LongTableBench: Benchmarking Long-Context Table Reasoning across Real-World Formats and Domains

Liyao Li[1*], **Jiaming Tian**[1*], **Hao Chen**[1], **Wentao Ye**[1], **Chao Ye**[1]
Haobo Wang[1], **Ningtao Wang**[2], **Xing Fu**[2], **Gang Chen**[1], **Junbo Zhao**[1†]
[1]Zhejiang University [2]Ant Group
{liliyao, hsxz2, j.zhao}@zju.edu.cn

Abstract

We introduce **LongTableBench**, a benchmark for evaluating long-context reasoning over semi-structured tables across diverse formats, tasks, and domains. It comprises 5,950 QA instances spanning 7 table formats (e.g., Markdown, HTML, SQL), 18 domains, and input lengths up to 128K tokens, including multi-turn and multi-table settings. To ensure data quality, we combine symbolic supervision, cross-model validation, and human review. Evaluating 52 LLMs—including general-purpose, table-specific, and reasoning-enhanced models—reveals that only the strongest models maintain robust performance under increasing context lengths and format diversity. We further show that end-to-end models outperform compression-based approaches, especially on tasks requiring semantic integration. LongTableBench provides a rigorous, scalable testbed for advancing long-context tabular understanding and highlights key limitations in current LLMs' structural and reasoning capabilities. The code and data are available at https://github.com/liyaooi/LongTableBench.

1 Introduction

Recent advancements in large language models (LLMs) have pushed the boundaries of context length from thousands of tokens (Bai et al., 2023a; Bi et al., 2024) to millions (Meta AI, 2025; Team, 2025a), enabling new capabilities in tasks like multi-document analysis (Pang et al., 2021; Wang et al., 2024) and long-form dialogue (Bai et al., 2024a). However, a critical question remains: ***Do the expanded contexts these models process truly unlock the reasoning potential needed for complex, real-world scenarios?***

Semi-structured tables, which are structured data represented in textual formats, provide an ideal playground for investigating this question. These tables—found in various markup languages and file formats like Markdown, HTML, LaTex, JSON, and CSV (Sui et al., 2024; Jaitly et al., 2023; Singha et al., 2023)—preserve a consistent logical structure despite varying textual representations. Examples include financial tables (Zhang et al., 2024a; Chen et al., 2021), healthcare records (Guo et al., 2024; Shi et al., 2024), and scientific data (Smock et al., 2022), all of which are prevalent in critical real-world applications. They offer unique challenges for LLMs to process and reason over complex, hierarchical structures within long contexts.

We identify three key challenges that make semi-structured tables valuable for evaluating long-context understanding, aspects that current evaluation benchmarks often fail to adequately address:

(i)-Structural Complexity: Tables embed information in a multi-dimensional space, where meaning emerges from vertical (rows), horizontal (columns), and hierarchical relationships (nested tables), including interconnected tables across multi-turn dialogues. This complexity tests an LLM's ability to reason about multi-dimensional data and distinguish structural meaning from syntactic variation. Traditional document-based QA benchmarks (Jacovi et al., 2025; Zou et al., 2024) focus on linear text comprehension, neglecting structural reasoning.

(ii)-Long-Range Dependencies: Table comprehension necessitates synthesizing information across expansive token distances, where relationships between headers, cells, and annotations must be maintained coherently. This tests an LLM's ability to maintain attention relationships across long contexts. While current TableQA benchmarks (Chen et al., 2019; Pasupat and Liang, 2015) often confine themselves to narrow contexts of simple, self-contained tables within 3K tokens, real-world applications demand sustained attention and consistency across much broader contexts.

(iii)-Semantic Integration: Mastering semi-

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11927–11965
November 4-9, 2025

structured tables requires a sophisticated fusion of pattern recognition and semantic understanding. Models must simultaneously process explicit patterns (like date formats "2025-01-01") while leveraging world knowledge to handle semantic variations (such as interpreting Roman numerals alongside Arabic numbers) within their proper context. Traditional approaches like Text-to-SQL (Pourreza and Rafiei, 2024a; Gao et al., 2023; Pourreza and Rafiei, 2024b) or retrieval-based methods (Chen et al., 2024; Kong et al., 2024; Mao et al., 2024) often falter when confronting real-world tables with irregular, mixed expressions (Biswal et al., 2024), where success depends not on mechanical parsing but on the nuanced integration of world knowledge, domain-specific information, and contextual awareness.

To facilitate further research in these directions, we introduce ***LongTableBench***—the first comprehensive multitask benchmark for long-context semi-structured table reasoning. Specifically, LongTableBench consists of 6 core task categories: *exact matching*, *basic conditional filtering*, *fuzzy conditional manipulation*, *fact retrieval*, *external knowledge fusion*, and *irregular numerical interpretation*. These tasks span common long-text applications and incorporate combinations of single/multi-turn dialogues and single/multi-table scenarios. Meanwhile, we offer 7 widely used semi-structured table formats—Markdown, HTML, JSON, LaTeX, SQL, and CSV, for each question to assess LLMs' format preferences. The benchmark includes 5,950 fully English test samples, with the task-wise length distribution presented in Figure 1. To ensure cost-effective automated evaluation, we use the F1 metric to compare output similarity with ground truth. LongTableBench provides a systematic evaluation of current LLMs' abilities on real-world semi-structured data, such as structural reasoning, long-range dependencies, and semantic integration.

Our contributions are summarized as follows:

- We propose **LongTableBench**, the first benchmark targeting long-context table reasoning across 7 real-world formats and up to 128K tokens, covering multi-turn dialogues, multi-table setups, and six reasoning task types.
- We introduce a *format-agnostic* table expansion pipeline with controlled token-length sampling, domain diversity, and symbolic verifiability to ensure realistic and scalable benchmark construction.
- We perform extensive evaluations over **52 LLMs**, revealing the impact of model scale, reasoning ability, and format robustness across structural, temporal, and semantic dimensions.
- We demonstrate that *end-to-end LLMs* outperform compression-based paradigms such as RAG and Text-to-SQL, especially on tasks requiring deep semantic integration.
- We publish all data, formats, and evaluation scripts to encourage reproducibility and facilitate future research in long-context structured data understanding.

2 Related Works

Long-context Benchmarks. Recent advancements have extended LLMs' context length to 128K tokens or more (Liu et al., 2024; Hurst et al., 2024; Anthropic, 2025; Meta AI, 2025; Team, 2025b; GLM et al., 2024), spurring diverse evaluation benchmarks. Comprehensive multi-task benchmarks include ZeroSCROLLS (Shaham et al., 2023), L-Eval (An et al., 2023), LongBench series (Bai et al., 2023b, 2024b), BAMBOO (Dong et al., 2023), LooGLE (Li et al., 2023a), ∞-Bench (Zhang et al., 2024c), Ruler (Hsieh et al., 2024), and HELMET (Yen et al., 2024). Specialized benchmarks target document QA (Pang et al., 2021; Wang et al., 2024), code tasks (Liu et al., 2023; Bogomolov et al., 2024), generation (Ye et al., 2025), factuality (Jacovi et al., 2025), and in-context learning (Li et al., 2024). Despite this progress, tabular data remains underexplored as a common structure in real-world applications. Our work addresses this gap with a dedicated long-table benchmark.

Table-related Benchmarks. Existing TableQA benchmarks like TabFact (Chen et al., 2019) and WikiTQ (Pasupat and Liang, 2015) feature small Wikipedia-derived tables, while HiTab (Cheng et al., 2021) and HybridQA (Chen et al., 2020) introduce more complex tables that still fit within early LLMs' context windows (4K-8K tokens). Text-to-SQL tasks (Yu et al., 2018; Li et al., 2023b) typically focus on schema rather than full tables and struggle with non-database formats and fuzzy semantics, as TAG (Biswal et al., 2024) demonstrates. Though TQA-bench (Qiu et al., 2024) begins exploring long-table QA, its 64K-token limit

and multiple-choice format constrain evaluation depth. With modern LLMs supporting 128K tokens, we introduce **LongTableBench** to systematically evaluate fine-grained LLM capabilities across full 128K-token contexts.

3 LongTableBench: Task and Design

3.1 Problem Definition

We define the challenge of QA with long-context semi-structured tables as follows: Given an input question sequence Q and a tabular context sequence T, the model is tasked with generating an output answer A. In LongTableBench, Q and A are typically short, while T can be a lengthy sequence, consisting of multiple tables and reaching up to thousands of tokens. A detailed breakdown of (Q, T, A) for each task can be found in Appendix A.

3.2 Task Overview

LongTableBench comprises 6 carefully designed tasks that evaluate LLMs' capabilities in reasoning over long-context semi-structured tables. These tasks are strategically aligned with 3 core challenges we identified in processing semi-structured tables, ensuring comprehensive coverage of the key competencies required for real-world applications.

3.2.1 Structural Complexity Challenges

Processing semi-structured tables requires models to correctly parse logical structure amid format-specific noise (e.g. HTML tags (Sui et al., 2023), LaTeX separators (Jaitly et al., 2023), Markdown delimiters (Fang et al., 2024)) while preserving semantic relationships between headers, rows, and columns. This challenge tests a model's ability to maintain consistent structural understanding across different representation formats.

Exact Matching (EM). This task evaluates precise information location and extraction capabilities while maintaining structural awareness. Models must navigate table hierarchies to find exact matches across multiple attributes, demonstrating their ability to understand both data positioning and inter-column relationships. Similar to the recent "needle-in-a-table" task (Wang et al., 2025), which requires locating information precisely in large and complex tabular data.

Basic Conditional Filtering (BCF). This task assesses the application of logical conditions across tabular data. Models must execute filtering operations based on numerical or categorical conditions while preserving structural awareness, such as finding transactions that satisfy multiple criteria across different columns. Similar to the fundamental tasks commonly found in Text-to-SQL benchmarks (such as Spider (Yu et al., 2018) and BIRD (Li et al., 2023b)), except that here the model is not required to generate SQL but instead directly answers questions in natural language.

3.2.2 Long-Range Dependencies Challenges

Real-world tables often present relevant information sparsely distributed across thousands of tokens (Sauber-Cole and Khoshgoftaar, 2022). This challenge tests a model's ability to maintain attention and synthesize information across extensive contexts, particularly crucial in multi-turn dialogues and multi-table scenarios where information must be gathered from various sources.

Fuzzy Conditional Manipulation (FCM). This task examines complex operations like sorting, aggregation, and fuzzy matching across extensive contexts (Biswal et al., 2024; Chang et al., 2023). Models must maintain coherence while synthesizing information from multiple table sections, demonstrating their ability to handle operations that span long token distances.

Fact Retrieval (FR). This task tests fact verification and retrieval capabilities across lengthy contexts. Models must determine whether claimed facts are supported by table data, requiring sustained attention and accurate relationship verification across extensive sequences. Our task setting differs from traditional table-based fact checking (Thorne et al., 2018; Chen et al., 2019) in that, for fact-supported statements, the model is required to explicitly output the supporting evidence.

3.2.3 Semantic Integration Challenges

The final challenge evaluates models' ability to understand domain-specific semantics and contextual expressions within tabular data. This tests end-to-end semantic understanding, including the interpretation of specialized terminology and conventional expressions within their specific contexts.

External Knowledge Fusion (EKF). This task assesses the integration of external knowledge with tabular information. Models must combine structural understanding with domain knowledge to answer questions requiring both factual recall and inferential reasoning (Li et al., 2023b).

Irregular Numerical Interpretation (INI). Some studies have found that LLMs exhibit certain vulnerabilities when handling mathematical problems; performing operations that involve numerical fragility can lead to performance degradation (Yu et al., 2025). We aim to investigate whether the way data is represented in the TableQA domain affects LLM performance, especially when dealing with large tables. This task evaluates the processing of non-standard numerical representations. Models must handle diverse number formats while maintaining semantic understanding and calculation accuracy, demonstrating their ability to work with mixed numerical expressions in context.

3.3 Data Collection

Our benchmark construction adopts an *expansion-reduction* methodology as shown as Figure 1: we first generate a large pool of candidate examples through multiple pathways and then apply rigorous filtering to ensure data quality. This section details the expansion phase—how we construct diverse table contexts and question-answer pairs—while the verification process is discussed in Section 3.4. Task-specific data generation logic is further elaborated in Appendix A.

3.3.1 Table Context Length Control

To evaluate LLMs under varying table context lengths, we build upon existing multi-domain Text-to-SQL datasets—BIRD (Li et al., 2023b), Spider (Yu et al., 2018), and Dr.Spider (Chang et al., 2023)—as seed resources. These datasets are preferred over typical Wikipedia-based TableQA corpora due to their larger relational databases and the availability of executable SQL queries, which facilitate symbolic verification.

We serialize each table into 7 text formats—Markdown, HTML, JSON, LaTeX, SQL, XML, and CSV—to assess LLM robustness against format variations. All formats[1] are programmatically derived from structured `pandas.DataFrame` objects using native APIs of Pandas (pandas development team, 2020) (e.g., `to_markdown`, `to_html`, `to_json`) or custom converters (e.g., for LaTeX, SQL). These formats reflect diverse real-world use cases, such as documentation, web display, data analysis scripts, or direct SQL interfaces (see Appendix B for format examples).

To construct tables at different length tiers (8K, 32K, 128K tokens), we compute the token count of the serialized tables across all formats using the *tiktoken* tokenizer [2], and choose the maximum as the effective length. For single-table cases, we iteratively sample rows until the desired token threshold is reached. For multi-table settings, we adapt the topological sampling approach from (Qiu et al., 2024): we first determine a topological order of tables based on their foreign key graph to ensure that dependent tables are sampled after their referenced parents. Then, during row sampling, only rows that preserve referential consistency are included. This pipeline ensures relational validity and supports scalable length-controlled table generation. The resulting benchmark contains 4,512 tables at 8K, 992 at 32K, and 1,296 at 128K token lengths.

3.3.2 Task-specific QA Pairs Collection

We construct question-answer pairs using a three-pronged strategy: transformation, synthesis, and generation.

Question-based Transformation. We leverage non-training splits of BIRD, Spider, and Dr.Spider to obtain SQL queries and natural questions. Each query is executed to retrieve answers, and GPT-4o (Hurst et al., 2024) is used to verify and categorize the questions into our benchmark's six task types. For quality control, we filter for queries returning 0–10 rows, and for Dr.Spider we exclude SQLs containing complex clauses such as `GROUP BY`, `HAVING`, or nested `AVG` operations. Answers are reformatted into lists or labels as required, and queries returning no result are reframed into unsupported factual assertions for fact verification tasks.

Template-based Synthesis. For Exact Matching (EM), Basic Conditional Filtering (BCF), and Fact Retrieval (FR) tasks in single-table single-turn settings, we use controlled synthesis, following data construction and generation strategies similar to those used in TQA-Bench (Qiu et al., 2024). We define template-based answer generation scripts in Python that select meaningful column combinations (e.g., low-frequency rows for EM, inequality filters for BCF, unique facts for FR), then prompt GPT-4o to generate corresponding questions. This approach ensures both task alignment and answer verifiability. For Irregular Numerical Interpretation (INI), we introduce numerical perturbations: 70% of numeric cells are randomly transformed into alternative formats such as Roman numerals,

[1] These formats can also be obtained via the API provided by `https://tableconvert.com/`.

[2] `https://github.com/openai/tiktoken`

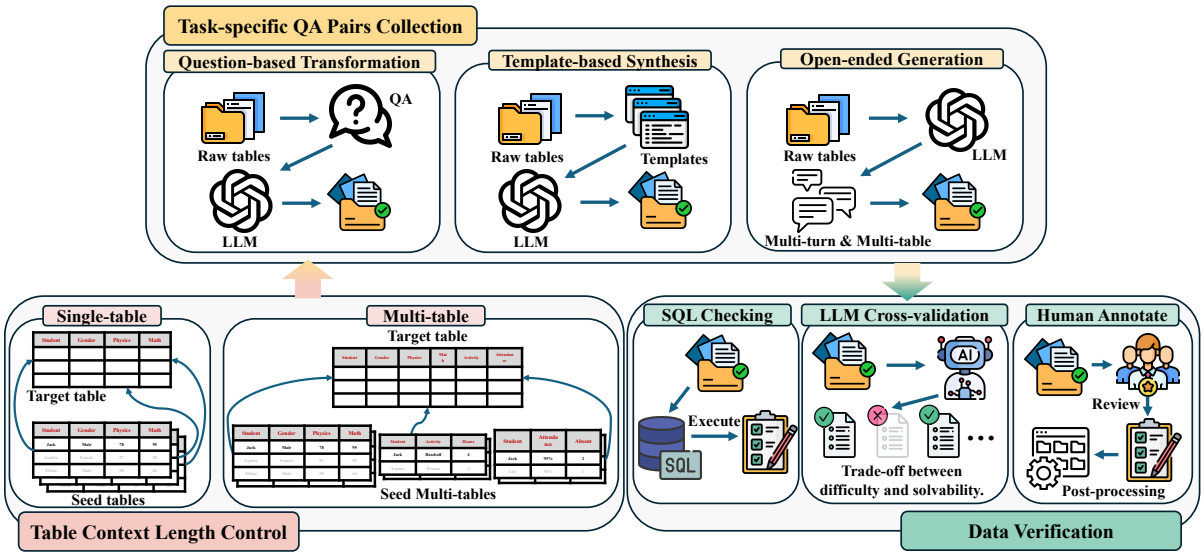

Figure 1: Overview of the LongTableBench construction pipeline. Our method follows an expansion–reduction paradigm: (i) diverse table contexts are first created through single-table and multi-table expansion strategies, and task-specific QA pairs are generated via question-based transformation, template-based synthesis, and open-ended LLM generation; (ii) data quality is subsequently ensured through a multi-stage verification process, including SQL execution checking, LLM-based cross-validation, and human annotation. This design balances task diversity with reliability, enabling high-quality benchmark construction.

English words, or scientific notation, while the remaining 30% are preserved to provide natural variance. In addition, we apply diverse formatting schemes to date values at similar proportions, which are randomly rewritten into daily common date formats such as `YYYY-MM-DD`, `DD/MM/YYYY`, or `MM.DD.YYYY`.

Open-ended Generation. To complement coverage, especially for multi-turn or multi-table scenarios where transformation or synthesis is insufficient, we directly prompt GPT-4o with full table contexts to generate plausible questions and corresponding answers. In some cases, we include seed QA examples as few-shot prompts to guide generation. This route enables high-complexity examples and conversational interactions beyond the scope of traditional datasets.

3.4 Data Verification

To ensure the highest standards of data quality and answer consistency, we implement a comprehensive verification framework that combines symbolic checks, LLM-based cross-validation, and human oversight.

SQL-based Consistency Checking. For transformation-based data derived from Text-to-SQL datasets, we treat the original SQL queries as symbolic ground truth. Whenever a natural question is rewritten or reformatted, we maintain a parallel mapping between the modified question and the original SQL. This allows us to re-execute the query and compare the actual database outputs with the expected answers. By ensuring consistency between logical execution and question semantics, we preserve the validity of transformed examples and reduce noise introduced by language variations.

LLM-based Cross-validation. For synthesis and generation-based data where no SQL supervision is available, we adopt a three-model cross-validation protocol to assess both answer correctness and difficulty level, following the data processing approach used in S1 (Muennighoff et al., 2025). Specifically, we run each QA instance through three LLMs with varying capabilities—Qwen2.5-72B-Instruct (Team, 2024), GPT-4o (Hurst et al., 2024), and DeepSeek-V3 (Liu et al., 2024)—and collect their predicted answers. F1 scores are computed for each model's prediction against the ground truth. Instances where all three models achieve high F1 scores are considered overly simple and excluded to prevent performance inflation. Conversely, instances where all models fail are removed due to potential ambiguity or annotation noise. We retain only the middle 50% of samples ranked by average F1 scores across the three models, ensuring a balance between difficulty and solvability. These selected examples then proceed to human verifica-

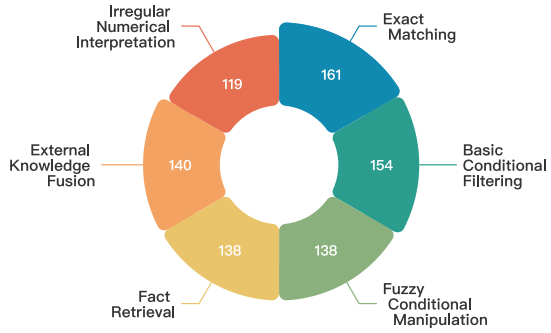

Figure 2: The number distribution of seed questions for each type of task.

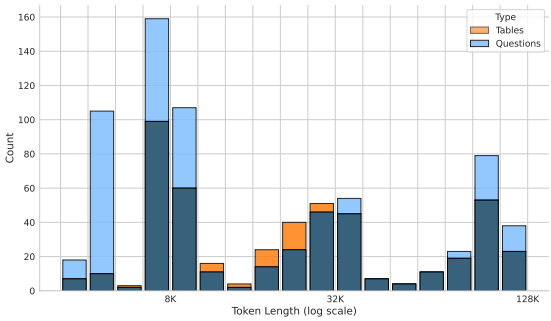

Figure 3: Token count distributions for table tokens and full input sequences across all samples.

tion.

Human Review and Post-processing. To further ensure semantic alignment and factual consistency, we perform human verification on the filtered subset of QA instances. The reviewers—comprising the authors of this paper—evaluate whether the question faithfully represents the intended reasoning logic, whether the answer is supported by the table content, and whether the language remains natural and unambiguous. No external annotators are involved. Additional post-processing steps, including deduplication, answer format normalization, and metadata cleanup, are applied before releasing the final benchmark.

This multi-stage verification pipeline enables LongTableBench to maintain both structural rigor and semantic integrity across a wide range of question types, table formats, and length conditions.

3.5 Data Statistics

LongTableBench comprises 5,950 test samples derived from 850 seed questions, with each seed question rendered into seven semi-structured table formats: Markdown, HTML, JSON, LaTeX, SQL, XML, and CSV (see Figure 2). This ensures format diversity for robust evaluation across representation schemes. Detailed format specifications are provided in Appendix B.

The dataset spans 18 domains and 54 subdomains, covering fields such as medical, education, and entertainment. Domain distribution is shown in Figure 4.

Samples are stratified by context length: 40% short (0–8K), 35% medium (8K–32K), and 25% long (32K–128K), measured using *tiktoken*. Token-level statistics, including both table content and surrounding prompts, are visualized in Figure 3. A detailed analysis of token distribution across different formats is provided in Appendix C.

Tasks are balanced across core capabilities: 35% focus on Structural Complexity (EM, BCF), 35% on Long-Range Dependencies (FCM, FR), and 30% on Contextual Semantic Integration (EKF, INC).

Manual review led to the rejection of 5% of submissions due to task misalignment, 8% for insufficient complexity, and 4% for answer inconsistency, ensuring high data quality.

Figure 4: Domain and subdomain coverage in LongTableBench, spanning 18 domains and 54 subdomains.

4 Experiments

4.1 Setup

We comprehensively evaluate 52 large language models (LLMs) across two primary categories: vanilla models and reasoning-enhanced models with built-in chain-of-thought capabilities (Wei et al., 2022) (e.g., Gemini-2.0-thinking (AI, 2025), DeepSeek R1 (Guo et al., 2025)). The evaluated models encompass a diverse ecosystem, includ-

Models	Task Category							Format Variation									Context Length		
	EM	BCF	FCM	FR	EKF	INI	Avg.	MD	HTML	JSON	SQL	XML	LaTeX	CSV	Avg.	Range	8K	32K	128K
Vanilla Models																			
Llama-3.2-3B-Instruct	24.30	14.78	13.76	2.99	11.99	9.68	13.94	14.35	13.36	11.72	16.10	13.14	12.35	15.23	13.75	31.83	21.76	11.92	8.15
Qwen2.5-3B-Instruct	17.14	7.07	9.40	33.74	36.44	18.63	20.40	21.04	20.19	22.46	24.66	21.00	20.11	19.25	21.25	25.49	27.20	18.21	15.78
Phi-3-mini-128k-instruct	28.66	9.74	8.29	33.34	19.05	17.39	20.25	20.12	21.75	22.65	20.83	20.78	19.52	16.54	20.31	30.09	25.26	21.24	14.24
Phi-4-mini-instruct	28.36	6.94	7.41	22.85	21.22	13.67	17.49	18.22	17.15	19.94	17.61	19.60	15.33	13.63	17.35	36.35	21.55	19.09	11.82
Qwen3-4B	23.15	11.07	9.88	24.59	14.80	13.30	16.71	17.69	16.77	16.94	17.51	17.00	17.32	14.64	16.84	18.10	21.70	15.66	12.77
Gemma-3-4B-it	31.60	13.40	14.29	21.98	19.08	19.97	20.56	21.98	23.24	18.92	20.87	20.30	22.29	20.74	21.19	20.37	28.67	19.95	13.07
Llama-3.1-8B-Instruct	47.73	19.09	21.29	15.11	22.41	22.97	25.69	26.94	24.68	22.07	27.51	25.29	26.56	28.83	25.98	26.04	34.66	22.66	19.74
Qwen2.5-7B-Instruct	35.85	15.18	20.58	50.78	39.11	30.50	32.05	31.93	35.30	36.89	32.43	34.74	33.72	30.55	33.65	18.82	40.84	29.35	25.95
Qwen2.5-7B-Instruct-1M	47.52	20.13	23.79	52.77	24.88	24.65	33.61	34.88	35.29	35.77	34.44	35.64	33.73	31.53	34.47	12.31	40.55	32.43	27.83
TableGPT2-7B	41.10	17.68	18.80	45.34	39.82	28.64	32.26	34.46	32.52	34.10	32.40	36.73	33.33	30.48	33.43	18.70	42.15	29.11	25.50
TableLLM-Qwen2-7B	18.00	4.59	5.34	10.51	6.69	5.94	8.98	10.55	8.63	8.18	10.95	7.52	9.52	8.87	9.17	37.39	13.46	8.76	4.73
TableLLM-Llama3.1-8B	27.72	1.87	2.38	11.65	5.15	7.78	10.31	11.12	10.81	10.53	10.76	10.39	10.50	7.24	10.19	38.06	11.88	12.14	6.90
TableLlama	3.36	1.38	1.41	2.25	1.61	0.72	1.83	3.25	0.47	0.89	2.51	0.35	2.80	1.43	1.67	173.23	4.97	0.36	0.17
Mistral-7B-Instruct-v0.3	29.32	14.16	15.59	19.89	15.46	14.85	19.19	20.56	19.67	16.81	18.97	19.06	18.56	20.69	19.19	20.27	29.15	19.47	8.95
Ministral-8B-Instruct-2410	23.98	9.37	14.82	25.61	12.70	14.25	17.26	18.72	17.48	14.53	18.26	15.38	17.13	19.68	17.31	29.73	32.90	17.33	1.55
Qwen3-8B	35.97	14.18	17.02	46.68	27.87	25.29	28.55	29.54	28.11	28.76	30.65	27.46	29.32	29.06	28.99	11.01	39.73	25.11	20.80
GLM-4-9B-Chat	46.25	18.82	18.79	32.86	21.74	20.72	27.92	27.84	29.95	27.12	28.52	31.20	29.13	27.22	28.71	14.21	34.79	27.07	21.89
GLM-4-9B-Chat-1M	46.81	17.79	17.69	31.21	21.78	20.87	27.24	28.64	27.76	27.00	29.49	27.26	30.11	27.05	28.19	11.03	32.70	26.95	22.09
GLM-4-9B-0414	31.90	17.74	20.17	35.61	28.19	21.78	26.51	28.41	27.46	23.89	27.11	24.82	28.64	30.09	27.20	22.77	44.61	23.58	11.35
Gemma-3-12B-it	46.34	21.51	27.48	46.77	42.90	29.93	36.90	39.40	37.47	34.02	39.84	34.67	39.61	37.95	37.57	15.49	49.83	34.71	26.16
Mistral-Nemo-Instruct-2407	35.32	18.02	18.30	49.10	53.69	32.40	34.99	36.29	33.47	33.73	35.36	34.82	37.20	37.19	35.44	10.52	52.13	35.00	17.85
Qwen2.5-14B-Instruct	47.26	20.17	26.62	54.34	42.86	34.93	38.25	40.98	41.64	37.09	39.23	39.93	41.77	39.32	39.99	11.71	51.75	34.58	28.43
Qwen2.5-14B-Instruct-1M	56.51	22.16	29.39	63.84	37.37	37.15	42.37	43.93	44.94	39.14	44.65	44.27	46.20	43.80	43.85	16.12	50.98	40.11	36.03
Qwen3-14B	53.18	22.13	25.89	58.49	52.69	40.16	42.62	46.66	45.36	38.78	46.74	45.11	45.76	42.50	44.42	17.93	54.51	40.54	32.81
Mistral-Small-3.1-24B-Instruct-2503	62.33	28.98	29.01	41.93	48.95	42.40	43.06	46.66	44.98	41.37	48.02	44.59	48.30	45.47	45.63	15.19	56.60	41.20	31.38
Qwen3-30B-A3B	48.73	16.54	22.67	55.53	44.91	33.66	37.74	40.10	38.85	36.40	38.45	37.86	41.95	39.78	39.05	14.21	51.40	35.31	26.50
Qwen3-32B	51.48	24.62	26.92	51.37	46.19	42.67	40.72	44.27	42.78	40.79	43.28	45.57	43.64	43.82	43.45	10.99	52.09	37.79	32.29
GLM-4-32B-0414	43.71	25.43	26.39	55.39	51.14	35.24	40.48	47.21	39.62	35.61	41.88	36.39	44.51	45.79	41.57	27.91	58.09	43.04	20.31
Llama-3.1-70B-Instruct	62.58	31.90	29.15	56.69	36.26	37.67	44.35	49.85	46.14	39.74	46.16	45.18	46.16	45.78	47.99	7.73	52.62	46.84	33.57
Llama-3.3-70B-Instruct	60.33	33.37	33.47	41.31	41.67	37.13	42.74	47.88	41.49	35.83	46.87	38.57	49.11	48.72	44.07	30.12	56.65	40.96	30.60
Qwen2.5-72B-Instruct	59.43	32.15	33.72	63.55	57.17	45.96	49.94	52.63	52.16	45.10	51.74	52.57	52.23	53.53	51.42	16.40	63.97	46.51	39.36
Mistral-Large-Instruct-2411	55.96	29.69	29.10	57.10	46.99	44.20	44.99	49.84	45.37	40.19	49.13	42.21	50.55	49.32	46.66	22.20	63.87	45.23	25.89
DeepSeek-V3	69.63	44.51	43.65	66.36	57.44	54.18	57.09	61.39	62.37	53.67	62.42	58.96	61.71	59.71	60.03	14.57	70.80	54.67	45.80
GPT-4o-mini-2024-07-18	60.43	31.73	33.17	49.55	35.26	36.10	42.76	41.78	40.85	37.53	43.19	41.04	41.46	41.42	41.04	13.79	54.44	38.61	35.22
Gemini-2.0-flash	71.19	47.51	42.16	50.05	60.76	53.39	55.03	58.66	57.69	58.14	58.95	58.20	57.33	58.44	58.20	2.79	65.13	55.25	44.72
GPT-4o-2024-08-06	78.95	52.40	49.97	72.76	68.79	62.11	65.60	67.36	64.44	58.67	65.62	60.99	65.89	66.18	64.16	13.53	78.57	63.33	54.92
Claude-3.5-sonnet-20241022	76.47	46.13	42.26	68.45	65.37	58.91	60.63	64.09	60.70	52.03	58.88	59.05	61.16	61.26	59.60	20.24	75.59	58.46	47.84
Reasoning Models																			
Phi-4-mini-reasoning	3.45	2.39	1.35	11.79	3.10	4.51	4.25	5.18	3.95	3.74	4.73	5.20	3.96	3.36	4.30	42.78	9.12	2.47	1.15
Qwen3-4B-thinking	30.85	7.06	8.62	21.96	24.78	19.87	18.92	18.26	17.75	19.37	21.35	22.89	20.31	16.95	19.55	30.38	28.56	15.49	12.70
DeepSeek-R1-Distill-Qwen-7B	7.96	4.00	4.52	13.78	13.13	6.61	8.28	8.99	5.50	5.58	7.51	5.63	8.22	7.73	7.02	49.72	17.26	5.06	2.52
DeepSeek-R1-Distill-Llama-8B	33.25	25.24	15.19	38.07	33.58	20.53	29.28	30.71	30.12	23.38	28.28	27.34	34.75	28.54	29.02	39.20	39.95	31.87	16.02
Qwen3-8B-thinking	34.20	8.61	10.86	32.18	29.97	24.04	23.32	24.94	24.82	22.87	24.51	27.03	24.23	22.29	24.38	19.43	34.00	18.83	17.13
GLM-Z1-9B-0414	42.27	33.97	20.53	44.88	46.19	30.39	37.89	42.52	31.79	38.41	41.63	33.86	41.49	40.38	38.58	27.81	59.57	39.25	14.84
Qwen3-14B-thinking	46.66	14.54	18.38	51.10	42.99	31.72	35.31	38.05	35.64	32.39	37.88	38.37	36.80	35.64	36.39	16.43	47.94	30.40	27.60
Qwen3-30B-A3B-thinking	40.10	11.91	16.28	37.86	35.36	27.16	28.63	29.81	29.40	29.25	30.86	31.47	28.94	27.94	29.67	11.88	42.79	24.05	19.04
QwQ-32B	59.44	40.07	28.73	59.33	58.68	45.93	50.31	54.12	54.04	47.39	53.57	48.58	53.99	52.80	52.07	12.91	65.43	49.72	35.77
DeepSeek-R1-Distill-Qwen-32B	50.73	39.06	25.43	52.48	57.16	44.55	45.60	52.13	46.52	41.33	49.58	44.12	49.71	50.45	47.69	22.66	66.13	47.76	22.91
Qwen3-32B-thinking	48.29	14.86	17.70	48.90	41.80	34.08	34.94	38.95	36.09	30.94	36.56	38.58	35.94	36.84	36.27	22.08	48.09	30.15	26.57
GLM-Z1-32B-0414	47.70	37.50	25.13	39.45	57.02	34.67	41.81	43.61	38.26	44.90	42.23	39.84	45.69	42.71	42.46	17.49	58.74	41.11	25.59
DeepSeek-R1-Distill-Llama-70B	56.02	42.10	27.57	51.87	52.61	44.58	47.17	51.20	46.90	43.42	53.20	43.82	52.08	51.07	48.81	20.03	62.88	49.64	28.98
DeepSeek-R1	71.24	66.21	42.41	62.68	65.28	57.87	62.82	67.92	67.84	61.26	66.55	64.21	65.81	65.29	65.55	10.15	74.63	64.37	49.45
Gemini-2.0-flash-thinking-exp-01-21	69.95	54.74	39.76	51.48	64.26	54.04	56.98	58.61	61.44	61.25	60.69	59.59	59.47	60.13	60.17	4.70	70.16	56.05	44.74

Table 1: Representative results on **LongTableBench**, aggregated across three evaluation dimensions: task categories, format variations, and context lengths. All scores (except *Avg.* and *Range*) are F1, where higher is better. *Avg.* is the mean score over the corresponding metrics. *Range* is the normalized format score range, defined as $(\max_f s_f - \min_f s_f)/\text{Avg}_f(s_f)$, where s_f is the F1 score under format f; lower values indicate stronger robustness across formats. Complete results for all 52 models are provided in Appendix G.

ing commercial closed-source models accessed via API (Gemini (AI, 2025), GPT (Hurst et al., 2024), and Claude series (Anthropic, 2024)), table-specialized models (TableGPT2 (Zha et al., 2023), TableLlama (Zhang et al., 2023), and TableLLM (Zhang et al., 2024b) variants with 7-8B parameters), and general-purpose instruction-following models of various sizes. For closed-source commercial models, precise parameter counts remain undisclosed by their providers. We prioritized instruction-following models for our evaluation since LongTableBench focuses on real-world open-domain question answering scenarios that require strong instruction adherence.

All models in our evaluation support context windows of at least 128,000 tokens, with certain models (such as those in the Gemini series) handling up to one million tokens. For our assessment methodology, we employ a strictly zero-shot setting without demonstration examples. When input sequences exceed a model's context window capacity, we apply middle truncation following established protocols from prior work. To ensure experimental reproducibility, we utilize greedy decoding with temperature fixed at 0 during response generation and evaluate performance consistently across all tasks using the F1 metric, which balances precision and recall in answer evaluation.

4.2 Results on LongTableBench

The complete experimental records are documented in Appendix G, while Table 1 presents a consolidated view of some models, synthesizing our comprehensive evaluation results across three critical dimensions: task categories, format variations, and context lengths. Several key findings emerge from our analysis:

Task-wise Performance. Our benchmark effectively differentiates model capabilities across distinct table understanding dimensions: in Structural Complexity tasks (EM, BCF), table-specialized models demonstrate advantages over general-purpose counterparts of similar size, e.g., TableGPT2-7B yields a 3.88% average gain over its base model Qwen2.5-7B-Instruct; Long-Range Dependencies tasks (FCM, FR) establish clear correlation between model size and performance reliability, with smaller variants showing progressive degradation at longer distances; most notably, Semantic Integration tasks (EKF, INI) emerge as the strongest differentiator between standard and reasoning-focused models, where GPT-4o achieves 68.79% on EKF tasks, dramatically outperforming comparable general models like Llama-3.3-70B-Instruct (41.67%), confirming that robust semantic reasoning capabilities are essential for complex table understanding beyond simple pattern matching.

Format Impact. Table formats significantly influence model performance, revealing important insights about model adaptability: most models consistently achieve optimal results with Markdown and HTML formats (GPT-4o: 67.36% with Markdown vs. 58.67% with JSON), likely reflecting these formats' prevalence in pretraining corpora and their inherent structural clarity; notably, performance disparities (i.e. range) across formats diminish as model size increases (DeepSeek-V3 shows only 14.57% variance across formats while Phi-4-mini-instruct shows 36.35%), indicating larger models develop more format-agnostic understanding capabilities; surprisingly, table-specialized models show unexpectedly strong format preferences rather than improved format robustness, suggesting general reasoning ability may be more important than format-specific optimization for real-world table understanding.

Length-wise Analysis. Context length emerges as a critical performance determinant, revealing fundamental architectural limitations: in short contexts (8K), most models maintain robust performance (top performers: 75-80% F1); medium contexts (32K) introduce moderate challenges with 10-15% performance reduction, particularly in reasoning-intensive tasks; the most significant challenges emerge at long contexts (128K), where only the largest models maintain F1 scores above 45% (DeepSeek-V3: 45.80%, GPT-4o: 54.92%), while smaller models collapse dramatically; performance at 128K scales almost linearly with model size (3-4B: 12-13% F1; 7-8B: 20-27%; 14-32B: 30-40%; 70B+: 45-55%); notably, reasoning-focused models demonstrate superior long-context performance (DeepSeek-R1: 49.45% at 128K), suggesting enhanced reasoning capabilities help maintain coherent understanding across extended contexts—a frontier challenge requiring significant advances in attention mechanisms and reasoning capabilities.

4.3 Effects of Context Compression

Models	Vanilla RAG	TableRAG	Text-to-SQL	End-to-End QA
Qwen2.5-7B-Instruct	20.54	26.88	37.42	42.59
TableGPT2-7B	24.73	30.06	41.78	42.17
Llama-3.3-70B-Instruct	29.36	31.56	30.24	54.70

Table 2: F1 scores comparing compression techniques vs. end-to-end QA across three models.

We investigated how context compression techniques impact model performance by comparing three approaches (Vanilla RAG, TableRAG (Chen et al., 2024), and Text-to-SQL) against end-to-end QA across three representative models: Qwen2.5-7B-Instruct, TableGPT2-7B, and Llama-3.3-70B-Instruct. All experiments used 100 randomly sampled questions from the 32K context length tier to ensure fair comparison.

The results in Table 2 show that end-to-end LLM approaches outperform compression techniques in some certain aspects, with performance gaps ranging from 5.2% to 23.1%. While TableRAG improves upon vanilla RAG, it still struggles with irregular table structures and semantic integration tasks. These findings highlight that current compression methods remain insufficient for real-world table understanding, where structural irregularities often exceed the capabilities of traditional parsing-based approaches.

4.4 Context Memorization Analysis

Given that some table content originates from real-world scenarios, we investigate whether models rely on pre-training memorization rather than active comprehension. Following Bai et al. (2023b,

Models	Average
Qwen2.5-7B-Instruct	42.59
w/o context	11.21
TableGPT2-7B	42.17
w/o context	8.21
Llama-3.3-70B-Instruct	54.7
w/o context	7.22

Table 3: Model performance with vs. without table context, showing minimal memorization effects.

2024b), we evaluate model performance on questions without their accompanying table contexts. We evaluated performance on 100 randomly sampled questions in two conditions: with complete table contexts and with empty tables. This approach forces models to rely solely on parametric knowledge in the second condition, revealing any memorization effects across different task categories.

Our results in Table 3 demonstrate that LLM performance collapses without table context, dropping from 42.59-54.70% to 7.22-11.21% F1 scores (an average reduction of 43.5 percentage points). Task-specific analysis revealed minimal memorization effects across all categories, with only slight advantages (2-3% above random) for EM and EKF tasks. These findings confirm that LongTableBench effectively evaluates active reasoning capabilities rather than the recall of memorized patterns from pretraining data.

5 Conclusion

We present **LongTableBench**, a comprehensive benchmark for evaluating long-context reasoning over semi-structured tables across diverse formats, domains, and tasks. By combining realistic table construction, multi-turn and multi-table settings, and extensive evaluations across 52 LLMs, LongTableBench uncovers key limitations in current models' structural, temporal, and semantic understanding. Our findings highlight the need for robust, format-agnostic reasoning capabilities and suggest future research directions in long-context tabular comprehension. We hope this benchmark serves as a valuable resource for the community and drives progress toward more capable and generalizable LLMs.

6 Limitations

Our benchmark has several limitations worth noting. While LongTableBench covers diverse domains, it cannot capture all specialized table formats found in industry-specific applications. Our evaluation relies primarily on F1 scores, which may not fully capture nuanced aspects of reasoning like explanation quality or uncertainty handling. The benchmark's focus on English-language content limits its utility for evaluating multilingual table reasoning capabilities. Additionally, as LLM capabilities advance rapidly, some benchmark aspects may become less challenging over time, potentially requiring periodic updates to maintain discriminative power. Despite these limitations, LongTableBench provides a valuable foundation for systematically evaluating and advancing long-context table reasoning research.

7 Ethical Considerations

LongTableBench is built entirely from publicly available or synthetically generated data, containing no personally identifiable or sensitive information. All verification was conducted directly by the authors rather than external annotators, combining automated checks with manual review to ensure quality and appropriateness. To mitigate environmental impact, we provide scalable context lengths so that experiments can be tailored to available resources. While the benchmark is intended to advance research in long-context table reasoning, its application to sensitive domains (e.g., healthcare, finance) should be accompanied by domain-specific safeguards. In preparing this work, large language models were only used for limited purposes such as polishing the manuscript text and rephrasing certain dataset questions. All code, data, and evaluation protocols are openly released to promote transparency, reproducibility, and responsible use.

8 Acknowledgments

This paper is supported by the National Regional Innovationand Development Joint Fund (No. U24A20254). Haobo Wang is also supported by the Fundamental Research Funds for the Zhejiang Provincial Universities (No. 226-2025-00065). Additionally, this work receives support from MYbank, Ant Group.

References

Google AI. 2025. Gemini 2.0 flash thinking (experimental). https://ai.google.dev/gemini-api/docs/thinking. Accessed: 2025-05-20.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and

Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

Anthropic. 2024. Claude 3.5 sonnet.

Anthropic. 2025. Claude 3.7 sonnet and claude code.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024a. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2023b. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Asim Biswal, Liana Patel, Siddarth Jha, Amog Kamsetty, Shu Liu, Joseph E Gonzalez, Carlos Guestrin, and Matei Zaharia. 2024. Text2sql is not enough: Unifying ai and databases with tag. *arXiv preprint arXiv:2408.14717*.

Egor Bogomolov, Aleksandra Eliseeva, Timur Galimzyanov, Evgeniy Glukhov, Anton Shapkin, Maria Tigina, Yaroslav Golubev, Alexander Kovrigin, Arie van Deursen, Maliheh Izadi, and 1 others. 2024. Long code arena: a set of benchmarks for long-context code models. *arXiv preprint arXiv:2406.11612*.

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. In *The Eleventh International Conference on Learning Representations*.

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

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. *arXiv preprint arXiv:1909.02164*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. 2020. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2021. Hitab: A hierarchical table dataset for question answering and natural language generation. *arXiv preprint arXiv:2108.06712*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*.

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

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. *arXiv preprint arXiv:2308.15363*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in

llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

YiQiu Guo, Yuchen Yang, Ya Zhang, Yu Wang, and Yanfeng Wang. 2024. Dictllm: Harnessing key-value data structures with large language models for enhanced medical diagnostics. *arXiv preprint arXiv:2402.11481*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, and 1 others. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*.

Sukriti Jaitly, Tanay Shah, Ashish Shugani, and Razik Singh Grewal. 2023. Towards better serialization of tabular data for few-shot classification. *arXiv preprint arXiv:2312.12464*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kezhi Kong, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Chuan Lei, Christos Faloutsos, Huzefa Rangwala, and George Karypis. 2024. Opentab: Advancing large language models as open-domain table reasoners. *arXiv preprint arXiv:2402.14361*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023a. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Ma Chenhao, Guoliang Li, Kevin Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023b. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. In *Advances in Neural Information Processing Systems*, volume 36, pages 42330–42357. Curran Associates, Inc.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tianyang Liu, Canwen Xu, and Julian McAuley. 2023. Repobench: Benchmarking repository-level code auto-completion systems. *arXiv preprint arXiv:2306.03091*.

Qingyang Mao, Qi Liu, Zhi Li, Mingyue Cheng, Zheng Zhang, and Rui Li. 2024. Potable: Programming standardly on table-based reasoning like a human analyst. *arXiv preprint arXiv:2412.04272*.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Accessed: 2025-04-21.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and 1 others. 2021. Quality: Question answering with long input texts, yes! *arXiv preprint arXiv:2112.08608*.

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

Mohammadreza Pourreza and Davood Rafiei. 2024a. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *Advances in Neural Information Processing Systems*, 36.

Mohammadreza Pourreza and Davood Rafiei. 2024b. Dts-sql: Decomposed text-to-sql with small large language models. *arXiv preprint arXiv:2402.01117*.

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

Rick Sauber-Cole and Taghi M Khoshgoftaar. 2022. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, 9(1):98.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. 2024. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in llms. *arXiv preprint arXiv:2310.10358*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2023. Evaluating and enhancing structural understanding capabilities of large language models on tables via input designs. *arXiv preprint arXiv:2305.13062*.

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*, pages 645–654.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025a. Qwen2.5-1m: Deploy your own qwen with context length up to 1m tokens.

Qwen Team. 2025b. Qwen3.

Qwen Team. 2025c. Qwq-32b: Embracing the power of reinforcement learning.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Lanrui Wang, Mingyu Zheng, Hongyin Tang, Zheng Lin, Yanan Cao, Jingang Wang, Xunliang Cai, and Weiping Wang. 2025. Needleinatable: Exploring long-context capability of large language models towards long-structured tables. *arXiv preprint arXiv:2504.06560*.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, and 1 others. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*.

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.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, and 1 others. 2025. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. 2025. Longproc: Benchmarking long-context language models on long procedural generation. *arXiv preprint arXiv:2501.05414*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

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.

Tong Yu, Yongcheng Jing, Xikun Zhang, Wentao Jiang, Wenjie Wu, Yingjie Wang, Wenbin Hu, Bo Du, and Dacheng Tao. 2025. Benchmarking reasoning robustness in large language models. *arXiv preprint arXiv:2503.04550*.

Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, and 1 others. 2023. Tablegpt: Towards unifying tables, nature language and commands into one gpt. *arXiv preprint arXiv:2307.08674*.

Chao Zhang, Yuren Mao, Yijiang Fan, Yu Mi, Yunjun Gao, Lu Chen, Dongfang Lou, and Jinshu Lin. 2024a. Finsql: Model-agnostic llms-based text-to-sql framework for financial analysis. *arXiv preprint arXiv:2401.10506*.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2023. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*.

Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, and 1 others. 2024b. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024c. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2024. Docbench: A benchmark for evaluating llm-based document reading systems. *arXiv preprint arXiv:2407.10701*.

A Task-specific QA Construction Details

This section details the construction strategies used for each of the six task categories in LongTableBench. For clarity, we group them by their data construction source—transformation, synthesis, or generation.

I. Exact Matching (EM)

Source: Template-based synthesis.
Construction Method: Select a column with low row-level duplication (frequency ≤ 5), extract its distinct values as answer candidates, and use the corresponding primary key as the lookup target. Prompt GPT-4o to generate a question that explicitly queries the target value.
Task Description: Evaluates precise information retrieval and structural comprehension. Requires locating exact matches across attributes, emphasizing awareness of table hierarchies and inter-column relationships.
Example Questions: 1. What is the unique id identifying the ruling dated 2018-10-05 with the description "When calculating a spell's total cost, include..."? 2. What is the raceId of the object from the year 1980 and round 4?

II. Basic Conditional Filtering (BCF)

Source: Template-based synthesis.
Construction Method: Randomly select a numeric or categorical column and define a comparison condition (e.g., greater than, less than, equal to). Identify rows satisfying the condition and extract their primary key values as the answer. GPT-4o generates a filtering-based question.
Task Description: Assesses the execution of logical conditions over tabular data. Requires filtering based on numerical or categorical criteria while maintaining structural awareness, such as identifying records that meet multi-column constraints.
Example Questions: 1. How many customers who paid in euro have a monthly consumption of over 1000? 2. What are the SalesIDs of transactions where the trading quantity is greater than or equal to 987?

III. Fuzzy Conditional Manipulation (FCM)

Source: Open-ended generation.
Construction Method: Given the complexity of FCM questions (e.g., fuzzy matching, sorting, aggregation), GPT-4o is prompted with full table context and few-shot examples to generate natural language questions involving fuzzy operations. Answers are returned as ranked lists or partial matches.
Task Description: Evaluates the ability to perform complex operations such as sorting, aggregation, and fuzzy matching over large contexts. Requires synthesizing information across distant table regions, testing the model's capacity for long-range reasoning and structural consistency.
Example Questions: 1. Which schools located in counties with names starting with 'S' have a 'Charter' status of 1? 2. List the names of schools that are charter schools and have a free meal count (K-12) greater than 40.

IV. Fact Retrieval (FR)

Source: Mixed.
Construction Method: Two construction paths are used. In the transformation-based path, SQL queries that return 0 rows are reframed as unsupported claims, while those returning exactly 1 row are treated as supported claims; GPT-4o then reformulates these into natural language questions. In the synthesis-based path, ground-truth rows are first defined, and both supporting and contradicting facts are manually constructed based on them; GPT-4o subsequently generates the corresponding claims.
Task Description: Evaluates fact verification and retrieval over long contexts. Requires determining the validity of claims based on table data, with emphasis on sustained attention and accurate verification of relationships across extended sequences.
Example Questions: Introducing the Sicilian Deep Dish Pizza, a delightful creation brought to you by The Florida Tomato Committee. This recipe, identified uniquely by the recipe ID 711, promises a rich and flavorful experience. To prepare this dish, you will need approximately 30 minutes of preparation time. The directions are straightforward: Preheat your oven to 350 degrees Fahrenheit. Use tomatoes generously to enhance the flavor profile of this deep dish pizza. Although the introduction and subtitle for this recipe are not provided, the detailed directions ensure that you can recreate this delicious pizza with ease. The recipe does not specify the yield unit, allowing you to adjust the quantity based on your needs. Once the pizza is prepared and cooked, it requires no additional standing time, making it a quick and convenient option for any meal. Enjoy the authentic taste of Sicilian Deep Dish Pizza, a testament to the quality and expertise

of The Florida Tomato Committee. Please determine whether the given table supports the above statement. Output the full row corresponding to the fact using a Python dict if it exists; otherwise, output 'Unsupported'.

V. External Knowledge Fusion (EKF)

Source: Open-ended generation.
Construction Method: Prompt GPT-4o to incorporate external knowledge (e.g., country capitals, unit conversions) when interpreting cell content. Questions are designed to require reasoning beyond direct table lookup, simulating real-world semantic integration.
Task Description: Assesses the integration of external knowledge with tabular data. Requires combining structural understanding with domain-specific facts to support both factual recall and inferential reasoning.
Example Questions: 1. preferred_foot = 'left';. List the IDs of players whose preferred foot is left. 2. adults refer to users where Age BETWEEN 19 and 65;. Please list the display names of users who are adults.

VI. Irregular Numerical Interpretation (INI)

Source: Augmented synthesis.
Construction Method: Introduce controlled perturbations to numeric columns in synthesized EM/BCF/FR examples—70% are converted into alternative formats such as Roman numerals ("III"), verbal expressions ("one hundred"), or scientific notation ("1.1×10^2"). Date formats are also varied (e.g., "YYY-MM-DD", "MM/DD/YYYY", "DD.MM.YYYY").
Task Description: Evaluates the ability to interpret non-standard numerical and date formats. Requires accurate semantic understanding and reasoning over heterogeneous numeric expressions within tabular contexts.
Example Questions: 1. What are the create dates of the votes that were for the contestant named 'Tabatha Gehling'? 2. What is the part number of the object with a size of 9 and manufactured by Manufacturer#1?

B Table Format Examples

To evaluate model robustness under diverse table serialization schemes, each table is rendered into seven formats using programmatic conversion. Below we show examples of the same table rendered in three representative formats.

Figure 5 displays an example table. The results of converting this example table into other formats are: Markdown Format (Figure 6), HTML Format (Figure 8), JSON Format (Figure 7), LaTeX Format (Figure 11), SQL Format (Figure 10), XML Format (Figure 12), and CSV Format (Figure 9).

Original DataFrame

```
  country     population
0 Germany         83200000
1 Japan          125800000
2 Canada          38900000
```

Figure 5: Original DataFrame Example.

Markdown Format

```
| country | population |
|---------|------------|
| Germany | 83200000   |
| Japan   | 125800000  |
| Canada  | 38900000   |
```

Figure 6: Markdown Format Example.

JSON Format

```
{
  "data": [
    {
      "index": 0,
      "country": "Germany",
      "population": 83200000
    },
    {
      "index": 1,
      "country": "Japan",
      "population": 125800000
    },
    {
      "index": 2,
      "country": "Canada",
      "population": 38900000
    }
  ]
}
```

Figure 7: JSON Format Example.

C Token Distribution Analysis Across Formats

The token counts for tables and questions discussed in the main text represent the maximum

HTML Format

```
<table>
  <thead>
   <tr><th>country</th><th>population</th>
</tr>
  </thead>
  <tbody>
   <tr><td>Germany</td><td>83200000</td></tr>
   <tr><td>Japan</td><td>125800000</td></tr>
   <tr><td>Canada</td><td>38900000</td></tr>
  </tbody>
</table>
```

Figure 8: HTML Format Example.

CSV Format

```
country,population
Germany,83200000
Japan,125800000
Canada,38900000
```

Figure 9: CSV Format Example.

SQL Format

```
CREATE TABLE test (country, population);
INSERT INTO test (country, population) VALUES
('Germany', 83200000),
('Japan', 125800000),
('Canada', 38900000);
```

Figure 10: SQL Format Example.

LaTeX Format

```
\begin{tabular}{ll}
    \hline
    country & population \\
    \hline
    Germany & 83200000 \\
    Japan & 125800000 \\
    Canada & 38900000 \\
    \hline
\end{tabular}
```

Figure 11: LaTeX Format Example.

values across all formats. In this section, we provide a detailed analysis of token distributions across the seven semi-structured formats used in LONGTABLEBENCH: Markdown, HTML, JSON, LaTeX, SQL, XML, and CSV. Unlike the aggregated histograms presented in the main text, here we present token distribution density plots separated by format to better illustrate the subtle differences in tokenization behavior and length characteristics across representation schemes.

Figure 13 depicts the kernel density estimates (KDE) of token counts calculated specifically over the question portions for each format. Each curve corresponds to a format, revealing how question length varies depending on the encoding style and formatting conventions.

Similarly, Figure 14 shows the KDE curves of token counts derived from the table content itself, again separated by format. This visualization highlights differences in token density related to how tabular data is represented, including markup verbosity and syntactic overhead.

From the density plots in Figures 13 and 14, it is evident that the token counts vary notably across formats. In particular, HTML, JSON, and XML consistently exhibit higher token densities, reflecting their relatively verbose markup syntax. Among these, XML shows the largest token counts overall. Conversely, the remaining formats—Markdown, LaTeX, SQL, and especially CSV—tend to have lower token counts, with CSV being the most compact representation. These differences highlight the impact of format choice on input length and, potentially, model efficiency and performance.

D Prompt Examples

Since LLMs participate in many tasks, the total number of prompts used is very large. However, considering that many prompts share similarities, we list some representative prompt templates, including evaluation prompts (Figure 15), prompts for question-based transformation (Figure 16), prompts for template-based synthesis (Figure 17), and prompts for open-ended generation (Figure 18).

It is important to note that prompts may vary across different models. For instance, some models require prompts to be specifically designed to activate their reasoning mode, necessitating corresponding adjustments to the prompt formulation.

E Dataset Construction and License

Our benchmark is derived from modifications to three existing datasets: **BIRD**, **Spider**, and **Dr.Spider**, all licensed under the **Creative Commons Attribution 4.0 International License (CC-BY-4.0)**. To ensure compliance and promote open research, we adopt the same **CC-BY-4.0** license for

XML Format

```
<?xml version="1.0" ?>
<root>
  <row>
    <country>Germany</country>
    <population>83200000</population>
  </row>
  <row>
    <country>Japan</country>
    <population>125800000</population>
  </row>
  <row>
    <country>Canada</country>
    <population>38900000</population>
  </row>
</root>
```

Figure 12: XML Format Example.

our derived benchmark.

Attribution Requirements. As mandated by CC-BY-4.0, users must:

- Cite the original datasets: **BIRD** (Li et al., 2023b), **Spider** (Yu et al., 2018), and **Dr.Spider** (Chang et al., 2023) in any public use or derivatives.
- Acknowledge their work as the source of the modified benchmark.
- Include a link to CC-BY-4.0 in distributions.

Recommended Attribution. For proper attribution, use this wording in documentation or publications:

> "This work uses data derived from **BIRD** (Li et al., 2023b), **Spider** (Yu et al., 2018), and **Dr.Spider** (Chang et al., 2023) (CC-BY-4.0). Our modified benchmark is also CC-BY-4.0 licensed."

Modifications. We document all changes (e.g., schema adjustments, question rewrites, error fixes) in our repository for reproducibility.

License Rationale. CC-BY-4.0 ensures:

- Compatibility with the original datasets' terms.
- Community reuse while protecting authorship.
- Flexibility: Users may further modify our benchmark if they preserve the attribution chain.

F Models Used and Citation Information

The models utilized in our experiments include the GPT series (Hurst et al., 2024), Gemini series (AI, 2025), Claude series (Anthropic, 2024), DeepSeek series (Liu et al., 2024; Guo et al., 2025), LLama3 series (Grattafiori et al., 2024), Qwen2.5 series (Team, 2024), Qwen2.5-1M series (Team, 2025a), QwQ-32B (Team, 2025c), Qwen3 series (Team, 2025b), phi series (Xu et al., 2025), GLM series (GLM et al., 2024), gemma3 series (Team et al., 2025), and mistral series (Jiang et al., 2023). Due to the large number of models employed, we provide citation information only for representative models rather than individual citations for each model.

G Complete Experimental Results

We provide the full benchmark results evaluated on 52 large language models (LLMs), covering both vanilla models and reasoning models with built-in chain-of-thought (CoT) capabilities. Table 4 reports results under the **8K** length setting, Table 5 corresponds to the **32K** setting, and Table 6 shows performance under the **128K** setting. These comprehensive evaluations form the foundation for the aggregated views presented in above sections.

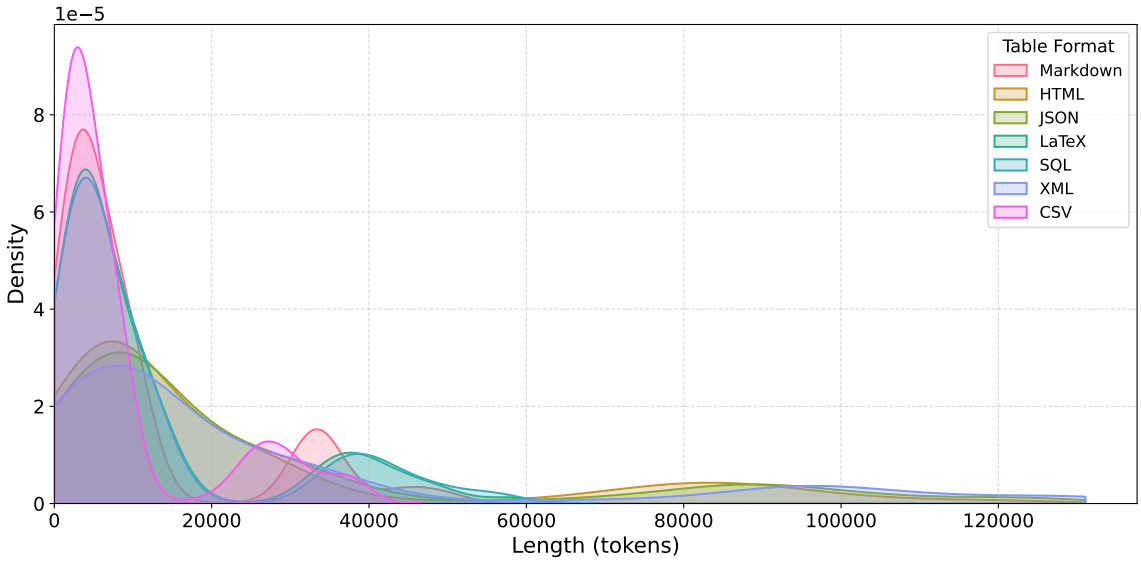

Figure 13: Kernel density estimates of question token counts across the seven formats. Each line corresponds to one format.

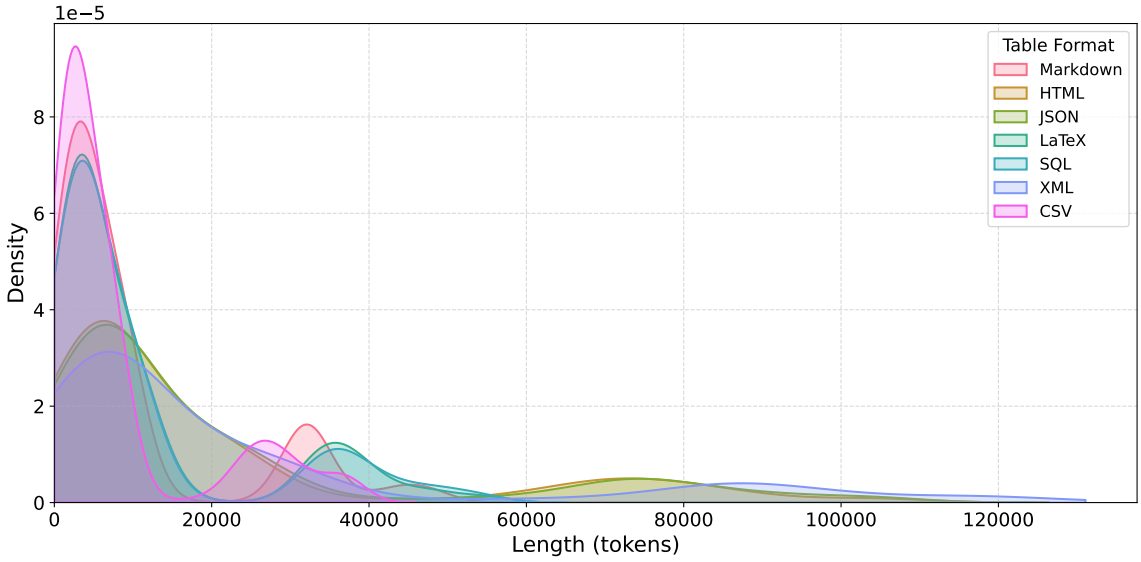

Figure 14: Kernel density estimates of table token counts across the seven formats. Each line corresponds to one format.

Evaluation Prompts

Requirements:
Please read the following table and then answer the questions based on the table.
Organize your answers into a list of strings, with each element being an answer item. If the question includes a special requirement, such as outputting a dictionary, please fulfill that specific request. Otherwise, always output a list of strings, even if there is only one answer item.
Please place your answers between the ```and ```.

Notes:
The table may include non-standard formats for numbers or dates.
For numbers, formats may include Roman numerals, English words, or scientific notation. Whenever possible, please convert these into Arabic numerals in your response.
For dates, the following formats may be encountered:
%Y-%m-%d <other time elements> (in Arabic numerals)
%d/%m/%Y <other time elements> (in Arabic numerals)
%m.%d.%Y <other time elements> (in Arabic numerals)
%Y.%m.%d <other time elements> (in English words for numbers)
%Y-%m-%d <other time elements> (in Roman numerals)
Please convert all dates to the format %Y-%m-%d <other time elements> (in Arabic numerals) in your response.

Additionally, if there are duplicate answer_items, they should be output repeatedly (i.e., no deduplication).

Example Output Format:
list[str]:
```
['answer_item_1', 'answer_item_2', ..., 'answer_item_n']
```

dict:
```
{'column_1': value_1, 'column_2': value_2, ..., 'column_n': value_n}
```

Please do not generate any text after outputting the final answer.

Table information:
{table_infos}

Question: {query}
Answer:

Figure 15: Evaluation Prompts.

Prompts for Question-based Transformation

CONTEXT
The following is a text-to-SQL dataset. We have the table information, the question, and the SQL execution result for this data. Our task is to rewrite the question into a TableQA format. The rewriting rules are:

1.The answer to a TableQA question typically consists of a single column. If the returned values contain multiple columns and include a column with unique values, that column is usually chosen. Otherwise, the most relevant column is selected based on context. For example, the question "Please find the ID, name, and age of people with a monthly income of 5000" should be rewritten as "Please find the ID of people with a monthly income of 5000."
2.Based on the first step, rewrite the question without changing its meaning.
3.According to the SQL execution result, rewrite the answer as a Python list of strings. For example, given the result:
ID Name Age
1 003 Tom 25
2 004 John 30
3 009 Mike 33
If we select only the 'ID', the answer would be rewritten as ['003', '004', '009'].

# Input #
{
"question":"Original question",
"result":"The result of executing the SQL statement"
}

# RESPONSE #
{
"new_question":"Questions that are modified and rewritten according to the rules",
"new_answer":['<Answer 1>','<Answer 2>',...,'<Answer n>']"
}

Answer Required
* Please strictly follow the format above for your response, i.e., generate a dict when modification or creation is possible. This dict should contain two key-value pairs.
* Please generate only one dict.
* Do not generate any additional text.

Input
{input_str}

RESPONSE

Figure 16: Prompts for Template-based Synthesis.

Prompts for Question-based Transformation

CONTEXT # I will now provide you with a template. Please rewrite it according to the specified requirements.{specific_task_requirements}

Please note: {specific_task_considerations}

Answer Required
* Do not generate any additional text.

{filled_template}

RESPONSE

Figure 17: Prompts for Question-based Transformation.

Prompts for Open-ended Generation

CONTEXT
Please generate a question and corresponding answer based on the information provided in the given table. {specific_task_requirements}
1.If only one table is provided, the question should be answerable using just that table.
2.If multiple tables are provided, the question must require information from at least two tables along with the given knowledge to answer.

The question should be specific enough that the answer can be clearly derived from the tables. The answer should be organized in the form of a Python list, with the length of the list being greater than 0 and less than 6. Please generate a multi-round dialogue where each round's question builds upon the previous one, adding additional conditions or details in each subsequent round.

Note: You can generate questions that require exact matching or numerical comparisons, and you may include calculations involving several columns of values. However, you should not generate questions that involve statistical calculations, such as those requiring the computation of averages or variances.

# Input #
{
"table_infos":"Tabular information"
}

# RESPONSE #
{
"highlighted_table":['<table 1>','<table 2>',...],#The core table where the answer is located.
"QA":[
{
round: 1,
"question":"question 1",
"answer":<Answer 1>,
},
...
,
{
round: n,
"question":"Quesrion in round n",
"answer": <Answer n>
}...#n>1
]#list[dict]
}
Where <Answer i> represents either a Python list[str], 0 < len(<Answer n>) < 6.

Answer Required
* Please strictly follow the format above for your response, i.e., generate a dict when modification or creation is possible. This dict should contain two key-value pairs.
* Please generate only one dict.
* The number of answer terms should be between 1 and 5.
* Do not generate any additional text.
* Do not generate a judgment question or an open-ended question

Input
{input_str}

RESPONSE

Figure 18: Prompts for Open-ended Generation.

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
Vanilla Models								
LLaMA-3.2-3B-Instruct	Markdown	41.73	20.61	27.35	2.00	24.38	13.60	21.61
	HTML	45.91	19.28	29.12	3.02	21.11	13.99	22.07
	JSON	40.49	20.03	27.13	2.04	20.20	13.84	20.62
	LaTex	38.77	21.92	29.95	4.99	21.13	9.68	21.07
	SQL	45.10	22.43	32.09	9.97	22.96	17.37	24.99
	XML	44.15	21.08	34.48	12.11	25.23	13.22	25.04
	CSV	40.88	18.47	25.76	12.16	21.26	13.24	21.96
Qwen2.5-3B-Instruct	Markdown	26.89	12.09	11.37	46.43	45.22	25.76	27.96
	HTML	25.92	13.54	14.05	42.20	48.74	19.79	27.37
	JSON	27.50	15.86	27.12	34.74	50.13	22.64	29.67
	LaTex	25.28	12.47	14.64	41.30	51.84	21.31	27.81
	SQL	40.02	15.74	15.98	42.67	58.35	28.54	33.55
	XML	23.68	13.93	22.69	47.14	50.43	23.63	30.25
	CSV	21.24	11.93	10.07	45.47	53.28	18.61	26.77
Phi-3-mini-128k-instruct	Markdown	35.64	6.22	8.58	38.22	20.91	16.70	21.05
	HTML	47.14	21.32	25.77	50.15	21.91	23.51	31.63
	JSON	49.03	22.04	32.74	42.66	27.42	23.85	32.96
	LaTex	36.75	7.56	10.54	39.76	24.34	19.51	23.08
	SQL	29.55	9.45	11.50	41.34	21.49	21.09	22.40
	XML	48.34	16.04	31.12	52.30	32.10	25.46	34.23
	CSV	24.42	2.92	4.15	19.58	17.13	10.81	13.17
Phi-4-mini-instruct	Markdown	24.12	3.12	2.55	21.92	24.54	10.27	14.42
	HTML	37.11	11.80	14.25	50.45	36.96	25.97	29.42
	JSON	44.87	15.40	25.42	42.80	40.98	28.30	32.96
	LaTex	20.03	3.97	3.44	18.90	17.85	12.34	12.76
	SQL	24.26	4.51	8.42	22.87	26.51	14.89	16.91
	XML	42.06	11.31	21.31	54.38	43.64	23.52	32.70
	CSV	19.53	2.64	5.33	26.34	15.69	2.52	12.01
Qwen3-4B	Markdown	37.96	13.68	17.39	34.23	17.00	18.03	23.05
	HTML	34.88	9.22	22.17	28.54	18.83	16.93	21.76
	JSON	28.32	9.36	16.29	25.83	21.90	20.39	20.35
	LaTex	30.68	10.28	14.77	33.87	22.11	13.77	20.91
	SQL	37.73	11.80	14.02	38.13	22.85	20.48	24.17
	XML	39.62	9.65	26.22	30.05	26.96	24.46	26.16
	CSV	29.92	12.37	12.62	35.09	15.95	13.93	19.98
Gemma-3-4B-it	Markdown	50.65	25.35	20.57	30.71	25.91	16.96	28.36
	HTML	47.19	24.26	28.15	33.94	33.41	17.40	30.72
	JSON	44.57	20.43	37.90	23.78	27.79	23.01	29.58
	LaTex	51.90	23.21	23.28	29.85	30.31	16.90	29.24
	SQL	48.01	24.80	27.81	26.79	27.71	18.08	28.86
	XML	51.63	24.78	36.30	28.72	33.58	21.44	32.74
	CSV	46.82	23.59	25.69	23.53	25.73	14.36	26.62
Llama-3.1-8B-Instruct	Markdown	68.92	27.03	42.08	8.33	33.22	22.69	33.71
	HTML	70.61	29.49	38.80	11.46	30.48	23.52	34.06
	JSON	52.95	23.19	44.68	20.15	32.20	21.75	32.49
	LaTex	70.69	30.12	45.67	7.82	32.14	18.08	34.09
	SQL	69.98	28.05	38.41	20.75	32.50	25.10	35.80
	XML	70.13	32.88	48.13	11.98	34.15	25.29	37.09
	CSV	65.16	24.54	43.01	28.07	34.00	23.22	36.33
Qwen2.5-7B-Instruct	Markdown	57.02	24.27	32.63	58.78	49.78	33.06	42.59
	HTML	52.04	26.73	35.27	54.88	55.94	36.08	43.49
	JSON	53.59	21.84	49.03	55.53	67.85	40.66	48.08
	LaTeX	56.05	25.86	26.17	55.68	54.29	36.13	42.36
	SQL	52.23	24.59	34.06	57.59	46.89	32.13	41.25
	XML	49.30	25.54	35.67	55.89	64.50	40.26	45.19
	CSV	52.73	22.09	18.87	58.93	49.38	30.05	38.67
Qwen2.5-7B-Instruct-1M	Markdown	62.02	24.03	28.61	65.65	36.52	31.72	41.42
	HTML	64.61	31.00	35.26	59.22	45.99	34.53	45.10
	JSON	64.82	27.18	44.57	55.97	42.70	36.30	45.26
	LaTeX	60.87	27.63	35.68	55.11	36.71	32.87	41.48
	SQL	66.16	26.46	23.85	61.71	37.15	29.17	40.75
	XML	63.90	31.21	46.31	59.04	44.75	36.26	46.91

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	57.98	25.88	22.81	61.56	32.12	30.28	38.44
TableGPT2-7B	Markdown	64.14	23.43	28.26	52.29	51.87	33.04	42.17
	HTML	56.85	25.53	38.78	48.99	52.67	33.65	42.75
	JSON	61.74	28.05	55.64	52.74	59.63	42.47	50.04
	LaTeX	62.72	23.52	30.21	52.42	60.90	33.61	43.90
	SQL	64.26	23.37	29.80	47.44	54.61	35.48	42.49
	XML	62.62	27.28	46.81	56.46	68.29	39.59	50.17
	CSV	57.34	24.06	24.24	50.62	50.63	31.87	39.79
TableLLM-Qwen2-7B	Markdown	24.48	10.95	14.72	17.71	9.42	6.20	13.92
	HTML	36.16	9.83	7.65	15.86	11.43	8.89	14.97
	JSON	32.87	7.25	10.87	9.56	9.28	5.28	12.52
	LaTeX	28.68	8.45	12.00	18.56	10.85	7.00	14.26
	SQL	31.36	13.32	11.08	14.84	9.23	9.07	14.82
	XML	24.78	10.20	9.82	17.30	10.90	7.29	13.38
	CSV	23.50	6.39	11.27	15.52	8.81	6.37	11.98
TableLLM-Llama3.1-8B	Markdown	33.56	0.68	5.74	17.38	10.43	5.91	12.28
	HTML	36.27	0.68	7.07	14.54	10.43	5.27	12.38
	JSON	38.96	2.38	5.97	16.96	9.91	5.88	13.34
	LaTeX	33.26	0.68	3.63	13.27	7.36	3.41	10.27
	SQL	34.45	0.68	4.49	17.13	9.06	3.91	11.62
	XML	38.07	1.36	0.94	17.18	10.47	4.27	12.05
	CSV	25.28	0.68	2.30	8.59	6.17	3.91	7.82
TableLlama	Markdown	19.61	4.89	6.12	14.07	9.10	4.63	9.74
	HTML	1.28	2.05	1.03	0.60	2.10	0.66	1.28
	JSON	3.11	5.05	2.01	2.04	2.79	1.08	2.68
	LaTeX	20.01	5.13	5.27	9.78	6.88	3.24	8.39
	SQL	14.81	6.19	3.39	10.80	6.34	3.62	7.52
	XML	1.91	0.31	0.00	1.36	2.38	0.40	1.06
	CSV	8.16	2.66	3.17	8.12	2.86	0.72	4.28
Mistral-7B-Instruct-v0.3	Markdown	50.93	23.03	30.56	28.39	19.34	20.40	28.77
	HTML	48.90	23.65	30.31	35.32	23.40	22.83	30.73
	JSON	48.94	20.79	42.43	24.48	23.32	27.75	31.29
	LaTeX	45.96	22.94	29.66	20.37	20.84	20.85	26.77
	SQL	47.36	23.15	38.67	27.91	22.98	23.14	30.53
	XML	52.21	23.99	43.34	28.81	25.71	25.12	33.20
	CSV	46.87	23.36	31.11	32.47	24.21	18.02	29.34
Ministral-8B-Instruct-2410	Markdown	51.73	18.87	46.25	57.93	25.37	29.01	38.20
	HTML	53.78	21.12	35.13	48.20	26.73	26.76	35.29
	JSON	40.19	17.67	39.92	38.99	25.19	27.90	31.64
	LaTeX	53.81	13.21	38.24	53.57	23.46	26.81	34.85
	SQL	51.09	20.82	38.37	46.10	24.17	28.93	34.92
	XML	46.85	16.98	42.39	40.66	30.64	25.16	33.78
	CSV	47.53	22.67	46.84	42.56	25.24	26.03	35.15
Qwen3-8B	Markdown	60.45	21.31	29.30	58.04	41.26	36.54	41.15
	HTML	52.16	19.03	35.31	58.09	47.79	30.31	40.45
	JSON	53.28	21.17	48.60	52.95	41.66	30.06	41.29
	LaTeX	58.00	19.86	39.74	58.32	51.64	27.80	42.56
	SQL	63.67	21.92	23.15	64.93	52.05	29.55	42.54
	XML	62.27	19.43	42.50	55.61	50.71	36.82	44.56
	CSV	57.61	22.09	30.70	68.82	39.85	26.37	40.91
GLM-4-9B-Chat	Markdown	62.07	20.96	26.25	46.56	27.79	22.73	34.39
	HTML	64.82	20.63	46.00	48.54	32.10	23.77	39.31
	JSON	61.49	21.13	50.72	23.93	31.32	21.79	35.06
	LaTeX	61.80	19.58	37.14	44.34	29.78	21.39	35.67
	SQL	63.79	22.92	34.34	47.51	30.01	24.12	37.11
	XML	65.17	22.08	57.24	45.20	33.38	26.09	41.53
	CSV	58.94	18.53	27.16	43.84	27.86	21.47	32.97
GLM-4-9B-Chat-1M	Markdown	60.04	18.32	27.44	51.04	27.35	24.20	34.73
	HTML	63.39	19.52	41.47	40.60	32.02	20.11	36.19
	JSON	58.41	17.45	52.87	21.29	33.93	22.34	34.38
	LaTeX	63.32	21.25	24.36	43.99	32.06	25.79	35.13
	SQL	64.23	21.78	28.33	44.01	29.20	18.46	34.34
	XML	57.38	18.79	33.34	45.26	29.18	15.64	33.27

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	57.31	16.15	26.47	41.25	23.56	20.72	30.91
GLM-4-9B-0414	Markdown	56.45	31.21	42.12	60.23	51.69	34.68	46.06
	HTML	57.22	29.27	50.42	74.06	57.78	38.63	51.23
	JSON	55.85	30.30	50.34	50.92	53.09	34.98	45.92
	LaTeX	56.78	30.96	46.94	63.61	49.70	34.43	47.07
	SQL	56.82	33.80	35.85	58.77	48.97	32.61	44.47
	XML	61.58	34.67	52.91	69.82	63.79	36.38	53.19
	CSV	56.82	30.52	45.93	59.65	47.23	37.95	46.35
Gemma-3-12B-it	Markdown	73.04	32.08	62.34	54.53	55.55	38.14	52.61
	HTML	73.70	33.94	53.81	58.45	67.08	39.64	54.43
	JSON	67.02	31.13	50.58	48.25	57.74	39.41	49.02
	LaTeX	69.05	36.75	58.23	58.56	61.45	41.16	54.20
	SQL	74.41	30.24	44.20	65.31	62.09	38.43	52.45
	XML	71.87	33.81	47.19	57.45	65.03	42.93	53.05
	CSV	69.53	30.37	61.29	58.08	53.28	36.83	51.56
Mistral-Nemo-Instruct-2407	Markdown	69.34	29.39	46.68	61.82	75.47	46.63	54.89
	HTML	74.96	30.11	42.15	61.86	76.05	48.68	55.64
	JSON	73.80	25.32	50.44	56.93	76.29	43.62	54.40
	LaTeX	69.86	31.16	43.65	62.93	76.10	44.77	54.75
	SQL	73.16	28.26	45.85	61.49	74.10	45.89	54.79
	XML	72.97	30.40	53.49	63.63	76.72	47.27	57.41
	CSV	70.70	26.91	41.57	59.96	72.71	44.91	52.79
Qwen2.5-14B-Instruct	Markdown	68.94	27.17	49.95	71.16	63.14	45.94	54.38
	HTML	70.03	32.19	60.90	71.75	65.93	40.71	56.92
	JSON	65.43	24.92	57.43	47.63	67.88	37.82	50.19
	LaTeX	70.53	31.36	51.16	77.38	68.85	43.97	57.21
	SQL	68.28	33.98	43.12	72.16	61.48	41.53	53.43
	XML	67.74	32.07	61.48	65.62	72.25	42.41	56.93
	CSV	71.35	28.82	56.12	78.36	66.77	40.95	57.06
Qwen2.5-14B-Instruct-1M	Markdown	70.53	32.56	50.81	82.24	44.68	45.84	54.44
	HTML	72.96	27.60	52.76	83.04	57.93	43.91	56.37
	JSON	69.53	21.16	51.71	54.03	55.82	40.47	48.79
	LaTeX	68.66	31.67	56.29	84.26	41.94	36.50	53.22
	SQL	74.29	29.94	41.59	78.16	42.64	36.47	50.52
	XML	76.35	33.98	62.62	76.37	63.41	49.71	60.41
	CSV	66.77	26.08	42.43	83.19	52.24	35.34	51.01
Qwen3-14B	Markdown	75.66	32.38	54.33	78.06	76.76	41.19	59.73
	HTML	74.86	29.60	45.46	80.91	77.30	42.07	58.37
	JSON	69.25	23.71	46.54	51.95	79.37	36.42	51.21
	LaTeX	78.27	28.93	51.65	78.47	77.41	41.94	59.45
	SQL	75.74	32.28	46.00	76.10	79.01	41.82	58.49
	XML	75.06	29.72	48.38	72.53	75.14	44.76	57.60
	CSV	74.10	27.82	53.49	75.99	69.91	41.10	57.07
Mistral-Small-3.1-24B-Instruct-2503	Markdown	82.02	41.28	57.05	62.94	66.16	43.26	58.78
	HTML	85.63	43.48	54.71	64.87	64.81	44.93	59.74
	JSON	85.07	37.44	55.81	31.93	74.63	42.96	54.64
	LaTeX	83.40	42.10	57.67	60.65	68.96	47.33	60.02
	SQL	83.41	43.12	55.79	61.38	78.75	47.63	61.68
	XML	87.15	42.60	64.36	60.58	75.25	52.14	63.68
	CSV	83.88	39.40	53.80	58.67	64.87	43.71	57.39
Qwen3-30B-A3B	Markdown	76.16	31.64	54.16	75.95	63.82	43.13	57.48
	HTML	73.39	33.18	52.35	73.37	52.69	38.34	53.89
	JSON	71.38	23.28	45.39	57.53	61.49	41.82	50.15
	LaTeX	75.30	32.34	50.52	73.94	66.20	40.45	56.46
	SQL	76.98	33.43	47.13	75.83	55.40	34.94	53.95
	XML	76.91	34.64	48.49	70.74	62.96	43.41	56.19
	CSV	76.66	28.24	43.83	71.66	65.10	40.07	54.26
Qwen3-32B	Markdown	72.94	34.95	50.61	68.81	67.85	43.36	56.42
	HTML	73.21	31.51	43.83	68.31	62.05	37.12	52.67
	JSON	73.90	29.67	45.46	40.96	65.60	41.81	49.57
	LaTeX	74.90	36.31	48.56	70.37	67.48	43.49	56.85
	SQL	75.83	32.11	41.25	71.80	69.13	43.14	55.54
	XML	81.08	33.88	42.70	68.46	73.18	48.27	57.93

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	79.12	34.50	49.57	69.79	53.59	40.45	54.50
GLM-4-32B-0414	Markdown	76.20	37.66	53.55	90.64	81.38	52.81	65.37
	HTML	76.30	36.92	59.58	82.08	80.54	48.35	63.96
	JSON	68.77	30.92	60.91	55.08	79.19	44.27	56.52
	LaTeX	75.38	37.57	56.42	81.67	78.62	49.82	63.25
	SQL	75.32	34.84	44.48	74.98	80.10	45.44	59.19
	XML	77.65	35.41	52.02	72.98	81.72	50.81	61.77
	CSV	69.29	36.21	57.69	85.30	72.39	52.50	62.23
Llama-3.1-70B-Instruct	Markdown	85.69	41.17	42.40	75.80	43.67	39.47	54.70
	HTML	85.39	40.50	47.41	82.64	50.61	40.46	57.83
	JSON	80.05	35.53	49.59	45.89	46.62	34.04	48.62
	LaTeX	84.03	43.23	41.45	77.69	43.95	39.97	55.05
	SQL	84.87	41.34	54.29	72.33	46.24	39.44	56.42
	XML	87.83	43.30	60.91	73.09	56.84	42.98	60.82
	CSV	81.50	39.35	40.85	76.52	46.00	35.11	53.22
Llama-3.3-70B-Instruct	Markdown	83.50	50.31	47.85	75.61	66.25	43.40	61.15
	HTML	83.14	48.13	51.11	74.03	64.34	39.12	59.98
	JSON	80.27	40.64	49.54	46.52	61.54	40.24	53.12
	LaTeX	82.44	46.53	55.29	72.91	66.38	43.21	61.13
	SQL	83.05	49.94	55.72	76.70	67.02	47.02	63.24
	XML	86.82	47.52	71.73	48.28	61.83	39.66	59.31
	CSV	83.70	44.04	42.17	68.22	66.19	43.82	58.02
Qwen2.5-72B-Instruct	Markdown	81.63	46.36	72.98	83.46	78.88	52.30	69.27
	HTML	81.69	45.60	67.48	75.54	74.93	50.23	65.91
	JSON	79.09	38.39	67.30	52.25	78.77	49.87	60.94
	LaTeX	84.85	47.59	66.97	86.44	78.99	56.20	70.18
	SQL	84.51	47.14	67.64	81.47	84.86	57.11	70.45
	XML	85.67	47.76	70.19	77.77	79.82	56.33	69.59
	CSV	82.61	41.69	67.36	80.58	80.58	52.48	67.55
Mistral-Large-Instruct-2411	Markdown	85.92	46.55	65.01	87.74	85.93	53.86	70.84
	HTML	86.43	46.35	64.03	84.12	80.17	51.07	68.70
	JSON	86.71	40.44	63.48	54.77	79.90	51.37	62.78
	LaTeX	87.48	49.92	67.23	88.41	74.10	56.04	70.53
	SQL	84.83	45.94	57.89	85.35	81.50	61.01	69.42
	XML	84.97	44.95	55.26	81.25	84.70	55.09	67.70
	CSV	86.46	42.60	63.42	85.74	65.75	46.33	65.05
DeepSeek-V3	Markdown	91.49	56.05	74.37	84.88	79.65	60.05	74.42
	HTML	92.64	61.25	73.13	89.30	90.09	63.16	78.26
	JSON	88.53	48.87	71.44	55.68	84.25	54.19	67.16
	LaTeX	92.08	61.29	77.39	90.03	86.96	64.16	78.65
	SQL	92.08	57.31	77.08	85.98	90.60	61.17	77.37
	XML	92.99	58.33	76.96	80.23	85.15	62.81	76.08
	CSV	91.63	57.63	71.28	89.40	80.92	60.12	75.16
GPT-4o-mini-2024-07-18	Markdown	75.46	37.36	50.76	74.02	42.31	36.79	52.78
	HTML	77.03	36.64	58.55	67.21	44.88	37.31	53.60
	JSON	77.82	34.81	53.64	44.20	51.03	33.94	49.24
	LaTex	75.78	37.98	61.93	67.25	46.34	39.73	54.84
	SQL	77.82	40.55	56.01	72.90	43.20	38.59	54.84
	XML	81.90	38.19	47.51	70.62	48.70	41.77	54.78
	CSV	76.96	37.50	47.50	77.86	40.93	33.93	52.45
Gemini-2.0-flash	Markdown	90.51	58.94	73.40	51.61	86.06	56.99	69.58
	HTML	90.89	57.71	73.12	51.76	82.55	55.12	68.53
	JSON	91.06	59.10	71.00	51.88	85.54	54.49	68.85
	LaTeX	90.68	62.30	74.75	53.40	82.66	58.98	70.46
	SQL	89.85	62.01	71.81	55.76	78.21	55.26	68.82
	XML	90.80	63.13	64.04	52.89	86.66	56.70	69.04
	CSV	87.25	59.09	72.49	49.59	77.79	53.35	66.59
GPT-4o-2024-08-06	Markdown	94.83	61.20	75.47	87.64	91.16	69.22	79.92
	HTML	89.80	65.31	75.58	85.99	86.80	66.00	78.25
	JSON	92.69	58.95	72.57	55.34	86.43	56.88	70.48
	LaTex	93.16	61.75	73.91	87.39	90.00	69.63	79.31
	SQL	93.13	65.42	73.77	84.68	91.49	66.29	79.13
	XML	95.29	64.45	73.17	81.29	87.67	67.58	78.24

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	92.82	63.29	73.76	88.51	90.76	67.98	79.52
Claude-3.5-sonnet-20241022	Markdown	89.51	61.66	67.42	86.98	87.73	62.91	76.03
	HTML	90.31	66.91	67.37	85.56	79.03	66.12	75.88
	JSON	88.63	60.53	66.15	54.38	76.71	59.81	67.70
	LaTex	91.92	67.21	65.39	87.62	86.15	66.74	77.50
	SQL	93.56	64.55	64.22	81.44	85.18	67.02	75.99
	XML	93.02	64.79	67.45	80.61	84.37	60.05	75.05
	CSV	87.73	66.96	67.38	87.74	82.90	64.19	76.15
Reasoning Models								
Phi-4-mini-reasoning	Markdown	13.28	13.03	3.28	27.26	6.89	7.81	11.93
	HTML	8.61	3.34	1.34	26.79	10.66	8.79	9.92
	JSON	8.70	5.48	5.55	15.79	7.95	8.00	8.58
	LaTeX	10.53	5.30	3.93	22.84	10.59	8.18	10.23
	SQL	12.49	9.84	4.13	23.09	8.00	5.63	10.53
	XML	14.54	6.68	5.97	22.84	4.76	9.36	10.69
	CSV	5.81	5.29	1.29	14.70	4.34	5.91	6.22
Qwen3-4B-thinking	Markdown	51.72	6.92	19.96	24.53	41.25	20.46	27.47
	HTML	48.61	6.36	18.91	31.85	43.66	22.02	28.57
	JSON	44.08	9.22	27.99	28.79	51.14	28.70	31.65
	LaTeX	48.70	10.43	20.93	26.20	36.92	20.83	27.34
	SQL	51.61	8.98	21.26	33.21	50.45	20.95	31.08
	XML	47.72	7.06	28.96	35.50	55.94	27.00	33.70
	CSV	43.04	6.23	19.39	19.24	30.94	17.35	22.70
DeepSeek-R1-Distill-Qwen-7B	Markdown	27.01	9.74	15.45	23.75	31.51	20.67	21.36
	HTML	18.65	9.20	12.06	20.85	19.20	7.54	14.58
	JSON	16.77	10.49	9.49	14.47	20.31	10.18	13.62
	LaTeX	25.00	16.15	13.10	23.77	29.73	12.26	20.00
	SQL	27.44	13.01	11.93	22.67	20.61	18.75	19.07
	XML	18.53	7.86	12.54	12.61	21.26	8.64	13.57
	CSV	27.83	8.07	9.31	26.54	27.84	19.18	19.79
DeepSeek-R1-Distill-Llama-8B	Markdown	59.50	39.11	26.96	50.80	47.14	27.36	41.81
	HTML	58.32	34.46	29.92	55.52	49.42	26.53	42.36
	JSON	50.14	30.89	37.54	36.90	55.50	31.16	40.35
	LaTeX	63.44	33.01	28.96	57.12	47.36	32.74	43.77
	SQL	52.64	29.71	27.66	51.12	46.95	29.69	39.63
	XML	53.87	34.85	24.15	57.19	56.71	30.81	42.93
	CSV	58.80	31.21	26.50	50.72	45.41	19.42	38.68
Qwen3-8B-thinking	Markdown	52.07	11.03	29.70	38.91	53.83	26.93	35.41
	HTML	50.88	9.25	25.46	47.50	52.26	28.33	35.61
	JSON	47.07	8.07	34.83	39.01	53.34	29.71	35.34
	LaTeX	51.42	11.43	25.47	40.26	50.64	30.21	34.90
	SQL	54.08	7.27	29.61	47.67	54.52	29.99	37.19
	XML	51.00	10.51	26.75	47.99	63.96	26.54	37.79
	CSV	51.88	7.27	20.03	36.70	41.76	23.19	30.14
GLM-Z1-9B-0414	Markdown	83.27	60.10	51.85	76.23	73.90	50.07	65.90
	HTML	77.41	51.80	48.18	73.21	75.35	47.12	62.18
	JSON	84.88	69.59	54.55	52.58	79.51	51.22	65.39
	LaTeX	83.54	59.40	45.97	67.77	68.27	52.58	62.92
	SQL	79.55	60.71	49.56	64.62	76.09	51.30	63.64
	XML	80.75	58.35	57.62	64.12	77.43	51.49	64.96
	CSV	81.18	55.89	45.83	69.44	75.57	49.93	62.97
Qwen3-14B-thinking	Markdown	68.91	15.79	40.11	69.12	65.79	43.64	50.56
	HTML	65.48	20.80	43.27	68.32	66.89	41.27	51.00
	JSON	64.50	17.21	50.63	48.13	74.97	34.70	48.35
	LaTeX	69.19	21.88	44.35	69.41	68.29	43.08	52.70
	SQL	68.38	13.09	44.82	66.50	65.55	36.88	49.20
	XML	70.81	15.08	45.29	67.42	72.55	41.17	52.05
	CSV	70.06	17.25	42.53	72.45	64.01	39.72	51.00
Qwen3-30B-A3B-thinking	Markdown	71.28	13.63	44.70	45.63	56.76	29.12	43.52
	HTML	63.17	16.68	47.42	49.67	69.16	34.00	46.68
	JSON	66.15	20.60	53.21	43.98	71.85	37.23	48.84
	LaTeX	61.03	14.49	39.21	45.92	62.66	35.49	43.13
	SQL	64.05	11.86	44.44	49.00	59.73	32.56	43.61
	XML	66.66	15.59	40.19	55.60	70.77	40.56	48.23

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	61.87	12.91	42.97	42.61	48.69	30.00	39.84
QwQ-32B	Markdown	87.48	49.31	64.82	79.15	83.59	60.47	70.80
	HTML	83.51	57.50	65.97	79.76	84.22	57.69	71.44
	JSON	82.22	65.83	63.32	50.05	83.96	54.04	66.57
	LaTeX	87.61	52.11	67.00	79.02	84.76	56.88	71.23
	SQL	85.90	53.01	56.69	75.62	85.73	62.76	69.95
	XML	84.78	56.79	55.15	69.80	83.58	56.39	67.75
	CSV	86.99	49.62	66.66	80.73	84.87	53.54	70.40
DeepSeek-R1-Distill-Qwen-32B	Markdown	85.02	63.80	55.39	75.59	78.12	60.13	69.68
	HTML	81.79	70.22	51.90	82.17	82.91	60.28	71.54
	JSON	82.78	67.05	58.12	48.64	81.13	55.63	65.56
	LaTeX	83.47	70.42	58.72	77.78	80.41	61.79	72.10
	SQL	86.96	67.81	56.98	75.42	82.98	61.15	71.88
	XML	84.09	65.79	66.77	74.78	83.05	61.98	72.74
	CSV	84.87	65.06	55.08	77.76	74.60	57.28	69.11
Qwen3-32B-thinking	Markdown	69.70	17.12	46.06	68.07	68.96	42.82	52.12
	HTML	69.69	20.62	44.56	69.48	67.22	39.24	51.80
	JSON	66.31	17.45	51.29	46.93	69.48	38.39	48.31
	LaTeX	67.93	20.62	40.10	72.10	61.61	43.72	51.01
	SQL	66.05	14.58	37.83	67.35	64.69	44.48	49.16
	XML	70.30	19.97	43.98	63.14	73.08	40.28	51.79
	CSV	70.11	21.12	40.10	66.60	60.24	38.20	49.39
GLM-Z1-32B-0414	Markdown	81.95	59.36	46.90	49.75	81.07	47.74	61.13
	HTML	76.82	53.04	42.95	47.40	76.53	50.86	57.93
	JSON	83.86	72.72	61.73	48.30	85.06	56.52	68.03
	LaTeX	84.05	61.18	49.47	47.82	84.28	52.20	63.17
	SQL	83.54	60.12	46.50	47.23	81.05	51.34	61.63
	XML	75.81	55.92	43.33	51.71	81.84	50.77	59.90
	CSV	82.20	62.62	50.70	49.16	77.26	51.37	62.22
DeepSeek-R1-Distill-Llama-70B	Markdown	81.57	59.39	54.24	76.79	72.68	55.71	66.73
	HTML	80.98	59.90	51.63	76.50	75.96	60.40	67.56
	JSON	80.32	61.37	49.87	47.04	78.12	57.85	62.43
	LaTeX	82.47	68.50	51.61	71.07	71.16	58.50	67.22
	SQL	85.41	62.61	55.61	71.04	69.93	57.18	66.97
	XML	86.90	62.55	53.58	77.14	76.71	58.28	69.19
	CSV	84.75	66.04	56.06	71.25	67.32	51.09	66.09
DeepSeek-R1	Markdown	91.88	79.87	73.72	81.88	84.97	68.38	80.12
	HTML	94.00	84.14	71.81	80.03	84.15	66.77	80.15
	JSON	94.38	81.39	68.44	54.85	85.24	65.22	74.92
	LaTeX	94.49	82.77	73.75	86.65	87.09	69.21	82.33
	SQL	91.91	81.70	66.50	85.74	85.50	67.86	79.87
	XML	92.37	82.94	75.37	78.51	90.77	68.08	81.34
	CSV	91.14	80.89	71.01	84.24	86.81	67.89	80.33
Gemini-2.0-flash-thinking-exp-01-21	Markdown	91.14	74.26	74.65	56.03	89.11	64.37	74.93
	HTML	91.52	72.96	75.47	59.89	90.01	62.81	75.44
	JSON	90.83	72.25	74.86	53.88	88.27	65.80	74.32
	LaTeX	92.33	80.03	76.35	56.42	87.93	61.55	75.77
	SQL	90.31	71.53	74.82	57.97	89.73	64.09	74.74
	XML	93.34	64.52	66.12	59.30	88.16	62.23	72.28
	CSV	91.13	77.73	75.74	61.47	81.74	64.18	75.33

Table 4: Full experimental results under the **8K length setting** across 7 formats and 6 task types. Avg is the mean score.

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
Vanilla Models								
LLaMA-3.2-3B-Instruct	Markdown	22.06	21.72	8.69	0.00	11.36	5.59	11.57
	HTML	17.71	25.71	2.19	0.00	7.89	10.76	10.71
	JSON	14.66	27.75	3.62	2.08	7.53	3.81	9.91
	LaTeX	7.70	16.15	3.52	0.00	13.33	4.93	7.60
	SQL	24.55	27.74	1.92	0.00	13.36	12.71	13.38
	XML	5.64	28.06	0.51	1.56	7.73	6.27	8.30
	CSV	22.08	25.16	4.70	0.00	13.03	12.92	12.98
Qwen2.5-3B-Instruct	Markdown	4.60	7.98	3.35	39.93	30.74	25.19	18.63
	HTML	9.40	10.87	7.60	22.92	37.90	34.03	20.45
	JSON	7.90	11.51	11.48	31.77	38.13	32.78	22.26
	LaTeX	11.30	1.13	4.44	28.47	34.65	29.45	18.24
	SQL	22.16	9.42	10.63	21.88	40.97	34.75	23.30
	XML	11.94	4.48	1.78	31.48	36.86	32.35	19.82
	CSV	5.40	2.11	9.99	29.90	31.55	18.36	16.22
Phi-3-mini-128k-instruct	Markdown	24.54	16.97	2.28	31.49	20.87	29.69	20.97
	HTML	32.93	17.49	1.92	31.25	16.48	17.12	19.53
	JSON	35.03	21.35	6.65	22.73	16.98	30.54	22.21
	LaTeX	29.10	13.06	2.14	29.86	16.72	19.59	18.41
	SQL	34.14	18.47	5.34	30.56	18.02	29.31	22.64
	XML	34.48	15.06	1.40	16.21	22.37	22.89	18.73
	CSV	29.96	15.43	4.62	37.49	16.22	27.85	21.93
Phi-4-mini-instruct	Markdown	33.14	13.16	2.94	30.73	37.58	30.67	24.70
	HTML	35.14	9.36	0.11	7.22	19.18	14.02	14.17
	JSON	37.44	18.65	2.73	9.03	17.17	8.53	15.59
	LaTeX	30.88	12.20	7.31	20.31	27.73	14.54	18.83
	SQL	45.31	10.49	1.85	20.83	28.52	28.11	22.52
	XML	38.70	9.63	1.19	18.55	18.65	15.49	17.04
	CSV	34.82	8.84	6.33	10.59	31.02	11.78	17.23
Qwen3-4B	Markdown	17.66	16.52	5.57	22.98	14.89	10.99	14.77
	HTML	31.47	13.79	4.53	14.58	17.04	11.38	15.47
	JSON	13.41	20.09	6.55	12.50	20.94	21.20	15.78
	LaTeX	22.12	16.66	4.63	19.39	22.75	5.91	15.24
	SQL	24.18	17.85	2.91	13.00	12.99	23.44	15.73
	XML	22.36	14.32	6.85	7.81	25.09	6.50	13.82
	CSV	21.89	10.64	6.13	16.04	21.83	12.81	14.89
Gemma-3-4B-it	Markdown	31.49	13.22	8.41	22.02	20.74	26.73	20.43
	HTML	30.49	11.50	11.91	28.82	16.94	25.06	20.79
	JSON	28.69	16.87	7.50	9.44	15.39	17.92	15.97
	LaTeX	33.46	16.89	9.32	21.56	21.47	27.65	21.72
	SQL	33.19	11.64	5.61	16.07	16.80	22.65	17.66
	XML	31.50	13.17	7.74	22.74	17.96	18.34	18.57
	CSV	31.34	12.23	10.88	15.49	17.41	24.05	18.57
Llama-3.1-8B-Instruct	Markdown	65.70	25.65	4.10	0.00	25.59	12.77	22.30
	HTML	30.50	28.53	11.03	14.58	14.11	12.70	18.58
	JSON	43.84	12.92	9.00	1.56	17.60	9.31	15.70
	LaTeX	54.37	22.39	11.86	3.13	22.30	7.18	20.20
	SQL	63.53	25.80	7.58	3.18	22.94	18.21	23.54
	XML	51.92	17.44	9.25	1.56	17.00	8.54	17.62
	CSV	54.35	22.21	8.84	20.31	17.56	15.64	23.15
Qwen2.5-7B-Instruct	Markdown	29.79	11.14	13.90	44.81	34.81	37.16	28.60
	HTML	30.62	12.69	17.30	60.00	38.22	45.13	33.99
	JSON	34.36	19.02	20.21	43.87	47.08	41.06	34.27
	LaTeX	33.52	13.07	12.58	53.13	26.13	43.07	30.25
	SQL	31.03	21.22	14.60	52.97	33.06	31.15	30.67
	XML	33.36	16.29	17.49	44.71	36.72	42.73	31.88
	CSV	27.65	20.97	14.95	47.43	39.34	32.20	30.42
Qwen2.5-7B-Instruct-1M	Markdown	36.88	28.70	30.81	54.00	21.03	24.24	32.61
	HTML	48.76	22.99	27.26	52.20	22.25	20.46	32.32
	JSON	47.24	23.00	31.43	48.26	28.42	19.38	32.96
	LaTeX	46.95	25.85	8.64	51.98	26.15	24.37	30.66
	SQL	53.37	22.72	23.35	51.22	26.08	20.00	32.79
	XML	52.62	26.02	21.59	54.24	27.77	22.52	34.13

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	39.23	21.24	13.42	47.50	18.94	18.10	26.40
TableGPT2-7B	Markdown	36.20	20.78	8.75	52.43	31.31	44.91	32.40
	HTML	38.00	22.72	13.51	40.45	31.52	27.30	28.92
	JSON	37.66	24.65	13.08	34.03	30.91	23.95	27.38
	LaTeX	37.33	11.80	9.00	52.43	38.61	30.91	30.01
	SQL	32.25	24.23	7.01	45.49	32.86	26.62	28.08
	XML	35.63	17.82	10.64	51.86	44.22	44.28	34.07
	CSV	30.19	14.65	7.99	43.40	33.84	41.18	28.54
TableLLM-Qwen2-7B	Markdown	28.22	4.71	5.56	5.56	3.43	11.22	9.78
	HTML	17.83	4.27	5.56	5.56	10.92	11.81	9.32
	JSON	14.53	1.80	5.56	11.81	6.29	9.17	8.19
	LaTeX	21.02	6.13	5.56	4.17	9.10	6.11	8.68
	SQL	25.48	5.66	5.56	15.28	7.28	9.58	11.47
	XML	13.52	6.23	5.56	6.94	4.82	5.83	7.15
	CSV	13.11	0.94	5.56	5.56	4.96	8.33	6.41
TableLLM-Llama3.1-8B	Markdown	42.07	4.17	0.00	13.37	5.40	10.00	12.50
	HTML	39.43	1.75	0.00	22.92	4.58	15.83	14.08
	JSON	32.42	4.17	0.00	14.06	3.92	12.50	11.18
	LaTeX	37.60	4.44	0.00	22.92	4.66	14.17	13.96
	SQL	39.43	3.33	0.00	18.92	6.13	12.50	13.39
	XML	44.51	2.22	0.00	10.59	3.68	11.67	12.11
	CSV	27.95	3.06	0.00	2.95	3.43	6.67	7.34
TableLlama	Markdown	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	HTML	0.00	0.00	0.00	0.00	0.73	0.00	0.12
	JSON	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LaTeX	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SQL	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	XML	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CSV	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mistral-7B-Instruct-v0.3	Markdown	39.87	17.30	10.58	17.77	20.89	18.25	20.78
	HTML	40.66	23.98	14.49	16.54	15.39	13.55	20.77
	JSON	22.23	14.86	7.80	4.36	18.20	7.97	12.57
	LaTeX	30.92	16.10	9.24	13.24	21.07	18.13	18.12
	SQL	29.33	17.04	10.22	11.79	21.71	15.54	17.60
	XML	27.99	14.56	10.40	27.13	17.60	16.42	19.02
	CSV	37.93	16.69	8.00	13.59	22.82	27.02	21.01
Ministral-8B-Instruct-2410	Markdown	35.43	7.46	7.86	21.99	17.24	17.72	17.95
	HTML	31.80	0.99	6.66	30.09	14.21	19.11	17.15
	JSON	14.94	16.57	4.19	13.37	9.93	10.44	11.57
	LaTeX	25.87	2.28	7.48	27.14	16.08	20.28	16.52
	SQL	38.84	21.54	6.26	14.81	17.08	17.93	19.41
	XML	10.90	4.76	4.09	31.19	9.29	13.90	12.36
	CSV	33.21	21.50	5.15	42.36	11.91	21.24	22.56
Qwen3-8B	Markdown	31.75	11.48	3.77	40.10	24.10	31.69	23.81
	HTML	25.45	18.02	3.98	41.35	18.48	27.62	22.48
	JSON	26.86	24.66	13.94	30.56	17.16	15.04	21.37
	LaTeX	27.30	17.86	9.40	44.20	22.02	31.76	25.42
	SQL	30.63	18.38	17.84	35.75	25.24	42.84	28.45
	XML	30.49	23.19	2.44	36.63	15.23	14.79	20.46
	CSV	29.82	17.73	10.33	39.10	26.08	40.88	27.32
GLM-4-9B-Chat	Markdown	51.91	18.51	4.24	34.03	22.46	24.97	26.02
	HTML	49.72	27.20	7.57	25.69	20.76	23.22	25.69
	JSON	52.42	25.61	2.53	9.72	21.41	17.80	21.58
	LaTeX	50.02	22.89	16.18	36.11	20.55	24.64	28.40
	SQL	39.88	23.67	4.31	26.39	21.35	22.24	22.97
	XML	55.70	22.54	5.45	28.47	19.40	27.53	26.52
	CSV	51.01	22.33	4.48	27.78	22.77	25.97	25.73
GLM-4-9B-Chat-1M	Markdown	51.38	19.31	10.18	33.33	26.39	28.57	28.19
	HTML	48.68	23.85	15.78	13.19	25.01	19.06	24.26
	JSON	49.14	28.93	6.14	25.69	19.33	15.07	24.05
	LaTeX	49.57	22.14	12.10	38.19	17.88	28.20	28.02
	SQL	55.94	23.27	11.77	35.59	22.41	29.52	29.75
	XML	56.11	17.65	5.73	5.59	15.75	29.62	21.74

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	46.24	24.76	13.39	36.81	20.36	40.90	30.41
GLM-4-9B-0414	Markdown	28.57	14.22	20.36	43.98	25.71	18.76	25.27
	HTML	26.47	19.51	15.15	21.88	26.78	20.65	21.74
	JSON	26.20	25.56	8.62	18.58	13.16	10.05	17.03
	LaTeX	24.95	22.47	15.59	48.15	26.47	20.03	26.28
	SQL	30.96	19.37	10.05	42.82	25.76	21.51	25.08
	XML	15.34	20.35	4.49	21.88	15.38	6.38	13.97
	CSV	29.72	21.61	9.93	41.44	38.83	43.81	30.89
Gemma-3-12B-it	Markdown	45.79	13.46	21.00	49.17	41.37	26.07	32.81
	HTML	42.34	21.10	11.49	48.44	37.05	23.72	30.69
	JSON	46.17	24.20	20.65	42.00	41.60	25.02	33.27
	LaTeX	45.81	19.57	22.33	48.67	42.83	26.27	34.25
	SQL	39.59	24.60	17.30	45.64	42.25	41.42	35.13
	XML	48.72	24.19	29.67	42.36	39.76	24.68	34.90
	CSV	45.66	22.37	10.67	53.19	40.10	23.43	32.57
Mistral-Nemo-Instruct-2407	Markdown	38.16	19.80	8.18	54.35	56.24	46.56	37.22
	HTML	20.37	10.74	10.76	54.55	34.60	41.21	28.71
	JSON	24.90	16.49	4.24	47.92	41.76	40.94	29.38
	LaTeX	31.00	31.72	11.66	63.24	57.50	49.62	40.79
	SQL	32.57	26.64	5.05	51.09	57.51	49.42	37.05
	XML	23.56	21.63	5.56	44.65	49.49	46.89	31.96
	CSV	45.51	25.62	15.47	63.97	59.75	48.09	43.07
Qwen2.5-14B-Instruct	Markdown	48.36	13.16	14.08	37.08	40.09	47.13	33.32
	HTML	46.85	29.36	14.91	58.45	38.92	51.30	39.96
	JSON	40.64	29.74	11.75	45.49	42.29	35.56	34.25
	LaTeX	49.11	23.59	14.60	48.26	35.71	48.86	36.69
	SQL	45.89	23.31	14.10	40.63	40.42	36.65	33.50
	XML	41.12	29.02	14.23	55.03	29.62	40.91	34.99
	CSV	41.61	23.88	14.46	41.03	43.23	35.17	33.23
Qwen2.5-14B-Instruct-1M	Markdown	52.86	22.80	14.65	53.95	37.93	27.95	35.02
	HTML	60.80	27.53	20.42	59.12	40.55	44.24	42.11
	JSON	60.14	23.44	21.51	43.06	33.70	23.60	34.24
	LaTeX	47.80	24.49	20.69	66.45	42.61	53.19	42.54
	SQL	57.82	27.51	22.44	60.38	31.72	56.34	42.70
	XML	57.68	26.01	20.18	57.95	31.18	26.01	36.50
	CSV	48.93	20.82	22.82	58.69	35.26	53.30	39.97
Qwen3-14B	Markdown	49.03	29.44	12.44	55.73	53.65	53.85	42.36
	HTML	49.24	27.83	12.84	54.99	52.47	56.80	42.36
	JSON	45.33	30.10	13.21	40.00	52.72	48.27	38.27
	LaTeX	48.87	28.67	17.77	58.68	50.56	55.91	43.41
	SQL	51.08	28.19	19.55	49.85	50.74	51.47	41.81
	XML	49.44	39.20	16.95	48.88	50.87	51.69	42.84
	CSV	44.50	27.37	13.38	56.08	43.72	46.92	38.66
Mistral-Small-3.1-24B-Instruct-2503	Markdown	59.45	29.66	14.65	39.86	49.01	56.20	41.47
	HTML	63.06	29.83	18.68	34.88	51.27	49.69	41.24
	JSON	60.42	27.83	15.74	35.59	50.02	48.36	39.66
	LaTeX	60.98	33.72	18.84	40.33	48.43	60.18	43.75
	SQL	63.79	32.84	15.27	39.40	50.58	51.99	42.31
	XML	65.19	33.93	20.83	46.40	46.33	42.84	42.59
	CSV	61.74	38.63	16.87	40.52	43.45	55.41	42.77
Qwen3-30B-A3B	Markdown	56.03	23.45	14.79	53.70	42.07	34.54	37.43
	HTML	48.18	7.63	9.56	51.74	41.79	45.80	34.12
	JSON	33.65	27.92	9.62	51.62	40.27	42.70	34.30
	LaTeX	48.22	16.42	12.20	52.44	38.67	48.73	36.12
	SQL	44.24	22.29	16.67	49.30	44.75	37.53	35.80
	XML	47.00	14.00	4.50	55.21	43.92	19.74	30.73
	CSV	54.18	16.76	15.41	52.95	44.30	45.75	38.23
Qwen3-32B	Markdown	40.80	22.90	20.43	29.88	42.23	46.50	33.79
	HTML	42.37	37.24	22.34	55.21	43.49	54.42	42.51
	JSON	41.61	35.89	24.83	36.41	40.82	44.44	37.34
	LaTeX	41.10	27.98	17.51	45.35	43.50	52.19	37.94
	SQL	45.97	30.14	20.11	44.83	45.57	36.89	37.25
	XML	51.93	41.17	18.78	58.25	49.43	53.56	45.52

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	45.81	26.21	16.01	44.97	40.82	48.41	37.04
GLM-4-32B-0414	Markdown	46.96	42.93	20.65	67.31	58.23	55.40	48.58
	HTML	42.39	39.19	16.72	60.34	46.66	47.20	42.08
	JSON	52.47	31.46	12.44	47.51	48.42	40.36	38.78
	LaTeX	41.51	44.33	25.85	55.04	54.80	52.23	45.63
	SQL	46.62	40.27	26.07	62.54	45.82	52.38	45.62
	XML	37.53	43.60	19.59	42.65	44.81	43.80	38.66
	CSV	50.63	35.98	18.53	68.01	57.31	55.42	47.65
Llama-3.1-70B-Instruct	Markdown	60.73	48.27	26.76	60.45	42.74	47.62	47.76
	HTML	60.53	49.39	24.68	53.99	37.76	35.56	43.65
	JSON	62.23	45.67	20.27	48.04	42.70	41.42	43.39
	LaTeX	63.67	45.27	23.36	55.86	32.38	32.59	42.19
	SQL	60.41	47.69	28.63	58.19	30.07	30.12	42.52
	XML	59.13	49.71	21.47	55.45	42.55	46.55	45.81
	CSV	58.60	48.09	20.56	64.59	33.58	38.12	43.93
Llama-3.3-70B-Instruct	Markdown	63.95	50.90	30.13	28.51	37.29	28.12	39.82
	HTML	60.98	39.47	32.02	33.15	37.41	25.57	38.10
	JSON	50.21	43.21	30.26	24.31	34.71	16.35	33.18
	LaTeX	62.97	48.41	28.18	39.16	41.47	43.51	43.95
	SQL	54.64	42.48	21.90	36.94	30.10	26.43	35.41
	XML	62.47	51.67	32.48	22.74	42.71	27.04	39.85
	CSV	61.26	52.34	33.01	52.96	34.60	32.99	44.53
Qwen2.5-72B-Instruct	Markdown	58.06	36.86	16.80	60.36	44.36	50.55	44.50
	HTML	61.50	40.15	14.95	62.38	54.73	60.97	49.11
	JSON	62.27	39.59	13.62	46.60	49.77	31.25	40.52
	LaTeX	54.63	41.06	18.63	45.17	47.74	31.60	39.80
	SQL	52.48	34.44	12.10	52.76	44.02	45.49	40.22
	XML	64.89	45.41	17.54	62.21	56.16	60.71	51.15
	CSV	54.70	34.21	20.29	63.78	41.78	57.31	45.35
Mistral-Large-Instruct-2411	Markdown	59.95	38.35	15.88	66.42	46.15	59.12	47.64
	HTML	52.31	37.82	19.70	58.99	46.11	52.59	44.59
	JSON	37.86	27.50	9.37	49.14	43.98	44.04	35.31
	LaTeX	57.51	35.43	14.31	63.86	47.25	57.87	46.04
	SQL	61.01	43.09	15.95	66.39	47.29	59.39	48.85
	XML	44.01	42.41	12.53	58.22	50.53	47.22	42.49
	CSV	62.43	40.43	11.22	70.19	44.06	59.39	47.95
DeepSeek-V3	Markdown	68.71	58.70	31.18	63.70	49.62	59.73	55.27
	HTML	70.44	60.10	40.34	59.16	56.30	60.39	57.79
	JSON	62.29	48.92	28.80	45.41	56.84	55.82	49.68
	LaTeX	66.57	58.54	30.76	63.46	54.16	58.75	55.37
	SQL	70.76	56.63	36.46	64.48	55.47	65.85	58.27
	XML	70.24	56.50	26.91	60.97	47.08	66.24	54.66
	CSV	67.81	56.13	29.03	64.03	53.45	60.07	55.09
GPT-4o-mini-2024-07-18	Markdown	51.72	35.17	23.32	37.10	28.06	32.79	34.70
	HTML	48.56	33.90	20.15	33.07	29.91	30.54	32.69
	JSON	48.80	36.63	14.28	25.80	25.41	23.86	29.13
	LaTex	50.36	30.44	14.56	40.52	26.23	35.32	32.91
	SQL	52.91	37.27	24.17	36.35	29.16	30.17	35.01
	XML	51.91	43.82	26.57	39.22	37.36	29.29	38.03
	CSV	45.81	35.68	27.57	41.77	31.48	43.48	37.63
Gemini-2.0-flash	Markdown	71.77	64.86	27.85	49.90	65.74	46.70	54.47
	HTML	72.15	65.65	36.80	48.95	53.66	60.55	56.30
	JSON	68.83	62.92	33.48	43.82	67.01	57.44	55.58
	LaTeX	69.05	64.40	31.32	48.55	59.18	55.03	54.59
	SQL	73.02	67.12	39.22	48.93	64.01	61.73	59.00
	XML	74.18	66.47	37.19	44.75	57.78	48.65	54.84
	CSV	67.57	62.95	34.22	54.22	68.01	64.39	58.56
GPT-4o-2024-08-06	Markdown	69.70	61.60	33.04	66.39	66.95	67.43	60.85
	HTML	71.09	59.26	39.04	63.42	60.91	62.16	59.31
	JSON	64.42	65.09	32.48	49.72	64.99	61.13	56.31
	LaTex	67.89	58.59	36.26	65.12	63.78	65.26	59.49
	SQL	75.54	60.60	34.08	67.61	65.59	63.94	61.23
	XML	72.39	61.08	35.47	64.31	60.61	69.90	60.63

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	73.10	61.78	39.76	68.94	66.86	62.29	62.12
Claude-3.5-sonnet-20241022	Markdown	68.78	58.98	30.16	60.56	65.66	57.10	56.87
	HTML	69.54	45.04	29.87	61.18	70.49	67.34	57.24
	JSON	70.02	39.67	34.93	40.48	72.35	51.45	51.48
	LaTex	68.23	57.95	31.50	67.15	64.89	65.79	59.25
	SQL	70.40	50.39	22.24	61.56	62.48	68.22	55.88
	XML	69.11	33.96	30.10	55.44	64.65	55.01	51.38
	CSV	69.21	58.12	31.35	63.61	66.24	69.97	59.75
Reasoning Models								
Phi-4-mini-reasoning	Markdown	1.23	0.89	0.00	6.77	1.72	1.67	2.05
	HTML	0.81	0.00	0.00	6.25	0.98	0.00	1.34
	JSON	0.00	0.00	0.00	9.38	0.00	0.88	1.71
	LaTeX	0.00	0.00	1.39	4.69	0.00	0.43	1.08
	SQL	4.11	0.00	0.00	14.41	0.00	1.67	3.36
	XML	0.00	0.00	0.00	7.81	4.60	12.50	4.15
	CSV	1.22	1.65	0.00	5.21	0.06	15.00	3.86
Qwen3-4B-thinking	Markdown	23.48	15.94	0.00	22.78	16.77	24.39	17.22
	HTML	25.00	14.63	0.00	10.76	5.95	17.50	12.31
	JSON	23.07	14.31	0.00	12.33	12.78	31.67	15.69
	LaTeX	25.97	8.22	3.70	20.66	19.87	33.33	18.62
	SQL	33.16	15.45	0.00	12.15	25.44	24.39	18.43
	XML	29.78	14.83	0.00	13.89	15.30	28.33	17.02
	CSV	33.74	12.46	5.56	12.50	15.48	20.67	16.74
DeepSeek-R1-Distill-Qwen-7B	Markdown	0.00	1.30	0.00	10.94	12.07	0.00	4.05
	HTML	0.00	0.00	0.00	6.25	5.17	0.00	1.90
	JSON	0.00	0.44	0.00	12.50	3.45	0.00	2.73
	LaTeX	1.22	0.30	0.00	15.63	3.56	0.00	3.45
	SQL	1.22	0.00	0.00	10.94	8.62	0.00	3.46
	XML	0.00	0.00	0.79	17.01	1.72	0.00	3.26
	CSV	0.00	3.13	0.68	4.69	7.16	0.00	2.61
DeepSeek-R1-Distill-Llama-8B	Markdown	25.85	35.01	6.86	39.58	33.66	25.84	27.80
	HTML	50.31	36.35	25.21	34.12	32.99	25.32	34.05
	JSON	37.83	34.53	5.04	22.22	30.38	10.00	23.33
	LaTeX	34.63	41.25	10.25	38.64	40.70	48.94	35.73
	SQL	22.14	38.45	13.54	34.26	38.12	20.16	27.78
	XML	32.24	36.95	5.97	43.06	38.23	31.64	31.35
	CSV	36.83	35.12	16.05	31.94	27.86	17.22	27.51
Qwen3-8B-thinking	Markdown	30.39	16.53	0.24	19.97	20.37	23.06	18.42
	HTML	29.78	17.33	5.56	27.14	18.36	26.39	20.76
	JSON	28.72	24.85	1.27	15.28	16.96	23.83	18.49
	LaTeX	27.25	12.77	0.00	23.09	18.30	37.78	19.86
	SQL	28.73	19.44	0.00	25.87	24.14	13.33	18.58
	XML	27.74	20.41	2.38	26.04	14.81	33.33	20.79
	CSV	32.22	12.61	1.11	17.19	16.44	35.00	19.09
GLM-Z1-9B-0414	Markdown	55.79	58.96	20.57	45.20	41.24	45.29	44.51
	HTML	30.14	43.15	8.82	29.72	34.43	27.93	29.03
	JSON	38.75	51.56	10.25	37.33	42.74	36.93	36.26
	LaTeX	54.86	54.66	14.52	44.10	51.11	27.23	41.08
	SQL	55.98	49.49	19.43	51.15	47.52	39.70	43.88
	XML	33.26	39.53	0.87	35.63	38.52	38.93	31.12
	CSV	57.01	47.61	5.94	54.50	49.45	30.20	40.78
Qwen3-14B-thinking	Markdown	34.54	22.90	0.00	45.65	42.63	45.61	31.89
	HTML	38.19	23.98	0.00	44.79	29.41	31.39	27.96
	JSON	35.87	31.87	2.38	26.56	30.80	29.44	26.15
	LaTeX	34.94	29.09	5.56	43.58	31.86	21.13	27.69
	SQL	42.32	27.35	5.56	47.92	30.68	32.37	31.03
	XML	39.58	32.09	7.94	42.71	30.29	30.50	30.52
	CSV	31.79	27.47	0.00	37.85	32.41	33.17	27.11
Qwen3-30B-A3B-thinking	Markdown	37.35	24.80	5.56	35.06	24.48	23.61	25.14
	HTML	26.79	15.91	0.00	31.42	22.86	38.61	22.60
	JSON	24.87	26.61	0.00	30.42	19.80	30.00	21.95
	LaTeX	31.32	21.05	0.00	32.12	25.29	37.50	24.55
	SQL	39.02	25.32	5.56	32.64	21.20	43.61	27.89
	XML	37.46	26.05	0.00	32.56	24.99	43.06	27.35

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	29.57	20.71	5.56	27.43	24.41	26.67	22.39
QwQ-32B	Markdown	61.88	55.70	21.15	54.94	64.12	54.81	52.10
	HTML	58.56	55.00	12.35	54.58	58.57	56.49	49.26
	JSON	49.59	64.10	4.68	44.46	51.43	52.33	44.43
	LaTeX	60.30	62.52	7.97	55.91	61.94	57.75	51.07
	SQL	56.21	57.20	14.41	56.53	59.86	54.60	49.80
	XML	49.20	55.11	5.50	50.15	53.53	62.53	46.00
	CSV	62.98	57.55	13.90	57.05	66.00	56.49	52.33
DeepSeek-R1-Distill-Qwen-32B	Markdown	63.19	55.10	16.00	63.87	60.95	53.19	52.05
	HTML	60.93	54.36	17.79	44.91	62.68	62.00	50.44
	JSON	49.83	53.37	22.98	44.10	59.23	37.12	44.44
	LaTeX	53.48	53.84	11.52	57.77	64.32	42.23	47.19
	SQL	58.14	51.89	18.60	56.13	54.24	47.33	47.72
	XML	64.12	46.80	19.18	49.48	56.03	52.44	48.01
	CSV	49.98	49.06	12.62	61.28	54.39	41.84	44.86
Qwen3-32B-thinking	Markdown	39.82	32.73	0.00	47.09	32.60	43.52	32.63
	HTML	33.74	27.27	0.00	35.59	31.46	36.50	27.43
	JSON	31.50	25.14	2.38	25.00	33.91	24.17	23.68
	LaTeX	44.24	30.11	0.00	41.67	25.98	28.89	28.48
	SQL	35.65	22.26	5.56	37.33	30.99	48.56	30.06
	XML	44.68	30.49	0.00	47.15	30.13	41.39	32.31
	CSV	43.66	24.80	0.00	45.83	34.52	40.95	31.63
GLM-Z1-32B-0414	Markdown	48.06	43.60	17.23	38.10	65.03	34.80	41.14
	HTML	38.38	50.37	9.98	34.20	52.45	44.72	38.35
	JSON	40.81	59.26	23.48	36.04	62.52	42.98	44.18
	LaTeX	47.77	60.58	14.00	34.01	63.13	27.71	41.20
	SQL	49.77	49.20	15.16	29.34	55.12	30.84	38.24
	XML	39.57	51.58	10.45	37.27	57.34	52.72	41.49
	CSV	51.47	46.39	10.13	32.81	53.78	36.55	38.52
DeepSeek-R1-Distill-Llama-70B	Markdown	54.16	55.01	26.96	67.79	47.97	39.26	48.52
	HTML	57.44	49.62	23.65	60.74	58.01	37.25	47.79
	JSON	67.24	51.73	29.74	44.62	56.70	44.68	49.12
	LaTeX	56.50	56.03	14.67	46.09	55.25	44.27	45.47
	SQL	59.92	50.94	12.72	50.21	54.38	56.74	47.49
	XML	66.43	56.79	24.32	58.22	53.62	50.38	51.63
	CSV	60.14	59.60	17.80	59.84	52.01	32.81	47.03
DeepSeek-R1	Markdown	72.02	88.63	34.49	58.19	73.43	76.64	67.23
	HTML	70.17	88.02	32.99	60.60	78.79	62.84	65.57
	JSON	70.94	90.21	32.25	39.00	72.23	66.08	61.79
	LaTeX	72.25	87.50	30.39	48.91	65.36	75.69	63.35
	SQL	74.94	85.07	33.26	58.18	70.18	76.01	66.27
	XML	71.84	89.57	33.59	54.87	72.00	69.18	65.17
	CSV	71.24	92.12	31.33	56.10	65.59	75.34	65.29
Gemini-2.0-flash-thinking-exp-01-21	Markdown	66.93	77.86	22.98	51.11	61.68	51.49	55.34
	HTML	67.78	71.43	22.53	49.20	66.53	63.16	56.77
	JSON	64.94	70.68	33.09	46.77	69.27	65.33	58.34
	LaTeX	68.00	76.86	26.80	52.50	64.40	57.07	57.61
	SQL	63.64	73.23	27.26	49.24	66.28	51.81	55.24
	XML	68.66	66.85	30.15	52.15	52.15	65.53	55.92
	CSV	69.75	71.97	25.43	50.42	62.90	56.51	56.16

Table 5: Full experimental results under the **32K length setting** across 7 formats and 6 task types. Avg is the mean score.

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
Vanilla Models								
LLaMA-3.2-3B-Instruct	Markdown	18.34	3.84	3.77	2.27	7.62	23.32	9.86
	HTML	17.29	0.00	15.43	4.76	1.98	4.33	7.30
	JSON	27.75	0.00	0.00	0.00	0.00	0.00	4.63
	LaTeX	19.90	0.51	15.06	7.14	1.25	6.31	8.36
	SQL	23.23	1.81	7.96	0.00	13.03	13.49	9.92
	XML	8.10	0.42	8.64	4.76	4.79	6.37	5.51
	CSV	23.65	2.47	15.02	2.27	10.63	10.48	10.75
Qwen2.5-3B-Instruct	Markdown	19.31	3.03	9.04	27.38	31.77	8.61	16.52
	HTML	14.91	0.89	0.62	29.65	26.27	4.06	12.73
	JSON	20.10	2.64	8.82	25.11	25.52	10.54	15.46
	LaTeX	17.21	0.00	7.50	38.52	18.06	4.44	14.29
	SQL	21.05	0.95	8.16	29.84	28.61	14.28	17.15
	XML	16.15	0.00	5.07	29.55	23.55	3.33	12.94
	CSV	13.81	2.35	5.72	33.93	27.14	5.62	14.76
Phi-3-mini-128k-instruct	Markdown	27.91	0.00	3.77	41.18	16.57	20.59	18.34
	HTML	20.00	0.00	3.97	33.55	17.03	10.00	14.09
	JSON	23.75	0.00	2.93	19.21	24.52	6.33	12.79
	LaTeX	33.65	0.00	4.40	31.39	19.78	13.13	17.06
	SQL	32.55	0.00	8.30	31.76	19.46	12.67	17.46
	XML	10.00	0.00	0.00	20.89	20.33	5.00	9.37
	CSV	18.73	0.00	8.58	35.93	12.91	11.00	14.53
Phi-4-mini-instruct	Markdown	29.54	4.52	10.93	30.63	5.95	11.59	15.53
	HTML	16.53	1.67	2.12	21.10	2.38	3.33	7.86
	JSON	28.71	0.02	6.75	10.71	13.42	8.03	11.27
	LaTeX	32.88	1.82	10.68	25.38	6.01	9.61	14.40
	SQL	26.17	3.09	2.09	28.25	8.73	12.03	13.39
	XML	24.58	0.04	2.19	17.53	6.79	3.33	9.08
	CSV	25.57	2.15	4.57	21.10	6.22	10.31	11.65
Qwen3-4B	Markdown	21.06	8.81	9.75	26.92	6.03	18.99	15.26
	HTML	22.82	4.59	9.22	25.87	7.30	8.73	13.09
	JSON	18.58	15.49	8.32	25.97	5.44	14.29	14.68
	LaTeX	28.45	7.72	4.57	36.15	8.07	9.93	15.82
	SQL	19.91	7.33	8.92	27.54	2.94	9.23	12.65
	XML	23.15	0.46	8.43	16.67	7.92	9.46	11.01
	CSV	13.51	6.97	5.72	12.77	8.04	7.37	9.06
Gemma-3-4B-it	Markdown	25.01	4.43	7.04	29.65	10.18	26.50	17.14
	HTML	26.52	7.53	9.45	15.18	16.60	33.91	18.20
	JSON	21.37	0.03	4.90	10.01	17.63	13.33	11.21
	LaTeX	23.99	4.48	6.85	15.71	12.32	32.13	15.92
	SQL	23.52	4.80	9.02	22.85	11.90	24.35	16.07
	XML	20.75	1.67	2.88	11.09	11.53	9.56	9.58
	CSV	29.20	6.60	8.37	25.01	13.55	19.45	17.03
Llama-3.1-8B-Instruct	Markdown	37.75	10.80	15.17	25.97	19.52	39.54	24.79
	HTML	37.16	16.14	12.10	4.55	18.15	30.95	19.84
	JSON	40.18	14.00	15.48	8.33	12.86	17.17	18.00
	LaTeX	33.38	4.70	25.87	23.00	19.80	36.94	23.95
	SQL	39.59	3.77	17.29	24.68	13.79	38.59	22.95
	XML	41.85	2.96	10.89	3.57	18.32	46.04	20.61
	CSV	37.16	8.29	16.63	30.19	13.23	49.48	25.83
Qwen2.5-7B-Instruct	Markdown	27.21	5.60	12.27	49.68	28.68	24.16	24.60
	HTML	25.64	10.31	21.82	55.03	35.10	22.60	28.42
	JSON	31.42	14.55	24.49	49.78	29.11	20.51	28.31
	LaTeX	39.77	4.29	20.39	62.55	28.36	15.95	28.55
	SQL	32.43	5.84	19.61	46.00	24.62	23.66	25.36
	XML	31.87	2.75	11.44	47.40	39.72	29.62	27.13
	CSV	31.74	6.52	7.81	47.40	20.69	21.24	22.57
Qwen2.5-7B-Instruct-1M	Markdown	46.59	10.78	25.57	59.98	13.43	27.23	30.60
	HTML	47.05	13.14	20.10	56.40	9.37	24.58	28.44
	JSON	45.24	19.47	20.90	41.96	25.44	21.64	29.11
	LaTeX	48.14	10.00	23.60	53.84	11.89	26.84	29.05
	SQL	48.22	20.00	22.13	46.81	11.97	29.49	29.77
	XML	51.47	2.80	11.77	48.48	21.59	19.27	25.90

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	46.97	20.14	13.77	51.27	21.97	24.37	29.75
TableGPT2-7B	Markdown	41.74	8.81	10.81	49.32	38.36	23.76	28.80
	HTML	37.45	17.25	12.22	39.18	30.07	19.16	25.89
	JSON	35.31	16.04	17.14	36.09	27.80	16.92	24.89
	LaTeX	34.50	9.31	10.84	42.15	38.50	21.11	26.07
	SQL	41.97	16.27	9.95	35.71	34.21	21.70	26.64
	XML	33.48	3.69	14.07	42.15	35.49	26.84	25.95
	CSV	32.55	4.47	9.82	43.43	30.74	17.65	23.11
TableLLM-Qwen2-7B	Markdown	12.64	1.53	1.04	19.53	10.00	2.94	7.95
	HTML	4.17	1.25	0.00	0.00	1.43	2.67	1.59
	JSON	11.11	0.00	3.70	2.38	1.43	4.33	3.83
	LaTeX	11.22	0.28	0.00	13.96	4.29	4.06	5.64
	SQL	16.65	1.25	0.00	15.96	2.86	2.67	6.56
	XML	4.58	0.00	0.00	2.27	0.00	5.33	2.03
	CSV	23.41	1.25	1.04	13.96	8.57	1.10	8.22
TableLLM-Llama3.1-8B	Markdown	25.42	0.00	2.43	7.03	8.57	8.00	8.58
	HTML	14.17	1.67	2.78	8.82	1.43	7.00	5.98
	JSON	19.17	1.67	2.78	9.42	4.29	5.00	7.05
	LaTeX	17.08	0.00	2.78	9.52	3.81	10.33	7.25
	SQL	25.42	0.00	4.17	2.27	1.43	10.33	7.27
	XML	20.28	1.67	2.78	4.76	2.86	9.67	7.00
	CSV	15.83	2.78	2.78	9.42	1.43	7.11	6.56
TableLlama	Markdown	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	HTML	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	JSON	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LaTeX	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SQL	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	XML	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CSV	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mistral-7B-Instruct-v0.3	Markdown	23.54	7.55	5.35	15.67	5.71	15.01	12.14
	HTML	7.50	1.67	0.69	26.73	2.79	5.67	7.51
	JSON	9.44	0.00	1.04	20.40	6.59	1.89	6.56
	LaTeX	17.61	5.41	5.70	16.96	9.94	9.13	10.79
	SQL	18.44	6.09	6.56	13.41	3.94	4.13	8.76
	XML	2.00	0.00	1.59	17.42	5.47	3.35	4.97
	CSV	19.69	10.70	8.94	11.60	9.09	10.40	11.74
Ministral-8B-Instruct-2410	Markdown	0.00	0.06	0.00	0.00	0.00	0.00	0.01
	HTML	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	JSON	0.00	0.00	0.00	2.27	0.00	0.00	0.38
	LaTeX	0.00	0.10	0.00	0.00	0.00	0.00	0.02
	SQL	1.82	0.15	0.00	0.00	0.48	0.30	0.46
	XML	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CSV	3.33	0.00	0.00	4.55	0.06	0.00	1.32
Qwen3-8B	Markdown	31.29	7.87	9.98	52.28	20.20	20.32	23.66
	HTML	34.97	4.96	8.36	43.34	15.54	21.26	21.41
	JSON	27.28	12.54	7.44	52.76	19.24	22.53	23.63
	LaTeX	30.17	3.11	6.44	38.90	19.18	22.11	19.99
	SQL	29.70	6.59	12.94	44.66	15.36	16.49	20.96
	XML	29.81	0.00	9.85	36.23	13.05	15.21	17.36
	CSV	29.04	4.67	8.35	34.74	18.67	18.29	18.96
GLM-4-9B-Chat	Markdown	34.90	15.08	10.78	41.13	17.21	19.56	23.11
	HTML	39.43	17.63	14.56	38.47	19.28	19.67	24.84
	JSON	48.30	16.69	16.35	19.05	27.79	20.20	24.73
	LaTeX	39.60	16.63	14.16	33.13	16.40	19.97	23.31
	SQL	41.30	14.92	15.31	45.18	15.99	20.20	25.48
	XML	46.26	14.24	12.41	34.90	23.55	22.04	25.57
	CSV	34.73	16.44	8.68	43.72	15.40	18.80	22.96
GLM-4-9B-Chat-1M	Markdown	39.86	16.41	11.99	42.80	14.13	12.82	23.00
	HTML	43.62	15.60	11.99	26.46	23.57	15.81	22.84
	JSON	43.77	14.17	16.86	23.59	20.98	16.04	22.57
	LaTeX	43.12	16.77	16.69	38.15	24.29	24.09	27.18
	SQL	42.29	14.25	11.66	36.36	19.52	22.22	24.38
	XML	52.81	14.87	9.36	31.60	26.48	25.43	26.76

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	40.12	4.89	8.52	38.50	10.24	16.78	19.84
GLM-4-9B-0414	Markdown	20.34	9.22	10.78	13.85	15.08	14.16	13.90
	HTML	19.34	3.00	4.89	9.09	4.26	15.81	9.40
	JSON	22.83	0.42	4.13	0.00	11.56	13.51	8.74
	LaTeX	23.57	4.91	8.01	10.39	11.09	17.55	12.59
	SQL	23.74	2.50	8.29	14.34	8.78	13.06	11.78
	XML	22.14	0.00	3.51	6.82	3.89	7.38	7.29
	CSV	20.35	7.21	8.31	21.97	5.98	14.34	13.03
Gemma-3-12B-it	Markdown	42.12	17.41	19.22	46.75	36.12	35.03	32.77
	HTML	39.54	17.89	19.88	30.71	30.72	25.07	27.30
	JSON	34.17	2.66	7.67	28.73	24.38	20.98	19.77
	LaTeX	36.12	19.22	19.91	43.53	31.36	32.18	30.38
	SQL	36.30	19.40	21.57	48.43	31.18	34.77	31.94
	XML	25.58	5.29	5.31	24.68	20.53	15.07	16.08
	CSV	41.64	13.01	17.74	43.89	36.21	25.83	29.72
Mistral-Nemo-Instruct-2407	Markdown	13.28	4.62	0.61	35.42	38.29	8.40	16.77
	HTML	15.87	1.84	2.31	35.01	33.81	6.72	15.93
	JSON	13.44	4.37	2.45	33.71	39.17	10.22	17.23
	LaTeX	8.06	2.28	0.00	36.80	42.50	6.78	16.07
	SQL	5.48	4.69	0.61	33.25	34.44	4.50	13.83
	XML	13.44	1.89	3.17	30.87	22.50	11.00	13.81
	CSV	7.83	8.66	4.57	25.43	42.43	5.37	15.72
Qwen2.5-14B-Instruct	Markdown	39.92	15.57	18.41	56.38	33.61	47.49	35.23
	HTML	38.69	9.08	22.76	55.84	26.59	15.23	28.03
	JSON	36.90	13.06	18.78	46.13	22.86	23.22	26.82
	LaTeX	44.77	6.03	19.92	65.42	22.81	28.82	31.29
	SQL	42.28	12.62	24.91	50.38	28.23	26.16	30.76
	XML	50.35	3.88	13.82	47.91	17.46	33.88	27.88
	CSV	39.33	9.39	23.32	47.51	32.09	14.37	27.67
Qwen2.5-14B-Instruct-1M	Markdown	61.74	23.23	25.85	75.65	28.46	39.00	42.32
	HTML	60.97	13.23	17.65	73.27	29.60	23.33	36.34
	JSON	55.29	21.15	23.42	40.31	36.64	29.51	34.39
	LaTeX	66.61	17.78	26.05	70.29	35.71	40.67	42.85
	SQL	60.45	20.38	25.75	70.18	26.62	40.93	40.72
	XML	63.79	2.85	13.86	57.41	34.20	43.36	35.91
	CSV	59.09	19.34	23.55	66.56	28.43	45.51	40.41
Qwen3-14B	Markdown	53.26	11.15	19.47	65.68	36.70	41.15	37.90
	HTML	58.39	12.92	18.04	57.64	34.76	30.34	35.35
	JSON	36.25	14.60	20.31	40.62	27.10	22.31	26.87
	LaTeX	52.07	14.11	23.59	57.63	39.63	19.49	34.42
	SQL	60.26	15.21	18.83	61.69	32.62	50.95	39.93
	XML	55.86	0.95	8.67	61.90	38.96	43.00	34.89
	CSV	54.31	15.71	16.25	48.17	28.89	27.25	31.76
Mistral-Small-3.1-24B-Instruct-2503	Markdown	64.11	21.42	22.33	48.81	41.43	40.20	39.72
	HTML	62.93	21.97	17.67	31.02	29.93	39.72	33.87
	JSON	53.79	10.67	20.94	14.58	33.75	37.98	28.62
	LaTeX	60.74	22.96	22.09	53.68	42.38	45.02	41.14
	SQL	65.52	25.09	19.60	37.12	44.17	43.64	39.19
	XML	53.18	5.28	13.39	27.27	26.05	39.77	27.49
	CSV	57.76	20.18	20.51	36.85	42.34	39.91	36.26
Qwen3-30B-A3B	Markdown	41.00	2.31	10.43	47.40	33.72	17.43	25.38
	HTML	43.24	1.14	13.18	45.13	34.98	33.58	28.54
	JSON	31.45	6.64	15.55	47.40	28.67	18.82	24.76
	LaTeX	46.42	3.24	14.16	52.06	46.61	37.17	33.28
	SQL	37.00	8.35	10.10	47.40	34.23	16.44	25.59
	XML	35.03	0.00	10.98	45.73	32.45	35.68	26.64
	CSV	36.50	6.32	12.13	47.51	39.59	19.02	26.84
Qwen3-32B	Markdown	55.85	19.78	28.39	57.25	42.19	52.09	42.59
	HTML	49.85	17.40	12.44	53.08	24.32	41.79	33.15
	JSON	50.53	20.25	21.86	30.52	41.64	48.05	35.47
	LaTeX	41.78	13.60	23.85	49.73	34.91	52.99	36.14
	SQL	42.71	15.14	24.73	56.66	31.69	51.28	37.04
	XML	52.48	2.46	18.25	54.27	32.52	39.55	33.26

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	54.09	24.21	21.63	52.00	37.13	50.41	39.91
GLM-4-32B-0414	Markdown	31.43	9.11	10.48	62.18	39.99	12.92	27.68
	HTML	20.33	5.29	2.96	24.73	12.25	11.25	12.80
	JSON	20.51	2.60	5.62	18.40	15.24	6.78	11.53
	LaTeX	29.24	9.89	15.54	46.62	31.07	15.49	24.64
	SQL	20.43	2.59	11.15	44.88	29.35	16.58	20.83
	XML	22.24	0.13	5.79	6.93	10.21	7.16	8.74
	CSV	33.20	10.02	12.95	51.46	43.42	13.98	27.51
Llama-3.1-70B-Instruct	Markdown	71.81	19.14	34.18	67.53	39.47	50.38	47.08
	HTML	63.18	7.17	20.97	52.98	32.93	44.30	36.92
	JSON	60.26	0.98	12.88	24.78	31.24	33.12	27.21
	LaTeX	70.66	17.68	33.99	50.32	25.18	49.58	41.24
	SQL	66.22	18.26	28.86	47.73	28.03	48.20	39.55
	XML	45.58	3.88	13.14	37.23	31.62	41.90	28.89
	CSV	59.31	16.94	26.88	50.11	35.49	52.49	40.20
Llama-3.3-70B-Instruct	Markdown	66.92	27.79	35.59	32.47	34.02	59.16	42.66
	HTML	52.91	5.11	17.90	24.08	19.80	38.64	26.40
	JSON	47.05	5.65	12.82	13.37	18.56	29.77	21.20
	LaTeX	71.06	10.14	35.86	34.96	30.24	71.19	42.24
	SQL	66.44	24.43	35.20	35.44	30.24	60.03	41.96
	XML	22.75	3.32	12.88	13.85	16.79	29.72	16.55
	CSV	70.17	30.14	36.81	30.19	33.28	60.98	43.60
Qwen2.5-72B-Instruct	Markdown	57.38	24.87	23.49	65.64	44.36	48.98	44.12
	HTML	55.38	16.81	25.48	66.34	39.43	45.34	41.46
	JSON	45.97	27.21	21.42	40.42	41.42	26.56	33.83
	LaTeX	60.57	18.35	29.49	70.78	55.39	45.65	46.71
	SQL	56.02	17.42	28.64	66.68	50.49	48.02	44.54
	XML	46.87	3.81	17.93	65.85	38.09	49.19	36.96
	CSV	60.03	20.89	26.74	71.48	58.64	48.42	47.70
Mistral-Large-Instruct-2411	Markdown	50.23	16.37	20.41	39.61	19.52	40.07	31.03
	HTML	40.59	2.51	10.05	32.39	16.71	34.70	22.82
	JSON	30.80	6.80	16.32	13.77	27.38	39.85	22.49
	LaTeX	52.75	12.34	21.82	50.32	28.81	44.45	35.08
	SQL	48.04	21.12	19.90	41.34	20.36	23.87	29.10
	XML	25.48	2.65	13.36	15.15	9.64	32.29	16.43
	CSV	65.79	15.91	20.56	37.82	25.36	44.28	34.95
DeepSeek-V3	Markdown	71.90	31.66	44.02	77.54	47.42	54.44	54.50
	HTML	72.44	26.21	38.50	77.44	38.37	53.49	51.07
	JSON	70.66	25.77	34.45	30.52	43.50	60.13	44.17
	LaTex	71.63	30.68	39.18	70.40	42.24	52.54	51.11
	SQL	76.16	31.89	41.72	75.05	33.57	51.28	51.61
	XML	68.93	26.77	22.76	56.87	44.72	56.78	46.14
	CSV	68.82	29.96	35.82	68.13	42.42	48.05	48.87
GPT-4o-mini-2024-07-18	Markdown	48.83	25.36	29.01	49.95	33.03	41.05	37.87
	HTML	50.63	23.34	26.92	45.08	33.63	37.99	36.27
	JSON	55.09	24.84	20.60	32.52	30.81	41.47	34.22
	LaTex	50.41	25.39	25.33	47.56	27.45	43.73	36.65
	SQL	58.00	25.62	28.54	52.11	33.99	40.03	39.72
	XML	62.51	4.43	14.97	40.53	28.81	30.57	30.30
	CSV	50.60	21.44	24.61	44.97	27.72	35.70	34.17
Gemini-2.0-flash	Markdown	76.81	31.62	35.39	55.48	51.02	61.21	51.92
	HTML	76.67	39.10	35.76	49.03	39.19	49.81	48.26
	JSON	71.58	28.79	31.66	53.57	40.83	73.51	49.99
	LaTeX	76.90	30.66	27.75	55.23	41.20	49.85	46.93
	SQL	74.96	32.56	32.35	55.95	38.61	59.73	49.03
	XML	71.00	38.00	29.31	47.46	45.55	73.11	50.74
	CSV	74.85	34.24	31.80	50.99	45.30	63.82	50.17
GPT-4o-2024-08-06	Markdown	75.10	37.54	49.01	82.67	52.74	70.76	61.30
	HTML	69.31	36.35	42.31	77.82	54.52	54.23	55.76
	JSON	74.26	34.17	32.60	48.24	48.09	58.07	49.24
	LaTex	73.78	39.56	43.67	86.63	55.43	54.17	58.87
	SQL	76.25	36.44	43.52	75.52	55.29	52.04	56.51
	XML	73.50	13.28	22.60	57.34	52.34	45.46	44.09

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	69.92	34.73	47.25	83.46	52.14	54.00	56.92
Claude-3.5-sonnet-20241022	Markdown	72.80	42.25	36.25	86.74	64.00	54.21	59.38
	HTML	70.17	19.03	38.12	66.56	45.75	54.15	48.96
	JSON	63.64	12.94	25.54	30.57	39.82	48.98	36.92
	LaTex	69.35	22.50	30.08	62.07	44.88	51.49	46.73
	SQL	70.87	25.15	21.25	73.16	34.96	43.25	44.77
	XML	67.66	36.53	29.87	73.16	44.70	52.47	50.73
	CSV	71.31	13.66	30.80	70.78	49.88	50.89	47.89
Reasoning Models								
Phi-4-mini-reasoning	Markdown	0.00	0.00	0.00	6.82	2.50	0.00	1.55
	HTML	0.00	0.00	0.00	2.27	0.00	0.00	0.38
	JSON	0.00	0.00	0.00	2.27	2.50	0.00	0.80
	LaTeX	0.00	0.00	0.00	4.17	0.00	0.00	0.69
	SQL	0.00	0.00	0.00	1.79	0.00	0.00	0.30
	XML	0.00	0.00	0.00	4.55	0.00	0.00	0.76
	CSV	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen3-4B-thinking	Markdown	26.11	0.00	3.70	16.23	5.60	8.78	10.07
	HTML	20.00	0.00	2.31	20.02	16.43	15.44	12.37
	JSON	16.94	1.67	7.52	9.31	25.48	3.61	10.75
	LaTeX	34.09	0.00	3.70	26.95	13.93	11.11	14.96
	SQL	36.17	1.67	3.70	21.00	11.07	13.56	14.53
	XML	26.11	1.67	5.79	28.23	15.00	30.89	17.95
	CSV	24.17	1.67	2.31	22.67	3.57	14.11	11.42
DeepSeek-R1-Distill-Qwen-7B	Markdown	0.00	0.00	0.00	6.82	2.50	0.00	1.55
	HTML	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	JSON	0.00	0.00	0.00	2.27	0.00	0.00	0.38
	LaTeX	0.00	0.00	0.00	2.27	5.00	0.00	1.21
	SQL	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	XML	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CSV	0.00	0.00	0.00	2.27	2.50	0.00	0.80
DeepSeek-R1-Distill-Llama-8B	Markdown	20.88	19.69	12.33	29.79	15.06	27.69	20.91
	HTML	10.25	5.15	2.16	30.00	18.93	8.36	12.47
	JSON	3.17	0.00	3.47	13.74	10.00	2.33	5.45
	LaTeX	26.81	15.97	11.79	43.45	26.67	14.75	23.24
	SQL	13.07	15.24	6.63	36.72	11.43	9.14	15.37
	XML	2.50	0.00	0.93	24.17	17.14	0.00	7.46
	CSV	20.57	18.79	16.70	30.68	18.57	7.59	18.82
Qwen3-8B-thinking	Markdown	39.59	2.27	4.63	28.81	20.24	30.28	20.97
	HTML	34.44	0.48	4.96	36.63	17.02	15.07	18.10
	JSON	26.42	3.43	7.70	24.96	11.07	15.11	14.78
	LaTeX	28.73	1.67	4.40	35.17	18.81	18.67	17.91
	SQL	29.09	0.00	7.52	27.44	12.86	29.67	17.76
	XML	33.33	0.00	6.94	39.72	29.40	25.67	22.51
	CSV	38.61	1.67	1.39	21.59	14.29	28.33	17.65
GLM-Z1-9B-0414	Markdown	20.16	0.51	6.86	38.31	27.00	10.11	17.16
	HTML	2.78	2.11	3.01	4.55	8.10	4.44	4.16
	JSON	13.26	3.92	5.82	28.30	25.81	4.33	13.57
	LaTeX	22.33	2.21	12.52	38.31	35.99	11.37	20.46
	SQL	17.36	6.75	10.62	30.68	31.23	7.61	17.38
	XML	10.78	0.00	0.00	13.64	5.00	3.61	5.50
	CSV	8.61	5.49	11.11	38.10	14.60	26.44	17.39
Qwen3-14B-thinking	Markdown	58.62	3.98	14.56	51.52	24.64	36.92	31.71
	HTML	44.87	4.08	9.03	54.78	34.64	20.28	27.95
	JSON	40.20	4.11	13.82	36.04	20.71	21.11	22.67
	LaTeX	43.59	1.18	16.18	64.34	33.69	21.00	30.00
	SQL	61.65	1.67	11.94	55.74	34.52	34.83	33.39
	XML	60.01	1.67	8.85	53.68	38.57	32.51	32.55
	CSV	45.98	0.00	8.61	55.25	28.33	34.67	28.81
Qwen3-30B-A3B-thinking	Markdown	45.29	0.00	7.87	33.87	17.74	19.83	20.77
	HTML	36.67	1.11	4.17	42.59	15.60	13.33	18.91
	JSON	27.46	1.25	5.38	26.95	20.00	20.78	16.97
	LaTeX	38.95	1.55	0.00	36.15	24.13	14.10	19.15
	SQL	39.04	0.00	10.42	38.13	22.14	16.67	21.07
	XML	34.59	0.00	5.09	39.50	15.71	18.00	18.82

Model	Format	EM	BCF	FCM	FR	EKF	INI	Avg.
	CSV	36.33	1.03	6.48	42.80	28.45	14.44	21.59
QwQ-32B	Markdown	56.95	22.12	26.16	64.16	37.54	29.77	39.45
	HTML	61.62	6.60	15.06	65.75	42.42	57.12	41.43
	JSON	49.58	6.49	16.76	45.75	41.07	27.41	31.18
	LaTeX	52.31	25.47	25.19	66.13	27.23	41.69	39.67
	SQL	55.64	25.67	22.28	61.31	50.00	30.93	40.97
	XML	53.37	3.03	12.25	61.11	34.52	27.63	31.98
	CSV	46.20	30.68	18.22	55.25	40.24	23.45	35.67
DeepSeek-R1-Distill-Qwen-32B	Markdown	38.16	17.72	10.84	40.53	56.77	44.00	34.67
	HTML	11.94	1.91	3.83	37.77	31.45	18.56	17.58
	JSON	14.17	0.67	3.08	15.91	28.98	21.10	13.99
	LaTeX	36.25	6.47	13.31	38.26	38.98	45.69	29.83
	SQL	32.60	8.25	8.98	44.48	35.60	44.83	29.12
	XML	14.17	1.26	1.67	19.59	15.00	18.00	11.61
	CSV	41.77	16.98	11.55	46.97	52.63	54.43	37.39
Qwen3-32B-thinking	Markdown	58.06	3.44	14.74	49.91	32.74	33.67	32.09
	HTML	52.37	2.00	8.33	41.88	37.26	32.33	29.03
	JSON	41.87	3.33	13.73	28.14	19.88	18.00	20.82
	LaTeX	50.98	2.33	15.39	57.94	23.45	19.78	28.31
	SQL	57.40	1.67	13.44	52.98	21.67	35.67	30.47
	XML	57.01	0.00	10.42	55.84	33.21	33.44	31.65
	CSV	51.87	0.67	13.20	47.84	29.29	34.11	29.50
GLM-Z1-32B-0414	Markdown	35.42	12.55	23.74	42.37	38.05	19.28	28.57
	HTML	25.97	2.94	11.41	27.71	29.75	13.19	18.49
	JSON	27.83	10.04	15.27	42.26	31.55	7.94	22.48
	LaTeX	33.34	24.50	20.55	45.94	48.93	22.86	32.69
	SQL	31.47	9.26	22.03	33.77	38.14	26.32	26.83
	XML	27.51	4.78	13.21	26.95	26.64	9.67	18.13
	CSV	33.20	9.93	18.07	36.53	33.48	33.17	27.40
DeepSeek-R1-Distill-Llama-70B	Markdown	45.30	24.52	17.70	42.53	45.95	48.37	37.40
	HTML	30.42	2.47	10.92	34.90	34.88	35.26	24.81
	JSON	18.19	5.09	14.31	17.21	24.29	23.90	17.16
	LaTeX	49.53	23.61	19.44	50.54	37.89	59.37	40.06
	SQL	52.31	32.98	17.52	43.40	50.79	61.28	43.05
	XML	12.50	0.73	1.94	18.29	12.50	16.67	10.44
	CSV	53.77	30.46	17.05	42.05	33.25	48.00	37.43
DeepSeek-R1	Markdown	76.85	48.31	38.79	75.05	50.45	48.92	56.40
	HTML	75.05	49.21	39.98	67.02	50.47	65.83	57.93
	JSON	61.89	45.66	34.40	45.17	44.17	51.15	47.07
	LaTex	72.61	44.48	36.83	59.42	49.98	47.14	51.74
	SQL	72.21	53.97	39.06	65.75	45.16	44.83	53.49
	XML	71.05	41.02	19.53	58.41	41.63	44.97	46.10
	CSV	69.67	45.40	38.18	59.31	44.30	44.58	50.24
Gemini-2.0-flash-thinking-exp-01-21	Markdown	73.35	38.95	26.93	52.74	41.99	39.36	45.55
	HTML	76.63	38.44	33.59	57.75	51.60	54.58	52.10
	JSON	72.09	38.82	31.21	49.03	55.58	59.76	51.08
	LaTeX	76.48	32.92	17.68	46.66	48.15	48.36	45.04
	SQL	77.92	36.17	36.31	60.61	49.53	52.01	52.09
	XML	74.63	32.39	31.38	57.75	49.26	58.10	50.59
	CSV	71.69	34.71	28.44	46.16	52.11	60.19	48.88

Table 6: Full experimental results under the **128K length setting** across 7 formats and 6 task types. Avg is the mean score.