

RUST-BENCH: Benchmarking LLM Reasoning on Unstructured Text within Structured Tables

Nikhil Abhyankar¹, Purvi Chaurasia², Sanchit Kabra¹, Ananya Srivastava²,
Vivek Gupta³, Chandan K. Reddy¹

¹Virginia Tech, ²IGDTUW New Delhi, ³Arizona State University

Abstract

Existing tabular reasoning benchmarks mostly test models on small, uniform tables, underrepresenting the complexity of real-world data and giving an incomplete view of Large Language Models’ (LLMs) reasoning abilities. Real tables are long, heterogeneous, and domain-specific—mixing structured fields with free text and requiring multi-hop reasoning across thousands of tokens. To address this gap, we introduce RUST-BENCH, a benchmark of 7,966 questions from 2,031 real-world tables spanning two domains: (i) RB-Science (NSF grant records) and (ii) RB-Sports (NBA statistics). Unlike prior work, RUST-BENCH evaluates LLMs jointly across scale, heterogeneity, domain specificity, and reasoning complexity. Experiments with open-source and proprietary models show that LLMs struggle with heterogeneous schemas and complex multi-hop inference, revealing persistent weaknesses in current architectures and prompting strategies. RUST-BENCH establishes a challenging new testbed for advancing tabular reasoning research.¹

¹<https://github.com/tabular-reasoning/RUST-BENCH>

1 Introduction

Semi-structured tables containing free-form text embedded within structured fields are common across various domains (Gupta et al., 2020). Effective data analysis in science, finance, and sports requires reasoning over large, domain-specific tables that combine symbolic structure with textual context. However, existing benchmarks predominantly evaluate short, homogeneous Wikipedia-derived tables (Pasupat and Liang, 2015; Chen et al., 2019), which limits both model generalizability and robustness. Although Large Language Models (LLMs) have made tabular reasoning more accessible by allowing users to query tables directly in natural language (Cheng et al., 2022), systematic



Figure 1: **Illustration of a multi-step reasoning process for a complex question grounded in a sports table from RUST-BENCH.** The example shows that real-world tabular reasoning often demands multiple complementary reasoning skills (temporal, arithmetic, and contextual) and the coordinated use of heterogeneous evidence across long, domain-specific tables.

evaluation of their reasoning abilities over complex tables remain underexplored (Chen, 2023).

Real-world tabular reasoning introduces four major challenges for LLMs: **scale**, **multi-hop reasoning**, **heterogeneity**, and **domain specificity**. First, tables can be **long**, often spanning hundreds of rows and columns, and such long contexts are known to degrade LLM reasoning performance (Liu et al., 2023a). Similarly, model performance deteriorates as table size grows, even when the entire table fits within the context window, since only a small fraction of rows are typically relevant to a given query (Abhyankar et al., 2024). Second, many queries require **multi-hop reasoning**—locating relevant rows, integrating dispersed evidence, and composing it into an answer. Third, **heterogeneity** arises when tables mix structured fields with free-form text, requiring models to reason over diverse data modalities (Chen et al., 2020b; Zhu et al., 2021). Finally, **domain specificity** intro-

¹Correspondence: nikhilsa@vt.edu, vgupt140@asu.edu

Table 1: **Comparison of RUST-BENCH with other Table QA datasets.** RUST-BENCH contains a variety of complex question types over large, domain-specific tables containing semi-structured information. *Only the contents of the table are considered.

Dataset	Source	Complex Reasoning	Unanswerable Questions	Domain Specific	Semi Structured	Large Tables	# Avg. Rows	Context Length
WikiTQ (Pasupat and Liang, 2015)	Wikipedia (Wikipedia)	✗	✗	✗	✗	✗	6.3	1133.51
TabFact (Chen et al., 2019)	Wikipedia	✗	✗	✗	✗	✗	6.2	586.51
Hybrid-QA (Chen et al., 2020b)	Wikipedia	✓	✗	✗	✓	✗	15.7	372.14*
OTT-QA (Chen et al., 2020a)	Wikipedia	✓	✗	✗	✓	✗	15.7	372.14*
CRT-QA (Zhang et al., 2023)	Wikipedia	✓	✓	✗	✗	✗	12.6	257.12
TAT-QA (Zhu et al., 2021)	Financial Reports (Reports, 2025)	✓	✗	✓	✓	✗	9.4	378.31
FINQA (Chen et al., 2021)	FinTabNet (Zheng et al., 2021)	✓	✗	✓	✓	✗	6.4	687.51
SciTab (Lu et al., 2023)	SciGen (Moosavi et al., 2021)	✓	✗	✓	✗	✗	7.5	254.53
RUST-BENCH	NSF (NSF, 2024), Sportsett (Thomson et al., 2020)	✓	✓	✓	✓	✓	45.1	23040.68

duces specialized terminology and domain-specific reasoning patterns, as seen in finance (Chen et al., 2021) and science (Lu et al., 2023), which require specialized domain knowledge for effective inference. While existing benchmarks assess specific aspects of table reasoning, they often evaluate these challenges in isolation. *The absence of benchmarks that jointly incorporate scale, heterogeneity, and domain specificity constitutes a fundamental limitation*, constraining systematic progress toward generalizable tabular reasoning models. We therefore pose the question: *Can LLMs effectively reason over unstructured text embedded in long, domain-specific tables?*

To answer this, we introduce **RUST-BENCH**, a new benchmark explicitly designed to stress-test models across **four orthogonal axes** of real-world tabular reasoning: *domain specificity, table length, semi-structured information, and multi-hop reasoning*, offering a comprehensive and realistic evaluation framework. RUST-BENCH comprises 2,031 tables primarily sourced from two domains: (a) science and (b) sports, accompanied by 7,966 carefully curated question–answer pairs. We construct the dataset using an *LLM-driven hybrid symbolic–semantic generation pipeline*, that systematically constructs high-quality, multi-hop queries grounded in real-world semi-structured tables while reducing manual annotation costs. As illustrated in Figure 1, each question is designed to evaluate a wide spectrum of reasoning skills (including temporal, numerical, aggregation, verification, commonsense, counterfactual, and ambiguity resolution) with most requiring multi-hop reasoning that integrates information across multiple cells through both parallel and sequential inference. As shown in Table 1, existing benchmarks primarily rely on Wikipedia, which generally involves short contexts and relatively simple reasoning. These datasets often lack domain-specific information,

unanswerable queries, and large semi-structured tables, thereby limiting their capacity to appropriately reflect real-world complexity. In contrast, RUST-BENCH introduces domain-grounded tables, expands the range of reasoning types, and substantially scales up table size (averaging 45.1 rows and roughly 23000 tokens per table). This design offers a more realistic and challenging evaluation setting for LLMs. We evaluate RUST-BENCH using state-of-the-art proprietary and open-source LLMs, employing diverse prompting strategies and reasoning methods. Our findings expose systematic weaknesses in handling scale, heterogeneity, and reasoning composition, confirming the value of RUST-BENCH as a challenging and diagnostic benchmark for advancing research on LLM-based table reasoning. Our main contributions are:

- We introduce **RUST-BENCH**, a large-scale benchmark that jointly evaluates LLMs across four orthogonal dimensions (i.e., scale, heterogeneity, domain specificity, and complex reasoning) previously treated in isolation by existing datasets.
- We develop a hybrid dataset generation pipeline that leverages the complementary strengths of symbolic and semantic reasoning to construct diverse, multi-hop, domain-grounded QA pairs efficiently.
- Comprehensive evaluations of state-of-the-art open-source and proprietary models reveal that current LLMs struggle with large, heterogeneous tables and multi-step reasoning, exposing persistent gaps in table reasoning architectures and prompting strategies.

2 RUST-BENCH Dataset

2.1 Task Formulation

In table-based reasoning, each problem instance is represented as a triplet (T, Q, A) , where T denotes the tabular data, Q represents the associated query, and A signifies the anticipated response. Specifically, in the context of table-centric question-

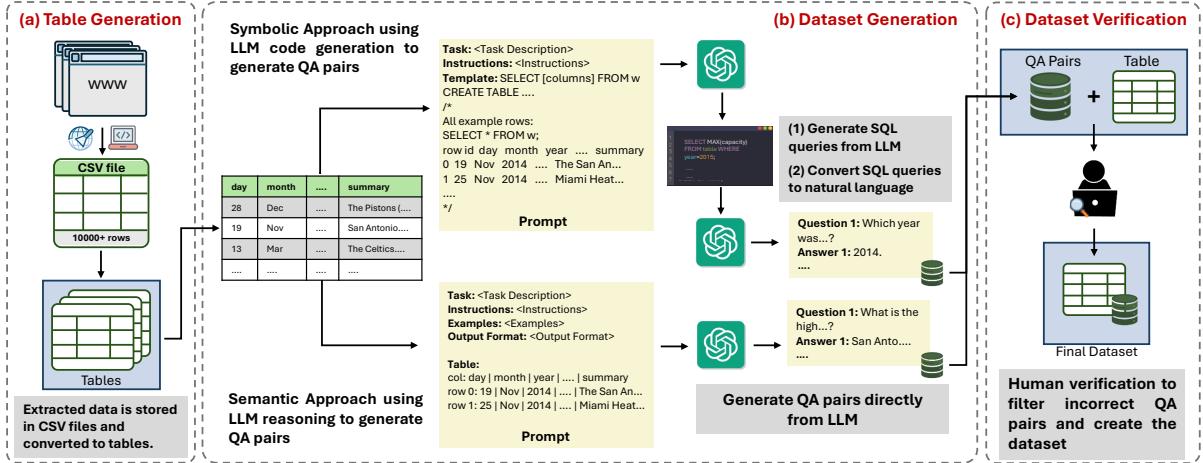


Figure 2: Overview of RUST-BENCH’s dataset generation and verification pipeline. (a) **Table Generation**: Raw data are extracted from public web sources and reorganized into tables containing at least 30 rows each. (b) **Dataset Generation**: Question–Answer pairs are created through two complementary methods: (i) a *symbolic approach*, which uses SQL-like logical forms to construct schema-intensive, reasoning-heavy queries, and (ii) a *semantic approach*, which employs LLMs to generate natural, inference-oriented questions from unstructured text. (c) **Dataset Verification**: All generated pairs undergo human verification to ensure factual correctness and annotation quality.

answering systems, both Q and A are in natural language. The primary objective is to derive a prediction a utilizing T and Q , which can be formally expressed as $a = \pi_\theta(T, Q)$, where π_θ symbolizes the predictive model.

2.2 RUST-BENCH Creation

Table Collection. We curate domain-grounded tables from two high-quality sources: the NSF Grants Database (NSF, 2024) for science and the SportSett:Basketball dataset (Thomson et al., 2020), an enhanced version of RotoWire (Wiseman et al., 2017), for sports. The raw data is cleaned and organized into domain-specific JSON tables, sampled by attributes (such as year and region) and by uniform random selection (Figure 2(a)). We focus on constructing large tables with more than 30 rows, consistent with the definition in (Chen, 2023). To ensure diversity and cross-domain comparability, we apply structured sampling to balance table sizes: 50% with 30–40 rows, 40% with 40–60, and 10% with 60–100. This stratification balances coverage and scale across the domains, yielding a representative mixture of table sizes and schema complexities.

QA Generation. Creating high-quality QA pairs for long, domain-specific tables is particularly challenging as manual annotation is slow, costly, and prone to errors when tables span thousands of tokens. Inspired by recent LLM-based data generation methods (Park et al., 2023; Zhang et al., 2023; Li and Zhang, 2024), we adopt in-

context learning and role-playing paradigms to enable scalable and diverse dataset construction at a lower annotation cost. However, only using LLMs’ textual (semantic) reasoning is inadequate as it captures natural-language inference but fails on structural and quantitative reasoning. Conversely, symbolic reasoning methods yield precise numerical manipulation and structural consistency but lack flexibility with unstructured text (Liu et al., 2023b). We therefore leverage their complementary strengths to design a hybrid symbolic–semantic pipeline (Figure 2(b)) comprising (a) a **symbolic approach**, which uses SQL-like logical forms to create schema-intensive, reasoning-heavy queries, and (b) a **semantic approach**, which generates natural, inference-oriented questions from unstructured text.

(a) Symbolic Approach. The symbolic approach exploits LLMs’ code-generation abilities to synthesize SQL queries over both structured and unstructured table components, to create questions involving numerical reasoning, aggregation, and logic. We construct a library of 75 SQL templates with placeholders (e.g., `SELECT [columns] FROM [table] WHERE [condition]`) covering diverse query patterns such as selection, aggregation, and conditional operations (Appendix A.1). During generation, a template is sampled and instantiated with table-specific values, providing a structural scaffold for producing valid SQL queries (Figure 2(b)). For example, a template may yield `SELECT MAX(attendance) FROM RB_Sports`

`WHERE city=='New York'`, which is then paraphrased into a natural language question ‘What is the highest attendance recorded in NYC?’ by prompting an LLM. To ensure fluency and avoid explicit SQL exposure, entity names are masked or rephrased (e.g., New York → NYC) during paraphrasing. This dual process enables coverage of multiple reasoning types, integrating structured computation with textual variation.

(b) Semantic Approach. The semantic component uses LLMs’ semantic reasoning to derive insights from unstructured text segments and generate diverse, inference-driven questions that go beyond surface-level lookups. However, LLMs struggle with long or complex inputs (Liu et al., 2023a), often producing (1) overly simplistic questions and (2) repetitive patterns, especially on large tables. To mitigate these issues, we restrict inputs to either: **Single Row-Based** method for focused intra-row reasoning, or a **Multi-Row-Based** method for multi-hop reasoning across a small subset of semantically related rows. This setup reduces contextual load and encourages inference beyond simple lookups while keeping questions easily verifiable by human annotators. To further enhance diversity, we maintain a pool of in-context exemplars spanning multiple reasoning types and randomly sample from them during generation. Combined with temperature variation, this encourages broader coverage and deeper reasoning. Details of the single-row and multi-row generation processes are in Appendix A.2.

2.3 RUST-BENCH Validation

Although LLMs can generate QA pairs at scale, their outputs often suffer from misalignment, limited diversity, and uneven reasoning depth (Zhang et al., 2023). To ensure high-quality supervision for RUST-BENCH, we adopt a rigorous human-in-the-loop verification pipeline. This process substantially improves quality by filtering out poor generations. We first discard malformed or duplicated QA pairs and those with empty or ill-formed answers. Eight Computer Science graduate students act as annotators to review each remaining pair using a custom web interface that displays the full semi-structured table alongside its question and answer (See Appendix A.3). Annotators rate *clarity*, *answer correctness*, and *reasoning complexity* and flag uncertain or incorrect cases for secondary review. They are also instructed to ensure that the

final answers are concise, self-contained, and free of redundant text to facilitate consistent automatic and human evaluation. Three expert reviewers then re-examine all pairs and consolidate the verified dataset. Low-quality or unverifiable examples are removed, while minor errors are corrected. As summarized in Table 2, this process yields a curated set of high-quality QA pairs supporting multi-hop reasoning over long, heterogeneous tables.

Table 2: Breakdown of QA pairs before and after human verification.

Dataset	Category	Original # QA	Final # QA	% Discarded
RB-Sports	Single Row	2886	2712	6.0%
	Multi Row	1222	838	31.4%
	Symbolic	1431	1338	6.5%
RB-Science	Single Row	915	805	12.0%
	Multi Row	1516	1101	27.3%
	Symbolic	1267	1172	7.5%

2.4 RUST-BENCH Statistics

Table 3 summarizes the RUST-BENCH dataset, comprising 2,031 tables spanning RB-Sports (1,326) and RB-Science (705). Although both domains contain tables of similar length, RB-Science shows greater structural complexity, with more columns and higher token counts per table. We include 5,674 questions in RB-Sports and 2,292 in RB-Science, averaging 4.28 questions per table in RB-Sports and 3.25 in RB-Science, plus a subset of unanswerable queries. For unstructured passages, RB-Science has higher average token counts (477.62 vs. 400.58; medians 469 vs. 368) and a larger token standard deviation (149.87 vs. 114.21), while RB-Sports has slightly more sentences per passage on average. To assess annotation quality, we conducted a human-rated complexity study following (Nan et al., 2022). Three experts rated 100 random examples on a 1–5 scale, with scores ≥ 4 indicating high-quality QA pairs. The study achieved 91.7% inter-annotator agreement, confirming the dataset’s reliability.

3 Experiments

LLM Backbones. We benchmark a diverse set of state-of-the-art large language models, spanning both open-source and proprietary families, as well as reasoning-optimized variants for complex problem-solving. Specifically, we evaluate Llama-3.3-70B-Instruct (Dubey et al., 2024), GPT-4o-mini (OpenAI, 2023),

Table 3: Summary statistics of RUST-BENCH across RB-Sports and RB-Science.

	RB-Sports	RB-Science
Tables		
# Tables	1326	705
Avg. Rows / Table	44.95	45.13
Avg. Columns / Table	12.0	28.0
Avg. Tokens / Table	18304.47	31948.79
Questions		
# Questions	5674	2292
Avg. Question Length (words)	26.92	27.48
# Questions / Table	4.28	3.25
# Unanswerable Questions	132	372
Unstructured Text		
Avg Tokens / Passage	400.58	477.62
Std Tokens	114.21	149.87
Median Tokens	368.00	469.00
Avg Sentences / Passage	16.22	14.34
Std Sentences	4.32	4.84
Median Sentences	15.00	14.00
Inter-Annotator Agreement	91.7%	

Gemini-2.0-Flash (Team et al., 2023), and Mistral-Small-3.2-24B-Instruct-2506 (Mistral, 2024). Beyond these general-purpose models, we also assess Qwen3-14B, Qwen-32B-QwQ (Yang et al., 2025), Qwen-Distill-32B, and Llama-Distill-70B (Guo et al., 2025), which are specialized for reasoning tasks. All models are evaluated using default hyperparameters and a fixed decoding temperature ($\tau = 0.1$) for consistency across runs. Following (Wang et al., 2023), each table is linearized into a pipe-separated format and concatenated with its query across models.

Baselines. We evaluate two baseline categories: (i) **prompting strategies** and (ii) **table reasoning methods** developed specifically for tabular data. For prompting, we adopt four standard paradigms: (i) **Zero-shot prompting**, where the model directly answers the table–question pair; (ii) **Few-shot prompting** (Chen, 2023), with four in-context examples; (iii) **Chain-of-Thought (CoT)** (Wei et al., 2022), encouraging intermediate reasoning steps; and (iv) **Program-of-Thought (PoT)** (Chen et al., 2023), which incorporates executable programs as intermediate reasoning. For table reasoning methods, we use GPT-4o-mini and Llama-3.3-70B as LLM backbones to evaluate six state-of-the-art approaches: BlendSQL (Glenn et al., 2024), a hybrid framework embedding SQL-style reasoning within natural prompts; Chain-of-Table (Wang et al., 2023), which performs stepwise table updates for interpretable reasoning; ProTrix (Wu and Feng, 2024), integrating SQL planning with compositional reasoning; TabSQLify (Nahid and Rafiei, 2024b), which uses SQL to partition

large tables into sub-tables for scalable inference; TableMaster (Cao and Liu, 2025), combining textual and symbolic reasoning via adaptive table verbalization; and NormTab (Nahid and Rafiei, 2024a), normalizing table structures and values to improve symbolic interpretability. Additional implementation details are provided in Appendix B.3.

Evaluation Metrics. For fairness and consistency, all models are evaluated under identical input and output constraints, focusing on accuracy and generation quality. Each model is instructed to produce concise, self-contained natural language answers; for SQL-based methods, query output is post-processed and verbalized in natural language for comparability. Following (Pasupat and Liang, 2015; Zhang et al., 2023), we report **Exact Match (EM)** as the primary metric. We further relax the evaluation with **BLEU** (Papineni et al., 2002) to capture n -gram overlap and an **LLM-as-a-Judge** (LLM-Score) evaluation using GPT-4o-mini to assess semantic equivalence. This combination provides complementary signals for lexical accuracy, surface fluency, and semantic faithfulness. For more details, see Appendix B.2.

3.1 Main Results

Table 4 reports the performance of different LLM backbones on RUST-BENCH using Exact Match (EM), BLEU, and LLM-score. Overall, Qwen-QwQ achieves the highest performance across all metrics, with an LLM-score reaching 54.1 and 55.7 for RB-Science and RB-Sports, respectively. Furthermore, it can be seen that CoT consistently outperforms Zero-Shot and Few-Shot for smaller models, highlighting the importance of explicit reasoning in this setting. In contrast, PoT exhibits the weakest performance across all models, likely due to the semi-structured nature of the data. In Table 5, we present a comparison of table reasoning baselines on RUST-BENCH, implemented using GPT-4o-mini and Llama-3.3-70B as the backbones. Among these, TableMaster achieves the best overall results, reaching 42.3% EM on RB-Science and 43.1% on RB-Sports. In contrast, symbolic or SQL-based methods such as TabSQLify and BlendSQL perform worse, achieving EM scores of 15.3% and 13.6%, respectively. These findings suggest that purely symbolic reasoning pipelines are insufficient for the flexible, context-driven inference required by RUST-BENCH, which is consistent with our findings.

Table 4: Comparison of LLM backbones using various prompting strategies on variants RB-Science and RB-Sports using: (a) Exact Match (EM), (b) BLEU, and (c) LLM-as-a-judge (LLM-score). Higher values indicate better performance.

Model	Strategy	RB-Science			RB-Sports		
		EM (%)	BLEU	LLM-score (%)	EM (%)	BLEU	LLM-score (%)
Large Language Models							
GPT-4o-mini	Zero-Shot	36.6	0.293	40.4	39.8	0.285	43.1
	Few-Shot	37.9	0.296	36.7	31.3	0.301	33.9
	CoT	44.4	0.378	48.8	42.1	0.365	45.2
	PoT	32.8	0.312	34.5	30.6	0.285	33.6
Llama-3.3-70B	Zero-Shot	38.8	0.301	47.1	39.2	0.311	44.3
	Few-Shot	41.7	0.347	46.4	46.7	0.350	48.9
	CoT	44.2	0.401	45.3	42.2	0.392	43.9
	PoT	27.7	0.299	30.6	31.1	0.289	33.0
Gemini-2.0-Flash	Zero-Shot	40.7	0.370	47.3	38.6	0.345	45.4
	Few-Shot	45.9	0.373	48.8	41.4	0.340	43.3
	CoT	47.3	0.454	50.8	44.1	0.419	48.7
	PoT	18.2	0.225	23.6	26.3	0.239	29.1
Mistral-Small-3.2	Zero-Shot	48.3	0.410	50.5	45.7	0.404	48.0
	Few-Shot	50.3	0.373	51.6	43.9	0.365	45.2
	CoT	52.6	0.454	53.1	51.5	0.446	51.7
	PoT	29.8	0.278	29.9	20.5	0.241	26.4
Large Reasoning Models							
Qwen3-14B		42.6	0.441	44.4	41.2	0.433	43.1
Qwen-QwQ		48.1	0.526	54.1	46.1	0.479	55.7
Qwen-Distill-32B		43.1	0.407	49.9	39.2	0.426	44.6
Llama-Distill-70B		44.6	0.483	52.4	40.5	0.455	50.9

3.2 Impact of Table Size

To investigate how table size affects reasoning accuracy in our setting, we analyze model performance across naturally occurring tables grouped by total token count, spanning from 10K to 85K tokens. As illustrated in Figure 3, GPT-4o-mini, Gemini-2.0-Flash, and LLaMA-3.3-70B exhibit a consistent, monotonic decline in Exact Match accuracy as table size increases, with degradation becoming particularly pronounced beyond the 35K–50K token threshold. Notably, this performance drop occurs well within the nominal context windows of modern LLMs (typically 128k+ tokens), suggesting that the bottleneck arises from reasoning and attention limitations rather than raw context length. This degradation can be attributed to LLMs’ difficulty in retrieving and integrating dispersed evidence across long sequences (Liu et al., 2023a), difficulty in locating relevant information, and increased multi-hop reasoning complexity. Unlike existing benchmarks that predominantly feature concise tables under 5000 tokens (Pasupat and Liang, 2015; Chen et al., 2019), RUST-BENCH includes substantially longer and more heterogeneous tables where critical information is often scattered across extensive contexts. These findings highlight the need for improved query-specific data extraction mechanisms to effectively handle large-scale tabular reasoning tasks.

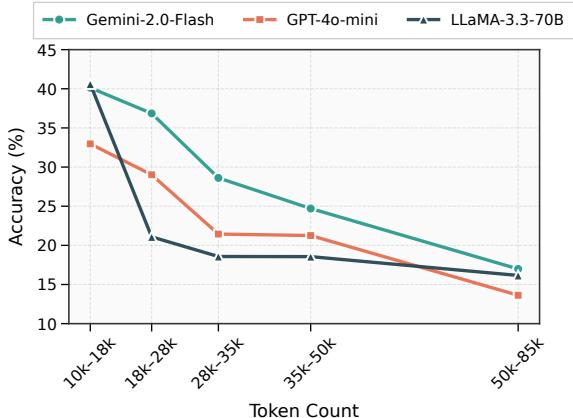


Figure 3: Accuracy comparison of LLMs across varying token count bins. The x-axis represents token length ranges, while the y-axis shows accuracy in percentage.

3.3 Impact of Real-World Table Complexity

To assess how the combination of real-world structural complexity and multi-hop reasoning affects model performance, we compare two proprietary LLMs GPT-4o-mini and Gemini-2.0-Flash across WikiTQ (a general-knowledge benchmark) and RUST-BENCH. We evaluate both models under zero-shot and Chain-of-Thought (CoT) prompting settings. As shown in Figure 4, both models demonstrate strong performance on WikiTQ, with GPT-4o-mini achieving 59.4% accuracy in zero-shot and 64.5% with CoT, while Gemini-2.0-Flash reaches 69.7% and 80.4%, re-

Table 5: Comparison of baselines on RUST-BENCH using GPT-4o-mini and Llama-3.3-70B using: (a) Exact Match (EM), (b) BLEU, and (c) LLM-as-a-judge (LLM-score), with higher values indicating better performance.

Method	GPT-4o-mini			Llama-3.3-70B		
	EM (%)	BLEU	LLM-score (%)	EM (%)	BLEU	LLM-score (%)
TabSQLify	15.3	0.206	22.3	14.4	0.120	18.6
BlendSQL	13.6	0.186	20.2	11.7	0.145	13.6
ProTrix	32.6	0.319	33.9	28.3	0.265	31.5
Chain-of-Table	30.1	0.247	35.1	33.2	0.358	36.9
NormTab	33.9	0.338	36.8	30.9	0.279	34.9
TableMaster	42.3	0.431	44.2	43.1	0.386	45.4

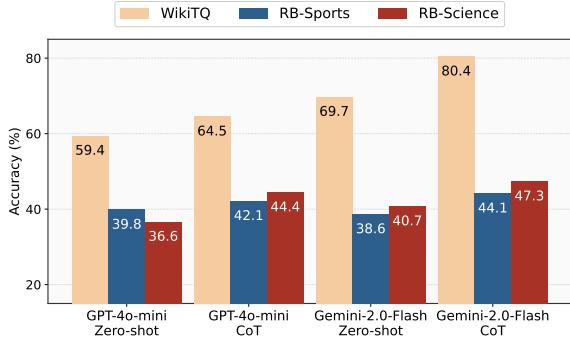


Figure 4: Performance comparison of LLM backbones on RUST-BENCH and WikiTQ using EM accuracy. Unlike WikiTQ, RUST-BENCH tests LLMs with more challenging questions and tables, resulting in a reduced LLM performance.

spectively. In contrast, performance on RUST-BENCH drops sharply to roughly 20-30% across all prompting strategies for both models. This substantial gap reveals the compounding challenges introduced by domain-specific reasoning, heterogeneous table schemas, long contexts, and multi-hop inference. Unlike WikiTQ’s short, homogeneous tables dominated by direct lookup queries, RUST-BENCH captures the full spectrum of real-world tabular reasoning, where multiple factors interact to create harder reasoning problems. Such a dramatic decline underscores the limits of current LLMs in generalizing beyond simplified benchmarks and highlights the pressing need for more robust and compositional reasoning mechanisms.

3.4 Impact of Heterogeneous Data

While multi-hop evaluation has been extensively studied as a driver of task difficulty, the influence of data heterogeneity and structure on reasoning performance remains less explored. To investigate how the underlying data influences reasoning performance, we conduct controlled experiments on a subset of randomly sampled 100 RB-Sports tables in two settings: structured and unstructured. We convert the semi-structured tables while keeping the underlying content identical in both its variants. In the structured setting, information

is normalized into explicit columns, minimizing free-form text; in the unstructured setting, each table row is verbalized into natural-language sentences and appended to the textual field, simulating highly heterogeneous inputs. We first evaluate symbolic reasoning methods, specifically Program-of-Thought (PoT) prompting, on the structured and semi-structured variants. As shown in Figure 5, PoT consistently achieves higher accuracy on the structured version across all models except Llama-3.3-70B, which performs comparably on both. This pattern indicates that symbolic reasoning benefits from explicit schema structure and reduced textual noise, confirming its reliance on syntactic regularity. Next, we assess text-based reasoning methods using Chain-of-Thought (CoT) prompting on the unstructured and semi-structured variants.

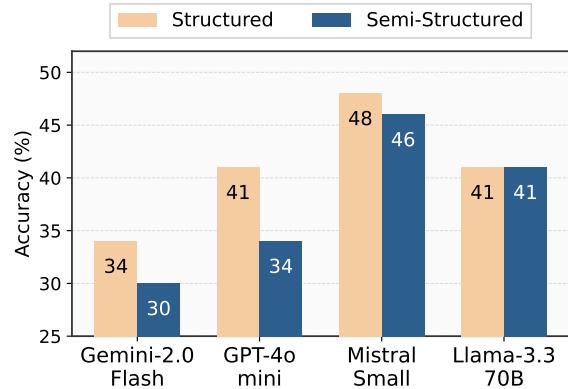


Figure 5: Performance comparison on structured and semi-structured variants for different LLM backbones using Program-of-Thought (PoT) prompting.

Figure 6 shows that CoT yields higher accuracy on the unstructured representation, indicating that natural-language continuity facilitates stepwise reasoning when explicit structure is absent. Overall, these results show that semi-structured data presents the greatest reasoning challenge, as it combines the ambiguity of free-text with the rigidity of tabular schema, while purely structured or un-

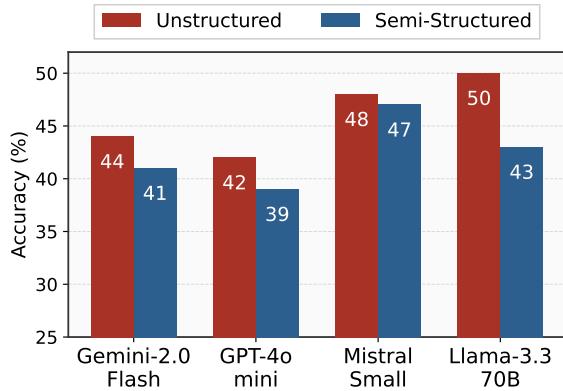


Figure 6: **Performance comparison on unstructured and semi-structured variants** for different LLM backbones using Chain-of-Thought (CoT) prompting.

structured formats better align with the respective strengths of symbolic and semantic reasoning. We further perform an in-depth error analysis to characterize common failure modes and provide qualitative examples of reasoning diversity in RUST-BENCH in Appendices C, D, and E.

4 Related Work

General Table Reasoning. Table reasoning tasks typically involve well-structured, short tables, often derived from Wikipedia-based sources. Datasets such as WikiTQ (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), WikiLeaks (Zhong et al., 2017), and Spider (Yu et al., 2018) focus on question answering or text-to-SQL tasks that test reasoning over such tables. While WikiTQ and SQA include complex questions, WikiLeaks pairs natural language questions with SQL queries, and Spider offers a large-scale, cross-domain collection with diverse databases and complex SQL. Beyond question answering, fact-verification datasets like TabFact (Chen et al., 2019) and Infotabs (Gupta et al., 2020) evaluate claim verification over Wikipedia data, while FetaQA (Nan et al., 2022) targets free-form question answering requiring reasoning over entity relations. However, these datasets primarily rely on short, factual tables with limited query diversity and shallow reasoning depth.

Semistructured and Complex Reasoning. Datasets such as FEVEROUS (Aly et al., 2021), Hybrid-QA (Chen et al., 2020b), and OTT-QA (Chen et al., 2020a) extend table reasoning to open-domain contexts combining text and tables, yet still exhibit limited diversity in reasoning types and structural variation. In contrast, reasoning-

focused datasets like TempTabQA (Gupta et al., 2023) and TABMWP (Lu et al., 2022) emphasize specific reasoning skills like temporal and numerical reasoning, respectively, but lack semi-structured contexts. CRT-QA (Zhang et al., 2023) covers a broader range of reasoning types but remains constrained by structured-only, open domain data. Our dataset bridges these gaps by combining domain-specific, semi-structured tables with diverse, multi-hop reasoning tasks that span both structured and unstructured modalities.

Domain-Specific Datasets. Datasets tailored to specific domains typically require specialized background knowledge and retrieval mechanisms to answer domain-grounded questions. In the finance domain, FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), and MultiHieert (Zhao et al., 2022) emphasize numerical and logical reasoning, often integrating heterogeneous data sources. SemTabFacts (Wang et al., 2021) and SciTAB (Lu et al., 2023) focus on claim verification using tables from scientific articles, while SciTabQA (Lu et al., 2023) extends this to question answering over mixed textual and tabular evidence. Despite their domain focus, these datasets generally contain small, homogeneous tables with limited semi-structured context, thereby constraining the study of complex, multi-hop reasoning. As illustrated in Figure 7, RUST-BENCH differs by unifying large-scale, heterogeneous, and domain-specific tables—capturing the full spectrum of real-world reasoning challenges.

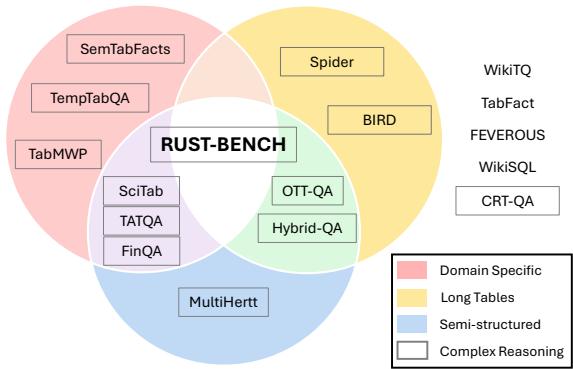


Figure 7: **Overview of table reasoning datasets categorized by key challenges:** on (a) domain-specific, (b) long, (c) semi-structured tables, and (d) complex queries. RUST-BENCH integrates datasets that span multiple dimensions of real-world complexity. In contrast, existing benchmarks satisfy only a subset or none of these criteria (e.g., WikiTQ, TabFact, etc.), limiting their applicability to practical, heterogeneous information systems.

5 Conclusion

We presented RUST-BENCH, the first benchmark that jointly evaluates LLMs on tabular reasoning across four fundamental challenges of real-world data: scale, heterogeneity, domain specificity, and multi-hop inference. Our experiments demonstrate that even the strongest proprietary and open-source models systematically fail under these conditions, as accuracy drops sharply with increasing table length, and multi-hop reasoning over semi-structured, domain-specific tables frequently breaks down. RUST-BENCH provides a robust evaluation framework and a foundation for advancing research in symbolic and structured reasoning, which is an essential step toward reliable real-world deployment. Future work on RUST-BENCH will emphasize **broader coverage** by adding diverse domains (e.g., healthcare, finance, climate), multilingual settings, and more complex table structures (hierarchical, nested, evolving) to better test cross-domain generalization. We will also introduce **real-world noise**, i.e., missing cells, typos, schema drift, and conflicting units—to assess robustness, calibration, and recovery under imperfect data. Finally, we will pair LLMs with tools for retrieval, schema induction, and execution, aiming for verifiable, scalable reasoning over semi-structured data.

Limitations

While RUST-BENCH marks a step forward in evaluating LLMs on realistic tabular reasoning, it could further incorporate multi-table and relational reasoning, introduce training splits to support fine-tuning and adaptation, and explore richer evaluation protocols that better capture semantic correctness in complex answers. These developments can help create robust and generalizable approaches to tabular reasoning in real-world applications.

Ethics Statement

We, the authors, affirm that our work adheres to the highest ethical standards in research and publication. We have carefully considered and addressed various ethical issues to ensure the responsible and fair use of computational linguistics methodologies. To facilitate reproducibility, we provide detailed information, including code, datasets (all publicly available and in compliance with their respective ethical standards), and other relevant resources. Our claims align with the experimental results, though some stochasticity is expected with

black-box large language models, which we minimize by maintaining a fixed temperature. We provide comprehensive details on annotations, dataset splits, models used, and prompting methods, ensuring our work can be reliably reproduced.

Acknowledgments

This research was partially supported by the U.S. National Science Foundation (NSF) under Grant No. 2416728. We also extend our gratitude to the Complex Data Reasoning and Analysis Lab (CoRAL) at Arizona State University for providing essential computational resources, mentorship, and a collaborative research environment that greatly contributed to the progress of this work. We sincerely thank Beenaa Salian and Preethi Suresh for their assistance in data annotation, verification, and code implementation, which played a key role in ensuring the accuracy and reliability of our results. Finally, we appreciate the thoughtful and constructive feedback provided by the reviewers, which helped strengthen the quality and presentation of this research.

References

- Nikhil Abhyankar, Vivek Gupta, Dan Roth, and Chandan K Reddy. 2024. H-star: Llm-driven hybrid sql-text adaptive reasoning on tables. *arXiv preprint arXiv:2407.05952*.
- Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and verification over unstructured and structured information (feverous) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13.
- Lang Cao and Hanbing Liu. 2025. Tablemaster: A recipe to advance table understanding with language models. *arXiv preprint arXiv:2501.19378*.
- Wenhu Chen. 2023. Large language models are few (1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130.
- Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W Cohen. 2020a. Open question answering over tables and text. In *International Conference on Learning Representations*.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

- Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. 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.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021. 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.
- Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, and 1 others. 2022. Binding language models in symbolic languages. In *The Eleventh International Conference on Learning Representations*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Parker Glenn, Parag Pravin Dakle, Liang Wang, and Preethi Raghavan. 2024. Blendsql: A scalable dialect for unifying hybrid question answering in relational algebra. *arXiv preprint arXiv:2402.17882*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Sriku-mar. 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.
- Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Sriku-mar. 2020. Infotabs: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2324.
- 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.
- Kunze Li and Yu Zhang. 2024. Planning first, question second: An llm-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4715–4729.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.
- Tianyang Liu, Fei Wang, and Muhaoo Chen. 2023b. Rethinking tabular data understanding with large language models. *arXiv preprint arXiv:2312.16702*.
- Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.
- Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. 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.
- Mistral. 2024. Mistral-small-3.2-24b-instruct-2506. <https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506>. Hugging Face model card; 24B parameters; updated version of Small-3.1.
- Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Md Mahadi Hasan Nahid and Davood Rafiei. 2024a. Normtab: Improving symbolic reasoning in llms through tabular data normalization. *arXiv preprint arXiv:2406.17961*.
- Md Mahadi Hasan Nahid and Davood Rafiei. 2024b. Tabsqlyfy: Enhancing reasoning capabilities of llms through table decomposition. *arXiv preprint arXiv:2404.10150*.
- Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, and 1 others. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.
- NSF. 2024. National science foundation(nsf).
- OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*, pages 1–22.
- 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.
- Financial Reports. 2025. [Annual reports](#).
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Craig Thomson, Ehud Reiter, and Somayajulu Sripada. 2020. Sportsett: basketball-a robust and maintainable data-set for natural language generation. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 32–40.
- Nancy XR Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. 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.
- Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and 1 others. 2023. 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, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Wikipedia. [Wikipedia](#).
- Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.
- Zirui Wu and Yansong Feng. 2024. Protrix: building models for planning and reasoning over tables with sentence context. *arXiv preprint arXiv:2403.02177*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, and 1 others. 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.
- 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.
- 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.
- Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706.
- Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287.

A More Details on Dataset Generation

A.1 Symbolic Approach

To enable QA pair generation using the symbolic approach, we curate a diverse collection of approximately 75 SQL query templates. These templates are designed to cover a broad spectrum of SQL constructs, including basic SELECT statement, conditional logic (AND, OR), aggregation (MAX, SUM, etc.), sorting (ORDER BY), grouping (GROUP BY), and joins. As shown in Figure 8, each template includes placeholder tokens for table names, columns, and filter conditions, allowing for broad applicability across different schemas. To instantiate these templates, we adopt a prompt-based generation approach leveraging large language models (LLMs). Specifically, we sample a template at random and prompt the LLM with task instructions and in-context exemplars to replace the template placeholders using schema-specific information derived from a target semi-structured table. This results in a fully instantiated SQL query tailored to the table (Figure 9). The generated SQL is then executed on the underlying table to obtain the corresponding answer. In a subsequent step, we prompt the LLM with the SQL query and its result to generate a natural language question that semantically aligns with the query logic but obscures the clauses. The final output is a question-answer pair, where the answer is grounded in the execution result of the SQL, and the question is a fluent natural language version reflecting the underlying semantics. This pipeline supports scalable QA dataset generation grounded in executable symbolic programs, enabling evaluation of models on structured reasoning tasks.

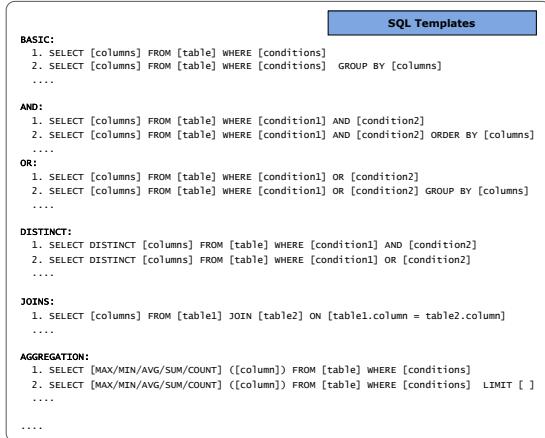


Figure 8: Example of SQL templates used for QA generation.

A.2 Semantic Approach

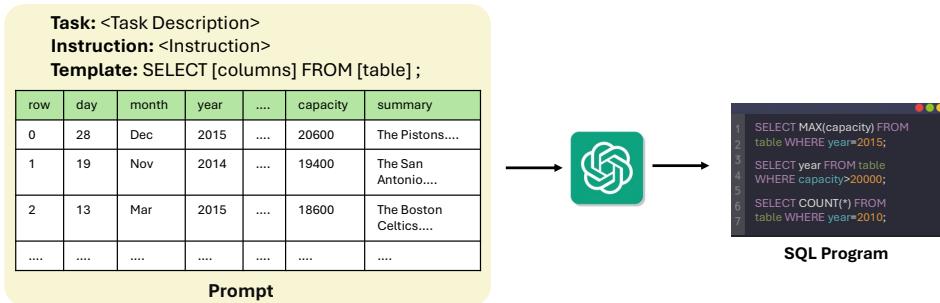
As outlined in Section 2, we employ two prompting strategies: Single Row-Based and Multi-Row-Based to improve the quality, diversity, and verifiability of LLM-generated questions over large tabular data. Figure 10 illustrates both approaches. In the Single Row-Based method, we randomly sample one row from the table and use it as the entire input context. This localization helps the LLM focus on intra-row reasoning, such as retrieving or interpreting structured and unstructured cell content. It also simplifies verification, as each question-answer (QA) pair depends on a well-defined and constrained context. In contrast, the Multi-Row Based method is designed to enable multi-row reasoning by selecting a subset of rows that are semantically connected via a shared entity in a specific column. By narrowing the input to only a few rows, these strategies, as shown in Figure 10 (bottom), help overcome LLM limitations with long inputs by explicitly controlling context size and composition. They allow generating QA pairs that are diverse in type, grounded in the table content, and more easily verifiable.

A.3 More Details on Data Validation

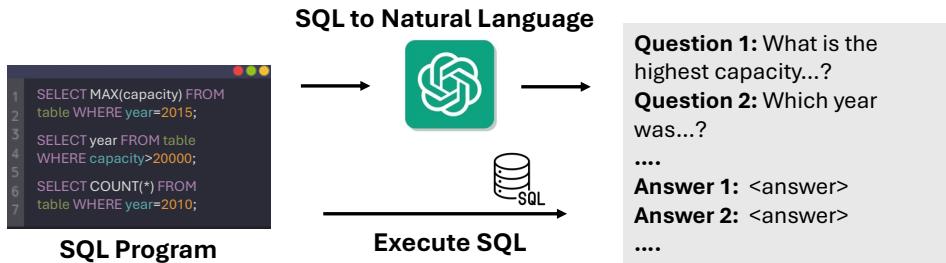
Figure 11 illustrates the custom verification interface used during the human-in-the-loop annotation process. Each screen presents a question, its predicted answer, and a detailed explanation generated by the model, alongside an interactive table view displaying the relevant semi-structured data. Annotators could validate the question-answer pair using tools such as column-specific filters, row-level sorting, and a search bar to locate supporting evidence quickly. The interface also includes input fields for correcting errors and a checkbox for discarding invalid questions. This setup ensured that annotators had full contextual access while verifying QA pairs, improving both accuracy and efficiency. After one round of annotations, the samples were further verified by expert verifiers to ensure high-quality question-answer pairs. The entire process was conducted by annotators and reviewed by graduate students in Computer Science.

B Implementation Details

In this section, we describe the prompting strategies, evaluation metrics, and LLM-based table reasoning baselines used in our study, along with their implementation details.



(a) Using code generation capabilities of LLMs to generate SQL queries.



(b) Converting the SQL queries to natural language question-answer pairs.

Figure 9: **QA pair generation using symbolic approach.** We leverage LLMs' code generation capabilities to generate SQL queries, which are then converted to natural language questions and answers by executing the SQL queries on the table data.

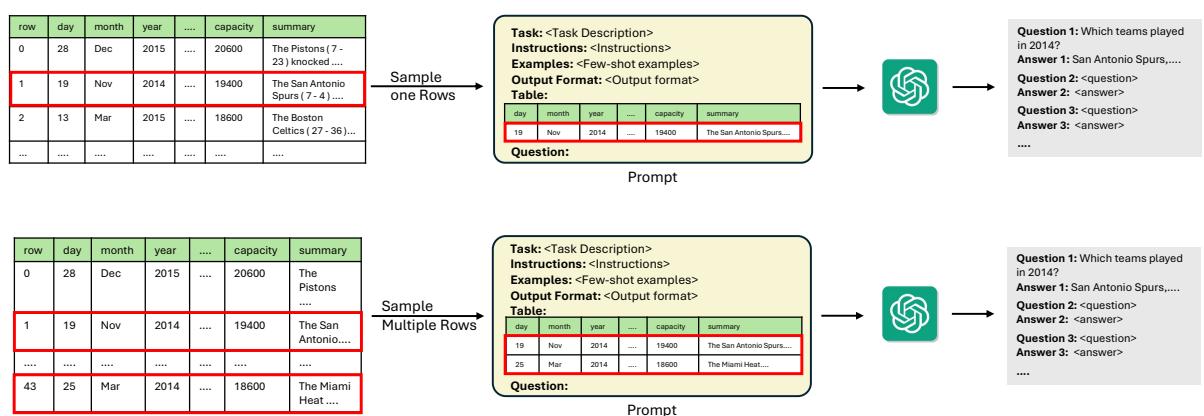


Figure 10: **QA pair generation using semantic approach:** (a) Single-Row Approach (top); (b) Multi-Row Approach (bottom), which forms questions on a subset of the table.

Question 186 of 837

Question: Which player struggled offensively while facing a team with a high shooting percentage in a late November 2014 game at Smoothie king ?

Answer: Anthony Davis

Row ID: undefined

Prediction: The question asks about a player who struggled offensively in a late November 2014 game at Smoothie King Center against a team with a high shooting percentage. The only game played in late November 2014 at the Smoothie King Center is in row 4. In row 4, the New Orleans Pelicans played the Sacramento Kings. The Kings shot 48% from the field and 36% from the three-point line. Anthony Davis of the Pelicans "wasn't his normal self tonight" and "struggled" offensively. Therefore, the player who struggled offensively is Anthony Davis. Answer: Anthony Davis

Discard Question: Correct Question: Correct Answer:

Choose File: SPORTSET..._FINAL.json Save Next

Show 10 entries Search:

Day	Month	Year	Dayname	Season	Stadium	City	State	Attendance	Capacity	Gan
All	All	All	All	All	All	All	All	All	All	All
31	December	2014	Wednesday	2014	Bankers Life Fieldhouse	Indianapolis	Indiana	18200	17900	754
6	December	2014	Saturday	2014	AT&T Center	San Antonio	California	18600	18400	951
4	December	2014	Thursday	2014	Oracle Arena	Oakland	California	19600	19600	1156
18	December	2014	Thursday	2014	Toyota Center	Houston	Texas	18100	18100	915

Figure 11: Annotation Platform - User Interface.

B.1 Prompting Techniques

We implement four reasoning techniques to use LLMs to perform tabular reasoning. Figures 12, 13, 14, and 15 highlight the direct prompting (zero-shot), few-shot, chain-of-thought (CoT) and program-of-thought (PoT) prompts for the LLMs respectively.

B.2 Evaluation Metrics

Exact Match (EM). Following WikiTQ (Pasupat and Liang, 2015), we implement exact match (EM) as the metric for evaluating model performance. EM assigns a score of 1 if the predicted answer is exactly the same as the gold answer, and 0 if otherwise. The final EM accuracy is calculated by adding the individual exact match scores divided by the total number of samples in the set. However, despite ignoring regex, punctuations, and case-sensitivity, EM penalizes semantically correct generations that do not exactly match the ground truth. It becomes increasingly challenging to evaluate longer answers that contain short phrases or multiple entities as the answer. We thus explore more relaxed metrics that do not penalize semantically correct generations.

BLEU Score. BLEU score (Papineni et al., 2002) is a metric used in machine translation to compare the quality of machine-translated text with a set of reference translations. It measures the n-gram overlap between the reference text and the prediction, assigning a score of 0-1 depending on the amount of overlap. Despite being better than EM at longer phrases, the BLEU score measures the word overlap, missing out on the semantic relevance between the prediction and the reference.

LLM-score. To correctly measure the generation quality and take the semantic similarity between the outputs and the predictions, we use an LLM as a judge to evaluate and score the generated outputs. As illustrated in Figure 16, the LLM is tasked to assign a score on a scale of 0-5 based on the correctness of the prediction. With a score of 4 representing less than 5% error between the ground truth and the prediction, the final accuracy is calculated by summing the total number of samples reporting a score of 4 or more, divided by the total samples. This enables us to gauge the answers semantically and return a better metric to evaluate the answers semantically.

B.3 Baselines

BlendSQL (Glenn et al., 2024) is a unified dialect that integrates SQL logic with large language model (LLM) reasoning across semi-structured data. It serves as a superset of SQLite, enabling complex hybrid question answering tasks involving multi-hop reasoning. The implementation utilizes the open-source repository blendsql², with dataset-specific in-context examples and default parameters.

Chain-of-Table (Wang et al., 2023) is a prompting framework that extends Chain-of-Thought by incorporating tabular data explicitly in the reasoning chain. It guides LLMs using in-context learning to iteratively generate operations and update the table to represent a tabular reasoning chain. The implementation follows the official GitHub repository chain-of-table³ with the in-context examples tailored to our dataset.

²<https://github.com/parkervg/blendsql>

³<https://github.com/google-research/chain-of-table>

Zero-Shot

```
"""
### Task
Your task is to read the provided table and answer the question.

### Table
{table}

### Question
{question}

### Output
The answer should be short and simple. It can be a number, a word, or a
phrase in the table, but not a full sentence.
Answer the question directly.
Answer:
"""
```

Figure 12: Prompt for Direct prompting.

Few Shot

```
f"""
### Task
Your task is to read the provided table and answer the question.

Here are a few examples that can help you answer the question.

```
Examples

Table:
/*
col : day | month | year | dayname | season | stadium | city | state | attendance |
capacity | game_id | summary
row 1 : 10 | January | 2015 | Saturday | 2014 | Madison Square Garden | New York City | New York | 19800 | 19800 | 512 | The Charlotte Hornets (15-24) defeated the New York Knicks (5-35) 110-82 on Saturday....
row 2 : 30 | March | 2016 | Wednesday | 2015 | Target Center | Minneapolis | Minnesota | 12300 | 19000 | 2334 | The Los Angeles Clippers (47-27) defeated the Minnesota Timberwolves (35-50) 99-79 on wednesday....
row 3 : 30 | December | 2014 | Tuesday | 2014 | Moda Center | Portland | Oregon | 20100 | 19400 | 1124 | The Portland Trail Blazers (26-7) defeated the Toronto Raptors (24-8) 102-97 on Tuesday....
*/
Question: Which away team played the March game hosted in Minnesota?
Answer: Los Angeles Clippers.

###

...
...

```
### Table
{table}

### Question
{question}

### Output
The answer should be short and simple. It can be a number, a word, or a phrase in the table, but not a full sentence.
Answer the question directly.
Answer:
"""
```

Figure 13: Prompt for Few Shot reasoning.

Chain of Thought

```
f"""
### Task
Your task is to read the table and answer the question based on the
provided table.
Think step-by-step and answer the question.

### Table
{table}

### Question
{question}

### Response Format
The answer should be short and simple. It can be a number, a word, or
a phrase in the table, but not a full sentence.
Your response should end with `Answer: xxx` (answer to the question).
Answer the question step by step:
"""

```

Figure 14: Chain-of-Thought reasoning prompt.

Program of Thought

```
f"""### Task
Your task is to read the table and then generate a Python code to answer the
corresponding question.
Perform step-by-step reasoning using Python code to answer the question.

### Examples
Table:
/*
Day | Month | Year | Day Name | Season | Stadium | City | State | Attendance
| Capacity | Game ID | Summary
12 | March | 2018 | Monday | 2017 | Chase Center | San Francisco |
California | 18300 | 18500 | 3081 | The Golden State Warriors (52-14)
outlasted the Denver Nuggets (40-26) 118-112.
...
*/
Question: what percentage of the stadiums' capacities were filled?
Answer:
```python
Attendance and Capacity from each game
attendances, capacities = [18300, 20000, 18900], [18500, 20000, 20500]
Total attendance and capacity
total_attendance, total_capacity = sum(attendances), sum(capacities)
Percentage of capacity filled
percentage_filled = (total_attendance / total_capacity) * 100
ans = percentage_filled
ans
```
### Table
{table}

### Question
{question}

### Response Format
Provide the Python code to answer the question.
Answer:
"""

```

Figure 15: Program-of-Thought reasoning prompt.

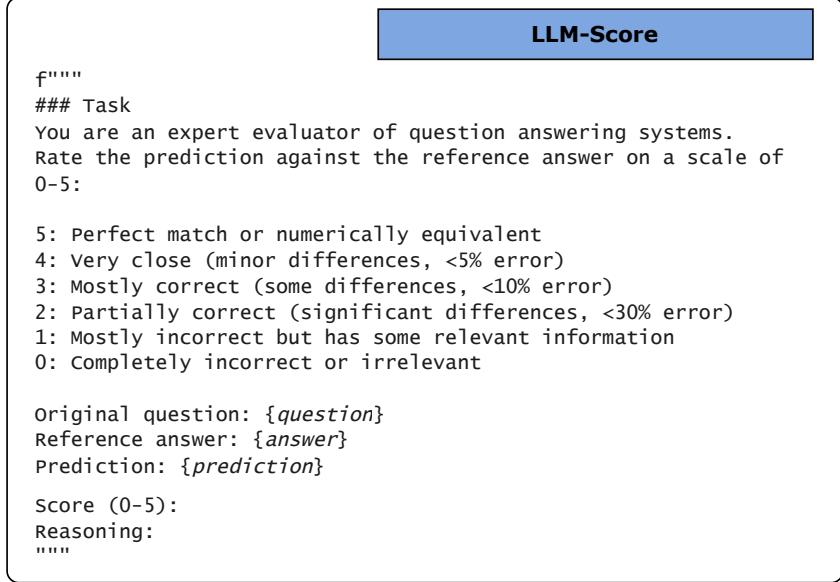


Figure 16: Prompt for using LLM-as-a-judge to output LLM-score.

ProTrix (Wu and Feng, 2024) introduces a Plan-then-Reason framework that plans the reasoning path using the query and context, then assigns each step to either textual or program-based reasoning to arrive at the final answer. We modify their official repository⁴ in-context examples to suit RUST-BENCH and use their default hyperparameters.

TabSQLify (Nahid and Rafiei, 2024b) is a semantic parsing-based method that translates natural language questions into executable SQL queries over structured tables. It leverages text-to-SQL generation to decompose tables into smaller, relevant sub-tables containing only essential information for answering questions or verifying statements. We utilize tabsqlify⁵ with updated in-context examples for inference.

TableMaster (Cao and Liu, 2025) is a unified framework that combines multiple techniques for table reasoning. The method first retrieves relevant table content and enriches it with semantic verbalizations, and employs adaptive reasoning to flexibly choose between textual and symbolic reasoning depending on each query. We adopt the official repository TableMaster⁶, retaining their default hyperparameters for fair comparison.

NormTab (Nahid and Rafiei, 2024a) focuses on improving symbolic interpretability by normalizing table structures and values prior to reasoning. It standardizes heterogeneous column names and

formats, reducing schema variance and enabling more consistent SQL-based reasoning across diverse tables. We utilize the public normtab⁷ repository, following default parameters and adapting the prompts to our dataset.

C Reasoning Diversity in RUST-BENCH

Distribution of Question Types. To understand the reasoning diversity in RUST-BENCH, we adopt and extend the taxonomy proposed in CRT-QA (Zhang et al., 2023), which builds on the BIG-bench framework (Srivastava et al., 2022). As shown in Table 6, our annotation covers a broad spectrum of reasoning types—from high-frequency operations such as filtering and temporal reasoning to more complex forms including multi-hop, implicit, and counterfactual reasoning. This diversity underscores the layered cognitive demands required for real-world table understanding. Filtering and temporal reasoning are the most common types, reflecting the frequent need to locate relevant records and interpret time-dependent relationships. However, a significant proportion of questions also require multi-hop reasoning (26.18%), numerical computation (26.83%), and logical composition (27.85%), highlighting the dataset’s emphasis on compositional and quantitative reasoning. Although rarer, counterfactual, commonsense, and causal reasoning further test model generalization beyond surface-level retrieval.

⁴<https://github.com/WilliamZR/ProTrix>

⁵<https://github.com/mahadi-nahid/TabSQLify>

⁶<https://github.com/zzlang-c/TableMaster>

⁷<https://github.com/mahadi-nahid/NormTab>

Table 6: Distribution of reasoning types. Categories are non-exclusive; percentages may not sum to 100%.

| Reasoning Type | Percentage (%) |
|---------------------------|----------------|
| Filtering / Selection | 75.89 |
| Temporal Reasoning | 39.33 |
| Logical Reasoning | 27.85 |
| Numerical | 26.83 |
| Multi-hop Reasoning | 26.18 |
| Aggregation | 23.97 |
| Comparison | 17.43 |
| Implicit Reasoning | 11.36 |
| Unanswerable | 6.83 |
| Sorting / Ranking | 5.47 |
| Causal Reasoning | 5.20 |
| Commonsense Reasoning | 0.44 |
| Spatial Reasoning | 0.24 |
| Counterfactual / Negative | 0.19 |

Unanswerable Questions. In practical table reasoning, not all queries are grounded in the available data. Distinguishing answerable from unanswerable questions is therefore crucial for reliable model deployment in domains such as finance and science. To evaluate this capability, RUST-BENCH incorporates explicitly unanswerable questions following (Zhang et al., 2023)—queries that cannot be resolved using the table content alone. Examples include those that require external knowledge or contain logical contradictions. A model is considered correct only if it abstains by responding with phrases such as “cannot answer” or “not enough information.” We manually verify outputs to measure accuracy. As shown in Figure 17, models struggle considerably with this task: even under Chain-of-Thought prompting, Gemini-2.0-Flash achieves only 52.27% accuracy in RB-Sports and 26.97% in RB-Science, indicating the persistent challenge of reliable unanswerable detection in table QA.

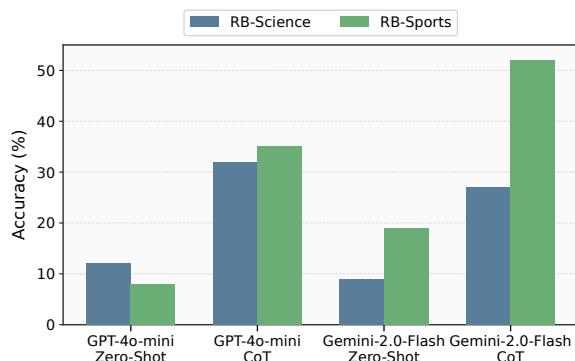


Figure 17: Accuracy of GPT-4o-mini and Gemini-2.0-Flash models on RB-Science and RB-Sports datasets, evaluated on questions that include unanswerable/ambiguous cases.

D Qualitative Analysis

Semi-structured tables in RUST-BENCH pose a unique challenge for LLMs, as they require reasoning that spans both structured schema elements (e.g., categorical or numeric fields) and unstructured text (e.g., summaries or descriptions). Such inputs expose the limitations of models that excel in either symbolic precision or semantic understanding, but not both. As illustrated in Figure 18, answering ‘How many projects focus on children and how many children did the earliest project address?’ requires scanning abstracts for child-related projects, counting across rows, and applying temporal reasoning to identify the earliest award. Crucially, the abstract of the 2016 brain connectivity project mentions developmental trajectories without specifying participant numbers, so the correct response must acknowledge the absence of detail. Similarly, for the question ‘In a March game at TD Garden, which player from the losing team had the highest points and what was the point difference between him and the leading scorer of the winning team?’ (Figure 19) requires filtering structured fields to locate the relevant March 2019 Celtics–Nuggets game, extracting top scorers from the unstructured summary, aligning them with their teams, and performing arithmetic to compute the score difference. This case exemplifies hybrid reasoning across structured and unstructured inputs, combined with entity disambiguation and grounded numeric comparison. These cases underscore how RUST-BENCH questions move beyond single-field lookup, requiring schema filtering, semantic interpretation, aggregation, and handling capabilities that remain fragile in current LLMs.

E Error Analysis

To analyze the sources of performance degradation, we manually examined 100 randomly sampled erroneous predictions from Gemini-2.0-Flash (CoT). Errors were grouped into four major categories reflecting distinct failure modes: (i) **Interpretation Error:** counting or lookup mistakes caused by complex table structures and increased token load from unstructured fields; (ii) **Logical Inconsistency Errors:** contradictory or incomplete reasoning chains, particularly in multi-hop settings; (iii) **Misalignment Errors:** outputs that deviate from the expected answer schema or pro-

vide only partial results; and (iv) **Extraction Errors**: incorrect or missed retrievals from structured or unstructured regions of the table. The breakdown in Table 7 shows that no single type dominates; instead, errors stem from the interaction between structural complexity, multi-step reasoning, and representational inconsistencies introduced by semi-structured inputs.

Table 7: Breakdown of 100 randomly sampled erroneous predictions from Gemini-2.0-Flash (CoT).

| Error Type | Percentage |
|-------------------------|------------|
| Interpretation Error | 22% |
| Logical Inconsistencies | 31% |
| Misalignment Error | 27% |
| Extraction Error | 20% |

Extraction Error. These involve failures to retrieve key information from structured fields or unstructured text. The model may skip valid rows or miss implicit cues, such as differences in project counts across years (Figure 20) or mentions of child-related studies buried in abstracts (Figure 21).

Logical Inconsistency Error. These occur when the model generates an apparently coherent reasoning chain but produces a final answer inconsistent with its intermediate analysis. For example, as shown in Figure 22, the model may identify both *Standard Grant* and *Continuing Grant* as valid answers but report only one, revealing a collapse between reasoning and final output generation.

Interpretation Error. Here, the model misreads the scope of the question or the table structure, overlooking relevant rows or applying filters incorrectly. As illustrated in Figure 23, it may compute time gaps based on a single record while ignoring other valid entries, leading to incomplete evidence gathering and erroneous conclusions.

Misalignment Error. In some cases, the model’s reasoning is correct, but the output format deviates from the expected answer schema—for instance, returning a sum instead of individual attendance values (Figure 24). Collectively, these patterns show that while LLMs can perform multi-step reasoning, they often lose alignment between reasoning, evidence retrieval, and output generation particularly when operating on semi-structured data that demands both symbolic precision and semantic understanding.

| Award Title | Award Expiration Date | Award Amount | Award Instrument | Abstract Narration | ... | Institution State Code | Institute State | year |
|---|-----------------------|--------------|------------------|---|-----|------------------------|-----------------|------|
| NeTS: Medium: Object-Centric, View-Adaptive... | 07/24/2023 | 884042 | Standard Grant | Most videos streamed on the Internet are sequences of flat two-dimensional (2D)... | ... | NY | New York | 2023 |
| Incorporating professional science writing into high school STEM... | 07/15/2020 | 313906 | Standard Grant | This exploratory project addresses important challenge of incorporating... | ... | GA | Georgia | 2020 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| RAPID: Navigating Social Distancing with DeafBlind Children... | 08/01/2020 | 196564 | Standard Grant | This project will contribute to the national need for well-educated scientists... | ... | MO | Missouri | 2020 |
| Heavy Metal Recognition and Separation... | 07/21/2023 | 457729 | Standard Grant | The goal of this project is to develop a separation system that can.... | ... | CA | California | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Collaborative Research: The Role of Brain Connectivity in Reasoning Development | 03/15/2016 | 295937 | Standard Grant | Understanding the patterns of communication between brain regions, and how they develop across childhood... | ... | CA | California | 2016 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Collaborative Research: EAGER: SaTC-EDU: Secure and Privacy-Preserving... | 07/31/2023 | 10626 | Continuing Grant | In recent years, researchers have applied artificial intelligence (AI) to.... | ... | TX | Texas | 2023 |

Question: How many projects focus on children and how many children did the earliest project address?

-
1. Scan the table for mentions of “children” in the **unstructured fields**.
 2. Identify and count the number of projects where children are the explicit focus.
 3. Identify the earliest project among them by using the **structured field** `Award Expiration Date`.
 4. Read the unstructured field for the earliest project to determine how many children it addressed.
 5. Return the number of projects and the number of children.
-

Correct Answer: 3 projects, 400 children

Figure 18: **Example from the RB-Science subset.** The question requires understanding data from *unstructured fields*, aggregation across rows, temporal reasoning to identify the earliest project, and recognition of underspecified information, highlighting challenges beyond surface retrieval.

| day | month | year | stadium | ... | state | attendance | capacity | summary |
|-----|----------|------|---------------------------|-----|---------------|------------|----------|--|
| 21 | January | 2019 | TD Garden | ... | Massachusetts | 18600 | 18600 | The Boston Celtics defeated the visiting Miami Heat... |
| 18 | March | 2019 | TD Garden | ... | Massachusetts | 18600 | 18600 | The Denver Nuggets defeated the host Boston Celtics , 114 - 105 , at TD Garden... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2 | March | 2016 | Target Center | ... | Minnesota | 11300 | 19000 | The Washington Wizards (30 - 30) earned a road win over the Minnesota Tim... |
| 13 | April | 2016 | BMO Harris Bradley Center | ... | Florida | 16600 | 18700 | The Miami Heat (45 - 37) closed out their season and head into the playoffs... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 13 | February | 2019 | TD Garden | ... | Massachusetts | 18600 | 18600 | The Boston Celtics defeated the Detroit Pistons , 118 - 110 , at TD Garden on Wednesday... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 3 | January | 2018 | Capital One Arena | ... | Washington | 17200 | 20400 | The Washington Wizards defeated the visiting Sacramento Kings , 121 -... |

Question: In a March game at TD Garden, which player from the losing team had the highest points and what was the point difference between him and the leading scorer of the winning team?

Reasoning Steps Involved:

1. Filter the table rows using the structured fields mentioned in the question (e.g., by month, stadium, year, or other constraints).
 2. Identify the relevant game(s) that match these conditions.
 3. Read the unstructured “summary” field of the selected row(s) to extract information about the winning and losing teams, as well as their top scorers and points.
 4. Locate the top scorer of the losing team and record their points.
 5. Locate the leading scorer of the winning team and record their points.
 6. Perform the required arithmetic operation (in this case, subtract the winning scorer’s points from the losing scorer’s points).
 7. State the answer by reporting the losing team’s highest scorer and the computed point difference.
-

Correct Answer: Kyrie Irving, 9 points

Figure 19: **Example from the RB-Sports subset.** Answering the question requires filtering by *structured fields* (month, stadium), extracting top scorers from *unstructured summaries*, and performing arithmetic comparison, illustrating hybrid multi-hop reasoning across modalities.

| Award Title | Min Amd Letter Date | Award Amount | Award Instrument | Abstract Narration | ... | Investigator LastName | Institute State | year |
|---|---------------------|--------------|-----------------------|--|-----|---|-----------------|------|
| NeTS: Medium: Object-Centric, View-Adaptive... | 07/24/2023 | 884042 | Standard Grant | Most videos streamed on the Internet are sequences of flat two-dimensional (2D)... | ... | ['Liu', 'Wang', 'DuBois'] | New York | 2023 |
| Dark Matter on the Smallest Scales | 07/19/2023 | 270000 | Standard Grant | This award funds the research activities of Professor.... | ... | Erickcek | North Carolina | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Creating Pathways to Computing Careers... | 07/17/2023 | 2322436 | Standard Grant | This project will contribute to the national need for well-educated scientists... | ... | ['Gomez', 'Hou', 'Sun', 'Zhang', 'Dajani'] | California | 2023 |
| Heavy Metal Recognition and Separation... | 07/21/2023 | 457729 | Standard Grant | The goal of this project is to develop a separation system that can.... | ... | ['Zhong', 'Hooley'] | California | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Center for Advanced Materials & Manufacturing (CAMM) | 06/26/2023 | 3000000 | Cooperative Agreement | Nontechnical Abstract: The Center for Advanced Materials... | ... | ['Dadmun', 'Zinkle', 'Delmaestro', 'Tennant', 'Rawn'] | Tennessee | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Collaborative Research: EAGER: SaTC-EDU: Secure and Privacy-Preserving... | 07/31/2023 | 10626 | Continuing Grant | In recent years, researchers have applied artificial intelligence (AI) to.... | ... | Lipsmeyer | Texas | 2023 |

Question: How many weeks apart are the letter dates for five-member project grants?

The question asks about the time difference in weeks between the earliest and latest amendment letter dates (MinAmdLetterDate and MaxAmdLetterDate) for projects with exactly five investigators.

1. **Identify projects with five investigators:

** The Investigator_FirstName column lists the first names of the investigators.

We need to find rows where this list has five elements. Row 2 has five first names.

2. **Extract MinAmdLetterDate and MaxAmdLetterDate for the identified projects:**

For row 2, MinAmdLetterDate is 07/31/2023 and MaxAmdLetterDate is 07/31/2023.

3. **Compute the difference in dates:** The difference between 07/31/2023 and 07/31/2023 is 0 days.

4. **Convert to weeks:** 0 days is equal to 0 weeks.

Answer: 0

Correct Answer: 3 weeks

Figure 20: **Extraction Error.** The LLM fails to extract the relevant information from the structured table. Instead of identifying the number of projects sanctioned in October 2022 and comparing it with October 2023, it wrongly concludes that no “previous October” exists.

| Award Title | Award Expiration Date | Award Amount | Award Instrument | Abstract Narration | ... | Institution State Code | Institute State | year |
|---|-----------------------|--------------|------------------|---|-----|------------------------|-----------------|------|
| NeTS: Medium: Object-Centric, View-Adaptive... | 07/24/2023 | 884042 | Standard Grant | Most videos streamed on the Internet are sequences of flat two-dimensional (2D)... | ... | NY | New York | 2023 |
| Incorporating professional science writing into high school STEM... | 07/15/2020 | 313906 | Standard Grant | This exploratory project addresses important challenge of incorporating... | ... | GA | Georgia | 2020 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| RAPID: Navigating Social Distancing with DeafBlind Children... | 08/01/2020 | 196564 | Standard Grant | This project will contribute to the national need for well-educated scientists... | ... | MO | Missouri | 2020 |
| Heavy Metal Recognition and Separation... | 07/21/2023 | 457729 | Standard Grant | The goal of this project is to develop a separation system that can.... | ... | CA | California | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Collaborative Research: The Role of Brain Connectivity in Reasoning Development | 03/15/2016 | 295937 | Standard Grant | Understanding the patterns of communication between brain regions, and how they develop across childhood... | ... | CA | California | 2016 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Collaborative Research: EAGER: SaTC-EDU: Secure and Privacy-Preserving... | 07/31/2023 | 10626 | Continuing Grant | In recent years, researchers have applied artificial intelligence (AI) to.... | ... | TX | Texas | 2023 |

Question: How many projects focus on children and how many children did the earliest project address?

To determine how many projects in the US focus on children and how many children the earliest project addressed, we will analyze the provided table step by step.

1. **Identify Projects Focused on Children**:

- We need to look for keywords related to children in the project titles or abstracts.

- The relevant project found is:

- **RAPID: Navigating Social Distancing with DeafBlind Children:....

2. **Count the Projects**:

- There is **1 project** that specifically focuses on children.

3. **Determine the Number of Children Addressed by the Earliest Project**:

- The project mentioned in row 9 addresses **DeafBlind children** but does not specify...

Conclusion:

- Number of projects focusing on children:

1- Number of children addressed by the earliest project:

Not specified

Final answer:

Answer: 1

Correct Answer: 3 projects, 400 children

Figure 21: **Extraction Error.** The LLM fails to extract relevant information from the unstructured portion of the table. While only one project explicitly mentions the term *children* in its title, two additional projects are related but require a deeper comprehension of the unstructured content to be correctly identified and extracted.

| Award Title | Award Expiration Date | Award Amount | Award Instrument | Abstract Narration | ... | Investigator Role Code | Institute State | year |
|--|-----------------------|--------------|------------------|---|-----|---|----------------------|------|
| EAGER: EEG-based Cognitive-state Decoding for Interactive Virtual... | 10/01/2019 | 209996 | Standard Grant | The increasing availability of affordable, high-performance... | ... | PI | Virginia | 2019 |
| Nucleon Structure and Hadronization Beyond the Collinear... | 11/15/2019 | 211036 | Continuing Grant | It has been a long-standing question in nuclear physics of how... | ... | ['Former PI', 'PI', 'Former PI'] | New Mexico | 2019 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Institute for Pure and Applied Mathematics | 09/01/2020 | 20000000 | Continuing Grant | The mission of the Institute for Pure and Applied Mathematics (IPAM) is... | ... | ['Co-PI', 'Co-PI', 'Former Co-PI', 'PI', 'Co-PI'] | California | 2019 |
| Belmont Forum Collaborative Research: Biodiversity Scenarios... | 04/15/2019 | 179965 | Continuing Grant | Innovative research on the complex interaction of socio-economic and global... | ... | PI | New York | 2019 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Next Steps for Functional Genomics: A Workshop | 07/16/2019 | 327444 | Standard Grant | This award is to support a workshop convened by the National Academy of Sciences... | ... | ['Co-PI', 'Co-PI', 'Former PI', 'PI'] | District of Columbia | 2019 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| I-Corps: Tracking Cognitive Functions with Ear-worn Bio-sensing... | 07/01/2019 | 50000 | Standard Grant | The broader impact of this I-Corps project will be to enable bio-sensing... | ... | PI | Colorado | 2019 |

Question: Which types of award instruments were associated with more than one research grant, where each grant had an amount exceeding \$200,000 in the year 2019?

Let's analyze the table and find the award instruments associated with more than one research grant where each grant had an award amount exceeding \$200,000 in the year 2019.

The column `AwardInstrument_value` stores the award instrument type. The column `AwardAmount` stores the award amount and the column `year` stores the year of the grant. we need to find out the unique values in `AwardInstrument_Value` column that has `AwardAmount` > 200000 and `year` as 2019.

From row 0, AwardInstrument_Value = Standard Grant, AwardAmount = 209996 (less than 200000), year = 2019
 From row 1, AwardInstrument_Value = Continuing Grant, AwardAmount = 211036, year = 2019

From row 2,...

...

...

Count of Standard Grant with AwardAmount > 200000 and year = 2019 = 12

Count of Continuing Grant with AwardAmount > 200000 and year = 2019 = 10

So, both \"Standard Grant\" and \"Continuing Grant\" appear more than once.

Answer: Standard Grant

Correct Answer: Standard Grant, Continuing Grant

Figure 22: **Logical Inconsistency Error.** Owing to the large number of rows, the LLM engages in extensive reasoning and correctly identifies both *Standard Grant* and *Continuing Grant*. However, the final answer only lists *Standard Grant*, revealing a collapse between reasoning and output under heavy analysis.

| Award Title | Award Effective Date | Award Amount | Award Instrument | Abstract Narration | ... | Investigator LastName | Institute State | year |
|---|----------------------|--------------|------------------|---|-----|-------------------------------|-----------------|------|
| NeTS: Medium: Object-Centric, View-Adaptive... | 10/01/2023 | 884042 | Standard Grant | Most videos streamed on the Internet are sequences of flat two-dimensional (2D) images... | ... | ['Liu', 'Wang', 'DuBois'] | New York | 2023 |
| Dark Matter on the Smallest Scales | 08/01/2023 | 270000 | Standard Grant | This award funds the research activities of Professor.... | ... | Erickcek | North Carolina | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ADVANCE Adaptation: Leveraging Best.... | 10/01/2023 | 100000 | Standard Grant | The University of Texas at San Antonio's (UTSA) ADVANCE Adaptation project... | ... | ['Nash', 'Guerra', 'Shipley'] | Texas | 2023 |
| Heavy Metal Recognition and Separation... | 08/01/2023 | 457729 | Standard Grant | The goal of this project is to develop a separation system that can.... | ... | ['Zhong', 'Hooley'] | California | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| CAREER: Deep Learning Empowered.... | 10/01/2022 | 469704 | Continuing Grant | Moduli spaces pervade mathematics. Given a mathematical object.... | ... | Apisa | Wisconsin | 2023 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Collaborative Research: EAGER: SaTC-EDU: Secure and Privacy-Preserving... | 08/15/2023 | 10626 | Continuing Grant | In recent years, researchers have applied artificial intelligence (AI) to.... | ... | Lipsmeyer | Texas | 2023 |

Question: How many more projects were sanctioned in October 2023 compared to previous October?

The table contains data about projects sanctioned in the year 2023. The question asks about October 2023 compared to "previous October," but there's no prior October mentioned.

Because we only have project data for the year 2023, there is no data for any "previous October." Therefore, we cannot determine the difference.

Answer: Cannot determine

Correct Answer: 4

Figure 23: **Interpretation Error**, where the LLM misinterprets both the question and the table. While tasked with finding the week gap between the earliest and latest amendment dates for five-investigator projects, it only considers a single row and ignores other valid rows. This leads to an incorrect calculation, showing how errors in interpreting table structure and question scope can cascade into a wrong final answer.

| day | month | year | stadium | ... | state | attendance | capacity | summary |
|-----|----------|------|---------------------------|-----|--------------|------------|----------|--|
| 13 | April | 2016 | BMO Harris Bradley Center | ... | Wisconsin | 16600 | 18700 | The Indiana Pacers (45 - 37) closed out their season and head into the playoffs... |
| 4 | December | 2015 | Verizon Center | ... | Washington | 17300 | 20400 | The Washington Wizards (8 - 9) defeated the Phoenix Suns (8 - 12) 109... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2 | March | 2016 | Target Center | ... | Minnesota | 11300 | 19000 | The Washington Wizards (30 - 30) earned a road win over the Minnesota Tim... |
| 15 | March | 2018 | Madison Square Garden | ... | New York | 18900 | 19800 | The Philadelphia 76ers defeated the New York Knicks, 118 - 110 , at Madison... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 18 | January | 2017 | Wells Fargo Center | ... | Pennsylvania | 13700 | 20500 | Joel Embiid 's 26 points , nine rebounds , two assists , two blocks and a steal led the Sixers to their... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 3 | January | 2018 | Capital One Arena | ... | Washington | 17200 | 20400 | The Washington Wizards defeated the visiting Sacramento Kings , 121 - ... |

Question: What was the attendance figure for the games held at the venues in Washington and Philadelphia during the specified years?

First, we need to identify the rows corresponding to games played in Washington and Philadelphia. Then, filter these rows by the games played during the specified years.

Finally, read the attendance figures from those rows.

Row 7: Washington, 2015, attendance 17300

Row 20: Philadelphia, 2015, attendance 13700

Row 28: Philadelphia, 2014, attendance 13700

Row 51: Washington, 2017, attendance 17200

Total attendance = $17300 + 13700 + 13700 + 17200 = 61900$

Answer: 61900

Correct Answer: 17300, 13700, 13700, 17200

Figure 24: **Misalignment Error.** The reasoning correctly identifies the relevant rows and extracts the attendance figures. However, instead of listing these individual values as expected, the LLM sums them up. This misalignment between the required output format and the final answer leads to an incorrect response despite accurate intermediate reasoning.