WHEN REASONING CONTRADICTS THE TABLE: DIAGNOSING ERROR RECOVERY IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models often solve TableQA by generating reasoning traces that cite intermediate evidence. However, an incorrect table reference can make an otherwise coherent reasoning trace inconsistent with the evidence in the table. When such a conflict arises, can models return to the evidence and recover on their own? We investigate this with the *Evidence–Trace Conflict* (ETC) protocol: we hold the table, question, and continuation point fixed while varying whether the supplied reasoning prefix is correct or contains data referencing errors (DREs). Across model families and three TableQA benchmarks, conflicting prefixes cause accuracy drops of 49.4 to 78.7 percentage points. On WikiTableQuestions (WTQ), an explicit reconsideration cue placed after the conflict improves accuracy by 67.3 percentage points, from 18.0% to 85.3%, whereas a reminder in the original user prompt, before the conflicting prefix, provides little help. This contrast reveals a *self-initiation gap* under evidence–trace conflict: models often recover when reconsideration is triggered after the conflict, but rarely initiate it on their own. Guided by this diagnosis, we fine-tune models on self-generated ordinary and recovery continuations to internalize the externally elicited behavior. This training raises single-pass recovery from 18.0% to 88.0% on WTQ and achieves over 85% recovery on two unseen TableQA benchmarks. Together, these results show that when reasoning contradicts the table, models often follow erroneous traces despite having access to the correct evidence, but can learn to recover without additional inference-time intervention.

1 INTRODUCTION

Large language models increasingly solve TableQA tasks through natural-language reasoning traces, locating, comparing, and combining values from structured evidence (Jin et al., 2022; Zhou et al., 2026; Zhang et al., 2024; Wang et al., 2024). These traces expose how evidence is used, but they can also propagate data referencing errors (DREs): a step cites a value that is absent from the table or taken from the wrong cell, and later steps inherit that reference as if it were correct. A single DRE can therefore determine the final answer even while the correct evidence remains visible in the context.

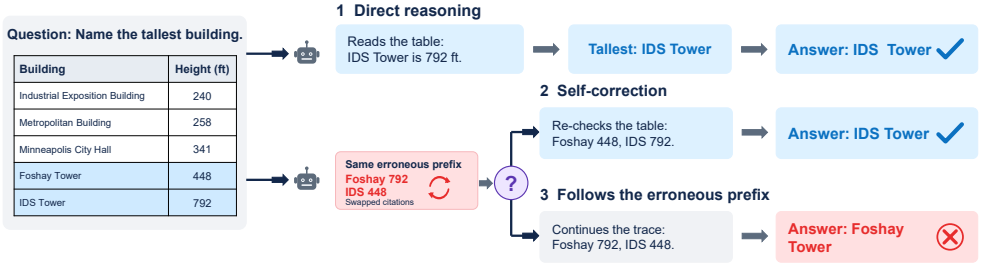


Figure 1: Three reasoning trajectories under the same table–question setting. Direct reasoning reads the table and answers correctly. Given the same conflicting prefix, the model either re-checks the table and self-corrects or follows the erroneous citations to an incorrect answer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Yang et al. (2026) show that DREs arise in model-generated table reasoning across models and tasks. Once a DRE enters an otherwise coherent trace, the model may re-read the cited cells and correct its reasoning or continue along the erroneous citation. Figure 1 illustrates these alternatives: direct reasoning yields the table-supported IDS Tower, while a prefix that swaps two heights can lead the model either to self-correct or to answer Foshay Tower by following the erroneous trace. Existing approaches use resampling, critics, verification chains, or DRE-triggered restart to improve table reasoning (Wang et al., 2023; Shaikh, 2026; Yu et al., 2025). These external interventions leave a narrower question open: **when reasoning contradicts the table, will the model reject that trace and return to the visible evidence on its own?**

To answer this question, we introduce the *Evidence–Trace Conflict* (ETC) protocol and construct matched evaluation sets (Section 2). Starting from a correct chain of thought, we create a conflicting version containing one or more incorrect evidence references while keeping the surrounding reasoning coherent and directed toward an incorrect answer. Holding the table, question, and continuation point fixed lets the matched pair separate ordinary continuation ability from recovery after DREs have entered the trace. This construction targets a failure pattern that can arise in natural reasoning: an evidence-reference mistake becomes embedded in an otherwise fluent trajectory, leaving the model to choose between the accumulated trace and the original evidence. The ETC protocol varies DRE type, source, position, and count, and places reconsideration interventions before the conflict, after it, or in a separate review turn.

Experiments across models and three table benchmarks reveal a consistent evidence–trace failure (Section 3). Replacing a correct prefix with its conflicting counterpart causes accuracy drops of 49.4 to 78.7 percentage points across the tested settings (Section 3.1). The model nevertheless retains substantial recovery ability: on WTQ, an explicit post-conflict table-check cue improves accuracy by 67.3 percentage points, from 18.0% to 85.3%, whereas a reminder placed before the conflicting prefix provides little improvement (Section 3.2). Together, these results reveal a *self-initiation gap* under ETC: models often recover once externally induced to reconsider, but rarely do so autonomously. This suggests that failure to initiate recovery is a major bottleneck in this controlled setting. Error-stratified analysis further identifies the DREs for which recovery is most difficult (Section 3.3).

This diagnosis suggests turning externally elicited recovery into supervision for an unaided continuation. We call this strategy *recovery-pair supervised fine-tuning* (recovery-pair SFT). From the target model's own generations, we pair a correct prefix with its ordinary continuation and a conflicting prefix with a table-grounded recovery elicited through a separate review turn. We remove the review turn and train directly on these continuations: correct-prefix examples preserve normal continuation behavior, while conflicting-prefix examples teach the transition back to table evidence. At inference time, the model receives no reconsideration cue, critic output, or additional turn; it must decide whether recovery is needed (Sections 4–5). Recovery-pair SFT raises Qwen3-8B recovery on WTQ from 18.0% to 88.0% and transfers above 85% to HiTab and TableBench, converting the diagnostic contrast into a single-pass mitigation. Training-data ablations further separate the roles of the two supervision sources: conflicting-prefix examples teach the model how to recover, while correct-prefix examples calibrate when recovery should be initiated. Their combination reduces unnecessary re-checking on already-correct traces.

Our contributions are:

- We introduce the ETC protocol for diagnosing error recovery under evidence–trace conflict. Matched correct and conflicting prefixes isolate recovery from ordinary continuation, while controlled variations in DRE type, source, position, and intervention timing expose when models return to table evidence.
- We identify a widespread evidence–trace failure across model families and TableQA benchmarks. Models often follow a coherent but conflicting trace despite visible source evidence, while post-conflict reconsideration substantially improves recovery. Within ETC, this contrast reveals a self-initiation gap in evidence-grounded recovery.
- Guided by this diagnosis, we introduce recovery-pair SFT, pairing correct prefixes with ordinary continuations and conflicting prefixes with elicited recovery continuations. Training on WTQ transfers to HiTab and TableBench while retaining high correct-prefix accuracy, without a critic, resampling, or an extra inference turn.

2 THE EVIDENCE–TRACE CONFLICT (ETC) PROTOCOL

Task and matched-prefix formulation. For item i, let T_i be the table, q_i the question, and a_i^* the gold answer. We construct a matched pair

$$x_i^B = (T_i, q_i, p_i^B), \qquad x_i^C = (T_i, q_i, p_i^C), \tag{1}$$

where p_i^B is a truncated correct derivation and p_i^C is an internally consistent derivation containing one or more incorrect evidence references. The pair holds the table, question, prompt interface, conversational role, and continuation point fixed. Only the supplied reasoning trace is altered to introduce and consistently propagate the conflicting evidence. The model must continue either prefix and recover a_i^*.

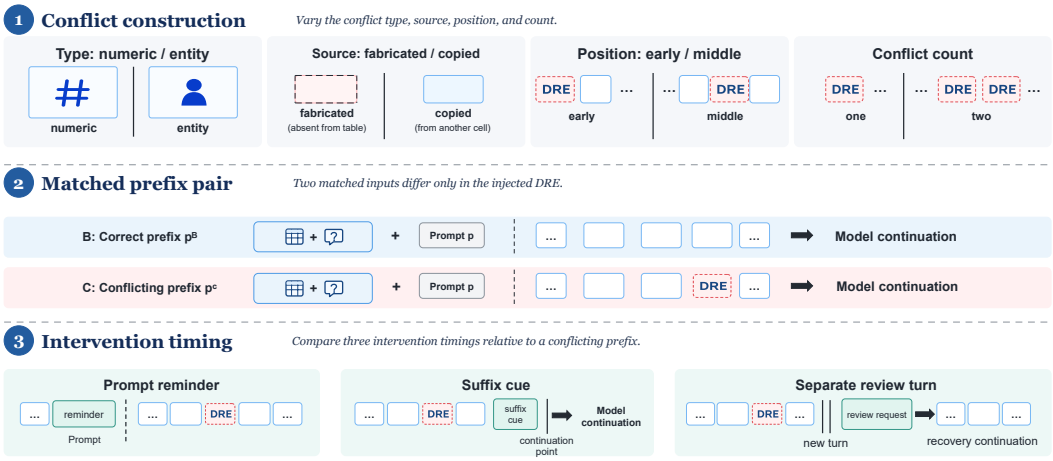


Figure 2: Overview of the ETC protocol. Conflict construction varies DRE type, source, position, and count while preserving matched correct and conflicting prefixes. Timing interventions are placed before the conflict, after it as a suffix, or in a separate review turn.

Conflict construction. As illustrated in Figure 2, we inject numeric or entity DREs at controlled positions (early / middle) while keeping the table and question fixed. Fabricated errors use a value absent from the table. Copied errors take a value from another cell, mimicking a plausible "wrong cell" read. A rewriting pipeline keeps the prefix internally consistent and ensures that each rewritten trajectory leads to an incorrect answer.

Constructing the evaluation sets. For each table–question pair, a stronger model generates a chain of thought (CoT), and we retain derivations that reach the gold answer with table-consistent evidence. We truncate each derivation at an early or middle point to form the correct prefix in condition B. To form its matched conflicting prefix in condition C, rule-based edits replace one or two earlier cited numbers or entities with fabricated values or values copied from another table cell. A stronger model checks whether the result remains logically natural and, when needed, lightly rewrites only the affected span from the injection step to the truncation point.

Each conflicting prefix then passes three stronger-model checks: consistency across reasoning steps; disagreement with the table, with the erroneous step and correct value identified; and answer impact, requiring a non-gold answer. We manually spot-check retained items. The final sets contain 150 matched WTQ items, 112 TableBench items, and 221 HiTab items.

Diagnostic axes and condition family. The ETC protocol's core conditions compare full generation (A), continuation from a correct prefix (B), and continuation from a conflicting prefix (C). Construction axes vary DRE type (numeric/entity), source (fabricated/copied), position (early/middle), and conflict count. Intervention axes vary when reconsideration is requested: in the prompt before the conflicting prefix, as a suffix after it, or in a separate review turn. Together, these controls test the breadth of evidence–trace failures, externally elicited recoverability, and variation in recovery difficulty. All conditions use the same Markdown-table template and temperature-0 decoding. Full prompts and results are in Appendices A.3 and A.7.

Evaluation targets and datasets. Answer-level recovery requires the condition-C continuation to produce a_i^* under our normalized exact-match rule. Evidence-level recovery further requires observable table re-grounding through explicit correction or implicit rereading and use of the supported value. Answer accuracy is the primary metric; output audits assess evidence-level recovery. We summarize the matched accuracy drop as

$$\Delta_{\text{rec}} = \text{Acc}(C) - \text{Acc}(B), \tag{2}$$

where a more negative value indicates stronger anchoring to the conflicting trace. The evaluation uses 1:1 matched B/C designs on WTQ (150), TableBench (112), and HiTab (221). Appendix A.6 gives the normalization rules; §5.3 reports the evidence-level audit.

3 Diagnosing Error Recovery under Evidence–Trace Conflict

We diagnose error recovery in three steps. We first measure the breadth and magnitude of evidence–trace failures across models and datasets. We then use intervention timing to compare unaided recovery with externally elicited recoverability, and finally characterize the DREs for which recovery is hardest.

3.1 Evidence–Trace Conflicts across Models and Datasets

Replacing a correct prefix with a conflicting one under the ETC protocol consistently shifts the continuation away from table evidence and toward the supplied trace. For Qwen3-8B, this replacement reduces accuracy by 78.7, 70.6, and 72.3 percentage points on WTQ, HiTab, and TableBench, respectively. Table 1(b) shows the same pattern across model families and sizes on WTQ. Even the strongest recovery in this comparison, Qwen3-4B thinking mode, remains 49.4 percentage points below its correct-prefix condition. The thinking result is only a capability reference because its prefix appears inside the model's thinking block; we do not treat it as a same-protocol baseline.

Table 1: Accuracy (%) under correct and conflicting reasoning traces. Negative Δ_{rec} values show that an erroneous trace can override visible table evidence across datasets and model families. Thinking mode is a capability reference because prefix placement differs.

(a) Across datasets: Qwen3-8B

Dataset	Correct prefix (B)	Conflict prefix (C)	Δ_{rec}
WTQ	96.7	18.0	−78.7
HiTab	90.5	19.9	−70.6
TableBench	95.5	23.2	−72.3

(b) Across models: WTQ

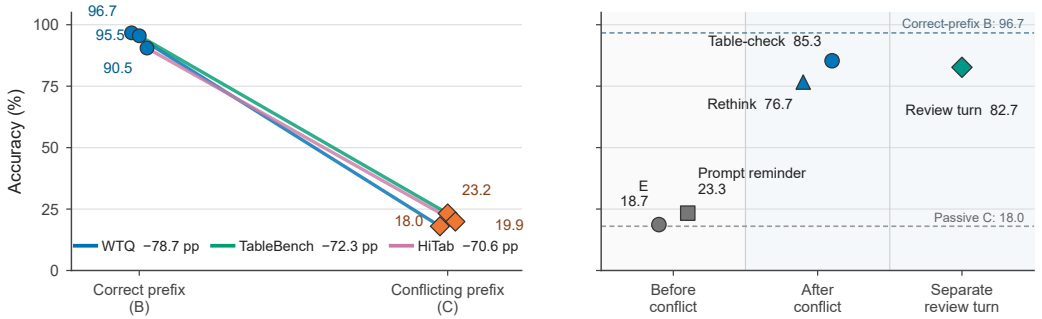
Model	Correct prefix (B)	Conflict prefix (C)	Δ_{rec}
Qwen3-4B	97.3	25.3	−72.0
Qwen3-4B thinking	96.7	47.3	−49.4
Qwen3-8B	96.7	18.0	−78.7
Llama-3.1-8B-Instruct	84.0	24.7	−59.3
Qwen3-30B-A3B	97.3	25.3	−72.0
DeepSeek-V4	98.0	22.7	−75.3

The effect is therefore not tied to one model or one table format. Even with the table visible, a fluent and internally consistent prefix containing incorrect evidence references can strongly anchor the continuation.

3.2 Cue Timing and Recovery Initiation

Starting from the conflicting-prefix condition, we request a return to table evidence at the three points shown in Figure 2: before the prefix in the prompt, after it as a suffix, or in a separate review turn. Appendices A.3 and A.7 provide the full conditions and results.

Pre-conflict reminders do not prevent trace following. The conflicting-prefix baseline (C) reaches 18.0%, compared with 18.7% for the original instruction from Yang et al. (2026) (E) and 23.3% for



(a) Conflicting prefixes cause 70.6–78.7 pp accuracy drops **(b) Post-conflict reconsideration restores recovery**

Figure 3: Conflicting traces sharply reduce accuracy across datasets, whereas post-conflict reconsideration restores most of the loss. (a) Matched $B \rightarrow C$ changes. (b) WTQ interventions by cue timing; dashed lines mark the unaided C and correct-prefix B references.

Prompt Reminder. These small gains show that an advance warning does not prevent the model from following the later conflict.

Post-conflict suffixes induce recovery. An explicit table-check suffix raises accuracy from 18.0% to 85.3%, a gain of 67.3 percentage points. Matched Qwen3-8B controls rule out a generic continuation restart or suffix-length explanation: "Now, let me continue." reaches 30.0%, whereas replacing *continue* with *rethink* reaches 76.7%; the corresponding "Wait" cues reach 64.7% and 76.7% (Appendix Table 11). Reconsideration wording therefore drives recovery even without an explicit table-check instruction, although table-check performs best. Most newly correct "Wait, let me continue." outputs also show explicit correction or implicit table re-grounding, suggesting that the interruption indirectly prompts a return to the table.

A separate review turn provides an external restart. Adding a user review request after the conflicting prefix raises accuracy to 82.7%, a gain of 64.7 percentage points over C, while retaining 92.7% on correct prefixes versus the 96.7% baseline. Thus, both suffixes and review turns restore recovery at different interaction points.

Figure 3 connects the large matched drops across datasets with the recovery induced by post-conflict interventions. We call this contrast the *self-initiation gap*: **the model often recovers once externally induced to interrupt an erroneous trajectory and reconsider, but rarely does so autonomously.** This diagnosis motivates our training objective.

3.3 Recovery Difficulty across DRE Types

The ETC protocol varies error type (numeric vs. entity), error source (fabricated vs. copied from another table cell), and error position (early vs. middle). Table 2 summarizes the resulting recovery rates.

Table 2: Recovery from evidence–trace conflicts on the WTQ constructed set with DeepSeek-V4. Copied, numeric, and middle-position DREs produce the lowest recovery rates.

Dimension	Setting	Acc. (%)	Dimension	Setting	Acc. (%)
Source	fabricated	31.0	Source	copied cell	19.4
Type	entity	30.2	Type	numeric	17.2
Position	early	27.2	Position	middle	12.8
Conflict count	1	19.8	Conflict count	2	28.6
Cross-strata: error type × source					
Numeric	copied	13.8	Numeric	fabricated	27.3
Entity	copied	27.9	Entity	fabricated	35.0

Recovery is lowest for copied rather than fabricated values (19.4% vs. 31.0%), numeric rather than entity DREs (17.2% vs. 30.2%), and middle-position rather than early-position DREs (12.8% vs. 27.2%). Cross-stratification confirms that numeric/copied errors are hardest (13.8%), while entity/-

fabricated errors are easiest (35.0%). Two-conflict examples recover more often than one-conflict examples (28.6% vs. 19.8%), possibly because repeated contradictions make the trace easier to question. **Copied, numeric, and middle-position DREs are hardest in this controlled set.**

4 From Diagnosis to Lightweight Mitigation

The multi-turn condition in §3.2 elicits table-grounded recovery after a conflicting prefix. We remove its review turn and pair the resulting continuation directly with that prefix. Figure 4 summarizes the construction.

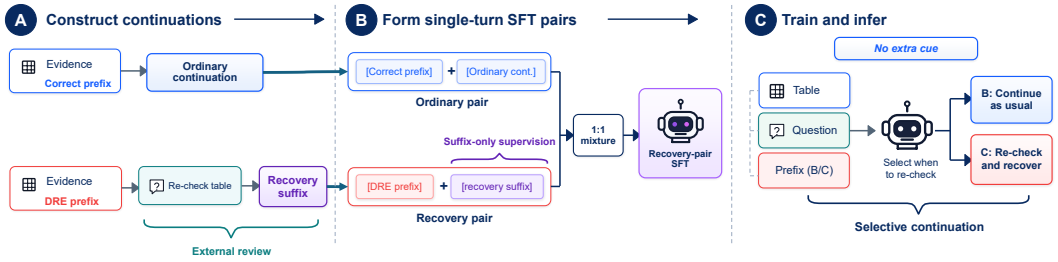


Figure 4: Constructing recovery-pair supervision. Matched correct and conflicting prefixes expose the behavioral gap. A post-conflict review turn elicits table-grounded recovery continuations; recovery-pair SFT removes that turn and trains the continuation directly from its prefix.

Training data. We build both sides of the training mixture from the target model's own generations. A correct chain of thought supplies the correct-prefix example and its ordinary continuation; its conflicting counterpart is constructed as described in §2.

Figure 4 shows how the Review Turn asks the same model to re-check table evidence after a conflicting prefix. The resulting continuation identifies or bypasses the wrong reference, returns to the table, and re-derives the answer. No larger teacher or external recovery trace is used.

Training objective. Training removes the second user turn. For each table T_i and question q_i, the model learns to continue according to the supplied prefix:

$$(T_i, q_i, p_i) \mapsto \begin{cases} y_i^{\mathrm{ord}}, & p_i = p_i^B \quad \text{(correct prefix)}, \\ y_i^{\mathrm{rec}}, & p_i = p_i^C \quad \text{(conflicting prefix)}. \end{cases} \tag{3}$$

Here y_i^{ord} continues the correct reasoning, while y_i^{rec} returns to table evidence and repairs the conflicting trace; both continuations lead to the gold answer a_i^*. We mix the two example types 1:1. The table, question, and prefix remain visible as context but are excluded from the loss; only the continuation is supervised. Thus, the model observes both table evidence and DRE tokens while learning the transition to recovery. We use standard supervised fine-tuning without introducing a special recovery marker. Training and evaluation use disjoint tables.

Inference. Inference recreates the unaided evidence–trace conflict in condition C. The model receives the table, question, and supplied prefix, then produces one continuation without a review prompt, critic output, or edited prefix. A recovery may explicitly name the earlier mistake or silently return to the table. Both count when the continuation clearly uses the correct evidence, so training targets table-grounded recovery rather than a fixed announcement.

5 Experiments

5.1 Setup

We train two backbones under the same recovery-pair SFT setup. Qwen3-8B receives approximately 400 recovery pairs (800 SFT instances) constructed from WTQ training tables, with correct and conflicting prefixes mixed 1:1. Llama-3.1-8B-Instruct is trained on 189 retained correct/conflicting-prefix pairs (378 SFT instances) from WTQ. We evaluate both models on WTQ

Table 3: Accuracy (%) under correct (B) and conflicting (C) reasoning traces. The comparison tests whether the recovery behavior exposed by our diagnosis can be internalized without inference-time cues, critics, or resampling. Each Avg. column is the arithmetic mean of B and C; the paired evaluations use a 1:1 ratio. For C and Avg., the best and second-best values within each backbone block are bolded and underlined, respectively; B values are not ranked.

Method	WTQ			TableBench			HiTab		
	B	C	Avg.	B	C	Avg.	B	C	Avg.
Qwen3-8B									
Direct	96.7	18.0	57.4	95.5	23.2	59.4	90.5	19.9	55.2
Few-shot	99.3	16.0	57.7	97.3	17.9	57.6	92.3	15.8	54.1
SC (N=8)	96.7	16.7	56.7	94.6	17.0	55.8	92.3	17.7	55.0
Yang-style	96.7	20.0	58.4	94.6	25.0	59.8	93.2	27.2	60.2
SSC	97.3	<u>74.0</u>	<u>85.7</u>	95.5	64.3	79.9	90.1	39.8	65.0
Table-Critic	96.0	66.0	81.0	95.5	<u>65.2</u>	<u>80.4</u>	83.7	<u>50.7</u>	<u>67.2</u>
Self-Refine	53.3	45.3	49.3	67.9	50.9	59.4	38.9	17.2	28.1
Recovery-pair SFT	98.7	**88.0**	**93.4**	96.4	**86.6**	**91.5**	90.5	**85.5**	**88.0**
Llama-3.1-8B-Instruct									
Direct	84.0	24.7	54.4	86.6	22.3	54.5	67.9	17.7	42.8
Few-shot	93.3	18.7	56.0	87.5	19.6	53.6	59.7	19.5	39.6
SC (N=8)	94.7	36.0	65.4	90.2	33.9	62.1	71.0	28.1	49.6
Yang-style	94.7	40.7	67.7	90.2	34.8	62.5	73.8	32.6	53.2
SSC	70.7	<u>71.3</u>	<u>71.0</u>	66.1	**68.8**	67.5	46.2	<u>50.7</u>	48.5
Table-Critic	87.3	50.7	69.0	90.2	54.5	**72.4**	71.0	42.5	<u>56.8</u>
Self-Refine	68.0	46.7	57.4	61.6	48.2	54.9	54.8	41.6	48.2
Recovery-pair SFT	89.3	**75.3**	**82.3**	75.0	<u>67.9</u>	<u>71.5</u>	69.2	**74.7**	**72.0**

(150), TableBench (112), and HiTab (221); the latter two datasets are never used for training. For TableBench (Wu et al., 2025), we follow Yang et al. (2026) and use its Fact Checking and Numerical Reasoning categories.

The trained and untrained models are evaluated locally. All methods use the same normalized exact-match rule described in Appendix A.6. We compare Direct decoding, Few-shot prompting, self-consistency (SC, N=8) (Wang et al., 2023), SSC (Shaikh, 2026), Table-Critic (Yu et al., 2025), and Self-Refine (Madaan et al., 2023) on the same B/C evaluation sets. We additionally report a Yang-style SC-plus-critic pipeline inspired by Yang et al. (2026); it is not a component-for-component reproduction, so we label it Yang-style rather than the original method. Detailed baseline settings and information-access assumptions are given in Appendix A.1; prompts are listed in Appendix A.3.

5.2 Cue-Free Recovery and Cross-Dataset Transfer

We test cue-free, single-pass recovery and transfer to unseen table formats. Table 3 reports B, C, and their arithmetic mean on paired evaluation sets. For Qwen3-8B, recovery-pair SFT raises C accuracy from 18.0% to 88.0% on WTQ, from 23.2% to 86.6% on TableBench, and from 19.9% to 85.5% on HiTab. These changes correspond to absolute gains of 63.4–70.0 percentage points across the three datasets. B remains high at 98.7%, 96.4%, and 90.5%, respectively.

Few-shot prompting, SC, and the Yang-style pipeline provide limited improvement because their continuations remain constrained by the erroneous prefix. By contrast, SSC, Table-Critic, and Self-Refine generally improve C accuracy relative to Direct through multi-pass verification, refinement, or multi-agent critique. Self-Refine also sharply reduces B accuracy, suggesting that repeated refinement can over-correct traces that already agree with the table.

Llama-3.1-8B shows the same transfer pattern: C rises from 24.7% to 75.3% on WTQ, from 22.3% to 67.9% on TableBench, and from 17.7% to 74.7% on HiTab, yielding absolute gains of 45.6–57.0 percentage points. Its B accuracy falls from 86.6% to 75.0% on TableBench, so its improvement on conflicting prefixes comes with a larger loss on correct prefixes than for Qwen3-8B. Together, the two runs show that elicited recovery can be internalized across model families without a cue, critic, resampling procedure, or additional inference turn.

5.3 Evidence Re-grounding in Recovered Outputs

Each conflicting prefix is checked to yield a wrong answer if followed, so a correct C answer implies departure from the erroneous trajectory but does not reveal how recovery occurred. We conduct an exploratory manual audit of the 132 correct C continuations from the trained Qwen3-8B on

Table 4: Qwen3-8B training-data ablation. Conflicting-prefix examples provide the supervision for recovery, while adding correct prefixes preserves high recovery and shortens correct-prefix continuations. All models use WTQ and 2,400 instance presentations; Tokens is the average total per question.

Dataset	Training data	B		C	
		Acc.	Tokens	Acc.	Tokens
WTQ	Correct-only	98.7	573.0	18.0	579.2
WTQ	Wrong-only	98.7	753.9	92.7	763.0
WTQ	Correct + wrong (1:1)	98.7	719.2	88.0	743.3
TableBench	Wrong-only	95.5	817.3	88.4	819.3
TableBench	Correct + wrong (1:1)	96.4	766.2	86.6	809.0
HiTab	Wrong-only	93.7	1006.2	86.4	991.4
HiTab	Correct + wrong (1:1)	90.5	991.7	85.5	993.0

WTQ. They either explicitly identify the earlier error (e.g., "not X" or "the previous reasoning is incorrect") or reread the table and use the correct evidence without explicitly naming the error; none jumps directly to the answer without re-grounding. The untrained thinking reference shows the same patterns, but explicit correction appears less common. Thus, training appears to make re-grounding more visible, although the objective does not require explicit re-check wording.

5.4 Accuracy–Cost Trade-off

We compare the token cost of recovery-pair SFT against multi-call baselines. Figure 5 summarizes the C-condition accuracy–cost trade-off. On Qwen3-8B, recovery-pair SFT uses 743 tokens on WTQ, 809 on TableBench, and 993 on HiTab; the Llama-3.1-8B model uses 675, 713, and 995 tokens, respectively. It is substantially cheaper than the competing sampling, critic, and multi-stage baselines. Appendix Table 7 reports the complete B/C token statistics, including the direct baselines.

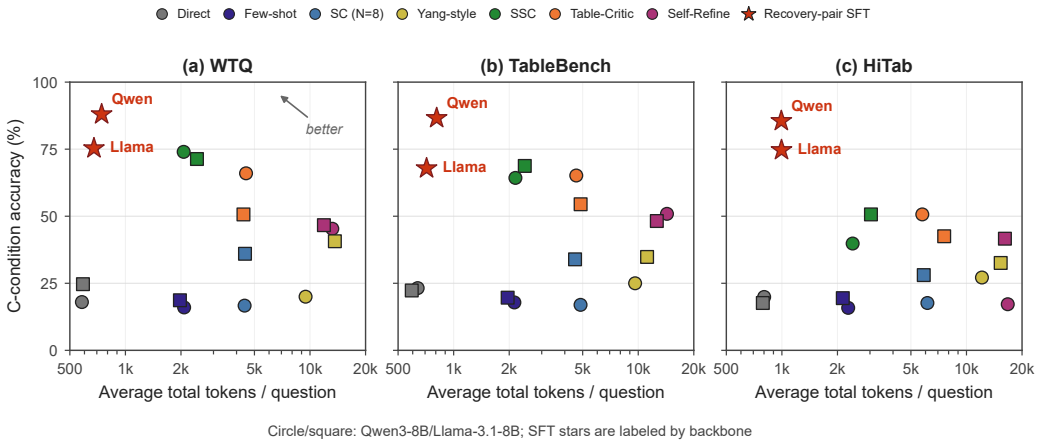


Figure 5: Conflicting-prefix recovery accuracy versus average total tokens per question (log-scaled x-axis). Recovery-pair SFT occupies the high-accuracy, low-cost region because it returns to table evidence without resampling or an additional inference turn. Color denotes method; circles and squares denote Qwen3-8B and Llama-3.1-8B, and stars denote recovery-pair SFT.

5.5 Roles of Correct and Conflicting Training Prefixes

We test which examples teach recovery and which calibrate its initiation. The single-source models use 400 instances for six epochs, while the 1:1 model uses 800 instances for three epochs, giving each setting 2,400 instance presentations. Table 4 reports this matched-budget comparison.

Correct-only training leaves WTQ C at 18.0%, whereas wrong-only training reaches 92.7% and transfers to TableBench and HiTab. Conflict-recovery examples therefore teach *how* to recover. Correct-prefix examples instead improve selectivity: the 1:1 mixture reduces unnecessary re-checking and completion length on already-correct traces while retaining high conflicting-prefix recovery. Appendix Table 8 reports the full audit.

6 Discussion and Limitations

Interpretation. Coherent traces can override conflicting table evidence, yet post-conflict reconsideration often returns models to the table; matched controls rule out a generic continuation effect. This contrast identifies failure to initiate recovery as a major source of error for controlled DREs without specifying a unique internal mechanism. Recovery-pair SFT internalizes this behavior and transfers from WTQ to TableBench and HiTab, indicating that it is not tied to answer formatting. ETC could extend to document QA, fact verification, and chart or image QA, where recovery entails re-grounding in text or visual evidence.

Exploratory natural-DRE evaluation. We additionally evaluate a small filtered set of naturally occurring DREs, where errors are not restricted to the controlled substitutions used by ETC (Appendix Table 9). Unaided recovery is only 3.4%, while an explicit review turn and self-check intervention raise recovery to 27.1% and 35.6%, respectively. Recovery-pair SFT, trained only on controlled conflicts, improves unaided recovery to 15.3%. These results are consistent with recovery initiation remaining relevant beyond synthetic conflicts, but also show that natural DREs introduce additional challenges such as error localization and repair. We therefore use ETC to isolate the initiation component rather than claim that initiation alone explains natural-error recovery.

Limitations. Our main experiments train Qwen3-8B and Llama-3.1-8B-Instruct on WTQ examples and evaluate controlled numeric and entity substitutions. The natural-DRE pilot remains small, and its recovery accuracy remains substantially below the controlled ETC setting. Because natural errors mix initiation with error localization and repair, this pilot should not be interpreted as a controlled estimate of the self-initiation gap. Recovery from omissions also remains untested. Because the training conflicts are synthetic, the distribution-shift concern raised by EIFT (Wu et al., 2026) remains. Thinking mode has a prefix-position confound, and output audits do not establish the model's internal causal mechanism.

7 Related Work

Table reasoning, verification, and self-correction. Table reasoning methods use parsing, decomposition, resampling, or external verification. DRE-oriented work detects faulty references and uses critic-based filtering or rejection sampling (Yang et al., 2026). SSC builds table-specific verification chains (Shaikh, 2026), while Table-Critic uses multi-agent critique (Yu et al., 2025). Self-Refine and Reflexion use multi-turn feedback (Madaan et al., 2023; Shinn et al., 2023), while other methods internalize retrieval or reflection signals (Asai et al., 2024; Zelikman et al., 2024; Dhuliawala et al., 2024). ContraTable creates conflicts by editing the table (Wang & Liu, 2026). In contrast, ETC holds the table fixed and intervenes on a reasoning prefix containing one or more DREs.

Reasoning intervention and recovery. Controlled chain-of-thought studies show that continuations depend strongly on earlier reasoning prefixes (von Recum et al., 2026; Iwase et al., 2026). Recovery methods use external restarts or local repair (Chenkang et al., 2026; Cheng et al., 2026c; Samanta et al., 2026), reinforcement learning over noisy prefixes (Xu et al., 2026), or fixed recognition-and-correction patterns for synthetic errors (Wu et al., 2026). Vision-language work contrasts spontaneous reconsideration with turn-boundary interventions and repairs trajectories using visual constraints, tools, or retries (Shi et al., 2026; Cheng et al., 2026a;b). Tsui (2026) isolates a self-correction blind spot by varying whether the same error is user- or model-attributed. ETC instead holds conversational role fixed and varies evidence–trace consistency with visible table evidence. We train full recovery continuations without a fixed marker, critic, or extra inference turn.

8 Conclusion

When an otherwise coherent reasoning trace conflicts with visible table evidence, language models often continue the error, yet a brief post-conflict reconsideration cue restores much of the lost accuracy by returning them to the table. Separating unaided from externally elicited recovery reveals a behavioral self-initiation gap in evidence-grounded recovery: models often recover after an external trigger but do not reliably initiate recovery themselves. Recovery-pair SFT internalizes this behavior and transfers from WTQ to two unseen table benchmarks while preserving high correct-prefix accuracy and requiring no additional inference-time mechanism.

AI USE STATEMENT

Large language models are intrinsic to the study rather than being used only as writing aids. During ETC data construction, a stronger model generates candidate correct reasoning traces, checks the internal consistency, table conflict, and answer impact of edited prefixes, and locally rewrites affected spans when direct substitution would make a trace unnatural. The target models also generate the ordinary and recovery continuations used for recovery-pair SFT, and model-based critics or evaluators are used by several comparison methods. All retained evaluation items follow the filtering and manual spot-checking procedure described in Section 2. Generative AI tools were also used to assist language editing and LaTeX preparation. The authors designed the protocol, selected and verified the data, ran the experiments, audited the outputs, and take responsibility for the manuscript and its conclusions.

ETHICS STATEMENT

This study uses existing TableQA benchmarks and model-generated reasoning traces; it does not involve human subjects or the collection of new personal data. The conflicting prefixes intentionally contain incorrect table references and are used only as controlled research artifacts paired with their source tables and gold answers. They should not be interpreted as factual annotations outside this evaluation setting. The study evaluates a specific form of evidence-grounded recovery and does not establish that the tested models are reliable in safety-critical or open-domain deployments.

REPRODUCIBILITY STATEMENT

The paper and appendix report the ETC construction procedure, dataset sizes, condition definitions, prompts, model and decoding settings, training hyperparameters, random seed, baseline configurations, normalization rules, and token accounting. Training tables are disjoint from all evaluation tables. Anonymized reproduction code is included in the Supplementary Material; we will publicly release it upon acceptance.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11511. arXiv:2310.11511.

Yu Cheng, Arushi Goel, and Hakan Bilen. Think, look, and revise: Inconsistency-aware visual self-correction in MLLMs, 2026a. URL https://arxiv.org/abs/2608.29374. Accepted at EMNLP 2026.

Zehua Cheng, Wei Dai, and Jiahao Sun. Constraint-anchored reasoning traces, 2026b. URL https://arxiv.org/abs/2607.16727. ACM Multimedia 2026.

Zehua Cheng, Wei Dai, Jiahao Sun, and Thomas Lukasiewicz. Imbuing large language models with bidirectional logic for robust chain repair, 2026c. URL https://arxiv.org/abs/2606.05030.

Chenkang, Fan Yu, Junjie Nian, Sihan Zhao, Zhuoka Feng, Zijun Yao, Wang Heng, Yu Minshen, and Yixin Cao. Thinking traps in long chain-of-thought: A measurable study and trap-aware adaptive restart. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38746–38760, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1930. URL https://aclanthology.org/2026.findings-acl.1930/.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.

findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/. arXiv:2309.11495.

Naoto Iwase, Yuki Ichihara, Mohammad Atif Quamar, and Junpei Komiyama. Reliable chain-of-thought via prefix consistency, 2026. URL https://arxiv.org/abs/2605.07654.

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. A survey on table question answering: Recent advances, 2022. URL https://arxiv.org/abs/2207.05270.

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. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2303.17651. arXiv:2303.17651.

Ankur Samanta, Akshayaa Magesh, Ayush Jain, Kavosh Asadi, Youliang Yu, Daniel Jiang, Boris Vidolov, Kaveh Hassani, Paul Sajda, Jalaj Bhandari, and Yonathan Efroni. Structure enables effective self-localization of errors in LLMs. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.02416. arXiv:2602.02416.

Farseen Shaikh. Can LLMs self-correct table reasoning errors? In *Proceedings of the First Workshop on Structured Understanding, Retrieval, and Generation in the LLM Era (SURGeLLM 2026)*, pp. 298–312, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.surgellm-1.19. URL https://aclanthology.org/2026.surgellm-1.19/.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs seeing or just saying? uncovering the illusion of visual re-examination. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.15864. Oral presentation; arXiv:2605.15864.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377. URL https://papers.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Ken Tsui. Self-correction bench: Uncovering and addressing the self-correction blind spot in large language models. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2507.02778. arXiv:2507.02778.

Alexander von Recum, Leander Girrbach, and Zeynep Akata. Are reasoning LLMs robust to interventions on their chain-of-thought? In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a03037317560b8c5f2fb4b6466d4c439-Abstract-Conference.html. arXiv:2602.07470.

Xinzhi Wang and Chunwei Liu. The table says otherwise: Testing llms with counterfactual relational data, 2026. URL https://arxiv.org/abs/2606.23667.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. arXiv:2203.11171.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=4L0xnS4GQM.

David X. Wu, Shreyas Kapur, Anant Sahai, and Stuart Russell. Synthetic error injection fails to elicit self-correction in language models. In *I Can't Believe It's Not Better: Where Large Language Models Need to Improve (ICBINB) Workshop at ICLR 2026*, 2026. URL https://iclr.cc/virtual/2026/10014635. ICLR 2026 Workshop ICBINB; arXiv:2512.02389.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. TableBench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25497–25506, 2025. doi: 10.1609/aaai.v39i24.34739. URL https://doi.org/10.1609/aaai.v39i24.34739.

Caijun Xu, Changyi Xiao, Zhongyuan Peng, and Yixin Cao. Denoiserl: Bootstrapping reasoning models to recover from noisy prefixes, 2026. URL https://arxiv.org/abs/2605.28421.

Yuqing Yang, Qi Zhu, Zhen Han, Boran Han, Zhengyuan Shen, Shuai Wang, Vassilis N. Ioannidis, and Huzefa Rangwala. When LLMs read tables carelessly: Measuring and reducing data referencing errors. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16734–16752, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.762. URL https://aclanthology.org/2026.acl-long.762/.

Peiying Yu, Guoxin Chen, and Jingjing Wang. Table-critic: A multi-agent framework for collaborative criticism and refinement in table reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17432–17451, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.853. URL https://aclanthology.org/2025.acl-long.853/.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-star: Language models can teach themselves to think before speaking, 2024. URL https://arxiv.org/abs/2403.09629. arXiv:2403.09629.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. A survey of table reasoning with large language models, 2024. URL https://arxiv.org/abs/2402.08259.

Wei Zhou, Bolei Ma, Annemarie Friedrich, and Mohsen Mesgar. Table question answering in the era of large language models: A comprehensive survey of tasks, methods, and evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12161–12189, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.557. URL https://aclanthology.org/2026.acl-long.557/.

A Appendix

The appendix contains:

- training and baseline configurations (§A.1);
- complete token costs (§A.2);
- prompts for the diagnostic and intervention conditions (§A.3);
- the recovery selectivity audit (§A.4);
- the exploratory natural-DRE evaluation (§A.5);
- evaluation and answer-normalization details (§A.6);
- full intervention results and matched suffix controls (§A.7);
- the tagged-pilot diagnostic (§A.8).

A.1 Training and Baseline Configuration

ETC construction model. The stronger model used to generate candidate reasoning traces, check edited prefixes, and rewrite affected spans is DeepSeek-V4. It is not used as the main model for recovery evaluation or SFT.

Shared data card and evaluation. We evaluate the same English table–question pairs from WTQ (150 items), TableBench (112), and HiTab (221) under two paired conditions. In condition B, the supplied assistant prefix contains correct table evidence; in condition C, one cited value in that prefix is replaced by a conflicting value while the table and question remain unchanged. All runs use the full Markdown table, Transformers inference in bfloat16 with `device_map=auto`, SDPA attention, and `trust_remote_code`; thinking mode and caching are disabled. We score the last "Answer:" using the shared normalized exact-match rule described in §A.6. No test questions are used for training, and the training data contain no external teacher answers.

Recovery-pair SFT. Each final SFT instance pairs a table/question prompt with a correct or conflicting prefix and an ordinary-language continuation. The final objective is marker-free next-token SFT: no `<reground>` token is inserted. The prompt and supplied prefix remain visible in the input but are excluded from the loss; only the continuation is supervised. Correct and conflicting-prefix instances are mixed approximately 1:1. The `<reground>` variant is retained only as a separate tagged pilot that motivated this marker-free formulation. The common recipe is summarized below; the adapter is merged with the base model at inference.

Table 5: Compact Recovery-pair SFT recipe.

Setting	Configuration
Training data	Qwen3-8B: approximately 400 correct/conflicting pairs, yielding 800 SFT instances (400+400; `qwen8b_en_train_400_clean.jsonl`); Llama-3.1-8B: approximately 200 pairs, with 189 retained pairs yielding 378 SFT instances (189+189; `llama_en_train_200.jsonl`); an optional Llama 432-instance no-critic adapter uses the same recipe.
Optimization	3 epochs; learning rate 2×10^{-4} with cosine decay; 50-step warm-up; weight decay 0.01; maximum length 4096.
Batching	Per-device batch size 1, gradient accumulation 16 (effective batch size 16); gradient checkpointing on; packing off; seed 42.
LoRA	Rank 16, $\alpha = 32$, dropout 0.05, bias `none`; inject into `q/k/v/o_proj` and `gate/up/down_proj`; no 4-bit QLoRA.
Precision	bfloat16.

Completed baselines. Table 6 records the settings that determine the completed comparisons. Multi-call methods count every realized call in their token totals. Critic, reviewer, and judge prompts receive the table, question, and generated candidate(s), but no reference answer; gold answers are used only for post-hoc scoring. Internal voting uses exact equality after string normalization, whereas final accuracy uses the shared normalized exact-match rule.

Table 6: Core settings for every completed comparison method.

Method	Core configuration
Direct CoT	Untrained base model; greedy decoding ($T = 0$), 1 call, maximum 1,024 new tokens.
Few-shot	Fixed four WTQ demonstrations (two normal and two correction examples; IDs `7819, 6875, 3197, 5524`); greedy, 1,024 new tokens, one call.
Self-Consistency (SC) (Wang et al., 2023)	Eight independent candidates at $T = 0.7$, 1,024 new tokens each; majority vote. Seeds are $42 + i \times 8 + j$ for item i and candidate j.
Yang-style DRE filtering (Yang et al., 2026)	Eight candidates with temperatures $0.00, 0.17, 0.34, 0.51, 0.69, 0.86, 1.03, 1.20$ and 4,096 new tokens each; a DRE critic (2,048-token budget) counts incorrect citations and omitted information by `Step N`, using only the candidate continuation and no gold answer; retain the minimum-count set and vote. Both Qwen and Llama use the same DeepSeek critic (`deepseek-v4-flash`).
Table Self-Correction (SSC) (Shaikh, 2026)	Init and reviewer, followed by a conditional refinement when judged incorrect; each stage has a 2,048-token budget, $T = 0$. The reviewer checks CELL, MATH, LOGIC, and COMPLETENESS.
Table-Critic (Yu et al., 2025)	Init → Judge; if incorrect, Tree → Critic → Refine, up to five iterations. Init/Judge allow 1,024 new tokens and later agents 2,048; no cache, one worker.
Self-Refine (Madaan et al., 2023)	Init → Feedback → Refine for at most three rounds; budgets 1,024/2,048/2,048 new tokens; greedy, two generic toy-table demonstrations, no cache.
Recovery-pair SFT	LoRA model trained with the recipe in Table 5; one greedy continuation per item, maximum 1,024 new tokens.

The implementation entry points are `train_sft.py` and the Qwen/Llama training wrappers for SFT; `sc_n8.py` for SC; `fewshot_eval.py` for Few-shot; and the scripts un-

der `table_referencing_Yang`, `table_self_correction_SSC`, `table_critic`, and `self_refine` for the remaining baselines.

A.2 Complete Token Costs

Table 7 reports B/C token statistics for the completed comparison methods. Figure 5 uses the C columns, including the direct baselines.

Table 7: Average total tokens per question for the completed baseline runs. Yang-style reports the complete SC-plus-critic pipeline.

	WTQ		TableBench		HiTab	
Method	B	C	B	C	B	C
Qwen3-8B						
Direct	574	581	631	638	790	802
Few-shot	2,077	2,080	2,131	2,133	2,285	2,287
SC (N=8)	4,371	4,423	4,844	4,869	6,094	6,142
Yang-style	8,973	9,463	9,416	9,617	11,616	12,128
SSC	1,544	2,068	1,694	2,161	1,935	2,417
Table-Critic	2,284	4,503	1,866	4,616	2,601	5,763
Self-Refine	13,329	13,209	14,912	14,309	16,979	16,733
Recovery-pair SFT	**719**	**743**	**766**	**809**	**992**	**993**
Llama-3.1-8B-Instruct						
Direct	567	589	584	593	792	788
Few-shot	1,929	1,974	1,953	1,961	2,116	2,138
SC (N=8)	4,251	4,445	4,440	4,546	5,919	5,869
Yang-style	10,177	13,637	9,767	11,166	15,226	15,326
SSC	2,417	2,441	2,659	2,428	2,758	3,038
Table-Critic	2,377	4,360	2,425	4,874	4,759	7,572
Self-Refine	11,822	11,894	13,253	12,610	15,671	16,152
Recovery-pair SFT	**663**	**675**	**730**	**713**	**1,154**	**995**

A.3 Prompts

All conditions share the base user template:

```
[TABLE]
{markdown table}

[QUESTION]
{question}

Please reason step by step, and finally give the final answer with "
    Answer:".

Please reason step by step, and label each step with Step N (e.g., Step
    1: ...).
```

Table-Check Suffix (appended to a correct or conflicting assistant prefix):

```
Wait, let me first verify the table data referenced in the reasoning
    above:
```

Rethink Suffix (appended to the conflicting assistant prefix):

```
Wait, let me rethink.
```

Matched suffix controls (appended to the assistant prefix):

```
Now, let me continue.
Wait, let me continue.
Now, let me rethink.
Wait, let me rethink.
```

Pre-conflict table-check reminder (appended to the user instruction):

```
Before continuing, check whether all numbers, names, and table structures
     mentioned in each step above exactly match the table; if you find
    any inconsistency, point it out and correct it first, then continue
    reasoning.
```

Prompt Reminder (appended to the user instruction):

```
During reasoning, continuously self-check data and facts for every step.
    If you spot contradictions or errors, correct them prior to further
    derivation.
```

Review Turn (added after a conflicting or correct prefix):

```
Please re-check whether every number, name, and table structure mentioned
     in your reasoning above is consistent with the table; if you find
    any inconsistency, point it out and correct it, then continue
    reasoning. Finally, give the final answer with "Answer:".
```

E instruction from Yang et al. (2026) (appended to the user instruction):

```
Use only the table. Do not omit, miscite, or fabricate information.
    Ensure all cited values exactly match the table.
```

A.4 Recovery Selectivity Audit

Table 8 compares wrong-only training with the 1:1 recovery-pair mixture on correct-prefix examples. The mixture reduces explicit re-checking and average completion length across all three datasets.

Table 8: Recovery selectivity on correct (B) prefixes. Compared with wrong-only training, the 1:1 recovery-pair mixture reduces explicit re-checking and completion length when the supplied trace already agrees with the table.

	Any explicit re-check		Avg. completion tokens	
Dataset	Wrong-only	Recovery-pair SFT	Wrong-only	Recovery-pair SFT
WTQ	41/150 (27.3%)	35/150 (23.3%)	204.8	170.1
HiTab	50/221 (22.6%)	28/221 (12.7%)	256.5	242.0
TableBench	25/112 (22.3%)	9/112 (8.0%)	219.5	168.4

On the WTQ no-prefix condition (A), the Qwen3-8B recovery-pair LoRA model obtains 130/150 (86.7%), compared with 132/150 (88.0%) before fine-tuning.

A.5 Exploratory Natural-DRE Evaluation

Table 9 reports an exploratory evaluation on a small filtered set of naturally occurring DREs. Unlike ETC, these errors may require error localization and repair in addition to recovery initiation.

Table 9: Exploratory recovery accuracy on naturally occurring DREs. Gains are absolute percentage-point changes over Base.

Setting	Natural-DRE accuracy	vs. Base
Base	3.39%	—
Review Turn	27.12%	+23.73 pp
Self-check	35.59%	+32.20 pp
Recovery-pair SFT	15.25%	+11.86 pp

A.6 Evaluation Details

We first compare the normalized final answer with the reference answer under a normalized exact-match rule. Numeric answers are matched after normalization with numeric equivalence; for entity answers, parenthetical annotations are removed and the first candidate before "or" is used. Unless noted otherwise, the main tables report this normalized final-answer accuracy. The recovery audit additionally checks whether a correct continuation uses the table evidence and whether it explicitly identifies the earlier DRE.

A.7 Full Intervention Results

Table 10 gives the complete diagnostic run. The main text reports only the rows needed to establish the timing contrast. For reproducibility, the internal run identifiers map to the reader-facing names as follows: `B_self` = B + Table-Check Suffix, `D_C_selfcheck` = C + Table-Check Suffix, `D_C_rethink` = C + Rethink Suffix, `D_C_prompt` = C + Prompt Reminder, `D_C_multi` = C + Review Turn, and `D_B_multi` = B + Review Turn.

Table 10: Full condition-level accuracy on WTQ (150 items, Qwen3-8B, temperature 0).

Condition	Configuration	Accuracy
A	no prefix	88.00%
B	correct prefix	96.67%
B + Table-Check Suffix	correct prefix + check suffix	94.67%
C	conflicting prefix	18.00%
C + Table-Check Suffix	conflicting prefix + check suffix	85.33%
C + Rethink Suffix	conflicting prefix + rethink suffix	76.67%
C + Prompt Reminder	pre-conflict reminder	23.33%
C + Review Turn	review turn after conflicting prefix	82.67%
B + Review Turn	review turn after correct prefix	92.67%
E	original instruction from Yang et al. (2026)	18.67%

Table 11: Qwen3-8B matched controls on the same 150 WTQ items. The no-suffix C baseline is 27/150 (18.0%). A post-hoc audit of "Wait, let me continue." finds 66 explicit corrections (49 correct), 56 implicit table re-groundings (46 correct), and 28 outputs without sufficient re-grounding (2 correct); 70/72 newly correct outputs fall in the first two groups.

Matched suffix controls

Discourse marker	Continue	Rethink
Now	45/150 (30.0%)	115/150 (76.7%)
Wait	97/150 (64.7%)	115/150 (76.7%)

A.8 Tagged-Pilot Diagnostic

The initial tagged Qwen3-4B pilot makes the learned trigger observable, although the final method is marker-free. With plain-text ⟨`reground`⟩ participating in the loss, the marker is emitted for 7/75 correct prefixes (all seven answers correct) and 57/75 conflicting prefixes (56 correct); the corresponding numbers of correct answers without emission are 67 and 8.