Syllogism-Inspired TableQA: Evidentialization Makes Decomposition Reasoning and Answer Verification More Reliable

Zhe Zhang, Lili Bai, Chaopeng Guo, Jie Song*

Software College, Northeastern University, Shenyang, China
2390171@stu.neu.edu.cn, 2210533@stu.neu.edu.cn, guochaopeng@swc.neu.edu.cn, songjie@mail.neu.edu.cn

Abstract

Existing large language model (LLM)-based table question answering (TableQA) methods primarily involve decomposition reasoning and answer verification processes. However, decomposing questions solely at the semantic level, without considering the factual evidence in tables, fails to significantly reduce the difficulty for LLMs in understanding the key information in questions. Furthermore, reasoning and verification without supporting factual evidence are often arbitrary and unreliable. In light of these issues, this paper proposes a **S**yllogism-**I**nspired **R**easoning and **V**erification method (SIRV), which performs reliable decomposition reasoning and answer verification based on the evidential concept of syllogism. Specifically, SIRV extracts question-relevant factual evidence from the table to construct the premises. Based on the constructed premises, SIRV plans reasoning paths and generates sub-questions that explicitly indicate relevant factual evidence, performing evidence-centered reasoning. Additionally, SIRV examines the consistency between the premises and the table to focus on factual evidence, thereby reliably identifying and correcting errors in the reasoning process. Compared to state-of-the-art methods, SIRV achieves performance improvements of up to 5.24% in single-mode and 2.89% in joint reasoning, while also demonstrating excellent generalization ability and efficiency.

Extended version — https://github.com/zzxxzjydnx/SIRV

Introduction

Tables are among the most commonly used formats for storing structured information. Table Question Answering (TableQA) refers to reasoning the answers to questions by understanding and analyzing table content, playing a crucial role in tasks such as judgment (Huang et al. 2025), analysis (Zhang, Gao, and Lou 2024), and decision-making (Yang et al. 2025; Wang et al. 2024b; Chen et al. 2020). Large Language Models (LLMs) can answer questions through carefully designed prompts (Wang et al. 2024a; Jin et al. 2025; Ma et al. 2024; Sui et al. 2025), and the use of LLMs in TableQA has become a mainstream research direction (Zhang et al. 2024b, 2025a; Wang et al. 2025b; Lu et al. 2025).

*Corresponding author.

Most existing LLM-based TableQA methods primarily involve two processes: decomposition reasoning (Fu et al. 2023; Sui et al. 2024) and answer verification (Ni et al. 2023; Wang, Gan, and Qi 2025). Specifically, decomposition reasoning involves decomposing the question and gradually reasoning to obtain the final answer. Answer verification involves checking reasoning processes and generated answers to reduce the occurrence of errors. Direct Prompting (DP) and Agent are the primary answer modes in LLM-based TableQA methods. In the DP mode, LLMs perform decomposition reasoning and answer verification through a few example prompts (Zhang et al. 2023b; Guan, Huang, and Zhang 2024; Zhang et al. 2023a). In the Agent mode, LLMs generate decomposition reasoning strategies based on the current state and context, leveraging self-correction mechanisms to verify the reasoning process and answers (Li et al. 2024; Jiang et al. 2023).

However, current LLM-based methods encounter several challenges. As shown in Figure 1(a), in the decomposition reasoning process, existing methods primarily focus on the semantic splitting of the question while overlooking the factual evidence (Definition 1) present in tables. They fail to significantly reduce the difficulty for LLMs in understanding key information (i.e., "country" or "Top 10") in questions, and cannot guide LLMs towards reliable reasoning. Additionally, in the answer verification process, the lack of support from factual evidence makes it difficult for LLMs to effectively identify and correct systematic or factual errors in the reasoning process. Research indicates that simply prompting self-correction may bias the model away from generating optimal responses to the initial prompt (Huang et al. 2024). Therefore, there is an urgent need for a new evidence-centered method in TableQA, which incorporates factual evidence into every process of reasoning and verification. Then, two key questions arise: **How can we automatically extract question-relevant factual evidence from tables for different question types? How can we construct a normative and efficient, evidence-centered process for decomposition reasoning and answer verification?** A natural way is syllogism (Wan et al. 2024; Deng et al. 2023), which relies on evidential premises for reliable reasoning and verification in practical applications (Definition 2).

This paper proposes a **S**yllogism-**I**nspired **R**easoning and

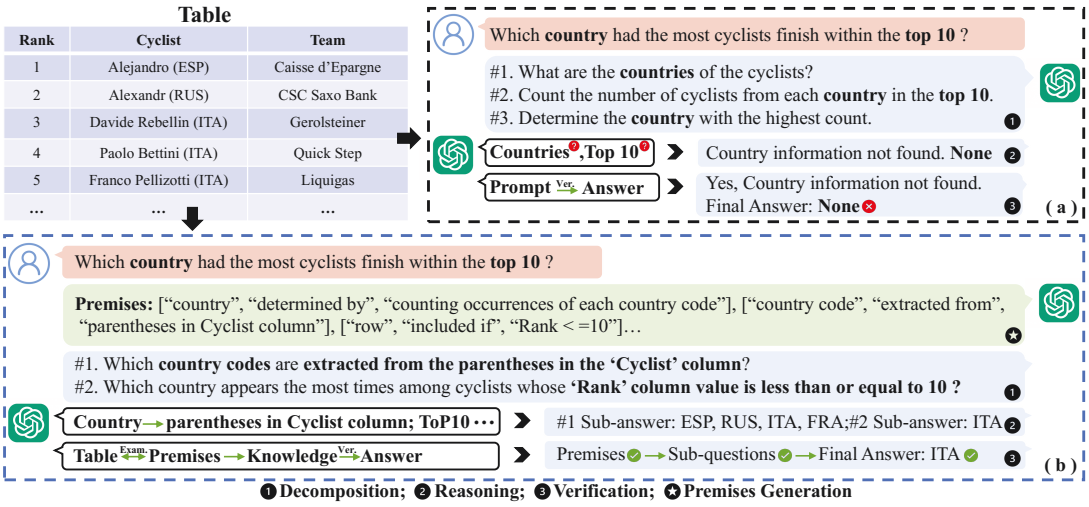

Figure 1: (a): Directly generating sub-questions through semantic splitting fails to significantly reduce the difficulty for LLMs in understanding key information. The lack of support from factual evidence makes it difficult for LLMs to identify and correct errors in the reasoning process. (b): Evidence-centered question decomposition and reasoning can guide LLMs to understand key information and focus on critical table content, thus facilitating more reliable reasoning. Meanwhile, LLMs can verify answers reliably with the support of factual evidence.

Verification (SIRV) method. As shown in Figure 1(b), unlike existing methods that directly decompose the question semantically, SIRV first extracts question-relevant factual evidence from the table to construct premises. These premises explicitly state the reasoning rules related to the questions and clearly describe the correspondence between key information and the table content. Based on premises, SIRV plans reasoning paths and generates sub-questions that explicitly indicate relevant factual evidence. Finally, SIRV integrates the premises and sub-questions to understand the reasoning rules and focus on critical table content, performing reliable step-by-step reasoning. In the answer verification process, SIRV critically examines the consistency between the premises and the table content to construct an evidential knowledge background, thereby facilitating reliable answer verification.

SIRV can perform TableQA independently using either the DP mode or the Agent mode, or by combining both to further enhance performance. We evaluate SIRV on the WikiTableQuestions (Pasupat and Liang 2015) and TabFact (Chen et al. 2020) datasets. Compared to state-of-the-art (SOTA) methods, our method achieves performance improvements of 5.24% in single-mode and 2.89% in joint reasoning, while also demonstrating excellent generalization ability and efficiency. The main contributions of this study are summarized as follows:

- We propose an evidence-centered decomposition reasoning method that can extract question-relevant factual evidence from tables and generate sub-questions explicitly indicating the relevant evidence. This method prompts LLMs to understand reasoning rules and focus on critical table content, enabling them to perform step-by-step reasoning based on factual evidence.
- We design an evidential answer verification method. This method guides LLMs to critically examine the consistency between premises and table content, emphasizing the evidential knowledge background, thereby enabling them to effectively identify and correct errors in the reasoning process.
- We are the first study to leverage the evidential concept of syllogism to promote reliable decomposition reasoning and answer verification. The proposed method applies to various answering modes and achieves SOTA performance on multiple TableQA benchmark datasets.

Preliminaries

In this section, we present the TableQA task, factual evidence, and syllogism.

TableQA. Given a question Q and a table T, LLMs are required to reason about the final answer A. This work focuses on two core TableQA tasks: question answering and fact verification. For question answering, the result may be a numerical value, a specific entry from the table, or another form of output. In fact verification, the answer is restricted to either "Yes" or "No". The process can be represented as:

$$LLMs(Q, T) \rightarrow A. \tag{1}$$

Definition 1. ***Factual Evidence*** *refers to observable or verifiable table content, or reasoning rules derived from tables that are relevant to the question. Formally, the role of factual evidence in our study can be expressed as:* $E(T) \wedge Operation^{(i)}(Input^{(i)}) \rightarrow Output^{(i)}$*, where* $E(T)$ *denotes the factual evidence in the table,* $Operation^{(i)}$ *represents the operation for each corresponding task (e.g., decomposition, reasoning, examination, verification), and*

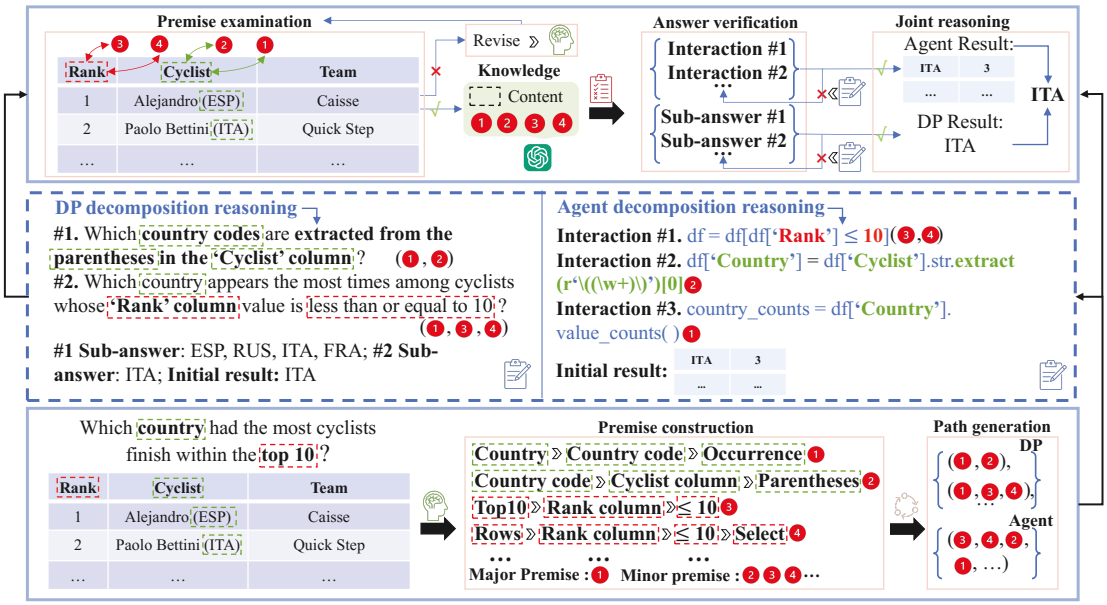

Figure 2: The overall architecture of SIRV. Premise construction, path generation, and decomposition reasoning constitute the decomposition reasoning step, which is responsible for answering the question. Premise examination and answer verification constitute the answer verification step, which is responsible for verifying the outputs. The joint reasoning step is responsible for selecting the optimal answer.

$Input^{(i)}$ and $Output^{(i)}$ refer to the input and output results for different tasks.

Definition 2. ***Syllogism*** *is a well-recognized form of logical reasoning, consisting of major premises, minor premises, and conclusions. It can be formally expressed as:* $\forall x\,(A(x) \rightarrow B(x)) \wedge A(C) \rightarrow B(C)$*, where the major premise* $\forall x\,(A(x) \rightarrow B(x))$ *states that for all* x*, if* x *has property* A*, then* x *also has property* B*. The minor premise* $A(C)$ *indicates that* C *possesses property* A*, and the conclusion* $B(C)$ *denotes that* C *possesses property* B*.*

Syllogisms observe the following principles in practical applications: (1) The conclusion must be strictly derived from the premises. (2) The reliability of the conclusion depends on the premises. Only when the premises are true and support the conclusion can the conclusion be considered reliable.

Method

Figure 2 shows the overall architecture of SIRV, which comprises three core steps: decomposition reasoning, answer verification, and joint reasoning. Specifically, the decomposition reasoning step includes constructing premises, planning reasoning paths, decomposing questions, and performing reasoning. The answer verification step involves examining the consistency between the premises and the table content and verifying the answer. Finally, the joint reasoning step integrates the reasoning results from different modes and selects the optimal answer through a voting mechanism.

Decomposition Reasoning Step

Inspired by syllogism, this study proposes an evidence-centered decomposition reasoning step to achieve more reliable reasoning. Drawing on syllogistic knowledge, SIRV treats the table as general facts and extracts factual evidence to construct premises. Subsequently, SIRV integrates the different factual evidence contained in the premises to plan the reasoning path and generates sub-questions that explicitly indicate the relevant factual evidence. Finally, SIRV performs evidence-centered reasoning based on the premises and the generated sub-questions.

Premise Construction. As shown in Figure 2, SIRV extracts reasoning rules, data, or structural information from the tables that are relevant to the question, and utilizes them as factual evidence to construct corresponding premises. By defining the table as the factual background, SIRV focuses on identifying table-oriented factual evidence, which facilitates LLMs in understanding the relationship between key information in the question and the table content. This process is formalized as:

$$LLMs(Q, T, Prompt_{P.g}) \rightarrow P_a, P_i, \tag{2}$$

where $Prompt_*$ is the prompt related to the current task, and P_a and P_i represent major and minor premises, respectively.

For the example shown in Figure 2, the constructed major premise refers to: **Countries are determined by counting the occurrences of each country code.**

The minor premise refers to: **(1) Rows are only included when the rank is less than or equal to 10. (2) The country codes come from the parentheses in the Cyclist column.**

(3) The count of occurrences for each country is based on the filtered rows. (4) The country with the highest number of occurrences is considered to have the most cyclists.

The major premise primarily clarifies the foundational reasoning rules required to answer the question, providing overarching guidance for the reasoning process. The minor premise elaborates on more specific table content, reasoning rules, or criteria for judgment. The constructed premises explain the meaning and reasoning rules of the key information, further clarifying the relationship between this information and the table content, thereby reducing the difficulty of understanding for LLMs.

Path Generation and Decomposition. Based on the constructed premises, SIRV plans reasoning paths and generates sub-questions. As illustrated in Figure 2, each sub-question consists of multiple premises, and there are explicit logical sequential relationships among these sub-questions, forming a specific reasoning path. Meanwhile, the generated sub-questions not only preserve the essential semantic information but also explicitly indicate the relevant factual evidence. They further practically guide LLMs to locate key table content and perform step-by-step reasoning according to specific reasoning rules. The process can be represented as:

$$LLMs(P_a, P_i, Prompt_D) \rightarrow S_i^Q, \tag{3}$$

where $S = \{S_1^Q, \ldots, S_i^Q, \ldots, S_k^Q\}$ is the set of generated sub-questions, and S_i^Q denotes the i-th sub-question.

SIRV emphasizes decomposing questions based on premises rather than simple semantic splitting. This process reduces the complexity of the questions and the difficulty for LLMs to understand key information, prompting LLMs to reason based on factual evidence.

Reasoning. SIRV is applicable in both the DP and Agent modes for answering questions. In the DP mode, SIRV reasons about sub-questions step by step based on premises, thereby obtaining the initial result. In the Agent mode, SIRV leverages the factual evidence in the premises as guidance for interacting with the table. It then continuously observes the current execution state and generates corresponding code to interact with the table, thus obtaining the initial result. It is crucial to note that the factual background of the premises is the table. The explicit descriptions of factual evidence in the premises can provide guidance for the DP mode. Meanwhile, the reasoning rules contained in the premises typically involve strategies for interacting with the table, offering significant support for code generation in the Agent mode. The process can be represented as:

$$LLMs(S_i^Q, P_a, P_i, Prompt_{DP}) \rightarrow A_i^{DP}, \tag{4}$$

$$LLMs(Q, P_a, P_i, Prompt_{Agent}) \rightarrow A_i^{Agent}, \tag{5}$$

where $A^{DP} = \{A_1^{DP}, \ldots, A_i^{DP}, \ldots, A_k^{DP}\}$ is the set of sub-question answers in the DP mode, and A_i^{DP} denotes the i-th sub-question and sub-answer. $A^{Agent} = \{A_1^{Agent}, \ldots, A_i^{Agent}, \ldots, A_k^{Agent}\}$ is the set of interaction results in the Agent mode, and A_i^{Agent} denotes the i-th interaction and result.

SIRV performs reliable reasoning in different modes based on premises. This evidence-centered reasoning form is similar to syllogism, where the conclusion must be strictly derived from the premises. Meanwhile, this form also lays the foundation for reliable answer verification.

Answer Verification Step

This study proposes an evidence-centered answer verification step to enhance the reliability of answer verification. Specifically, SIRV stimulates thinking and constructs an evidential knowledge background by critically examining the consistency between the premises and the table content, and then performs answer verification based on factual evidence.

Premise Examination. As shown in Figure 2, SIRV first examines the premises. Given that the premises are composed of explicit factual evidence, SIRV can easily examine their consistency with the table content. If the premises are consistent with the data, SIRV integrates both the validated premises and the critical table content into its knowledge background, ensuring reliable verification. If discrepancies are found, SIRV revises the premises by re-examining the table content, updates its knowledge background accordingly, and then verifies the answer. This examination process enables SIRV to focus on factual evidence and critical table content, thereby reliably verifying the answer. The process can be represented as:

$$LLMs(P_a, P_i, Prompt_{P.v}) \rightarrow V^P, \tag{6}$$

$$P_a, P_i = \begin{cases} P_a, P_i & \text{if } V^P = \text{True}, \\ LLMs(Q, T, Prompt_{P.r}) & \text{otherwise}, \end{cases} \tag{7}$$

where V^P is the result of premise examination.

Answer Verification. Once the premises are validated and factual evidence, along with critical table content, are integrated into the knowledge background, SIRV subsequently verifies the sub-questions, sub-answers, interaction results, and the initial results.

$$LLMs(P_a, P_i, A_i^{DP}, Prompt_{DP.v}) \rightarrow V_i^{DP}, \tag{8}$$

$$LLMs(P_a, P_i, A_i^{Agent}, Prompt_{Agent.v}) \rightarrow V_i^{Agent}, \tag{9}$$

where $V^{DP} = \{V_1^{DP}, \ldots, V_i^{DP}, \ldots, V_k^{DP}\}$ denotes the set of verification results in the DP mode, and V_i^{DP} denotes the verification result of the i-th sub-question. $V^{Agent} = \{V_1^{Agent}, \ldots, V_i^{Agent}, \ldots, V_k^{Agent}\}$ denotes the set of verification results in the agent mode, and V_i^{Agent} denotes the verification result of the i-th interaction.

SIRV will output the final answer when all outputs have been confirmed to be correct. $\forall i \in \{1, \ldots, k\}$, the process can be represented as:

$$A_i^{DP} = \begin{cases} A_i^{DP} & \text{if } V_i^{DP} = \text{True}, \\ LLMs(S_i^Q, P_a, P_i, Prompt_{DP.r}) & \text{otherwise}, \end{cases} \tag{10}$$

$$A_i^{Agent} = \begin{cases} A_i^{Agent} & \text{if } V_i^{Agent} = \text{True}, \\ LLMs(A_i^{Agent}, P_a, P_i, Prompt_{Agent.r}) & \text{otherwise}, \end{cases} \tag{11}$$

$$A_{Final}^{DP}, A_{Final}^{Agent} = A_k^{DP}, A_k^{Agent}. \tag{12}$$

	Methods	WTQ	TableFact	Mean
PLM	TAPEX-large	59.10	84.20	71.65
	T5-3B	50.60	83.68	67.14
	TabLaP	76.60	-	-
LLM	StructGPT	57.00	87.30	72.15
	BINDER	64.60	85.10	74.85
	DATER	65.90	85.60	75.75
	ReAcTable	68.00	86.10	77.05
	CHAIN-OF-TABLE	59.94	80.20	70.07
	S2L	66.00	86.20	76.10
	Norm-DP&Agent	73.65	88.50	81.08
	TIDE-DP&Agent	75.00	89.82	82.41
Our	SIRV-DP&Agent	**77.89** 2.89 ↑	**91.85** 2.03 ↑	**84.87** 2.46 ↑

Table 1: Exact match accuracy on the two datasets. The boldface indicates the best results, and the arrows indicate performance changes compared to the LLM-based baseline.

Methods	Exact Match Accuracy		
	WTQ	TabFact	Mean
Norm-DP	66.99	-	-
TIDE-DP	66.51	81.32	73.91
SIRV-DP	**67.38** 0.39 ↑	**85.97** 4.65 ↑	**76.67** 2.76 ↑
Norm-Agent	63.77	-	-
TIDE-Agent	68.72	88.19	78.45
SIRV-Agent	**73.96** 5.24 ↑	**90.86** 2.67 ↑	**82.41** 3.96 ↑

Table 2: Comparison results on the DP and Agent modes. The boldface indicates the best results, and the arrows indicate performance changes.

Joint Reasoning Step

To effectively integrate results from different modes and avoid potential errors that may arise from single-pass reasoning, SIRV employs a majority voting mechanism to select the optimal final answer. Specifically, SIRV generates five candidate answers in both DP and Agent modes. Subsequently, SIRV randomly selects a specified number of answers from the candidates of both modes according to predefined hyperparameters. Finally, SIRV uses the majority voting mechanism to choose the answer with the highest frequency as the final result.

$$A_{\text{final}} = \text{Majority Vote}(A^{DP}_{Final}, A^{Agent}_{Final}). \quad (13)$$

Experiment

Datasets and Evaluation

Datasets. We evaluate the performance of SIRV using two widely recognized TableQA datasets: WikiTableQuestions and TabFact. WikiTableQuestions (WTQ) contains a large number of question-answer pairs based on structured tables, covering various complex question types, with a test set consisting of 4,344 samples. TabFact is mainly used for fact verification tasks on structured tables, with a test set consisting of 2,024 samples. We conduct relevant experiments on the test sets of both WTQ and TabFact.

Evaluation. We use exact match accuracy as the evaluation metric, which is widely used in most LLM-based TableQA methods (Yang et al. 2025; Liu, Wang, and Chen 2024; Cheng et al. 2023). This metric effectively measures whether the answers generated by the model exactly match the true answers.

Implementation Details

Following previous studies (Cheng et al. 2023; Ye et al. 2023; Liu, Wang, and Chen 2024; Yang et al. 2025), SIRV adopts GPT-3.5 as the backbone model in both the DP and Agent modes. SIRV generates five answers for each test sample in both modes. We compare SIRV with several baseline methods, including TAPEX-LARGE (Liu et al. 2022), T5-3B (Xie et al. 2022), STRUCTGPT (Jiang et al. 2023), BINDER (Cheng et al. 2023), DATER (Ye et al. 2023), REACTABLE (Zhang et al. 2024b), CHAIN-OF-TABLE (Wang et al. 2024b), NORM-DP&AGENT (Liu, Wang, and Chen 2024), TIDE-DP&AGENT (Yang et al. 2025), TABLAP (Wang, Qi, and Gan 2025), and S2L (Wang et al. 2025a).

Main Results

Table 1 presents the comparison results between SIRV and baseline methods on the WTQ and TabFact datasets. Specifically, SIRV demonstrates better performance than PLM methods. Meanwhile, SIRV outperforms the SOTA LLMs method TIDE in accuracy by 2.89% and 2.03% on the two datasets, respectively, with a mean improvement of 2.46%. Since the best performance on TabFact is already high, the improvement of 2.03% should be interpreted as a proportion of the scope of further improvement possible (Yang et al. 2025), 2.03/(100 - 89.82), which is $\approx$ 19.94%. In fact, the above improvement surpasses all baseline methods. These experimental results validate the critical role of evidence-centered decomposition reasoning and answer verification steps in enhancing TableQA performance.

Table 2 reports the accuracy of SIRV when using DP or Agent mode for reasoning separately. As shown, SIRV achieves the best results in both modes. In particular, SIRV-Agent improves the accuracy on the WTQ dataset by 5.24%, while SIRV-DP achieves 4.65/(100 - 81.32) $\approx$ 24.89% improvement in accuracy on the TabFact dataset. Additionally, SIRV-Agent shows more significant improvements on both datasets. This can be attributed to two key factors. First, the premises typically contain table-oriented reasoning rules, which can guide the model in generating code and performing reasoning accurately. Second, SIRV-Agent interacts with tables through code, making it more suitable for handling structured tables.

Ablation Study

In this subsection, we again scrutinize the impact of the various components of SIRV. We conduct ablation experiments

Core components				WTQ (Exact Match Accuracy)			TabFact (Exact Match Accuracy)		
SIRV-DR$_{DP}$	SIRV-AV$_{DP}$	SIRV-DR$_{Agent}$	SIRV-AV$_{Agent}$	DP	Agent	DP&Agent	DP	Agent	DP&Agent
✓	✓	✓	✓	**67.38**	**73.96**	**77.89**	**85.97**	**90.86**	**91.85**
✓	✗	✓	✗	65.20(2.18↓)	71.43(2.53↓)	74.31(3.58↓)	83.20(2.77↓)	88.45(2.41↓)	87.70(4.15↓)
✗	✓	✗	✓	62.11(5.27↓)	68.01(5.95↓)	72.39(5.50↓)	82.25(3.72↓)	86.31(4.55↓)	87.26(4.59↓)
✗	✗	✓	✓	61.35(6.03↓)	73.96(0.00↓)	73.17(4.72↓)	80.45(5.52↓)	90.86(0.00↓)	89.97(1.88↓)
✓	✓	✗	✗	67.38(0.00↓)	66.84(7.12↓)	71.91(5.98↓)	85.97(0.00↓)	84.95(5.91↓)	87.42(4.43↓)

Table 3: Ablation study results. The boldface indicates the best results, and the arrows indicate performance changes.

on the WTQ and TabFact datasets, evaluating the effects of using only the evidence-centered decomposition reasoning step (SIRV-DR), using only the evidence-centered answer verification step (SIRV-AV), and using only the standard chain-of-thought (Wei et al. 2022).

As shown in Table 3, the proposed method achieves better results across different modes and datasets. When the SIRV-AV is removed, there is a noticeable decline in accuracy in both the single-mode and joint reasoning. Similarly, the removal of the SIRV-DR has a substantial negative impact on performance, particularly in the Agent mode, where the WTQ accuracy decreases by 5.95%. Furthermore, when using only the standard chain-of-thought, the overall performance further declines, with the Agent mode experiencing particularly pronounced effects. The above results further validate the essential contributions of each core component to improving accuracy and reveal the importance of factual evidence in decomposing reasoning and answer verification.

Decomposition Reasoning Analysis

In this subsection, we evaluate the premise construction capability of SIRV and analyze its reasoning efficiency. Specifically, we further analyze the content of the premises and the reasoning process. First, we count the distinct table columns mentioned in the premises and sub-questions. The mentioned columns serve as explicit evidence elements, reflecting the connection between the question and the table content. Secondly, we analyze the number of actions performed by SIRV during the reasoning process, including both the number of answers and interactions. Fewer actions indicate that SIRV can reason about the question more efficiently.

Number of mentioned columns. Figure 3 shows the number of distinct columns mentioned in premises and sub-questions. Given that the sub-questions are generated based on the premises, the columns mentioned in the sub-questions also reflect the factual evidence contained in the premises. For most questions in WTQ and TabFact, the number of columns explicitly mentioned in the premises is ≥ 2. Similarly, the number of columns mentioned in the sub-questions is typically ≥ 1. These results demonstrate the generality of our method in constructing premises. SIRV can construct premises containing factual evidence for most questions.

Number of actions. Figure 4 shows the mean number of answers and interactions of SIRV across the five reasoning processes. SIRV performs better with fewer actions than the previous SOTA method, TIDE. In DP mode, the mean number of answers in a complete reasoning process of SIRV is ≤ 3. In Agent mode, the mean number of interactions of SIRV with the table is typically ≤ 3. This indicates that SIRV can perform concise and efficient reasoning based on premises and sub-questions containing factual evidence.

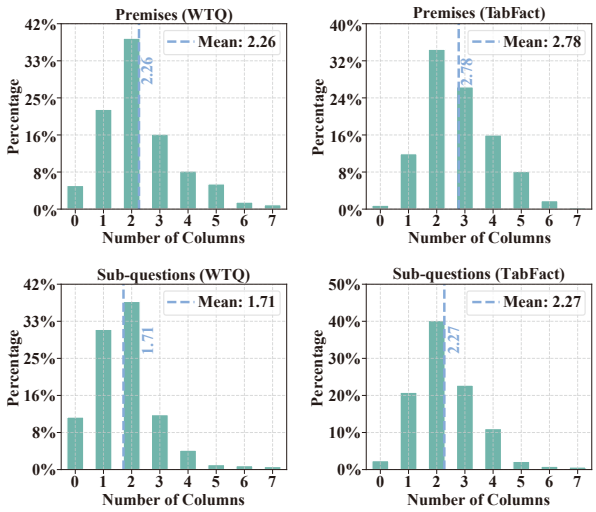

Figure 3: Number of mentioned columns in the premises and sub-questions.

Answer Verification Analysis

The evidence-centered answer verification step is a crucial component that enables SIRV to achieve SOTA performance. To further validate the reliability of this step, we analyze its capability to correct erroneous answers.

Table 4 shows the capability of SIRV-AV to correct erroneous answers. Specifically, the answer correction rates of SIRV-AV on the two datasets are 39.25% and 72.17%, respectively. When only prompting the LLMs to perform self-correction, the mean answer correction rate of SIRV decreases by 8.32%. This phenomenon demonstrates that SIRV can indeed achieve a more reliable answer verification based on factual evidence.

Joint Reasoning Analysis

SIRV is capable of joint reasoning to enhance performance. In this subsection, we further analyze how different combinations of candidate answer quantities affect accuracy.

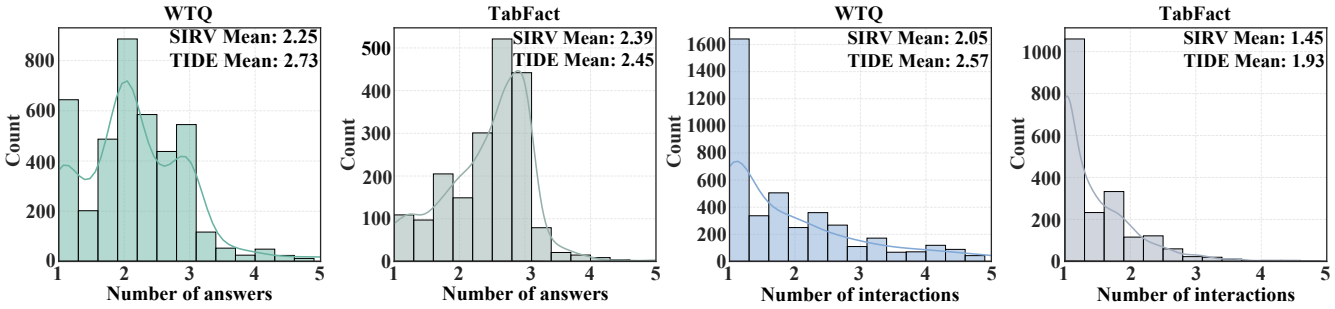

Figure 4: The mean number of actions in the five reasoning processes. The number of answers refers to the quantity of answers involved in the DP mode, while the number of interactions refers to the quantity of interaction steps involved in the Agent mode.

Methods	Answer Correction Rate (%)		
	WTQ	TabFact	Mean
SIRV-AV	**39.25**	**72.17**	**55.71**
Self-Correction	29.53	65.25	47.39
	9.72 ↓	6.92 ↓	8.32 ↓

Table 4: Answer verification capability test results. The boldface indicates the best results, and the arrows indicate performance changes.

Answer Count		WTQ	TabFact	Mean
DP	Agent			
1	1	70.07	87.53	78.80
3	3	76.70	90.91	83.81
5	5	**77.89**	91.21	84.55
1	3	75.83	91.60	83.72
1	5	76.15	91.75	83.95
3	1	70.63	87.74	79.19
3	5	77.72	**91.85**	**84.79**
5	1	70.73	87.83	79.28
5	3	75.30	89.13	82.22

Table 5: Exact match accuracy of different combinations of candidate answers. The boldface indicates the best results.

As shown in Table 5, increasing the number of candidate answers improves the performance of SIRV in both modes. When the number of candidate answers increases from 1 to 5, the accuracy improves by 5.75%. Meanwhile, we find that the optimal combination ratio of answers for SIRV on TabFact is 3:5, indicating that an excessive number of candidate answers leads to answer dispersion, reducing the effectiveness of the majority voting mechanism.

Related Work

TableQA. LLM-based TableQA methods (Zhao et al. 2024; Kong et al. 2024; Wu et al. 2025a; Dong, Hu, and Cao 2025; Contalbo et al. 2025; Zhang et al. 2024a; Fang et al. 2024) typically employ DP mode and Agent mode to reason the answers to questions. Meanwhile, LLM-based TableQA methods mainly involve decomposition reasoning and answer verification processes.

Decomposition reasoning. Decomposition reasoning involves decomposing the original question into multiple subquestions, enabling step-by-step reasoning to reduce complexity (Chen 2023; Cheng et al. 2023; Ye et al. 2023; Zhang et al. 2024b). Recently, a series of decomposition reasoning methods have been proposed and developed, such as Least-to-Most (Zhou et al. 2023), DecomP (Khot et al. 2023), and CHAIN-OF-TABLE (Wang et al. 2024b). However, existing methods overlook factual evidence during the decomposition process, failing to guide LLMs in understanding key information within the question and conducting reliable reasoning. Drawing inspiration from syllogism, we propose an evidence-centered decomposition reasoning method.

Answer verification. Answer verification involves checking the correctness of reasoning processes and generated answers, making necessary corrections to improve the overall accuracy and reliability (Yang et al. 2025; Zhang et al. 2025b). LLM-based TableQA methods typically perform answer verification and self-correction based on specific task prompts or state changes (Wang, Gan, and Qi 2025; Wu et al. 2025b). Research (Huang et al. 2024) indicates that simply prompting LLMs to self-correct may bias the model away from generating optimal responses to the initial prompt, leading to a decline in verification performance. Meanwhile, verification without supporting evidence is often unreliable. Therefore, we propose a syllogism-inspired answer verification method that can verify answers based on evidence.

Conclusion

This paper proposes a TableQA approach called SIRV. Specifically, SIRV can better understand key information and focus on critical table content based on factual evidence, enabling reliable step-by-step reasoning. Additionally, SIRV constructs an evidential knowledge background by examining premises, thereby performing answer verification based on factual evidence. Experimental results indicate that SIRV achieves SOTA performance in TableQA. We believe that the idea of being evidence-centered is practically meaningful for TableQA task, and we hope that this work can inspire further research on TableQA from an evidence-centered perspective.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62302086.

References

Chen, W. 2023. Large Language Models are few (1)-shot Table Reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, 1120–1130.

Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; and Wang, W. Y. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations*.

Cheng, Z.; Xie, T.; Shi, P.; Li, C.; Nadkarni, R.; Hu, Y.; Xiong, C.; Radev, D.; Ostendorf, M.; Zettlemoyer, L.; et al. 2023. Binding Language Models in Symbolic Languages. In *The Eleventh International Conference on Learning Representations*.

Contalbo, M. L.; Pederzoli, S.; Del Buono, F.; Valeria, V.; Guerra, F.; and Paganelli, M. 2025. GRI-QA: a Comprehensive Benchmark for Table Question Answering over Environmental Data. In *Findings of the Association for Computational Linguistics: ACL 2025*, 15764–15779.

Deng, W.; Pei, J.; Kong, K.; Chen, Z.; Wei, F.; Li, Y.; Ren, Z.; Chen, Z.; and Ren, P. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13997–14009.

Dong, H.; Hu, Y.; and Cao, Y. 2025. Reasoning and Retrieval for Complex Semi-structured Tables via Reinforced Relational Data Transformation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1382–1391.

Fang, X.; Xu, W.; Tan, F. A.; Zhang, J.; Hu, Z.; Qi, Y.; Nickleach, S.; Socolinsky, D.; Sengamedu, S.; and Faloutsos, C. 2024. Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding – A Survey. arXiv:2402.17944.

Fu, Y.; Peng, H.; Sabharwal, A.; Clark, P.; and Khot, T. 2023. COMPLEXITY-BASED PROMPTING FOR MULTI-STEP REASONING. In *11th International Conference on Learning Representations, ICLR 2023*.

Guan, C.; Huang, M.; and Zhang, P. 2024. Mfort-qa: Multi-hop few-shot open rich table question answering. In *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence*, 434–442.

Huang, J.; Chen, X.; Mishra, S.; Zheng, H. S.; Yu, A. W.; Song, X.; and Zhou, D. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *The Twelfth International Conference on Learning Representations*.

Huang, S.; Gu, Y.; Li, Z.; Hu, X.; Qing, L.; and Xu, G. 2025. StructFact: Reasoning Factual Knowledge from Structured Data with Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 7521–7552.

Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, W. X.; and Wen, J.-R. 2023. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9237–9251.

Jin, R.; Li, Y.; Qi, G.; Hu, N.; Li, Y.-F.; Chen, J.; Wang, J.; Chen, Y.; Min, D.; and Bi, S. 2025. Hegta: Leveraging heterogeneous graph-enhanced large language models for few-shot complex table understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24294–24302.

Khot, T.; Trivedi, H.; Finlayson, M.; Fu, Y.; Richardson, K.; Clark, P.; and Sabharwal, A. 2023. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. In *The Eleventh International Conference on Learning Representations*.

Kong, K.; Zhang, J.; Shen, Z.; Srinivasan, B.; Lei, C.; Faloutsos, C.; Rangwala, H.; and Karypis, G. 2024. OpenTab: Advancing Large Language Models as Open-domain Table Reasoners. In *The Twelfth International Conference on Learning Representations*.

Li, P.; He, Y.; Yashar, D.; Cui, W.; Ge, S.; Zhang, H.; Rifinski Fainman, D.; Zhang, D.; and Chaudhuri, S. 2024. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proceedings of the ACM on Management of Data*, 2(3): 1–28.

Liu, Q.; Chen, B.; Guo, J.; Ziyadi, M.; Lin, Z.; Chen, W.; and Lou, J.-G. 2022. TAPEX: Table Pre-training via Learning a Neural SQL Executor. arXiv:2107.07653.

Liu, T.; Wang, F.; and Chen, M. 2024. Rethinking Tabular Data Understanding with Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 450–482.

Lu, W.; Zhang, J.; Fan, J.; Fu, Z.; Chen, Y.; and Du, X. 2025. Large language model for table processing: A survey. *Frontiers of Computer Science*, 19(2): 192350.

Ma, Z.; Zhang, B.; Zhang, J.; Yu, J.; Zhang, X.; Zhang, X.; Luo, S.; Wang, X.; and Tang, J. 2024. Spreadsheetbench: Towards challenging real world spreadsheet manipulation. *Advances in Neural Information Processing Systems*, 37: 94871–94908.

Ni, A.; Iyer, S.; Radev, D.; Stoyanov, V.; Yih, W.-t.; Wang, S.; and Lin, X. V. 2023. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*, 26106–26128. PMLR.

Pasupat, P.; and Liang, P. 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)*, 1470–1480.

Sui, Y.; He, Y.; Liu, N.; He, X.; Wang, K.; and Hooi, B. 2025. FiDeLiS: Faithful Reasoning in Large Language Model for Knowledge Graph Question Answering. arXiv:2405.13873.

Sui, Y.; Zou, J.; Zhou, M.; He, X.; Du, L.; Han, S.; and Zhang, D. 2024. TAP4LLM: Table Provider on Sampling, Augmenting, and Packing Semi-structured Data for Large

Language Model Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10306–10323.

Wan, Y.; Wang, W.; Yang, Y.; Yuan, Y.; Huang, J.-t.; He, P.; Jiao, W.; and Lyu, M. 2024. LogicAsker: Evaluating and Improving the Logical Reasoning Ability of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2124–2155.

Wang, Y.; Chen, L.; Cai, S.; Xu, Z.; and Zhao, Y. 2024a. Revisiting automated evaluation for long-form table question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14696–14706.

Wang, Y.; Cheng, S.; Sun, Z.; Li, P.; and Liu, Y. 2025a. Leveraging Language-based Representations for Better Solving Symbol-related Problems with Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, 5544–5557.

Wang, Y.; Gan, J.; and Qi, J. 2025. TabSD: Large Free-Form Table Question Answering with SQL-Based Table Decomposition. arXiv:2502.13422.

Wang, Y.; Qi, J.; and Gan, J. 2025. Accurate and regret-aware numerical problem solver for tabular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12775–12783.

Wang, Z.; Yang, W.; Zhou, K.; Zhang, Y.; and Jia, W. 2025b. Retqa: A large-scale open-domain tabular question answering dataset for real estate sector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25452–25460.

Wang, Z.; Zhang, H.; Li, C.-L.; Eisenschlos, J. M.; Perot, V.; Wang, Z.; Miculicich, L.; Fujii, Y.; Shang, J.; Lee, C.-Y.; et al. 2024b. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The 12th International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, J.; Yang, L.; Li, D.; Ji, Y.; Okumura, M.; and Zhang, Y. 2025a. MMQA: Evaluating LLMs with multi-table multi-hop complex questions. In *The Thirteenth International Conference on Learning Representations*, 1.

Wu, X.; Yang, J.; Chai, L.; Zhang, G.; Liu, J.; Du, X.; Liang, D.; Shu, D.; Cheng, X.; Sun, T.; et al. 2025b. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25497–25506.

Xie, T.; Wu, C. H.; Shi, P.; Zhong, R.; Scholak, T.; Yasunaga, M.; Wu, C.-S.; Zhong, M.; Yin, P.; Wang, S. I.; Zhong, V.; Wang, B.; Li, C.; Boyle, C.; Ni, A.; Yao, Z.; Radev, D.; Xiong, C.; Kong, L.; Zhang, R.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. arXiv:2201.05966.

Yang, Z.; Du, Z.; Zhang, M.; Du, W.; Chen, J.; Duan, Z.; and Zhao, S. 2025. Triples as the Key: Structuring Makes Decomposition and Verification Easier in LLM-based TableQA. In *The Thirteenth International Conference on Learning Representations*.

Ye, Y.; Hui, B.; Yang, M.; Li, B.; Huang, F.; and Li, Y. 2023. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 174–184.

Zhang, L.; Wu, Y.; Mo, F.; Nie, J.-Y.; and Agrawal, A. 2023a. MoqaGPT: Zero-Shot Multi-modal Open-domain Question Answering with Large Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1195–1210.

Zhang, W.; Jin, L.; Zhu, Y.; Chen, J.; Huang, Z.; Wang, J.; Hua, Y.; Liang, L.; and Chen, H. 2025a. Trustuqa: A trustful framework for unified structured data question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25931–25939.

Zhang, W.; Wang, Y.; Song, Y.; Wei, V. J.; Tian, Y.; Qi, Y.; Chan, J. H.; Wong, R. C.-W.; and Yang, H. 2024a. Natural language interfaces for tabular data querying and visualization: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 6699–6718.

Zhang, X.; Luo, S.; Zhang, B.; Ma, Z.; Zhang, J.; Li, Y.; Li, G.; Yao, Z.; Xu, K.; Zhou, J.; Zhang-Li, D.; Yu, J.; Zhao, S.; Li, J.; and Tang, J. 2025b. TableLLM: Enabling Tabular Data Manipulation by LLMs in Real Office Usage Scenarios. arXiv:2403.19318.

Zhang, Y.; Henkel, J.; Floratou, A.; Cahoon, J.; Deep, S.; and Patel, J. M. 2024b. ReAcTable: Enhancing ReAct for Table Question Answering. *Proceedings of the VLDB Endowment*, 17(8): 1981–1994.

Zhang, Z.; Gao, Y.; and Lou, J.-G. 2024. e5: Zero-shot hierarchical table analysis using augmented LLMs via explain, extract, execute, exhibit and extrapolate. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1244–1258.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023b. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhao, Y.; Chen, L.; Cohan, A.; and Zhao, C. 2024. TaPERA: enhancing faithfulness and interpretability in long-form table QA by content planning and execution-based reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12824–12840.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; et al. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.