Table question answering in the era of large language models: a comprehensive survey of tasks, methods, and evaluation

Wei Zhou, Bolei Ma, Annemarie Friedrich, Mohsen Mesgar

Angaben zur Veröffentlichung / Publication details:

Zhou, Wei, Bolei Ma, Annemarie Friedrich, and Mohsen Mesgar. 2026. "Table question answering in the era of large language models: a comprehensive survey of tasks, methods, and evaluation." In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026) - volume 1: long papers, July 2-7, 2026*, edited by Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, 12161–89. Stroudsburg, PA: Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2026.acl-long.557.

Table Question Answering in the Era of Large Language Models: A Comprehensive Survey of Tasks, Methods, and Evaluation

Wei Zhou[1,3] **Bolei Ma**[2] **Annemarie Friedrich**[3] **Mohsen Mesgar**[1]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]LMU Munich & Munich Center for Machine Learning, Germany
[3]University of Augsburg, Germany
{wei.zhou3|mohsen.mesgar}@de.bosch.com
bolei.ma@lmu.de annemarie.friedrich@uni-a.de

Abstract

Table Question Answering (TQA) aims to answer natural language questions about tabular data, often accompanied by additional contexts such as text passages. The task spans diverse settings, varying in table representation, question/answer complexity, modality involved, and domain. While recent advances in large language models (LLMs) have led to substantial progress in TQA, the field still lacks a systematic organization and understanding of task formulations, core challenges, and methodological trends, particularly in light of emerging research directions such as reinforcement learning. This survey addresses this gap by providing a comprehensive and structured overview of TQA research with a focus on LLM-based methods. We provide a comprehensive categorization of existing benchmarks and task setups. We group current modeling strategies according to the challenges they target, and analyze their strengths and limitations. Furthermore, we highlight underexplored but timely topics that have not been systematically covered in prior research. By unifying disparate research threads and identifying open problems, our survey offers a consolidated foundation for the TQA community, enabling a deeper understanding of the state of the art and guiding future developments in this rapidly evolving area.

1 Introduction

Tables are a ubiquitous data format in daily life (Cafarella et al., 2008). Automatically processing and understanding tabular data with (multi-modal) large language models ((M)LLMs) has recently attracted considerable attention from both industry (Katsis et al., 2022; Su et al., 2024) and academia (Pasupat and Liang, 2015; Wolff and Hulsebos, 2025), emerging as a prominent research direction.

Among the various tasks involving tables, including table generation (Gulati and Roysdon, 2023) and table-to-text (Parikh et al., 2020), table question answering (TQA) stands out as one of the most widely studied (Wu et al., 2025d). The goal of TQA is to answer questions based on tabular data, optionally augmented with additional context such as text passages or images. TQA can be instantiated in diverse settings. As illustrated in Figure 1, the input table may be provided directly or retrieved from a large corpus; tables vary in format, size, and structural complexity; and questions may target specific cells or require multi-step reasoning. These variations reflect real-world applications and necessitate different modeling strategies, driving rapid growth in the field. This survey consolidates resources and modeling approaches, and distills insights into promising directions for future research.

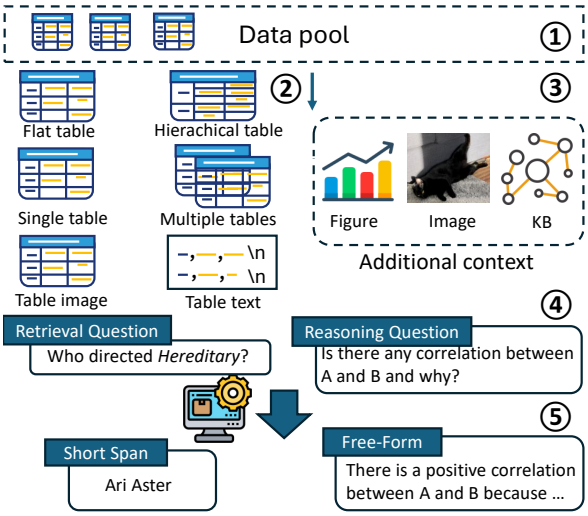


Figure 1: Five table question answering task setups.① **Domain:** Either the inputs need to be retrieved from a data pool (open-domain) or directly given (closed-domain). ② **Table**: Input tables can vary in structure, numbers, and representations. ③ **Additional Context**: Charts, images, and knowledge graphs can also be involved as inputs. ④ **Question Complexity**: A question can involve retrieving certain cells from a table or require reasoning and analysis to be solved. ⑤ **Answer Format**: Answers can be in short text spans, consisting only of numbers and entities, or in free-form natural language, with no limitation on types and length.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12161–12189
July 2-7, 2026

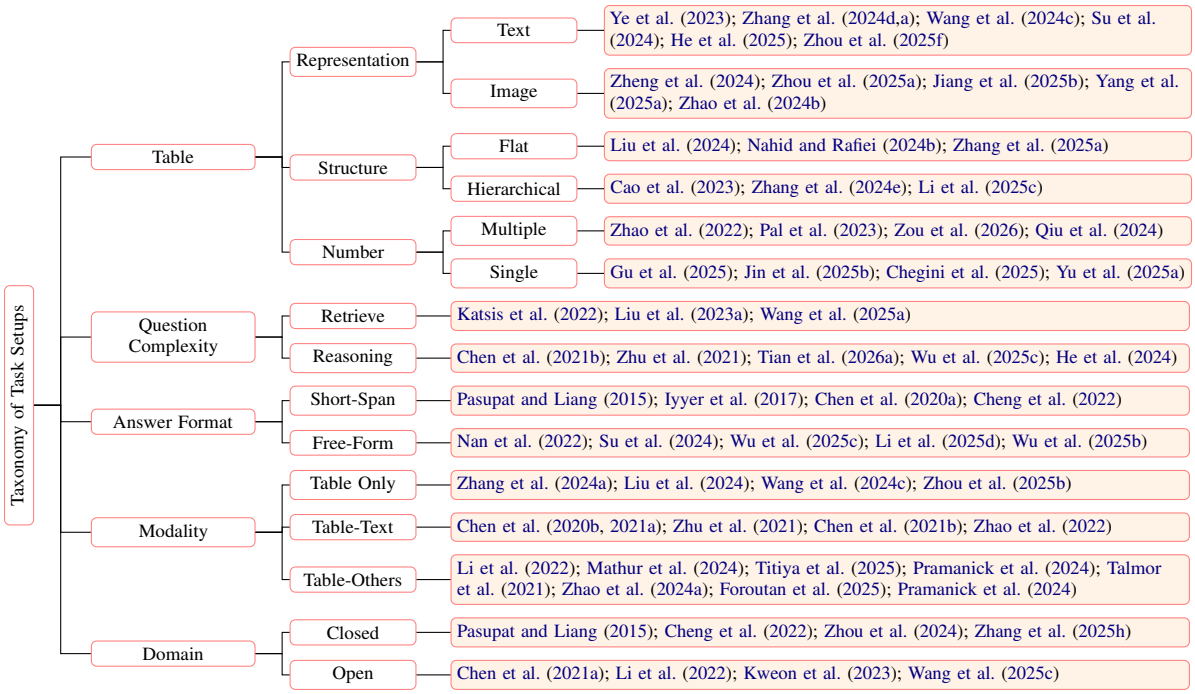


Figure 2: A taxonomy of TQA task setups. We list representative papers for each setup.

Comparing with Existing Surveys. Most prior surveys relevant to TQA focus solely on textual tables (Dong et al., 2022; Jin et al., 2022; Zhang et al., 2024c; Fang et al., 2024; Ren et al., 2025; Xu et al., 2026a). Among those addressing both textual and image tables, Wu et al. (2025d) concentrate on table representations without discussing modeling approaches, while Tian et al. (2026a) emphasize agentic setups and overlook fine-tuning methods. Crucially, no existing survey provides an overview of TQA task setups or covers LLM-era advances such as reinforcement learning, interpretability, and novel evaluation paradigms (see detailed comparisons in Appendix A.1).

Scope. We include both TQA and table fact verification (TFV), in which models must determine if a given statement is valid based on tabular data. TFV can be reformulated as TQA (Lu et al., 2023b). We also incorporate Text-to-SQL datasets where applicable, as SQL queries can produce final answers. Our survey covers 277 papers published mostly after 2022, when LLMs were used. Details on the collection and paper statistics are in Appendix A.2.

Structure. Section 2 outlines TQA task setups and benchmarks. Section 3 presents modeling approaches grouped by challenge. Section 4 reviews evaluation methodologies, and Section 5 discusses emerging topics and future research directions.

2 Task Setups and Resources

We dissect TQA from five perspectives. These are illustrated in Figure 2. Table 2 provides existing TQA datasets categorized by the characteristics of their task setups.

Table. Tables may appear as text or images, giving rise to two distinct task setups: those operating over textual tables (Zhang et al., 2024d,a; Wang et al., 2024c; Su et al., 2024; He et al., 2025; Zhou et al., 2025f) and those over table images (Zheng et al., 2024; Zhou et al., 2025a; Jiang et al., 2025b; Yang et al., 2025a; Zhao et al., 2024b). Beyond format, table structure and quantity further shape task difficulty. Hierarchical tables (Cao et al., 2023; Zhang et al., 2024e; Li et al., 2025c) pose greater structural challenges than flat ones (Liu et al., 2024; Nahid and Rafiei, 2024b; Zhang et al., 2025a). Multi-table settings (Zhao et al., 2022; Pal et al., 2023; Zou et al., 2026; Qiu et al., 2024) additionally require understanding inter-table relationships and handling longer inputs, compared to single-table setups (Gu et al., 2025; Jin et al., 2025b; Chegini et al., 2025; Yu et al., 2025a).

Question Complexity. TQA questions vary in the capabilities they require. Following Zhou et al. (2024), they can be divided into *retrieval* questions, solvable by locating relevant cells, and *reasoning* questions, which demand additional infer-

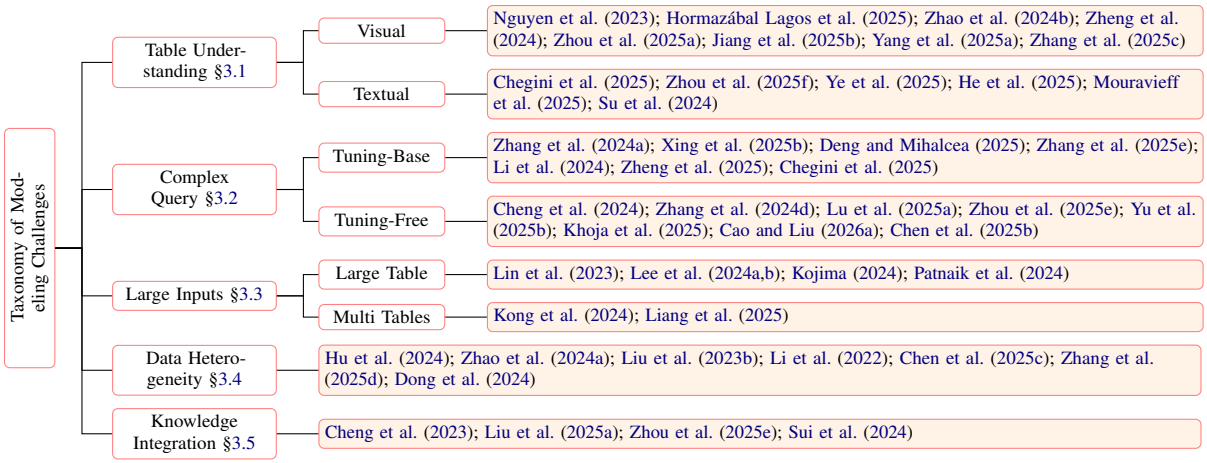


Figure 3: A taxonomy of methods categorized by challenges. We list representative papers for each challenge.

ence. Accordingly, TQA benchmarks range from simple setups involving only retrieval (Katsis et al., 2022; Liu et al., 2023a; Wang et al., 2025a) to complex ones requiring numerical (Chen et al., 2021b; Zhu et al., 2021; Lu et al., 2023a; Tian et al., 2026a), commonsense (Zhang et al., 2023), temporal (Gupta et al., 2023; Shankarampeta et al., 2025) reasoning, data analysis and plotting (Wu et al., 2025c; He et al., 2024). More recently, Shen et al. (2025) introduce probabilistic TQA, requiring models to reason under uncertainty.

Answer Formats. Most TQA setups feature short span answers composed of a few words or numbers (Pasupat and Liang, 2015; Iyyer et al., 2017; Chen et al., 2020a; Cheng et al., 2022; Wu et al., 2025a), which allow straightforward evaluation by exact match against reference answers. Osés Grijalba et al. (2024) further categorize answers by type (e.g., Boolean, numerical), finding that Boolean questions tend to be the easiest. However, real-world queries often demand free-form answers spanning multiple sentences, as in data analysis tasks. This format is receiving growing attention for its alignment with practical use cases (Nan et al., 2022; Su et al., 2024; Wu et al., 2025c; Li et al., 2025d; Wu et al., 2025b).

Modality. Beyond the standard table-question setup (Zhang et al., 2024a; Liu et al., 2024; Wang et al., 2024c; Zhou et al., 2025b), many works introduce additional input modalities such as passages (Chen et al., 2020b, 2021a; Zhu et al., 2021; Chen et al., 2021b; Zhao et al., 2022), images (Li et al., 2022; Mathur et al., 2024; Titiya et al., 2025; Pramanick et al., 2024; Talmor et al., 2021), charts (Zhao et al., 2024a; Foroutan et al., 2025; Pramanick et al., 2024), and knowledge graphs (Christmann et al., 2024; Hu et al., 2024; Huang et al., 2025a), better reflecting real-world settings where tables co-occur with heterogeneous data. Among these, the combination of tables and passages (table-text QA) is the most widely studied.

Domains. TQA can be divided into open-domain (Chen et al., 2021a; Li et al., 2022; Kweon et al., 2023; Strich et al., 2026) and closed-domain (Pasupat and Liang, 2015; Cheng et al., 2022; Zhou et al., 2024; Zhang et al., 2025h) settings, depending on whether the relevant inputs are provided. In open-domain TQA, a system must first retrieve relevant tables, either textual (Chen et al., 2021a) or image-based (Xu et al., 2026b), from a large pool before reasoning over them, posing additional challenges compared to the closed-domain setting where target inputs are given directly.

3 Modeling

We categorize modeling methods based on the challenges they try to address. In addition, we also analyze their strengths and limitations. Figure 3 provides an overview.

3.1 Table Understanding

Visual Table Modeling. Visual table understanding requires comprehending both table content and structure. A common approach is to *parse table images into text* using MLLMs (Nguyen et al., 2023; Hormazábal Lagos et al., 2025; Si et al., 2025). However, Xia et al. (2024) find that despite promising OCR performance on tabular data, MLLMs still struggle with spatial and formatting recognition, particularly for large or structurally complex

tables (Zheng et al., 2024; Bindini et al., 2025).

Another line of work focuses on *pre-training and/or fine-tuning MLLMs* (Zhao et al., 2024b; Zheng et al., 2024; Zhou et al., 2025a; Jiang et al., 2025b; Yang et al., 2025a; Zhang et al., 2025c; Yu et al., 2026). Training datasets of table images have been constructed from existing tabular datasets (Zhao et al., 2024b), by converting textual tables into images (Zheng et al., 2024; Zhou et al., 2025a; Jiang et al., 2025b), or built from scratch (Yang et al., 2025a; Zhang et al., 2025c).

In terms of training strategies, Zhu et al. (2026) decouple structure learning from content grounding during training. Kim et al. (2026) enhance visual table encoding through progressive question conditioning: injecting questions into vision transformers, along with a pruning mechanism to remove redundant background regions and a training recipe to mitigate the resulting information loss. Reinforcement learning has also been shown to improve performance in this setting (Kang et al., 2025). In terms of model architecture, dual vision encoders are employed to capture information at different levels of granularity (Zhao et al., 2024b; Zhou et al., 2025a), and Zhang et al. (2025c) train a mixture-of-experts model to jointly capture layout and semantic information.

Textual Table Modeling. Prior work has explored three main directions for table structure understanding: table representations, architecture modifications, and specialized training tasks. For *representation*, tables have been modeled as relational databases or Pandas DataFrames, where operations are expressed in code (Chegini et al., 2025; Zhou et al., 2025f), or as spreadsheets with formula-based operations (Cao et al., 2025; Wang et al., 2025d). Compared to database and DataFrame formats, which require a pre-defined schema, spreadsheet representations offer greater flexibility. Recent work improves the generalization of database representations by providing a processing module that converts a raw table into a SQL-ready table (Najpande et al., 2026). Another approach models tables as graphs (Hu et al., 2020; Chen et al., 2023; Huang et al., 2025b; Li et al., 2025b; Jin et al., 2025a; Wang et al., 2026e), where nodes represent cells and edges encode positional relationships, enabling explicit structural learning. For instance, Jin et al. (2025a) model tables as heterogeneous graphs with four node types: `TABLE` nodes, `ROW` nodes, `Header` nodes and `Data` nodes. Edges are defined based on node relationships, such as `Table-Header`. Tables have also been expressed as natural language tuples (Zhao et al., 2023a; Yang et al., 2025d) or trees (Wang et al., 2021; Cao et al., 2026; Tang et al., 2025; Qu et al., 2026)

Some methods *encode structure within the model* (He et al., 2025; Mouravieff et al., 2025; Su et al., 2024). For instance, He et al. (2025) serialize tables with special tokens and applies 2D LoRA (Hu et al., 2022) to capture low-rank positional information, while Mouravieff et al. (2025) design a sparse attention mask for tabular data.

Finally, *task-specific objectives* have been proposed to enhance structural reasoning, such as layout transformation inference (Jin et al., 2025b), where models detect changes between original and altered layouts, and cell position generation (Cho et al., 2025), where models predict cell locations.

Discussion. Visual table understanding is generally more challenging than textual table understanding (Deng et al., 2024; Zheng et al., 2024), possibly because it requires additional content interpretation. Notably, OCR-based pipelines do not outperform models directly fine-tuned on table images (Zheng et al., 2024). For details, see Appendix A.3. For textual table understanding, the need to define a fixed table schema may be a drawback for database tables. When using more fine-grained textual representations such as JSON or Markdown, there seems to be no optimal format across datasets and models (Zhang et al., 2024b).

3.2 Complex Query

Tuning-Based. Methods in this group fine-tune (M)LLMs to improve reasoning over tabular data. Diverse training datasets covering a wide range of reasoning types are crucial for handling complex queries, and are either collected from existing benchmarks (Zhang et al., 2024a; Xing et al., 2025b; Deng and Mihalcea, 2025) or synthesized (Zhang et al., 2025e; Li et al., 2024; Zheng et al., 2025; Chegini et al., 2025). For instance, Zheng et al. (2025) select training samples based on identified model weaknesses and progressively fine-tune the model, while Chegini et al. (2025) distill Python programs generated by large closed-source models to train a smaller open-weight LLM. Combining fine-tuned models with tools at inference time has also been explored (Wu and Feng, 2024; Mouravieff et al., 2024; Vinayagame et al., 2025), and reinforcement learning can further boost per-

formance (Nahid and Rafiei, 2024b; Zhou et al., 2025b; Stoisser et al., 2025; Chi et al., 2025; Zhang et al., 2026a; Jiang et al., 2026a; Li et al., 2026).

Rather than fine-tuning the policy model, some work instead fine-tune a reward model that ranks candidate solutions (Zou et al., 2025; Tang et al., 2026). Specifically, a process reward model is trained to score each intermediate reasoning step, with step scores aggregated for solution ranking and optionally used to guide next-step search. An alternative use of training data is to extract reusable experience from it (Gao et al., 2025), which can then be retrieved and appended to test instances at inference time.

Tuning-Free. In tuning-free approaches, agentic workflows that integrate tools are widely adopted to enable complex query handling and to reduce hallucinations (Cheng et al., 2024; Zhang et al., 2024d; Lu et al., 2025a; Zhou et al., 2025e; Yu et al., 2025b; Khoja et al., 2025; Cao and Liu, 2026a; Chen et al., 2025b; Cao and Liu, 2026b). These methods typically employ multi-step reasoning to decompose complex questions into simpler sub-problems (Mao et al., 2026; Zhou et al., 2025e; Ji et al., 2024; Zhang et al., 2025b; Deng et al., 2025; Zhao et al., 2024c; Nguyen et al., 2025c; Jiang et al., 2025a, 2026b). Some also incorporate memory modules to store and reuse past reasoning experiences (Bai et al., 2025; Gu et al., 2025; Cheng et al., 2026).

As for the tooling part, models generate and execute Python (Cao et al., 2023; Zhang et al., 2024e; Yu et al., 2025b; Pyo et al., 2026) or SQL (Abhyankar et al., 2025b; Nahid and Rafiei, 2024b; Khoja et al., 2025; Thanga et al., 2025; Sui et al., 2025) code to obtain reasoning results. To improve code generation, error feedback can be provided to the LLM as a revision signal (Cheng et al., 2024; López Gude et al., 2025; Site et al., 2025). To alleviate the reliance on keyword matching inherent in code-based approaches, Nguyen et al. (2025b) propose a framework that leverages LLMs with entity-oriented search. Other commonly used tools include calculator (Zhu et al., 2024) and Wikipedia search (Zhou et al., 2025e).

In terms of prompting strategies, agentic-flow systems often adopt ReAct-style prompting (Zhou et al., 2025e; Yu et al., 2025b; Bai et al., 2025; Wang et al., 2026d). Zhang et al. (2025j) show that prompting models to iterate over rows can reduce hallucination, while Rajgaria et al. (2025) find that no single prompting technique consistently outperforms others in TQA involving temporal reasoning. Yu et al. (2025c) propose a self-evolving framework for prompt optimization.

Verification modules have also been introduced to check the correctness of intermediate reasoning (Wang et al., 2025b; Yu et al., 2025a; Luo et al., 2026b; Hyeon et al., 2026; Tang et al., 2026). For instance, Tang et al. (2026) train process reward models to assess the correctness of each reasoning step, and adopt Best-of-N selection to identify the final answer from multiple sampled reasoning chains. Other similar methods involving test-time scaling include self-consistency (Liu et al., 2024) and uncertainty-aware sampling (Cheng et al., 2026).

Discussion. Tuning-free methods require no training data, but incur longer inference times and higher token costs (Zhou et al., 2025b). In contrast, tuning-based approaches are more efficient at inference but are prone to out-of-domain performance degradation (Deng and Mihalcea, 2025; Deng et al., 2026). Both methods demonstrate state-of-the-art performance (Yang et al., 2025e; Abhyankar et al., 2025c; Cao and Liu, 2026a), but involve trade-offs. A promising intermediate strategy might be to fine-tune models for general reasoning capabilities while delegating specific table operations, such as retrieval, to tools (Wu and Feng, 2024; Zhu et al., 2024).

3.3 Large Inputs

The main challenge with large inputs is efficiently identifying relevant information, as processing full tables is often infeasible or ineffective. We review methods for handling large and multiple tables.

Large Tables. A common approach is to fine-tune retrievers to identify the most relevant cells for a given question (Lin et al., 2023; Lee et al., 2024a,b; Kojima, 2024; Patnaik et al., 2024). Another strategy leverages LLMs to select pertinent table content or relevant headers (Ye et al., 2023; Jiang et al., 2023, 2025c; Sui et al., 2024; Sun et al., 2026). Some methods define atomic operations for table manipulation (Wang et al., 2024a,c), while others embed both queries and tables for semantic matching (Sui et al., 2024; Yu et al., 2025b; Shirafuji et al., 2025). A further line of work employs code generation to execute table-filtering operations (Gemmell and Dalton, 2023; Zhou et al., 2025c; Vyatkin and Oliseenko, 2025).

Multiple Tables. LLMs can assist retrieval by enhancing table semantics. For instance, Liang et al. (2025) augment table snippets with LLM-generated questions to produce richer table representations. LLMs can also directly facilitate retrieval. Kong et al. (2024) generate SQL queries to identify relevant tables. Rather than treating each table as an independent document, Zou et al. (2026) represent the table corpus as a hypergraph and select the most relevant subgraph using a multi-stage coarse-to-fine process. Similarly, Luo et al. (2026a) capture table relationships using graphs. They design a question decomposer and a coverage-aware retriever to identify relevant tables. Many works adopt dense passage retrieval (Karpukhin et al., 2020) or language model embeddings to encode tables, passages, and questions (Guan et al., 2024; Bardhan et al., 2024).

Discussion. Directly using LLMs to retrieve relevant cells or tables can lead to information loss (Zhou et al., 2025c). Fine-tuned sub-table retrievers have shown effectiveness, but their generalizability to diverse table formats remains limited. For both scenarios, retrieval-augmented generation (RAG) offers a viable alternative: tables or cells are embedded into a vector database, and questions are issued as queries (Chen et al., 2024).

3.4 Data Heterogeneity

To handle different modalities, existing methods typically follow two directions: (1) employing specialized retrievers and reasoners (Li et al., 2022; Zhao et al., 2024a; Hu et al., 2024; Liu et al., 2023b), or (2) designing unified representations (Dong et al., 2024; Chen et al., 2025c; Zhang et al., 2025d; Park et al., 2025). In the first category, Hu et al. (2024) use a multi-stage knowledge-graph retriever, Zhao et al. (2024a) employ multi-agent retrieval from charts and tables, and Liu et al. (2023b) generate image captions to capture salient visual content; atomic retrieval functions can also target both passages and tables (Shi et al., 2024; Zhou et al., 2025e). In the second category, unified structures integrate heterogeneous sources: Chen et al. (2025c) use DataFrames to jointly represent tabular and textual data, Zhang et al. (2025d) propose Condition Graphs combining tables and knowledge graphs, and Agarwal et al. (2025) build hybrid graphs from linked entities; tabular content may also be summarized into text for downstream tasks (Bardhan et al., 2024). For reasoning over both tables and text, LLMs can align references across modalities (Luo et al., 2023; Zhang et al., 2025h; Sharafath et al., 2026) or be fine-tuned for domain-specific reasoning, as in Zhu et al. (2024)'s financial QA system, which processes both modalities to produce multi-step reasoning chains combining evidence extraction, logical or equation formulation, and execution. In summary, which method to choose depends on the modalities involved. Constructing graphs is straightforward for tables and text, whereas employing separate retrievers is preferable when modalities are harder to unify, e.g., if information from charts and tables needs to be combined.

3.5 Knowledge Integration

External knowledge is often required to answer TQA problems. For example, a question in DataBench (Osés Grijalba et al., 2024) asks: *"What is the total number of rebounds recorded in the dataset where the ball didn't change possession?"* Answering this question requires knowing that *OREB* in the table header denotes *offensive rebounds*, a case where ball possession does not change. To handle such cases, prior work has integrated external resources such as Wikipedia into the reasoning process (Zhou et al., 2025e; Sui et al., 2024). For instance, Zhou et al. (2025e) design an atomic function *Search (arg)*, which returns the first few lines from the Wikipedia page of a specified argument. The retrieved content is then stored in the system's memory for subsequent reference. An alternative strategy is to elicit factual knowledge directly from LLMs (Cheng et al., 2023; Liu et al., 2025a). However, this approach is susceptible to hallucinations, as LLMs may generate factually incorrect information.

4 Evaluation

In this section, we discuss evaluation in terms of task performance, system robustness, and model-generated reasoning as explanations.

4.1 Task Performance

Current TQA evaluation primarily focuses on performance, measured by automatic metrics such as Exact Match (EM) and ROUGE. EM suits short, span-based answers, whereas ROUGE, BLEU, or F1 scores are better for free-form responses. While efficient, these metrics often miss subtle mismatches between predicted and gold answers (Wolff and Hulsebos, 2025), e.g., EM may wrongly

mark answers as incorrect due to formatting differences (Jan 1 vs. 01-01). To address such issues, outputs are normalized to a canonical form before applying EM (Khoja et al., 2025). This "relaxed EM" improves robustness but can cause inconsistencies, as systems may adopt different normalization rules, leading to misleading cross-system comparisons (Hormazábal Lagos et al., 2025).

Beyond traditional metrics, some studies employ LLMs as judges (Wu and Feng, 2024; Zhou et al., 2025e; Jiang et al., 2025b; Zhang et al., 2025f) or use human evaluation (Zhao et al., 2024c; Ye et al., 2025; Khoja et al., 2025). Rajgaria et al. (2025) propose the Hybrid Correctness Score, combining F_1 with LLM judgments. Wolff and Hulsebos (2025) show that, when calibrated with human annotations, LLM-as-a-judge can offer a reliable evaluation signal for tasks requiring reasoning over tables.

4.2 Robustness Evaluation

Robustness to structural or content variations in tables and questions is a key property of TQA systems (Yang et al., 2022; Bhandari et al., 2025). Zhao et al. (2023b) present a benchmark for adversarial attacks on table structure/content and question perturbations, showing that state-of-the-art models still struggle. Zhou et al. (2024) define three robustness dimensions: (1) resilience to table structure changes, (2) resistance to shortcut exploitation, and (3) robustness in numerical reasoning. Their benchmark indicates that pipeline models handle value and positional changes best, whereas LLM-based models are more vulnerable to table shuffling, a trend also observed in other works (Ashury-Tahan et al., 2025; Liu et al., 2024; Yang et al., 2022). Wolff and Hulsebos (2025) further evaluate LLM robustness on real-world tables with missing or duplicated values, underscoring the need for robustness-oriented evaluation.

4.3 Evaluating Explanations and Reasoning

An underexplored aspect of TQA evaluation is assessing explanations and reasoning processes. Model-generated chains of thought (Wei et al., 2022) are often treated as natural language explanations (Zhao et al., 2024c; Zhou et al., 2025f; Lu et al., 2025b), either elicited directly (Zhao et al., 2024c; Zhou et al., 2025f) or derived from executable program outputs (Lu et al., 2025b). Nguyen et al. (2025a) represent explanations as chains of attribution maps, showing intermediate relevant tables alongside reasoning steps, and propose three evaluation tasks: (1) preference ranking, where judges rank explanation quality; (2) forward simulation, where judges answer using only the explanation; and (3) verification, where judges assess prediction correctness based on the explanation. Both human annotators and LLMs serve as judges. Zhou et al. (2025b) take a different approach, estimating the probability of reaching the correct answer from a reasoning step to quantify each step's contribution to the final outcome.

Discussion. Popular datasets like WTQ (Pasupat and Liang, 2015) and TabFact (Chen et al., 2020a) may suffer from data contamination (Zhou et al., 2025d), leading to overly optimistic performance estimates. Wang (2026) design a light-weight evaluation protocol taking the issues into consideration. Beyond task performance, dimensions such as robustness and reasoning correctness should be systematically assessed to ensure the development of trustworthy TQA systems.

5 Discussion and Future Directions

We discuss emerging topics for future exploration, including table representation, multilinguality, reinforcement learning, multi-modal modeling, interpretability and human-centric setups.

Table Representation. Tables appear in various formats, including structured databases, plain text, images, and graphs. When tables are stored in databases, they can be directly queried using SQL. Zhang et al. (2026c) find that modeling tables as databases achieves better results than using LLMs directly when tables feature clear schemas. However, tables in textual or image form are often noisier due to inconsistent formatting (Zhou et al., 2024), mixed data types (Nahid and Rafiei, 2024a), missing values (Wolff and Hulsebos, 2025), and implicit or incomplete schemas (Zheng et al., 2023), posing challenges for building models that generalize to real-world noisy tables.

A growing line of work investigates leveraging both textual and visual representations of tables (Deng et al., 2024; Zhou et al., 2025d; Borisova et al., 2025; Liu et al., 2025b; Xing et al., 2026), capitalizing on the fact that one modality can often be converted to the other (e.g., image-to-text via OCR, or text-to-image via HTML rendering). However, most existing approaches adopt ensemble strategies that select the optimal representation based on specific problem features such as table

size (Deng et al., 2024; Zhou et al., 2025d), which lack flexibility when new features emerge or when interactions among multiple features need to be considered.

To overcome this limitation, some work instead lets the model learn which modality to use (Liu et al., 2025b; Xing et al., 2026). For instance, Xing et al. (2026) train a gating network to select among text-based, vision-based, or fused representations. An alternative approach is to process textual and visual inputs through dedicated encoders and integrate them within the model, a strategy widely adopted in vision-language tasks (Zhao et al., 2024b; Zhou et al., 2025a). Separate encoders can capture complementary signals, such as layout structure from images and semantic content from text, potentially leading to more robust and generalizable representations.

Multilinguality and Low-Resource Settings. Tables in real-world applications can be text-heavy and may contain content in one or more languages, as in user or product information tables. Nevertheless, most existing TQA datasets and studies focus on English (Pasupat and Liang, 2015; Zhang et al., 2023, 2024a) or other high-resource languages with large speaker populations, such as Chinese (Zheng et al., 2023; Liu et al., 2023a; Zhao et al., 2024a), leaving low-resource and multilingual scenarios underexplored. These settings introduce additional challenges, including right-to-left text processing for languages such as Persian and Hebrew, as well as out-of-distribution lexical patterns for LLMs predominantly trained on high-resource languages. Although modern LLMs support multilingual processing, directly applying them to low-resource table reasoning tasks typically results in degraded performance. For instance, Shu et al. (2025) construct a multilingual table reasoning dataset and report that LLMs achieve their highest performance on Indo-European languages and their lowest on Niger-Congo languages, likely reflecting disparities in pre-training data distribution. Their study also finds that multilingual fine-tuning does not consistently improve table question answering performance. Translation-based approaches, which convert tables from the target language into English prior to reasoning, are similarly limited, as their effectiveness depends heavily on translation quality (Zhang et al., 2025i).

Addressing table understanding and reasoning in multilingual and low-resource contexts is both practically important and ethically motivated: such tasks arise frequently in real-world applications, and equitable digital access requires robust performance across languages. Recent efforts have begun to introduce datasets tailored to these scenarios (Minhas et al., 2022; Zhang et al., 2025i; Shu et al., 2025), but the field remains nascent. We argue that effective table reasoning systems for multilingual and low-resource settings will require more than translation pipelines and off-the-shelf LLMs, calling for dedicated methods that explicitly address the unique linguistic and resource constraints of these environments.

Reinforcement Learning in TQA. Reinforcement learning with verifiable rewards (RLVR) has gained increasing attention due to its success in developing reasoning-oriented models such as DeepSeek-R1 (DeepSeek-AI et al., 2025). Recent studies have also explored RLVR in TQA (Jin et al., 2025b; Yang et al., 2025e; Jiang et al., 2025b; Lei et al., 2025; Cao et al., 2025; Stoisser et al., 2025; Liu et al., 2025b). Common rewards include answer correctness (Yang et al., 2025e; Jiang et al., 2025b; Stoisser et al., 2025; Liu et al., 2025b), program executability (Jin et al., 2025b; Cao et al., 2025), output formatting (Yang et al., 2025e; Jiang et al., 2025b), positional alignment (Lei et al., 2025), and length constraints (Jin et al., 2025b). Jiang et al. (2026a) incorporate tool use into the reward design and propose RAPO, a rank-aware policy optimization algorithm that modifies GRPO by removing KL divergence and employing token-level policy gradients to improve training stability.

Beyond final outcome rewards, several studies explore richer supervision signals. Zhou et al. (2025b) propose a process supervision framework and demonstrate that models trained with intermediate rewards outperform those trained on final rewards alone.Yang et al. (2025c) similarly use process rewards to update policy models. Zhao et al. (2026) employ Monte Carlo Tree Search to generate pseudo-gold trajectories for optimizing agents with reinforcement learning. For table image understanding, Kang et al. (2025) apply continuous Tree-Edit-Distance Similarity rewards to train models for recognizing table structures and contents.

RLVR has been shown to improve LLMs' reasoning capabilities over tabular data. Compared to supervised fine-tuning (SFT), RLVR-trained models exhibit better generalizability (Yang et al., 2025e; Cao et al., 2025) and greater robustness to

row and column perturbations (Lei et al., 2025). Nevertheless, initializing models with SFT prior to RLVR training remains crucial for achieving strong performance (Cao et al., 2025).

Diverse and Multi-Modal Data Modeling. Many existing TQA studies focus on table-only settings with relatively simple queries (Ye et al., 2023; Ni et al., 2023; Nahid and Rafiei, 2024b; Wang et al., 2024c; Liu et al., 2024; Zhou et al., 2025c), a setup well-suited for evaluating LLMs' ability to understand table structures. However, it falls short in capturing more complex scenarios that involve multiple modalities and open-domain settings. An increasing body of work has recognized these limitations and proposed more challenging benchmarks (He et al., 2024; Qiu et al., 2024; Wu et al., 2025a; Osés Grijalba et al., 2024; Wu et al., 2025c,b; Zhu et al., 2025). Beyond information retrieval and aggregation, real-world table interactions are considerably more diverse. Queries can be ambiguous (Grijalba et al., 2026), and users also engage in tasks such as interpreting markup tables (Wang et al., 2026c) and performing table manipulations (Xing et al., 2025a). We argue that future research should extend modeling to these real-world scenarios.

Interpretability and Faithfulness. Instead of directly producing a final answer (Herzig et al., 2020; Liu et al., 2022; Jiang et al., 2022; Zhang et al., 2025f), an increasing number of TQA approaches also return a reasoning process (Zhao et al., 2024c; Chegini et al., 2025; Nguyen et al., 2025a). Such reasoning not only improves system performance but also provides human-understandable justifications for how an answer is derived. However, despite appearing plausible, these explanations may not faithfully represent the model's actual decision-making process (Turpin et al., 2023; Chen et al., 2025a). Building trustworthy TQA systems is particularly important in high-stakes domains such as medicine (Bardhan et al., 2022). Achieving this requires that output reasoning accurately reflect a model's table understanding capabilities, e.g., faithfully responding with "I don't know" when a table is beyond the model's ability to interpret. We argue that much of the current work on interpretability in TQA focuses on generating post-hoc justifications for answers, rather than explanations that transparently reveal the underlying reasoning process.

A complementary direction involves probing the internal mechanisms by which LLMs process tabular data. Zhang et al. (2026b) find that LLMs go through three stages when retrieving table cells: semantic binding (aligning the query with table headers), coordinate localization (navigating structural indices), and information extraction (retrieving target values). During coordinate localization in particular, models locate the target cell by counting discrete delimiters. Wang et al. (2026b) extend this line of inquiry from cell retrieval to general TQA, analyzing model attention patterns during question answering. They find that early layers broadly encode table structure, middle layers localize relevant cells, and later layers generate the final answer.

Human-Centric and Socially-Aware Setups. Current research in TQA has largely focused on improving system performance, often overlooking the role of human interaction. However, the ultimate aim of such systems is to empower humans to more effectively interact with tabular data, analogous to the broader goal of developing NLP systems for human and socially aware uses (Hovy and Yang, 2021; Ziems et al., 2024; Yang et al., 2025b). This highlights the importance of human-centric and socially aware design. We identify two complementary research directions: (1) human-centric modeling, which emphasizes representations and benchmarks that capture the unique characteristics of tabular data (e.g., Hu et al., 2024; Ahmad et al., 2025); and (2) socially grounded applications, where TQA systems serve downstream tasks for social good, such as analyzing environmental sustainability reports (Dimmelmeier et al., 2024; Beck et al., 2025), advancing biomedical research (Luo et al., 2022), and supporting decision-making in financial domains (Strich et al., 2026; Malarkkan et al., 2026).

6 Conclusion

In this survey, we have reviewed recent advances in TQA with (M)LLMs, covering representative TQA setups, key challenges, and corresponding solutions, along with a discussion of promising future research directions. Looking ahead, we envision a stronger synergy between methods from NLP and related fields, and anticipate that TQA research, both in modeling and evaluation, will increasingly adopt more comprehensive settings. Such settings should account for diverse factors, including human model interaction, system robustness, and real-world applicability.

Limitations

In this study, we present a survey on TQA with LLMs. Related surveys are discussed in Appendix A.1. Despite our best efforts, certain limitations remain.

References and Methods. We collected papers published before April 2026, which means that works appearing after this time are not included. The majority of papers were retrieved from venues such as ACL, EMNLP, NAACL, NeurIPS, ICLR, and arXiv, using English-language queries. This approach may have led to the omission of works published in other languages. Furthermore, due to space constraints, we are unable to provide exhaustive technical details for all methods covered in this survey.

Survey Scope. This survey focuses exclusively on table question answering. Our objective is to provide a comprehensive and detailed overview of the task's characteristics, challenges, corresponding modeling methods, as well as emerging topics for future research. This focus excludes other table-related tasks, such as table generation and summarization.

References

Nikhil Abhyankar, Purvi Chaurasia, Sanchit Kabra, Ananya Srivastava, Vivek Gupta, and Chandan K. Reddy. 2025a. Rust-bench: Benchmarking llm reasoning on unstructured text within structured tables. *Preprint*, arXiv:2511.04491.

Nikhil Abhyankar, Vivek Gupta, Dan Roth, and Chandan K. Reddy. 2025b. H-STAR: LLM-driven hybrid SQL-text adaptive reasoning on tables. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8841–8863, Albuquerque, New Mexico. Association for Computational Linguistics.

Nikhil Abhyankar, Vivek Gupta, Dan Roth, and Chandan K. Reddy. 2025c. H-STAR: LLM-driven hybrid SQL-text adaptive reasoning on tables. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8841–8863, Albuquerque, New Mexico. Association for Computational Linguistics.

Ankush Agarwal, Chaitanya Devaguptapu, and Ganesh S. 2025. Hybrid graphs for table-and-text based question answering using LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 858–875, Albuquerque, New Mexico. Association for Computational Linguistics.

Chaitanya Agarwal, Vivek Gupta, Anoop Kunchukuttan, and Manish Shrivastava. 2022. Bilingual tabular inference: A case study on Indic languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4018–4037, Seattle, United States. Association for Computational Linguistics.

Mohammad S. Ahmad, Zan A. Naeem, Michaël Aupetit, Ahmed Elmagarmid, Mohamed Eltabakh, Xiasong Ma, Mourad Ouzzani, and Chaoyi Ruan. 2025. Hct-qa: A benchmark for question answering on human-centric tables. *Preprint*, arXiv:2504.20047.

Mubashara Akhtar, Chenxi Pang, Andreea Marzoca, Yasemin Altun, and Julian Martin Eisenschlos. 2024. Tanq: An open domain dataset of table answered questions. *Trans. Assoc. Comput. Linguistics*, 13:461–480.

Iñigo Alonso, Imanol Miranda, Eneko Agirre, and Mirella Lapata. 2026. TABLET: A large-scale dataset for robust visual table understanding. In *The Fourteenth International Conference on Learning Representations*.

Rana Alshaikh, Israa Alghanmi, and Shelan S. Jeawak. 2025. Aratable: Benchmarking llms' reasoning and understanding of arabic tabular data. *ArXiv*, abs/2507.18442.

Shir Ashury-Tahan, Yifan Mai, C Rajmohan, Ariel Gera, Yotam Perlitz, Asaf Yehudai, Elron Bandel, Leshem Choshen, Eyal Shnarch, Percy Liang, and Michal Shmueli-Scheuer. 2025. The mighty torr: A benchmark for table reasoning and robustness. *ArXiv*, abs/2502.19412.

Ye Bai, Minghan Wang, and Thuy-Trang Vu. 2025. MAPLE: Multi-agent adaptive planning with long-term memory for table reasoning. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 132–158, Sydney, Australia. Association for Computational Linguistics.

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. DrugEHRQA: A question answering dataset on structured and unstructured electronic health records for medicine related queries. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1083–1097, Marseille, France. European Language Resources Association.

Jayetri Bardhan, Bushi Xiao, and Daisy Zhe Wang. 2024. Ttqa-rs- a break-down prompting approach for multi-hop table-text question answering with reasoning and summarization. *ArXiv*, abs/2406.14732.

Jacob Beck, Anna Steinberg, Andreas Dimmelmeier, Laia Domenech Burin, Emily Kormanyos, Maurice Fehr, and Malte Schierholz. 2025. Addressing data gaps in sustainability reporting: A benchmark dataset for greenhouse gas emission extraction. *Scientific Data*, 12(1):1497.

Kushal Raj Bhandari, Sixue Xing, Soham Dan, and Jianxi Gao. 2025. Exploring the robustness of language models for tabular question answering via attention analysis. *Transactions on Machine Learning Research*.

Luca Bindini, Simone Giovannini, Simone Marinai, Valeria Nardoni, and Kimiya Noor Ali. 2025. Hierarchical structure understanding in complex tables with vllms: a benchmark and experiments. In *Document Analysis and Recognition – ICDAR 2025 Workshops: Wuhan, China, September 20–21, 2025, Proceedings, Part II*, page 3–16, Berlin, Heidelberg. Springer-Verlag.

Ekaterina Borisova, Fabio Barth, Nils Feldhus, Raia Abu Ahmad, Malte Ostendorff, Pedro Ortiz Suarez, Georg Rehm, and Sebastian Möller. 2025. Table understanding and (multimodal) LLMs: A cross-domain case study on scientific vs. non-scientific data. In *Proceedings of the 4th Table Representation Learning Workshop*, pages 109–142, Vienna, Austria. Association for Computational Linguistics.

Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. Webtables: exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1):538–549.

Bin Cao, Huixian Lu, Chenwen Ma, Ting Wang, Ruizhe Li, and Jing Fan. 2026. Orthogonal hierarchical decomposition for structure-aware table understanding with large language models. *Preprint*, arXiv:2602.01969.

Lang Cao and Hanbing Liu. 2026a. Tablemaster: A recipe to advance table understanding with language models. In *The Fourteenth International Conference on Learning Representations*.

Lang Cao and Hanbing Liu. 2026b. Tablemaster: A recipe to advance table understanding with language models. In *The Fourteenth International Conference on Learning Representations*.

Lang Cao, Jingxian Xu, Hanbing Liu, Jinyu Wang, Mengyu Zhou, Haoyu Dong, Shi Han, and Dongmei Zhang. 2025. Fortune: Formula-driven reinforcement learning for symbolic table reasoning in language models. *ArXiv*, abs/2505.23667.

Yihan Cao, Shuyi Chen, Ryan Liu, Zhiruo Wang, and Daniel Fried. 2023. API-assisted code generation for question answering on varied table structures. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14536–14548, Singapore. Association for Computational Linguistics.

Atoosa Chegini, Keivan Rezaei, Hamid Eghbalzadeh, and Soheil Feizi. 2025. RePanda: Pandas-powered tabular verification and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32200–32212, Vienna, Austria. Association for Computational Linguistics.

Pei Chen, Soumajyoti Sarkar, Leonard Lausen, Balasubramaniam Srinivasan, Sheng Zha, Ruihong Huang, and George Karypis. 2023. Hytrel: Hypergraph-enhanced tabular data representation learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 32173–32193. Curran Associates, Inc.

Si-An Chen, Lesly Miculicich, Julian Martin Eisenschlos, Zifeng Wang, Zilong Wang, Yanfei Chen, Yasuhisa Fujii, Hsuan-Tien Lin, Chen-Yu Lee, and Tomas Pfister. 2024. Tablerag: Million-token table understanding with language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 74899–74921. Curran Associates, Inc.

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

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

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson E. Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vladimir Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025a. Reasoning models don't always say what they think. *ArXiv*, abs/2505.05410.

Yibin Chen, Yifu Yuan, Zeyu Zhang, Yan Zheng, Jinyi Liu, Fei Ni, Jianye Hao, Hangyu Mao, and Fuzheng Zhang. 2025b. Sheetagent: Towards a generalist agent for spreadsheet reasoning and manipulation via large language models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 158–177, New York, NY, USA. Association for Computing Machinery.

Yongrui Chen, Junhao He, Linbo Fu, Shenyu Zhang, Rihui Jin, Xinbang Dai, Jiaqi Li, Dehai Min, Nan Hu, Yuxin Zhang, Guilin Qi, Yi Huang, and Tongtong Wu. 2025c. Pandora: A code-driven large language model agent for unified reasoning across diverse structured knowledge. *ArXiv*, abs/2504.12734.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021b. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mingyue Cheng, Shuo Yu, Chuang Jiang, Xiaoyu Tao, Qingyang Mao, Jie Ouyang, Qi Liu, and Enhong Chen. 2026. Tablemind++: An uncertainty-aware programmatic agent for tool-augmented table reasoning. *Preprint*, arXiv:2603.07528.

Sitao Cheng, Ziyuan Zhuang, Yong Xu, Fangkai Yang, Chaoyun Zhang, Xiaoting Qin, Xiang Huang, Ling Chen, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. 2024. Call me when necessary: LLMs can efficiently and faithfully reason over structured environments. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4275–4295, Bangkok, Thailand. Association for Computational Linguistics.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland. Association for Computational Linguistics.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding language models in symbolic languages. In *The Eleventh International Conference on Learning Representations*.

Ce Chi, Xing Wang, Zhendong Wang, Xiaofan Liu, Ce Li, Zhiyan Song, Chen Zhao, Kexin Yang, Boshen Shi, Jingjing Yang, Chao Deng, and Junlan Feng. 2025. Jt-da: Enhancing data analysis with tool-integrated table reasoning large language models. *Preprint*, arXiv:2512.06859.

Sanghyun Cho, Hye-Lynn Kim, Jung-Hun Lee, and Hyuk-Chul Kwon. 2025. Swing: Weakly supervised table question answering with self-training via reinforcement learning. In *2025 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 273–278.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2024. Compmix: A benchmark for heterogeneous question answering. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1091–1094, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Irwin Deng, Kushagra Dixit, Dan Roth, and Vivek Gupta. 2025. Enhancing temporal understanding in LLMs for semi-structured tables. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4951–4970, Albuquerque, New Mexico. Association for Computational Linguistics.

Naihao Deng and Rada Mihalcea. 2025. Rethinking table instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21757–21780, Vienna, Austria. Association for Computational Linguistics.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as texts or images: Evaluating the table reasoning ability of LLMs and MLLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 407–426, Bangkok, Thailand. Association for Computational Linguistics.

Naihao Deng, Sheng Zhang, Henghui Zhu, Shuaichen Chang, Jiani Zhang, Alexander Hanbo Li, Chung-Wei Hang, Hideo Kobayashi, Yiqun Hu, and Patrick Ng. 2026. What really matters for table LLMs? a meta-evaluation of model and data effects. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3755–3782, Rabat, Morocco. Association for Computational Linguistics.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: Towards proactive conversational question answering over tabular and textual data in finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6970–6984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andreas Dimmelmeier, Hendrik Doll, Malte Schierholz, Emily Kormanyos, Maurice Fehr, Bolei Ma, Jacob Beck, Alexander Fraser, and Frauke Kreuter. 2024. Informing climate risk analysis using textual information - a research agenda. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 12–26, Bangkok, Thailand. Association for Computational Linguistics.

Haoyu Dong, Zhoujun Cheng, Xinyi He, Mengyu Zhou, Anda Zhou, Fan Zhou, Ao Liu, Shi Han, and Dongmei Zhang. 2022. Table pre-training: A survey on model architectures, pre-training objectives, and downstream tasks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5426–5435. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Haoyu Dong, Haochen Wang, Anda Zhou, and Yue Hu. 2024. Ttc-quali: A text-table-chart dataset for multimodal quantity alignment. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 181–189, New York, NY, USA. Association for Computing Machinery.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.

Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. 2025. WikiMixQA: A multimodal benchmark for question answering over tables and charts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24941–24958, Vienna, Austria. Association for Computational Linguistics.

Chufan Gao, Jintai Chen, and Jimeng Sun. 2025. Utilizing training data to improve llm reasoning for tabular understanding. *Preprint*, arXiv:2508.18676.

Somraj Gautam, Abhishek Bhandari, and Gaurav Harit. 2025. TabComp: A dataset for visual table reading comprehension. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5773–5780, Albuquerque, New Mexico. Association for Computational Linguistics.

Carlos Gemmell and Jeff Dalton. 2023. ToolWriter: Question specific tool synthesis for tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16137–16148, Singapore. Association for Computational Linguistics.

Akash Ghosh, Venkata Sahith Bathini, Niloy Ganguly, Pawan Goyal, and Mayank Singh. 2024. How robust are the QA models for hybrid scientific tabular data? a study using customized dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8258–8264, Torino, Italia. ELRA and ICCL.

Jorge Osés Grijalba, L. Alfonso Ureña, Eugenio Martínez-Cámara, and Jose Camacho-Collados. 2026. The problem of ambiguity in table question answering. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3835–3848, Rabat, Morocco. Association for Computational Linguistics.

Jiawei Gu, Ziting Xian, Yuanzhen Xie, Ye Liu, Enjie Liu, Ruichao Zhong, Mochi Gao, Yunzhi Tan, Bo Hu, and Zang Li. 2025. Toward structured knowledge reasoning: Contrastive retrieval-augmented generation on experience. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23891–23910, Vienna, Austria. Association for Computational Linguistics.

Che Guan, Mengyu Huang, and Peng Zhang. 2024. Mfort-qa: Multi-hop few-shot open rich table question answering. In *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence*, ICCAI '24, page 434–442, New York, NY, USA. Association for Computing Machinery.

Manbir Gulati and Paul Roysdon. 2023. Tabmt: Generating tabular data with masked transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 46245–46254. Curran Associates, Inc.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453, Singapore. Association for Computational Linguistics.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Xinyi He, Yihao Liu, Mengyu Zhou, Yeye He, Haoyu Dong, Shi Han, Zejian Yuan, and Dongmei Zhang. 2025. TableLoRA: Low-rank adaptation on table structure understanding for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22376–22391, Vienna, Austria. Association for Computational Linguistics.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18206–18215.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Maximiliano Hormazábal Lagos, Álvaro Bueno Sáez, Pedro Alonso Doval, Jorge Alcalde Vesteiro, and Héctor Cerezo-Costas. 2025. Explicit-qa: Explainable code-based image table question answering. In *Progress in Artificial Intelligence: 24th EPIA Conference on Artificial Intelligence, EPIA 2025, Faro, Portugal, October 1–3, 2025, Proceedings, Part II*, page 339–351, Berlin, Heidelberg. Springer-Verlag.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 588–602, Online. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mengkang Hu, Haoyu Dong, Ping Luo, Shi Han, and Dongmei Zhang. 2024. KET-QA: A dataset for knowledge enhanced table question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9705–9719, Torino, Italia. ELRA and ICCL.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, WWW '20, page 2704–2710, New York, NY, USA. Association for Computing Machinery.

Sirui Huang, Yanggan Gu, Zhonghao Li, Xuming Hu, Li Qing, and Guandong Xu. 2025a. StructFact: Reasoning factual knowledge from structured data with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7521–7552, Vienna, Austria. Association for Computational Linguistics.

Sirui Huang, Hanqian Li, Yanggan Gu, Xuming Hu, Qing Li, and Guandong Xu. 2025b. Hyperg: Hypergraph-enhanced llms for structured knowledge. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1218–1228, New York, NY, USA. Association for Computing Machinery.

Sieun Hyeon, Jusang Oh, Sunghwan Steve Cho, and Jaeyoung Do. 2026. Mata: Multi-agent framework for reliable and flexible table question answering. *Preprint*, arXiv:2602.09642.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. 2016. Tables as semi-structured knowledge for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 474–483, Berlin, Germany. Association for Computational Linguistics.

Deyi Ji, Lanyun Zhu, Siqi Gao, Peng Xu, Hongtao Lu, Jieping Ye, and Feng Zhao. 2024. Tree-of-table: Unleashing the power of llms for enhanced large-scale table understanding. *ArXiv*, abs/2411.08516.

Changjiang Jiang, Fengchang Yu, Haihua Chen, Wei Lu, and Jin Zeng. 2025a. TabDSR: Decompose, sanitize, and reason for complex numerical reasoning in tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3172–3196, Suzhou, China. Association for Computational Linguistics.

Chuang Jiang, Mingyue Cheng, Xiaoyu Tao, Qingyang Mao, Jie Ouyang, and Qi Liu. 2026a. Tablemind: An autonomous programmatic agent for tool-augmented table reasoning. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM '26, page 260–270, New York, NY, USA. Association for Computing Machinery.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Jun-Peng Jiang, Yu Xia, Hai-Long Sun, Shiyin Lu, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2025b. Multimodal tabular reasoning with privileged structured information. *ArXiv*, abs/2506.04088.

Ruya Jiang, Chun Wang, and Weihong Deng. 2025c. Seek and solve reasoning for table question answering. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yangfan Jiang, Fei Wei, Ergute Bao, Yaliang Li, Bolin Ding, Yin Yang, and Xiaokui Xiao. 2026b. Accurate table question answering with accessible llms. *Preprint*, arXiv:2601.03137.

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

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: Recent advances. *Preprint*, arXiv:2207.05270.

Rihui Jin, Yu Li, Guilin Qi, Nan Hu, Yuan-Fang Li, Jiaoyan Chen, Jianan Wang, Yongrui Chen, Dehai Min, and Sheng Bi. 2025a. Hegta: Leveraging heterogeneous graph-enhanced large language models for few-shot complex table understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24294–24302.

Rihui Jin, Zheyu Xin, Xing Xie, Zuoyi Li, Guilin Qi, Yongrui Chen, Xinbang Dai, Tongtong Wu, and Gholamreza Haffari. 2025b. Table-r1: Self-supervised

and reinforcement learning for program-based table reasoning in small language models. *ArXiv*, abs/2506.06137.

Changwook Jun, Jooyoung Choi, Myoseop Sim, Hyun Kim, Hansol Jang, and Kyungkoo Min. 2022. Korean-specific dataset for table question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6114–6120, Marseille, France. European Language Resources Association.

Xiaoqiang Kang, Shengen Wu, Zimu Wang, Yilin Liu, Xiaobo Jin, Kaizhu Huang, Wei Wang, Yutao Yue, Xiaowei Huang, and Qiufeng Wang. 2025. Can GRPO boost complex multimodal table understanding? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12631–12644, Suzhou, China. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2022. AIT-QA: Question answering dataset over complex tables in the airline industry. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 305–314, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Rohit Khoja, Devanshu Gupta, Yanjie Fu, Dan Roth, and Vivek Gupta. 2025. Weaver: Interweaving SQL and LLM for table reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28282–28308, Suzhou, China. Association for Computational Linguistics.

Jongha Kim, Minseong Bae, Sanghyeok Lee, Jinsung Yoon, and Hyunwoo J. Kim. 2026. Tabflash: Efficient table understanding with progressive question conditioning and token focusing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22573–22581.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *ArXiv*, abs/2404.19205.

Atsushi Kojima. 2024. Sub-table rescorer for table question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15422–15427, Torino, Italia. ELRA and ICCL.

Kezhi Kong, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Chuan Lei, Christos Faloutsos, Huzefa Rangwala, and George Karypis. 2024. Opentab: Advancing large language models as open-domain table reasoners. In *The Twelfth International Conference on Learning Representations*.

Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. 2023. Open-WikiTable : Dataset for open domain question answering with complex reasoning over table. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8285–8297, Toronto, Canada. Association for Computational Linguistics.

Chia-Hsuan Lee, Oleksandr Polozov, and Matthew Richardson. 2021. KaggleDBQA: Realistic evaluation of text-to-SQL parsers. 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 2261–2273, Online. Association for Computational Linguistics.

Wonjin Lee, Kyumin Kim, Sungjae Lee, Jihun Lee, and Kwang In Kim. 2024a. Piece of table: A divide-and-conquer approach for selecting sub-tables in table question answering. *ArXiv*, abs/2412.07629.

Younghun Lee, Sungchul Kim, Ryan A. Rossi, Tong Yu, and Xiang Chen. 2024b. Learning to reduce: Towards improving performance of large language models on structured data. *ArXiv*, abs/2407.02750.

Fangyu Lei, Jinxiang Meng, Yiming Huang, Tinghong Chen, Yun Zhang, Shizhu He, Jun Zhao, and Kang Liu. 2025. Reasoning-table: Exploring reinforcement learning for table reasoning. *ArXiv*, abs/2506.01710.

Ce Li, Xiaofan Liu, Zhiyan Song, Ce Chi, Chen Zhao, Jingjing Yang, Zhendong Wang, Kexin Yang, Boshen Shi, Xing Wang, Chao Deng, and Junlan Feng. 2025a. Treb: A comprehensive benchmark for evaluating table reasoning capabilities of large language models. *ArXiv*, abs/2506.18421.

Fengyu Li, Junhao Zhu, Kaishi Song, Lu Chen, Zhongming Yao, Tianyi Li, and Christian S. Jensen. 2026. Replacing multi-step assembly of data preparation pipelines with one-step llm pipeline generation for table qa. *Preprint*, arXiv:2602.22721.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C.C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Liyao Li, Chao Ye, Wentao Ye, Yifei Sun, Zhe Jiang, Haobo Wang, Gang Chen, and Junbo Zhao. 2025b. Injecting learnable table features into LLMs.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proc. ACM Manag. Data*, 2(3).

Qianlong Li, Chen Huang, Shuai Li, Yuanxin Xiang, Deng Xiong, and Wenqiang Lei. 2025c. GraphOTTER: Evolving LLM-based graph reasoning for complex table question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5486–5506, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiao Li, Yawei Sun, and Gong Cheng. 2021. Tsqa: Tabular scenario based question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13297–13305.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. MMCoQA: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4220–4231, Dublin, Ireland. Association for Computational Linguistics.

Zheng Li, Yang Du, Mao Zheng, and Mingyang Song. 2025d. MiMoTable: A multi-scale spreadsheet benchmark with meta operations for table reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2548–2560, Abu Dhabi, UAE. Association for Computational Linguistics.

Hsing-Ping Liang, Che-Wei Chang, and Yao-Chung Fan. 2025. Improving table retrieval with question generation from partial tables. In *Proceedings of the 4th Table Representation Learning Workshop*, pages 217–228, Vienna, Austria. Association for Computational Linguistics.

Weizhe Lin, Rexhina Blloshmi, Bill Byrne, Adria de Gispert, and Gonzalo Iglesias. 2023. An inner table retriever for robust table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9909–9926, Toronto, Canada. Association for Computational Linguistics.

Chuang Liu, Junzhuo Li, and Deyi Xiong. 2023a. TabCQA: A tabular conversational question answering dataset on financial reports. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 196–207, Toronto, Canada. Association for Computational Linguistics.

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

Tianyang Liu, Fei Wang, and Muhao Chen. 2024. Rethinking tabular data understanding with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 450–482, Mexico City, Mexico. Association for Computational Linguistics.

Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. 2023b. Mmhqa-icl: Multimodal in-context learning for hybrid question answering over text, tables and images. *ArXiv*, abs/2309.04790.

Yujian Liu, Jiabao Ji, Tong Yu, Ryan A. Rossi, Sungchul Kim, Handong Zhao, Ritwik Sinha, Yang Zhang, and Shiyu Chang. 2025a. Augment before you try: Knowledge-enhanced table question answering via table expansion. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20769–20786, Suzhou, China. Association for Computational Linguistics.

Zhenghao Liu, Haolan Wang, Xinze Li, Qiushi Xiong, Xiaocui Yang, Yu Gu, Yukun Yan, Qi Shi, Fangfang Li, Ge Yu, and Maosong Sun. 2025b. Hippo: Enhancing the table understanding capability of large language models through hybrid-modal preference optimization. *ArXiv*, abs/2502.17315.

Boammani Aser Lompo and Marc Haraoui. 2025. Visual-tableqa: Open-domain benchmark for reasoning over table images. *Preprint*, arXiv:2509.07966.

Adrián López Gude, Roi Santos Ríos, Francisco Prado Valiño, Ana Ezquerro, and Jesús Vilares. 2025. LyS at SemEval 2025 task 8: Zero-shot code generation for tabular QA. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1282–1288, Vienna, Austria. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023a. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*.

Weizheng Lu, Jiaming Zhang, Jing Zhang, and Yueguo Chen. 2024. Large language model for table processing: A survey. *Frontiers Comput. Sci.*, 19:192350.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023b. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Xinyuan Lu, Liangming Pan, Yubo Ma, Preslav Nakov, and Min-Yen Kan. 2025a. TART: An open-source tool-augmented framework for explainable table-based reasoning. In *Findings of the Association*

for Computational Linguistics: NAACL 2025, pages 4323–4339, Albuquerque, New Mexico. Association for Computational Linguistics.

Xinyuan Lu, Liangming Pan, Yubo Ma, Preslav Nakov, and Min-Yen Kan. 2025b. TART: An open-source tool-augmented framework for explainable table-based reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4323–4339, Albuquerque, New Mexico. Association for Computational Linguistics.

Feng Luo, Hai Lan, Hui Luo, Zhifeng Bao, Xiaoli Wang, J. Shane Culpepper, and Shazia Sadiq. 2026a. Decomposition-driven multi-table retrieval and reasoning for numerical question answering. *Preprint*, arXiv:2603.07950.

Man Luo, Sharad Saxena, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Biotabqa: Instruction learning for biomedical table question answering. *Preprint*, arXiv:2207.02419.

Tongxu Luo, Fangyu Lei, Jiahe Lei, Weihao Liu, Shihu He, Jun Zhao, and Kang Liu. 2023. Hrot: Hybrid prompt strategy and retrieval of thought for table-text hybrid question answering. *ArXiv*, abs/2309.12669.

Zhizhao Luo, Zhaojing Luo, Meihui Zhang, and Rui Mao. 2026b. Tabtracer: Monte carlo tree search for complex table reasoning with large language models. *Preprint*, arXiv:2602.14089.

Arun Vignesh Malarkkan, Manan Roy Choudhury, Guangwei Zhang, Vivek Gupta, Qingyun Wang, Yanjie Fu, and Denghui Zhang. 2026. Finrule-bench: A benchmark for joint reasoning over financial tables and principles. *Preprint*, arXiv:2603.11339.

Qingyang Mao, Qi Liu, Zhi Li, Mingyue Cheng, Zheng Zhang, and Rui Li. 2026. PoTable: Towards Systematic Thinking via Plan-then-Execute Stage Reasoning on Tables . *IEEE Transactions on Knowledge & Data Engineering*, (01):1–13.

Suyash Vardhan Mathur, Jainit Sushil Bafna, Kunal Kartik, Harshita Khandelwal, Manish Shrivastava, Vivek Gupta, Mohit Bansal, and Dan Roth. 2024. Knowledge-aware reasoning over multimodal semi-structured tables. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14054–14073, Miami, Florida, USA. Association for Computational Linguistics.

Bhavnick Minhas, Anant Shankhdhar, Vivek Gupta, Divyanshu Aggarwal, and Shuo Zhang. 2022. XInfoTabS: Evaluating multilingual tabular natural language inference. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 59–77, Dublin, Ireland. Association for Computational Linguistics.

Raphaël Mouravieff, Benjamin Piwowarski, and Sylvain Lamprier. 2024. Learning relational decomposition of queries for question answering from tables. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10471–10485, Bangkok, Thailand. Association for Computational Linguistics.

Raphaël Mouravieff, Benjamin Piwowarski, and Sylvain Lamprier. 2025. Structural deep encoding for table question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2389–2402, Vienna, Austria. Association for Computational Linguistics.

Md Mahadi Hasan Nahid and Davood Rafiei. 2024a. NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3569–3585, Miami, Florida, USA. Association for Computational Linguistics.

Md Mahadi Hasan Nahid and Davood Rafiei. 2024b. TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5725–5737, Mexico City, Mexico. Association for Computational Linguistics.

Gaurav Najpande, Tampu Ravi Kumar, Manan Roy Choudhury, Neha Valeti, Yanjie Fu, and Vivek Gupta. 2026. Quiett: Query-independent table transformation for robust reasoning. *Preprint*, arXiv:2602.20017.

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

Giang Nguyen, Ivan Brugere, Shubham Sharma, Sanjay Kariyappa, Anh Totti Nguyen, and Freddy Lecue. 2025a. Interpretable LLM-based table question answering. *Transactions on Machine Learning Research*.

Phuc Nguyen, Nam Tuan Ly, Hideaki Takeda, and Atsuhiro Takasu. 2023. Tabiqa: Table questions answering on business document images. *ArXiv*, abs/2303.14935.

Thi-Nhung Nguyen, Hoang Ngo, Dinh Phung, Thuy-Trang Vu, and Dat Quoc Nguyen. 2025b. Improving table understanding with LLMs and entity-oriented search. In *Second Conference on Language Modeling*.

Thi-Nhung Nguyen, Hoang Ngo, Dinh Phung, Thuy-Trang Vu, and Dat Quoc Nguyen. 2025c. Planning for success: Exploring LLM long-term planning capabilities in table understanding. In *Proceedings of the 29th Conference on Computational Natural*

Language Learning, pages 81–92, Vienna, Austria. Association for Computational Linguistics.

Van-Quang Nguyen and Takayuki Okatani. 2026. CoReTab: Improving multimodal table understanding with code-driven reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6498–6523, Rabat, Morocco. Association for Computational Linguistics.

Ansong Ni, Srini Iyer, Dragomir Radev, Ves Stoyanov, Wen-tau Yih, Sida I. Wang, and Xi Victoria Lin. 2023. Lever: learning to verify language-to-code generation with execution. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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.

Vaishali Pal, Evangelos Kanoulas, Andrew Yates, and Maarten de Rijke. 2024. Table question answering for low-resourced Indic languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 75–92, Miami, Florida, USA. Association for Computational Linguistics.

Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. MultiTabQA: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6322–6334, Toronto, Canada. Association for Computational Linguistics.

Changzai Pan, Jie Zhang, Kaiwen Wei, Chenshuo Pan, Yu Zhao, Jingwang Huang, Jian Yang, Zhenhe Wu, Haoyang Zeng, Xiaoyan Gu, Weichao Sun, Yanbo Zhai, Yujie Mao, Zhuoru Jiang, Jiang Zhong, Shuangyong Song, Yongxiang Li, and Zhongjiang He. 2026. Reasontabqa: A comprehensive benchmark for table question answering from real world industrial scenarios. *Preprint*, arXiv:2601.07280.

Chaoxu Pang, Yixuan Cao, Chunhao Yang, and Ping Luo. 2024. Uncovering limitations of large language models in information seeking from tables. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1388–1409, Bangkok, Thailand. Association for Computational Linguistics.

Ankur 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)*, pages 1173–1186, Online. Association for Computational Linguistics.

Sungho Park, Jueun Kim, and Wook-Shin Han. 2026. SPARTA: Scalable and principled benchmark of tree-structured multi-hop QA over text and tables. In *The Fourteenth International Conference on Learning Representations*.

Sungho Park, Joohyung Yun, Jongwuk Lee, and Wook-Shin Han. 2025. HELIOS: Harmonizing early fusion, late fusion, and LLM reasoning for multi-granular table-text retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32424–32444, Vienna, Austria. Association for Computational Linguistics.

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

Sohan Patnaik, Heril Changwal, Milan Aggarwal, Sumit Bhatia, Yaman Kumar, and Balaji Krishnamurthy. 2024. CABINET: Content relevance-based noise reduction for table question answering. In *The Twelfth International Conference on Learning Representations*.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: a dataset for multimodal question answering on scientific papers. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Seho Pyo, Jiheon Seok, and Jaejin Lee. 2026. Reasoning by commented code for table question answering. *Preprint*, arXiv:2602.00543.

Zipeng Qiu, You Peng, Guangxin He, Binhang Yuan, and Chen Wang. 2024. Tqa-bench: Evaluating llms for multi-table question answering with scalable context and symbolic extension. *ArXiv*, abs/2411.19504.

Jinxiu Qu, Zirui Tang, Hongzhang Huang, Boyu Niu, Wei Zhou, Jiannan Wang, Yitong Song, Guoliang Li, Xuanhe Zhou, and Fan Wu. 2026. St-raptor: An agentic system for semi-structured table qa. *Preprint*, arXiv:2602.07034.

Abhishek Rajgaria, Kushagra Dixit, Mayank Vyas, Harshavardhan Kalalbandi, Dan Roth, and Vivek Gupta. 2025. No universal prompt: Unifying reasoning through adaptive prompting for temporal table reasoning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2800–2821, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Weijieying Ren, Tianxiang Zhao, Yuqing Huang, and Vasant Honavar. 2025. Deep learning within tabular data: Foundations, challenges, advances and future directions. *ArXiv*, abs/2501.03540.

Abhilash Shankarampeta, Harsh Mahajan, Tushar Kataria, Dan Roth, and Vivek Gupta. 2025. TRANSIENTTABLES: Evaluating LLMs' reasoning on temporally evolving semi-structured tables. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6526–6544, Albuquerque, New Mexico. Association for Computational Linguistics.

Mohamed Sharafath, Aravindh Annamalai, Ganesh Murugan, and Aravindakumar Venugopalan. 2026. N2n-gqa: Noise-to-narrative for graph-based table-text question answering using llms. *Preprint*, arXiv:2601.06603.

Chen Shen, Sajjadur Rahman, and Estevam Hruschka. 2025. Towards probabilistic question answering over tabular data. *Preprint*, arXiv:2506.20747.

Qi Shi, Han Cui, Haofeng Wang, Qingfu Zhu, Wanxiang Che, and Ting Liu. 2024. Exploring hybrid question answering via program-based prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11035–11046, Bangkok, Thailand. Association for Computational Linguistics.

Daiki Shirafuji, Koji Tanaka, and Tatsuhiko Saito. 2025. A hybrid search for complex table question answering in securities report. *Preprint*, arXiv:2511.09179.

Daixin Shu, Jian Yang, Zhenhe Wu, Xianjie Wu, Xianfu Cheng, Xiangyuan Guan, Yanghai Wang, Pengfei Wu, Tingyang Yang, Hualei Zhu, Wei Zhang, Ge Zhang, Jiaheng Liu, and Zhoujun Li. 2025. M3tqa: Massively multilingual multitask table question answering. *Preprint*, arXiv:2508.16265.

Jacob Si, Mike Qu, Michelle Lee, and Yingzhen Li. 2025. TabRAG: Tabular document retrieval via structured language representations. In *EurIPS 2025 Workshop: AI for Tabular Data*.

Anshul Singh, Chris Biemann, and Jan Strich. 2025. MTabVQA: Evaluating multi-tabular reasoning of language models in visual space. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19866–19891, Suzhou, China. Association for Computational Linguistics.

Atakan Site, Emre Erdemir, and Gülşen Eryiğit. 2025. ITUNLP at SemEval-2025 task 8: Question-answering over tabular data: A zero-shot approach using LLM-driven code generation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1504–1514, Vienna, Austria. Association for Computational Linguistics.

Josefa Lia Stoisser, Marc Boubnovski Martell, and Julien Fauqueur. 2025. Sparks of tabular reasoning via Text2SQL reinforcement learning. In *Proceedings of the 4th Table Representation Learning Workshop*, pages 229–240, Vienna, Austria. Association for Computational Linguistics.

Jan Strich, Enes Kutay Isgorur, Maximilian Trescher, Chris Biemann, and Martin Semmann. 2026. T2-RAGBench: Text-and-table aware retrieval-augmented generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 165–191, Rabat, Morocco. Association for Computational Linguistics.

Aofeng Su, Aowen Wang, Chaonan Ye, Chengcheng 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. *ArXiv*, abs/2411.02059.

Songyuan Sui, Hongyi Liu, Serena Liu, Li Li, Soo-Hyun Choi, Rui Chen, and Xia Hu. 2025. Chain-of-query: Unleashing the power of LLMs in SQL-aided table understanding via multi-agent collaboration. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 957–986, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. 2024. TAP4LLM: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10306–10323, Miami, Florida, USA. Association for Computational Linguistics.

Chaojie Sun, Bin Cao, Tiantian Li, Chenyu Hou, Ruizhe Li, and Jing Fan. 2026. Fgtr: Fine-grained multi-table retrieval via hierarchical llm reasoning. *Preprint*, arXiv:2603.12702.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Lei Tang, Wei Zhou, and Mohsen Mesgar. 2026. Exploring generative process reward modeling for semi-structured data: A case study of table question answering. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 407–420, Rabat, Morocco. Association for Computational Linguistics.

Zirui Tang, Boyu Niu, Xuanhe Zhou, Boxiu Li, Wei Zhou, Jiannan Wang, Guoliang Li, Xinyi Zhang, and Fan Wu. 2025. St-raptor: Llm-powered semi-structured table question answering. *Proc. ACM Manag. Data*, 3(6).

Ashish Thanga, Vibhu Dixit, Abhilash Shankarampeta, and Vivek Gupta. 2025. Evidence-guided schema normalization for temporal tabular reasoning. *Preprint*, arXiv:2512.00329.

Jiaming Tian, Liyao Li, Wentao Ye, Haobo Wang, Lingxin Wang, Lihua Yu, Zujie Ren, Gang Chen, and Junbo Zhao. 2026a. Toward real-world table agents: capabilities, workflows, and design principles for llm-based table intelligence. *World Wide Web*, 29(2).

Shi-Yu Tian, Zhi Zhou, Wei Dong, Kun-Yang Yu, Ming Yang, Zi-Jian Cheng, Lan-Zhe Guo, and Yu-Feng Li. 2026b. Tabularmath: Understanding math reasoning over tables with large language models. *Preprint*, arXiv:2505.19563.

Prasham Titiya, Jainil Trivedi, Chitta Baral, and Vivek Gupta. 2025. Mmtbench: A unified benchmark for complex multimodal table reasoning. *ArXiv*, abs/2505.21771.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Vishnou Vinayagame, Gregory Senay, and Luis Martí. 2025. Matata: Weakly supervised end-to-end mathematical tool-augmented reasoning for tabular applications. In *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025, Proceedings, Part I*, page 428–467, Berlin, Heidelberg. Springer-Verlag.

A. A. Vyatkin and V. D. Oliseenko. 2025. Generating pandas code for big table question answering using large language models. In *2025 XXVIII International Conference on Soft Computing and Measurements (SCM)*, pages 164–166.

Hanjun Wang, Wenda Liu, Qun Wang, Jiaming Xu, Shuai Nie, Yuchen Liu, and Runyu Shi. 2024a. Light-table-chain: Simplifying and enhancing chain-based table reasoning. In *2024 10th International Conference on Computer and Communications (ICCC)*, pages 269–275.

Hexuan Wang, Yaxuan Ren, Srikar Bommireddypalli, Shuxian Chen, Adarsh Prabhudesai, Rongkun Zhou, Elina Baral, and Philipp Koehn. 2026a. Scitarc: Benchmarking qa on scientific tabular data that requires language reasoning and complex computation. *Preprint*, arXiv:2603.08910.

Jia Wang, Chuanyu Qin, Mingyu Zheng, Qingyi Si, Peize Li, and Zheng Lin. 2026b. A closer look into llms for table understanding. *Preprint*, arXiv:2603.15402.

Lanrui Wang, Mingyu Zheng, Hongyin Tang, Zheng Lin, Yanan Cao, Jingang Wang, Xunliang Cai, and Weiping Wang. 2025a. NeedleinATable: Exploring long-context capability of large language models towards long-structured tables. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lexin Wang, Shenghua Liu, Yiwei Wang, Yujun Cai, Yuyao Ge, Jiayu Yao, Jiafeng Guo, and Xueqi Cheng. 2026c. Highlightbench: Benchmarking markup-driven table reasoning in scientific documents. *Preprint*, arXiv:2603.26784.

Qizhi Wang. 2026. Glean: Grounded lightweight evaluation anchors for contamination-aware tabular reasoning. *Preprint*, arXiv:2603.02212.

Tong Wang, Chi Jin, Yongkang Chen, Huan Deng, Xiaohui Kuang, and Gang Zhao. 2026d. Datafactory: Collaborative multi-agent framework for advanced table question answering. *Information Processing & Management*, 63(6):104723.

Yuqi Wang, Lyuhao Chen, Songcheng Cai, Zhijian Xu, and Yilun Zhao. 2024b. Revisiting automated evaluation for long-form table question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14696–14706, Miami, Florida, USA. Association for Computational Linguistics.

Yuxiang Wang, Junhao Gan, Shengxiang Gao, Shenghao Ye, Zhengyi Yang, and Jianzhong Qi. 2026e. Beyond linearization: Attributed table graphs for table reasoning. *Preprint*, arXiv:2601.08444.

Yuxiang Wang, Jianzhong Qi, and Junhao Gan. 2025b. Accurate and regret-aware numerical problem solver for tabular question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):12775–12783.

Zhensheng Wang, Wenmian Yang, Kun Zhou, Yiquan Zhang, and Weijia Jia. 2025c. Retqa: A large-scale open-domain tabular question answering dataset for real estate sector. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25452–25460.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. Tuta: Tree-based transformers for generally structured table pre-training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1780–1790, New York, NY, USA. Association for Computing Machinery.

Zhongyuan Wang, Richong Zhang, and Zhijie Nie. 2025d. General table question answering via answer-formula joint generation. *ArXiv*, abs/2503.12345.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Cornelius Wolff and Madelon Hulsebos. 2025. How well do LLMs reason over tabular data, really? In *Proceedings of the 4th Table Representation Learning Workshop*, pages 241–250, Vienna, Austria. Association for Computational Linguistics.

Jian Wu, Linyi Yang, Dongyuan Li, Yuliang Ji, Manabu Okumura, and Yue Zhang. 2025a. MMQA: Evaluating LLMs with multi-table multi-hop complex questions. In *The Thirteenth International Conference on Learning Representations*.

Pengzuo Wu, Yuhang Yang, Guangcheng Zhu, Chao Ye, Hong Gu, Xu Lu, Ruixuan Xiao, Bowen Bao, Yijing He, Liangyu Zha, Wentao Ye, Junbo Zhao, and Haobo Wang. 2025b. RealHiTBench: A comprehensive realistic hierarchical table benchmark for evaluating LLM-based table analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7105–7137, Vienna, Austria. Association for Computational Linguistics.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. 2025c. Tablebench: A comprehensive and complex benchmark for table question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25497–25506.

Xiaofeng Wu, Alan Ritter, and Wei Xu. 2025d. Tabular data understanding with llms: A survey of recent advances and challenges. *Preprint*, arXiv:2508.00217.

Xueqing Wu, Rui Zheng, Jingzhen Sha, Te-Lin Wu, Hanyu Zhou, Mohan Tang, Kai-Wei Chang, Nanyun Peng, and Haoran Huang. 2024. Daco: Towards application-driven and comprehensive data analysis via code generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 90661–90682. Curran Associates, Inc.

Zirui Wu and Yansong Feng. 2024. ProTrix: Building models for planning and reasoning over tables with sentence context. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4378–4406, Miami, Florida, USA. Association for Computational Linguistics.

Shiyu Xia, Junyu Xiong, Haoyu Dong, Jianbo Zhao, Yuzhang Tian, Mengyu Zhou, Yeye He, Shi Han, and Dongmei Zhang. 2024. Vision language models for spreadsheet understanding: Challenges and opportunities. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 116–128, Bangkok, Thailand. Association for Computational Linguistics.

Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Lingjiao Chen, Dongmei Zhang, Surajit Chaudhuri, and H. V. Jagadish. 2025a. MMTU: A massive multi-task table understanding and reasoning benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Dongmei Zhang, and Surajit Chaudhuri. 2025b. Table-LLM-specialist: Language model specialists for tables using iterative fine-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35443–35460, Suzhou, China. Association for Computational Linguistics.

Xiaobo Xing, Wei Yuan, Tong Chen, Quoc Viet Hung Nguyen, Xiangliang Zhang, and Hongzhi Yin. 2026. TableDART: Dynamic adaptive multi-modal routing for table understanding. In *The Fourteenth International Conference on Learning Representations*.

Weiqiang Xu, Yang Liu, Lingfeng Lu, Huakang Li, and Guozi Sun. 2026a. A comprehensive survey on table question answering: Datasets, methods and future directions. *Engineering Applications of Artificial Intelligence*, 174:114459.

Zhuoyan Xu, Haoyang Fang, Boran Han, Bonan Min, Bernie Wang, Cuixiong Hu, and Shuai Zhang. 2026b. Efficient table retrieval and understanding with multimodal large language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4327–4340, Rabat, Morocco. Association for Computational Linguistics.

Cui Yakun, Yanting Zhang, Zhu Lei, Jian Xie, Zhizhuo Kou, Hang Du, Zhenghao Zhu, and Sirui Han. 2026. Mmfctub: Multi-modal financial credit table understanding benchmark. *Preprint*, arXiv:2601.04643.

Bohao Yang, Yingji Zhang, Dong Liu, Andr'e Freitas, and Chenghua Lin. 2025a. Does table source matter? benchmarking and improving multimodal scientific table understanding and reasoning. *ArXiv*, abs/2501.13042.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025b. Socially aware language technologies: Perspectives and practices. *Computational Linguistics*, 51:689–703.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. TableFormer: Robust transformer modeling for table-text encoding. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 528–537, Dublin, Ireland. Association for Computational Linguistics.

Saisai Yang, Qingyi Huang, Jing Yuan, Liangyu Zha, Kai Tang, Yuhang Yang, Ning Wang, Yucheng Wei, Liyao Li, Wentao Ye, Hao Chen, Tao Zhang, Junlin Zhou, Haobo Wang, Gang Chen, and Junbo Zhao. 2025c. Tablegpt-r1: Advancing tabular reasoning through reinforcement learning. *Preprint*, arXiv:2512.20312.

Zhen Yang, Ziwei Du, Minghan Zhang, Wei Du, Jie Chen, Zhen Duan, and Shu Zhao. 2025d. Triples as the key: Structuring makes decomposition and verification easier in LLM-based tableQA. In *The Thirteenth International Conference on Learning Representations*.

Zheyuan Yang, Lyuhao Chen, Arman Cohan, and Yilun Zhao. 2025e. Table-r1: Inference-time scaling for table reasoning tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20605–20624, Suzhou, China. Association for Computational Linguistics.

Junyi Ye, Mengnan Du, and Guiling Wang. 2025. DataFrame QA: A universal llm framework on dataframe question answering without data exposure. In *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pages 575–590. PMLR.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 174–184, New York, NY, USA. Association for Computing Machinery.

Kun-Yang Yu, Zhi Zhou, Shi-Yu Tian, Xiao-Wen Yang, Zi-Yi Jia, Ming Yang, Zi-Jian Cheng, Lan-Zhe Guo, and Yu-Feng Li. 2026. Thinking with tables: Enhancing multi-modal tabular understanding via neuro-symbolic reasoning. *Preprint*, arXiv:2603.24004.

Peiying Yu, Guoxin Chen, and Jingjing Wang. 2025a. Table-critic: A multi-agent framework for collaborative criticism and refinement in table reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17432–17451, Vienna, Austria. 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.

Xiaohan Yu, Pu Jian, and Chong Chen. 2025b. TableRAG: A retrieval augmented generation framework for heterogeneous document reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14063–14082, Suzhou, China. Association for Computational Linguistics.

Yaoning Yu, Kai-Min Chang, Ye Yu, Kai Wei, Haojing Luo, and Haohan Wang. 2025c. Synthetic data-driven prompt tuning for financial qa over tables and documents. *Preprint*, arXiv:2511.06292.

Han Zhang, Yuheng Ma, and Hanfang Yang. 2025a. ALTER: Augmentation for large-table-based reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 179–198, Albuquerque, New Mexico. Association for Computational Linguistics.

Han Zhang, Yuheng Ma, and Hanfang Yang. 2025b. ALTER: Augmentation for large-table-based reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 179–198, Albuquerque, New Mexico. Association for Computational Linguistics.

Huajian Zhang, Mingyue Cheng, Yucong Luo, and Xiaoyu Tao. 2026a. Star: Towards effective and stable table reasoning via slow-thinking large language models. *Preprint*, arXiv:2511.11233.

Junwen Zhang, Pu Chen, and Yin Zhang. 2025c. Tablemoe: Neuro-symbolic routing for structured expert reasoning in multimodal table understanding. *ArXiv*, abs/2506.21393.

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

Wen Zhang, Long Jin, Yushan Zhu, Jiaoyan Chen, Zhiwei Huang, Junjie Wang, Yin Hua, Lei Liang, and Huajun Chen. 2025d. Trustuqa: a trustful framework for unified structured data question answering. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Xiaokang Zhang, Sijia Luo, Bohan Zhang, Zeyao Ma, Jing Zhang, Yang Li, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, Jifan Yu, Shu Zhao, Juanzi Li, and Jie Tang. 2025e. TableLLM: Enabling tabular data manipulation by LLMs in real office usage scenarios. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10315–10344, Vienna, Austria. Association for Computational Linguistics.

Xiaokang Zhang, Sijia Luo, Bohan Zhang, Zeyao Ma, Jing Zhang, Yang Li, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, Jifan Yu, Shu Zhao, Juanzi Li, and Jie Tang. 2025f. TableLLM: Enabling tabular data manipulation by LLMs in real office usage scenarios. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10315–10344, Vienna, Austria. Association for Computational Linguistics.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Baoxin Wang, Dayong Wu, Qingfu Zhu, and Wanxiang Che. 2024b. Flextaf: Enhancing table reasoning with flexible tabular formats. *ArXiv*, abs/2408.08841.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. 2024c. A survey of table reasoning with large language models. *Frontiers Comput. Sci.*, 19:199348.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. 2025g. A survey of table reasoning with large language models. *Front. Comput. Sci.*, 19(9).

Xuanliang Zhang, Dingzirui Wang, Baoxin Wang, Longxu Dou, Xinyuan Lu, Keyan Xu, Dayong Wu, and Qingfu Zhu. 2025h. SCITAT: A question answering benchmark for scientific tables and text covering diverse reasoning types. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3859–3881, Vienna, Austria. Association for Computational Linguistics.

Xuanliang Zhang, Dingzirui Wang, Keyan Xu, Qingfu Zhu, and Wanxiang Che. 2025i. MULTITAT: Benchmarking multilingual table-and-text question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 626–647, Suzhou, China. Association for Computational Linguistics.

Xuanliang Zhang, Dingzirui Wang, Keyan Xu, Qingfu Zhu, and Wanxiang Che. 2025j. RoT: Enhancing table reasoning with iterative row-wise traversals. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 559–579, Suzhou, China. Association for Computational Linguistics.

Xuanliang Zhang, Dingzirui Wang, Keyan Xu, Qingfu Zhu, and Wanxiang Che. 2026b. How do language models understand tables? a mechanistic analysis of cell location. *Preprint*, arXiv:2602.08548.

Yue Zhang, Seiji Maekawa, and Nikita Bhutani. 2026c. Same content, different representations: A controlled study for table QA. In *The Fourteenth International Conference on Learning Representations*.

Yuji Zhang, Qingyun Wang, Cheng Qian, Jiateng Liu, Chenkai Sun, Denghui Zhang, Tarek F. Abdelzaher, ChengXiang Zhai, Preslav Nakov, and Heng Ji. 2025k. Atomic reasoning for scientific table claim verification. *ArXiv*, abs/2506.06972.

Yunjia Zhang, Jordan Henkel, Avrilia Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M. Patel. 2024d. Reactable: Enhancing react for table question answering. *Proc. VLDB Endow.*, 17(8):1981–1994.

Zhehao Zhang, Yan Gao, and Jian-Guang Lou. 2024e. e^5: Zero-shot hierarchical table analysis using augmented LLMs via explain, extract, execute, exhibit and extrapolate. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1244–1258, Mexico City, Mexico. Association for Computational Linguistics.

Zhehao Zhang, Xitao Li, Yan Gao, and Jian-Guang Lou. 2023. CRT-QA: A dataset of complex reasoning question answering over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2131–2153, Singapore. Association for Computational Linguistics.

Bowen Zhao, Tianhao Cheng, Yuejie Zhang, Ying Cheng, Rui Feng, and Xiaobo Zhang. 2024a. Ct2c-qa: Multimodal question answering over chinese text, table and chart. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3897–3906, New York, NY, USA. Association for Computing Machinery.

Bowen Zhao, Changkai Ji, Yuejie Zhang, Wen He, Yingwen Wang, Qing Wang, Rui Feng, and Xiaobo Zhang. 2023a. Large language models are complex table parsers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14786–14802, Singapore. Association for Computational Linguistics.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Shu Wei, Binghong Wu, Lei Liao, Yongjie Ye, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2024b. Tabpedia: Towards comprehensive visual table understanding with concept synergy. In *Advances in Neural Information Processing Systems*, volume 37, pages 7185–7212. Curran Associates, Inc.

Yilun Zhao, Lyuhao Chen, Arman Cohan, and Chen Zhao. 2024c. TaPERA: Enhancing faithfulness and interpretability in long-form table QA by content planning and execution-based reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12824–12840, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. 2023b. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6081, Toronto, Canada. Association for Computational Linguistics.

Yujie Zhao, Lanxiang Hu, Yang Wang, Minmin Hou, Hao Zhang, Ke Ding, and Jishen Zhao. 2026. Stronger-MAS: Multi-agent reinforcement learning for collaborative LLMs. In *The Fourteenth International Conference on Learning Representations*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

Mingyu Zheng, Zhifan Feng, Jia Wang, Lanrui Wang, Zheng Lin, Hao Yang, and Weiping Wang. 2025. TableDreamer: Progressive and weakness-guided data synthesis from scratch for table instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7290–7315, Vienna, Austria. Association for Computational Linguistics.

Mingyu Zheng, Yang Hao, Wenbin Jiang, Zheng Lin, Yajuan Lyu, QiaoQiao She, and Weiping Wang. 2023. IM-TQA: A Chinese table question answering dataset with implicit and multi-type table structures. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5074–5094, Toronto, Canada. Association for Computational Linguistics.

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

Bangbang Zhou, Zuan Gao, Zixiao Wang, Boqiang Zhang, Yuxin Wang, Zhineng Chen, and Hongtao Xie. 2025a. SynTab-LLaVA: Enhancing Multimodal Table Understanding with Decoupled Synthesis . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24796–24806, Los Alamitos, CA, USA. IEEE Computer Society.

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. 2024. FREB-TQA: A fine-grained robustness evaluation benchmark for table question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2479–2497, Mexico City, Mexico. Association for Computational Linguistics.

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. 2025b. p²-TQA: A process-based preference learning framework for self-improving table question answering models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 217–231, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. 2025c. RITT: A retrieval-assisted framework with image and text table representations for table question answering. In *Proceedings of the 4th Table Representation Learning Workshop*, pages 86–97, Vienna, Austria. Association for Computational Linguistics.

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. 2025d. Texts or images? a fine-grained analysis on the effectiveness of input representations and models for table question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2307–2318, Vienna, Austria. Association for Computational Linguistics.

Wei Zhou, Mohsen Mesgar, Annemarie Friedrich, and Heike Adel. 2025e. Efficient multi-agent collaboration with tool use for online planning in complex table question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 945–968, Albuquerque, New Mexico. Association for Computational Linguistics.

Wei Zhou, Mohsen Mesgar, Annemarie Friedrich, and Heike Adel. 2025f. G-MACT at SemEval-2025 task 8: Exploring planning and tool use in question answering over tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 726–742, Vienna, Austria. Association for Computational Linguistics.

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.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning

over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 310–318, New York, NY, USA. Association for Computing Machinery.

Junnan Zhu, Jingyi Wang, Bohan Yu, Xiaoyu Wu, Junbo Li, Lei Wang, and Nan Xu. 2025. TableEval: A real-world benchmark for complex, multilingual, and multi-structured table question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7146, Suzhou, China. Association for Computational Linguistics.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Youcheng Pan, Xiaoqiang Zhou, and Min Zhang. 2026. Decoupling skeleton and flesh: Efficient multimodal table reasoning with disentangled alignment and structure-aware guidance. *Preprint*, arXiv:2602.03491.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Jiaru Zou, Dongqi Fu, Sirui Chen, Xinrui He, Zihao Li, Yada Zhu, Jiawei Han, and Jingrui He. 2026. RAG over tables: Hierarchical memory index, multi-stage retrieval, and benchmarking. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*.

Jiaru Zou, Soumya Roy, Vinay Kumar Verma, Ziyi Wang, David Wipf, Pan Lu, Sumit Negi, James Zou, and Jingrui He. 2025. Tattoo: Tool-grounded thinking prm for test-time scaling in tabular reasoning. *Preprint*, arXiv:2510.06217.

A Appendix

A.1 Comparing with Other Surveys

We compare our survey with recent work on table question answering (TQA) in Table 1. *TQA Setups* indicates whether a given survey proposes a TQA-specific task taxonomy or discusses multiple task setups within TQA. *Input Modality* specifies the types of input modalities considered in TQA. *Lan* denotes the languages of the benchmarks covered in the survey, and *Eval* indicates whether evaluation methodologies are discussed. *RLVR* and *ITPT* refer to reinforcement learning with verifiable rewards and interpretability, respectively. As shown in Table 1, our survey differs from prior work in the following ways: (1) We provide a fine-grained discussion of diverse TQA task setups and the modalities involved. (2) We include benchmarks covering languages beyond English. (3) We present an up-to-date review of modeling approaches with (M)LLMs and evaluation paradigms. (4) We discuss recent advances and emerging themes in the era of LLMs.

A.2 Survey Scope

We clarify the scope of this survey and describe the process used to collect the papers reviewed.

Tasks. This survey primarily focuses on TQA. We also include table fact verification, as this task can be readily reformulated into TQA. For example, by appending a question such as "Is the statement true or false?" to the statement being validated. Another related task is text-to-SQL, for which we mainly discuss relevant datasets, since they can also serve as benchmarks for TQA. We further consider SQL generation as an approach to solving TQA, but we do not provide a detailed review of methods specifically targeting SQL generation. Tasks that are not explicitly covered in this survey include table prediction, table generation, and table summarization, as these differ from TQA in terms of problem formulation and objectives.

Models. Given our focus on TQA in the era of LLMs, we primarily include work that leverages these models. For modeling papers, we restrict our collection to works published after 2022, as earlier studies are comprehensively reviewed in the survey by Jin et al. (2022).

Paper Collection. We search for papers using the keywords table/tabular reasoning, table/tabular question answering, and table/tabular understanding on both arXiv and the ACL Anthology. We include works published up to April 1st, 2026 and expand the collection by identifying relevant papers cited in the related work sections of the retrieved publications. In total, we compile a corpus of 277 papers. Figure 4 presents the distribution of papers by year and theme, illustrating a clear upward trend in research on table question answering.

A.3 Method Comparison

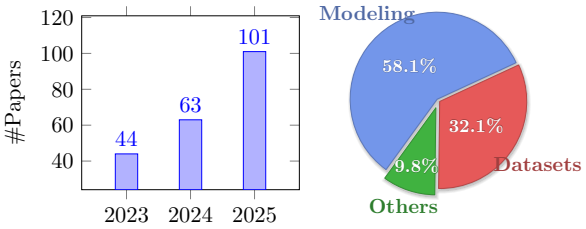
Figure 5 compares the performance of (multimodal) large language models ((M)LLMs)) in textual and image-based table understanding. The results are drawn from Deng et al. (2024) and Zheng et al. (2024).

A.4 TQA Datasets

Table 2 shows existing TQA datasets, categorized by features introduced in Section 2. Licenses are

Survey	Publication Year	TQA Setups	Input Modality	Lan	Modeling with LLMs	Eval	RLVR	ITPT	Summary
Dong et al. (2022)	2022	✗	text	en	✗	✗	✗	✗	table-pretraining
Jin et al. (2022)	2022	✓	text	en	✗	✗	✗	✗	benchmarks & pre-LLM modeling
Zhang et al. (2025g)	2024	✗	text	en	✓	✗	✗	✗	LLM-based modeling
Fang et al. (2024)	2024	✗	text	en	✓	p	✗	✓	table prediction, generation and understanding
Lu et al. (2024)	2024	✗	text, image	en	✓	p,r	✗	✗	tasks discussed via a data lifecycle aspect
Wu et al. (2025d)	2025	✗	text, image	en	✗	p,r	✗	✗	table representations and tasks
Tian et al. (2026a)	2025	✗	text, image	en	✓	p,r	✗	✗	LLM agents
Ren et al. (2025)	2025	✗	text	en	✓	✗	✗	✗	general table modeling with deep learning
Xu et al. (2026a)	2026	✗	text, image	en	✓	p	✓	✓	TQA from semantic parsing to PLM and to LLM
Ours	2026	✓	text, image, KB various		✓	p,r,e	✓	✓	fine-grained TQA task taxonomy, up-to-date modeling and discussion

Table 1: Comparing this work with recent surveys pertinent to table question answering from various perspectives. *TQA Setups*: if a fine-grained TQA task taxonomy is given. *Lan*: language of sourced benchmarks. *Eval*: Evaluation discussions. p,r,e stands for performance, robustness and explanation, respectively. *RLVR*: reinforcement learning with verifiable reward. *ITPT*: interpretability.



(a) #collected paper per year. (b) Paper types distribution.

Figure 4: Statistics of the collected paper. We show the number of collected paper by year as well as the distribution of different types of paper.

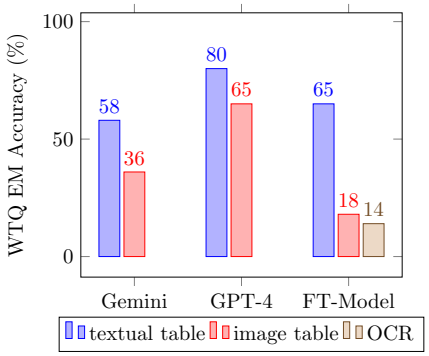


Figure 5: Performance of (M)LLMs in textual and image-based table understanding. FT-Model denotes fine-tuned models, specifically TableLlaVA-7B (Zheng et al., 2024) and TableLlaMA-7B (Zhang et al., 2024a). OCR refers to configurations in which image tables are first converted to text via optical character recognition (OCR) and then processed using TableLlaMA-7B.

subject to each individual dataset. Please refer to the original datasets for more information.

Table 2: Existing TQA Datasets. *MC* denotes multiple-choice. *ret* and *rea* refer to retrieval and reasoning, respectively. *LAN* indicates the language of the dataset. *Q* and *T* represent question and table, respectively. The value *both* in the table suggests that the dataset contains both single and multiple tables, or includes both flat and hierarchical tables.*DB* stands for database.

Datasets	Source/ Domain	Open domain	Answer format	Reasoning	LAN	Size		table features			additional inputs			
						#Q	#T	Single	Flat	Format	text	image	chart	KB
WTQ (Pasupat and Liang, 2015)	Wikipedia	✗	spans	ret, rea	EN	22k	2.1k	✓	✓	text	✗	✗	✗	✗
SQA (Iyyer et al., 2017)	WTQ	✗	spans	ret, rea	EN	17k		✓	✓	text	✗	✗	✗	✗
TabMCQ (Jauhar et al., 2016)	science exam	✗	MC		EN	9k	68	✓	✓	text	✗	✗	✗	✗
TabFact (Chen et al., 2020a)	Wikipedia	✗	MC	ret, rea	EN	118k	16k	✓	✓	text	✗	✗	✗	✗
WikiSQL (Zhong et al., 2017)	Wikipedia	✗	SQL	ret, rea	EN	80k	26k	✓	✓	text	✗	✗	✗	✗
Spider (Yu et al., 2018)	WikiSQL, Internet	✗	SQL	ret, rea	EN	10k	200	both	✓	DB	✗	✗	✗	✗
INFOTABS (Gupta et al., 2020)	Wikipedia	✗	MC	ret, rea	EN	23k	2.5k	✓	✓	text	✗	✗	✗	✗
KaggleDBQA (Lee et al., 2021)	Kaggle	✗	SQL	ret,rea	EN	272	8	both	✓	DB	✗	✗	✗	✗
HiTab (Cheng et al., 2022)	statistical report	✗	spans	ret, rea	EN	10k	3.6k	✓	✗	text	✗	✗	✗	✗
AIT-QA (Katsis et al., 2022)	airline	✗	spans	ret	EN	515	116	✓	both	text	✗	✗	✗	✗
FeTaQA (Nan et al., 2022)	ToTTo	✗	free-form	ret, rea	EN	10k	10k	✓	✓	text	✗	✗	✗	✗
TabMWP (Lu et al., 2023a)	websites	✗	MC, spans	numerical	EN	38k	37k	✓	✓	text, image	✗	✗	✗	✗
XINFOTABS (Minhas et al., 2022)	INFOTABS	✗	MC	ret, rea	MLT	23k	2.5k	✓	✓	text	✗	✗	✗	✗
EI-INFOTABS (Agarwal et al., 2022)	INFOTABS	✗	MC	ret, rea	HI	23k	2.5k	✓	✓	text	✗	✗	✗	✗
KorWikiTQ (Jun et al., 2022)	Wikipedia	✗	spans	ret, rea	KO	70k		✓	✓	text	✗	✗	✗	✗
TempTabTQA (Gupta et al., 2023)	Wikipedia	✗	spans	temporal	EN	11k	1.2k	✓	✓	text	✗	✗	✗	✗
RobuT (Zhao et al., 2023b)	WTQ, SQA, WikiSQL	✗	spans	robustness	EN	143k		✓	✓	text	✗	✗	✗	✗
IM-TQA (Zheng et al., 2023)	reports	✗	spans	ret,rea	ZH	5k	1.2k	✓	✗	text	✗	✗	✗	✗
Text2Analysis (He et al., 2024)		✗	code	ret,rea	EN	2.2k	347	✓	✓	text	✗	✗	✗	✗
SciTab (Lu et al., 2023b)	SciGen	✗	MC	ret,rea	EN	1.2k		✓	✓	text	✗	✗	✗	✗
CRT (Zhang et al., 2023)	TabFact	✗	spans	ret,rea	EN	728	423	✓	✓	text	✗	✗	✗	✗
Tab-CQA (Liu et al., 2023a)	reports	✗	spans	ret	ZH	10k	7k	✓		text	✗	✗	✗	✗
BIRD (Li et al., 2023)	internet	✗	SQL	ret,rea	EN	12.7k	95	both	✓	DB	✗	✗	✗	✗
LF-TQA (Wang et al., 2024b)	FeTAQA, QTSUMM	✗	free-form	ret,rea	EN	2.9k		✓	✓	text	✗	✗	✗	✗
Indic-TQA (Pal et al., 2024)	Wikipedia	✗	spans	ret,rea	BN,Hi	2m	21k	✓	✓	text	✗	✗	✗	✗
TableBench (Wu et al., 2025c)	existing datasets	✗	spans, free-form	ret,rea	EN	20k	3.6k	✓	✓	text	✗	✗	✗	✗
Databench (Osés Grijalba et al., 2024)	internet	✗	spans	ret,rea	EN	1.8k	65	✓	✓	DB	✗	✗	✗	✗
DACO (Wu et al., 2024)	Spider, Kaggle	✗	free-form	ret,rea	EN	1.9k	440	✗	✓	DB	✗	✗	✗	✗
TabIS (Pang et al., 2024)	ToTTo HiTab	✗	MC	ret,rea	EN	61k		✓	✓	DB	✗	✗	✗	✗
FREB-TQA (Zhou et al., 2024)	existing datasets	✗	spans	robustness	EN	8.5k		✓	✓	DB	✗	✗	✗	✗
RealTabBench (Su et al., 2024)	existing datasets	✗	free-form	robustness	EN,ZH	6k	360	✓	both	csv	✗	✗	✗	✗

Datasets	Source/ Domain	Open domain	Answer format	Reasoning	LAN	Size		table features			additional inputs			
						#Q	#T	Single	Flat	Format	text	image	chart	KB
TabularMath (Tian et al., 2026b)	GSM8K	✗	number	ret,rea	EN	6.7k	3.3k	✓	✓	text image	✗	✗	✗	✗
NIAT (Wang et al., 2025a)	WTQ, HiTab AIT-QA	✗	spans	ret	EN	287k	750	✓	both	text	✗	✗	✗	✗
HCT-QA (Ahmad et al., 2025)	internet	✗	spans	ret,rea	EN,AR	77k	6.7k	✓	✗	text image	✗	✗	✗	✗
TableEval (Zhu et al., 2025)	reports	✗	spans	ret,rea	EN,ZH	2.3k	617	✓	both	csv	✗	✗	✗	✗
SciAtomicBench (Zhang et al., 2025k)	PubTables MatSciTable	✗	MC	ret,rea	EN	2.5k		✓	both	text	✗	✗	✗	✗
AraTable (Alshaikh et al., 2025)	Internet	✗	spans	ret,rea	AR	615	41	✓	✓	text	✗	✗	✗	✗
RealHiTBench (Wu et al., 2025b)	online platforms	✗	free-form	ret,rea	EN	3.7k	708	both	✗	text image	✗	✗	✗	✗
TReB (Li et al., 2025a)	existing datasets	✗	free-form	ret,rea	EN,ZH	7.8k		✓	both	text	✗	✗	✗	✗
M3TQA (Shu et al., 2025)	reports	✗	spans	ret,rea	MLT	46k	50	✓	both	text	✗	✗	✗	✗
TableDreamer (Zheng et al., 2025)	synthesized	✗	free-form	ret,rea	EN	27k		✓	✓	text	✗	✗	✗	✗
RUST-BENCH (Abhyankar et al., 2025a)	NSF, Sportsett	✗	spans	ret,rea	EN	7.9k	2k	✓	✓	text	✗	✗	✗	✗
SciTaRC (Wang et al., 2026a)	arXiv	✗	free–form	ret,rea	EN	370k	3.7k	✓	both	LaTex	✗	✗	✗	✗
MMQA (Wu et al., 2025a)	Spider	✗	spans	ret,rea	EN	3.3k	3.3k	✗	✓	DB	✗	✗	✗	✗
MultiTableQA (Zou et al., 2026)	existing datasets	✗	spans	ret,rea	EN	23k	57k	✗	✓	text	✓	✗	✗	✗
TRANSIENTTABLES (Shankarampeta et al., 2025)	Wikipedia	✗	spans	temporal	EN	3.9k	14k	✗	✓	text	✗	✗	✗	✗
MiMoTable (Li et al., 2025d)	internet	✗	free-form	ret,rea	EN,ZH	1.7k	428	both	both	csv	✗	✗	✗	✗
TQA-Bench (Qiu et al., 2024)	world-Bank BIRD	✗	MC	ret,rea	EN	56k	10	✗	✓	DB	✗	✗	✗	✗
ReasonTabQA (Pan et al., 2026)	Internet	✗	spans	ret,rea	EN,ZH	5.5k	1.9k	✗	✗	text	✗	✗	✗	✗
ComTQA (Zhao et al., 2024b)	FinTabNet PubTab1M D	✗	spans	ret,rea	EN	9k	1.5k	✓	both	image	✗	✗	✗	✗
MMTab (Zheng et al., 2024)	existing datasets	✗	spans	ret,rea	EN	49k	23k	✓	both	image	✗	✗	✗	✗
MTabVQA (Singh et al., 2025)	existing datasets	✗	spans	ret,rea	EN	3.7k	8.5k	✗	✓	image	✗	✗	✗	✗
MMSci (Yang et al., 2025a)	SciGen	✗	spans	ret,rea	EN	15k	52k	✓	✓	image	✗	✗	✗	✗
TabComp (Gautam et al., 2025)	DocVQA	✗	free-form	ret,rea	EN	30k	10k	✓	both	image	✗	✗	✗	✗
TableVQA-Bench (Kim et al., 2024)	WTQ, TabFact FinTabNet	✗	spans	ret,rea	EN	1.5k	894k	✓	✓	image	✗	✗	✗	✗
TABLET (Alonso et al., 2026)	existing datasets	✗	spans	ret,rea	EN	4000k	2000k	✓	both	image	✗	✗	✗	✗
MMFCTUB (Yakun et al., 2026)	synthesized	✗	MC	ret,rea	EN	25k	19k	✓	✓	image	✗	✗	✗	✗
Visual-TableQA (Lompo and Haraoui, 2025)	synthesized	✗	free-form	ret,rea	EN	6k	2.5k	✓	both	image	✗	✗	✗	✗
CoReTab (Nguyen and Okatani, 2026)	MMTab	✗	spans	ret,rea	EN	115k		✓	both	image	✗	✗	✗	✗
OTT-QA (Chen et al., 2021a)	Wikipedia	✓	spans	ret,rea	EN	45k			✓	text	✓	✗	✗	✗
TANQ (Akhtar et al., 2024)	QAMPARI Wikipedia	✓	table	ret,rea	EN	43k			✓	text	✓	✗	✗	✗
RETQA (Wang et al., 2025c)	QAMPARI Wikipedia	✓	free-form	ret,rea	ZH	20k	4.9k		✓	DB	✓	✗	✗	✗
Open-WikiTable (Kweon et al., 2023)	WTQ WikiSQL	✓	spans	ret,rea	EN	67k	24k		✓	text	✗	✗	✗	✗

Datasets	Source/ Domain	Open domain	Answer format	Reasoning	LAN	Size		table features			additional inputs			
						#Q	#T	Single	Flat	Format	text	image	chart	KB
CompMix (Christmann et al., 2024)	CONVMIX	✓	spans	ret,rea	EN	9.4k		✓	✓	text	✓	✗	✗	✓
T^2-RAGBench (Strich et al., 2026)	existing datasets	✓	spans	ret,rea	EN	32k				text	✓	✗	✗	✗
MMCoQA (Li et al., 2022)	MMQA	✓	spans	ret,rea	EN	5.7k	10k			text	✓	✓	✗	✗
MMTBENCH (Titiya et al., 2025)	Internet	✗	spans	ret,rea	EN	4k	500	✓	both	csv image	✗	✓	✓	✗
KET-QA (Hu et al., 2024)	HybridTQA	✗	spans	ret,rea	EN	9.4k	5.7k	✓	✓	text	✗	✗	✗	✓
CT2C-QA (Zhao et al., 2024a)	reports	✗	spans	ret,rea	ZH	9.9k	369			html	✓	✗	✓	✗
mmtabqa (Mathur et al., 2024)	existing datasets	✗	spans	ret,rea	EN	69k	259	✓	✓	text	✗	✓	✗	✗
StructFact (Huang et al., 2025a)	existing datasets	✗	MC	ret,rea	EN	13k		✓	✓	text	✓	✗	✗	✓
WikiMixQA (Foroutan et al., 2025)	WTabHTML	✗	MC	ret,rea	EN	1k		both	✓	html	✓	✓	✓	✗
SPIQA (Pramanick et al., 2024)	arXiv	✗	free-form	ret,rea	EN	270k	117k	both	both	image	✓	✓	✓	✗
MMQA (Talmor et al., 2021)	Wikipedia	✗	spans	ret,rea	EN	30k		✓	✓	text	✓	✓	✗	✗
SciTabQA (Ghosh et al., 2024)	SciGen	✗	spans	ret,rea	EN,AR	822	198	✓	✓	text	✓	✗	✗	✗
MULTITAT (Zhang et al., 2025i)	existing datasets	✗	spans	ret,rea	MLT	250		✓	✓	text	✓	✗	✗	✗
SciTAT (Zhang et al., 2025h)	arXiv	✗	spans free-form	ret,rea	EN	13k		✓	✓	text	✓	✗	✗	✗
PACIFIC (Deng et al., 2022)	TATQA	✗	spans	ret,rea	EN	19k		✓	✓	text	✓	✗	✗	✗
TAT-QA (Zhu et al., 2021)	reports	✗	spans	ret,rea	EN	16k		✓	✓	text	✓	✗	✗	✗
GeoTSQA (Li et al., 2021)	exams	✗	MC	ret,rea	ZH	1k		✗		text	✓	✗	✗	✗
HybridQA (Chen et al., 2020b)	Wikipedia	✗	spans	ret	EN	70k	13k	✓	✓	text	✓	✗	✗	✗
FinQA (Chen et al., 2021b)	reports	✗	spans	ret,rea	EN	8k		✓	✓	text	✓	✗	✗	✗
MultiHiertt (Zhao et al., 2022)	FinTabNet	✗	spans	ret,rea	EN	10k		✗	✗	text	✓	✗	✗	✗
SPARTA (Park et al., 2026)	ROTOWIRE	✗	spans	ret,rea	EN	3.3k		✗	✓	DB	✓	✗	✗	✗