Table-Critic: A Multi-Agent Framework for Collaborative Criticism and Refinement in Table Reasoning

Anonymous ACL submission

Abstract

Despite the remarkable capabilities of large language models (LLMs) in various reasoning tasks, they still struggle with table reasoning tasks, particularly in maintaining consistency throughout multi-step reasoning processes. While existing approaches have explored various decomposition strategies, they often lack effective mechanisms to identify and correct errors in intermediate reasoning steps, leading to cascading error propagation. To address these issues, we propose Table-Critic, a novel multi-agent framework that facilitates collaborative criticism and iterative refinement of the reasoning process until convergence to correct solutions. Our framework consists of four specialized agents: a Judge for error identification, a Critic for comprehensive critiques, a Refiner for process improvement, and a Curator for pattern distillation. To effectively deal with diverse and unpredictable error types, we introduce a self-evolving template tree that systematically accumulates critique knowledge through experience-driven learning and guides future reflections. Extensive experiments have demonstrated that Table-Critic achieves substantial improvements over existing methods, achieving superior accuracy and error correction rates while maintaining computational efficiency and lower solution degradation rate.

1 Introduction

Despite significant advances in various reasoning tasks (Plaat et al., 2024; Yu et al., 2024; Chen et al., 2024a,b; Guo et al., 2025), large language models (LLMs) (Yang et al., 2024a; Dubey et al., 2024; Anthropic, 2024; Mesnard et al., 2024; Hurst et al., 2024) face substantial challenges in handling semi-structured data, such as table reasoning tasks, as they require both understanding of tabular structures and precise localization of relevant entries in redundant and noisy information (Zhao et al., 2024; Chen et al., 2024c; Zhang et al., 2025).

Existing approaches address these challenges through various decomposition strategies. For example, Binder (Cheng et al., 2022) decomposes complex questions into executable sub-programs (i.e., SQL or Python), while approaches such as Dater (Ye et al., 2023) and Chain-of-Table (Wang et al., 2024) focus on dynamic table decomposition for context-aware reasoning. Although these decomposition-based methods have demonstrated promising performance, they suffer from a critical limitation: the lack of effective mechanisms to criticize and refine the intermediate reasoning steps. This deficiency inevitably leads to error propagation throughout the reasoning process, significantly affecting the accuracy of final predictions.

However, recent studies (Madaan et al., 2023; Yang et al., 2024b) have revealed that while LLMs possess self-reflection capabilities to some extent, their self-reflection often lacks reliability and consistency. Simply forcing LLMs to engage in self-reflection may introduce additional biases, especially in table reasoning tasks, wherein models tend to either rationalize their previous erroneous reasoning or over-criticize correct steps, rather than identifying genuine errors (Zheng et al., 2024; Chen et al., 2025).

To address these issues, we propose Table-Critic, a multi-agent framework that introduces specialized agents to collaboratively criticize and refine the reasoning process in a step-by-step manner. Specifically, our Table-Critic simulates human-like reflective behaviors through four targeted agents: a Judge that identifies potential errors, a Critic that provides detailed suggestions, a Refiner that refines the entire reasoning process, and a Curator that distills critique patterns to guide future reflection. The collaborative strategy among multiple agents is motivated by our two insights: **(1) LLMs demonstrate proficiency in identifying and refining the first erroneous steps, yet tend to make other mistakes in subsequent steps, particularly**

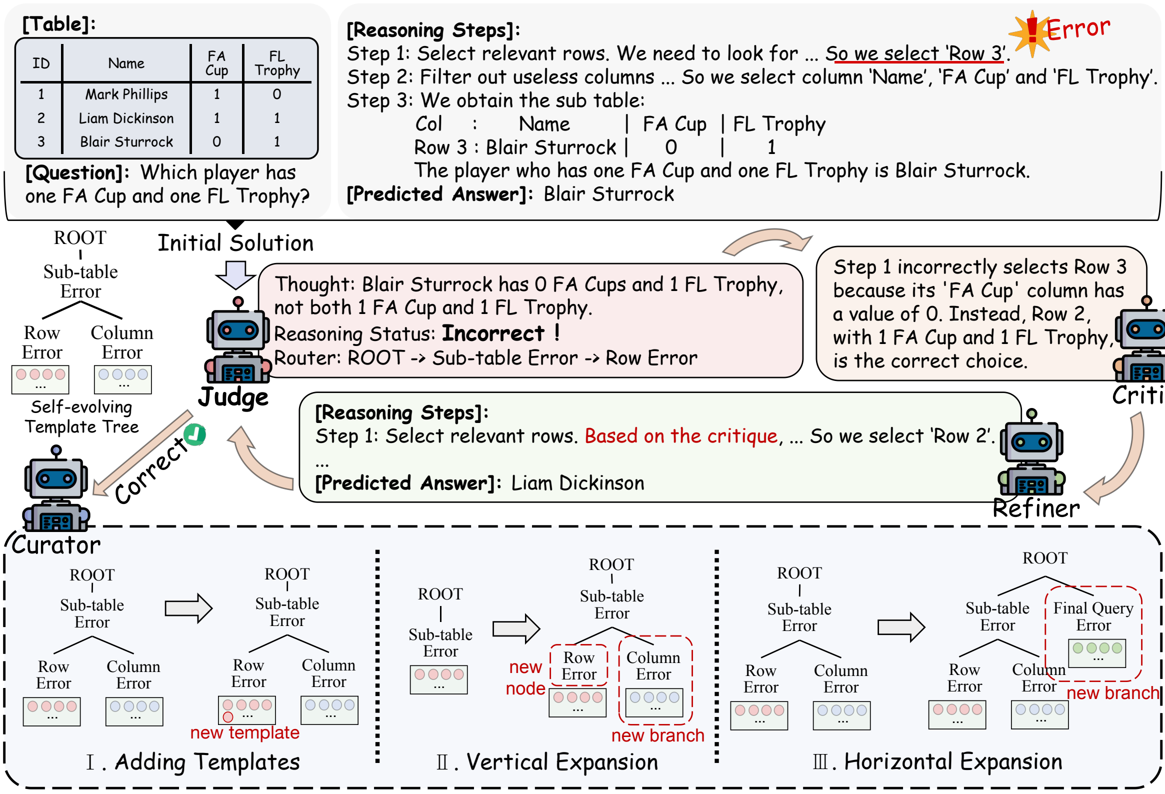


Figure 1: An illustration of Table-Critic, a multi-agent framework for table reasoning tasks, where the Judge identifies errors, the Critic provides detailed critique, the Refiner corrects the reasoning process, and the Curator updates a self-evolving template tree to accumulate critique knowledge and improve future performance.

when dealing with complex problems. This observation motivates our multi-turn design where different agents continuously monitor and refine the reasoning process until the Judge verifies its correctness. **(2) The diversity and unpredictability of error types in the reasoning process make it challenging for LLMs to effectively identify them based solely on their inherent knowledge.** This insight motivates the development of a dynamic template repository that categorizes and stores critique templates by error types, allowing our multi-agent system to systematically accumulate critique knowledge. Specifically, the Curator maintains a self-evolving template tree by expanding branches or adding templates after the entire reflection, while the Judge routes through the tree based on the identified reasoning errors to locate appropriate templates for assisting the Critic in generating high-quality critique, thereby facilitating subsequent refinement. Through this self-evolving template tree mechanism, our system continuously accumulates and distills critique patterns from previous experiences, enabling more effective error identification beyond LLMs' inherent capabilities. This experience-driven approach ensures continuous improvement in the quality and consistency.

Our contributions are summarized as follows:

- We introduce Table-Critic, a novel multi-agent framework where specialized agents collaboratively criticize and refine the reasoning process for complex table reasoning tasks.
- We design a multi-turn refinement mechanism where different agents continuously monitor and improve the reasoning process, effectively mitigating error propagation in multi-step reasoning.
- We introduce a self-evolving template tree that systematically accumulates and organizes critique knowledge, enabling our system to effectively handle emerging error types through experience-driven learning.
- Extensive experiments demonstrate that Table-Critic significantly outperforms existing methods and exhibits substantial advantages over majority voting under comparable or even superior computational costs.

2 Related Work

Table Reasoning. Table reasoning, which requires joint understanding of semi-structured tables and questions, has evolved through several paradigms. Early approaches focused on developing specialized models through fine-tuning (Yin et al., 2020; Liu et al., 2021; Gu et al., 2022), while recent work has shifted towards leveraging large language models (LLMs) in few-shot learning (Chen et al., 2024c; Zhao et al., 2024). To handle complex reasoning tasks, decomposition-based methods have emerged as a promising direction. These methods break down complex tasks into manageable steps, either through program execution (Cheng et al., 2022) or context-aware table partitioning (Ye et al., 2023; Wang et al., 2024). However, a critical limitation of existing approaches is their inability to effectively critique and refine intermediate reasoning steps, leading to error propagation. In contrast, our Table-Critic framework addresses this limitation by introducing systematic critique and refinement mechanisms throughout the reasoning process.

Self-Reflection. Recent studies have revealed that while LLMs possess inherent self-reflection capabilities, they often suffer from reliability and consistency issues (Madaan et al., 2023; Yang et al., 2024b). Simply enforcing self-reflection can be counterproductive, as models tend to either rationalize their errors or excessively critique correct reasoning steps (Zheng et al., 2024; Chen et al., 2025). To address these limitations, our Table-Critic introduces a structured approach through: (1) a multi-agent framework where specialized agents collaborate to provide targeted critiques, and (2) a self-evolving template tree that systematically accumulates and organizes critique knowledge. This design effectively overcomes the inherent limitations of LLMs' reflection capabilities while maintaining reliable and consistent error identification.

3 Table-Critic

3.1 Overview

To effectively implement the human-like correction process in multi-step reasoning, we propose a collaborative multi-agent framework, Table-Critic. As illustrated in Figure 1, this framework decomposes the complex reasoning refinement task into four specialized functions: error detection (Judge), critique generation (Critic), reasoning refinement (Refiner), and experience learning (Curator). These agents work in concert to progressively improve reasoning quality while accumulating valuable correction experiences. Specifically, given a table $\mathbb{T}$ and a question q, these agents iteratively refine the initial reasoning chain $\tau = \{s_1, s_2, ..., s_n\}$ until reaching a satisfactory solution.[1] The refinement process is guided by a self-evolving template tree $\mathcal{T}$ that systematically accumulates critique patterns from past experiences.

3.2 Multiple Agents

Inspired by human-like correction behavior, we design four specialized agents—Judge, Critic, Refiner, and Curator—to facilitate criticizing and refining in multi-step reasoning. We use specific instructions to prompt LLM (π) to execute the corresponding operations. Formally, we define each agent as follows:

Judge ($\mathcal{A}^j$). The Judge agent is responsible for identifying potential errors in the reasoning process. Given a table $\mathbb{T}$, question q, current reasoning chain τ, and the template tree $\mathcal{T}$, it analyzes each reasoning step and determines the specific error type if any exists. Based on the identified error type (if exists), the Judge routes through the template tree $\mathcal{T}$ to locate appropriate templates for guiding the subsequent critic agent. Formally, the Judge agent operates as:

$$E, P, R = \pi(\mathbb{T}, q, \tau, \mathcal{T}, \text{instruction}^{\mathcal{A}^j}), \quad (1)$$

where E denotes the error analysis for each reasoning step, $P \in \{\text{Correct}, \text{Incorrect}\}$ indicates the overall reasoning status, and R represents the routing path in the template tree that guides template selection. Based on the routing path, we sample relevant critique templates $\mathcal{T}_s$ from the template tree $\mathcal{T}$ to guide the Critic agent in generating targeted and high-quality critiques for the identified errors. Notably, due to the self-evolving nature of our template tree, when the Judge identifies an error type not yet present in the tree, we randomly sample various error types from existing templates to guide the Critic in generating helpful critique.

Critic ($\mathcal{A}^c$). The Critic agent serves as a crucial component in our framework, responsible for generating detailed and constructive critiques for the identified errors. With the guidance of sampled critique templates $\mathcal{T}_s$, the Critic agent locates the first error step in the reasoning chain τ, analyzes error details, and provides specific suggestions for subsequent refinement. Formally, the Critic agent operates as:

$$\mathcal{C}, I = \pi(\mathbb{T}, q, \tau, \mathcal{T}_s, \text{instruction}^{\mathcal{A}^c}), \quad (2)$$

[1] We use Chain-of-Table (Wang et al., 2024) for initial chains, though our framework is applicable to other methods.

where $\mathcal{C}$ denotes the generated critique and I indicates the index of the first error step in τ. The effectiveness of the Critic agent directly impacts the Refiner's ability to correct reasoning errors, which motivates our design of the template tree to enhance critique quality.

Refiner ($\mathcal{A}^r$). The Refiner agent is tasked with correcting the reasoning chain based on the critique provided by the Critic. Given the critique $\mathcal{C}$, the table $\mathbb{T}$, question q, and the partial reasoning chain up to the first error step (i.e., $\tau_p = \{s_1, ..., s_I\}$), the Refiner first rectifies the identified error and then completes the remaining reasoning steps to generate a full refined chain. Formally, the Refiner agent operates as:

$$\tau' = \pi(\mathbb{T}, q, \tau_p, \mathcal{C}, \text{instruction}^{\mathcal{A}^r}), \tag{3}$$

where τ' represents the newly generated complete reasoning chain.

Curator ($\mathcal{A}^{cu}$). The Curator agent serves as an experience-driven learning component that distills valuable critique templates from current refinement processes. It is activated only after the complete refinement process concludes, specifically when the Judge agent verifies that the final reasoning chain is error-free (P = Correct), as shown in Figure 1. Through reviewing each refinement iteration and the existing template tree $\mathcal{T}$, the Curator autonomously distills meaningful critique templates from effective refinement experiences. These newly distilled templates are then incorporated into $\mathcal{T}$ to enhance future critique generation. Formally, the Curator operates as:

$$\mathcal{T}' = \pi(\mathcal{T}, H, \text{instruction}^{\mathcal{A}^{cu}}), \tag{4}$$

where H represents the complete refinement history, and $\mathcal{T}'$ denotes the updated template tree. The detailed update strategy will be delineated in subsequent sections.

3.3 Multi-turn Refinement

As discussed in the Introduction, the multi-turn refinement in Table-Critic is motivated by our observation that LLMs often excel at identifying and correcting the first error in reasoning chains, but may introduce new errors in subsequent steps. To address this challenge, we implement an iterative refinement process where multiple agents collaboratively monitor and improve the reasoning chain until reaching a satisfactory solution.

Specifically, given an initial reasoning chain τ, our framework operates through the following steps in each iteration: (1) The Judge agent first analyzes the entire reasoning chain to identify potential errors and determine their types. If no errors are detected (P = Correct), the process terminates. Otherwise, the Judge routes through the template tree to locate relevant critique templates. (2) With the guidance of sampled templates $\mathcal{T}_s$, the Critic agent generates detailed critiques $\mathcal{C}$ focusing on the first identified error at step I. This strategy ensures that each refinement iteration addresses errors sequentially, preventing the introduction of cascading errors. (3) The Refiner agent then generates a new reasoning chain τ' by incorporating the critique. Importantly, the Refiner only receives the partial chain τ_p up to the error step I, forcing it to reconstruct the remaining steps with the help of critique. This design prevents the Refiner from being biased by previous erroneous chain. (4) The above process continues iteratively until one of the following conditions is met: the Judge determines the current reasoning chain is correct (P = Correct) or the maximum number of iterations K is reached.

Through this multi-turn design, Table-Critic effectively manages the complexity of multi-step reasoning refinement while maintaining the quality of each correction step. The iterative nature of our approach, combined with specialized agent roles and strategic process control, enables robust and efficient reasoning improvement.

3.4 Self-evolving Template Tree

To address the challenge of identifying diverse and unpredictable error types in table reasoning, we introduce a self-evolving template tree that systematically accumulates and organizes critique knowledge. This dynamic structure enables our system to effectively handle both common and emerging error patterns through experience-driven learning.

Tree Structure. The template tree $\mathcal{T}$ represents a hierarchical structure that captures the relationships among different error types. As shown in Figure 1, each node in the tree represents a specific type of error, where: (1) Internal nodes represent broader error categories (e.g., Sub-table Error) that can be further subdivided into more specific error types. (2) Leaf nodes represent specific error types (e.g., Row Error, Column Error) and maintain a repository of critique templates associated with that particular error type.

Self-evolving Mechanism. The template tree evolves dynamically through the Curator agent, which manages two primary operations: adding templates to existing leaf nodes and expanding tree branches. As illustrated in Figure 1, the evolution process includes:

(1) *Template Enhancement*. When new effective critique patterns are identified, the Curator adds them to the corresponding leaf node's template repository. This operation enriches existing error type categories without changing the tree structure. For instance, when a new effective template for Row Error is discovered, it is directly added to the corresponding template repository.

(2) *Branch Expansion.* The Curator expands the tree structure in two ways when new error types are identified:

- *Vertical Expansion*: When a new error type is discovered that requires more fine-grained categorization, the Curator performs a vertical split. This operation transforms an existing leaf node into an internal node with two new child nodes. Specifically, the Curator first categorizes the existing templates in the leaf node with an appropriate name (e.g., Row Error), creating one new leaf node. Then, it creates another leaf node with a different name (e.g., Column Error) to accommodate the newly discovered error type and its corresponding templates. This process ensures that each leaf node maintains a cohesive collection of templates for a specific error type.

- *Horizontal Expansion*: When a completely new error type is identified that parallels existing categories, the Curator adds a new branch at the same level. This operation preserves the existing structure while accommodating new error types. As illustrated in the Figure 1 (bottom), the addition of the Final Query Error branch represents a horizontal expansion that complements the existing Sub-table Error category.

Through these evolution mechanisms, our template tree maintains a dynamic balance between preserving accumulated knowledge and incorporating new error patterns. The vertical expansion enables more precise error categorization, while horizontal expansion ensures comprehensive coverage of diverse error types. This adaptive structure allows the system to continuously improve its critique capabilities while maintaining organized and efficient template management. The detailed pipeline of our Table-Critic is presented in Appendix B.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our approach on two standard benchmarks: (1) WikiTableQuestions (WikiTQ) (Pasupat and Liang, 2015): A table reasoning benchmark with 4,344 test samples from 421 tables. (2) TabFact (Chen et al., 2020): A fact verification benchmark in table reasoning with 2,024 test samples from 298 tables.

Baselines. We conduct comprehensive experiments comparing Table-Critic against three categories of baselines: **(1) Standard Reasoning.** End-to-End QA directly generates answers using table and question as input. Few-Shot QA extends this by incorporating exemplar (Table, Question, Answer) triplets from the training set. **(2) Decomposition-Based Reasoning.** Binder (Cheng et al., 2022) decomposes questions into executable SQL/Python sub-programs. Dater (Ye et al., 2023) employs parsing-execution-filling strategy with sub-table decomposition. Chain-of-Table (Wang et al., 2024) generates intermediate tables through dynamic operations. **(3) Critic-Based Reasoning.** Critic-CoT (Zheng et al., 2024) implements self-reflection for error identification.

Implementation Details. To ensure comprehensive evaluation, we conduct experiments across three LLMs: Qwen2.5-72B-Instruct (Yang et al., 2024a), LLaMA3.3-70B-Instruct (Dubey et al., 2024), and GPT-4o-mini (Hurst et al., 2024). For all baseline methods, we follow their original settings to ensure optimal performance. For fair comparison, both Critic-CoT (Zheng et al., 2024) and our Table-Critic framework are implemented upon Chain-of-Table (Wang et al., 2024). For our Table-Critic, the template tree is initialized with only 2 templates that demonstrate basic critique patterns. From this minimal starting point, the tree evolves autonomously through our self-evolving mechanism, continuously learning and incorporating new critique patterns. For all experiments, we set the maximum refinement iterations K to 5 and use temperature 0.0 for greedy decoding. The detailed prompts and instructions for each agent in our framework are provided in Appendix E.

4.2 Main Results

We report the performance on different table reasoning benchmarks across different LLMs in Table 1. Our comprehensive evaluation reveals several key findings: **First,** Table-Critic consistently outper-

Method	Qwen2.5-72B		LLaMA3.3-70B		GPT-4o-mini		Average	
	WikiTQ	TabFact	WikiTQ	TabFact	WikiTQ	TabFact	WikiTQ	TabFact
End-to-End QA	56.6	85.1	51.1	81.0	52.6	73.5	53.4	79.9
Few-Shot QA	61.7	85.0	62.0	80.7	57.6	75.1	60.4	80.3
Binder (Cheng et al., 2022)	57.0	82.2	52.2	80.5	54.8	83.3	54.7	82.0
Dater (Ye et al., 2023)	63.8	90.0	59.5	87.6	65.8	83.6	63.0	87.1
Chain-of-Table (Wang et al., 2024)	68.3	89.7	62.1	89.9	67.5	88.9	66.0	89.5
Critic-CoT (Zheng et al., 2024)	69.0	89.8	66.8	88.0	66.3	86.9	67.4	88.2
Table-Critic (ours)	**77.2**	**92.6**	**70.1**	**91.5**	**73.9**	**91.1**	**73.7**	**91.7**
	↑8.2	↑2.6	↑3.3	↑1.6	↑6.4	↑2.2	↑6.3	↑2.2

Table 1: Table reasoning results on WikiTQ and TabFact with Qwen2.5-72B, LLaMA3.3-70B, and GPT-4o-mini. Bold denotes the best performance and underline denotes the second-best performance.

forms all baseline methods across both datasets and all three LLMs. On average, our method achieves 73.7% accuracy on WikiTQ and 91.7% on TabFact, representing significant improvements of 6.3% and 2.2% respectively over the strongest baselines. **Second,** the improvements are robust across different model architectures. With Qwen2.5-72B-Instruct, we achieve the highest absolute performance (77.2% on WikiTQ, 92.6% on TabFact), showing substantial gains of 8.2% and 2.6% respectively. Similar patterns are observed with LLaMA3.3-70B-Instruct and GPT-4o-mini, demonstrating the framework's generalizability across different foundation models. **Third,** the performance variations between WikiTQ and TabFact provide insights into our method's strengths. Table-Critic shows larger improvements on WikiTQ (average +6.3%) compared to TabFact (average +2.2%), indicating its particular effectiveness in handling complex, multi-step reasoning tasks. This aligns with our framework design, as WikiTQ's compositional questions benefit more from our multi-turn refinement and self-evolving template tree mechanism than TabFact's binary verification tasks. Nevertheless, the consistent improvements on TabFact demonstrate our method's capability even in simpler scenarios. **Finally,** comparing against different baseline categories reveals the advancement of our approach. While recent methods like Chain-of-Table (Wang et al., 2024) and Critic-CoT (Zheng et al., 2024) have made notable progress through decomposition and criticism mechanisms, Table-Critic achieves substantially larger improvements over these strong baselines. This suggests that our multi-agent framework, combining multi-turn refinement with self-evolving template tree, provides a more effective solution for complex table reasoning tasks.

4.3 Analysis of Critic Effectiveness

As shown in Table 2, we conduct a detailed analysis of different critic mechanisms by comparing Table-Critic with Chain-of-Table (Wang et al., 2024) and Critic-CoT (Zheng et al., 2024). Our analysis focuses on four key metrics: **(1) Overall Accuracy (Acc)**: The percentage of correctly solved questions; **(2) Error Correction Rate ($\Delta^{i\to c}$):** The percentage of questions incorrectly solved by Chain-of-Table but corrected by different Critic methods; **(3) Solution Degradation Rate ($\Delta^{c\to i}$):** The percentage of questions correctly solved by Chain-of-Table but degraded by different Critic methods; **(4) Net Performance Gain (Δ):** The overall improvement relative to Chain-of-Table, calculated as $\Delta = \Delta^{i\to c} + \Delta^{c\to i}$.

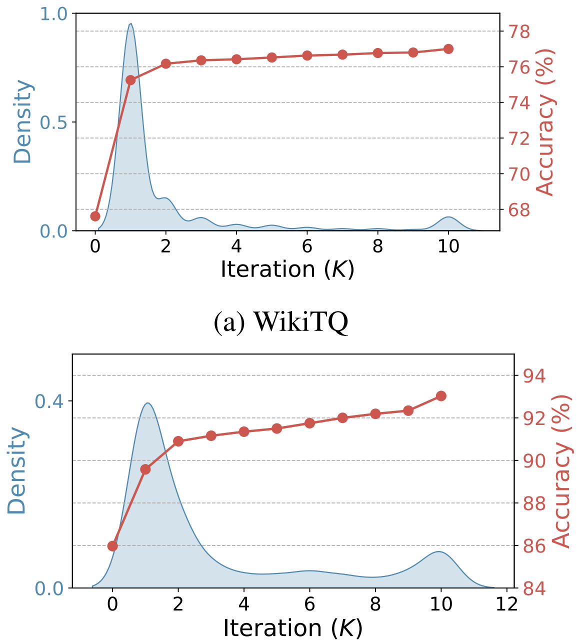
Error Correction vs. Solution Degradation. Table-Critic demonstrates superior error correction capabilities while minimizing solution degradation. On WikiTQ, it successfully corrects 9.6% of Chain-of-Table's errors while only degrading 0.7% of correct solutions, resulting in a substantial net performance gain (+8.9%). In contrast, Critic-CoT shows a less effective pattern, with a 5.6% correction rate offset by a high degradation rate (-4.9%), yielding only a marginal improvement (+0.7%).

Task-Specific Performance. The effectiveness of critique mechanisms varies across different tasks. On WikiTQ, which involves complex multi-step reasoning, Table-Critic achieves a higher error correction rate (+9.6% vs +5.6%) and maintains a observably lower degradation (-0.7% vs -4.9%) compared to Critic-CoT. For TabFact's simpler verification tasks, while the improvements are more modest, Table-Critic still maintains better stability with lower degradation rates (-0.5% vs -2.8%).

Critic Stability. A key advantage of Table-Critic is its stability in maintaining correct solutions. The

Dataset	Chain-of-Table	Critic-CoT				Table-Critic			
	Acc	Acc	$\Delta^{i\to c}$	$\Delta^{c\to i}$	Δ	Acc	$\Delta^{i\to c}$	$\Delta^{c\to i}$	Δ
WikiTQ	68.3	69.0	+5.6	-4.9	+0.7	77.2	+9.6	-0.7	+8.9
TabFact	89.7	89.8	+2.9	-2.8	+0.1	92.6	+3.4	-0.5	+2.9

Table 2: Critic performance comparison of different critic methods. $\Delta^{i\to c}$, $\Delta^{c\to i}$, and Δ measure the error correction rate, solution degradation rate, and net performance gain relative to Chain-of-Table, respectively.



(a) WikiTQ

(b) TabFact

Figure 2: Analysis of Model Convergence and Iteration Requirements on WikiTQ and TabFact Datasets.

consistently low degradation rates (-0.7% for WikiTQ and -0.5% for TabFact) suggest that our self-evolving template tree effectively preserves valid reasoning patterns while identifying and correcting errors. This contrasts with Critic-CoT's higher degradation rates (-4.9% for WikiTQ and -2.8% for TabFact), indicating potential instability in its critique process.

4.4 Analysis of Multi-Turn Mechanism

To understand the effectiveness of our multi-turn refinement mechanism, we analyze how model performance evolves with the number of iterations K and the distribution of required iteration counts (set maximal $K = 10$), as shown in Figure 2.

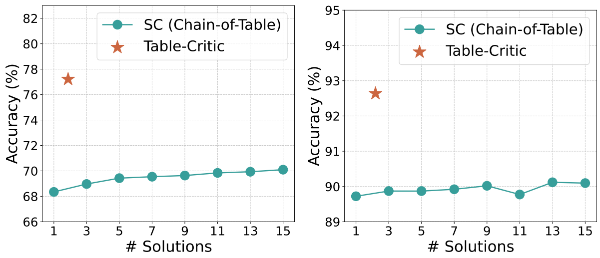
Performance Evolution. On both datasets, we observe a consistent pattern of rapid initial improvement followed by gradual convergence. For WikiTQ, the accuracy increases sharply from 67.6% to 76.5% within the first three iterations and stabilizes around 77% after six iterations. Similarly, on TabFact, the performance improves significantly in early iterations and plateaus at approximately 92% after five iterations. This pattern suggests that our multi-turn mechanism effectively refines solutions through iterative improvements.

Iteration Distribution. The density plots reveal interesting insights about the complexity of different tasks. On WikiTQ, we observe a broader distribution with multiple peaks, indicating that questions require varying numbers of iterations for resolution. The main peak occurs at 1-2 iterations, with smaller peaks extending up to 10 iterations, reflecting the diverse complexity of multi-step reasoning questions. TabFact also shows a concentrated distribution with two distinct peaks: a primary peak at 1-2 iterations and a secondary peak around 10 iterations. This bimodal pattern suggests that TabFact tend to fall into two categories: (1) straightforward cases that can be verified quickly within 1-2 iterations, and (2) complex cases that require extensive refinement to reach a conclusive verification. This distribution aligns with the inherent nature of fact verification tasks, where statements are either relatively simple to verify or require careful step-by-step examination.

Convergence and Stability Analysis. The results suggest that while our method allows for up to 10 iterations, most improvements are achieved within the first 5 iterations. This efficient convergence, combined with our early termination mechanism, helps maintain computational efficiency while ensuring thorough reasoning. Notably, as evidenced in Table 2, Table-Critic maintains stable performance across iterations without the degradation typically seen in iterative approaches, demonstrating the effectiveness of our Critic agent and self-evolving template tree mechanism.

4.5 Analysis of Computational Cost

To ensure a fair comparison with Chain-of-Table (Wang et al., 2024) in terms of computational cost, we conduct an analysis of the cost-effectiveness trade-off, as shown in Figure 3. Since



(a) WikiTQ (b) TabFact

Figure 3: Computational cost and Effectiveness Comparison between SC (Self-Consistency Based on Chain-of-Table) and our Table-Critic.

Table-Critic builds upon Chain-of-Table by incorporating additional critique mechanisms, we align the computational costs by allowing Chain-of-Table to generate multiple solutions (majority voting) through Self-consistency (Wang et al., 2023) (with temperature 0.8) and compare the performance under equivalent or even superior computational budgets.

Efficiency Comparison. Our method requires approximately 1.8-2.2× computational cost compared to the basic Chain-of-Table. The complete derivation process of computational cost is provided in Appendix C. However, as illustrated in Figure 3, Table-Critic achieves substantially higher accuracy (77.2% on WikiTQ and 92.6% on TabFact) compared to Chain-of-Table's performance even with 15 solution attempts. Notably, Chain-of-Table shows only marginal improvements as the number of solutions increases, reaching 70.0% on WikiTQ and 90.1% on TabFact with 15 solutions.

Cost-Effectiveness Analysis. The results demonstrate that simply increasing the number of solution attempts in Chain-of-Table fails to achieve comparable performance to Table-Critic, despite consuming similar or even greater computational resources. This suggests that our multi-agent refinement mechanism provides a more effective approach to improving reasoning accuracy than traditional majority voting strategies. The superior performance of Table-Critic justifies its additional computational overhead by offering substantially better reasoning capabilities.

4.6 Analysis of Self-evolving Template Tree

To investigate the effectiveness of our self-evolving mechanism, we conduct an ablation study comparing Table-Critic with and without the dynamic template evolution capability, as shown in Table 3. In the static setting (w/o Self-evolving), the template tree remains fixed with its initial two templates, while our full Table-Critic allows the Curator agent to dynamically maintain and evolve the template tree throughout the reasoning process.

Method	**WikiTQ**	**Tabfact**
Table-Critic	77.2	92.6
w/o Self-evolving	76.1	90.8
	$\downarrow$1.1	$\downarrow$1.8

Table 3: The impact of self-evolving mechanism on our template tree.

Performance Impact. The results demonstrate the clear benefits of the self-evolving mechanism. Without template evolution, performance drops by 1.1% on WikiTQ (from 77.2% to 76.1%) and 1.8% on TabFact (from 92.6% to 90.8%). The more substantial performance gap on TabFact suggests that template evolution is particularly beneficial for fact verification tasks, where diverse verification patterns may be needed.

Mechanism Analysis. These results highlight the importance of dynamic adaptation in our framework. The self-evolving mechanism allows the template tree to expand beyond its initial state, accommodating diverse reasoning patterns encountered during the critique process. This flexibility enables more effective error detection and correction compared to a static template approach. The performance gains validate our design choice of incorporating dynamic template evolution, showing that the ability to adapt and expand the template structure is crucial for robust table reasoning. For reference, we provide visualizations of both the initial template tree and its evolved state in Appendix D, illustrating how the tree structure adapts to accommodate different reasoning patterns.

5 Conclusion

In this paper, we propose Table-Critic, a novel multi-agent framework that enhances table reasoning through collaborative criticism and refinement. Our approach introduces four specialized agents working in concert with a self-evolving template tree, effectively addressing the challenges of error identification and correction in complex table reasoning tasks. Extensive experiments demonstrate that our method significantly outperforms existing approaches, achieving substantial improvements across different datasets while maintaining robust performance stability.

Limitations

Our Table-Critic framework has demonstrated strong performance in enhancing table reasoning through multi-agent collaboration and systematic refinement. While our current implementation focuses primarily on textual table reasoning, the proposed multi-agent critique framework is inherently flexible and can potentially be extended to various other scenarios. For instance, the framework could be adapted to handle multimodal reasoning tasks where tables are combined with images, graphs, or other visual elements. We believe the core principles of our approach—collaborative criticism, iterative refinement, and self-evolving template tree—could contribute to broader applications in complex reasoning tasks beyond the current textual domain.

Ethics Statement

Our research focuses on improving table reasoning capabilities through multi-agent collaboration, without involving any sensitive personal data or potentially harmful applications. The datasets used in our experiments are publicly available and widely used in the research community. Our framework is designed to enhance the reliability and transparency of AI systems in table understanding tasks, potentially benefiting various real-world applications such as data analysis and information retrieval. We acknowledge that while our method improves reasoning accuracy, it should be used as a complementary tool rather than a complete replacement for human judgment in critical decision-making processes.

References

Anthropic. 2024. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. Anthropic Blog.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Alphamath almost zero: Process supervision without process. In Advances in Neural Information Processing Systems, volume 37, pages 27689–27724. Curran Associates, Inc.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024b. Step-level value preference optimization for mathematical reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7889–7903, Miami, Florida, USA. Association for Computational Linguistics.

Guoxin Chen, Zhong Zhang, Xin Cong, Fangda Guo, Yesai Wu, Yankai Lin, Wenzheng Feng, and Yasheng Wang. 2025. Learning evolving tools for large language models. In The Thirteenth International Conference on Learning Representations.

Si-An Chen, Lesly Miculicich, Julian Eisenschlos, Zifeng Wang, Zilong Wang, Yanfei Chen, Yasuhisa Fujii, Hsuan-Tien Lin, Chen-Yu Lee, and Tomas Pfister. 2024c. Tablerag: Million-token table understanding with language models. In Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. arXiv preprint arXiv:2210.02875.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. CoRR, abs/2407.21783.

Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: table-operations aware fact verification via sentence-table cloze pre-training. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 4971–4983. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024, pages 8048–8057. ijcai.org.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. Gpt-4o system card. CoRR, abs/2410.21276.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. arXiv preprint arXiv:2107.07653.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, 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-Cristian 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, and et al. 2024. Gemma: Open models based on gemini research and technology. CoRR, abs/2403.08295.

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

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. arXiv preprint arXiv:2407.11511.

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. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. Proceedings of the 12th International Conference on Learning Representations,ICLR 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2024b. Confidence v.s. critique: A decomposition of self-correction capability for llms. CoRR, abs/2412.19513.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 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, pages 174–184.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. arXiv preprint arXiv:2005.08314.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. ACM Computing Surveys, 56(12):1–39.

Bin Zhang, Hangyu Mao, Lijuan Li, Zhiwei Xu, Dapeng Li, Rui Zhao, and Guoliang Fan. 2024. Sequential asynchronous action coordination in multi-agent systems: A stackelberg decision transformer approach. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 59559–59575. PMLR.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. 2025. A survey of table reasoning with large language models. Frontiers of Computer Science, 19(9):199348.

Yilun Zhao, Lyuhao Chen, Arman Cohan, and Chen Zhao. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 12824–12840. Association for Computational Linguistics.

Xin Zheng, Jie Lou, Boxi Cao, Xueru Wen, Yuqiu Ji, Hongyu Lin, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. 2024. Critic-cot: Boosting the reasoning abilities of large language model via chain-of-thoughts critic. CoRR, abs/2408.16326.

A Additional Related works

Multi-agent Systems. Multi-agent systems have recently demonstrated promising potential in complex reasoning tasks by enabling collaborative problem-solving through specialized agents (Guo et al., 2024; Zhang et al., 2024). These systems typically leverage the complementary strengths of different agents to achieve more robust and effective solutions than single-agent approaches. While existing work has explored multi-agent frameworks in various domains, their application to table reasoning tasks remains largely unexplored. To our knowledge, our Table-Critic presents the first attempt to introduce a multi-agent framework for table reasoning, where specialized agents collaborate to identify, critique, and refine reasoning steps, offering a novel perspective on addressing the challenges in complex table reasoning tasks.

B More Implementation Details

In this section, we provide a comprehensive implementation details of our proposed method. For additional insights and more intricate details, we refer the reader to our **supplementary materials**.

B.1 Overall Pipeline of Table-Critic

Table-Critic employs an iterative process to critique and refine the reasoning chain and predicted answer for table reasoning tasks. As described in Algorithm 1, the process begins with an input table $\mathbb{T}$, a question q, an initial reasoning chain τ, and a template tree $\mathcal{T}$. The `Judge` agent is first invoked to evaluate the correctness of the reasoning chain (Line 2). This evaluation yields the reasoning status P, an error analysis E, and a routing path R in the template tree.

When the reasoning chain is deemed incorrect (P = Incorrect), Table-Critic proceeds by sampling relevant critique templates $\mathcal{T}_s$ from the template tree using the routing path R (Line 4). These templates are then used by the `Critic` agent to generate a detailed critique $\mathcal{C}$ and identify the index of the first error step I in the reasoning chain (Line 5). To address the identified errors, the `Refiner` agent retains the reasoning steps up to step I and refines the chain starting from step I, guided by the critique $\mathcal{C}$ (Line 6). The refined reasoning chain τ' is subsequently re-evaluated by the `Judge` agent to determine if it is now correct (Line 7).

This refinement loop continues iteratively until the reasoning chain is verified as correct ($P =$

Correct). Once a correct reasoning chain is established, the `Curator` agent updates the template tree $\mathcal{T}$ by incorporating new critique templates distilled from the refinement history. This update enhances the template tree's ability to support future refinement processes (Line 9).

The final output of Table-Critic is the refined reasoning chain τ', which represents the accurate and improved solution to the table reasoning task. By systematically identifying and addressing errors in a collaborative multi-step process, Table-Critic ensures robust and precise refinement of reasoning chains and answers.

B.2 LLM Servers

Our approach implements agent behaviors through in-context learning, requiring no extensive training procedures. We deploy multiple LLM servers, including Qwen2.5-72B-Instruct and LLaMA3.3-70B-Instruct through the `SGLang` inference engine, and GPT-4o-mini via its provided API service. While the choice between fine-tuning and in-context learning remains an open question, it is not the primary focus of our work. Following prior studies (Wang et al., 2024; Zheng et al., 2024), we adopt in-context learning as our implementation strategy for its simplicity and effectiveness.

C Detailed Computational Cost Analysis

This appendix evaluates the computational cost of Table-Critic relative to the baseline Chain-of-Table method. The computational cost is analyzed for two datasets, WikiTQ and TabFact, based on the number of input and output tokens required. All token counts are expressed in millions (M), and the cost ratio reflects the relative cost of Table-Critic compared to Chain-of-Table.

C.1 Computational Cost Definition

The computational cost of a prompt-based method is defined as follows:

$$N_{\text{total}} = N_{\text{in}} \cdot \left(\frac{P_{\text{in}}}{P_{\text{in}} + P_{\text{out}}} \right) + N_{\text{out}} \cdot \left(\frac{P_{\text{out}}}{P_{\text{in}} + P_{\text{out}}} \right), \tag{5}$$

where N_{in} and N_{out} represent the number of input and output tokens, and P_{in} and P_{out} denote the costs per token for input and output, respectively. Based on the pricing model of Qwen2.5-72B-Instruct, $P_{\text{in}} = 0.004$ CNY per thousand tokens and $P_{\text{out}} = 0.012$ CNY per thousand tokens.

Using the above values, the normalized cost weights are:

$$\text{Input Weight} = \frac{P_{\text{in}}}{P_{\text{in}} + P_{\text{out}}} = 0.25,$$

$$\text{Output Weight} = \frac{P_{\text{out}}}{P_{\text{in}} + P_{\text{out}}} = 0.75.$$

Substituting these weights, the formula simplifies to:

$$N_{\text{total}} = 0.25 \cdot N_{\text{in}} + 0.75 \cdot N_{\text{out}}. \tag{6}$$

C.2 Dataset-Specific Computational Cost Analysis

The computational cost of Table-Critic is compared against Chain-of-Table for the WikiTQ and TabFact datasets. Detailed token counts and cost ratios are shown in Table 4.

On the WikiTQ dataset, Chain-of-Table incurs a total computational cost of 19.6M, with 73.5M input tokens and 1.6M output tokens. In contrast, Table-Critic requires 135.5M input tokens and 3.8M output tokens, resulting in a total cost of 36.7M. This corresponds to a cost ratio of $1.87\times$, indicating that Table-Critic is approximately 1.87 times more computationally expensive than Chain-of-Table on this dataset.

On the TabFact dataset, Chain-of-Table incurs a total computational cost of 7.8M, with 29.3M input tokens and 0.6M output tokens. Table-Critic, on the other hand, requires 62.1M input tokens and 20.4M output tokens, resulting in a total cost of 17.1M. This corresponds to a cost ratio of $2.19\times$, indicating that Table-Critic is approximately 2.19 times more computationally expensive than Chain-of-Table.

D Self-evolving Template Tree

Figure 4 illustrates the Self-evolving process of the Template Tree. In the initial stage (Figure 4a), the tree contains only two broad categories of errors: Sub-table Error and Final Query Error, each representing a high-level abstraction of error types. Through the self-evolving mechanism, the tree dynamically expands and refines its structure to accommodate more fine-grained error types, as shown in the evolved tree (Figure 4b).

It is important to note that the Evolved Tree is considerably larger in practice, containing a more extensive hierarchy of error types. However, for clarity, only a subset of the evolved structure is displayed here.

Algorithm 1 The overall pipeline of Table-Critic

Input: Table $\mathbb{T}$, question q, initial reasoning chain τ, the template tree $\mathcal{T}$.

Output: The refined reasoning chain τ'.

$H \leftarrow \emptyset$ ▷ Initialize refinement history.

$P, E, R \leftarrow$ `Judge`$(\mathbb{T}, q, \tau, \mathcal{T})$

while $P =$ Incorrect **do**

  $\mathcal{T}_s \leftarrow$ Sample Templates using R in the $\mathcal{T}$

  $\mathcal{C}, I \leftarrow$ `Critic`$(\mathbb{T}, q, \tau, \mathcal{T}_s)$ ▷ Generating critique $\mathcal{C}$ and identify the index of first error step I.

  $\tau_p \leftarrow \tau[: I]$ ▷ Retain the partial reasoning steps up to step I

  $\tau' \leftarrow$ `Refiner`$(\mathbb{T}, q, \tau_p, \mathcal{C})$ ▷ Refine the reasoning chain.

  $H \leftarrow H \cup \{\mathbb{T}, q, \tau, \tau', \mathcal{C}\}$ ▷ Update history.

  $P, E, R \leftarrow$ `Judge`$(\mathbb{T}, q, \tau', \mathcal{T})$ ▷ Re-evaluates the updated reasoning chain

end while

$\mathcal{T} \leftarrow$ `Curator`$(\mathcal{T}, H)$ ▷ Update the template tree $\mathcal{T}$ to facilitate future refinement.

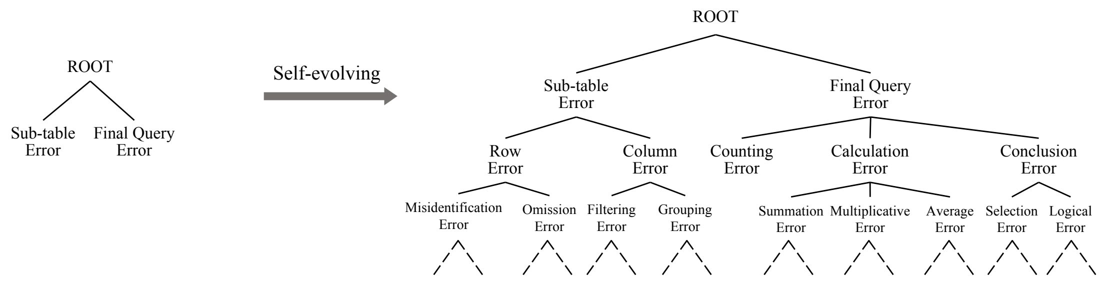
return Refined reasoning chain τ' and new template tree $\mathcal{T}$.

Dataset	Chain-of-Table			Table-Critic			Cost Ratio
	Input (M)	Output (M)	Total (M)	Input (M)	Output (M)	Total (M)	(TC/CoT)
WikiTQ	73.5	1.6	19.6	135.5	3.8	36.7	1.87×
TabFact	29.3	0.6	7.8	62.1	20.4	17.1	2.19×

Table 4: Computational Cost Comparison Between Chain-of-Table and Table-Critic (Token Counts in Millions)

E Prompts and Case Study

This appendix provides comprehensive instructions and illustrative examples for three intelligent agents: the Judge Agent, the Critic Agent, and the Refiner Agent. These agents are designed to collaboratively evaluate and refine reasoning processes applied to table-based questions. Figures 5 and 6 offer detailed guidance for the Judge Agent, including step-by-step procedures to assess the validity of reasoning steps, pinpoint errors, and categorize conclusions (e.g., correct, incorrect with identified error route, or random error). Figures 7 and 8 explain how the Critic Agent systematically evaluates each reasoning step, highlights the first incorrect step, and provides constructive critiques. Additionally, Figure 9 introduces the Refiner Agent, demonstrating how critiques are utilized to refine reasoning steps, ensuring accurate and complete solutions.



(a) Initial Tree

(b) Evolved Tree

Figure 4: An example of self-evolving mechanism in our Template Tree.

You are an intelligent judge tasked with evaluating the correctness of a given Prediction Answer. If the Prediction Answer is incorrect, identify which step within the reasoning process is incorrect and subsequently locate the corresponding error type within the error tree:
1. Original Table: The raw table data.
2. Question: The question pertaining to the table data.
3. Reasoning Steps: A step-by-step process of sub-table transformations and extractions based on the following functions.
  - f_add_column(): Adds a new column to the table.
  - f_select_row(): Selects specific rows based on the question.
  - f_select_column(): Removes irrelevant columns from the table.
  - f_group_column(): Groups rows based on the values in a specific column.
  - f_sort_column(): Sorts rows based on the values in a specified column.
4. Prediction Answer: The answer derived from the final sub-table.

Instruction:
1. **Explanation:** Conduct an explanation of why the Prediction Answer is correct or incorrect. If it is incorrect, then conduct an analysis of each reasoning step's validity.
2. **Conclusion:**
  - If the Prediction Answer is correct, conclude with 'Conclusion: [Correct]'.
  - If the Prediction Answer is incorrect, conclude with either 'Conclusion: [Incorrect] (ERROR ROUTE)' or 'Conclusion: [Incorrect] (random)'.
    - Use '(ERROR ROUTE)' to indicate the specific path in the error tree that represents the error.
  - If no such route can be identified, use '(random)' instead.

Figure 5: Instructions for the Judge Agent. These instructions outline the procedure for verifying the correctness of a predicted answer and identifying errors within the reasoning process.

```
Original Table:
/*
col : res. | record | opponent | method | event | date | round | time | location | notes
row 1 : win | 12-3 | mike hayes | ko (punch) | ksw 25: khalidov vs. sakurai | december 7, 2013
| 1 | 1:12 | wrocław, poland
row 2 : win | 11-3 | nick moghadden | tko (punches) | bellator 99 | september 13, 2013 | 1 |
3:22 | temecula, california, united states | bellator debut
row 3 : loss | 10-3 | guto inocente | decision (unanimous) | strikeforce: barnett vs. cormier |
may 19, 2012 | 3 | 5:00 | san jose, california, united states | light heavyweight debut
row 4 : win | 10-2 | brett albee | tko (strikes) | strikeforce: diaz vs. daley | april 9, 2011
| 1 | 1:46 | san diego, california, united states
row 5 : loss | 9-2 | lavar johnson | ko (punches) | strikeforce challengers: bowling vs. voelker
| october 22, 2010 | 1 | 2:17 | fresno, california, united states
row 6 : win | 9-1 | eddie sapp | submission (rear-naked choke) | native fighting championship 6
| august 14, 2010 | 1 | 2:01 | campo, california, united states
row 7 : loss | 8-1 | cody goodale | decision (unanimous) | gladiator challenge: maximum force |
april 25, 2010 | 3 | 5:00 | san jacinto, california, united states
*/
Question:
how many loses did virgil zwicker have?
Reasoning Steps:
Step 1: Select relevant rows.
The question aims to find out the number of losses Virgil Zwicker had based on the provided fight
records. We need to count the rows where the "res." column shows "loss".
So we use f_select_row(row 3, row 5, row 7).
Step 2: Filter out useless columns.
similar words of the question link to columns:
loses -> record
column value of the question link to columns:
10-3 -> record
semantic sentence of the question link to columns:
None
So we use f_select_column(record).
Step 3: After using f_select_row(row 3, row 5, row 7) and f_select_column(record), we obtain the
sub table:
/*
col : record
row 1 : 10-3
row 2 : 9-2
row 3 : 8-1
*/
The provided table shows records of wins and losses but does not specify who the records belong
to. However, if we assume that the table represents Virgil Zwicker's records, then we can deduce
the number of losses from each record. The losses in the records are 3, 2, and 1 respectively.
If we consider each row as a separate season or period, then the total number of losses would be
the sum of these losses: 3 + 2 + 1 = 6.
Prediction Answer:
6

Explanation:
Step 1 correctly identifies the relevant rows where the "res." column indicates a loss. The rows
selected are indeed the ones where the outcome is a loss. Step 1 is correct.
Step 2 incorrectly filters out the columns. The question asks for the number of losses, but
the filtering step selects only the 'record' column, which combines wins and losses in a single
string (e.g., "10-3"). This does not directly provide the number of losses. Instead, the 'res.'
column should be used to count the losses directly. Step 2 is incorrect.

Conclusion: [Incorrect] (sub-table error -> column error -> <END>)
```

Figure 6: Example of Judge Agent's Analysis and Error Detection. This example illustrates how the Judge Agent evaluates reasoning steps, identifies errors, and determines the correctness of a predicted answer.

You are an intelligent critic tasked with determining which step of the table reasoning is incorrect based on the following information:
1. Original Table: The raw table data.
2. Question: The question pertaining to the table data.
3. Reasoning Steps: A step-by-step process of sub-table transformations and extractions based on the following functions.
  - f_add_column(): Adds a new column to the table.
  - f_select_row(): Selects specific rows based on the question.
  - f_select_column(): Removes irrelevant columns from the table.
  - f_group_column(): Groups rows based on the values in a specific column.
  - f_sort_column(): Sorts rows based on the values in a specified column.
4. Prediction Answer: Final derived answer following the reasoning chain.

Instruction:
1. **Step-wise Analysis**: Conduct an evaluation of each reasoning step's validity. The step that is unnecessary but does not affect the answer is considered correct.
2. **Analysis Categories**:
  - For correct steps: Provide validation reasoning and mark as Step <NUM> is correct.
  - For incorrect steps: Detail the logical flaws and mark as Step <NUM> is incorrect.
  - You should stop at the first incorrect step.
3. **Conclude this critique**: Summarize this critique with an explicit conclusion.
4. **Conclusion Categories**:
  - Conclude with 'Conclusion: [Incorrect] Step <NUM>'.

Figure 7: Instructions for the Critic Agent. These instructions guide stepwise evaluation of reasoning steps.

```
Original Table:
/*
col : tie | home team | score | away team | attendance
row 1 : 1 | aylesbury united | 2-2 | windsor & eton | 847
row 2 : 2 | burscough | 5-0 | wakefield & emley | 437
row 3 : 3 | dover athletic | 0-3 | forest green roves | 932
row 4 : 4 | farnborough town | 2-0 | halifax town | 863
row 5 : 5 | gloucester city | 1-1 | southport | 1,237
row 6 : 6 | havant & waterlooville | 3-0 | hayes | 456
row 7 : 7 | margate | 0-2 | tamworth | 971
row 8 : 8 | yeovil town | 2-1 | northwich victoria | 4,469
*/
Question:
how many games attendance was under 1000?
Reasoning Steps:
Step 1: Select relevant rows.
The question aims to find out how many games had an attendance of less than 1000 people. To answer this, we need to count the number of rows where the attendance column shows a number less than 1000. So we use f_select_row(row 1, row 2, row 3, row 4, row 6, row 7).
Step 2: Filter out useless columns.
similar words of the question link to columns :
under 1000 -> attendance
column value of the question link to columns :
1000 -> attendance
semantic sentence of the question link to columns :
None
So we use f_select_column(attendance).
Step 3: After using f_select_row(row 1, row 2, row 3, row 4, row 6, row 7) and f_select_column(attendance), we obtain the sub table:
/*
col : attendance
row 1 : 847
row 2 : 437
row 3 : 932
row 4 : 863
row 5 : 456
row 6 : 971
*/
The games with attendance under 1000 were rows 1, 2, 4, and 5. Therefore, there were 4 games with attendance under 1000.
Prediction Answer:
4
```

Critique:
Step 1 correctly selects the rows that are relevant for further analysis regarding the attendance being under 1000. It includes all the rows where the attendance values might potentially be less than 1000, which are row 1, row 2, row 3, row 4, row 6, and row 7. This step makes the proper initial identification based on the condition stated in the question. Step 1 is correct.
Step 2 filters out the columns effectively by choosing the "attendance" column. Since the question is focused on the attendance values to determine which are under 1000, this step narrows down the data to just what's needed for the subsequent steps. Step 2 is correct.
Step 3 makes an error in counting the number of games with attendance under 1000. Looking at the sub-table from Step 3, all the rows shown (row 1 with 847, row 2 with 437, row 3 with 932, row 4 with 863, row 5 with 456, row 6 with 971) have attendance values that are less than 1000. Step 3 is incorrect.

Conclusion: [Incorrect] Step 3

Figure 8: Example of Critic Agent's Critique. This example demonstrates how to evaluate reasoning steps and identify errors in the reasoning chain.

```
Now, we have produced part of the Function Chain, but gained a critique.
Function Chain: f_select_row(row 1)
After step 1 (f_select_row(row 1)), we obtain the sub-table:
/*
  col : date introduced | class 1 (e.g. motorbike) | class 2 (e.g. car) | class 3 (e.g. car with
trailer) | class 4 (e.g. van) | class 5 (e.g. hgv)
  row 1 : 23 july 2004 | £1.00 | £2.00 | £5.00 | £5.00 | £6.00
*/
Question: on what date did the toll for class 1 first go above 2.00?
Critique:
Step 1 is incorrect. The selected row (row 2) has the toll for class 1 set at £1.00, which is not
above £2.00. The first row where the toll for class 1 exceeds £2.00 is row 3, dated 16 august
2004. Therefore, the selection of row 2 is incorrect.

Based on the critique, please continue to produce a complete and correct Function Chain.
/*
  col : date introduced | class 1 (e.g. motorbike) | class 2 (e.g. car) | class 3 (e.g. car with
trailer) | class 4 (e.g. van) | class 5 (e.g. hgv)
  row 1 : 9 december 2003 | £1.00 | £2.00 | £5.00 | £5.00 | £10.00
  row 2 : 23 july 2004 | £1.00 | £2.00 | £5.00 | £5.00 | £6.00
  row 3 : 16 august 2004 | £2.00 | £3.00 | £6.00 | £6.00 | £6.00
  row 4 : 14 june 2005 | £2.50 | £3.50 | £7.00 | £7.00 | £7.00
  row 5 : 1 january 2008 | £2.50 | £4.50 | £8.00 | £9.00 | £9.00
  row 6 : 1 january 2009 | £2.70 | £4.70 | £8.40 | £9.40 | £9.40
  row 7 : 1 march 2010 | £2.70 | £5.00 | £9.00 | £10.00 | £10.00
  row 8 : 1 march 2011 | £3.00 | £5.30 | £9.60 | £10.60 | £10.60
  row 9 : 1 march 2012 | £3.00 | £5.50 | £10.00 | £11.00 | £11.00
*/
Question: on what date did the toll for class 1 first go above 2.00?
The next operation must be one of f_add_column(), f_select_row(), f_select_column(),
f_group_column(), or f_sort_column().

Function Chain:
f_select_row(row 3)
```

Figure 9: Example of Refiner Agent's refinement. This example demonstrates how the critique is used to refine the Function Chain to accurately answer the question.