accuracy and reliability. Extensive validation demonstrates that the evaluation system achieves strong alignment with human judgment.
- We evaluate the ability of 25 strong methods on T2R-Bench. The experiments show that the best performed model Deepseek-R1 achieves only 62.71% overall score, which suggests great challenges in satisfying real-world table-based report generation needs.

2 Related Work

Tabular Benchmarks. With the development of deep learning (Wei et al., 2021, 2023b,a; Xing et al., 2025; Wang et al.; Zhao et al., 2025; Wu et al., 2025c; Wang et al., 2024a, 2025a; Dai et al., 2025a,b; Wang et al., 2024d; Li et al., 2025; Team et al., 2025), recent advances in table reasoning research have driven the development of diverse benchmarks covering TableQA, Table2Text, and advanced data analysis tasks, incorporating various table types including large-size tables, multiple tables, and complex structures. TableQA benchmarks (Zhong et al., 2017; Chen et al., 2020; Nan et al., 2021; Osés Grijalba et al., 2024) dominate the landscape, with TableBench (Wu et al., 2024) emerging as a representative benchmark that captures real-world tabular reasoning challenges. For Table2Text tasks (Lebret et al., 2016), ToTTo (Parikh et al., 2020) constructs table-description pairs from Wikipedia snippets, while DATATALES (Yang et al., 2024b) generates financial narratives from tabular data. Advanced analysis benchmarks like DAEval (Hu et al., 2024a) and Text2Analysis (He et al., 2024) focus on programmatic table manipulation. However, as evidenced in Table 1, current solutions remain limited in their coverage of diverse table types (including large-scale, multi-table, and complex layouts) and are constrained to sentence-level outputs that fail to meet industrial requirements for comprehensive report generation.

Recent research has placed growing emphasis on complex table structure understanding (Cheng et al., 2022; Katsis et al., 2021; Tang et al., 2024; Mathur et al., 2024), yielding specialized benchmarks like MiMoTable (Li et al., 2024b) for multidimensional spreadsheets, DataBench (Grijalba et al., 2024) for containing a limited number of large-size tables, and SPREADSHEETBENCH (Ma et al., 2024) for multiple tables manipulation. However, these works focused on TableQA and manipulation tasks, overlooking comprehensive report generation needs.

Text quality Evaluation. Established metrics like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002b), and BERTScore (Zhang et al., 2020) have been widely adopted, complemented by emerging LLM-as-judge approaches (Li et al., 2024a). For Table2Text tasks, Text2Analysis employs code generation metrics, while (Wiseman et al., 2017) designs three new dataset-adapted evaluation metrics for text generation. ToTTo (Li et al., 2024b) adapts PARENT (Dhingra et al., 2019) alongside BLEU. DATATALES introduces domain-specific criteria including factual accuracy, insightfulness, and stylistic quality, demonstrating the necessity for task-aligned evaluation frameworks. However, those methods typically neglect to evaluate numerical accuracy and table topic coverage (Szymanski et al., 2025), hindering the evaluation applicability.

3 Construction of T2R-bench

Table-to-report is the task of automatically converting a structured table T into a fluent article-level report R. To evaluate existing approaches, we introduce T2R-bench, whose construction pipeline consists of three key components: table data collection, table question annotation, and report reference annotation, as detailed in Figure 2.

3.1 Table Data Collection

The tables of T2R-bench are collected from publicly available internet resources. The primary sources encompass municipal open data platforms, the National Bureau of Statistics, industrial association official websites and open-source tabular datasets (refer to Appendix C.7 for details). We collect tables to cover as many real-world scenarios as possible, including single table with individual and multiple sheets, multiple tables, extremely large-size tables, tables with simple and complex header structures.

Specifically, we leverage a two-stage selecting method. Firstly, tables are pre-screened based on industry-specific topics to ensure domain relevance. Subsequently, to ensure each table has sufficient information density for statistical analysis, we remove files with obvious garbled text or cell blank values exceeding 60%. To ensure the quality and legality of the collected tables, we manually review each table and anonymize any potential private and sensitive information. Ultimately, we collect 252 Chinese tables and 205 English tables across 6 distinct domains and 19 secondary classes based on topics to fit diverse industrial fields.

3.2 Table Question Annotation

We adopt a semi-automatic heuristic method to efficiently generate diverse and high-quality questions. The specific steps are shown as follows:

Seed Question and Prompt Preparing. To improve the precision and relevance of the generated questions, we employ 24 annotators with expertise in data analysis and report writing in diverse domains (see Appendix C.1 for annotator qualifications). They carefully curate 10 seed questions, and meticulously design the prompt template library with 5 diverse prompt templates (prompt templates and seed questions are provided in Appendix C.4).

Self-Instruct to Generate Questions. We employ self-instruct(Wang et al., 2023) by using GPT-4o to efficiently generate a pool of questions. Two

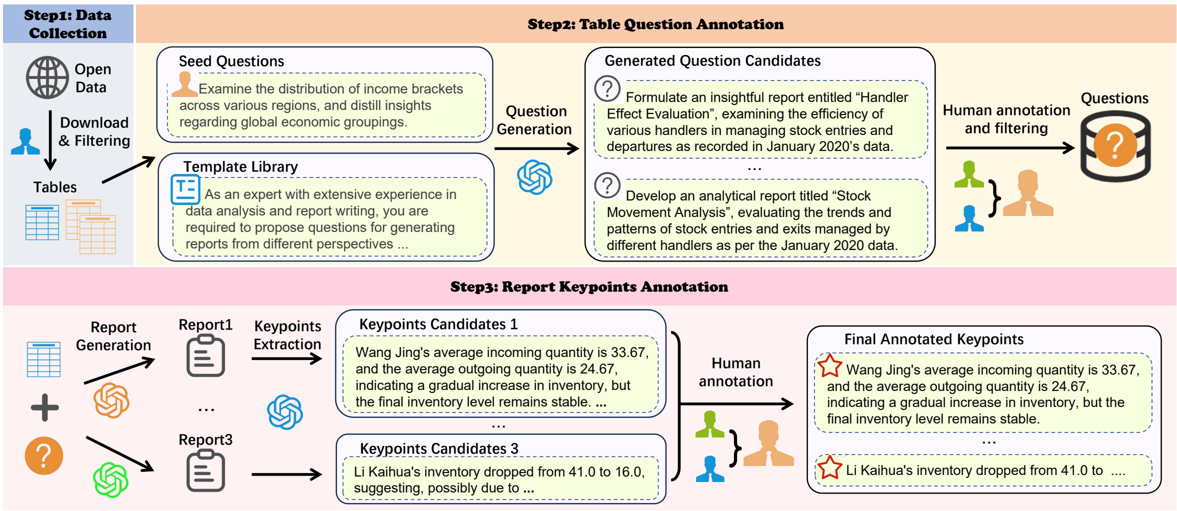

Step1: Data Collection
Open Data
Download & Filtering
Tables
Step2: Table Question Annotation
Seed Questions
Examine the distribution of income brackets across various regions, and distill insights regarding global economic groupings.
Template Library
As an expert with extensive experience in data analysis and report writing, you are required to propose questions for generating reports from different perspectives ...
Question Generation
Generated Question Candidates
Formulate an insightful report entitled "Handler Effect Evaluation", examining the efficiency of various handlers in managing stock entries and departures as recorded in January 2020's data.
...
Develop an analytical report titled "Stock Movement Analysis", evaluating the trends and patterns of stock entries and exits managed by different handlers as per the January 2020 data.
Human annotation and filtering
Questions
Step3: Report Keypoints Annotation
Report Generation
Report1
...
Report3
Keypoints Extraction
Keypoints Candidates 1
Wang Jing's average incoming quantity is 33.67, and the average outgoing quantity is 24.67, indicating a gradual increase in inventory, but the final inventory level remains stable. ...
...
Keypoints Candidates 3
Li Kaihua's inventory dropped from 41.0 to 16.0, suggesting, possibly due to ...
Human annotation
Final Annotated Keypoints
Wang Jing's average incoming quantity is 33.67, and the average outgoing quantity is 24.67, indicating a gradual increase in inventory, but the final inventory level remains stable.
...
Li Kaihua's inventory dropped from 41.0 to ....

Figure 2: An overview of the construction pipeline for T2R-bench.

prompt templates are randomly selected from the prompt template library for each table. Each template incorporates 2-5 seed questions as in-context demonstrations, with instructions to generate 3 relevant questions.

Human Annotation and Filtering. We randomly assign each question to two annotators, whose selection criteria and qualifications are detailed in Appendix C.1. Annotators evaluate question candidates based on three criteria: 1) tabular answer ability, where questions must be answerable solely using table data without external knowledge; 2) focused conclusions, where questions should target single analytical dimensions for definitive conclusions; and 3) complementary uniqueness, where questions from the same table must address distinct aspects. In cases where the evaluation results of the two annotators are inconsistent, the results will be handed over to a third senior annotator with extensive domain expertise and experience for the final judgment (For detailed annotation procedure, please refer to Appendix C.2). Through this rigorous quality assurance procedure, we obtain 910 high-quality, comprehensive questions.

3.3 Report Reference Annotation

Unlike summarization tasks, which often yield a single optimal summary, table-to-report tasks exhibit significant variability due to differences in expression, stylistic preferences, structural choices among annotators, and the inherent complexity of tabular data. Consequently, using entire reports as reference standards proves impractical.

To this end, we observe that professionally authored reports on the same tabular content and report topic consistently share core elements, including central viewpoints, analytical conclusions, recommendations, critical supporting data, despite variations in phrasing or presentation. This consistency motivates our introduction of report keypoints: distilled representations of a report's essential content, encompassing its analytical backbone and evidentiary support (See Figure 2 for keypoint examples). These invariant keypoints provide a robust basis for evaluating generated reports.

Based on this finding, we design the report reference annotation process, which consists: 1) **Report Generation**. We leverage three distinct LLMs(Qwen-3-32B, Moonshot-V1-32K and Deepseek-R1) to generate different reports for each <table, report question> pair, resulting in three distinct reports (see Appendix E.1 for the prompt template). 2) **Keypoints Extraction**. Then, we prompt GPT-4o to distill the most crucial information from each report, extracting 5-10 keypoints, resulting three groups of keypoints for each <table, question> pair (see Appendix C.5 for the prompt template). 3) **Human Annotation**. Mirroring the question annotation procedure, we implement a rigorous dual-annotator verification protocol for key point refinement, with discrepancies resolved by senior annotators with data analysis and domain-specific report writing experience. Please see Appendix C.1 and C.3 for full qualifications and annotation procedure details.

3.4 Dataset Statistics

Through the construction process, T2R-bench comprises 910 high-quality questions originating from 457 unique tables, along with 4,320 annotated re-

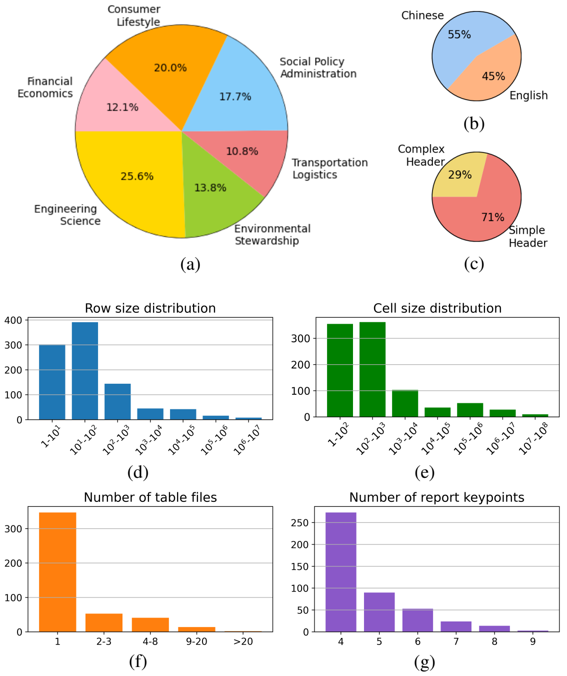


Figure 3: Distribution of different types of tables in T2R-bench. (a) Domain distribution. (b) Proportion of Chinese and English tables. (c) Proportion of complex header tables. (d-e) The row and cell size distribution for all tables in T2R-bench. (f) Number of table files for single tables and multiple tables. (g) Distribution of report reference keypoints for each report question.

port keypoints. These meticulously annotated keypoints of the report will serve as the gold reference to evaluate report in Section 4.2.

Table Statistics. Table 2 and Figure 3 show the key statistics and distribution of tables in T2R-bench. Specifically, the global statistics reveal that T2R-bench contains over 8.3% extremely large-size tables containing more than 50K cells; 28.9% complex structured tables with hierarchical indexing, merged cells, and non-uniform cell structures; and 23.6% multiple tables comprising interdependent files or sheets. A key feature distinguishing our benchmark is its substantial number of extremely large-size tables.

Domain Distribution. As shown in Figure 3a, T2R-bench covers six main industry domains, which can be further divided into 19 more specific sub-domains, including engineering, manufacturing, finance, education, healthcare, telecommunications, transportation (detailed sub-categories refers to Table 10 of Appendix C.6), ensuring that abundant types of tables in the dataset encompass as many real-world scenarios as possible.

Property	Value
Number of Tables	457
Avg Table Files or Sheets for Multi-Tables	5.04
Avg Rows per Table	30,183
Avg Cells per Table	490,308
Number of Extremely Large-size Tables	38
Avg Rows for Extremely Large-size Tables	721,882
Avg Cells for Extremely Large-size Tables	11,895,814
Number of Questions	910
Avg Questions per Table	1.99
Avg Report Reference Keypoints per Question	4.75

Table 2: Key Statistics of T2R-bench.

Questions and Report Keypoints. Following the human annotation process, T2R-bench comprises a total of 910 questions. Notably, the number of questions is pruned from an initial range of 3.00 to 1.99 per table during the expert annotation phase. As illustrated in Figure 2, the number of report keypoints per question is reduced to an average of 4.75 after expert verification and filtering. These rigorous annotation and verification processes enhance the quality of the benchmark.

4 Evaluation Criteria

To address the challenges encountered in automated evaluation for the table-to-report task, we propose a comprehensive evaluation system from three aspects: numerical accuracy, information coverage, and general quality.

4.1 Numerical Accuracy Criterion

Generated reports frequently incorporate numerical values, some directly extracted from source tables and others derived through data synthesis (e.g., aggregations like column averages). To ensure the fidelity of such numerical claims, we propose the Numerical Accuracy Criterion (NAC), a self-consistency mechanism for validating numerical facts against their tabular sources.

Specifically, we first segment sentences in the target report using NLTK[2] (for English) and Jieba[3] (for Chinese). We then apply regular expressions to extract clusters of sentences containing numerical statements (integers or floating-point numbers). For each cluster, we generate corresponding verification questions, treating the extracted numerical values as ground-truth answers (see Appendix D.1 for prompt).

To resolve these questions robustly, we employ three specialized code-generation LLMs (i.e., Qwen2.5-32B-Coder-Instruct, Deepseek-Coder,

[2] https://www.nltk.org/

[3] https://github.com/fxsjy/jieba

and CodeLlama-70B-Instruct), capable of interpreting and executing numerical operations (see Appendix D.1 for details). NAC enforces consensus by requiring agreement from at least two models; discordant results (including execution failures) are discarded to minimize noise. The final NAC score is computed by systematically comparing the validated solutions against the original numerical assertions in each sentence cluster.

4.2 Information Coverage Criterion

To address the challenges of incomplete coverage and irrelevant content in LLM-generated reports, we propose the Information Coverage Criterion (ICC), a quantitative measure of semantic alignment between generated reports and reference keypoints. Inspired by the successful application of mutual information (MI) in machine translation for evaluating alignment quality, ICC assesses how effectively a report preserves essential information from the source table.

Specifically, for each generated report, we define $K = \{k_1, k_2, \ldots, k_M\}$ as the set of annotated keypoints, where M represents the total keypoint number. Then, the generated report is segmented into multiple sentence clusters $S = \{s_1, s_2, \ldots, s_N\}$ by NLTK toolkit (English reports) and Jieba toolkit (Chinese reports). After that, we construct a semantic similarity matrix S, where each element S_{ij} represents the semantic similarity of keypoints-sentence pair(k_i, s_j) calculated by BERTScore(Zhang et al., 2020):

$$S_{ij} = BERTScore(k_i, s_j)$$

Given the similarity matrix S, the ICC is defined as normalized MI:

$$ICC = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} P(k_i, s_j) \log \frac{P(k_i, s_j)}{P(k_i)P(s_j)}}{-\sum_{i=1}^{M} P(k_i) \log P(k_i)} \tag{1}$$

where the joint and marginal probabilities are derived from similarity matrix S as follows:

$$\begin{aligned} P(k_i, s_j) &= \frac{S(k_i, s_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} S(k_i, s_j)} \\ P(k_i) &= \frac{\sum_{j=1}^{N} S(k_i, s_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} S(k_i, s_j)} \\ P(s_j) &= \frac{\sum_{i=1}^{M} S(k_i, s_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} S(k_i, s_j)} \end{aligned}$$

Eq. (1) provides an information-theoretic measure scaled to [0,1] by dividing the keypoint entropy $H(K)$, enabling consistent comparison across reports with varying numbers of keypoints. The final evaluation aggregates ICC scores across all reports, with higher values indicating better preservation of critical information in the generated outputs.

4.3 General Evaluation Criterion

Inspired by evaluation methodologies for long-context generation (Lee et al., 2024), we propose the General Evaluation Criterion (GEC) to holistically assess report quality using LLMs as judges. GEC focuses on five key dimensions that most effectively discriminate report quality: reasoning depth, human-like style, practicality, content completeness and logical coherence. The final GEC score is computed as the average across these dimensions. Detailed evaluation criteria and prompts are provided in Appendix D.3.

5 Experiments

5.1 Experimental Setup

Baselines and Evaluation. We evaluate 25 strong methods on T2R-Bench, including both open-source and closed-source foundation models. The open-source models comprise Qwen series (Bai et al., 2023; Yang et al., 2024a; Qwen et al., 2025; Hui et al., 2024), Llama family (Dubey et al., 2024; Roziere et al., 2023), Mistral (Jiang et al., 2023), Deepseek models (DeepSeek-AI et al., 2024; Guo et al., 2024; DeepSeek-AI et al., 2025), and TeleChat (Wang et al., 2024b,c, 2025b), while the closed-source models include GPT series (OpenAI, 2023), OpenAI o1-mini, Claude-3.5-Sonnet2, Doubao, and Moonshot. We also incorporate table-specific LLMs in our evaluation such as TableGPT2(Su et al., 2024). However, we exclude TableLLaMA from the main results, as it consistently generates abbreviated outputs rather than the comprehensive reports required by our task.

For closed-source models, we used the default parameters of the API. We employed model inference with the transformers version 4.51.22 and the vllm version 0.6.3.

The evaluation covers 4 practical industrial scenarios: single tables, multiple tables, complex structured tables, and extremely large-size tables. We assess all models using the proposed metrics: Numerical Accuracy Criterion (NAC), Information Coverage Criterion (ICC), and General Evaluation Criterion (GEC), and we report both overall and

Model	Overall				Single			Multiple			Complex Structure			Extremely Large-Size		
	NAC	ICC	GEC	AVG	NAC	ICC	GEC	NAC	ICC	GEC	NAC	ICC	GEC	NAC	ICC	GEC
Open-Source Models																
TableGPT2-7B (Su et al., 2024)	34.24	25.70	81.28	47.07	49.54	35.91	83.09	40.99	31.29	81.76	30.12	27.40	79.87	16.33	8.20	80.42
Qwen1.5-14B-Chat (Bai et al., 2023)	36.03	26.29	83.83	48.72	50.67	46.62	82.61	38.31	27.06	82.72	40.02	25.67	85.17	15.12	5.81	84.82
Qwen2.5-72B-Instruct (Qwen et al., 2025)	47.82	42.28	88.68	59.59	67.29	58.23	87.82	54.15	46.23	89.65	43.58	47.40	**90.42**	26.18	17.24	86.84
Qwen2-72B (Yang et al., 2024a)	44.64	33.76	87.76	55.39	66.23	50.36	88.55	46.96	39.35	88.71	39.92	29.93	88.53	25.47	15.40	85.25
Qwen2.5-32B (Qwen et al., 2025)	42.91	35.85	84.54	54.43	54.36	44.45	79.45	50.03	46.53	86.82	45.84	40.21	88.64	21.41	12.21	83.26
Qwen2.5-Coder-32B-Instruct (Hui et al., 2024)	43.82	32.33	86.25	54.13	61.97	46.36	86.18	44.57	40.82	86.24	46.23	28.53	86.13	22.52	13.62	86.43
Qwen3-30B-A3B (Qwen et al., 2025)	49.46	42.27	88.02	59.90	70.32	56.46	90.35	54.35	47.35	88.36	47.82	45.65	87.02	25.36	19.63	86.24
Qwen3-32B (Qwen et al., 2025)	53.01	45.01	89.12	61.37	73.21	59.34	**91.21**	58.24	50.53	**90.23**	51.24	48.82	88.53	29.35	**21.25**	88.82
Qwen3-8B (Qwen et al., 2025)	36.60	31.65	78.38	48.87	51.26	42.36	77.27	40.56	34.55	81.46	39.27	35.13	76.54	15.34	14.54	78.23
CodeLlama-70B-Instruct (Roziere et al., 2023)	40.04	29.72	80.80	50.19	47.34	36.82	85.64	50.93	42.18	79.53	42.67	34.07	80.81	19.22	5.80	77.21
Deepseek-Chat-V3 (DeepSeek-AI et al., 2024)	51.47	42.26	**89.63**	61.12	68.58	59.64	90.18	55.64	49.47	89.18	52.25	39.31	89.42	**29.43**	20.63	**89.72**
Deepseek-Coder (Guo et al., 2024)	50.96	40.93	87.07	59.65	71.52	55.51	87.45	56.32	47.06	88.35	50.17	46.53	88.21	25.83	14.62	84.28
Deepseek-R1 (DeepSeek-AI et al., 2025)	**53.51**	**45.12**	89.51	**62.71**	**74.58**	**60.64**	90.18	57.64	48.47	89.18	**53.39**	**50.32**	89.07	28.43	21.05	89.62
Llama3.1-70B (Dubey et al., 2024)	40.33	34.40	76.52	50.42	54.05	52.36	81.82	43.56	32.71	75.76	46.12	40.32	77.25	17.57	12.20	71.23
Llama3.1-8B (Dubey et al., 2024)	34.09	28.61	72.82	45.17	49.26	40.36	72.18	38.84	30.35	77.29	36.01	33.53	67.33	12.25	10.20	74.46
Llama3.3-70B (Dubey et al., 2024)	42.25	31.19	78.07	50.50	56.05	49.26	82.32	46.56	31.23	78.31	48.62	31.13	78.23	18.57	13.12	73.42
Mistral-Large-Instruct-2407 (Jiang et al., 2023)	44.28	35.86	79.86	53.33	59.15	51.36	86.23	53.26	37.72	82.63	49.42	43.12	78.25	15.27	11.24	72.32
Qwen2.5-7B-instruct (Qwen et al., 2025)	35.52	30.43	75.73	46.45	50.63	41.63	76.84	39.62	33.25	79.36	39.27	34.45	74.36	13.25	14.21	76.32
Telechat2.5-35B (Wang et al., 2024b)	45.18	34.71	86.56	55.48	66.45	49.98	88.32	47.12	38.12	85.23	41.02	35.07	85.84	26.13	15.67	86.85
Closed-Source Models[a]																
Moonshot-V1-32K	42.41	36.05	87.11	55.19	60.25	46.55	84.36	50.35	42.24	88.35	39.72	40.20	87.20	19.33	15.21	88.54
Claude-3.5-Sonnet	47.62	36.31	88.61	57.51	62.18	44.36	88.43	54.28	48.53	87.59	47.64	41.13	89.60	26.39	11.24	88.83
Doubao-Pro-128K	49.14	31.28	82.98	54.47	65.01	29.07	82.91	56.04	39.41	84.47	50.57	43.80	84.40	24.94	12.83	80.13
Doubao-Pro-32K	44.58	31.47	81.21	52.42	61.83	34.18	79.64	51.02	44.29	82.59	48.80	37.53	83.37	16.67	9.86	79.25
GPT-4o (OpenAI, 2023)	49.35	41.91	88.72	59.29	73.35	54.91	87.82	54.10	56.35	88.47	42.27	49.53	89.32	27.69	16.83	89.26
OpenAI o1-mini	51.59	41.19	89.07	60.62	69.41	53.36	88.36	**60.94**	**66.29**	89.53	47.98	35.27	90.17	28.04	18.84	88.21

Table 3: Overall performance of LLMs on T2R-bench. For each criterion, the best result is marked in bold, and the second best result is underlined.

Model	Languages	
	Chinese	English
Qwen3-32B	62.43	60.07
Qwen2.5-72B-Instruct	60.43	58.56
Deepseek-R1	63.74	61.45
Llama3.3-70B	48.26	53.24
GPT-4o	59.59	60.48

Table 4: Performance of LLMs on bilingual tables. The indicators in the table are based on the average values of NAC, ICC, and GEC.

average performance scores.

Implementation Details. We design a uniform style prompt template to ensure the fairness of the evaluation. To ensure compatibility across diverse file types and encodings, input tables are initially converted to a standardized CSV file using UTF-8 encoding. These tables are then converted into Markdown format, which is optimally structured for the LLM to use directly for generation. For tables whose content exceeds the LLM's context length limit, the content will be truncated. For closed-source models, we utilize official APIs with default parameters to generate complete reports, with detailed website information provided in Table 14 from Appendix E.4. For open-source models, we use 16 A100 40G GPUs for inference with the transformers version 4.51.22 and the vllm version 0.6.3. All models use the official default parameters. The uniform style prompt template can be found in Appendix E.1.

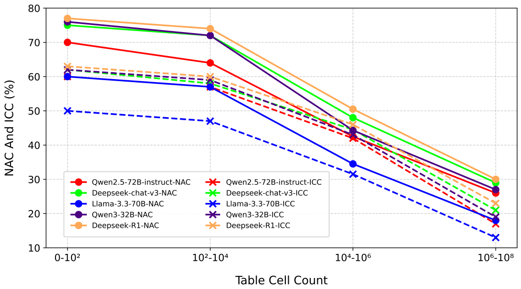


Figure 4: The performance of different LLMs on NAC and ICC criteria across varying numbers of table cell.

5.2 Main Results

Overall Performance. As shown in the Table 3, we conduct a comparative analysis of advanced LLMs on the proposed T2R-Bench. We could find: (1) The Deepseek series demonstrates superior performance across single table, multiple table, and complex table tasks, establishing its leading capability in Table-to-Report applications. (2) Notably, Qwen3-32B achieves the highest NAC score, showcasing exceptional numerical computation abilities and outperforming even the larger Qwen2.5-72B-Instruct model. (3) While the GPT series maintains strong performance with an ICC score of 66.29% on multiple table tasks, we observe significant performance degradation across most models when transitioning from single to multiple table tasks, suggesting limitations in cross-table comprehension. (4) The benchmark proves particularly challenging for extremely large-size tables, where all

	Markdown	Json	Html
Qwen2.5-72B-Instruct	59.59	55.82	54.91
Deepseek-R1	62.71	58.12	60.02
OpenAI-o1-mini	60.62	59.43	59.67

Table 5: Average performance of NAC, ICC and GEC of three different models across markdown, json and html table input formats.

models show substantially reduced performance across all evaluation criteria. The top-performing Deepseek-R1 achieves an average overall score of 62.71%, highlighting the considerable room for improvement in current approaches for comprehensive table understanding tasks.

Analysis of Table Cell Count. We conduct experiment to investigate how the number of cells in input tables affects the performance. As shown in the Figure 4 , we can see that as table size increases, all evaluated LLMs exhibit a sharp performance decline, particularly when processing extremely large-size tables. This finding provides the first empirical evidence in table-related benchmarks that current models face fundamental limitations in comprehending large-scale tabular data, mirroring known challenges in long-text understanding.

Analysis of Bilingual Capability. We conduct the English and Chinese experiment on T2R-Bench, have them processed by the five LLMs for report generation, and subsequently assess using averaged score of the proposed automated evaluation criteria. The Table 4 shows that nearly all models exhibit similar performance in both languages, highlighting their consistent ability to generate bilingual reports. However, Llama-3.3-70B's performance in generating Chinese reports lags significantly behind its English capabilities, indicating a need for further fine-tuning.

Analysis of Input Formatting. Table 5 demonstrates that among the three most representative table input formats (Markdown, HTML, and JSON), the Markdown format achieves the highest average performance, followed by HTML, while JSON exhibits the lowest performance.

5.3 Human Evaluation

As table-to-report is a newly formulated task, we establish human baseline for comparison. Given the substantial time commitment required for human report generation, we randomly select a subset of 50 questions (denoted as D_{val}) from the dataset by stratified sampling, covering single tables, multiple tables, complex-structure tables, and extremely large-size tables. To mitigate confirmation bias, six independent expert annotators with substantial data analysis experience (and no prior involvement in dataset creation) were recruited to generate reference reports, ensuring unbiased evaluations.

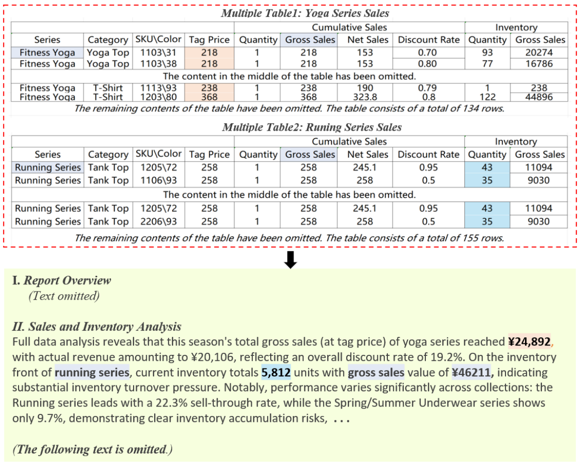
Models	Our Evaluation Criteria	Human Evaluation
Qwen2.5-72B-Instruct	59.59	61.06
Deepseek-R1	62.71	65.58
Llama3.3-70B	50.50	55.09
GPT-4o	59.29	62.56
Qwen3-32B-Instruct	61.37	63.02
Human baseline	89.32	96.52

Table 6: A consistency test of evaluation methods between the proposed evaluation criteria and human evaluation on average performance of NAC, ICC and GEC.

We conducted rigorous validation studies to assess the correlation between our proposed metrics and human evaluation. Another six independent annotators evaluated reports generated by five representative models (Qwen2.5-72B-Instruct, Llama3.3-70B, GPT-4o, DeepSeek-R1, Qwen3-32B-Instruct) alongside human-written reports on D_{val}. Evaluations followed criteria (NAC, ICC, GEC from Section 4), achieving excellent inter-rater reliability (Fleiss' k = 0.85 (Fleiss and Cohen, 1973)). As shown in Table 6, while systematically more stringent, those metrics demonstrated strong correlation with human judgments (Pearson's r = 0.908 (Cohen et al., 2009)), validating the framework's reliability despite absolute score differences.

5.4 Case Study

Our manual analysis of 50 randomly selected error cases from T2R-Bench reveals persistent challenges in LLMs' table-to-report capabilities. As shown in Figure 5, even the top-performing Deepseek-R1 model exhibits critical failures when processing multiple tables, such as numerical hallucinations (e.g., incorrect summation of "Tag Price" in Table 1) and table selection errors (e.g., mistakenly referencing "Gross Sales" from Table 1 instead of Table 2). These errors, along with challenges posed by complex table structures, descriptive hallucinations, and variable misinterpretations, reveal fundamental reasoning limitations despite the models' ability to generate superficially fluent, human-like reports. Comprehensive case study and error analysis are provided in Appendices E.2 and E.3.



Multiple Table1: Yoga Series Sales

				Cumulative Sales				Inventory	
Series	Category	SKU\Color	Tag Price	Quantity	Gross Sales	Net Sales	Discount Rate	Quantity	Gross Sales
Fitness Yoga	Yoga Top	1103\31	218	1	218	153	0.70	93	20274
Fitness Yoga	Yoga Top	1103\38	218	1	218	153	0.80	77	16786
The content in the middle of the table has been omitted.									
Fitness Yoga	T-Shirt	1113\93	238	1	238	190	0.79	1	238
Fitness Yoga	T-Shirt	1203\80	368	1	368	323.8	0.8	122	44896

The remaining contents of the table have been omitted. The table consists of a total of 134 rows.

Multiple Table2: Runing Series Sales

				Cumulative Sales				Inventory	
Series	Category	SKU\Color	Tag Price	Quantity	Gross Sales	Net Sales	Discount Rate	Quantity	Gross Sales
Running Series	Tank Top	1205\72	258	1	258	245.1	0.95	43	11094
Running Series	Tank Top	1106\93	258	1	258	258	0.5	35	9030
The content in the middle of the table has been omitted.									
Running Series	Tank Top	1205\72	258	1	258	245.1	0.95	43	11094
Running Series	Tank Top	2206\93	258	1	258	258	0.5	35	9030

The remaining contents of the table have been omitted. The table consists of a total of 155 rows.

I. *Report Overview*

(Text omitted)

II. Sales and Inventory Analysis

Full data analysis reveals that this season's total gross sales (at tag price) of yoga series reached **¥24,892**, with actual revenue amounting to ¥20,106, reflecting an overall discount rate of 19.2%. On the inventory front of **running series**, current inventory totals **5,812** units with **gross sales** value of **¥46211**, indicating substantial inventory turnover pressure. Notably, performance varies significantly across collections: the Running series leads with a 22.3% sell-through rate, while the Spring/Summer Underwear series shows only 9.7%, demonstrating clear inventory accumulation risks, ...

(The following text is omitted.)

Figure 5: An example illustrating an original table and its corresponding report generated by DeepSeek-R1, with critical error highlighting.

6 Conclusion

To meet practical industrial requirements, we introduce the **table-to-report** task and present **T2R-bench**, which requires models to generate article-level reports from tabular data. T2R-bench comprises 457 real-world tables spanning 19 diverse domains, with coverage of 4 industrial table types. In addition, we develop an adapted framework to rigorously evaluate model performance and conduct experiments on 25 state-of-the-art LLMs. Experimental results demonstrate that the top-performing model, Deepseek-R1, achieves suboptimal performance, revealing great room for advancing LLMs' capabilities in table-to-report generation.

Limitations

While our benchmark represents a significant step forward, several challenges remain. The current best-performing open-source model (Deepseek-R1) achieves suboptimal performance, with both Numerical Accuracy (NAC) and Information Coverage (ICC) scores below 65% on the proposed evaluation framework. This performance gap highlights two critical needs: (1) the expansion of our benchmark dataset to cover more diverse table types and domains, and (2) the development of specialized models specifically designed for the table-to-report task. These limitations underscore the pressing demand for methodological innovations that can bridge the gap between current capabilities and real-world application requirements.

Furthermore, more comprehensive evaluation protocols, particularly for style assessment of reports generated by different models, are worth exploring. Future research could also explore multimodal table-to-report tasks, for instance, by comparing the performance of specialized models like TabPedia (Zhao et al., 2024) against other multimodal approaches. These excited directions will pave the way for more robust, general-purpose systems capable of interpreting and communicating complex tabular information across diverse real-world scenarios.

Ethics Statement

In the construction and evaluation of the T2R-Bench, we rigorously adhered to established ethical guidelines for responsible AI research.

Data Collection and Privacy. All datasets utilized in this study were sourced from publicly available repositories with potential private and sensitive information eliminated.

Annotator Compensation and Instruction. Our annotation team comprises 24 annotators, with 12 native English speakers and 12 native Chinese speakers, selected from individuals with extensive experience in analyzing tabular data and demonstrated proficiency in writing analytical reports in relevant fields. We ensure fair compensation for all human annotators, paying each annotator a compensation of $40 per day, with specialized experts receiving an additional 20% premium in recognition of their advanced skills. All annotation work is conducted voluntarily with informed consent, and participants were fully aware of the research objectives and data usage policies.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (62176029, 62506050), China Postdoctoral Science Foundation Funded Project (2024M763867), Chongqing Higher Education Teaching Reform Research Project (No. 242009).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi

Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. *Preprint*, arXiv:2210.06710.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *CoRR*.

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. *Preprint*, arXiv:1909.02164.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. Hitab: A hierarchical table dataset for question answering and natural language generation. *Preprint*, arXiv:2108.06712.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Muzhi Dai, Shixuan Liu, Zhiyuan Zhao, Junyu Gao, Hao Sun, and Xuelong Li. 2025a. Secure tug-of-war (sectow): Iterative defense-attack training with reinforcement learning for multimodal model security. *arXiv preprint arXiv:2507.22037*.

Muzhi Dai, Jiashuo Sun, Zhiyuan Zhao, Shixuan Liu, Rui Li, Junyu Gao, and Xuelong Li. 2025b. From captions to rewards (carevl): Leveraging large language model experts for enhanced reward modeling in large vision-language models. *arXiv preprint arXiv:2503.06260*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 19 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 37 others. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 15 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Jorge Osés Grijalba, L Alfonso Urena Lopez, Eugenio Martínez-Cámara, and Jose Camacho-Collados. 2024. Question answering over tabular data with databench: A large-scale empirical evaluation of llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13471–13488.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming – the rise of code intelligence. *Preprint*, arXiv:2401.14196.

Xinyi He, Mengyu Zhou, Xinrun Xu, Xiaojun Ma, Rui Ding, Lun Du, Yan Gao, Ran Jia, Xu Chen, Shi Han, Zejian Yuan, and Dongmei Zhang. 2024. Text2analysis: A benchmark of table question answering with advanced data analysis and unclear queries. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18206–18215. AAAI Press.

Xueyu Hu, Ziyu Zhao, Shuang Wei, Ziwei Chai, Qianli Ma, Guoyin Wang, Xuwu Wang, Jing Su, Jingjing Xu, Ming Zhu, Yao Cheng, Jianbo Yuan, Jiwei Li, Kun Kuang, Yang Yang, Hongxia Yang, and Fei Wu. 2024a. Infiagent-dabench: Evaluating agents on data analysis tasks. *Preprint*, arXiv:2401.05507.

Xueyu Hu, Ziyu Zhao, Shuang Wei, Ziwei Chai, Guoyin Wang, Xuwu Wang, Jing Su, Jingjing Xu, Ming Zhu, Yao Cheng, Jianbo Yuan, Kun Kuang, Yang Yang, Hongxia Yang, and Fei Wu. 2024b. Infiagent-dabench: Evaluating agents on data analysis tasks. *CoRR*, abs/2401.05507.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, and 5 others. 2024. Qwen2.5-coder technical report. *Preprint*, arXiv:2409.12186.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *CoRR*.

Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2021. Ait-qa: Question answering dataset over complex tables in the airline industry. *Preprint*, arXiv:2106.12944.

Remi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *Preprint*, arXiv:1603.07771.

Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim, and Jaewoo Kang. 2024. ETHIC: evaluating large language models on long-context tasks with high information coverage. *CoRR*, abs/2410.16848.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv: 2411.16594*.

Zheng Li, Yang Du, Mao Zheng, and Mingyang Song. 2024b. Mimotable: A multi-scale spreadsheet benchmark with meta operations for table reasoning.

Zhongqiu Li, Shiquan Wang, Ruiyu Fang, Mengjiao Bao, Zhenhe Wu, Shuangyong Song, Yongxiang Li, and Zhongjiang He. 2025. Mr-uie: Multi-perspective reasoning with reinforcement learning for universal information extraction. *arXiv preprint arXiv:2509.09082*.

Zhongqiu Li, Zhenhe Wu, Mengxiang Li, Zhongjiang He, Ruiyu Fang, Jie Zhang, Yu Zhao, Yongxiang Li, Zhoujun Li, and Shuangyong Song. 2024c. Scalable database-driven kgs can help text-to-sql. In *Proceedings of the ISWC 2024 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 23nd International Semantic Web Conference (ISWC 2024), Hanover, Maryland, USA, November 11-15, 2024*, volume 3828 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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

Weizheng Lu, Jiaming Zhang, Jing Zhang, and Yueguo Chen. 2024. Large language model for table processing: A survey. *CoRR*, abs/2402.05121.

Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. InsightPilot: An LLM-empowered automated data exploration system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 346–352.

Zeyao Ma, Bohan Zhang, Jing Zhang, Jifan Yu, Xiaokang Zhang, Xiaohan Zhang, Sijia Luo, Xi Wang, and Jie Tang. 2024. Spreadsheetbench: Towards challenging real world spreadsheet manipulation. *Preprint*, arXiv:2406.14991.

Puneet Mathur, Alexa Siu, Nedim Lipka, and Tong Sun. 2024. MATSA: Multi-agent table structure attribution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 250–258, Miami, Florida, USA. Association for Computational Linguistics.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir Radev. 2021. Fetaqa: Free-form table question answering. *Preprint*, arXiv:2104.00369.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jorge Osés Grijalba, L. Alfonso Ureña-López, Eugenio Martínez Cámara, and Jose Camacho-Collados. 2024. Question answering over tabular data with DataBench: A large-scale empirical evaluation of LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13471–13488, Torino, Italia. ELRA and ICCL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1173–1186. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2015a. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015,* *Beijing, China, Volume 1: Long Papers*, pages 1470–1480. The Association for Computer Linguistics.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, and 1 others. 2023. Code llama: Open foundation models for code. *CORR*.

Jiawei Shao and Xuelong Li. 2025. Ai flow at the network edge. *IEEE Network*.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, and 14 others. 2024. Tablegpt2: A large multimodal model with tabular data integration. *Preprint*, arXiv:2411.02059.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets LLM: can large language models understand structured table data? A benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, pages 645–654. ACM.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI 2025, Cagliari, Italy, March 24-27, 2025*, pages 952–966. ACM.

Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gerstein. 2024. Struc-bench: Are large language models good at generating complex structured tabular data? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 12–34, Mexico City, Mexico. Association for Computational Linguistics.

MiniCPM Team, Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, and 1 others. 2025. Minicpm4: Ultra-efficient llms on end devices. *arXiv preprint arXiv:2506.07900*.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2024a. Boosting LLM agents with recursive contemplation for effective deception handling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9909–9953, Bangkok, Thailand. Association for Computational Linguistics.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Shiquan Wang, Ruiyu Fang, Mengxiang Li, Zhongjiang He, and Shuangyong Song. When less is more: Minimal prompts with lora for llm text detection. In *The 14th CCF International Conference on Natural Language Processing and Chinese Computing*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, Yan Wang, Xin Wang, Luwen Pu, Huihan Xu, Ruiyu Fang, Yu Zhao, Jie Zhang, Xiaomeng Huang, Zhilong Lu, and 17 others. 2024b. Telechat technical report. *Computing Research Repository*.

Zihan Wang, XinZhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Mengxiang Li, Zhongjiang He, Yongxiang Li, Luwen Pu, Huinan Xu, Chao Wang, and Shuangyong Song. 2024c. Telechat: An open-source billingual large language model. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*.

Zihan Wang, Xinzhang Liu, Yitong Yao, Chao Wang, Yu Zhao, Zhihao Yang, Wenmin Deng, Kaipeng Jia, Jiaxin Peng, Yuyao Huang, and 1 others. 2025b. Technical report of telechat2, telechat2. 5 and t1. *arXiv preprint arXiv:2507.18013*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and 1 others. 2024d. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Kaiwen Wei, Xian Sun, Zequn Zhang, Li Jin, Jingyuan Zhang, Jianwei Lv, and Zhi Guo. 2023a. Implicit event argument extraction with argument-argument relational knowledge. *IEEE Trans. Knowl. Data Eng.*, 35(9):8865–8879.

Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Zhi Guo, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. 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 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4672–4682. Association for Computational Linguistics.

Kaiwen Wei, Yiran Yang, Li Jin, Xian Sun, Zequn Zhang, Jingyuan Zhang, Xiao Li, Linhao Zhang, Jintao Liu, and Zhi Guo. 2023b. Guide the many-to-one assignment: Open information extraction via iou-aware optimal transport. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4971–4984. Association for Computational Linguistics.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. *Preprint*, arXiv:1707.08052.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, and 1 others. 2024. Tablebench: A comprehensive and complex benchmark for table question answering. *arXiv preprint arXiv:2408.09174*.

Zhenhe Wu, Zhongqiu Li, Mengxiang Li, Jie Zhang, Zhongjiang He, Jian Yang, Yu Zhao, Ruiyu Fang, Yongxiang Li, Zhoujun Li, and Shuangyong Song. 2025a. MR-SQL: multi-level retrieval enhances inference for llm in text-to-sql. *DASFAA*.

Zhenhe Wu, Zhongqiu Li, Jie Zhang, Zhongjiang He, Jian Yang, Yu Zhao, Ruiyu Fang, Bing Wang, Hongyan Xie, Shuangyong Song, and Zhoujun Li. 2025b. UCS-SQL: uniting content and structure for enhanced semantic bridging in text-to-sql. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 8156–8168. Association for Computational Linguistics.

Zhenhe Wu, Jian Yang, Jiaheng Liu, Xianjie Wu, Changzai Pan, Jie Zhang, Yu Zhao, Shuangyong Song, Yongxiang Li, and Zhoujun Li. 2025c. Table-r1: Region-based reinforcement learning for table understanding. *arXiv preprint arXiv:2505.12415*.

Hongrui Xing, Xinzhang Liu, Zhuo Jiang, Zhihao Yang, Yitong Yao, Zihan Wang, Wenmin Deng, Chao Wang, Shuangyong Song, Wang Yang, and 1 others. 2025. Llmsr@ xllm25: A language model-based pipeline for structured reasoning data construction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 342–350.

Sishi Xiong, Ziyang He, Zhongjiang He, Yu Zhao, Changzai Pan, Jie Zhang, Zhenhe Wu, Shuangyong Song, and Yongxiang Li. 2025a. Tablezoomer: A collaborative agent framework for large-scale table question answering. *arXiv preprint arXiv:2509.01312*.

Sishi Xiong, Mengxiang Li, Dakai Wang, Yu Zhao, Jie Zhang, Changzai Pan, Haowei He, Xiangyu Li, Wenhan Chang, Zhongjiang He, and 1 others. 2025b. Teleai at semeval-2025 task 8: Advancing table reasoning framework with large language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1828–1841.

Sishi Xiong, Dakai Wang, Yu Zhao, Jie Zhang, Changzai Pan, Haowei He, Xiangyu Li, Wenhan Chang, Zhongjiang He, Shuangyong Song, and 1 others. 2025c. Tablereasoner: Advancing table reasoning framework with large language models. *arXiv preprint arXiv:2507.08046*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *CoRR*, abs/2407.10671.

Yajing Yang, Qian Liu, and Min-Yen Kan. 2024b. DataTales: A benchmark for real-world intelligent data narration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10764–10788, Miami, Florida, USA. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 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*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Deji Zhao, Donghong Han, Jia Wu, Zhongjiang He, Bo Ning, Ye Yuan, Yongxiang Li, Chao Wang, and Shuangyong Song. 2025. Enhancing math reasoning ability of large language models via computation logic graphs. *Knowledge-Based Systems*, page 113905.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao

Liu, Wengang Zhou, and 1 others. 2024. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37:7185–7212.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 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 (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Examples of T2R-bench

A.1 English Table Example with Report Generated by the Single LLM

This subsection shows an example of a <question, table, report keypoints> triple, with report generated through the single LLM method of Qwen2.5-72B-Instruct. The incorrect parts in the report have been highlighted in red.

Question

Formulate an insightful report entitled "Handler Effect Evaluation", examining the efficiency of various handlers in managing stock entries and departures as recorded in January 2020's data.

Table

Serial Number	Date	Product Name	ID	Stock In		Stock Out		Real-time Inventory	Handler
				Quantity	Supplier	Quantity	Supplier		
1	1/10/20	Name1	mg-022	16	Supplier01	8	Supplier01	8	Wang Jing
2	1/11/20	Name2	mg-023	62	Supplier02	51	Supplier02	11	Wang Jing
3	1/12/20	Name3	mg-024	23	Supplier03	15	Supplier03	8	Wang Jing
4	1/13/20	Name4	mg-025	126	Supplier04	85	Supplier04	41	Li Kaihua
5	1/14/20	Name5	mg-026	103	Supplier05	55	Supplier05	48	Li Kaihua
6	1/15/20	Name6	mg-027	85	Supplier06	27	Supplier06	58	Li Kaihua
7	1/16/20	Name7	mg-028	54	Supplier07	38	Supplier07	16	Li Kaihua

Table 7: Inventory Management Table

Report Reference Keypoints

Keypoint 1: Wang Jing's average incoming quantity is 33.67, and the average outgoing quantity is 24.67, indicating a gradual increase in inventory, but the final inventory level remains stable.

Keypoint 2: Li Kaihua's average incoming quantity is 92.0, and the average outgoing quantity is 51.25. Despite the significantly higher incoming quantity compared to the outgoing quantity, the actual inventory level has decreased in practice.

Keypoint 3: Wang Jing's inventory management is more volatile, with an average real-time inventory of 9.0, indicating a lower and relatively stable inventory level.

Keypoint 4: Li Kaihua's inventory activities are more active, with an average real-time inventory of 40.75, indicating a higher inventory level.

Keypoint 5: Wang Jing's real-time inventory data for the earliest and latest dates processed is 8, indicating that the inventory level remained stable during the processing period with no significant changes.

Keypoint 6: Li Kaihua's initial inventory was 41.0, and the final inventory decreased to 16.0, indicating a reduction in actual inventory levels, possibly due to unrecorded outgoing activities or other factors.

Keypoint 7: Wang Jing's inventory management is relatively stable, but the inventory level is low, indicating a more conservative management strategy.

Keypoint 8: Li Kaihua's inventory management strategy is more active, with a higher average incoming quantity, but the actual inventory level has decreased, indicating a potentially more aggressive management strategy.

Generated Report

Title: Handler Effect Evaluation

Macro Description
This report aims to evaluate the efficiency of different handlers in managing inventory in and out during the period from January 10, 2020 to January 16, 2020. The data covers 7 inventory records involving 7 different products, and each record includes information such as date, product name, product ID, inbound quantity, outbound quantity, real-time inventory, and handler. By analyzing these data, we can understand the performance of different handlers in inventory management, thereby providing a basis for optimizing the inventory management process.

Overall Situation of Inbound and Outbound
During the period from January 10, 2020 to January 16, 2020, a total of 7 inventory operations were recorded, involving 7 different products. Overall, in these 7 operations, the total inbound quantity was 460 pieces, the total outbound quantity was 274 pieces, and the final real-time inventory was 186 pieces. Specifically, the inbound quantity on January 10 was 16 pieces, and the outbound quantity was 8 pieces; the inbound quantity on January 11 was 62 pieces, and the outbound quantity was 51 pieces; the inbound quantity on January 12 was 23 pieces, and the outbound quantity was 15 pieces; the inbound quantity on January 13 was 126 pieces, and the outbound quantity was 85 pieces; the inbound quantity on January 14 was 103 pieces, and the outbound quantity was 55 pieces; the inbound quantity on January 15 was 85 pieces, and the outbound quantity was 27 pieces; the inbound quantity on January 16 was 54 pieces, and the outbound quantity was 38 pieces. These data indicate that, overall, the inbound quantity is slightly higher than the outbound quantity, and the inventory has increased.

Handler Efficiency Analysis
In the 7 inventory operations, the handlers are mainly divided into two: Wang Jing and Li Kaihua. Wang Jing handled 3 operations, and Li Kaihua handled 4 operations. Specifically, in the 3 operations handled by Wang Jing, the total inbound quantity was 91 pieces, the total outbound quantity was 74 pieces, and the final real-time inventory was 17 pieces. In the 4 operations handled by Li Kaihua, the total inbound quantity was 369 pieces, the total outbound quantity was 200 pieces, and the final real-time inventory was 169 pieces. From these data, it can be seen that the inbound and outbound quantities handled by Li Kaihua are much higher than those handled by Wang Jing, and the final real-time inventory is also significantly higher than that of Wang Jing. This indicates that Li Kaihua is more efficient in handling a large number of inventory operations.

Specific Performance of Handlers
Further analysis of the specific performance of Wang Jing and Li Kaihua reveals some interesting phenomena. In the 3 operations handled by Wang Jing, the inbound and outbound quantities of each operation are relatively small, and the real-time inventory changes after each operation are not significant. For example, the inbound quantity on January 10 was 16 pieces, the outbound quantity was 8 pieces, and the real-time inventory was 8 pieces; the inbound quantity on January 11 was 62 pieces, the outbound quantity was 51 pieces, and the real-time inventory was 11 pieces; the inbound quantity on January 12 was 23 pieces, the outbound quantity was 15 pieces, and the real-time inventory was 8 pieces. This indicates that Wang Jing performs stably when handling small-batch inventory operations, but has relatively weak ability to handle large-batch inventory operations.

In contrast, in the 4 operations handled by Li Kaihua, the inbound and outbound quantities of each operation are relatively large, and the real-time inventory changes after each operation are obvious. For example, the inbound quantity on January 13 was 126 pieces, the outbound quantity was 85 pieces, and the real-time inventory was 41 pieces; the inbound quantity on January 14 was 103 pieces, the outbound quantity was 55 pieces, and the real-time inventory was 48 pieces; the inbound quantity on January 15 was 85 pieces, the outbound quantity was 27 pieces, and the real-time inventory was 58 pieces; the inbound quantity on January 16 was 54 pieces, the outbound quantity was 38 pieces, and the real-time inventory was 16 pieces. This indicates that Li Kaihua is more efficient in handling large-batch inventory operations and can complete tasks quickly and accurately.

Comparison of Handler Efficiency
In order to more intuitively compare the efficiency of Wang Jing and Li Kaihua, we can calculate the processing time of each operation. Assuming that the processing time of each operation is the same, the average inbound quantity of Wang Jing in handling 3 operations is 30.33 pieces, and the average outbound quantity is 24.67 pieces; the average inbound quantity of Li Kaihua in handling 4 operations is 92.25 pieces, and the average outbound quantity is 50 pieces. From these data, it can be seen that the average inbound and outbound quantities of Li Kaihua in each operation are significantly higher than those of Wang Jing, which further confirms the efficiency of Li Kaihua in handling large-batch inventory operations.

In addition, we can also calculate the real-time inventory change after each operation. In the 3 operations handled by Wang Jing, the real-time inventory changes after each operation are 8 pieces, -41 pieces, and -7 pieces, respectively, with an average change of -13.33 pieces; in the 4 operations handled by Li Kaihua, the real-time inventory changes after each operation are 41 pieces, -7 pieces, 31 pieces, and -22 pieces, respectively, with an average change of 10.75 pieces. This indicates that Li Kaihua can better maintain the stability of inventory after each operation, while Wang Jing, although performing stably when handling small-batch inventory operations, may cause greater inventory fluctuations when handling large-batch inventory operations.

Summary and Recommendations
In summary, through the analysis of inventory management data from January 10, 2020 to January 16, 2020, we can draw the following conclusions: Li Kaihua is more efficient in handling large-batch inventory operations, can complete tasks quickly and accurately, and can better maintain the stability of inventory after each operation. In contrast, Wang Jing performs stably when handling small-batch inventory operations, but may cause greater inventory fluctuations when handling large-batch inventory operations.

Based on the above analysis, we make the following recommendations:
1. **Optimize personnel allocation**: When handling large-batch inventory operations, prioritize Li Kaihua for operation to improve overall efficiency and inventory stability. For small-batch inventory operations, Wang Jing can continue to be responsible to maintain the stability and accuracy of the operation.
2. **Training and improvement**: In view of the shortcomings of Wang Jing in handling large-batch inventory operations, special training can be arranged to improve her ability to handle large-batch inventory operations. At the same time, regular performance evaluations can also be used to motivate handlers to continuously improve their skills.
3. **Process optimization**: In the inventory management process, more automated tools and technologies can be introduced to reduce the error of manual operations and improve overall efficiency. For example, an inventory management system can be used to automatically record and update inventory data to reduce the burden of manual operations.

Through the above measures, the inventory management process can be further optimized, the overall efficiency and inventory stability can be improved, and better support can be provided for the operation of the enterprise.

A.2 Chinese Table Example with Report Generated by the Single LLM

This subsection shows an example of a <question, table, report keypoints> triple, with report generated through the single LLM of Qwen2.5-72B-Instruct. The incorrect parts in the report have been highlighted in red.

Question

统计不同地区农村家庭卫生设施差异，形成《农村卫生设备区域对比》，重点探讨水冲式厕所在各区域的普及程度。

Table

	饮用水情况 单位：% 全国	饮用水情况 单位：% 东部地区	饮用水情况 单位：% 中部地区	饮用水情况 单位：% 西部地区	饮用水情况 单位：% 东北地区
获取饮用水困难住户的比重	10.3	2.8	9.6	22.2	1.3
使用管道水住户的比重	48.6	71.1	28.5	42.9	41.3
净化处理过的饮用水	23.1	44.2	8.8	13.3	15.0
深井水	41.8	37.6	53.4	27.5	75.9
浅井水	27.8	15.3	31.7	43.6	9.0
江河湖水	2.8	1.4	2.0	6.0	0.1
池塘水	1.4	0.4	1.6	2.6	0.0
雨水	1.4	0.1	1.1	3.7	0.0
其他水	1.7	1.0	1.4	3.3	0.0

Table 8: 农村生活条件与设备普及率统计。

Report Reference Keypoints

关键点1：全国农村家庭中，仅有12.8%的家庭拥有水冲式厕所，旱厕和简易厕所或无厕所的家庭占比87.2%。
关键点2：东部地区水冲式厕所普及率最高，为26.0%，而东北地区最低，仅为1.3%。
关键点3：中部地区旱厕最为普遍，占54.1%。
关键点4：西部地区农村家庭中，简易厕所或无厕所的比例最高，达到53.2%，水冲式厕所比例仅为6.2%。

Generated Report

《农村卫生设备区域对比》

宏观描述
本报告基于全国及各地区农村家庭卫生设施的数据，重点探讨了水冲式厕所在不同区域的普及程度。数据显示，全国范围内水冲式厕所的普及率仅为12.8%，而不同地区的普及率差异显著。东部地区的普及率最高，达到26.0%，而东北地区的普及率最低，仅为1.3%。这一差异反映了各地区在卫生设施建设和管理上的不同水平，也揭示了农村卫生设施改善的迫切需求。

东部地区：水冲式厕所普及率最高
东部地区的水冲式厕所普及率达到了26.0%，远高于全国平均水平。这一高普及率的背后，是东部地区相对发达的经济条件和较高的居民生活水平。东部地区的农村家庭不仅在卫生设施上投入更多，而且在其他生活设施方面也表现出较高的普及率。例如，东部地区的彩电普及率为97.5台/百户，固定电话普及率为68.2部/百户，手机普及率为86.1部/百户，电脑普及率为4.8台/百户，摩托车普及率为50.9辆/百户，生活用汽车普及率为5.1辆/百户。这些数据表明，东部地区的农村家庭在基础设施和生活条件方面整体较好，为水冲式厕所的普及提供了良好的基础。

此外，东部地区的饮用水条件也较为优越。使用管道水的住户比例为71.1%，远高于全国平均水平的48.6%。净化处理过的饮用水的使用比例为44.2%，同样高于全国平均水平的23.1%。这些数据进一步说明，东部地区的农村家庭在卫生和生活条件方面具有明显优势，为水冲式厕所的普及提供了有力支持。

中部地区：水冲式厕所普及率较低
中部地区的水冲式厕所普及率仅为6.0%，远低于全国平均水平。这一低普及率反映了中部地区在卫生设施建设和管理上的不足。尽管中部地区的经济条件相对较好，但卫生设施的普及率仍然较低。中部地区的彩电普及率为85.1台/百户，固定电话普及率为45.9部/百户，手机普及率为68.7部/百户，电脑普及率为0.9台/百户，摩托车普及率为36辆/百户，生活用汽车普及率为2.6辆/百户。这些数据表明，中部地区的农村家庭在基础设施和生活条件方面虽然有所改善，但仍存在较大差距。

中部地区的饮用水条件也较为一般。使用管道水的住户比例为28.5%，低于全国平均水平的48.6%。净化处理过的饮用水的使用比例为8.8%，远低于全国平均水平的23.1%。这些数据说明，中部地区的农村家庭在卫生和生活条件方面仍需进一步提升，特别是在水冲式厕所的普及方面。

西部地区：水冲式厕所普及率略高于中部
西部地区的水冲式厕所普及率为6.2%，略高于中部地区的6.0%，但仍低于全国平均水平。西部地区的经济条件相对较为落后，基础设施建设相对滞后。西部地区的彩电普及率为74.5台/百户，固定电话普及率为35.1部/百户，手机普及率为52.3部/百户，电脑普及率为0.6台/百户，摩托车普及率为25.6辆/百户，生活用汽车普及率为2.2辆/百户。这些数据表明，西部地区的农村家庭在基础设施和生活条件方面存在较大差距，特别是在卫生设施方面。

西部地区的饮用水条件也较为落后。使用管道水的住户比例为42.9%，低于全国平均水平的48.6%。净化处

理过的饮用水的使用比例为13.3%，略高于中部地区的8.8%，但仍低于全国平均水平的23.1%。这些数据说明，西部地区的农村家庭在卫生和生活条件方面仍需进一步改善，特别是在水冲式厕所的普及方面。

东北地区：水冲式厕所普及率最低
东北地区的水冲式厕所普及率仅为1.3%，是全国最低的。这一低普及率反映了东北地区在卫生设施建设和管理上的严重不足。尽管东北地区的经济条件相对较好，但卫生设施的普及率仍然较低。东北地区的彩电普及率为97.1台/百户，固定电话普及率为64.4部/百户，手机普及率为63.7部/百户，电脑普及率为1台/百户，摩托车普及率为34.3辆/百户，生活用汽车普及率为2.6辆/百户。这些数据表明，东北地区的农村家庭在基础设施和生活条件方面虽然有所改善，但仍存在较大差距。

东北地区的饮用水条件也较为一般。使用管道水的住户比例为41.3%，略低于全国平均水平的48.6%。净化处理过的饮用水的使用比例为15.0%，略高于全国平均水平的23.1%。这些数据说明，东北地区的农村家庭在卫生和生活条件方面仍需进一步提升，特别是在水冲式厕所的普及方面。

B Comparison of T2R bench and Other Tabular Datasets

A more detailed version of Table 1 for comparison with existing datasets on table types, answer lengths, domains and language is shown in Table B. As shown in the table, our proposed T2R dataset covers diverse table types, including those with large-scale, multi-table, and complex layouts, as well as report-level responses.

C Implementation Details for Benchmark Construction

C.1 Details for Annotation Team Composition

We recruit a total of 24 annotators in three batches, evenly split between native Chinese and English speakers. All annotators hold Master's degrees and have at least one year of experience in data analysis and report writing.

The first group of 12 annotators focuses on dataset construction and annotation introduced in Section 3, including six senior specialists with domain-specific report writing experience across six distinct fields in the dataset. These senior members serve as quality control reviewers, conducting final verification of annotations to ensure accuracy and consistency throughout the dataset development process.

The second group comprises six evaluators responsible for human evaluation of generated reports introduced in Section 5.3. This team receive comprehensive training through virtual meetings to establish unified evaluation criteria, enabling them to systematically annotate and score reports based on predefined quality metrics while maintaining inter-rater reliability.

The third group contains six independent report writers who manually create reference reports serving as the human baseline introduced in Section 5.3. This isolated team operates without exposure to the dataset construction details or evaluation protocols, ensuring an objective performance baseline by preventing any potential information leakage that might influence their writing outputs.

All annotators work eight hours a day and earned a wage of $40 per day on average, with specialized experts receiving an additional 20% premium. All annotators are trained through videos or online meetings provided with annotation guidelines that explains the data usage for academic research purposes.

C.2 Details of Procedure for Question Annotation

We randomly assign each question to two annotators, whose selection criteria and qualifications are detailed in Section C.1.

Each annotator assesses the quality of question candidate based on the following aspects: **a) scope compliance**: the question must be answerable using tabular data, without requiring any extraneous domain knowledge. Temporal and spatial references must be strictly confined within the boundaries of the dataset. **b) thematic focus**: the question should concentrate on a single analytical dimension to derive evidence-bound conclusions, rather than enabling the generation of multi-thematic reports across divergent analytical directions. **c) conceptual distinctiveness**: multiple questions derived from the same table must address non-overlapping thematic aspects with clearly differentiated analytical objectives.

In cases where the evaluation results of the two annotators are inconsistent, the results will be handed over to a third annotator for the final judgment. Through this rigorous quality assurance procedure, we obtained 910 high-quality, comprehensive questions.

C.3 Details of Procedure for Keypoints Annotation

Similarly to question annotation, each <table, question> pair and corresponding three groups of extracted report key points is assigned to two independent annotators for revision. However, more com-

Task and Benchmark	Multiple Table	Complex Structure Table	Extremely Large-Size Table	Answer Lengths	Labelled Domains	Language Category
TableQA						
WikiSQL (Zhong et al., 2017)	×	×	×	1.9	-	en
WTQ (Pasupat and Liang, 2015b)	×	×	×	10.39	-	en
TAT-QA (Zhu et al., 2021)	×	×	×	20.3	1	en
FeTaQA (Nan et al., 2021)	×	×	×	18.9	1	en
AIT (Katsis et al., 2021)	×	×	×	1.1	1	en
TabFact (Chen et al., 2020)	×	×	×	18.3	-	en
TableBench (Wu et al., 2024)	×	×	×	8.5	18	en
HiTab (Cheng et al., 2022)	×	✓	×	12.9	9	en
DataBench (Grijalba et al., 2024)	×	×	✓	3.2	8	en
Mimo (Li et al., 2024b)	✓	✓	×	44.2	7	zh, en
Spider (Yu et al., 2018)	✓	×	✓	35.5	138	en
Table2Text						
ToTTo (Parikh et al., 2020)	×	×	✓	17.4	44	en
DAE-val (Hu et al., 2024a)	×	×	✓	3.6	9	en
DataTales (Yang et al., 2024b)	×	×	✓	108.0	1	en
Text2Analysis (He et al., 2024)	×	×	×	-	-	en
Table2Report						
T2R-Bench (ours)	✓	✓	✓	950.2	20	zh, en

Table 9: Complete version of Table 1 for comparison with existing datasets on table types and answer lengths. Since Text2Analysis benchmark dose not provide the publicly accessible download links, the average length could not be calculated. Datasets without an explicit domain classification are denoted by the value '-' in the table.

plicated than binary validity judgments in question annotation, key point annotation requires multi-dimensional modifications including summarization, deletion, insertion and polishing based on AI-generated report keypoints. The annotation of key points adheres to the following criteria: 1) Factual Accuracy: The keypoints must be derived from and accurately reflect the data presented in the tables. 2) Relevance: The keypoints must align with the question of the report generation. 3) Essentiality: The key points should encompass the core content necessary to address the report's objectives. 4) Consistency: The key points should be logically coherent, non-repetitive, and form a cohesive narrative.

The results of two annotators are assigned to the third annotator for justification. If the third annotator finds the two annotations to be consistent or very similar, they will make minor adjustments and approve it as the final core point. However, if the third annotator identifies significant discrepancies between the two annotations, the issue will be documented and discussed during the daily meeting to reach a consensus with the other two annotators.

C.4 Prompts Library and Seed Questions for Question Generation

The five prompt templates in the prompt library for question generation are shown below:

```
As an expert with extensive experience in data analysis and report writing, you are required to propose questions for generating reports from multiple different perspectives along with specific requirements, based on the table description uploaded. The questions must be detailed enough and ensure there is sufficient differentiation among the questions.

## Response Format:

    Question 1:...
    Question 2:...
    Question 3:...

## Input Table Descriptions:
[TABLE DESCRIPTION]

Please directly output the generated 3 questions, do not include any additional explanations or comments.
```

```
As an expert in table structure comprehension and narrative synthesis, develop three questions that cover different investigative angles, such as ratio and share analysis, comparative growth rates, and anomaly flagging—detailing the fields involved, the analytical approach.

## Response Format:

    Question 1:...
    Question 2:...
    Question 3:...

## Input Table Descriptions:
[TABLE DESCRIPTION]

Please directly output the generated 3 questions, do not include any additional explanations or comments.
```

Drawing on your ability to unpick multidimensional tables and deliver actionable insights, craft three report questions that each emphasize a unique focus—layered segmentation, extreme-value exploration, and temporal dynamics—while clarifying which metrics to calculate, over what time or dimension range, and the expected outcome for managerial decision support.

## Response Format:

Question 1:...
Question 2:...
Question 3:...

## Input Table Descriptions:
[TABLE DESCRIPTION]

Please directly output the generated 3 questions, do not include any additional explanations or comments.

As a specialist skilled in dissecting complex tables and translating data into clear narratives, design three probing questions that focus respectively on trend analysis, distributional characteristics, and comparative benchmarking; for each, indicate the key metric, the comparison group or baseline, the required data granularity, and the decision-making context.

## Response Format:

Question 1:...
Question 2:...
Question 3:...

## Input Table Descriptions:
[TABLE DESCRIPTION]

Please directly output the generated 3 questions, do not include any additional explanations or comments.

From the perspective of an analyst with deep expertise in interpreting tabular datasets and crafting concise reports, propose three questions that each target a different analytical dimension, such as time series trends, category comparisons, or geographic breakdowns—while specifying the exact fields to use, the calculations or aggregations required.

## Response Format:

Question 1:...
Question 2:...
Question 3:...

## Input Table Descriptions:
[TABLE DESCRIPTION]

Please directly output the generated 3 questions, do not include any additional explanations or comments.

The 10 Seed Questions are shown below:

Seed Question 1:Produce a report entitled ’Analysis of Stock Market Trading Trends in May 2006’, providing a comprehensive examination of monthly fluctuations in both trading volume and transaction amounts.

Seed Question 2:Develop ’Q3 2008 Metals and Fuel Oil Market Dynamics Report’, investigating annual trading value fluctuations and market trends for copper, aluminum, and zinc contracts on SHFE.

Seed Question 3:Analyze September 2023 food and alcohol price variations across China’s major cities, with particular focus on how grain and vegetable price movements impact composite indices.

Seed Question 4:Prepare an in-depth report on ’Model Differentiation Analysis for Shenbei Avenue 4S Stores (July)’, evaluating sales performance and customer preference across vehicle models.

Seed Question 5:Generate a trend analysis report on township hospital bed utilization rates from 2014 to 2022, utilizing comprehensive tabular data.

Seed Question 6:Conduct ’Historical Analysis of Healthcare Personnel Structure (2014-2023)’, tracking growth patterns across medical staff categories with special attention to licensed physicians and registered nurses.

Seed Question 7:Complete ’Human Resource Efficiency in Small and Micro Enterprises’, examining workforce allocation and revenue efficiency across industries using employment and income data.

Seed Question 8:Produce ’Study on Melon Cultivation Structure Transformation (2014-2022)’, detailing area changes for watermelon, muskmelon, strawberry and related crops.

Seed Question 9:Investigate productivity trends in petroleum and natural gas extraction, creating a detailed change analysis report for August 2023 through March 2024.

Seed Question 10:Compile a report titled ’Seasonal Fluctuation Analysis of Beijing Secondary Housing Prices’, conducting year-round data dissection with emphasis on seasonal influencing factors.

C.5 Prompt for Report Keypoints Extraction

As an expert with extensive experience in information extraction, you are required to summarize 5-10 report reference keypoints based on the report I provide.

## Response Format:
Keypoint 1:...
Keypoint 2:...

## Reference Reports:
[REPORTS]

Please directly output the generated 5-10 report reference keypoints, do not include any additional explanations or comments.

C.6 Domain and Sub-domain of T2R-bench

The 6 domains and 19 sub-domains in T2R-bench are shown in Table 10.

Domains	**Sub-domains**

Engineering Science	Electronics and Automation Manufacturing;
	Chemical Engineering and Advanced Materials;
	Energy Production and Power Systems;
	Automotive Manufacturing and Mobility Solutions

Environmental Stewardship	Environmental Protection;
	Agricultural Production and Forestry Management;
	Marine Resources and Fisheries Management

Transportation Logistics	Telecommunications and IT Infrastructure; Transportation Networks and Logistics Management

Social Policy Administration	Education and Scientific Research;
	Government Administration and Public Sector Services;
	Healthcare Systems and Public Health;
	Demographics and Social Development

Consumer Lifestyle	Retail Trade and E-commerce Platforms;
	Tourism and Hospitality Services;
	Food and Beverage Services;
	Business Management

Financial Economics	Economic Development and International Trade; Banking and Financial Services

Table 10: The 6 domains and 19 sub-domains in T2R-bench

C.7 Data Source of T2R-bench

The sources of tabular data in T2R-bench are shown in Table 11.

Sources	**Websites**
Open-source data platform	
Wolrd Bank Group	https://datacatalog.worldbank.org/
National Bureau of Statistics of China	https://www.stats.gov.cn/sj/
Kaggle	https://www.kaggle.com/datasets
China Association of Automobile Manufactures	http://www.caam.org.cn/
Beijing Public Data Open Platform	https://data.beijing.gov.cn/
The United States Government's Open Data Site	https://catalog.data.gov/dataset
China Securities Regulatory Commission Data Platform	http://www.csrc.gov.cn/csrc/tjsj/index.shtml
Shanghai Public Data Open Platform	https://data.sh.gov.cn/view/data-resource/index.html
CelesTrak	https://celestrak.org/
Tabular dataset	
MiMoTable(Li et al., 2024b)	https://github.com/jasonNLP/MiMoTable

Table 11: The data sources of T2R-bench Tables

D Implementation Details for Evaluation Criteria

D.1 Prompts for Numerical Accuracy Criterion

This subsection introduce the details of evaluating numerical accuracy criterion. Firstly, given the report to be evaluate, we extract clusters of sentences with numerical values through using regular expressions. Secondly, we transfer the extracted sentence clusters with numerical statements to inversely generate questions which take these sentence clusters as answers, following the prompt below:

> As an expert in language logic analysis and data recognition, you need to transform the given sentence into question addressing the numerical parts. The questions should inquire about all numerical values appearing in the paragraph, clearly specifying the objects and criteria based on the context.
>
> ## Example:
>
> Input: In July 2023, the Kangming Road 4S store sold 20 Accord, 116 Odyssey, 35 Vezel, 123 CR-V, 43 Lingpai, 163 Fit, and 39 Odyssey units.
>
> Output: How many units of Accord, Odyssey, Vezel, CR-V, Lingpai, Fit, and Odyssey were sold by the Kangming Road 4S store in July 2023?
>
> Input Sentence:
> [SENTENCE]
>
> Please directly output the question, do not include any additional explanations or comments.

Thirdly, we get the answer of each question by prompting three different LLMs' coder versions (Qwen2.5-32B-Coder-Instruct, Deepseek-Coder and CodeLlama-70B-Instruct) to generate python code and extract relative data through Python programming, following the ideas of previous research proposed for Table QA task. If the code execution fails, it will not be included in the final score. The code generation prompt is shown below:

> You are a data analysis assistant. Based on the user's provided analysis question, analytical approach, and file path, generate an efficient and robust Python code snippet to read the file from the specified path and perform data extraction.
>
> ## Data Description (Input):
> [DATA]
>
> ## Specified File Path:
> [FILE PATH]
>
> ## User Query:
> [QUERY]
>
> ## Requirements:
>
> File Reading:
>
> Efficiently read data from the specified path according to the file format and size (supporting CSV, Excel, etc.) and load it into a pandas DataFrame. For larger datasets, choose appropriate reading methods to ensure performance.
>
> Data Processing:
>
> 1. Process data based on the user's requirements and analytical approach, including column selection, conditional filtering, group calculations, etc., ensuring the code results meet the user's query. All computed results should retain two decimal places for precise representation.

2. Analyze the user's query and the execution process of the Python code in the Chain-of-Thought (COT) manner.

3. Limit the number of keys in the output dictionary to within 10, avoiding tuple data types in the output.

4. Avoid using object types and numpy operations; ensure correct computation types during calculations.

5. Ensure all keys and values in the answer conform to dictionary format requirements, with keys being string types and values being strings or dictionaries, not lists or tuples, and convert types as necessary.

6. The generated code should be robust, including error handling and file format compatibility. It should strictly match column names mentioned in the user's query, avoiding irrelevant or mismatched columns.

7. Return variable format: The final result should only include the 'answer' variable in dictionary format, without any other outputs.

## Example:
[PYTHON CODE EXAMPLES]

## Note: Ensure outputs are formatted compactly and effectively, allowing successful loading by Python scripts, and avoid explanatory content.

After obtaining the three sets of answers from Qwen2.5-32B-Coder-Instruct, Deepseek-Coder, and CodeLlama-70B-Instruct, we apply a majority-voting mechanism to aggregate these outputs into the single most reliable result, using the prompt below:

You are a model evaluator who rigorously applies consistency based assessment principles. You excel at analyzing, synthesizing, and summarizing multiple large language model outputs, and under a majority voting scheme to deriving the most coherent and reliable final answer.

## Task Data Description (Input):
[ANSWER 1], [ANSWER 2], [ANSWER 3]

## Requirements:

Answer Grouping:

1. Extract the core response from each model's output, then categorize the model's answers into groups of equivalent meaning. Ensure that stylistic or formatting differences (e.g., "the RPN range is 24 to 24" versus "24.0–24.0") do not lead to separate groupings when the semantic content is identical.

2. Accurately identify semantically consistent answers to prevent fragmentation due to superficial wording variations.

3. You may paraphrase or consolidate responses when summarizing each group.

4. Numeric equivalence rule: If two numeric answers round to the same value, consider them identical; adopt the value with the highest precision as the representative.

Final Answer Determination:

1.Apply a majority voting rule: select the answer supported by the greatest number of models as the final result.

2. If there is a tie for highest votes, output "Unable to Infer" without subjective judgment.

3.If the selected answer is empty, "nan", "0", "0.0", an empty list or empty dictionary, or otherwise non-informative, also output "Unable to Infer".

## Response Format:

The response must strictly follow JSON format, as shown below:

```
{
    "Answer Groups": {
        "Answer 1": ["LLMA", "LLMB"]
        "Answer 2": ["LLMC", "LLMD"]
        ...
    },
    "Final Answer": "The consolidated answer, or
'Unable to Infer'"
}
```

## Example:
[TASK ANSWER EXAMPLES]

## Note: Ensure outputs are formatted compactly and effectively, fully understand the requirements.

Finally, by comparing the derived answers with numerical statements within each sentence cluster, we obtain the final NAC score, using the prompt below:

As an expert in fact verification and logical analysis, your task is to compare a given factual statement against a standard factual answer to determine if there is any contradiction in their numerical data. You should provide a score between 0 and 1, where 1 indicates complete agreement and 0 indicates complete contradiction.

## Note: Focus solely on the numerical portions of the statements. Only output the final score without any intermediate steps or explanations.

```
## Response Format: {
    "score": 1.0,
    "reason": "Reason for the assigned score"
}
```

## Factual Statement to Verify:
[SENTENCE]

## Standard Factual Answer:
[ANSWER]

Please ensure that your response is strictly formatted as a valid JSON object and can be directly parsed by 'json.load()'. Do not include any additional characters, comments, or text outside of the JSON structure."'

D.2 Report Evaluation Aspects for General Evaluation Criterion

Existing evaluation criteria for reports or long texts typically encompass multiple aspects, including relevance, logical coherence, clarity, human-like style,

Aspect	Description
Reasoning Depth	Does the report demonstrate deep and multi-layered reasoning behind its claims? Does the analysis go beyond surface-level observations to reveal underlying mechanisms or causes?
Human-like Style	Does the writing style of the report resemble natural human expression rather than overly structured or mechanical language generated by machines? Do you think it even slightly resembles machine-generated content, or human written content?
Practicality	Are the analyses and recommendations provided in the report practically feasible? Can they offer valuable references to readers? Does the report demonstrate profound industry insights?
Content Completeness	Does the report provide a comprehensive overview of both current status and future opportunities? Are there areas where the report's depth of coverage is insufficient? Where could additional data or examples strengthen the report's coverage?
Logical Coherence	Is the report structured so that each point builds logically on the previous one? Are there any gaps in reasoning or sudden jumps between topics? Do all conclusions follow clearly from the evidence or analysis presented?

Table 12: Evaluation Aspects for General Evaluation Criterion.

innovation, and structural rationality, when using LLMs as a judge(Bai et al., 2024; Zheng et al., 2023; Li et al., 2024a). However, some evaluation aspects, such as linguistic standardization and logical coherence, don not show significant difference across various methods. Therefore, we concentrate on those aspects that can effectively distinguish the quality of different reports, as shown in Table 12.

D.3 Prompt for General Evaluation Criterion

```
You are a professional large-model evaluation expert specializing in assessing the quality of AI-generated reports. We will provide you with a user instruction and the AI-generated report. Your task is to evaluate the AI-generated report according to the following evaluation criteria and scoring rules.

## Evaluation Aspects:
[EVALUATION ASPECTS]

## Scoring Criteria:
    The score ranges from 0 to 10. The intermediate ranges are defined as follows:
    - 10 points: Fully meets requirements, outstanding performance, comprehensive content, no obvious defects.
    - 8–9 points: Strong performance, meets most requirements with only minor flaws, very close to perfect overall.
    - 6–7 points: Some shortcomings or areas that need improvement, yet still generally meets the requirements and provides valuable information or analysis.
    - 4–5 points: Noticeable flaws or omissions; certain requirements are not adequately addressed, negatively affecting overall quality.
    - 0–3 points: Poor quality; fails to satisfy core requirements. Contains serious errors, omissions, or logical confusion that prevent effective communication of information.

## Response Format:
    The response must strictly follow JSON format, as shown below:
    {
        "Evaluation Aspect 1": {
            "Reason": "Explanation for this aspect's rating",
            "Score": <numeric score>
        },
        ...
        "Evaluation Aspect N": {
            "Reason": "Explanation for this aspect's rating",
            "Score": <numeric score>
        }
    }

## Evaluation Steps:
    1.Understand the User Question: Carefully read the user's request. Identify each requirement and how they interrelate.
    2.Analyze the AI-Generated Report: Thoroughly review the report to ensure you understand its content and topic.
    3.Evaluate Each Dimension: Check the report against every dimension in the list. Your evaluation should be both strict and fair, to enable comparison across different models.
    4.Assign Scores and Provide Explanations: Give each dimension a score (0–10) and clearly state the reasons that justify your score.
    5.Output the Final Evaluation: Present your results in JSON format, double-checking for any formatting or syntax errors.

## Example:
[EXAMPLES]

## Input User Question:
[QUESTION]

## Input AI-Generated Report:
[REPORT]

Please ensure that your response is strictly formatted as a valid JSON object and can be directly parsed by ‘json.load()‘. Do not include any additional explanations, comments, or extraneous characters outside of the JSON
```

structure.

E Implementation Details for Experiments

E.1 Prompt Template for Generating Report by LLMs

As an expert with extensive experience in data analysis and report writing, Please generate a comprehensive report based on the provided data and analysis perspectives, and follow these guidelines:

## Report Standards:

> Objectivity: Ensure that the analysis is grounded in the actual data provided by the user. Avoid subjective judgments and ensure accuracy.
> Precision: Each conclusion should be supported by data. Ensure numbers and results are accurate and based solely on the data provided.
> Logic: The structure of the analysis should be clear and logically connected from problem definition to conclusions and recommendations.
> Readability: Present the analysis in a simple and straightforward manner, avoiding overly complex terminology for better understanding by non-specialists.
> Action-Oriented: Beyond just reviewing the data, provide specific suggestions or strategies to support decision-making.
> Variety and Pacing: Use varied language to maintain reader interest and enhance the professionalism and appeal of the analysis.

## Requirements:

> 1. The output report should be complete and well-structured, with a minimum length of 1000 words.
> 2. Ensure content is appropriately detailed, avoiding repetition and vague descriptions.
> 3. Adjust the analysis perspective if there are gaps or incomplete data, rather than mentioning "insufficient data."
> 4. Follow the analysis standards closely, avoid directly applying template references, and tailor the content according to the actual data for proper derivation and summary.

## Input Question for generating report:
[QUESTION]

## Input Table Data for generating report:
[TABLE DATA]

Please directly output the generated report, do not include any additional explanations or comments.

E.2 Analysis of Detailed Case Study

This subsection shows examples of a <question, table, report keypoints, case study> combination to display detailed case study , with report generated through the single LLM of Deepseek-R1 (i.e., it is the best-performing model in our benchmark.). The incorrect parts in the report have been highlighted in red. Since the generated report is quite lengthy, the part of report has been omitted, and only the content that requires case analysis is displayed.

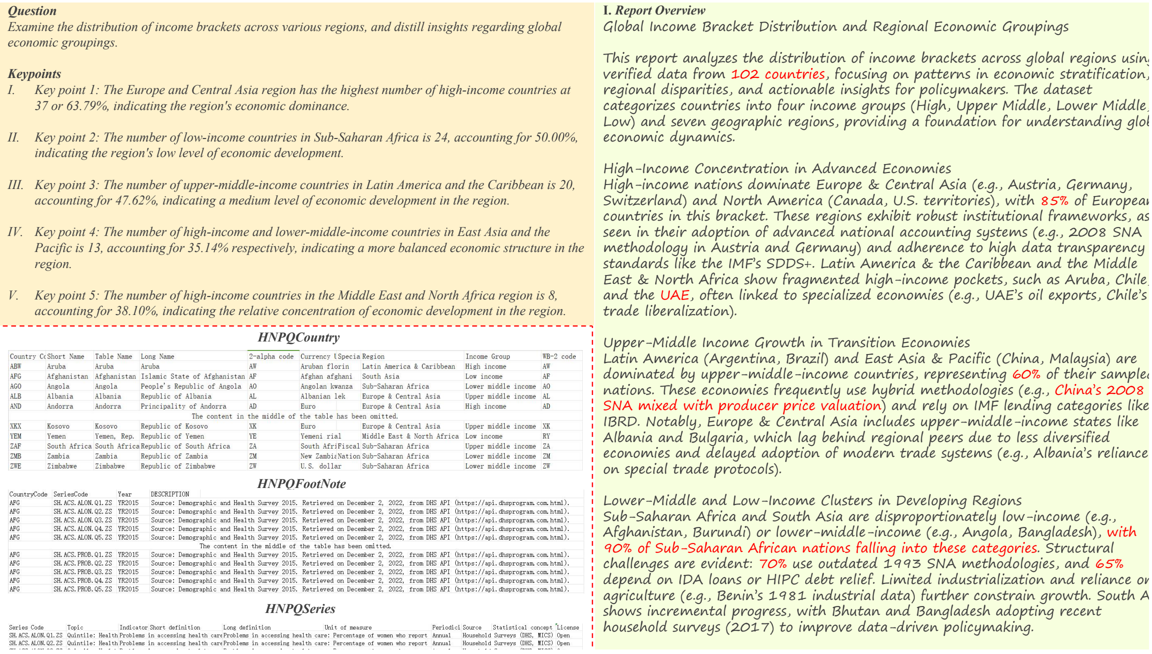
Case Study of English Extremely Large-size Table. As indicated by the highlighted text in red in Figure 6, the number of countries in the sentence "this report analyzes the distribution of income brackets across global regions using verified data from 102 countries" is indeed 217. This may be due to a truncation when inputting extremely large tabular data into LLM. Moreover, the second paragraph doesn't cover Keypoint 4 when analyzing high income concentration and be lack of correct supporting data. This directly reduces the ICC evaluation metric of the report.

Case Study of Chinese Complex Structured Table. For the aforementioned complex structured table shown in Figure 7, which features a complicated header and describes the comparison of sales between this year and last year from January to May, there have been numerous numerical hallucinations and incorrect conclusions. For example: "Data shows that Guangzhou (including Zone 1 and Zone 2) had total sales of ¥5.788 million, Shenzhen ¥4.529 million, and Nanning ¥158,000. The overall discount rates exhibited a 'higher in the south, lower in the north' pattern: Nanning's average discount rate was 0.77, Shenzhen 0.78, and Guangzhou 0.63." Here, Shenzhen's total sales figure was taken from the "January to May cumulative data" rather than the summation of the first quarter's, and the discount rates were also incorrect. These errors, along with challenges posed by complex table structures, descriptive hallucinations, and variable misinterpretations, reveal fundamental reasoning limitations.

E.3 Error Analysis of Samples

As described in the case study section, we conduct an error analysis by randomly selecting 50 samples (with each set of 10 samples representing the typical characteristics of a specific table type).

The primary error types identified are as follows: First, there are **hallucination errors**, which include numerical factual errors (such as incorrect numerical calculations or hallucinations of numbers from the table in the report), generation errors (such as generating content unrelated to the table, or producing incorrect or insufficiently supported conclusions or descriptions), and table structure understanding errors (e.g., column selection errors resulting from misinterpretation of table structures,



Question
Examine the distribution of income brackets across various regions, and distill insights regarding global economic groupings.

Keypoints

I. Key point 1: The Europe and Central Asia region has the highest number of high-income countries at 37 or 63.79%, indicating the region's economic dominance.

II. Key point 2: The number of low-income countries in Sub-Saharan Africa is 24, accounting for 50.00%, indicating the region's low level of economic development.

III. Key point 3: The number of upper-middle-income countries in Latin America and the Caribbean is 20, accounting for 47.62%, indicating a medium level of economic development in the region.

IV. Key point 4: The number of high-income and lower-middle-income countries in East Asia and the Pacific is 13, accounting for 35.14% respectively, indicating a more balanced economic structure in the region.

V. Key point 5: The number of high-income countries in the Middle East and North Africa region is 8, accounting for 38.10%, indicating the relative concentration of economic development in the region.

HNPQCountry

Country Cc	Short Name	Table Name	Long Name	2-alpha code	Currency U	Specia	Region	Income Group	WB-2 code
ABW	Aruba	Aruba	Aruba	AW	Aruban florin		Latin America & Caribbean	High income	AW
AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani		South Asia	Low income	AF
AGO	Angola	Angola	People's Republic of Angola	AO	Angolan kwanza		Sub-Saharan Africa	Lower middle income	AO
ALB	Albania	Albania	Republic of Albania	AL	Albanian lek		Europe & Central Asia	Upper middle income	AL
AND	Andorra	Andorra	Principality of Andorra	AD	Euro		Europe & Central Asia	High income	AD
The content in the middle of the table has been omitted.									
XKX	Kosovo	Kosovo	Republic of Kosovo	XK	Euro		Europe & Central Asia	Upper middle income	XK
YEM	Yemen	Yemen, Rep.	Republic of Yemen	YE	Yemeni rial		Middle East & North Africa	Low income	RY
ZAF	South Africa	South Africa	Republic of South Africa	ZA	South Afri	Fiscal	Sub-Saharan Africa	Upper middle income	ZA
ZMB	Zambia	Zambia	Republic of Zambia	ZM	New Zambia	Nation	Sub-Saharan Africa	Lower middle income	ZM
ZWE	Zimbabwe	Zimbabwe	Republic of Zimbabwe	ZW	U.S. dollar		Sub-Saharan Africa	Lower middle income	ZW

HNPQFootNote

CountryCode	SeriesCode	Year	DESCRIPTION
AFG	SH.ACS.ALON.Q1.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.ALON.Q2.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.ALON.Q3.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.ALON.Q4.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.ALON.Q5.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
The content in the middle of the table has been omitted.			
AFG	SH.ACS.PROB.Q1.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.PROB.Q2.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.PROB.Q3.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.PROB.Q4.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).
AFG	SH.ACS.PROB.Q5.ZS	YR2015	Source: Demographic and Health Survey 2015. Retrieved on December 2, 2022, from DHS API (https://api.dhsprogram.com.html).

HNPQSeries

Series Code	Topic	Indicator	Short definition	Long definition	Unit of measure	Periodici	Source	Statistical concept	License
SH.ACS.ALON.Q1.ZS	Quintile: Health	Problems in accessing health care		Problems in accessing health care: Percentage of women who report		Annual	Household Surveys (DHS, MICS)		Open
SH.ACS.ALON.Q2.ZS	Quintile: Health	Problems in accessing health care		Problems in accessing health care: Percentage of women who report		Annual	Household Surveys (DHS, MICS)		Open

I. ***Report Overview***
Global Income Bracket Distribution and Regional Economic Groupings

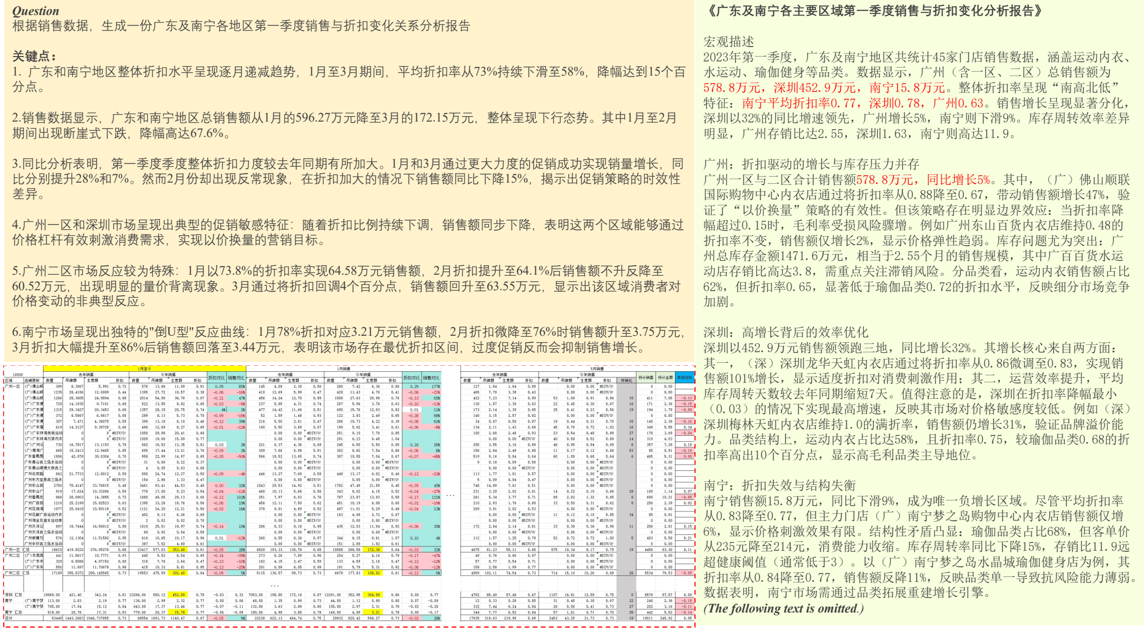
This report analyzes the distribution of income brackets across global regions usin verified data from 102 countries, focusing on patterns in economic stratification, regional disparities, and actionable insights for policymakers. The dataset categorizes countries into four income groups (High, Upper Middle, Lower Middle, Low) and seven geographic regions, providing a foundation for understanding glol economic dynamics.

High-Income Concentration in Advanced Economies
High-income nations dominate Europe & Central Asia (e.g., Austria, Germany, Switzerland) and North America (Canada, U.S. territories), with 85% of Europea countries in this bracket. These regions exhibit robust institutional frameworks, as seen in their adoption of advanced national accounting systems (e.g., 2008 SNA methodology in Austria and Germany) and adherence to high data transparency standards like the IMF's SDDS+. Latin America & the Caribbean and the Middle East & North Africa show fragmented high-income pockets, such as Aruba, Chile, and the UAE, often linked to specialized economies (e.g., UAE's oil exports, Chile's trade liberalization).

Upper-Middle Income Growth in Transition Economies
Latin America (Argentina, Brazil) and East Asia & Pacific (China, Malaysia) are dominated by upper-middle-income countries, representing 60% of their sample nations. These economies frequently use hybrid methodologies (e.g., China's 2008 SNA mixed with producer price valuation) and rely on IMF lending categories like IBRD. Notably, Europe & Central Asia includes upper-middle-income states like Albania and Bulgaria, which lag behind regional peers due to less diversified economies and delayed adoption of modern trade systems (e.g., Albania's reliance on special trade protocols).

Lower-Middle and Low-Income Clusters in Developing Regions
Sub-Saharan Africa and South Asia are disproportionately low-income (e.g., Afghanistan, Burundi) or lower-middle-income (e.g., Angola, Bangladesh), with 90% of Sub-Saharan African nations falling into these categories. Structural challenges are evident: 70% use outdated 1993 SNA methodologies, and 65% depend on IDA loans or HIPC debt relief. Limited industrialization and reliance o agriculture (e.g., Benin's 1981 industrial data) further constrain growth. South A shows incremental progress, with Bhutan and Bangladesh adopting recent household surveys (2017) to improve data-driven policymaking.

Figure 6: Case study of English extremely large-size table



Question
根据销售数据，生成一份广东及南宁各地区第一季度销售与折扣变化关系分析报告

关键点：
1. 广东和南宁地区整体折扣水平呈现逐月递减趋势，1月至3月期间，平均折扣率从73%持续下滑至58%，降幅达到15个百分点。

2.销售数据显示，广东和南宁地区总销售额从1月的596.27万元降至3月的172.15万元，整体呈现下行态势。其中1月至2月期间出现断崖式下跌，降幅高达67.6%。

3.同比分析表明，第一季度季度整体折扣力度较去年同期有所加大。1月和3月通过更大力度的促销成功实现销量增长，同比分别提升28%和7%。然而2月份却出现反常现象，在折扣加大的情况下销售额同比下降15%，揭示出促销策略的时效性差异。

4.广州一区和深圳市场呈现出典型的促销敏感特征：随着折扣比例持续下调，销售额同步下降，表明这两个区域能够通过价格杠杆有效刺激消费需求，实现以价换量的营销目标。

5.广州二区市场反应较为特殊：1月以73.8%的折扣率实现64.58万元销售额，2月折扣提升至64.1%后销售额不升反降至60.52万元，出现明显的量价背离现象。3月通过将折扣回调4个百分点，销售额回升至63.55万元，显示出该区域消费者对价格变动的非典型反应。

6.南宁市场呈现出独特的"倒U型"反应曲线：1月78%折扣对应3.21万元销售额，2月折扣微降至76%时销售额升至3.75万元，3月折扣大幅提升至86%后销售额回落至3.44万元，表明该市场存在最优折扣区间，过度促销反而会抑制销售增长。

《广东及南宁各主要区域第一季度销售与折扣变化分析报告》

宏观描述
2023年第一季度，广东及南宁地区共统计45家门店销售数据，涵盖运动内衣、水运动、瑜伽健身等品类。数据显示，广州（含一区、二区）总销售额为578.8万元，深圳452.9万元，南宁15.8万元。整体折扣率呈现“南高北低”特征：南宁平均折扣率0.77，深圳0.78，广州0.63。销售增长呈现显著分化，深圳以32%的同比增速领先，广州增长5%，南宁则下滑9%。库存周转效率差异明显，广州存销比达2.55，深圳1.63，南宁则高达11.9。

广州：折扣驱动的增长与库存压力并存
广州一区与二区合计销售额578.8万元，同比增长5%。其中，（广）佛山顺联国际购物中心内衣店通过将折扣率从0.88降至0.67，带动销售额增长47%，验证了“以价换量”策略的有效性。但该策略存在明显边界效应：当折扣率降幅超过0.15时，毛利率受损风险骤增。例如广州东山百货内衣店维持0.48的折扣率不变，销售额仅增长2%，显示价格弹性趋弱。库存问题尤为突出：广州总库存金额1471.6万元，相当于2.55个月的销售规模，其中广百百货水运动店存销比高达3.8，需重点关注滞销风险。分品类看，运动内衣销售额占比62%，但折扣率0.65，显著低于瑜伽品类0.72的折扣水平，反映细分市场竞争加剧。

深圳：高增长背后的效率优化
深圳以452.9万元销售额领跑三地，同比增长32%。其增长核心来自两方面：其一，（深）深圳龙华天虹内衣店通过将折扣率从0.86微调至0.83，实现销售额101%增长，显示适度折扣对消费刺激作用；其二，运营效率提升，平均库存周转天数较去年同期缩短7天。值得注意的是，深圳在折扣率降幅最小（0.03）的情况下实现最高增速，反映其市场对价格敏感度较低。例如（深）深圳梅林天虹内衣店维持1.0的满折率，销售额仍增长31%，验证品牌溢价能力。品类结构上，运动内衣占比达58%，且折扣率0.75，较瑜伽品类0.68的折扣率高出10个百分点，显示高毛利品类主导地位。

南宁：折扣失效与结构失衡
南宁销售额15.8万元，同比下滑9%，成为唯一负增长区域。尽管平均折扣率从0.83降至0.77，但主力门店（广）南宁梦之岛购物中心内衣店销售额仅增6%，显示价格刺激效果有限。结构性矛盾凸显：瑜伽品类占比68%，但客单价从235元降至214元，消费能力收缩。库存周转率同比下降15%，存销比11.9远超健康阈值（通常低于3）。以（广）南宁梦之岛水晶城瑜伽健身店为例，其折扣率从0.84降至0.77，销售额反降11%，反映品类单一导致抗风险能力薄弱。数据表明，南宁市场需通过品类拓展重建增长引擎。
(The following text is omitted.)

Figure 7: Case study of Chinese complex structured table

such as selecting wrong column names due to incorrect recognition of complex table headers and structures; cross-table selection errors where the model retrieves data from incorrect tables). Second, there is the issue of **missing key information**, where the generated reports do not fully cover the key points, directly leading to a low ICC evaluation metric. Third, there are columns **truncation errors** when the table content exceeds the context window length (e.g., for an extremely large-size table, miscalculating a column's mean value), which directly results in a low NAC evaluation metric. The statistics of the sampling error analysis are shown in the Table 13.

E.4 URLs of Closed-source Models

E.5 Analysis of Input Formatting

Table 15 demonstrates that among the three most representative table input formats (Markdown, HTML, and JSON), the Markdown format achieves

Error Types	Affected criteria	Ratio
Numerical Factual Errors	NAC	22%
Table Structure Understanding Errors	NAC, ICC	16%
Missing key points	ICC	17%
Generation Errors	NAC, ICC	20%
Truncation Errors	NAC, ICC	25%

Table 13: Error type distribution

Model	**URL**
Moonshot-V1-32k	https://kimi.moonshot.cn
Claude-3.5-Sonnet	https://www.anthropic.comt
Doubao-Pro-128k	https://www.volcengine.com
Doubao-Pro-32k	https://www.volcengine.com
GPT-4o	https://openai.com
OepnAI o1-mini	https://openai.com

Table 14: The URLs of closed-source models we used

the highest average performance, followed by HTML, while JSON exhibits the lowest performance.

	Markdown	JSON	HTML
Qwen2.5-72B-Instruct	59.59	55.82	54.91
Deepseek-R1	62.71	58.12	60.02
OpenAI-o1-1217	62.76	59.43	59.67

Table 15: Average performance of NAC, ICC and GEC of three different models across different input formats.

F Details for payment and GPU hours

We pay each annotator a daily remuneration of $40. We paid a total of $2500 for calling various LLMs API interfaces. We use 16 A100 40G GPUs for inference, which took a total of 25 hours.