Tabular Data Understanding with LLMs: A Survey of Recent Advances and Challenges

Anonymous ACL submission

Abstract

Tables have gained significant attention in large language models (LLMs) and multimodal large language models (MLLMs) due to their complex and flexible structure. Unlike linear text inputs, tables are two-dimensional, encompassing formats that range from well-structured database tables to complex, multi-layered spreadsheets, each with different purposes. This diversity in format and purpose has led to the development of specialized methods and tasks, instead of universal approaches—making navigation of table understanding tasks challenging. To address these challenges, this paper introduces key concepts through a taxonomy of tabular input representations and an introduction of table understanding tasks. We highlight several critical gaps in the field that indicate the need for further research: (1) the predominance of retrieval-focused tasks that require minimal reasoning beyond mathematical and logical operations; (2) significant challenges faced by models when processing complex table structures, large-scale tables, length context, or multi-table scenarios; and (3) the limited generalization of models across different tabular representations and formats.

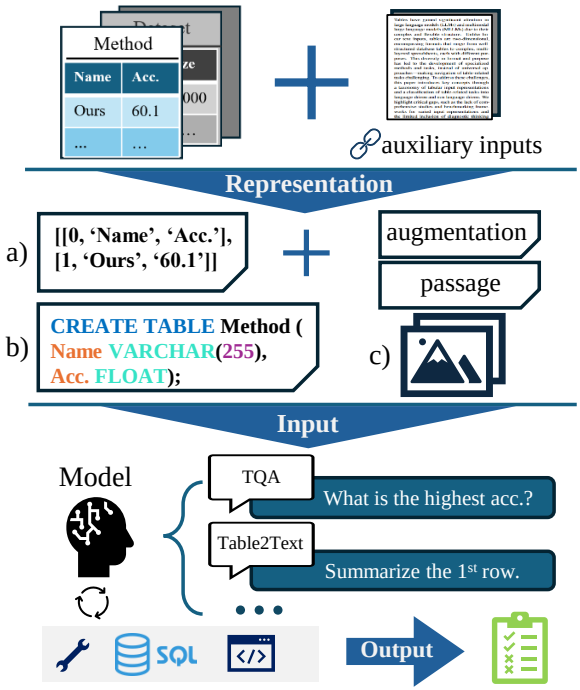

Figure 1: Workflow of table-related tasks in large models. Tables or databases, possibly accompanied by additional input data, are transformed into input representations, which could take the form of (a) serialization, (b) database schema, (c) images, or other format with optional augmentations. These inputs are then processed by models usually leveraging SQL, and other tools to generate task specific outputs.

1 Introduction

Tables have garnered increasing attention due to advances in large language models (LLMs) and multi-modal large language models (MLLMs), owing to the unique challenges they present. Unlike linear text, tabular data possess an inherently visual, two-dimensional format that requires specialized pipelines to be processed effectively, as shown in Figure 1. Additionally, tables exhibit structural flexibility, serving a wide range of purposes—from well-structured database tables to hierarchical, multi-layered spreadsheets and multimedia-linked info-boxes. These variations in purpose and structure have driven the development of diverse input representations, tasks, and specialized methods and datasets. However, such specialization often comes at the expense of universality (Zhang et al., 2024a), making it difficult for new researchers to navigate the field effectively. While existing surveys (Fang et al., 2024; Zhang et al., 2024b; Lu et al., 2024; Badaro et al., 2023; Ren et al., 2025) have explored various prompting, training, and transformer-based methods for table processing, there is a need for a comprehensive survey that uncovers new opportunities, focusing on tasks and benchmarks in tabular understanding.

To address the existing gap and assist researchers in navigating table-related tasks, this paper presents a systematic taxonomy of tabular data representations and introduces a broad range of both well-

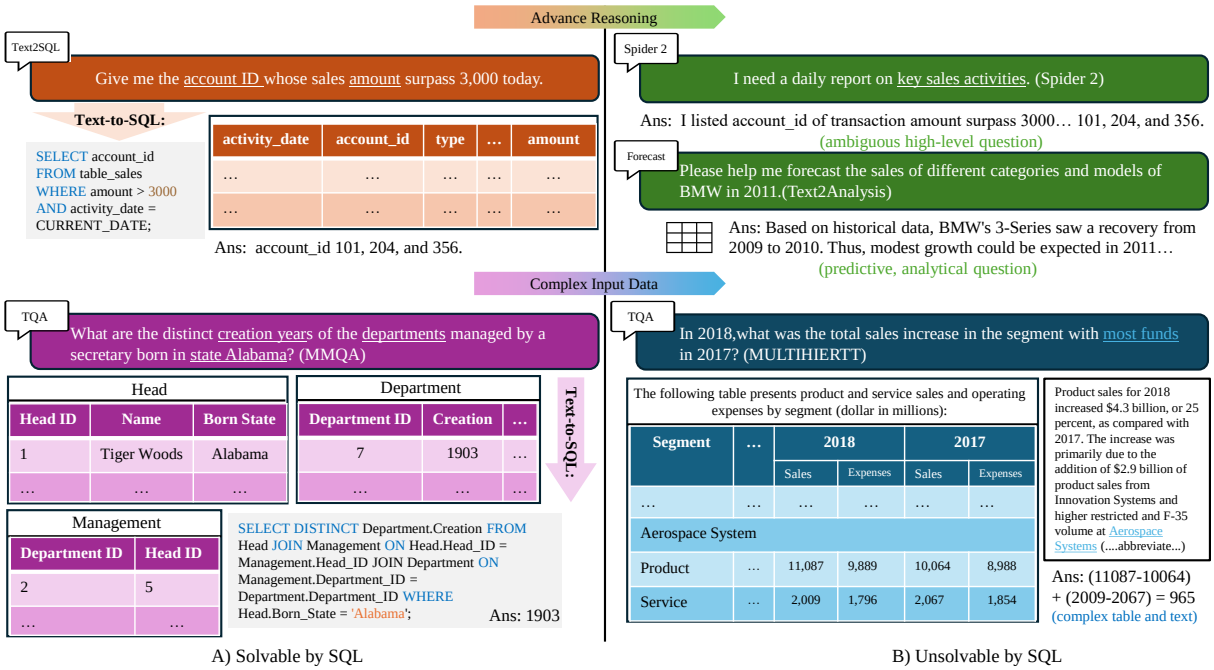

Figure 2: The left side illustrates examples of tasks that can be addressed with SQL-based methods such as typical Text-to-SQL task and a Table QA task from MMQA (Anonymous, 2024). In contrast, the right side presents tasks that demand advanced reasoning or involve complex inputs, such as those found in Spider 2 (Lei et al., 2024), Text2Analysis (He et al., 2024), and MULTIHIERTT (Zhao et al., 2022), which go beyond the capabilities of SQL-based approaches.

established and novel tasks. For instance, we examine *Table QA*, which focuses on answering natural language questions based on table content, and *Table-to-Text*, which involves generating natural language summaries from tabular data. We also highlight innovative tasks such as *leaderboard construction*, which aggregates result tables from scientific papers to provide a comprehensive comparison of methods in one specific field. For well-established tasks, we compile key benchmarks and their associated table formats, categorizing improvements in newer benchmarks relative to earlier ones to highlight emerging research trends.

Furthermore, our survey reveals new opportunities by focusing on tasks and challenges identified in widely used benchmarks. Despite significant progress in prompting and training methods—as highlighted in existing surveys (Lu et al., 2024; Badaro et al., 2023; Ren et al., 2025)—and the robust performance of recent tabular foundational models that integrate tabular data during the pre-training and fine-tuning stages of 72B base models (Su et al., 2024), current table processing benchmarks tend to concentrate on limited reasoning tasks and often rely on simplistic, synthetic tables with inconsistent input representations. While effective for initial evaluations, these benchmarks fall short in assessing the performance of more advanced methods and models in real-world scenarios that require higher-level reasoning and the processing of complex inputs, ultimately limiting their generalizability and broader applicability.

2 Findings and Future Direction

In this section, we outline three key findings that underscore the need for further investigation.

2.1 Limited Scope Beyond Mathematical Reasoning

Recent work has begun to saturate performance on many widely used benchmarks. For example, question-decomposition pipelines have yielded significant improvements (Gao et al., 2023; Ye et al., 2023; Wang et al., 2024b); the method proposed by Hussain (2025) achieved over 80% accuracy on the Wiki-Table Questions benchmark (Pasupat and Liang, 2015) and more than 93% on TabFact (Chen et al., 2020b), two popular datasets for table QA and fact verification. Moreover, the success of table foundation models—integrating specialized table encoders into large-scale language models pre-trained and fine-tuned on tabular data (Su et al., 2024)—signals a growing trend toward applying tabular methods to larger models. These advances suggest it is time to move beyond data retrieval-based tasks, as most benchmarks rely on detailed

queries that prompt models to extract specific information from tables using logical operations.

Many existing benchmarks are even constructed by first generating SQL queries or sequences of mathematical expressions, which are then translated into natural language query (Pasupat and Liang, 2015; Iyyer et al., 2017; Pal et al., 2023; Anonymous, 2024), or by framing questions whose answers can be fully derived using mathematical functions (Zheng et al., 2023; Zhang et al., 2023d; Zhao et al., 2022; Kweon et al., 2023). While efforts have focused on enhancing task complexity through additional reasoning steps or embedding complex mathematical functions, the core structure of these tasks remains fundamentally unchanged. As shown in Figure 2, such descriptive questions can be solved relatively easily by text-to-SQL methods when tables are well-structured.

Notably, recent work (Majumder et al., 2024) has further pushed the boundaries by emphasizing higher-order reasoning skills. For example, He et al. (2024) introduced tasks that extend beyond basic descriptive analysis, such as insight identification, similar to what is shown in Figure 3, which demands diagnostic thinking; forecasting, which requires predictive thinking; and chart creation from ambiguous queries, a task that requires prescriptive thinking—selecting the appropriate chart type and determining optimal intervals to produce visually appealing figures. In these tasks, models cannot simply rely on finding synonyms or related attributes in the table to perform data retrieval. Instead, they must understand the overall context of the table and the user's intent to address the query.

A similar direction is explored by Spider 2 (Lei et al., 2024), which introduces questions requiring higher levels of reasoning. Unlike benchmarks such as Spider (Yu et al., 2018) and its extensions, which introduce marginal difficulties by swapping explicit schema names with synonyms or rephrasing utterances (Deng et al., 2021; Gan et al., 2021a), Spider 2 presents high-level, intent-driven queries, as illustrated in Figure 2. For example, instead of asking explicitly (e.g., "*Give me the account ID whose sales surpass a threshold today*"), Spider 2 poses abstract, goal-oriented queries (e.g., "*I need a daily report on key sales activities*"). These queries challenge models to infer the user's intent, requiring a deep understanding of both the database schema and the query's broader context. Furthermore, Dong et al. (2025) introduce multi-turn conversations that teach models to seek clarification whenever a user's initial query is ambiguous, thereby better mirroring real-world interactions and mitigating the multiple-interpretation issue identified by Pourreza and Rafiei (2023b).

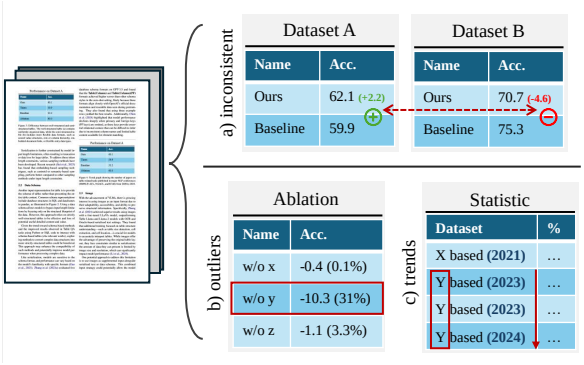

Figure 3: Illustration of the proposed task: Scientific Document Understanding with Tables which require diagnosing implicit knowledge embedded in tabular data, which may not be well addressed in text. Examples include: a) inconsistent results under conditions; b) outliers in values; and c) key trends.

2.2 Lack of Robustness on Input Complexity

Another area of opportunity in current table-related research is enhancing model robustness when processing complex input scenarios, including intricate table structures, long tables, lengthy texts, and multi-table contexts—challenges that have minimal impact on human performance (Anonymous, 2024; Pal et al., 2023). Benchmarks such as HiTab (Cheng et al., 2022) and MULTIHIERTT (Zhao et al., 2022) have been instrumental in highlighting these challenges. HiTab features hierarchical multidimensional tables, while MULTIHIERTT further incorporates lengthy texts where answers may be embedded, as well as multi-table scenarios. Both benchmarks report model performances below 50%, compared to a human accuracy of around 83% on MULTIHIERTT. Similarly, benchmarks like MultiTableQA (Pal et al., 2023) and MMQA (Anonymous, 2024), which focus on multi-table question answering from well-structured databases such as those in the Spider benchmark, provide valuable insights into current model limitations. For instance, in MMQA the strongest model evaluated, o1-preview (OpenAI, 2024), achieves an exact match score slightly above 50%, while human performance reaches approximately 89%.

Scientific Document Understanding with Tables. Scientific documents provide a rich test bed for information extraction and table extraction (Bai et al., 2024; Yang et al., 2022; Kardas et al., 2020). These papers typically contain complex ablation,

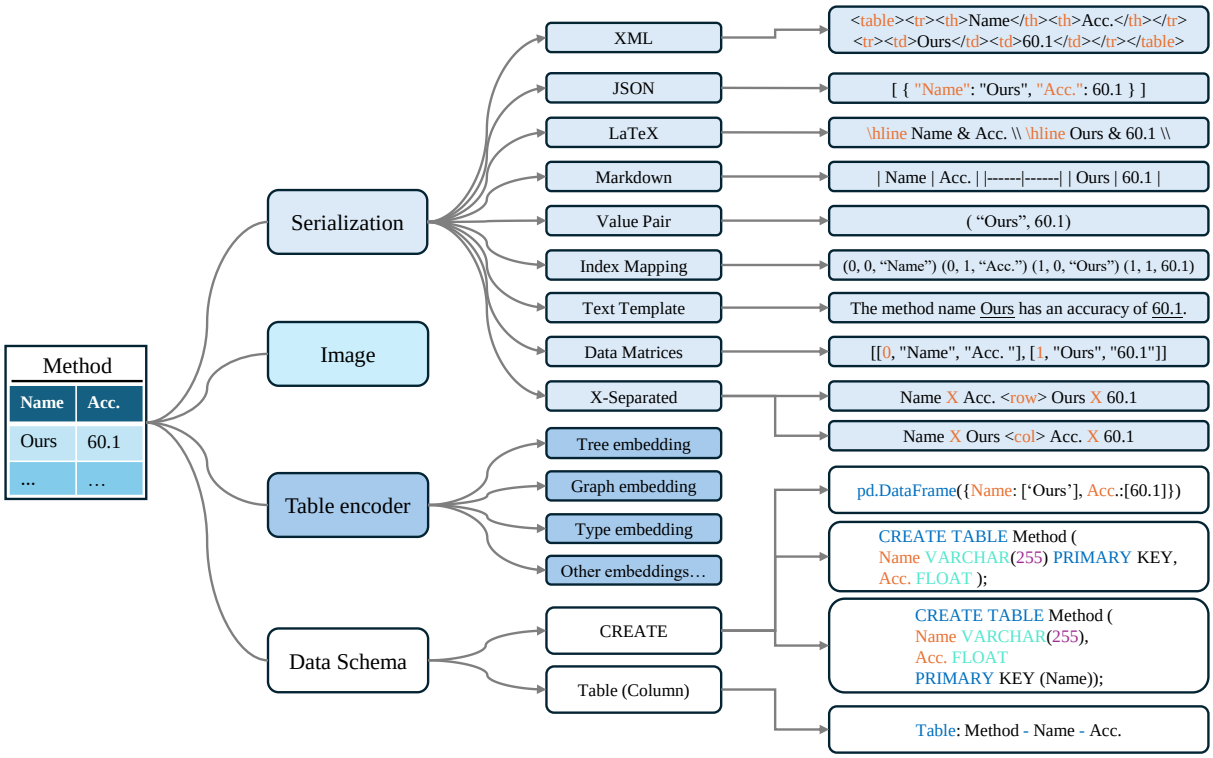

Figure 4: Taxonomy of table input representation methods, encompassing serialization, image, specialized table encoders, and data schema. Examples illustrating each representation type are shown on the right.

analysis, and method-comparison tables alongside extensive textual discussion, all of which demand sophisticated reasoning for accurate interpretation (Zhang et al., 2023c; Asai et al., 2024). Building on this foundation, future work can harness scientific-document data to develop higher-level table-reasoning systems that demand a broad repertoire of skills—such as trend detection, diagnostic assessment, and forecasting (see Figure 3).

2.3 Limited Generalization Across Tabular Representations

Despite recent advances, current models still struggle to generalize across diverse tabular representations. Their performance on commonly used benchmarks can vary by up to 5% depending on how closely input formats align with the data encountered during pretraining (Sui et al., 2024), as similarly observed by Gao et al. (2023) in the Text-to-SQL domain. Benchmarks highlight this issue by relying on a variety of input representations chosen based on convenience and accessibility. As demonstrated in our collection of major benchmarks (see Tables 1, 2, and 3), tabular representations for the same type of task lack universality. Even when categorized under the same format, such as JSON, the internal structures can vary greatly (Aly et al., 2021; Chen et al., 2020c), further complicating

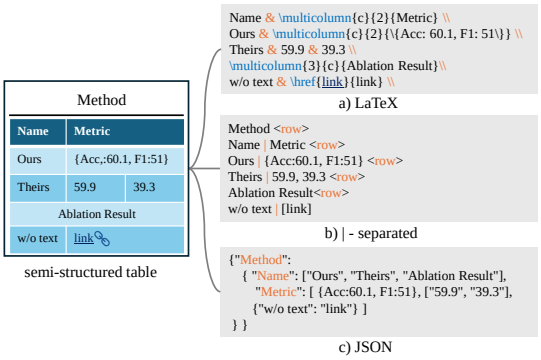

Figure 5: Comparison of serialization methods for semi-structured tables: a) LaTeX, b) X-separated, and c) JSON. Each method has its strengths and weaknesses in handling aspects such as nested value structures, row or column hierarchies, embedded document links, and flexible data types.

performance evaluations and introducing bias.

Efforts to address these inconsistencies are emerging. For example, Lei et al. (2023) provides standardized serialization options such as Markdown and flattened text, though additional formats remain underexplored. Another line of research (Zheng et al., 2024) focuses on visual representations of complex tables—such as Table Cell Locating and Merged Cell Detection—to generate serialized versions from images. Integrating these tasks into fine-tuning pipelines has proven beneficial.

Future research could explore serialization-to-serialization tasks, where models transform one

format (e.g., JSON) into another (e.g., LaTeX or Markdown). Integrating such task could enhance models' robustness to varied input styles and create opportunities for fine-tuning across multiple representations. Additionally, limited investigation has been conducted into the effectiveness of different representations for complex tables. For instance, LaTeX's `\multicolumn` command effectively captures hierarchical structures, whereas other formats may ignore this type of relationship during serialization process, as Figure 5 shown.

3 Modalities of Table Representation

In this section, we introduce key tabular representations that are essential for enabling large models to process table data effectively. Since these models require one-dimensional input formats, structured, two-dimensional tables must be converted accordingly. This transformation, however, often results in the loss of valuable structural information. To address these challenges, various methods have been developed, including serialization, database schema representations, image-based formats, and specialized table encoders, as illustrated in Figure 4. Recent studies (Sui et al., 2024; Zhang et al., 2023a) demonstrate that model performance is sensitive to the chosen input representation, underscoring the data-dependent nature of current approaches to processing tabular data. Unfortunately, many existing benchmarks rely on representations selected primarily for convenience (Sundararajan et al., 2024), lacking of robust, unbiased comparisons.

3.1 Serialization

Serialization has long been a common method for representing tabular data, transforming tables into serialized text. Its primary advantages lie in compatibility with standard models and ease of access to existing formats, such as HTML or Markdown tables on the web, LaTeX tables in PDF documents, and JSON or key-value pairs in code environments (see Figure 4). Most current benchmarks rely on serialization, as illustrated in Tables 1, 2, and 3. Below, we highlight several noteworthy papers:

Sensitivity of Input Design. Models are not only sensitive to different serialization formats, but variations in input design can also cause significant fluctuations in performance across table interaction tasks such as table partitioning, cell lookup, and reverse lookup (Sui et al., 2024). For example, omitting marked partitions or altering the input order has resulted in performance drops of up to 20%, while removing example shots has led to accuracy deteriorations of as much as 50%.

Sampling and Augmentation. Long or multi-table inputs pose challenges for serialization due to model input length limitations, often resulting in truncation or data loss. To address these constraints, researchers have developed methods for sampling rows or columns that capture the key information in a table. Recent research (Sui et al., 2023) demonstrates that embedding-based sampling techniques, such as centroid and semantic-based sampling, outperform other approaches. Furthermore, they show a balanced combination of augmentation data (e.g., table sizes and keyword explanations) and sampled table text has proven effective in achieving better overall performance within token limits.

3.2 Data Schema

Another input representation for table is to provide the schema of tables rather than presenting the entire table content. Common schema representations include database structures in SQL and dataframes in pandas, as illustrated in Figure 4. Using a data schema allows models to bypass input length limitations by focusing only on the structural blueprint of the data. However, this approach relies on strictly well-structured tables to be effective and loss of potential useful detailed content and value.

Sensitivity of Input Design. Like serialization, models are not only sensitive to the schema format, but also its designs: Zhang et al. (2023a) evaluated schema input designs on GPT-3.5 and found that using three example rows yielded the best results. Additionally, they highlighted that model performance declines sharply when primary and foreign keys (PF keys) in the data schema are omitted, which Chen et al. (2024) also mentioned.

Normalized structure. Given the trend toward schema-based methods and the improved results observed in Table QA tasks using Python or SQL code to interact with schema-based tables (Wang et al., 2024b; Pourreza and Rafiei, 2023a; Ye et al., 2023), exploring methods to convert complex data structures into more structured tables could be beneficial to enhance the compatibility of such methods.

3.3 Image

With the advancement of MLLMs, there is growing interest in using images as an input format

Benchmark	Sources / Domain	# Q	# T	Passage	Table Format	Output	Directions
WTQ (2015)	Wikipedia	22,033	2,108		HTML	cells	-
SQA (2017)	Wikipedia	17,553	6,066		HTML	cells	Input Complexity
HybridQA (2020c)	Wikipedia	69,611	13,000	✓	JSON	text-span	Input Complexity
FetaQA (2021a)	Wikipedia	-	10,330		Data Matrices	free-form	Answer Format
TAT-QA (2021)	Financial Reports	16,552	7,431	✓	Data Matrices	number	Domain, Input
OTT-QA (2021)	Wikipedia	-	45,841	✓	JSON	text-span	Input, Reasoning
AIT-QA (2022)	Airline Industry	515	113		Data Matrices	cells	Domain, Input
FinQA (2022)	Financial Report	8,281	2,789	✓	Data Matrices	number	Domain Knowledge
MMCoQA (2022)	MMQA (2018)	1,715	10,042	✓	JSON	text-span	Input Complexity
HiTab (2022)	Wikipedia, Statistic	10,672	3,597		Row-Separated	text-span	Input Complexity
MULTIHIERTT (2022)	Financial Report	10,440	2,513	✓	HTML	number	Input, Reasoning
Open-WikiTable (2023)	Wikipedia	67,023	24,680		Row-Separated	text-span, SQL	Answer Format
QTSUMM (2023)	Wikipedia	7,111	2,934		Data Matrices	free-form	Answer Format
TEMPTABQA (2023)	Wikipedia	11,454	1,208		JSON, HTML	text-span	Reasoning Difficulty
CRT-QA (2023d)	TabFact (2020b)	1,000	423		Row-Separated	text-span	Reasoning Difficulty
IM-TQA (2023)	Baidu Encyclopedia	5,000	1,200		Index Mapping	text-span	Input Complexity
TabCQA (2023a)	Financial Report	109,089	7,041		Text Template, Value Pair	text-span	Input Complexity
MultiTabQA (2023)	Spider (2018), Synthetic, TAPEX (2022) Corpus	136,461	-		Row-Separated	sub-table	Answer, Input
TABMWP (2023a)	Online Learning Web	38,431	37,544		Row-Seperated, SpreadSheet, Image	free-form	Reasoning Difficulty
FREB-TQA (2024)	WTQ, WikiSQL (2017), SQA, TAT-QA	75,205	8,590		Data Matrices	text-span	Input, Reasoning
Text2Analysis (2024)	Data Analysis Libraries	2,249	347		-	code, text	Reasoning Difficulty
MMQA (2024)	Spider (2018)	3,313	3,312		JSON	sub-table	Input Complexity

Table 1: Summary of benchmarks for Table-based Question Answering. **Sizes** shows the number of questions and tables. **Passage** indicates if an input passage is included. **Directions** categories each benchmark's primary focus compare to previous ones.

due to their adaptability, accessibility, and ability to preserve structural information (Wydmański et al., 2024). Specifically, Zheng et al. (2024) achieved superior results using images with a fine-tuned LLaVA model (Liu et al., 2023b), outperforming models with OCR and serialization settings. They found that additional training focused on table structure understanding—such as cell extraction and cell location—enhance the model's ability to accurately interpret tables.

Image resolution. While images offer the advantage of preserving the original table layout, they face constraints similar to serialization: the amount of data they can present is limited by image size and resolution, which can significantly impact model performance (Li et al., 2024). As tables grow larger, the information becomes blurred at a fixed resolution, leading to deteriorated performance. One potential approach is to use images as supplemental input alongside serialized text or data schema (Luo et al., 2023). This combined input strategy could potentially allow the model to receive structural information directly from the image while accessing detailed content from the text-based format. However, to the best of our knowledge, systematic evaluations of this approach remain lacking.

3.4 Table Encoder

Specific table encoder designs have been employed in smaller-scale language models to handle table-related tasks, utilizing various embeddings such as column-based (Iida et al., 2021), row-based (Herzig et al., 2020), tree-structured (Wang et al., 2021c), and graph-based embeddings (Wang et al., 2021a). Building on these approaches, recent work has demonstrated a trend toward employing specialized encoders in larger base models, effectively creating table foundation models (van Breugel and van der Schaar, 2024; Su et al., 2024; Ma et al., 2024). In particular, TableGPT2 leverages a specialized table encoder—with column- and row-wise attention—to integrate tabular data during the pretraining and fine-tuning stages of 7B and 72B base models (Su et al., 2024), outperforming other table generalist models across a range of tasks while remaining competitive with task-specific methods.

4 Table-Related Tasks

In this section, we introduce key table-related tasks such as Table Question Answering (TQA), Table-to-Text, and Table Fact Verification (TFV), along with other intriguing applications like leaderboard construction that actively utilize tables.

4.1 Table Question Answering

TQA[1] is one of the most common and well-studied table tasks, with various benchmarks developed as shown in Table 1. It typically involves a free-form question and a single table, sometimes accompanied by an optional passage or passage links, and

Benchmark	Sources / Domain	# Q	# T	Table Format	Focus	Text Input	Directions
Rotowire (2017)	NBA	-	4,853	JSON	N/A		Domain Knowledge
ToTTo (2020)	Wikipedia	134,161	83,141	Index Mapping	Highlight Span	Caption	-
Logic2Text (2020d)	WikiTable	10,800	5,600	Row-Separated	N/A		Logic Summarization
LogicNLG (2020a)	TabFact (2020b)	37,000	7,300	Data Matrices	N/A		Logic Comparison
SciGen (2021)	Scientific Paper	53,000	-	Row-Separated	N/A	Caption	Domain Knowledge
NumericNLG (2021)	Scientific Paper	1,300	1,300	JSON	N/A	Caption	Domain Knowledge
FetaQA (2021a)	ToTTo (2020)	-	10,330	Matrices	Text Query		Input Complexity
DART (2021b)	E2E (2020), WTQ WikiTable (2023) WebNLG (2019)	82,191	5,623	XML, JSON	N/A	Table Title	Table Structure
QTSUMM (2023)	Wikipedia	7,111	2,934	Data Matrices	Text Query		Input Complexity
FindSUM (2023c)	Company Report	-	21,125	Data Matrices	N/A	Long Text	Input Complexity

Table 2: Summary of benchmarks for Table-to-Text and Table Summarization. **Focus** specifies the subset of table content intended for natural language generation, while N/A indicates the entire table should be transformed to natural language.

Benchmark	Sources / Domain	# Q	# T	Table Format	Output	Directions
TabFact (2020b)	Wikipedia	117,843	18,000	Row-Separated	S, R	-
InfoTabs (2020)	Wikipedia	23,738	2,540	HTML, JSON	S, R, N	Output Format
FEVEROUS (2021)	Wikipedia	87,062	-	JSON / Mapping	S, R, N	Output Format
SEM-TAB-FACTS (2021b)	Science	5,715	2,961	XML	S, R, N, EC	Domain Knowledge
XInfoTabs (2022)	InfoTabs	23,738	2,540	JSON	S, R, N	Multi-Language
EI-InfoTabs (2022)	InfoTabs	23,738	2,540	JSON	S, R, N	Indic-Language
SciTab (2023b)	SciGen(Moosavi et al., 2021)	1,255	-	JSON / Mapping	S, R, N	Domain Knowledge

Table 3: Summary of benchmarks for Table-based Fact Verification. *S* in the output denotes Supported, *R* represents Refuted, *N* stands for Neither or Not Enough Evidence, and *EC* refers to Evidence Cells.

the output is expected to be information derived from the table or passage, generally presented as cell spans, calculated values, or minimal text spans.

TQA benchmarks have expanded significantly over the past two years, inspiring future work across multiple directions, including domain knowledge, answer format, input complexity, and reasoning difficulty. Domain-specific benchmarks now better reflect real-world scenarios in fields such as airlines (Katsis et al., 2022) and finance (Zhu et al., 2021; Chen et al., 2022). Answer formats have also diversified, with benchmarks requiring free-form responses (Nan et al., 2021a; Zhao et al., 2023; Wang et al., 2024a) and SQL queries (Kweon et al., 2023), beyond traditional cell values or text spans. Input complexity has increased through multi-table datasets (Pal et al., 2023; Zhao et al., 2022), hierarchical tables (Cheng et al., 2022), and semi-structured tables (Lu et al., 2023a), which challenge models to navigate intricate structures. Reasoning requirements have similarly intensified, incorporating hypothetical questions (Li et al., 2023b), implicit time-based inference (Gupta et al., 2023), and sequential or conversational queries (Iyyer et al., 2017; Li et al., 2022; Liu et al., 2023a). Overall, recent benchmarks generally demand more complex reasoning steps and operations to yield accurate answers.

[1]For a more comprehensive understanding of TQA, see this curated list of relevant papers: https://github.com/lfy79001/Awesome-Table-QA

4.2 Table-to-Text and Table Summarization

Table-to-Text and Table Summarization are table tasks initially developed to evaluate whether models could accurately interpret and describe table content. In these tasks, the input typically includes a table, sometimes with specified cell spans as shown in the *Focus* column in Table 2. If a span or region is provided, the model generates a textual description or summary of that specific area; if not, it summarizes the entire table. With advances in models' table understanding, this task has become less prominent, as the number of related publications has steadily decreased since 2021.

Query Focused Summarization. A recent, noteworthy benchmark in this area is QTSUMM (Zhao et al., 2023), which requires models to generate text-based summaries of specific table regions in response to questions. By integrating the aspect of table search based on textual queries from TQA with the descriptive demands of Table-to-Text, QTSUMM introduces new complexities that push models to move beyond simple fact retrieval. Notably, QTSUMM includes "why" questions, prompting models to reason about underlying causes or explanations—a shift that aligns more closely with human interests and highlights the importance of generating responses that incorporate causal understanding and contextual depth.

Lack of Multilingual Benchmarks. A notable gap in current research is the absence of multilin-

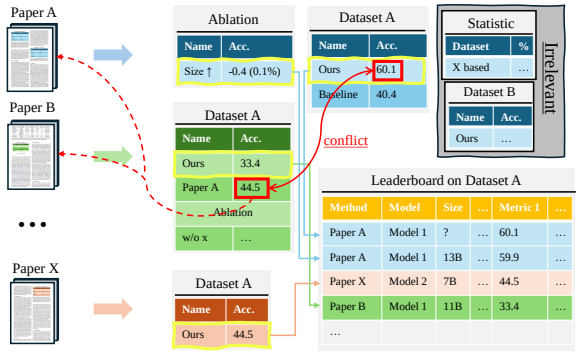

Figure 6: Illustration of automatic leaderboard construction pipeline. Results are extracted from ablation and performance tables in each paper. The red line highlights inconsistency across paper that may require examination across texts.

gual benchmarks for table-to-text tasks. As highlighted in (Osuji et al., 2024), to the best of our knowledge, no table-to-text benchmarks exist in languages other than English, significantly limiting the applicability and inclusivity of this task.

4.3 Table Fact Verification

Table Fact Verification (also referred to as Table Reasoning or Table Natural Language Inference) is a task designed to assess fact-searching and logic inference capabilities within tables. In this task, the input typically consists of a statement or claim alongside a reference table. The model's output is a verification label—such as "Supported," "Refuted," or "Not Enough Information"—indicating whether the claim aligns with the table content. Some benchmarks also require a justification for the answer, as shown in Table 3. Recent methods have enabled models to achieve over 80% accuracy on widely used benchmarks like TabFact and FEVEROUS (Sui et al., 2024; Ye et al., 2023; Wang et al., 2024b), demonstrating substantial progress in fact-checking within tabular data. However, scenarios involving longer contexts, multiple tables, or complex table structures remain unassessed.

4.4 Leaderboard Construction

Beyond the widely studied tasks, an intriguing direction proposed by Kardas et al. (2020) is leaderboard construction. This task aims to streamline the comparison of experimental results within a research domain through scientific papers, offering a concise and structured view of progress.

Existing methods, such as those proposed in (Kardas et al., 2020; Yang et al., 2022), have made notable strides in automating this process. These approaches typically employ pipelines that classify and extract data from performance and ablation tables in scientific papers, leveraging techniques like Named Entity Recognition (NER) or string matching to form tuples (Task, Dataset, Metric) or quadruples (Task, Dataset, Metric, Score). Such methods provide a foundational framework for building leaderboards and have proven effective in capturing basic performance comparisons across different methods and datasets. However, as scientific tasks and methodologies grow increasingly complex, these pipelines face limitations. Tasks often require varying schemas to account for unique aspects, and surface-level extraction may not fully capture the nuances of more intricate experiments or analyses. For instance, discrepancies in reported results between papers, as illustrated in Figure 6, often necessitate a deeper comparison and reasoning over both tables and textual content to resolve.

4.5 Other Tasks

Emerging new table-related tasks include innovations such as tabular synchronization across languages (Khincha et al., 2023) and column name abbreviation expansion (Zhang et al., 2023b). Among these, Text-to-Table has gained increasing attention in 2024 (Ramu et al., 2024; Jiang et al., 2024; Deng et al., 2024). The task was first formalized by Wu et al. (2022) as a sequence-to-sequence task by inversely applying table-to-text datasets. Recent studies have explored various methods, such as incorporating knowledge graphs (Jiang et al., 2024), to enhance its utility as a data integration task for field like finance, medicine, and law.

5 Further Reading

For readers seeking deeper insights into table-related research areas, several survey papers offer valuable perspectives. For methodologies aimed at improving table reasoning with LLMs, work by Zhang et al. (2024b) provides a detailed taxonomy and an analysis of emerging trends. Lu et al. (2024) explores prompting and training techniques for table-related tasks in the context of LLMs and VLMs. Meanwhile, Badaro et al. (2023); Ren et al. (2025) presents a focused analysis of transformer-based, smaller-scale models designed for tabular data. For an in-depth perspective, the comprehensive 30-page survey by Fang et al. (2024) provides an extensive overview of table understanding tasks, datasets, and corresponding fundamental methods.

Limitations

This study presents a comprehensive survey of table-related tasks with LLMs and MLLMs, highlighting key trends and emerging opportunities. While we have made our best effort to provide a thorough review, certain limitations remain. Due to space constraints, we focus on summarizing the main trends rather than providing exhaustive technical details for each approach. Our selection of works primarily draws from major NLP conferences, including ACL, EMNLP, NAACL, and ICLR, along with relevant studies from other domains and preprints. While we strive to incorporate the latest research, many new works continue to emerge during our submission of this paper. Given the rapid evolution of this field, our survey offers a snapshot of current progress rather than a definitive account. We will continue to track developments and refine our analysis in future updates.

References

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.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *Preprint*, arXiv:2106.05707.

Anonymous. 2024. MMQA: Evaluating LLMs with multi-table multi-hop complex questions. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, 11:227–249.

Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Mark Dredze, and Alan Ritter. 2024. Schema-driven information extraction from heterogeneous tables. *Preprint*, arXiv:2305.14336.

Ruisheng Cao, Fangyu Lei, Haoyuan Wu, Jixuan Chen, Yeqiao Fu, Hongcheng Gao, Xinzhuang Xiong, Hanchong Zhang, Yuchen Mao, Wenjing Hu, Tianbao Xie, Hongshen Xu, Danyang Zhang, Sida Wang, Ruoxi Sun, Pengcheng Yin, Caiming Xiong, Ansong Ni, Qian Liu, Victor Zhong, Lu Chen, Kai Yu, and Tao Yu. 2024. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *Preprint*, arXiv:2407.10956.

Shuaichen Chang, Jun Wang, Mingwen Dong, Lin Pan, Henghui Zhu, Alexander Hanbo Li, Wuwei Lan, Sheng Zhang, Jiarong Jiang, Joseph Lilien, Steve Ash, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, and Bing Xiang. 2023. Dr.spider: A diagnostic evaluation benchmark towards text-to-sql robustness. *Preprint*, arXiv:2301.08881.

Peter Baile Chen, Yi Zhang, and Dan Roth. 2024. Is table retrieval a solved problem? exploring join-aware multi-table retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv:2404.09889 [cs.IR], https://doi.org/10.48550/arXiv.2404.09889.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W. Cohen. 2021. Open question answering over tables and text. *Preprint*, arXiv:2010.10439.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020b. Tabfact: A large-scale dataset for table-based fact verification. *Preprint*, arXiv:1909.02164.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020c. 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.

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. 2022. Finqa: A dataset of numerical reasoning over financial data. *Preprint*, arXiv:2109.00122.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020d. Logic2text: High-fidelity natural language generation from logical forms. *Preprint*, arXiv:2004.14579.

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.

Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. 2021. Structure-grounded pretraining for text-to-sql. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9300–9322, Miami, Florida, USA. Association for Computational Linguistics.

Mingwen Dong, Nischal Ashok Kumar, Yiqun Hu, Anuj Chauhan, Chung-Wei Hang, Shuaichen Chang, Lin Pan, Wuwei Lan, Henghui Zhu, Jiarong Jiang, Patrick Ng, and Zhiguo Wang. 2025. PRACTIQ: A practical conversational text-to-SQL dataset with ambiguous and unanswerable queries. 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 255–273, Albuquerque, New Mexico. Association for Computational Linguistics.

Longxu Dou, Yan Gao, Mingyang Pan, Dingzirui Wang, Wanxiang Che, Dechen Zhan, and Jian-Guang Lou. 2022. Multispider: Towards benchmarking multilingual text-to-sql semantic parsing. *Preprint*, arXiv:2212.13492.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech and Language*, 59:123–156.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large language models(llms) on tabular data: Prediction, generation, and understanding – a survey. *Preprint*, arXiv:2402.17944.

Yujian Gan, Xinyun Chen, Qiuping Huang, Matthew Purver, John R. Woodward, Jinxia Xie, and Pengsheng Huang. 2021a. Towards robustness of text-to-sql models against synonym substitution. *Preprint*, arXiv:2106.01065.

Yujian Gan, Xinyun Chen, and Matthew Purver. 2021b. Exploring underexplored limitations of cross-domain text-to-sql generalization. *Preprint*, arXiv:2109.05157.

Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. 2023. Text-to-sql empowered by large language models: A benchmark evaluation. *Preprint*, arXiv:2308.15363.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vivek Gupta, Pranshu Kandoi, Mahek Bhavesh Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. Temptabqa: Temporal question answering for semi-structured tables. *Preprint*, arXiv:2311.08002.

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.

Moshe Hazoom, Vibhor Malik, and Ben Bogin. 2021. Text-to-sql in the wild: A naturally-occurring dataset based on stack exchange data. *Preprint*, arXiv:2106.05006.

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. *arXiv preprint arXiv:2312.13671*.

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.

Atin Sakkeer Hussain. 2025. Artemis-da: An advanced reasoning and transformation engine for multi-step insight synthesis in data analytics. *Preprint*, arXiv:2412.14146.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. TABBIE: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456, Online. Association for Computational Linguistics.

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.

Peiwen Jiang, Xinbo Lin, Zibo Zhao, Ruhui Ma, Yvonne Jie Chen, and Jinhua Cheng. 2024. TKGT: Redefinition and a new way of text-to-table tasks based on real world demands and knowledge graphs augmented LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16112–16126, Miami, Florida, USA. Association for Computational Linguistics.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, 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.

Siddharth Khincha, Chelsi Jain, Vivek Gupta, Tushar Kataria, and Shuo Zhang. 2023. InfoSync: Information synchronization across multilingual semi-structured tables. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2536–2559, Toronto, Canada. Association for Computational Linguistics.

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. *Preprint*, arXiv:2305.07288.

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.

Gyubok Lee, Woosog Chay, Seonhee Cho, and Edward Choi. 2024. Trustsql: Benchmarking text-to-sql reliability with penalty-based scoring. *Preprint*, arXiv:2403.15879.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2023. Ehrsql: A practical text-to-sql benchmark for electronic health records. *Preprint*, arXiv:2301.07695.

Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, Sida Wang, and Tao Yu. 2024. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. *Preprint*, arXiv:2411.07763.

Fangyu Lei, Tongxu Luo, Pengqi Yang, Weihao Liu, Hanwen Liu, Jiahe Lei, Yiming Huang, Yifan Wei, Shizhu He, Jun Zhao, and Kang Liu. 2023. Tableqakit: A comprehensive and practical toolkit for table-based question answering. *Preprint*, arXiv:2310.15075.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023a. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Preprint*, arXiv:2305.03111.

Moxin Li, Wenjie Wang, Fuli Feng, Hanwang Zhang, Qifan Wang, and Tat-Seng Chua. 2023b. Hypothetical training for robust machine reading comprehension of tabular context. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1220–1236, Toronto, Canada. Association for Computational Linguistics.

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.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ArXiv:2311.06607 [cs.CV], https://doi.org/10.48550/arXiv.2311.06607.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

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. *Preprint*, arXiv:2107.07653.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023c. Long text and multi-table summarization: Dataset and method. *Preprint*, arXiv:2302.03815.

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. *Preprint*, arXiv:2209.14610.

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

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. *Preprint*, arXiv:2305.13186.

Haohao Luo, Ying Shen, and Yang Deng. 2023. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9355–9367, Singapore. Association for Computational Linguistics.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony L. Caterini. 2024. Tabdpt: Scaling tabular foundation models. *Preprint*, arXiv:2410.18164.

Karime Maamari, Fadhil Abubaker, Daniel Jaroslawicz, and Amine Mhedhbi. 2024. The death of schema linking? text-to-sql in the age of well-reasoned language models. *Preprint*, arXiv:2408.07702.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models. *Preprint*, arXiv:2407.01725.

Qingkai Min, Yuefeng Shi, and Yue Zhang. 2019. A pilot study for chinese sql semantic parsing. *Preprint*, arXiv:1909.13293.

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.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Learning to reason for text generation from scientific tables. *Preprint*, arXiv:2104.08296.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. *Preprint*, arXiv:1904.03396.

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

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021b. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

OpenAI. 2024. Introducing openai o1 preview. https://openai.com/index/introducing-openai-o1-preview/. Accessed: 2025-02-03.

Chinonso Cynthia Osuji, Thiago Castro Ferreira, and Brian Davis. 2024. A systematic review of data-to-text nlg. *Preprint*, arXiv:2402.08496.

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.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *Preprint*, arXiv:2004.14373.

Junwoo Park, Youngwoo Cho, Haneol Lee, Jaegul Choo, and Edward Choi. 2021. Knowledge graph-based question answering with electronic health records. In *Proceedings of the 6th Machine Learning for Healthcare Conference (MLHC)*, volume 149, pages 36–53. PMLR.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *Preprint*, arXiv:1508.00305.

Xinyu Pi, Bing Wang, Yan Gao, Jiaqi Guo, Zhoujun Li, and Jian-Guang Lou. 2022. Towards robustness of text-to-sql models against natural and realistic adversarial table perturbation. *Preprint*, arXiv:2212.09994.

Mohammadreza Pourreza and Davood Rafiei. 2023a. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *Preprint*, arXiv:2304.11015.

Mohammadreza Pourreza and Davood Rafiei. 2023b. Evaluating cross-domain text-to-SQL models and benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1601–1611, Singapore. Association for Computational Linguistics.

Pritika Ramu, Aparna Garimella, and Sambaran Bandyopadhyay. 2024. Is this a bad table? a closer look at the evaluation of table generation from text. *Preprint*, arXiv:2406.14829.

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

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. 2024. Tablegpt2: A large multimodal model with tabular data integration. *Preprint*, arXiv:2411.02059.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465, Online. Association for Computational Linguistics.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*. ArXiv:2305.13062 [cs.CL], https://doi.org/10.48550/arXiv.2305.13062.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. 2023. TAP4LLM: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. *arXiv preprint arXiv:2312.09039*. https://doi.org/10.48550/arXiv.2312.09039.

Barkavi Sundararajan, Somayajulu Sripada, and Ehud Reiter. 2024. Improving factual accuracy of neural table-to-text output by addressing input problems in totto. *Preprint*, arXiv:2404.04103.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085, Online. Association for Computational Linguistics.

Boris van Breugel and Mihaela van der Schaar. 2024. Why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. ArXiv:2405.01147 [cs.LG], https://doi.org/10.48550/arXiv.2405.01147.

Fei Wang, Kexuan Sun, Muhao Chen, Jay Pujara, and Pedro Szekely. 2021a. Retrieving complex tables with multi-granular graph representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ArXiv:2105.01736 [cs.IR], https://doi.org/10.48550/arXiv.2105.01736.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021b. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

Yuqi Wang, Lyuhao Chen, Songcheng Cai, Zhijian Xu, and Yilun Zhao. 2024a. 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.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021c. TUTA: Tree-based transformers for generally structured table pre-training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*. ArXiv:2010.12537 [cs.IR], https://doi.org/10.48550/arXiv.2010.12537.

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. 2024b. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *Preprint*, arXiv:2401.04398.

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

Xueqing Wu, Jiacheng Zhang, and Hang Li. 2022. Text-to-table: A new way of information extraction. *Preprint*, arXiv:2109.02707.

Witold Wydmański, Ulvi Movsum-zada, Jacek Tabor, and Marek Śmieja. 2024. Vistabnet: Adapting vision transformers for tabular data. *Preprint*, arXiv:2501.00057.

Sean Yang, Chris Tensmeyer, and Curtis Wigington. 2022. TELIN: Table entity LINker for extracting leaderboards from machine learning publications. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 20–25, Online. Association for Computational Linguistics.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning. *Preprint*, arXiv:2301.13808.

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.

Hanchong Zhang, Ruisheng Cao, Lu Chen, Hongshen Xu, and Kai Yu. 2023a. Act-sql: In-context learning for text-to-sql with automatically-generated chain-of-thought. *arXiv preprint arXiv:2310.17342*. https://doi.org/10.48550/arXiv.2310.17342.

Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Shen Wang, Huzefa Rangwala, and George Karypis. 2023b. NameGuess: Column name expansion for tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13276–13290, Singapore. Association for Computational Linguistics.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024a. Tablellama: Towards open large generalist models for tables. *Preprint*, arXiv:2311.09206.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. 2024b. A survey of table reasoning with large language models. *Preprint*, arXiv:2402.08259.

Yi Zhang, Jan Deriu, George Katsogiannis-Meimarakis, Catherine Kosten, Georgia Koutrika, and Kurt Stockinger. 2023c. Sciencebenchmark: A complex real-world benchmark for evaluating natural language to sql systems. *Preprint*, arXiv:2306.04743.

Zhehao Zhang, Xitao Li, Yan Gao, and Jian-Guang Lou. 2023d. 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.

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, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Ruizhe Chen, Xiangru Tang, Yumo Xu, Dragomir Radev, and Arman Cohan. 2023. Qtsumm: Query-focused summarization over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Accepted at EMNLP 2023.

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 (ACL)*. ArXiv:2406.08100 [cs.CL], https://doi.org/10.48550/arXiv.2406.08100.

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. *Preprint*, arXiv:1709.00103.

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.

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. *Preprint*, arXiv:2105.07624.

A Text-to-SQL

Text-to-SQL is a semantic parsing task that is highly relevant to table-based applications: given a natural language question, the model must generate a SQL query that accurately captures the intent of the query. Over time, these tasks have

Benchmark	Sources / Domain	Sizes	Input Format	T / Q	Directions
WikiSQL (2017)	Wikipedia	80,654	Row Header, Row-Separated	1.0	-
Spider (2018)	Academic Databases, Online CSV, WikiSQL	10,181	Table(col), Type, PF	1.6	-
SEDE (2021)	Stack Exchange	12,023	Table(col), Type, PF	1.3	Noise Utterance
SpiderDK (2021b)	Spider	535	Table(col), Type, PF	> 1	Domain Knowledge
SpiderSyn (2021a)	Spider	8,034	Table(col), Type, PF	> 1	Query Perturbation
SpiderRealistic (2021)	Spider	508	Table(col), Type, PF	> 1	Query Perturbation
MIMICSQL (2021)	Electronic Medical Records	10,000	Row Header, Row-Separated	1.8	Domain Knowledge
KaggleDBQA (2021)	ATIS, GeoQuery, Restaurants, Yelp, Academic, IMDB, Scholar, Advising	272	Table(col), Type, PF, context	1.2	Domain Knowledge
ADVETA (2022)	Spider, WikiSQL, WTQ	-	Table(col), Type, PF	> 1	Table Perturbation
BIRD (2023a)	Kaggle, Machine Learning platform	12,751	Table(col), Type, PF, context	> 1	Table Size
Dr.Spider (2023)	Spider	15,000	Table(col), Type, PF	> 1	Table, Query Perturbation
EHRSQL (2023)	Electronic Medical Records	24,000	Table(col), Type, PF	2.4	Domain, Reasoning
ScienceBench (2023c)	CORDIS, SDSS, OncoMX	6,000	Table(col), Type, PF	> 1	Data Synthesis, Domain
TrustSQL (2024)	ATIS, Advising, EHRSQL, Spider	27,784	CREATE(EoT)	> 1	Reasoning
Spider2 (2024)	Cloud Data Warehouses	632	Table(col), PF	> 1	Reasoning, Table Size
Spider2V (2024)	Cloud Data Warehouses	494	Agent Workspace	> 1	Input Modality

Table 4: Summary of benchmarks for Text-to-SQL. **Sizes** refers to the number of SQL query pairs, and **T/Q** indicates the number of tables required to answer a single query.

evolved to incorporate additional contextual information—such as table schemas and optional sample rows—with the evaluation focus shifting from exact match (EM) to execution accuracy (EX) as the primary metric. A prominent benchmark in this area, Spider (Yu et al., 2018), significantly increased task complexity by introducing databases composed of multiple tables, foreign keys, and the requirement to employ a variety of functions.

Building on Spider, several adaptations and extensions have broadened the task's scope and complexity. Multilingual adaptations (Min et al., 2019; Tuan Nguyen et al., 2020; Dou et al., 2022) expanded Text-to-SQL to cross-lingual and multilingual settings, enabling SQL generation across diverse languages. Other extensions include Spider-DK (Gan et al., 2021b), which incorporates domain knowledge, and Spider-Syn (Gan et al., 2021a) and Spider-Realistic (Deng et al., 2021), which obscure schema-related words or column names to simulate noisy utterances and more realistic queries.

Text-to-SQL has been well-studied with question decomposition pipelines (Gao et al., 2023; Ye et al., 2023; Wang et al., 2024b), with current models nearing saturation on some commonly used benchmarks.

Effect of Noisy Input. Beyond evaluation issues, Text-to-SQL faces inherent challenges, especially when handling ambiguity, or on very large tables. As noted in (Chen et al., 2024), performance drops significantly without PF keys, as variations in column names across tables and limited sample rows complicate element matching. Moreover, as highlighted in (Lei et al., 2024; Maamari et al., 2024), model performance deteriorates sharply when processing extremely long database schema, a scenario prevalent in real-world industrial databases.

B Responsible NLP Miscellanea

B.1 AI Assistants

We acknowledge the use of GPT-4o and GPT-o3-mini for grammar checking and word polishing.