StructFact: Reasoning Factual Knowledge from Structured Data with Large Language Models

Sirui Huang[1,2*], **Yanggan Gu**[3*], **Zhonghao Li**[3], **Xuming Hu**[3†], **Qing Li**[2], **Guandong Xu**[1,4†]
[1]The University of Technology Sydney, [2]The Hong Kong Polytechnic University,
[3]The Hong Kong University of Science and Technology (Guangzhou),
[4]The Education University of Hong Kong
`sirui.huang@student.uts.edu.au`,
`xuminghu97@gmail.com`, `gdxu@eduhk.hk`

Abstract

Large language models (LLMs) have made significant strides in natural language processing by leveraging their ability to comprehend and reason with factual knowledge. However, a significant amount of factual knowledge is stored in structured data, which has unique characteristics not typically encountered in the unstructured texts used for pretraining LLMs. To evaluate the capability of LLMs in handling facts structurally stored, we introduce a benchmark called StructFact, which includes meticulously annotated factual questions, spanning five tasks that reflect the intrinsic properties of structured data. This benchmark aims to delineate the strengths and limitations of LLMs in reasoning with structured data for knowledge-intensive tasks in practical applications. Extensive experiments conducted on 10 common LLMs have yielded several insights, one notable finding being that these models struggle significantly with the heterogeneity of structured data during reasoning. Our data and codes are available[1].

1 Introduction

Large Language Models (LLMs) have revolutionized various downstream natural language processing (NLP) tasks with their impressive capabilities to comprehend and reason on textual data. Previous studies have demonstrated that factual knowledge can be stored within LLMs as a knowledge base, serving knowledge-sensitive tasks such as fact-checking and question-answering (Kojima et al., 2022; Tirumala et al., 2022; Hu et al., 2023). Compared to the traditional method of retrieving knowledge from knowledge bases, reasoning over factual knowledge with LLMs can introduce difficult-to-correct errors due to deviations in inference parameters (Sui et al., 2024). Additionally, LLMs are pretrained on serialized data, overlooking the structural nature of factual knowledge storage, e.g., tables (Hu et al., 2023; Cui et al., 2024). Therefore, effectively using structured data to infer factual knowledge with LLMs remains challenging.

Compared to unstructured data, certain unique characteristics of structured data affect the ability of LLMs to understand and reason about factual knowledge (Fang et al., 2024). These characteristics include: (1) *Heterogeneity*. structured knowledge consists of diverse data types (e.g., texts, numerics, dates). Misunderstandings or biases of any type can lead to inaccuracies in the factual knowledge. (2) *Topological Interdependencies*. Most LLMs are based on the Transformer architecture (Vaswani et al., 2017) and are trained with a next-word prediction loss objective, primarily designed to process continuous text data. Extracting relevant interdependencies from complex topological structures is a significant challenge for LLMs in understanding and reasoning about facts. (3) *Order Invariance*. A key assumption in pretraining is that the order of words significantly impacts their semantics (Chen et al., 2024b). However, in structured data, the permutation of entities (e.g., rows or columns in a table) does not alter the underlying factual knowledge. (4) *Sparsity*. To maintain the same performance in sparse structured data (e.g., missing values or incomplete descriptions) as in data-rich scenarios, LLMs need to accurately utilize the general knowledge learned during pretraining and avoid non-factual imputations. (5) *Lack of Prior Knowledge*. Structured data holds domain-specific knowledge not exposed during pretraining, challenging the accurate application of general reasoning to downstream tasks without distortion (Colon-Hernandez et al., 2021; Zhao et al., 2023a; Li et al., 2024a). These characteristics of structured data impact the ability of LLMs to reason about factual issues, limiting their real-world applications, especially in high-risk domains such

---

[*] Both authors contributed equally to this research.
[†] Corresponding authors.
[1] https://github.com/EganGu/StructFact

Findings of the Association for Computational Linguistics: ACL 2025, pages 7521–7552
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

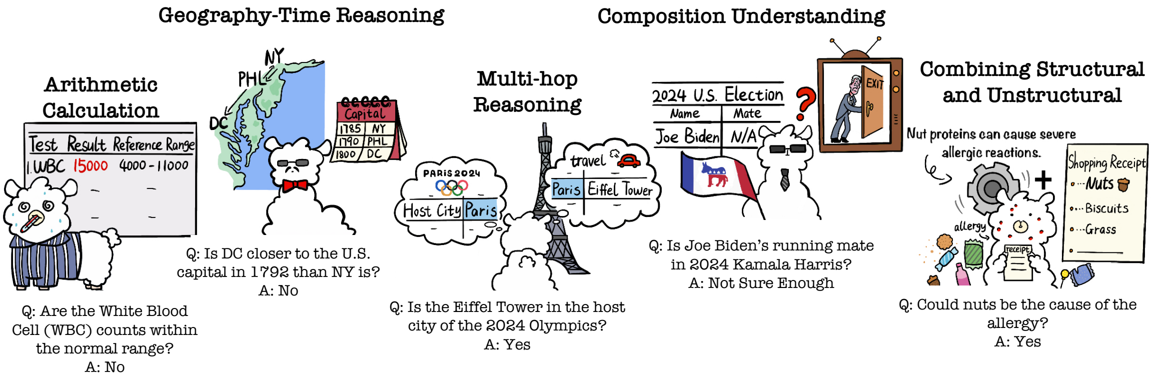

Geography-Time Reasoning
Composition Understanding
Arithmetic Calculation
Test Result Reference Range
1. WBC 15000 4000-11000
Q: Are the White Blood Cell (WBC) counts within the normal range?
A: No
NY
PHL
DC
Capital
1785 NY
1790 PHL
1800 DC
Q: Is DC closer to the U.S. capital in 1792 than NY is?
A: No
Multi-hop Reasoning
PARIS 2024
Host City Paris
travel
Paris Eiffel Tower
Q: Is the Eiffel Tower in the host city of the 2024 Olympics?
A: Yes
2024 U.S. Election
Name Mate
Joe Biden N/A
EXIT
Q: Is Joe Biden's running mate in 2024 Kamala Harris?
A: Not Sure Enough
Combining Structural and Unstructural
Nut proteins can cause severe allergic reactions.
Shopping Receipt
Nuts
Biscuits
Grass
allergy
receipt
Q: Could nuts be the cause of the allergy?
A: Yes

Figure 1: StructFact evaluates LLMs' reasoning ability over structured factual knowledge across five tasks.

as healthcare and finance. To enable LLMs to effectively utilize knowledge embedded in structured data and enhance reliable reasoning, it is essential to examine their capabilities based on the specific characteristics of structured data.

In light of these characteristics, we specifically focus on the structural challenge posed in previous works represented by the Pinocchio benchmark (Hu et al., 2023), and analyze the reasoning capabilities of LLMs on structured data from the perspective of five factual tasks: Arithmetic Calculation, Geography-Time Reasoning, Multi-hop Reasoning, Composition Understanding, Combining Structured and Unstructured. We develop StructFact, a benchmark comprising 13,407 factual queries and corresponding evidence in various structures (i.e., tables, lists, and graphs), covering diverse data types, knowledge domains, timeliness, and regions. We categorized these questions into five factual tasks and provided fine-grained difficulty annotations based on the specific focus of each task to facilitate a multifaceted analysis. Additionally, to specifically test the capability of LLMs to reason over fresh structured facts not encountered during pretraining in real-world applications, we have developed StructFact-Unseen.

Through explorations with StructFact, we examine how 10 commonly used LLMs understand and reason with factual knowledge stored in structured data. For instance, in tasks that rely on *heterogeneous* data, LLMs heavily depend on the order of information for understanding. We aim for StructFact to serve as a compass in exploring the boundaries of LLMs in knowledge-sensitive tasks involving structural facts, while also advancing their practical applications in real-world scenarios.

2 Dataset Construction

To assess LLMs' reasoning abilities over structured facts, we define five tasks based on characteristics of structured data and carefully curate factual questions supported by diverse data types.

2.1 Tasks

Arithmetic Calculation. Given the substantial amount of numerical facts stored in structural data, such as the health report in Figure 1, LLMs are required to perform arithmetic calculations over such *heterogeneity* (i.e., text and numeric) (IBM, 2024; Amazon Web Services, 2024). Models such as Graph Neural Networks (GNNs) seamlessly handle arithmetic calculations by inferring arithmetic rules from numerical patterns through their structural architecture, while LLMs are based on the transformer architecture which is designed for unstructured data. Moreover, structured data containing numbers (e.g., clinical sheets) is usually confidential and *lacks prior inclusion* in the LLMs' knowledge base.To evaluate the capability of handling the *lack of piror knowledge* and *heterogenity*, especially numerical types, StructFact includes factual questions from simple numerical matching to difficult computational analysis.

Geography-Time Reasoning. Geographic and temporal information often appears simultaneously in structured data and can be presented in various formats. Geographical data encompass geographic coordinates (such as latitude and longitude), city names, and country names, whereas temporal data include dates and time periods. This *heterogeneity* challenges LLMs to precisely interpret and align these diverse formats. For example, to determine whether Washington D.C. or New York was closer to the U.S. capital in 1792, as shown in Figure

1, LLMs must combine the year information of the U.S. capital with the geographic coordinates of these cities. To evaluate the proficiency of LLMs in geography-time reasoning, we integrate factual knowledge pertaining to geographical, temporal, and geography-time entities.

Multi-hop Reasoning. Factual knowledge in structural data involves entities dispersed across multiple sources (Li et al., 2023). In Figure 1, the query from the tourist llama involves structured knowledge about the Olympics and travel guides. However, language models typically generate answers by gathering factual knowledge separately, thereby overlooking the *topological interdependencies* (Yang et al., 2023). When gathering data from multiple sources, models should recognize the *order invariance* of structural data. Unlike textual data, which is order-dependent, the order of entities within a similar topological structure should not affect the inherent factual knowledge. In StructFact, we include questions where knowledge is spread across multiple discontinuous structured sources.

Composition Understanding. Reasoning about factual knowledge in structural data suffers from the *sparsity* issue due to missing values or incomplete descriptions. Beyond that, LLMs are expected to accurately reason through these sparse information while without misinterpreting *topological interdependencies*. As Figure 1 displays, to answer the question “Is Joe Biden’s running mate in 2024 Kamala Harris?”, LLMs have to comprehend the header “2024 U.S. Election”, which spans multiple columns with a missing value denoted by “N/A”. To assess whether LLMs comprehend structural compositions, StructFact includes factual questions about components with missing data, complex structures, and incomplete descriptions.

Combining Structured and Unstructured. Given the *sparsity* and *lack of prior knowledge* of the domain-specific information in structured data, LLMs needs to fully leverage the factual knowledge learned from textual contexts. The knowledge presented in unstructured data (e.g., table captions) often provides an important context for understanding the knowledge in structured data. Moreover, the general knowledge base of LLMs aids in reasoning domain-specific knowledge within structured data. As exemplified in Figure 1, general knowledge of nut proteins helps infer the cause of the allergy from the shopping receipt. To assess how structured data combined with their unstructured contexts in LLMs, StructFact includes factual questions that require factuality verification spanning both structured and unstructured evidences.

2.2 Data Construction and Annotation

Fact verification underpins practical applications across diverse fields such as political news and public health (Li et al., 2025a; Si et al., 2024). The knowledge contained within structured data is usually manually curated and is considered a reliable evidence source for verification (Kim et al., 2023). Verifying facts entails analyzing evidence from structured data to determine if the given statement is factual (Fact.), non-factual (Non-fact.) or not enough information (NEI). The evidence in StructFact includes freshly crawled structured knowledge and established structured data sources. Specifically, we include structured data from WiKipedia pages crawled with our developed tool[2], as well as tables from FEVEROUS (Aly et al., 2021), TabFact (Chen et al., 2020b), SQA (Iyyer et al., 2017), and ToTTo (Parikh et al., 2020), lists from FEVEROUS (Aly et al., 2021), and knowledge graphs from FactKG (Kim et al., 2023) and TGQA (Xiong et al., 2024). Additionally, we will periodically update StructFact-Unseen with fresh structured data in our released repository to evaluate newer LLMs.

Binary questions serve as an effective tool for probing the complex capabilities of LLMs (OpenAI, 2023; Burns et al.). For example, (Benara et al., 2024) leverage the yes/no question asked to LLMs to learn an interpretable embedding. With structured data serving as the evidences, we either had 12 English-specialized students craft binary questions or converted existing questions into a binary format. For each piece of structured data that requires newly crafted questions, annotators are first instructed to create one binary question for each of the five tasks. They are then required to assess these questions based on question clarity, answer correctness, and task alignment, scoring each aspect on a scale from 1 to 10. Lastly, we meticulously select questions that have a correct answer and an average score of over 7 for the other two criteria. The answer options provided—Yes, No, and Not Sure Enough—correspond to the categories Fact, Non-fact, and NEI, respectively.

Lastly, the binary questions derived from the structured evidence are annotated with one of five tasks: Arithmetic Calculation, Geography-Time Reasoning, Multi-hop Reasoning, Composition Un-

---

[2] https://anonymous.4open.science/r/WikiCrawler-1B83

derstanding, and Combining Structured and Unstructured Data. Annotators are paid according to the quality and quantity of their annotations. Detailed statistics for StructFact are in Table 2.

2.3 Quality Control

To ensure data quality, we conduct a two-phase verification. First, we divided the 12 annotators into three groups of four to cross-validate a random sample of 500 questions annotated by another group. We calculated the Cohen's Kappa score to quantify the consistency between each pair of annotation groups, resulting in an average score of 0.92. The Fleiss' Kappa score is also used to evaluate the consistency across multiple groups of annotators, yielding a final score of 0.83. Both scores indicate high-quality annotations. Then, two authors randomly select 50 questions per task and carefully review the questions, answers, and task annotations to eliminate duplicates, commonsense contradictions, and other inconsistencies.

3 Main Results

To examine the factual reasoning capabilities of LLMs on structured data, we conduct experiments with StructFact across 10 LLMs trained through pretraining, instruction tuning, and reinforcement learning with human feedback (RLHF). Given the bias of LLMs towards "Yes" answer (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024a,b), we replaced the Fact./Non-Fact./NEI options with A/B/C, respectively[3] and report each model's average performance with respect to weighted accuracy and F1 score over three runs with varied option orders. Detailed descriptions of the prompting strategies and selected LLMs can be found in Figure 13, Appendix F and K.1, respectively.

3.1 Different Prompts

In Table 1, our prompts (see Appendix K.1) adhere to the input formats used in previous studies (Singha et al., 2023; Sui et al., 2024; Wang et al., 2024b), where factual questions from StructFact are combined with corresponding structured data and fed into these LLMs, prompting the models to answer the questions. From the results reported in Table 1, we conclude the following findings.

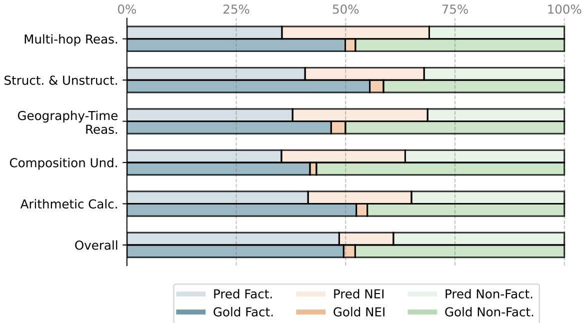


Figure 2: The distribution of three types of responses across five tasks, averaged across 10 LLMs.

- From an overall standpoint, **the ability to comprehend instructions, facilitated by either a larger parameter size or instruction tuning, contributes more significantly to a model's reasoning over structured data than pre-training knowledge alone.** GPT-3.5-turbo, despite its outdated knowledge cutoff in 2021, achieves competitive performance thanks to its large parameter size. Also, models with instruction tuning exhibit superior results compared to the pretrained models. The results obtained by LLaMA-3-8B Instruct, Gemma-2-9B, and Qwen2-7B Instruct outperform their corresponding pretrained models, with an average accuracy improvement of 22.93%. Those pretrained LLMs perform even below the level of random guessing (33.33%) in the zero-shot setting without CoT, as they struggle to follow instructions.
- **For pre-trained models, both the Chain of Thought (CoT; (Wei et al., 2022)) and few-shot strategies helps in utilizing their factual knowledge.** In a zero-shot setting without CoT, the performance of pre-trained models falls below random guessing (with a probability of 33%); incorporating few-shot learning and CoT results in an average F1 score improvement of 5.53%. **The CoT strategy has even negative impact on some instruction-tuned models (e.g., LLaMA-3-8B Instruct), and few-shot examples yield limited improvements.** More complex prompting strategies (Appendix L) also result in modest gains in instruction-tuned models.

3.2 Different Tasks

We further analyze the reasoning performance of the selected 10 LLMs on structured data from the standpoint of the five tasks. From the results in Table 3, we have the following observations.

---

[3] For pretrained LLMs, due to their weak instruction-following capabilities, outputs beyond 'A/B/C' are considered misclassifications, such as a response of 'None', and categorized as False Negatives.

Models	Zero-shot w/o CoT		Zero-shot w/ CoT		Few-shot w/o CoT		Few-shot w/ CoT		Overall	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen2-7B	29.94	37.73	46.77	50.61	44.69	49.51	52.74	56.63	43.53	48.62
LLaMA-3-8B	28.39	33.32	26.66	35.03	26.43	33.22	49.02	50.29	32.63	37.97
Gemma-2-9B	22.83	26.36	43.72	49.58	34.89	40.19	55.97	56.19	39.35	43.08
Qwen2-7B Instruct	46.67	52.40	43.22	51.47	44.18	50.24	43.40	51.95	44.37	51.52
LLaMA-3-8B Instruct	60.62	58.23	43.37	50.19	61.07	59.15	49.54	57.80	53.65	56.34
Gemma-2-9B It	42.66	49.64	47.63	56.97	45.92	53.50	49.78	58.87	46.50	54.75
GLM-4-9B Chat	56.23	59.04	48.44	57.36	54.91	58.15	52.37	60.11	52.99	58.67
Mistral-7B Instruct	47.63	50.45	40.52	49.75	55.78	57.41	46.96	55.49	47.72	53.28
GPT-3.5-turbo	59.84	60.76	59.90	63.94	52.30	56.84	63.91	66.52	59.00	62.02
GPT-4o-mini	**65.12**	**67.08**	**62.96**	**68.11**	**61.44**	**65.03**	**64.96**	**69.58**	**63.62**	**67.45**

Table 1: Performance of 10 LLMs on the StructFact benchmark using various prompts.

Tasks	Distribution			
	Fact.	Non-Fact.	NEI	Overall
Arithmetic Calc.	1,438	1,235	69	2,742
Geography-Time Reas.	1,602	1,717	112	3,431
Multi-hop Reas.	1,694	1,621	79	3,394
Composition Und.	683	925	24	1,632
Struct. and Unstruct.	1,226	913	69	2,208
Total	**6,643**	**5,411**	**353**	**13,407**

Table 2: The statistics of our StructFact benchmark.

- LLMs perform inferior on the tasks of geography-time reasoning and arithmetic calculation, with average weighted F1 scores of 44.93% and 46.85%, respectively. **This underperformance can be attributed to the limitations of LLMs in processing *heterogeneous* evidence**, i.e., reasoning collectively over dates, numbers, and/or texts within structured data sources.
- Among the five factual tasks, LLMs perform relatively well on the tasks of Composition Understanding and Combining Structured and Unstructured data. We have carried out case studies towards these two tasks, with examples provided in Appendix J. After a detailed examination of the cases, we conclude that this benefit stems from **LLMs utilizing their in-context learning and extensive commonsense knowledge to overcome the *sparsity* issue in the structured evidence for these two tasks**.
- We further analyze the distribution of three labels in Figure 2 and observed that the proportions of predicted NEI labels are generally higher than those of the gold labels across the five tasks, the proportions of factual and non-factual responses vary between different tasks. This indicates that, **akin to human behaviors (Tversky and Kahneman, 1981), LLMs demonstrate caution when accepting or rejecting factual queries, when comparing to the ambiguous answer.**

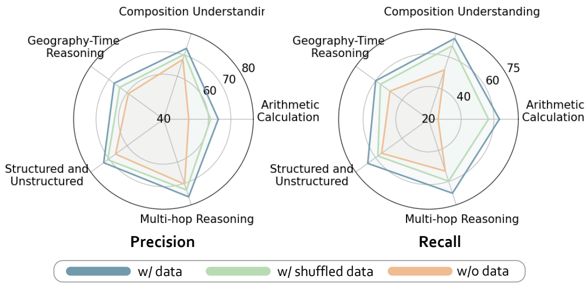


Figure 3: Performance of GPT-4o-mini under different settings of structured evidence.

4 Analysis

Building on the main results, we perform in-depth analyses from various perspectives to assess the LLMs' abilities in completing the five tasks, with GPT-4o-mini (which achieves best in experiments) as the representative model. Analyses of other LLMs are included in Appendix L.

4.1 Resilience to Evidence

We first investigate whether the ability of LLMs to answer factual questions is influenced by the presence of structured evidence. To this end, we categorize the resilience of LLMs to evidence into three levels, ranging from stringent to adaptable: (i) efficiently understanding and reasoning with the provided structured data as evidence, (ii) adapting to irrelevant interventions in the structure of the evidence data, and (iii) accurately recalling prior general knowledge without the support of structured data. We expect LLMs to maintain strong performance across all three levels, indicating remarkable resilience.

To this end, we assess the performance of GPT-4o-mini under three distinct conditions: (i) with structured data provided as corresponding evidence for the factual questions (denoted as "w/ data" in

Models	Arithmetic Calc.		Geography-Time Reas.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen2-7B	27.12	34.60	28.22	34.77	29.31	38.66	31.68	39.51	35.78	43.17
LLaMA-3-8B	27.77	32.25	28.12	31.99	28.07	33.98	28.51	33.91	30.00	34.95
Gemma-2-9B	17.01	23.47	21.30	24.00	25.92	27.56	26.39	27.83	25.10	28.82
Qwen2-7B Instruct	52.41	54.01	41.41	47.41	41.96	49.87	50.67	56.25	52.02	57.76
LLaMA-3-8B Instruct	53.63	46.20	54.66	50.51	**68.00**	<u>68.89</u>	<u>65.70</u>	<u>66.12</u>	<u>63.46</u>	58.73
Gemma-2-9B It	49.37	51.40	34.89	40.88	37.69	47.74	49.28	56.04	49.16	57.38
GLM-4-9B Chat	54.59	52.69	47.74	51.63	60.37	64.53	65.05	65.74	58.56	<u>61.99</u>
Mistral-7B Instruct	52.64	51.87	43.77	46.30	43.36	47.45	46.94	49.79	54.48	57.81
GPT-3.5-turbo	<u>57.70</u>	<u>57.90</u>	<u>57.42</u>	<u>58.82</u>	62.82	63.26	<u>63.79</u>	<u>64.09</u>	<u>58.79</u>	60.28
GPT-4o-mini	**63.93**	**64.15**	**59.92**	**62.96**	<u>67.56</u>	**69.66**	**71.88**	**72.44**	**65.94**	**68.50**
Overall	45.62	46.85	41.75	44.93	46.51	51.16	50.00	53.17	49.33	52.94

Table 3: Performance of 10 LLMs on the StructFact across five tasks in the zero-shot w/o CoT setting.

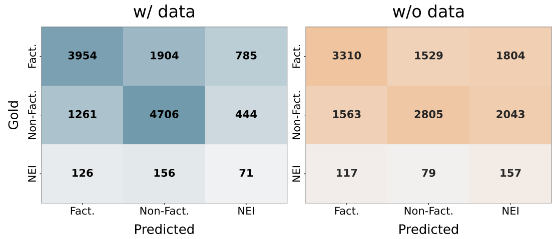


Figure 4: Confusion matrices of performance under the settings w/ and w/o structured data as evidence.

Figure 3), (ii) with the structured data shuffled (denoted as "w/ shuffled data" in Figure 3), and (iii) without any structured data as evidence (denoted as "w/o data" in Figure 3). The first condition aligns with the zero-shot without the CoT setting in the main results (Table 1). In the second condition, we exploit the *order invariance* property of structured data to introduce semantically irrelevant interventions by shuffling the rows and columns in tables and the elements in lists in our StructFact benchmark. For the third condition, since all factual questions in our framework are supported by structured data from Wikipedia, we anticipate that the LLM will rely on its pretraining knowledge to effectively handle scenarios where evidence is absent. We assess the resilience to evidence of the LLM across the five tasks under these three conditions, and depict in Figure 3 and Table 23.

Transitioning from the original structured data (w/ data) to shuffled structured data (w/ shuffled data) results in only a marginal performance decrease, with the largest decline of 3.2% in precision in the Multi-hop Reasoning task, where *order invariance* matters. **This demonstrates the strong adaptability of LLMs to the *order invariance* characteristic of structured data**, and also validates our task categorization.

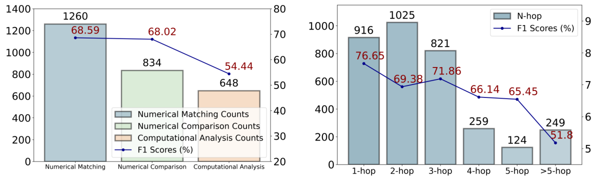
Transitioning from evidence-rich scenarios (w/ data) to the evidence-absence scenario (w/o data), the LLM's performance drops significantly, especially in the tasks of Arithmetic Calculation and Combining Structured and Unstructured, which underscores the *lack of prior knowledge* characteristic. We further analyze this performance drop with the confusion matrices in Figure 4 and attribute it to the uncertainty and a higher tendency for the model to classify facts or non-facts as NEI. This shows that **LLMs do not possess adequate *prior knowledge of facts* contained in real-world structured data.**

4.2 Fine-grained Studies of Different Tasks

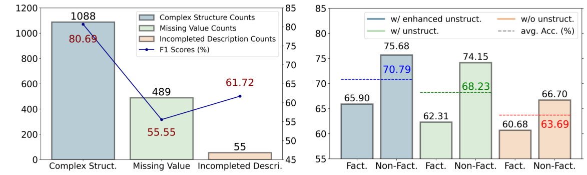
In this section, we fine-grainedly assess GPT-4o-mini's reasoning on structures across five tasks. See Appendix H for the fine-grained categorization defined for each task.

4.2.1 Arithmetic Calculation

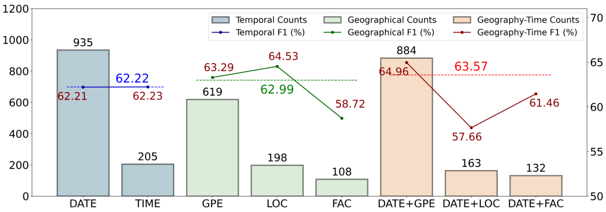
To assess whether large language models (LLMs) are capable of capturing and memorizing arithmetic rules, we categorize the questions in the arithmetic calculation task into three levels of mathematical problems depending on varying degrees of arithmetic difficulty: numerical matching, numerical comparison, and computational analysis. For instance, the factual question "Are the White Blood Cell (WBC) counts within the normal range?" as illustrated in Figure 1, falls under the numerical comparison category. Figure 5(a) presents the performance of GPT-4o-mini across these three categories of mathematical problems. This suggests that **LLMs manage basic numerical tasks like matching and comparison effectively, but struggle with complex computational analyses, such as statistics analysis.**



(a) Arithmetic Calculation (b) Multi-hop Reasoning



(c) Composition Understanding (d) Structured and Unstructured



(e) Geography-Time Reasoning

Figure 5: Fine-grained analysis of performance over the five tasks under zero-shot w/o CoT setting.

4.2.2 Geography-Time Reasoning

As shown in Table 3, LLMs exhibit inadqequate performance in the Geography-Time Reasoning task. We conducted a detailed analysis of GPT-4o-mini's performance across different named entity categories. In Figure 5(e), we classified the Geography-Time Reasoning questions in Struct-Fact into three categories: (i) temporal, which includes questions about dates (DATE), and times (TIME); (ii) geographical, encompassing questions related to political regions such as countries and cities (GPE), as well as locations such as mountains and rivers (LOC), and artificial landmarks (FAC); and (iii) geography-time, which involves questions containing both geographical and temporal entities (DATE+GPE, DATE+LOC, DATE+FAC). Overall, **the LLM performs consistently at understanding and reasoning with geography-time knowledge than with data that only involves temporal or geographical entities**. The varying performance across different entity types suggests that the LLM is more effective with entities that offer detailed granularity in geographical dimensions.

4.2.3 Multi-hop Reasoning

To investigate the capability of LLMs in recognizing and combining knowledge from various discontinuous sources of structured data, we categorized factual questions in the Multi-hop Reasoning task at a more fine-grained level based on the number of hops required to arrive at an answer. A "hop" refers to the step in which the LLM needs to infer knowledge by combining knowledge from two data sources. In particular, in our analysis, each source is defined as a Wikipedia element (e.g., cells, headers, captions in tables, or items in lists) that serves as evidence supporting the gold answer. Figure 5(b) reveals a clear trend: **as reasoning tasks become more complex, requiring an increasing number of hops, the LLMs' effectiveness in reasoning over factual knowledge from structured data diminishes.** Notably, there is a significant performance decline after 5-hop questions, with a 13.65% decrease in F1 score observed in questions with more than 5 hops.

4.2.4 Composition Understanding

To answer whether LLMs can accurately reason factual knowledge from challenging compositions in structured data, we categorize these compositions into three types of irregularities: (i) complex structure, where compositions involve intricate dependencies such as a single table cell spanning multiple columns; (ii) missing values, where cells contain unknown values; and (iii) incomplete descriptions, where cells have ambiguous or insufficient descriptions. Figure 5(c) illustrates illustrates that the primary bottleneck in enhancing LLMs' performance in understanding special composition within structures lies in addressing the challenges of missing values and incomplete descriptions. This challenge is associated with the characteristics of *lack of prior*, indicating that **accurately aligning general knowledge in LLMs with the domain-specific knowledge in structured data remains a significant obstacle for LLMs.**

4.2.5 Combine Structured and Unstructured

A prominent strength of LLMs in factual reasoning is their ability to comprehend knowledge in textual data. When extending this capability to tasks that involve structured data, it becomes imperative to assess whether LLMs can effectively combine factual knowledge extracted from unstructured contexts with reasoning applied to structured data. Therefore, beyond the original unstructured context provided as evidence in the Combining Structured and Unstructured task, we assess the capability of LLMs in scenarios with enhanced unstructured context, as well as in situations where unstructured context is absent. The results shown in Figure 5(d) illustrate that the performance of

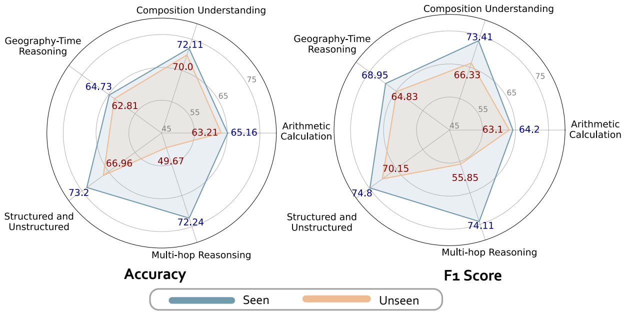


Figure 6: Performance of GPT-4o-mini with structured data seen and unseen during pretraining.

LLMs can be slightly improved by the availability of enhanced contexts when handling factual reasoning over structured data. It is noteworthy that in non-factual tasks, LLMs performed slightly better when provided with the enhanced unstructured context, compared to the original ones. The substantial decrease in performance when unstructured context is absent suggests that **LLMs are particularly dependent on this unstructured context for this task, especially in non-factual circumstances.**

4.3 StructFact-Unseen

To assess the performance of large language models (LLMs) on data not encountered during their pretraining, we constructed StructFact-Unseen to periodically crawl the most recent structured data from new Wikipedia pages. The current StructFact-Unseen dataset, used in this submission, covers the period from December 15, 2024, to January 14, 2025. Please see Table 8 for the pretraining knowledge cutoff dates of each selected LLM.

From Figure 6, it is clear that the language model (LLM) demonstrates a modest reduction in performance when processing structured data not previously encountered during its pretraining phase, compared to data it has been exposed to before. The LLM particularly maintains generalizable performance in tasks challenged by *heterogeneous* data, such as arithmetic calculation and geography-time reasoning. This generalization ability may be attributed to the domain-specific factual knowledge required for these tasks, which cannot be readily deduced using the general knowledge acquired from previously encountered data. Notably, there is a significant decline in performance on the StructFact-Unseen dataset for multi-hop reasoning. This suggests that **LLMs may rely more on their intrinsic knowledge base, rather than the structured evidence provided, to excel in multi-hop reasoning.**

5 Related Work

Extensive benchmarks and empirical studies conducted towards investigating the performance of LLMs over structured data. Form the perspective of instruction-tuning, researchers investigates different prompting strategies in understanding (Guo et al., 2023; Liu et al., 2023a; Tan et al., 2024) and question answering (Qi et al., 2022; Liu et al., 2023b; Slack and Singh, 2023; Bhandari et al., 2024). For example, Liu et al. (2023b) find that integrating textual and symbolic reasoning in TQA can be enhanced by the self-consistency strategy. Moreover, the performance of LLMs also varies across different formats of the structured data. Singha et al. (2023) transforms tables into eight formats, including but not limited to HTML, JSON, and Markdown. From the perspective of data, researchers designs analysis on the basis of self-supervised components in the structures (Singha et al., 2023; Zhao et al., 2023b; Sui et al., 2024). For example, Zhao et al. (2023b) conduct component-level perturbation (e.g, column adding), Sui et al. (2024) investigate via seven structural understanding tasks (e.g., cell lookup and row retrieval) based on different components. Additionally, LLM capability evaluations consider structured knowledge from the scientific domains (Ghosh et al., 2024; Wang et al., 2021) to the financial domain (Hollmann et al., 2025). Different from existing works, StructFact offers deeper insights by annotating tasks based on characteristics inherent in structured data. We further compare StructFact with other public datasets containing structured knowledge in Appendix B.

6 Conclusions

In this work, we present StructFact, a benchmark specifically developed to assess the factual reasoning abilities of LLMs over structured data. StructFact comprises 13,407 questions spanning five factual tasks built upon the characteristics of structured data. We tested 10 LLMs with StructFact and observed that they struggle with reasoning over *heterogeneous* structured data, especially in complex arithmetic operations. Additionally, their diminished resilience to evidence highlights the ongoing challenge of effectively leveraging their knowledge base for reasoning with structured facts. Our work underscores the pressing need to develop advanced techniques that help LLMs to better comprehend and reason over structured data. With the vast wealth of structured knowledge, we believe this will enhance reliable reasoning and pave the way for broader uses of LLMs in the future.

7 Limitations

While this study offers a valuable benchmark for evaluating LLMs' reasoning capabilities with structured data, it presents this data in a uniform format—markdown for tables and triplets for graphs. The omission of other structured formats such as JSON and XML could limit the generalization and applicability of StructFact. Second, the uneven distribution of fine-grained labels in each task could affect the analysis. Additionally, the lack of intentionally corrupted data could limit the persuasiveness of the resilience analysis in section 4.1.

Acknowledgement

This project has been supported by the Hong Kong Research Grants Council under the General Research Fund (project no. 15200023); Research Impact Fund (project no. R1015-23), the National Natural Science Foundation of China under Grants No.62072257; the Australian Research Council Grants DP22010371, LE220100078; the Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); CAAI-Ant Group Research Fund, Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (No.2025A03J3957), Education Bureau of Guangzhou Municipality.

References

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Inc. Amazon Web Services. 2024. What is structured data?

Vinamra Benara, Chandan Singh, John Xavier Morris, Richard Antonello, Ion Stoica, Alexander Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings for language neuroscience by asking LLMs questions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kushal Raj Bhandari, Sixue Xing, Soham Dan, and Jianxi Gao. 2024. On the robustness of language models for tabular question answering. *arXiv preprint arXiv:2406.12719*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Nuo Chen, Yuhan Li, Jianheng Tang, and Jia Li. 2024a. Graphwiz: An instruction-following language model for graph computational problems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 353–364, New York, NY, USA. Association for Computing Machinery.

Pei Chen, Soumajyoti Sarkar, Leonard Lausen, Balasubramaniam Srinivasan, Sheng Zha, Ruihong Huang, and George Karypis. 2024b. Hytrel: Hypergraph-enhanced tabular data representation learning. *Advances in Neural Information Processing Systems*, 36.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W Cohen. 2020a. Open question answering over tables and text. In *International Conference on Learning Representations*.

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

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

Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*.

Lingxi Cui, Huan Li, Ke Chen, Lidan Shou, and Gang Chen. 2024. Tabular data augmentation for machine learning: Progress and prospects of embracing generative ai. *arXiv preprint arXiv:2407.21523*.

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

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Akash Ghosh, B Venkata Sahith, Niloy Ganguly, Pawan Goyal, and Mayank Singh. 2024. How robust are the tabular qa models for scientific tables? a study using customized dataset. *arXiv preprint arXiv:2404.00401*.

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, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Jiayan Guo, Lun Du, and Hengyu Liu. 2023. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021a. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. *arXiv preprint arXiv:2103.12011*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

N. Hollmann, S. Müller, L. Purucker, et al. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, S Yu Philip, and Zhijiang Guo. 2023. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

IBM. 2024. Structured versus unstructured data: What's the difference?

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

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. Factkg: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025a. Loki: An open-source tool for fact verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 28–36, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and ZHAO-XIANG ZHANG. 2024a. Sheetcopilot: Bringing software productivity to the next level through large language models. *Advances in Neural Information Processing Systems*, 36.

Xiao Li, Yawei Sun, and Gong Cheng. 2021. Tsqa: tabular scenario based question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13297–13305.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Yuhan Li, Peisong Wang, Xiao Zhu, Aochuan Chen, Haiyun Jiang, Deng Cai, Victor Wai Kin Chan, and Jia Li. 2024b. GLBench: A comprehensive benchmark for graph with large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S Yu. 2023a. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *arXiv preprint arXiv:2303.13547*.

Tianyang Liu, Fei Wang, and Muhao Chen. 2023b. Rethinking tabular data understanding with large language models. *arXiv preprint arXiv:2312.16702*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

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

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

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Qianqian Qi, Zhenyun Deng, Yonghua Zhu, Lia Jisoo Lee, Michael Witbrock, and Jiamou Liu. 2022. TaKG: A new dataset for paragraph-level table-to-text generation enhanced with knowledge graphs. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 176–187, Online only. Association for Computational Linguistics.

Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal fact verification via argument structure. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15636–15659, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Dylan Slack and Sameer Singh. 2023. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188*.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web*

Search and Data Mining, WSDM '24, page 645–654, New York, NY, USA. Association for Computing Machinery.

Zhao Tan, Xiping Liu, Qing Shu, Xi Li, Changxuan Wan, Dexi Liu, Qizhi Wan, and Guoqiong Liao. 2024. Enhancing text-to-SQL capabilities of large language models through tailored promptings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6091–6109, Torino, Italia. ELRA and ICCL.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Haochen Wang, Kai Hu, Haoyu Dong, and Liangcai Gao. 2024a. Doctabqa: Answering questions from long documents using tables. In *International Conference on Document Analysis and Recognition*, pages 470–487. Springer.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin

Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024b. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang. 2023. Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5223–5234, New York, NY, USA. Association for Computing Machinery.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023a. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios. *arXiv preprint arXiv:2305.14987*.

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

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024a. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024b. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Seq2SQL: Generating structured queries from natural language using reinforcement learning.

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

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

Appendix

A Implementation Details

We use 32GB memory with Ubuntu 20.04 LTS (a open-source Operating System using the Linux kernel and based on Debian) and 4 Nvidia A800 with 80GB memory for inference. we adopt vllm (Kwon et al., 2023) 0.5.4 to speed up inference. All models share a set of hyperparameters, as detailed in Table 4.

Hyperparameter	**Value**
`top_p`	0.95
`temperature`	0.6
`max_generation_token (w/o CoT)`	10
`max_generation_token (w/ CoT)`	512
`max_evidence_token`	2500

Table 4: Hyperparameters of LLMs

B Comparisons to Other Datasets and Empirical Studies

We compare our proposed StructFact to a wide range of publicly available datasets that contain structured factual knowledge, summarized in Table 5. Additionally, we conduct a comparison with the analyses in previous empirical studies of LLMs from the perspective of five distinct characteristics of structured data in Table 6. It can be found in these two tables that our StructFact provides a comprehensive benchmark for testing the capabilities of LLMs on structured factual knowledge.

C Future Directions

Apart from our findings, we propose future directions to advance LLM use in knowledge-sensitive tasks involving structured data.

The performance of LLMs diminishes as the availability of structured evidence shifts from abundant to absent. Given the limited improvements achieved through prompt engineering on instruction-tuned models, incorporating an additional structure-aware module may offer a more effective approach for learning from structured data. Such specialized modules enable task-adaptive learning and knowledge transfer while keeping computational and time costs manageable. Furthermore, LLMs show considerable potential in leveraging unstructured knowledge to complement structured data. A key challenge in this process is avoiding the distortion of precise knowledge contained within the structured data. Future research could explore the use of reinforcement learning to iteratively correct distortions in reasoning.

D Generation Randomness

To assess the randomness of generation, we conducted zero-temperature experiments on the 10 LLMs and reported the results in Table 7.

E Evaluation Protocol

In this paper, we use six different metrics for evaluating the reasoning performance of LLMs on structured knowledge. We formulate all the evaluation metrics used in this section.

- Accuracy.

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, FN represent the number of true positive, true negative, false positive, and false negative, respectively.

- Weighted F1 score.

$$F1 = \sum_{i=1}^{N} \frac{n_i}{N} F1_i \tag{2}$$

where n_i is the number of samples in label i, N is the number of all samples, $F1_i$ is the F1 score for label i.

- Balanced accuracy.

$$BA = \frac{1}{N} \sum_{i=1}^{N} (TPR_i), TPR = \frac{TP}{TP + FN} \tag{3}$$

where TPR_i is the true positive rate of label i.

- Macro F1 score.

$$MacroF1 = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{4}$$

- Precision.

$$Prec. = \frac{TP}{TP + FP} \tag{5}$$

- Recall.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

F Detailed Introduction to selected LLMs

Meta's Llama series, including Llama 2 and Llama 3 (2023), released in 2023 and 2024, are designed for various tasks like text generation and programming. Llama3 is designed to be more intelligent,

Dataset	Task	Source	Evidence/Data Type	Answer Type	Domain
ToTTo (Parikh et al., 2020)	Generation	Wikipedia	Table		General
TaKG (Qi et al., 2022)	Generation	Wikipedia	Table, Graphs, Text		General
WebNLG (Gardent et al., 2017)	Generation	DBPedia	Graphs		General
DART (Nan et al., 2021)	Generation	Wikipedia	Table, Graphs		General
LOTNLG (Zhao et al., 2023a)	Generation	Wikipedia	Table		General
SQA (Iyyer et al., 2017)	Question Answering	Wikipedia	Table	Span	General
NQ-tables (Herzig et al., 2021)	Question Answering	Wikipedia	Table	Span	General
HybridQA (Chen et al., 2020c)	Question Answering	Wikipedia	Table, Text	Span	General
WikiTableQuestion(WTQ) (Pasupat and Liang, 2015)	Question Answering	Wikipedia	Table, Text	Span	General
FetaQA (Nan et al., 2022)	Question Answering	Wikipedia	Table, Text	Span	General
TAT-QA (Zhu et al., 2021)	Question Answering	Wikipedia	Table, Text	Span	General
Open-WikiTable (Kweon et al., 2023)	Question Answering	Wikipedia	Table, Text, SQL	Span	General
WebSRC (Chen et al., 2021)	Question Answering	Web pages	HTML	Span/Boolean	General
OTTQA (Chen et al., 2020a)	Question Answering	Wikipedia	Table, Text	Multiple Choice	General
MATH (Hendrycks et al., 2021a)	Question Answering	Exam	Text	Span	Mathematics
GSM8K (Cobbe et al., 2021)	Question Answering	Exam	Text	Span	Mathematics
TSQA (Li et al., 2021)	Question Answering	Exam	Table	Multiple Choice	Geography
ROBUT (Zhao et al., 2023b)	Question Answering	WTQ, WIKISQL-WEAK (Zhong et al., 2018), and SQA	Table	Span	General
SUC (Sui et al., 2024)	Question Answering	Wikipedia	Table	Span	General
MiMoTable (Li et al., 2025b)	Question Answering	Baidu Wenku, Google Search	Table	Span	General
TableBench (Wu et al., 2024)	Question Answering	Wikipedia	Table	Span	General
DocTabQA (Wang et al., 2024a)	Question Answering	Securities and Exchange Commission	Table	Span	Finance
FREB-TQA (Zhou et al., 2024)	Question Answering	WTQ,WikiSQL,SQA,TAT (Zhu et al., 2021)	Table	Span	General
SciTabQA (Ghosh et al., 2024)	Question Answering	SciGen (Moosavi et al., 2021)	Table	Span	General
GraphWiz (Chen et al., 2024a)	Question Answering	Self-developed	Graph	Span	N/A
GLBench (Li et al., 2024b)	Question Answering	citation networks, web links, and social networks	Graph	Span	General
FEVEROUS (Aly et al., 2021)	Fact-checking	Wikipedia	Table, List, Text	Fact/Non-Fact/NEI	General
TabFact (Chen et al., 2020b)	Fact-checking	Wikipedia	Table	Fact/Non-Fact	General
Infotabs (Gupta et al., 2020)	Fact-checking	Wikipedia	Table	Fact/Non-Fact/NEI	General
Fact-KG (Kim et al., 2023)	Fact-checking	WebNLG (Gardent et al., 2017),DBPedia	Graphs	Fact/Non-Fact	General
Semeval 2021 Task 9 (Wang et al., 2021)	Fact-checking	Scientific Articles	Table, Text	Fact/Non-Fact/NEI	Science
PubHealthTab (Akhtar et al., 2022)	Fact-checking	PubHealth (Kotonya and Toni, 2020), Wikipedia	HTML	Fact/Non-Fact	Healthcare
StructFact	Fact-checking	Wikipedia	Table, List, Graphs, Test	Fact/Non-Fact/NEI	General

Table 5: A comprehensive comparison of various **benchmarks** with structural facts.

Dataset	**Characteristics of Structured Data**					**Other Research Question**
	Heterogeneity	*Topological Interdependencies*	*Order Invariance*	*Sparsity*	*Lack of Prior Knowledge*	
Zhao et al. (2023b)			✓	✓		
Sui et al. (2024)		✓	✓			
Liu et al. (2023b)			✓			Symbolic reasoning
Zhou et al. (2024)	✓		✓			Question-related components
Ghosh et al. (2024)	✓					Scientific reasoning
Zhao et al. (2023a)	✓					Table-to-text capability
Li et al. (2025b)						reasoning difficulty, different domain
Wu et al. (2024)	✓	✓	✓			Prompting strategies
(Singha et al., 2023)			✓			Different formats
Guo et al. (2023)			✓	✓		Prompting strategies
Chen et al. (2024a)		✓	✓			Instruction following
Li et al. (2024b)	✓					Prompting strategies
StructFact	✓	✓	✓	✓	✓	

Table 6: Distribution of various characteristics of structured data across existing **empirical studies** that investigate the capabilities of LLMs.

Methods	Zero-shot w/o CoT		Zero-shot w/ CoT		Few-shot w/o CoT		Few-shot w/ CoT		Overall	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen2-7B	38.12	48.51	45.21	48.34	52.15	55.89	50.18	53.97	46.41	51.68
LLaMA-3-8B	27.89	32.01	25.98	33.82	28.76	35.04	47.33	48.71	32.49	37.40
Gemma-2-9B	31.45	34.77	42.06	47.12	43.15	47.92	53.14	54.32	42.45	46.03
Qwen2-7B Instruct	46.30	52.15	43.01	50.26	44.67	50.83	42.95	51.18	44.23	51.11
LLaMA-3-8B Instruct	60.88	58.55	42.05	49.73	60.32	58.97	49.12	57.32	53.09	56.14
Gemma-2-9B It	43.01	49.27	47.21	56.05	45.12	53.06	49.32	58.01	46.17	54.10
GLM-4-9B Chat	56.85	59.57	47.93	56.84	54.45	57.68	51.84	59.43	52.77	58.38
Mistral-7B Instruct	48.01	50.12	39.87	49.12	55.15	56.82	46.35	54.78	47.34	52.71
GPT-3.5-turbo	59.92	60.96	59.02	63.11	52.37	57.00	63.01	66.22	58.58	61.82
GPT-4o-mini	65.32	67.41	62.15	67.58	61.89	65.32	64.48	69.12	63.46	67.36

Table 7: Performance of 10 LLMs on StructFact using various prompts (temperature=0).

faster, and more versatile, making it suitable for a wide range of applications. Qwen2 (2024a) (2024b) is a strong language models developed by Alibaba Cloud, showing state-of-the-art performance in several benchmarks, especially in coding and mathematics. ChatGLM3 (2024) is the latest generation of pre-trained dialogue models developed by Zhipu AI in collaboration with Tsinghua University's Knowledge Engineering Group (KEG). Developed by OpenAI, GPT-4o-mini (2024) is its most cost-efficient small model in the GPT series, featuring enhanced context understanding and text generation capabilities, scoring 82% on MMLU (Hendrycks et al., 2021b). Gemma2 (2024) is Google's latest iteration of open large language models (LLMs), building on the success of the original Gemma series. Coming with two sizes, 9 billion and 27 billion parameters, each size has a base model (pre-trained) and an instruction-tuned version.

Model Name	**Company**	**Cut-off Date**	**Release Date**
Qwen2-7B	Alibaba	2023	February 2024
LLaMA-3-8B	Meta	March 2023	April 2024
Gemma-2-9B	Meta	-	July 2024
Qwen2-7B-Instruct	Alibaba	2023	February 2024
LLaMA-3-8B Instruct	Meta	March 2023	April 2024
Gemma-2-9B It	Meta	-	July 2024
GLM-4-9B Chat	Zhipu AI	-	June 2024
Mistral-7B Instruct	Mistral AI	-	September 2023
GPT-3.5 Turbo	OpenAI	September 2021	November 2022
GPT-4o mini	OpenAI	October 2023	May 2024

Table 8: Table of the selected LLMs, companies, knowledge cut-off dates, and released dates.

G Ethical Statement

We affirm that our StructFact benchmark is constructed using open-source datasets and adheres to the CC-BY-4.0 license. To uphold privacy and confidentiality, we have ensured that our dataset contains no direct or indirect sensitive personal information. Users accessing our StructFact should ensure that no personally identifiable information or toxic content is included.

Our research postulate that our StructFact benchmark is under an environment devoid of possible attacks. However, given that the structured data in our proposed benchmark is sourced from publicly editable WikiPedia pages, it is inherently vulnerable to various threats, including adversarial attacks. Intended attacks, such as data poisoning, involve malicious actors deliberately inserting false or misleading information or altering existing structured data. These actions can compromise the integrity of the data, distorting the knowledge within LLMs and undermining accurate factual reasoning. Unintentional attacks, such as accidental data deletion or incorrect data entry, also pose significant risks. These errors can degrade both the quality and structure of the data, potentially leading LLMs to draw incorrect inferences, thus might compromising the overall factuality of the benchmark.

Moreover, while the questions in our StructFact benchmark reflect real-world facts, they do not originate from practical applications. Therefore, we offer StructFact as a resource to guide users in their inferences, without claiming to provide absolute assertions. We advise against using StructFact as a basis for developing models intended to verify facts in real-world applications.

H Task Categorization for Fine-Grained Analysis

To perform a fine-grained analysis of LLMs' reasoning capabilities over structured data, we refined each major task into specific subcategories. This section outlines the methodologies employed for task refinement.

H.1 Arithmetic Calculation

Following Hendrycks et al. (2021a); Chen et al. (2020b), we refined the *Arithmetic Calculation* task into three distinct subcategories based on the complexity and nature of mathematical reasoning required:

1. **Numerical Matching**: Tasks involving the direct matching and verification of a specific value within the data, typically a precise, singular number, without any calculations or statistical analyses.
    - *Example*: "Is the number of gold medals won by China in the 2024 Olympics 22?"
2. **Numerical Comparison**: Tasks requiring the comparison of numerical values to determine size, quantity, or order, including comparisons involving superlatives.
    - *Example*: “Did China win more gold medals than any other country in the 2024 Olympics?”
3. **Computational Analysis**: Tasks necessitating calculations, statistical methods, or logical reasoning to perform aggregations, averages, percentage calculations, or other mathematical analyses. This includes basic arithmetic operations (addition, subtraction, multiplication, and division) and conditional judgments.
    - *Example 1*: “Does the percentage of gold medals won by China exceed 10% of the total in the 2024 Olympics?”
    - *Example 2*: “Did China win 4 more gold medals than France in the 2024 Olympics”

To annotate the dataset according to these refined categories, we established precise definitions and guidelines for each subcategory. We then employed GPT-4o for automatic labeling of the tasks, followed by independent reviews from three experts. Discrepancies identified by the majority of experts were manually corrected. This combination of automated labeling and expert validation ensured both efficiency and high accuracy, aligning with best practices in semi-automated annotation workflows.

H.2 Geography-time Reasoning

We focused on identifying and categorizing named entities related to temporal and spatial information within the questions. Utilizing the Named Entity Recognition (NER) tool from SpaCy, specifically the `en_core_web_sm` model[4], we automatically recognized entities in the text. Based on the NER tags assigned by SpaCy, we refined the questions into:

- **Temporal Entities**: Questions containing temporal entities such as dates (`DATE`) and times (`TIME`).
- **Spatial Entities**: Questions involving spatial entities, including geopolitical entities (`GPE`), natural locations like mountains and rivers (`LOC`), and facilities or artificial landmarks (`FAC`).
- **Spatiotemporal Entities**: Questions that contain both temporal and spatial entities, such as combinations of `DATE+GPE`, `DATE+LOC`, or `DATE+FAC`.

H.3 Multi-hop Reasoning

Multi-hop Reasoning refers to the process of linking and integrating information from multiple sources or steps to arrive at a final answer (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). It requires the system to go beyond simple, one-step reasoning (single-hop) by making logical connections across various pieces of information that are distributed across multiple documents, sentences, or structured data points.

In our work, we refined the *Multi-hop Reasoning* task by categorizing questions based on the number of reasoning steps, or "hops," required to derive the correct answer. Each "hop" is defined as a step where the model must integrate information from two distinct data sources within the structured data, such as table cells, headers, captions, or list items. Using a rule-based method, we classified the questions into six categories: 1-hop, 2-hop, 3-hop, 4-hop, 5-hop, and greater than 5 hops. This systematic classification allowed us to assess the complexity of the reasoning required and analyze the model's capacity to perform multi-hop reasoning over structured data.

H.4 Composition Understanding

Inspired by Singha et al. (2023), We examined the model's ability to reason over structured data compositions with irregularities. Specific categorization rules were defined to identify different types of compositional challenges:

---

[4] `https://spacy.io/usage/models`

1. **Complex Structures**: Compositions involving intricate dependencies, such as a single table cell spanning multiple columns or rows, nested tables, or hierarchical lists.
2. **Missing Values**: Instances where the structured data contains unknown or missing values, requiring the model to handle incomplete data.
3. **Incomplete Descriptions**: Cases where the structured data includes ambiguous or insufficient descriptions, such as unclear column headers or labels.

Two experts manually annotated the dataset according to these categories and conducted a cross-review to ensure consistency and resolve discrepancies. This rigorous annotation process enhanced the reliability of our categorization.

H.5 Combining Structured and Unstructured Data

Given the wide existence of semi-structured data (Chen et al., 2020c; Pasupat and Liang, 2015; Nan et al., 2022), the *Combining Structured and Unstructured Data* task aimed to assess the model's ability to integrate information from both structured data (e.g., tables) and accompanying unstructured textual context. We designed three experimental conditions:

- **Original Unstructured Context**: Using the original unstructured context provided in the dataset alongside the structured data.
- **Enhanced Unstructured Context**: Augmenting the unstructured context by generating additional descriptions of the structured data using GPT-4o, including details such as table formats and relevant contextual information to provide extra background knowledge.
- **No Unstructured Context**: Removing all unstructured context, leaving only the structured data for the model to process.

By varying the availability and richness of unstructured context, we aimed to analyze how different levels of contextual information affect the model's ability to reason over structured data.

I Model Confidence

Towards reliable reasoning outcomes, we also concern about the confidence of LLMs' outputs. Specifically, We used the probability of the model's first token output corresponding to the predicted label as the confidence score. Figure 7 illustrates that the model shows higher confidence with its predicted facts, compared to those are predicted as non-factual. For NEI responses, whether predicted correctly or not, the confidence remains low with only minor variations. Notably, the models exhibit uncertainty in their incorrect answers, as evidenced by a significant drop in confidence, especially for factual and non-factual predictions. This indicates that the model's confidence levels are somehow aligned with the accuracy of their predictions, which could be useful for gauging the reliability of the model's answer.

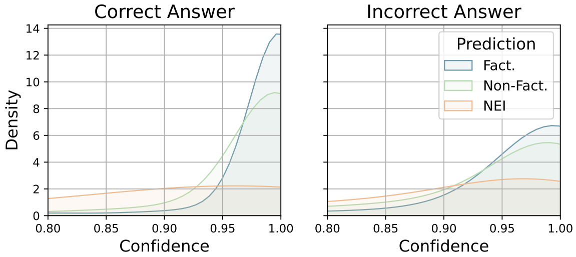


Figure 7: Confidence distribution of answers.

J Case Study

Please see figures 8 to 12 for case studies for each task and the responses from different LLMs.

K Prompt Strategies Analysis

K.1 Detailed Introduction to Employed Prompts

Each LLM in our main result decipted in Table 1 is experimented with different prompting strategies: Zero-shot without CoT (Kojima et al., 2022), Zero-shot with CoT, Few-shot with CoT, Few-shot with CoT. All the strategies used in this paper begin with an instruction denoted as $p =$ "*You will be given with a question. Please response with 'Yes', 'No', or 'Not Sure Enough'.*" For any input question $q_i \in \mathcal{Q}$, structural data $d_i \in \mathcal{D}$ the model $LLM(\cdot)$ is expected to generate an answer $y_i \in \mathcal{Y} = \{$`'Yes'`,`'No'`,`'Not Sure Enough'`$\}$. Each question is categorized into one task t from the five aforementioned reasoning tasks in $\mathcal{T}$. Examples of the prompts used in our experiments are shown in Figure 14.

K.1.1 Prompts in Main Results

Prompt with Zero-shot. In the prompting strategy with zero-shot setting, the LLM is expected to output the answer y_i to the question q_i directly, formally, $y_i = LLM(p, q_i, d_i)$. For example, the factual answer y_i = `"No"` should be responded from the LLMs when being asked with the question q_i = `"Is London-` `-the host city of the 2024 Olympic Games?"`, together with the table of Olympic Games host cities denoted by d_i.

Prompt with Few-shot. In the few-shot prompting strategy, to guide the LLM to correctly reason, we include an example question q_x and structural data d_x together with prompt p for question q_i, where the example question q_x and question q_i fall in the same task, i.e., $q_x, q_i \in t$. This process is formulated as $y_i = LLM(p \| q_x \| d_x, q_i, d_i)$. The LLM is expected to answer with y_i = `"Yes"` when given question q_i = `"Has Paris hosted the Olymp-` `-ic Games three times?"` and the table of Olympic Games host cities d_i.

Prompt with Chain of Thought (CoT). In the prompting strategy with CoT (Kojima et al., 2022), a two-stage prompt is employed to derive the reasoning process along with with the answer. To guide the LLM in carefully considering the process of determining the answer y_i, the prompting sentence s = `"Let's think step by step"` is added to the question q_i, formally, $y_i = LLM(p, q_i \| s, d_i)$.

K.1.2 Prompts in Evidence Resilience Analysis

Prompt with Shuffled Structured Data. To investigate the performances of LLMs towards different prompting context, we shuffle the structure of data. Specifically, we shuffle the rows/columns in tables, and the elements in lists. Formally, for question q_i, the output can be presented as $y_i = LLM(p, q_i, d_i')$, where d_i' denotes the shuffled data.

Prompt without Structured Data. Given that the structural data is sourced from Wikipedia, it is assumed that LLMs have been exposed to these data during their training phase. Therefore, we are also interested in the ability of LLMs to answer factual questions q_i without being provided with the contextual structural data d_i. The process under this strategy can be formulated as $f_5 : y_i = LLM(p, q_i)$.

Prompt with self-refinement. The self-refinement strategy is designed to enhance the performance of LLMs by prompting them to iteratively providing feedback to its previous responses. Formally, the process at n-th round of refinement can be presented as $y_i^n = LLM^n(p, q_i, d_i, r_i^{n-1})$, where r_i represents the LLM's response in the last round. In our experiments, due to constraints on computing resources and time, we set n=1.

Prompt with self-consistency. The self-consistency strategy is designed to enhance the performance of LLMs by employing majority voting on multiple rounds of queries. Assume the response from the model at the n-th round as y_n, the final prediction of LLM can be formualted as $y_{final} = argmax_{c_j} \sum_{j=1}^{k} counts(y_n = c_j)$, where c denotes the available choices of the prediction label, i.e., 'Fact.', 'Non-Fact.', and 'NEI' in this paper.

Prompt with format instructions. We also provide instructions of the formats of the structured data to the zero-shot prompts. Given format instructions as f, which illustrates how the structured data looks like, the process can be formulated as $y_i = LLM(p \| f, p_i, d_i)$.

K.2 Analysis towards Other Prompting Strategies

Given the successes of other CoT strategies and input data format instructions (Slack and Singh, 2023), we are interested in exploring their impact on reasoning about factual knowledge within structured data. We include three prompting strategies: (i) self-refinement (Madaan et al., 2024), which guides the LLM to iteratively evaluate and refine its previous responses to reach the correct answer, (ii) self-consistency (Wang et al., 2023), which mitigates hallucination through majority voting on multiple responses from the LLM, and (iii) format instructions, which prompts with descriptions of the format of the inputted structured data. There are the following notable observations from the results in Table 9. i) Self-consistency marginally improves performance across five tasks, with an overall enhancement of 0.23%, compared to the zero-shot results without CoT in Table 1. ii) Format descriptions help the LLM better interpret numerical compositions, leading to a 1.02% improvement in accuracy on Arithmetic Calculation tasks. Detailed results under these strategies please refer to Table 9.

L Supplementary Results

L.1 Comprehensive Results

Results for Different Prompts under Other Metrics:

Please refer to Tables 10 and 11.

Results for Different Tasks under Other Metrics:

Please refer to Tables 12 to 22.

L.2 Analysis of Other LLMs

Model Responses Distributions:

Please refer to Figure 15.

Model Resilience to Evidence:

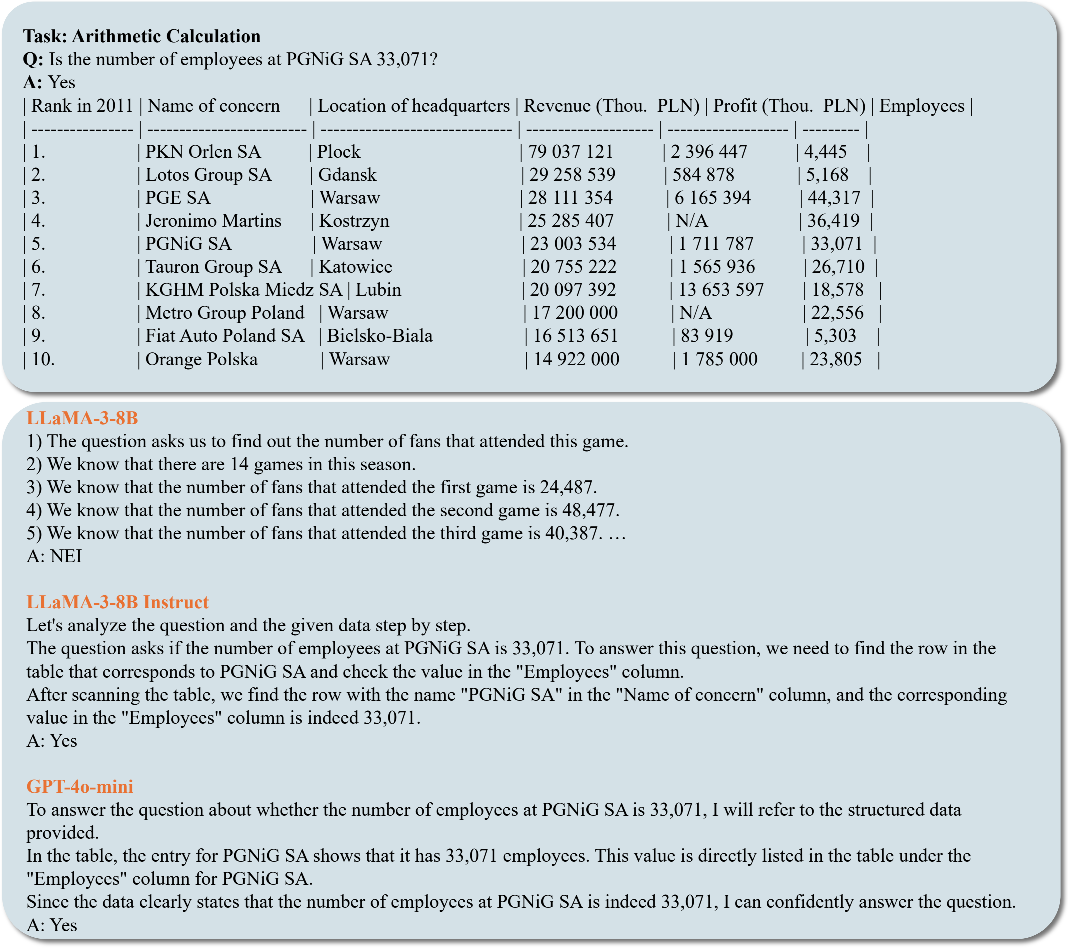
Please refer to Table 23 and Figure 16.

Fine-grained Studies of Different Tasks:

Please refer to Figures 17 to 20.

Model Confidence Analysis:

Please refer to Figure 21.



Task: Arithmetic Calculation
Q: Is the number of employees at PGNiG SA 33,071?
A: Yes

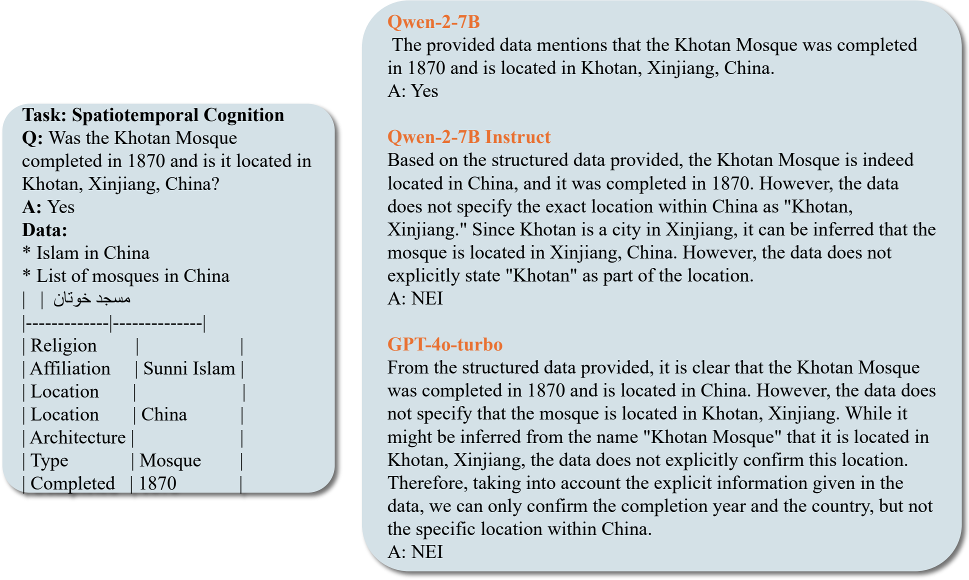
Rank in 2011	Name of concern	Location of headquarters	Revenue (Thou. PLN)	Profit (Thou. PLN)	Employees
1.	PKN Orlen SA	Plock	79 037 121	2 396 447	4,445
2.	Lotos Group SA	Gdansk	29 258 539	584 878	5,168
3.	PGE SA	Warsaw	28 111 354	6 165 394	44,317
4.	Jeronimo Martins	Kostrzyn	25 285 407	N/A	36,419
5.	PGNiG SA	Warsaw	23 003 534	1 711 787	33,071
6.	Tauron Group SA	Katowice	20 755 222	1 565 936	26,710
7.	KGHM Polska Miedz SA	Lubin	20 097 392	13 653 597	18,578
8.	Metro Group Poland	Warsaw	17 200 000	N/A	22,556
9.	Fiat Auto Poland SA	Bielsko-Biala	16 513 651	83 919	5,303
10.	Orange Polska	Warsaw	14 922 000	1 785 000	23,805

LLaMA-3-8B
1) The question asks us to find out the number of fans that attended this game.
2) We know that there are 14 games in this season.
3) We know that the number of fans that attended the first game is 24,487.
4) We know that the number of fans that attended the second game is 48,477.
5) We know that the number of fans that attended the third game is 40,387. …
A: NEI

LLaMA-3-8B Instruct
Let's analyze the question and the given data step by step.
The question asks if the number of employees at PGNiG SA is 33,071. To answer this question, we need to find the row in the table that corresponds to PGNiG SA and check the value in the "Employees" column.
After scanning the table, we find the row with the name "PGNiG SA" in the "Name of concern" column, and the corresponding value in the "Employees" column is indeed 33,071.
A: Yes

GPT-4o-mini
To answer the question about whether the number of employees at PGNiG SA is 33,071, I will refer to the structured data provided.
In the table, the entry for PGNiG SA shows that it has 33,071 employees. This value is directly listed in the table under the "Employees" column for PGNiG SA.
Since the data clearly states that the number of employees at PGNiG SA is indeed 33,071, I can confidently answer the question.
A: Yes

Figure 8: Case Study of Arithmetic Calculation.



Task: Spatiotemporal Cognition
Q: Was the Khotan Mosque completed in 1870 and is it located in Khotan, Xinjiang, China?
A: Yes
Data:
* Islam in China
* List of mosques in China

	مسجد خوتان
Religion	
Affiliation	Sunni Islam
Location	
Location	China
Architecture	
Type	Mosque
Completed	1870

Qwen-2-7B
The provided data mentions that the Khotan Mosque was completed in 1870 and is located in Khotan, Xinjiang, China.
A: Yes

Qwen-2-7B Instruct
Based on the structured data provided, the Khotan Mosque is indeed located in China, and it was completed in 1870. However, the data does not specify the exact location within China as "Khotan, Xinjiang." Since Khotan is a city in Xinjiang, it can be inferred that the mosque is located in Xinjiang, China. However, the data does not explicitly state "Khotan" as part of the location.
A: NEI

GPT-4o-turbo
From the structured data provided, it is clear that the Khotan Mosque was completed in 1870 and is located in China. However, the data does not specify that the mosque is located in Khotan, Xinjiang. While it might be inferred from the name "Khotan Mosque" that it is located in Khotan, Xinjiang, the data does not explicitly confirm this location. Therefore, taking into account the explicit information given in the data, we can only confirm the completion year and the country, but not the specific location within China.
A: NEI

Figure 9: Case Study of Geography-time Reasoning.

Task: Multi-hop Reasoning
Q: Did the 1952 Colgate Red Raiders football team win the game against the Buffalo Bulls football team, representing the University of Buffalo, on October 4?
A: No

```
|   Date|Opponent|Site|Result   |
|--------------|-----------------|--------------|---------|
| September 20 | at Cortland  | Cortland, NY | L 13–33 |
| September 27 | Connecticut |              | L 7–47  |
| October 4    | Colgate         |          | L 0–13  |
| October 11   | Lehigh         |          | L 7–26  |
| October 18   | Bucknell       |          | L 0–22  |
| October 25   | Western Reserve |          | L 13–35 |
| November 1   | Alfred         |          | L 7–19  |
| November 8   | St. Lawrence   |          | W 12–6  |
|   Date|Opponent|Site|Result   |
|--------------|---------------------|-------------|---------|
| September 27 | Cornell          |          | W 14–7  |
| October 4    | Buffalo          | Buffalo, NY | W 13–0  |
| October 11   | Rutgers          |          | W 13–7  |
| October 18   | Harvard          |          | L 20–21 |
| October 25   | Bucknell         |          | W 28–0  |
| November 1   | Mississippi College |          | W 53–12 |
| November 8   | Holy Cross       |          | L 7–13  |
| November 15  | Syracuse         |          | L 14–20 |
| November 29  | Brown            |          | W 33–27 |
|   1952 Buffalo Bulls football   |
|--------------|---------------|
| Conference   | Independent   |
| 1952 record  | 1–7           |
| Head coach   |               |
| Home stadium | Civic Stadium |
|   1952 Colgate Red Raiders football   |
|--------------|------------------------|
| Conference   | Independent            |
| 1952 record  | 6–3                    |
| Head coach   |                        |
| Captain      | Donald Main            |
| Home stadium | Colgate Athletic Field |
```

LLaMA-3-8B Instruct
To answer this question, we need to find the game between the 1952 Colgate Red Raiders football team and the 1952 Buffalo Bulls football team. We can do this by looking at the structured data provided.

The data shows that the Colgate Red Raiders football team played the following games:

```
|   Date|Opponent|Site|Result   |
|-------------|--------------------|-------------|---------|
| September 20 | at Cortland     | Cortland, NY | L 13–33 |
| September 27 | Connecticut     |          | L 7–47  |
| October 4    | Colgate         |          | L 0–13  |
| October 11   | Lehigh          |          | L 7–26  |
| October 18   | Bucknell        |          | L 0–22  |
| October 25   | Western Reserve |          | L 13–35 |
| November 1   | Alfred          |          | L 7–19  |
| November 8   | St. Lawrence    |          | W 12–6  |
```

The data also shows that the Buffalo Bulls football team played the following games:

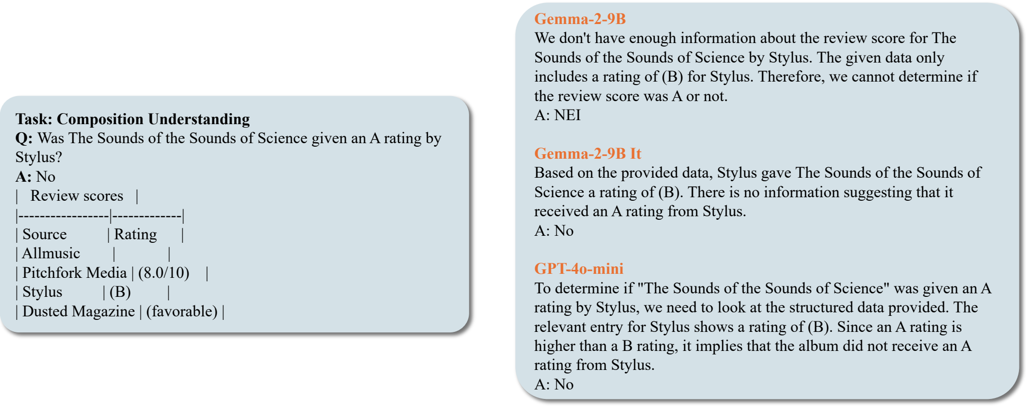
```
|   Date|Opponent|Site|Result   |
|-------------|--------------------|-------------|---------|
| September 27 | Cornell            |          | W 14–7  |
| October 4    | Buffalo            | Buffalo, NY | W 13–0  |
| October 11   | Rutgers            |          | W 13–7  |
| October 18   | Harvard            |          | L 20–21 |
| October 25   | Bucknell           |          | W 28–0  |
| November 1   | Mississippi College |          | W 53–12 |
| November 8   | Holy Cross         |          | L 7–13  |
| November 15  | Syracuse           |          | L 14–20 |
| November 29  | Brown              |          | W 33–27 |
```

By comparing the dates and opponents, we can see that the game between the Colgate Red Raiders football team and the Buffalo Bulls football team was not played on October 4. The Colgate Red Raiders football team played Colgate on October 4, not the Buffalo Bulls.
A: Yes

Gemma-2-9B It
Let's look at the provided data:
* **Colgate's Schedule:** The data shows Colgate played Buffalo on October 4th, 1952.
* **Colgate's Result:** The result for that game is listed as "W 13–0".
Therefore, the 1952 Colgate Red Raiders football team won the game against the Buffalo Bulls.
A: Yes

GPT-4o-turbo
The data provided shows the results of the games played by both the Colgate Red Raiders and the Buffalo Bulls during the 1952 football season. On October 4, the game between Colgate and Buffalo is specifically mentioned. According to the data, the result of the game was a win for Buffalo with a score of 13-0 over Colgate.
Therefore, the 1952 Colgate Red Raiders football team did not win against the Buffalo Bulls football team on October 4, 1952.
A: No

Figure 10: Case Study of Multi-hop Reasoning.



Task: Composition Understanding
Q: Was The Sounds of the Sounds of Science given an A rating by Stylus?
A: No

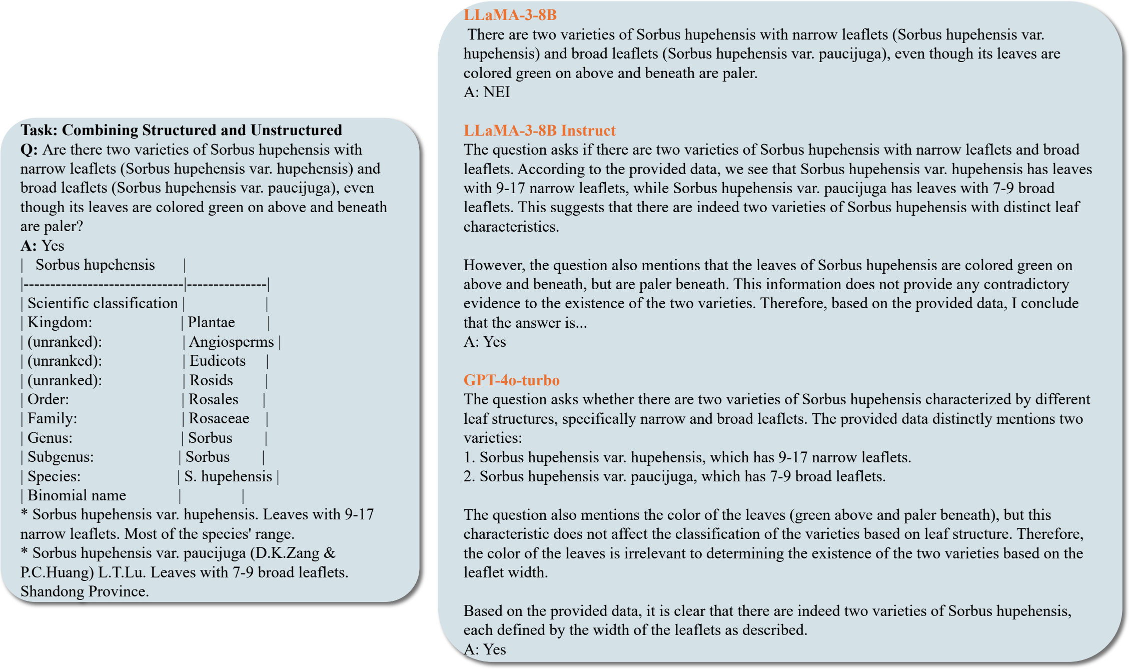
```
|   Review scores   |
|------------------|-------------|
| Source          | Rating     |
| Allmusic        |            |
| Pitchfork Media | (8.0/10)    |
| Stylus          | (B)        |
| Dusted Magazine | (favorable) |
```

Gemma-2-9B
We don't have enough information about the review score for The Sounds of the Sounds of Science by Stylus. The given data only includes a rating of (B) for Stylus. Therefore, we cannot determine if the review score was A or not.
A: NEI

Gemma-2-9B It
Based on the provided data, Stylus gave The Sounds of the Sounds of Science a rating of (B). There is no information suggesting that it received an A rating from Stylus.
A: No

GPT-4o-mini
To determine if "The Sounds of the Sounds of Science" was given an A rating by Stylus, we need to look at the structured data provided. The relevant entry for Stylus shows a rating of (B). Since an A rating is higher than a B rating, it implies that the album did not receive an A rating from Stylus.
A: No

Figure 11: Case Study of Composition Understanding.



Task: Combining Structured and Unstructured
Q: Are there two varieties of Sorbus hupehensis with narrow leaflets (Sorbus hupehensis var. hupehensis) and broad leaflets (Sorbus hupehensis var. paucijuga), even though its leaves are colored green on above and beneath are paler?
A: Yes

```
|   Sorbus hupehensis    |
|-------------------------------|---------------|
| Scientific classification |                |
| Kingdom:                  | Plantae       |
| (unranked):               | Angiosperms |
| (unranked):               | Eudicots     |
| (unranked):               | Rosids       |
| Order:                    | Rosales      |
| Family:                   | Rosaceae     |
| Genus:                    | Sorbus       |
| Subgenus:                 | Sorbus       |
| Species:                  | S. hupehensis |
| Binomial name             |               |
```

* Sorbus hupehensis var. hupehensis. Leaves with 9-17 narrow leaflets. Most of the species' range.
* Sorbus hupehensis var. paucijuga (D.K.Zang & P.C.Huang) L.T.Lu. Leaves with 7-9 broad leaflets. Shandong Province.

LLaMA-3-8B
There are two varieties of Sorbus hupehensis with narrow leaflets (Sorbus hupehensis var. hupehensis) and broad leaflets (Sorbus hupehensis var. paucijuga), even though its leaves are colored green on above and beneath are paler.
A: NEI

LLaMA-3-8B Instruct
The question asks if there are two varieties of Sorbus hupehensis with narrow leaflets and broad leaflets. According to the provided data, we see that Sorbus hupehensis var. hupehensis has leaves with 9-17 narrow leaflets, while Sorbus hupehensis var. paucijuga has leaves with 7-9 broad leaflets. This suggests that there are indeed two varieties of Sorbus hupehensis with distinct leaf characteristics.

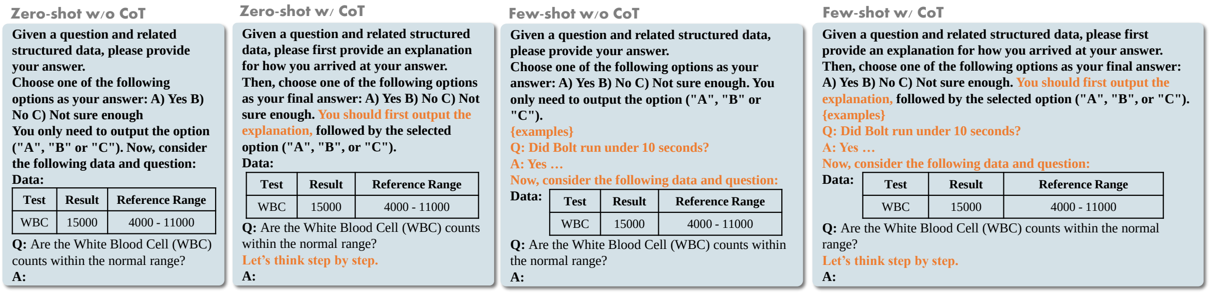
However, the question also mentions that the leaves of Sorbus hupehensis are colored green on above and beneath, but are paler beneath. This information does not provide any contradictory evidence to the existence of the two varieties. Therefore, based on the provided data, I conclude that the answer is...
A: Yes

GPT-4o-turbo
The question asks whether there are two varieties of Sorbus hupehensis characterized by different leaf structures, specifically narrow and broad leaflets. The provided data distinctly mentions two varieties:
1. Sorbus hupehensis var. hupehensis, which has 9-17 narrow leaflets.
2. Sorbus hupehensis var. paucijuga, which has 7-9 broad leaflets.

The question also mentions the color of the leaves (green above and paler beneath), but this characteristic does not affect the classification of the varieties based on leaf structure. Therefore, the color of the leaves is irrelevant to determining the existence of the two varieties based on the leaflet width.

Based on the provided data, it is clear that there are indeed two varieties of Sorbus hupehensis, each defined by the width of the leaflets as described.
A: Yes

Figure 12: Case Study of Combining Structured and Unstructured.



Zero-shot w/o CoT

Given a question and related structured data, please provide your answer.
Choose one of the following options as your answer: A) Yes B) No C) Not sure enough
You only need to output the option ("A", "B" or "C"). Now, consider the following data and question:
Data:

Test	Result	Reference Range
WBC	15000	4000 - 11000

Q: Are the White Blood Cell (WBC) counts within the normal range?
A:

Zero-shot w/ CoT

Given a question and related structured data, please first provide an explanation for how you arrived at your answer. Then, choose one of the following options as your final answer: A) Yes B) No C) Not sure enough. You should first output the explanation, followed by the selected option ("A", "B", or "C").
Data:

Test	Result	Reference Range
WBC	15000	4000 - 11000

Q: Are the White Blood Cell (WBC) counts within the normal range?
Let's think step by step.
A:

Few-shot w/o CoT

Given a question and related structured data, please provide your answer.
Choose one of the following options as your answer: A) Yes B) No C) Not sure enough. You only need to output the option ("A", "B" or "C").
{examples}
Q: Did Bolt run under 10 seconds?
A: Yes …
Now, consider the following data and question:
Data:

Test	Result	Reference Range
WBC	15000	4000 - 11000

Q: Are the White Blood Cell (WBC) counts within the normal range?
A:

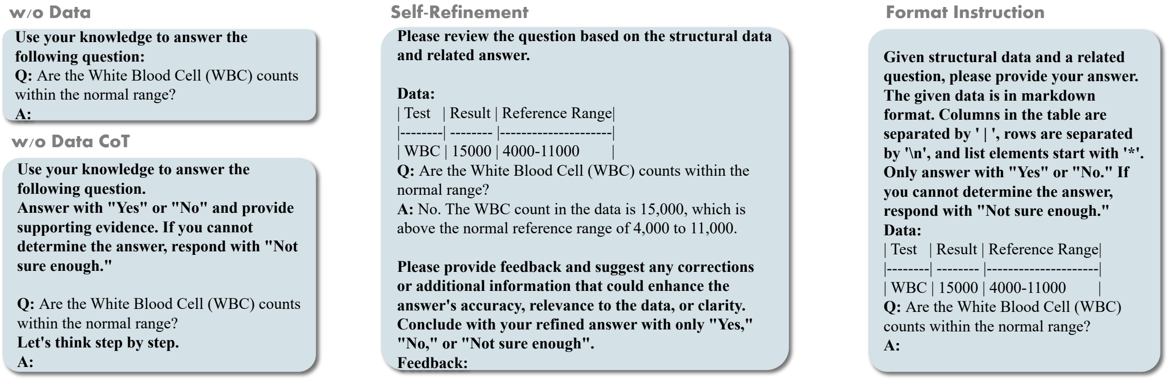
Few-shot w/ CoT

Given a question and related structured data, please first provide an explanation for how you arrived at your answer. Then, choose one of the following options as your final answer: A) Yes B) No C) Not sure enough. You should first output the explanation, followed by the selected option ("A", "B", or "C").
{examples}
Q: Did Bolt run under 10 seconds?
A: Yes …
Now, consider the following data and question:
Data:

Test	Result	Reference Range
WBC	15000	4000 - 11000

Q: Are the White Blood Cell (WBC) counts within the normal range?
Let's think step by step.
A:

Figure 13: Prompts used in different settings (main differences with zero-shot w/o CoT are marked in orange).



w/o Data

Use your knowledge to answer the following question:
Q: Are the White Blood Cell (WBC) counts within the normal range?
A:

w/o Data CoT

Use your knowledge to answer the following question.
Answer with "Yes" or "No" and provide supporting evidence. If you cannot determine the answer, respond with "Not sure enough."

Q: Are the White Blood Cell (WBC) counts within the normal range?
Let's think step by step.
A:

Self-Refinement

Please review the question based on the structural data and related answer.

Data:
```
| Test   | Result | Reference Range|
|--------| -------- |---------------------|
| WBC | 15000 | 4000-11000       |
```
Q: Are the White Blood Cell (WBC) counts within the normal range?
A: No. The WBC count in the data is 15,000, which is above the normal reference range of 4,000 to 11,000.

Please provide feedback and suggest any corrections or additional information that could enhance the answer's accuracy, relevance to the data, or clarity. Conclude with your refined answer with only "Yes," "No," or "Not sure enough".
Feedback:

Format Instruction

Given structural data and a related question, please provide your answer. The given data is in markdown format. Columns in the table are separated by ' | ', rows are separated by '\n', and list elements start with '*'. Only answer with "Yes" or "No." If you cannot determine the answer, respond with "Not sure enough."
Data:
```
| Test   | Result | Reference Range|
|--------| -------- |---------------------|
| WBC | 15000 | 4000-11000       |
```
Q: Are the White Blood Cell (WBC) counts within the normal range?
A:

Figure 14: Examples of different prompting strategies.

Models	Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.		Overall	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o-mini	Self-Refine.	58.03	60.73	54.44	60.08	51.61	58.81	66.91	68.17	56.51	63.37	55.36	60.38
	Self-Consis.	62.80	63.49	59.84	63.02	58.30	63.92	66.91	67.35	62.30	67.06	60.83	63.69
	Format Instruct.	62.82	63.44	58.53	61.95	57.52	63.10	66.67	67.16	61.37	66.38	60.03	63.03
LLaMA3-8B Instruct	Self-Refine.	58.05	57.74	54.87	55.40	64.71	66.75	61.20	59.46	69.83	70.94	60.23	60.44
	Self-Consis.	62.52	56.96	55.33	51.38	71.00	68.54	61.19	57.51	73.30	70.86	63.26	58.95
	Format Instruct.	61.95	55.85	51.35	46.00	68.41	66.38	60.20	54.04	71.21	68.72	60.75	55.76
Gemma-2-9B It	Self-Refine.	49.56	53.86	35.18	43.16	43.62	53.69	61.69	64.01	50.63	59.93	43.41	51.04
	Self-Consis.	51.31	53.81	33.09	39.19	44.81	54.58	60.20	62.91	52.19	60.74	43.52	49.87
	Format Instruct.	53.79	55.33	34.19	40.31	45.75	55.33	59.46	61.74	53.24	61.64	44.88	50.97
Qwen2-7B-Instruct	Self-Refine.	52.24	55.50	36.87	43.32	44.66	53.88	57.21	59.93	51.39	59.54	44.95	51.34
	Self-Consis.	55.18	57.62	40.95	47.56	47.61	56.30	57.46	61.66	53.88	61.58	48.20	54.24
	Format Instruct.	56.50	58.43	37.48	43.72	48.61	57.07	57.46	61.28	54.55	62.15	47.51	53.29

Table 9: Performance of GPT-4o-mini using different prompting strategies across five factual tasks.

Methods	Zero-shot w/o CoT		Zero-shot w/ CoT		Few-shot w/o CoT		Few-shot w/ CoT		Overall	
	BA	MF1	BA	MF1	BA	MF1	BA	MF1	BA	MF1
Qwen2-7B	23.06	19.97	39.69	26.00	35.85	25.69	38.86	28.70	34.37	25.09
LLaMA-3-8B	20.65	17.57	19.17	17.94	25.29	19.66	35.16	26.76	25.07	20.48
Gemma-2-9B	14.97	12.87	32.50	24.51	14.55	11.51	39.56	29.92	25.40	19.70
Qwen2-7B Instruct	43.51	27.29	40.87	24.53	43.81	29.25	41.45	24.66	42.41	26.43
LLaMA-3-8B Instruct	39.48	32.00	39.65	25.10	40.23	32.48	42.64	28.03	40.50	29.40
Gemma-2-9B It	44.34	27.74	44.58	25.94	45.11	34.52	44.51	26.81	44.64	28.75
GLM-4-9B Chat	42.88	38.49	43.25	26.78	42.47	38.17	44.65	28.41	43.31	32.96
Mistral-7B Instruct	39.77	26.97	41.31	24.16	42.46	30.52	43.20	27.17	41.69	27.20
GPT-4o-Mini	46.96	44.92	46.46	42.86	47.08	42.90	46.89	43.71	46.85	43.60
GPT-4-Turbo	48.02	45.19	48.14	43.13	47.72	43.63	48.59	43.15	48.12	43.77

Table 10: Balanced accuracy and Macro F1 of 10 LLMs on the StructFact benchmark using various prompts.

Methods	Zero-shot w/o CoT		Zero-shot w/ CoT		Few-shot w/o CoT		Few-shot w/ CoT		Overall	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Qwen2-7B	58.64	31.82	62.33	49.40	59.71	45.39	64.49	54.80	61.29	45.35
LLaMA-3-8B	52.78	29.72	55.78	27.65	53.25	32.13	58.83	55.64	55.16	36.28
Gemma-2-9B	51.77	22.67	58.18	42.76	53.72	17.31	61.93	61.14	56.40	35.97
Qwen2-7B Instruct	65.11	47.85	66.63	41.27	66.79	44.88	66.67	41.01	66.30	43.75
LLaMA-3-8B Instruct	62.84	62.92	64.04	43.01	63.44	63.39	67.71	45.43	64.51	53.69
Gemma-2-9B It	70.37	43.53	69.03	41.08	71.10	44.81	70.67	43.03	70.29	43.11
GLM-4-9B Chat	64.82	52.56	68.24	42.58	65.38	52.97	68.44	47.10	66.72	48.80
Mistral-7B Instruct	62.68	50.90	65.58	37.33	63.29	60.13	66.40	43.80	64.49	48.04
GPT-4o-Mini	68.00	60.80	70.27	54.20	68.82	55.06	70.43	56.35	69.38	56.60
GPT-4-Turbo	68.76	60.67	71.27	53.31	69.80	56.01	71.35	53.18	70.29	55.79

Table 11: Precision and recall of 10 LLMs on the StructFact benchmark using various prompts.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Qwen2-7B	58.11	28.30	55.32	28.24	68.87	34.05	65.94	38.31	69.78	41.26
LLaMA-3-8B	54.34	28.48	49.83	28.36	63.72	29.61	53.61	34.08	63.42	34.00
Gemma-2-9B	51.77	15.98	48.47	21.45	67.26	30.55	60.43	25.87	61.55	25.92
Qwen2-7B Instruct	61.48	54.58	66.00	40.52	71.62	47.91	70.20	57.46	73.95	53.33
LLaMA-3-8B Instruct	57.41	62.28	62.58	54.78	69.43	70.61	61.54	60.94	72.15	73.28
Gemma-2-9B It	60.51	51.36	73.96	33.03	81.38	44.73	71.58	59.95	82.41	52.46
GLM-4-9B Chat	63.06	59.27	63.75	46.70	70.01	50.67	67.78	63.93	73.41	56.80
Mistral-7B Instruct	61.48	55.37	59.07	43.44	70.98	52.03	62.60	54.98	73.77	59.07
GPT-4o-Mini	64.51	62.52	67.79	60.13	75.06	58.04	68.80	67.42	75.06	62.10
GPT-4-Turbo	64.16	61.76	68.71	61.93	75.03	54.90	70.78	70.15	77.18	61.59
Overall	59.68	47.99	61.55	41.86	71.34	47.31	65.33	53.31	72.27	51.98

Table 12: Precision and recall of 10 LLMs on the StructFact benchmark across five factual tasks under the zero-shot w/o CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	BA	MF1	BA	MF1	BA	MF1	BA	MF1	BA	MF1
Qwen2-7B	19.42	18.07	23.62	19.35	22.30	19.48	27.85	23.99	27.13	21.99
LLaMA-3-8B	20.35	17.12	21.08	17.39	18.61	16.23	32.55	21.59	22.01	18.34
Gemma-2-9B	11.08	10.58	15.57	12.84	18.50	14.40	16.96	14.42	15.33	12.74
Qwen2-7B Instruct	40.04	30.66	44.59	29.03	43.23	25.79	40.16	34.04	44.43	29.72
LLaMA-3-8B Instruct	37.14	35.27	39.92	31.00	38.34	37.48	38.26	37.28	40.23	39.96
Gemma-2-9B It	38.61	32.89	44.50	29.55	45.61	33.09	55.90	37.33	48.18	35.80
GLM-4-9B Chat	38.22	36.99	44.85	37.61	40.91	34.71	41.31	42.14	44.53	38.48
Mistral-7B Instruct	38.64	26.56	38.69	25.59	39.45	25.58	34.91	30.13	41.78	28.08
GPT-4o-Mini	41.68	41.14	49.20	46.20	46.31	41.15	52.58	52.69	46.86	42.68
GPT-4-Turbo	41.05	40.17	50.43	47.09	47.20	39.42	58.31	60.05	49.42	43.67
Overall	32.62	28.95	37.25	29.56	36.05	28.73	39.88	35.37	37.99	31.15

Table 13: Balanced accuracy and Macro F1 of 10 LLMs on the StructFact benchmark across five factual tasks under the zero-shot w/o CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen2-7B	57.65	57.56	39.63	42.68	52.11	59.05	59.21	59.50	55.22	61.00
LLaMA-3-8B	28.50	37.79	24.15	31.16	28.98	39.88	30.10	38.53	32.33	43.38
Gemma-2-9B	45.35	50.28	37.92	42.52	44.35	52.43	44.03	48.90	47.74	54.98
Qwen2-7B Instruct	53.92	56.87	31.66	39.43	39.74	49.90	51.24	55.54	45.18	55.02
LLaMA-3-8B Instruct	50.37	54.01	36.13	42.62	40.61	49.98	51.49	54.39	49.30	57.93
Gemma-2-9B It	48.75	53.66	33.91	43.30	40.61	51.25	57.46	60.31	44.87	55.46
GLM-4-9B Chat	53.57	58.14	35.52	45.05	39.00	49.17	56.22	59.62	45.09	55.43
Mistral-7B Instruct	43.93	50.87	32.00	40.89	34.94	44.83	50.00	55.88	40.95	51.28
GPT-4o-Mini	59.10	61.50	50.89	59.04	51.64	60.06	65.18	66.46	56.22	64.12
GPT-4-Turbo	58.44	61.04	51.71	60.26	49.48	57.78	64.93	65.93	52.64	61.56
Overall	49.96	54.17	37.35	44.70	42.15	51.43	52.99	56.51	46.95	56.02

Table 14: Accuracy and F1 score of 10 LLMs on the StructFact benchmark across five factual tasks under the zero-shot w/ CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Qwen2-7B	60.38	57.65	61.86	39.63	70.62	52.11	66.14	59.21	71.26	55.22
LLaMA-3-8B	57.51	28.50	52.43	24.15	66.28	28.98	57.02	30.10	67.36	32.33
Gemma-2-9B	58.46	45.35	55.55	37.92	67.74	44.35	57.02	44.03	69.66	47.74
Qwen2-7B Instruct	61.87	53.92	69.06	31.66	73.41	39.74	67.39	51.24	75.92	45.18
LLaMA-3-8B Instruct	60.99	50.37	63.89	36.13	69.48	40.61	63.20	51.49	74.34	49.30
Gemma-2-9B It	61.11	48.75	71.34	33.91	77.50	40.61	66.14	57.46	78.25	44.87
GLM-4-9B Chat	64.25	53.57	69.49	35.52	74.36	39.00	65.64	56.22	76.37	45.09
Mistral-7B Instruct	62.24	43.93	64.62	32.00	72.72	34.94	65.04	50.00	74.37	40.95
GPT-4o-Mini	64.24	59.10	72.71	50.89	76.01	51.64	68.24	65.18	76.66	56.22
GPT-4-Turbo	64.22	58.44	74.82	51.71	76.52	49.48	67.30	64.93	77.91	52.64
Overall	61.53	49.96	65.58	37.35	72.46	42.15	64.31	52.99	74.21	46.95

Table 15: Precision and recall of 10 LLMs on the StructFact benchmark across five factual tasks under the zero-shot w/ CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	BA	MF1	BA	MF1	BA	MF1	BA	MF1	BA	MF1
Qwen2-7B	36.98	26.80	40.23	23.88	39.31	26.03	37.19	28.95	42.11	25.92
LLaMA-3-8B	19.80	18.25	18.81	17.00	17.94	16.90	18.92	18.46	20.44	19.13
Gemma-2-9B	31.21	24.49	33.22	23.59	31.07	23.50	28.77	24.23	34.96	25.27
Qwen2-7B Instruct	36.21	26.54	41.80	22.07	41.48	23.18	35.79	27.41	43.30	23.86
LLaMA-3-8B Instruct	34.25	25.21	41.18	23.77	37.35	22.84	32.38	26.36	45.42	26.09
Gemma-2-9B It	36.29	25.70	46.34	24.21	46.42	25.20	51.45	35.83	46.82	25.81
GLM-4-9B Chat	38.35	28.25	42.97	24.81	45.16	24.60	47.18	31.43	46.62	26.48
Mistral-7B Instruct	36.19	24.77	40.53	22.61	43.70	22.41	36.69	28.33	45.62	24.15
GPT-4o-Mini	40.26	40.38	48.84	42.82	46.16	40.34	46.80	46.44	47.69	41.38
GPT-4-Turbo	42.68	41.29	50.38	43.69	48.17	39.82	53.91	49.44	48.18	41.04
Overall	35.22	28.17	40.43	26.84	39.68	26.48	38.91	31.69	42.12	27.91

Table 16: Balanced accuracy and Macro F1 of 10 LLMs on the StructFact benchmark across five factual tasks under the zero-shot w/ CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen2-7B	50.19	54.03	39.36	44.78	46.28	54.34	50.75	53.96	50.23	57.83
LLaMA-3-8B	30.30	37.95	29.45	34.63	36.58	46.16	30.35	37.67	35.85	45.21
Gemma-2-9B	17.43	23.45	15.98	20.19	18.65	25.42	18.41	25.58	18.46	24.95
Qwen2-7B Instruct	54.35	57.82	36.45	43.11	44.58	53.43	57.21	60.90	49.14	57.50
LLaMA-3-8B Instruct	63.37	58.22	55.42	53.08	70.17	68.47	63.68	60.77	73.30	72.09
Gemma-2-9B It	55.12	56.73	34.22	41.49	45.62	55.37	63.19	64.72	50.75	60.11
GLM-4-9B Chat	60.39	59.73	44.69	49.17	53.96	60.11	63.93	63.60	58.41	64.52
Mistral-7B Instruct	61.35	60.78	52.10	52.54	66.71	68.99	63.68	61.93	68.38	70.15
GPT-4o-Mini	60.38	62.24	52.50	58.28	52.55	59.80	66.42	67.15	54.84	62.19
GPT-4-Turbo	60.38	61.88	55.38	61.06	51.63	58.24	66.42	66.91	54.98	63.15
Overall	51.33	53.28	41.55	45.83	48.67	55.03	54.40	56.32	51.43	57.77

Table 17: Accuracy and F1 score of 10 LLMs on the StructFact benchmark across five factual tasks under the few-shot w/o CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Qwen2-7B	59.91	50.19	57.40	39.36	68.94	46.28	61.30	50.75	69.95	50.23
LLaMA-3-8B	53.94	30.30	49.07	29.45	66.43	36.58	55.05	30.35	65.39	35.85
Gemma-2-9B	54.89	17.43	48.64	15.98	64.74	18.65	65.32	18.41	66.72	18.46
Qwen2-7B Instruct	64.18	54.35	67.43	36.45	73.35	44.58	71.91	57.21	74.64	49.14
LLaMA-3-8B Instruct	58.19	63.37	63.33	55.42	68.70	70.17	64.32	63.68	73.42	73.30
Gemma-2-9B It	62.17	55.12	73.25	34.22	80.08	45.62	71.86	63.19	81.23	50.75
GLM-4-9B Chat	62.74	60.39	65.13	44.69	70.68	53.96	70.20	63.93	74.36	58.41
Mistral-7B Instruct	61.39	61.35	61.66	52.10	71.83	66.71	65.70	63.68	73.17	68.38
GPT-4o-Mini	65.42	60.38	67.54	52.50	76.16	52.55	68.91	66.42	76.72	54.84
GPT-4-Turbo	64.26	60.38	70.50	55.38	76.40	51.63	67.73	66.42	78.20	54.98
Overall	60.71	51.33	62.40	41.55	71.73	48.67	66.23	54.40	73.38	51.43

Table 18: Precision and recall of 10 LLMs on the StructFact benchmark across five factual tasks under the few-shot w/o CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	BA	MF1	BA	MF1	BA	MF1	BA	MF1	BA	MF1
Qwen2-7B	36.48	26.20	34.27	24.48	35.19	24.32	57.21	30.75	36.01	25.16
LLaMA-3-8B	23.39	18.58	26.25	19.28	26.55	20.46	22.82	18.56	25.00	19.54
Gemma-2-9B	14.01	11.48	14.54	10.91	14.74	11.74	12.36	12.65	15.24	11.31
Qwen2-7B Instruct	42.22	31.10	43.03	32.05	43.52	32.24	54.20	42.54	45.12	32.96
LLaMA-3-8B Instruct	36.66	35.01	41.74	31.90	37.46	36.98	40.16	39.54	41.87	42.13
Gemma-2-9B It	40.02	35.09	45.79	31.18	46.99	34.38	61.71	47.96	44.61	35.25
GLM-4-9B Chat	38.28	37.16	44.50	36.40	39.18	35.47	44.21	43.11	45.03	39.22
Mistral-7B Instruct	39.92	35.92	43.07	29.12	42.08	34.00	43.79	43.40	43.94	34.62
GPT-4o-Mini	42.76	41.79	47.27	42.27	46.73	39.86	55.97	53.10	49.30	40.46
GPT-4-Turbo	41.90	41.22	49.54	44.33	48.96	39.64	45.09	44.99	47.29	41.07
Overall	35.56	31.36	39.00	30.19	38.14	30.91	43.75	37.66	39.34	32.17

Table 19: Balanced accuracy and Macro F1 of 10 LLMs on the StructFact benchmark across five factual tasks under the few-shot w/o CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen2-7B	57.98	58.50	47.25	49.93	58.91	64.52	61.19	62.20	61.66	66.44
LLaMA-3-8B	56.03	55.56	48.52	46.85	60.50	63.71	54.97	54.24	65.22	66.91
Gemma-2-9B	60.02	58.96	53.72	52.45	68.17	69.11	62.94	62.50	70.94	71.31
Qwen2-7B Instruct	51.68	55.78	31.88	39.99	41.24	51.29	52.74	58.31	44.62	54.55
LLaMA-3-8B Instruct	52.21	57.44	40.64	49.51	42.44	52.08	53.23	57.34	48.74	58.37
Gemma-2-9B It	53.86	57.52	34.86	44.45	42.03	52.95	61.44	63.33	45.07	55.99
GLM-4-9B Chat	56.36	59.70	39.08	47.86	46.19	55.65	60.45	63.31	51.28	60.48
Mistral-7B Instruct	48.49	54.83	38.30	46.84	43.53	52.71	58.21	61.53	48.05	57.49
GPT-4o-Mini	62.36	63.52	52.88	60.49	53.66	61.60	70.15	70.40	57.11	64.72
GPT-4-Turbo	60.03	62.38	50.98	59.20	48.95	57.88	66.42	67.97	51.70	61.01
Overall	55.90	58.42	43.81	49.76	50.56	58.15	60.17	62.11	54.44	61.73

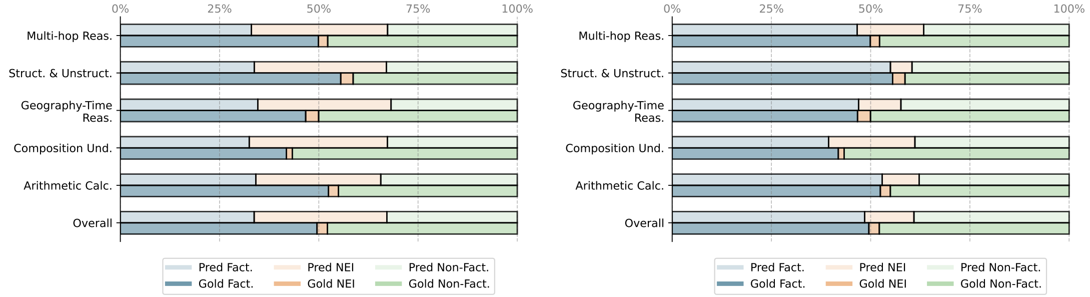
Table 20: Accuracy and F1 score of 10 LLMs on the StructFact benchmark across five factual tasks under the few-shot w/ CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Qwen2-7B	61.01	57.98	63.75	47.25	72.32	58.91	69.20	61.19	73.67	61.66
LLaMA-3-8B	56.82	56.03	59.04	48.52	67.47	60.50	60.35	54.97	69.66	65.22
Gemma-2-9B	58.77	60.02	61.26	53.72	70.30	68.17	65.10	62.94	72.25	70.94
Qwen2-7B Instruct	62.31	51.68	68.47	31.88	73.24	41.24	67.83	52.74	75.82	44.62
LLaMA-3-8B Instruct	64.28	52.21	68.66	40.64	73.38	42.44	63.49	53.23	75.91	48.74
Gemma-2-9B It	63.13	53.86	72.82	34.86	78.73	42.03	67.84	61.44	79.02	45.07
GLM-4-9B Chat	63.87	56.36	68.98	39.08	75.07	46.19	69.23	60.45	76.90	51.28
Mistral-7B Instruct	64.16	48.49	65.57	38.30	72.79	43.53	65.79	58.21	74.63	48.05
GPT-4o-Mini	64.77	62.36	72.79	52.88	76.08	53.66	71.00	70.15	76.38	57.11
GPT-4-Turbo	65.18	60.03	72.94	50.98	78.59	48.95	70.02	66.42	78.80	51.70
Overall	62.43	55.90	67.43	43.81	73.80	50.56	66.99	60.17	75.30	54.44

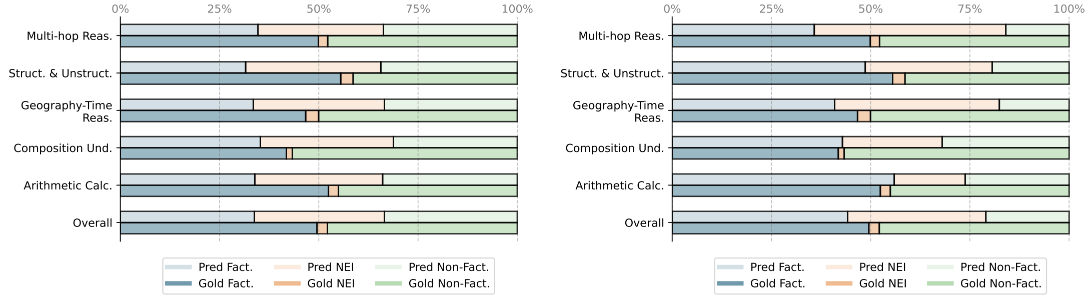
Table 21: Precision and recall of 10 LLMs on the StructFact benchmark across five factual tasks under the few-shot w/ CoT setting.

Methods	Arithmetic Calc.		Spatiotemporal Cogn.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
	BA	MF1	BA	MF1	BA	MF1	BA	MF1	BA	MF1
Qwen2-7B	36.38	27.26	39.40	27.52	38.25	28.76	38.74	30.34	41.33	29.97
LLaMA-3-8B	33.82	25.66	36.34	25.64	33.19	26.29	44.65	29.37	36.41	28.65
Gemma-2-9B	36.45	27.61	40.59	28.89	38.58	29.72	47.59	33.37	41.84	32.09
Qwen2-7B Instruct	37.95	26.50	40.97	22.27	41.58	23.80	37.52	29.18	46.30	24.26
LLaMA-3-8B Instruct	36.26	27.89	45.20	27.38	42.27	25.74	34.38	28.20	45.90	27.54
Gemma-2-9B It	38.08	27.32	45.93	24.73	46.63	26.56	50.37	37.23	45.33	26.06
GLM-4-9B Chat	37.39	28.48	44.61	26.39	48.13	27.40	46.04	36.53	48.93	31.90
Mistral-7B Instruct	38.88	27.36	43.15	25.80	45.75	25.82	38.37	34.12	44.68	26.94
GPT-4o-Mini	41.51	41.51	49.39	43.89	47.84	41.21	46.62	47.22	45.37	41.44
GPT-4-Turbo	43.31	42.05	49.59	42.99	49.79	40.50	55.17	50.34	50.68	41.24
Overall	38.00	30.16	43.52	29.55	43.20	29.58	43.95	35.59	44.68	31.01

Table 22: Balanced accuracy and Macro F1 of 10 LLMs on the StructFact benchmark across five factual tasks under the few-shot w/ CoT setting.



(a) GPT-4O Mini (b) Llama3-8B Instruct



(c) Gemma2-9B It (d) Qwen2-7B-Instruct

Figure 15: Responses Distributions of Different Models.

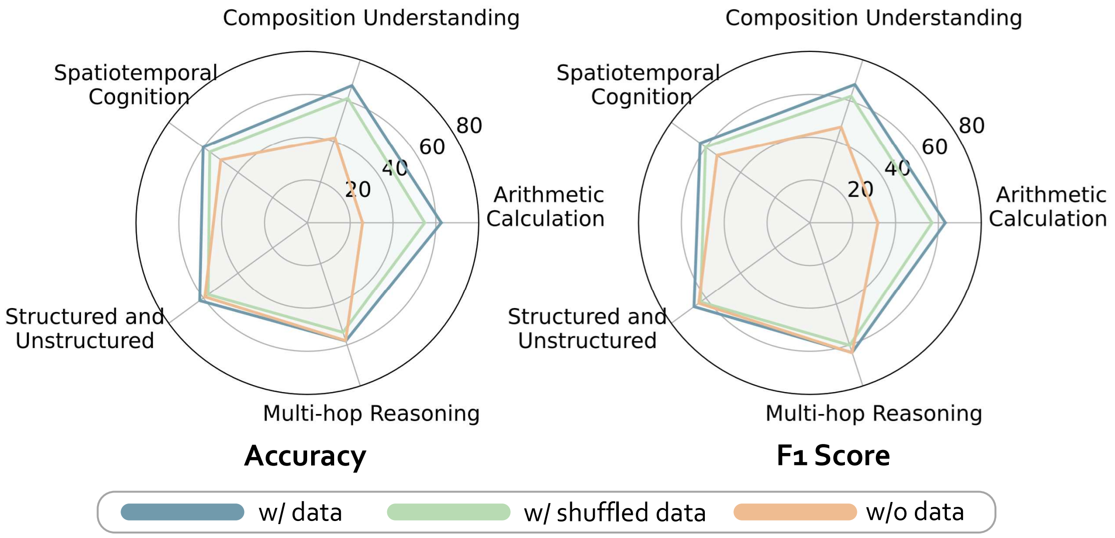
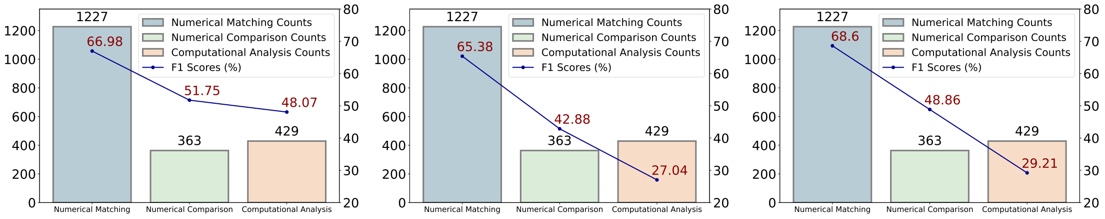


Figure 16: Accuracy and F1 score of GPT-4o-mini under different settings of structured evidence.

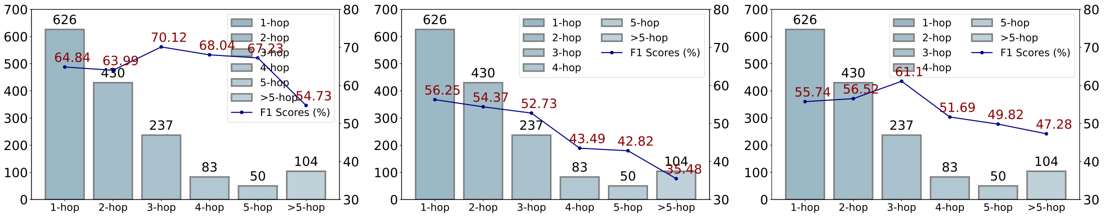
Table 23: GPT-4o-mini's evidence resilience across different factual tasks under zero-shot settings without CoT prompts. The percentage of decrease with respect to the setting with original structured data (w/ data) is shown in brackets.

Settings	Metrics	Arithmetic Calc.		Geography-Time Reas.		Multi-hop Reas.		Composition Und.		Struct. & Unstruct.	
		Score (%)	Decr.(%)	Score	Decr.(%)	Score	Decr.(%)	Score	Decr.(%)	Score	Decr.(%)
w/ data	Acc	63.93	-	59.92	-	67.56	-	71.88	-	65.94	-
	F1	64.15	-	62.96	-	69.66	-	72.44	-	68.50	-
	Prec.	64.39	-	67.16	-	76.29	-	73.10	-	72.89	-
	Recall	63.93	-	59.92	-	67.56	-	71.88	-	65.94	-
w/ shuffled data (rows)	Acc	62.68	(-1.99)	58.39	(-2.57)	64.81	(-4.12)	70.62	(-1.75)	63.18	(-4.25)
	F1	63.43	(-1.17)	61.73	(-1.96)	67.12	(-3.66)	71.16	(-1.77)	66.38	(-3.04)
	Prec.	64.77	(+0.58)	66.44	(-1.09)	75.76	(-0.70)	71.87	(-1.66)	73.31	(+0.57)
	Recall	62.68	(-1.99)	58.39	(-2.57)	64.81	(-4.12)	70.62	(-1.75)	63.18	(-4.25)
w/ shuffled data (columns)	Acc	61.19	(-4.22)	58.35	(-2.63)	65.04	(-3.73)	68.50	(-4.73)	63.87	(-3.12)
	F1	61.94	(-3.41)	61.66	(-2.08)	67.52	(-3.03)	69.42	(-4.19)	66.82	(-2.46)
	Prec.	63.79	(-0.93)	66.31	(-1.28)	76.11	(-0.23)	70.62	(-3.39)	72.63	(-0.35)
	Recall	61.19	(-4.22)	58.35	(-2.63)	65.04	(-3.73)	68.50	(-4.73)	63.87	(-3.12)
w/ shuffled data (rows and columns)	Acc	54.78	(-14.31)	56.30	(-6.04)	58.86	(-12.88)	67.19	(-6.53)	59.25	(-10.15)
	F1	59.01	(-8.01)	57.21	(-9.13)	62.18	(-10.73)	67.26	(-7.15)	63.14	(-7.83)
	Prec.	60.98	(-5.30)	64.07	(-4.60)	73.09	(-4.20)	70.33	(-3.79)	70.64	(-3.09)
	Recall	56.78	(-11.18)	56.30	(-6.04)	59.86	(-11.40)	67.19	(-6.53)	58.25	(-11.66)
w/o data	Acc	25.89	(-38.04)	49.11	(-10.81)	53.33	(-14.23)	51.65	(-20.23)	55.43	(-10.51)
	F1	32.59	(-31.56)	52.85	(-10.11)	60.23	(-9.43)	57.79	(-14.65)	59.67	(-8.83)
	Prec.	51.25	(-13.14)	59.30	(-7.86)	70.38	(-5.91)	67.96	(-5.14)	66.36	(-6.53)
	Recall	25.89	(-38.04)	49.11	(-10.81)	53.33	(-14.23)	51.65	(-20.23)	55.43	(-10.51)



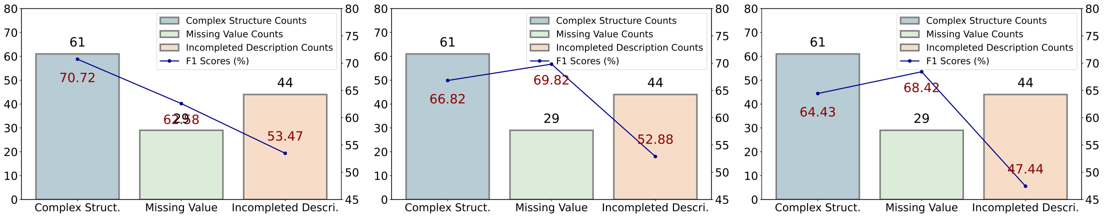
(a) Llama3-8B Instruct (b) Gemma2-9B It (c) Qwen2-7B-Instruct

Figure 17: Performance of Different Models on Arithmetic Calculation.



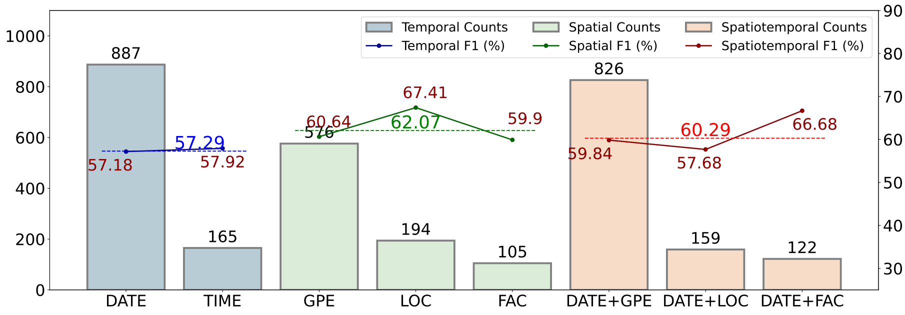
(a) Llama3-8B Instruct (b) Gemma2-9B It (c) Qwen2-7B-Instruct

Figure 18: Performance of Different Models on Multi-hop Reasoning.

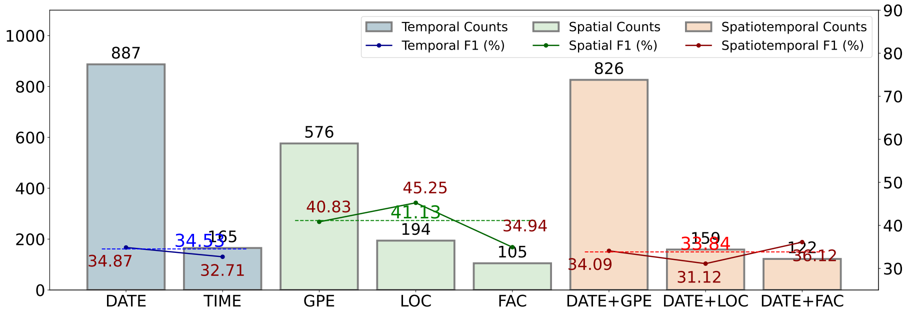


(a) Llama3-8B Instruct (b) Gemma2-9B It (c) Qwen2-7B-Instruct

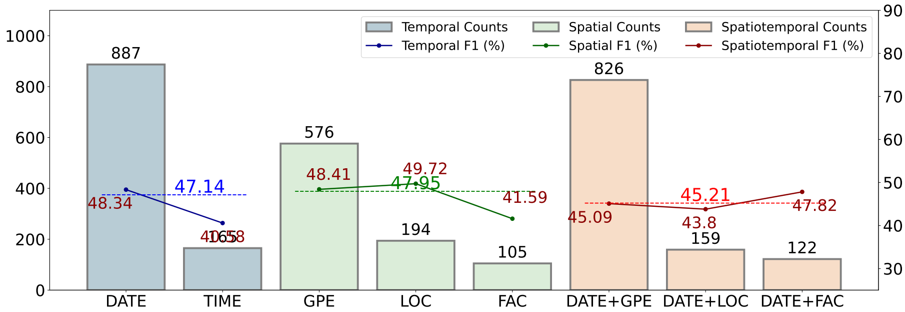
Figure 19: Performance of Different Models on Composition Understanding.



(a) Llama3-8B Instruct

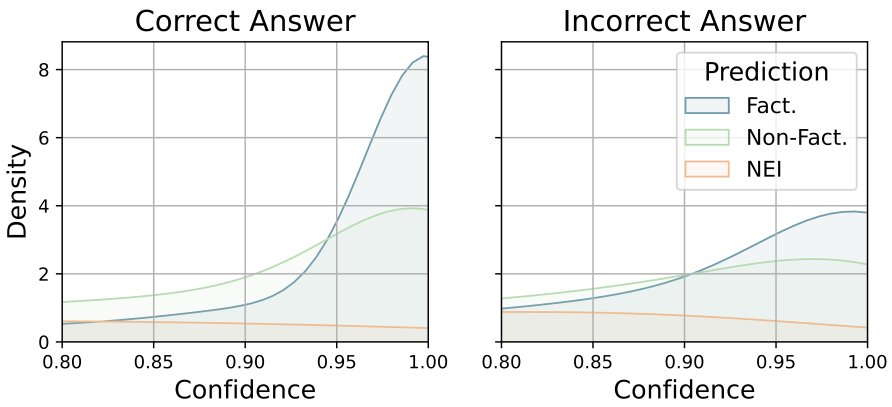


(b) Gemma2-9B It

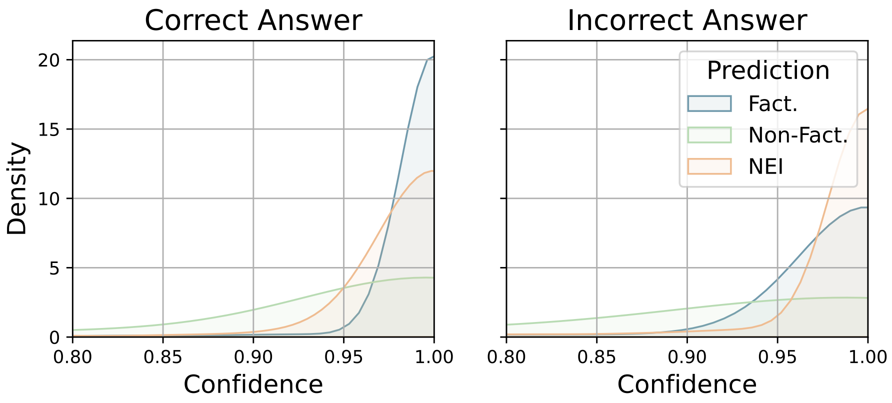


(c) Qwen2-7B-Instruct

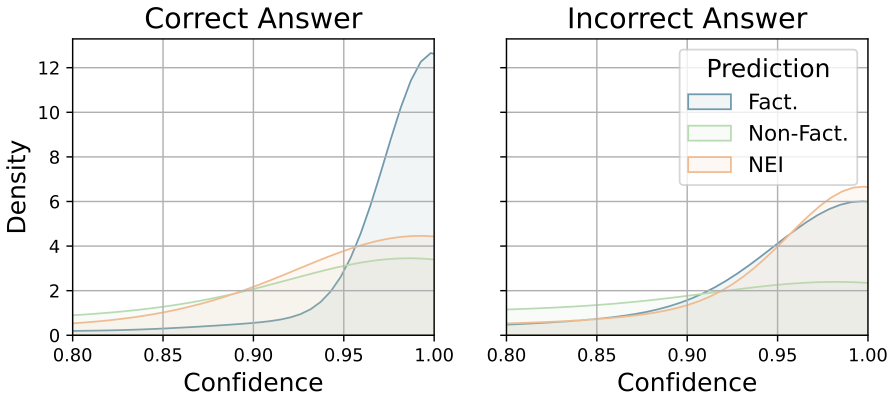
Figure 20: Performance of Different Models on geography-time Reasoning.



(a) Llama3-8B Instruct



(b) Gemma2-9B It



(c) Qwen2-7B-Instruct

Figure 21: Confidence of Different Models.