

MIRAGE: Reliability Assessment via Grounded Evidence for Table Reasoning

Jinlong Tian¹ Ruochun Jin^{1,†} Shixuan Liu^{1,‡} Yanfang Zhou²
Mengmeng Li² Haotian Wang¹ Xinhai Xu^{2,†} Yuhua Tang¹

¹National University of Defense Technology, Hunan, China

²Academy of Military Sciences, Beijing, China

Abstract

Uncertainty estimation for language models often treats reliability as a property of the answer distribution: high confidence, stable samples, or low semantic entropy are taken as signs that an answer can be trusted. Table reasoning exposes a fundamentally different failure mode. A model may consistently return the same answer while linking the wrong entity, selecting the wrong column, applying the wrong operation, or relying on incomplete or insufficient evidence. We introduce MIRAGE—**Model Inference Reliability Assessment via Grounded Evidence**, a structure-grounded reliability framework for selective table reasoning. MIRAGE estimates residual structural uncertainty from test-time traces, combining outcome consistency with grounding and computation signals, and produces calibrated answer-or-abstain decisions with lightweight reliability certificates. On a controlled four-family diagnostic benchmark, structural reliability signals substantially reduce selective risk compared with generic confidence, sampling, semantic-entropy, and hallucination-detection baselines.

1 Introduction

Large language models are increasingly used for table question answering and table reasoning tasks, where systems must answer natural-language questions using structured tabular data (Liu et al., 2024b). Unlike free-form text generation, correct table reasoning often depends on resolving implicit structural decisions, such as identifying the relevant entity or row, selecting the correct schema target, applying the intended aggregation, and combining evidence across tables (Zhang et al., 2024; Liu et al., 2024a). Incorrect entity grounding, schema selection, or aggregation can

still produce answers that appear stable across reasoning traces, even when the underlying table support is wrong, making reliability estimation particularly challenging (Tian et al., 2023; Srinivasan et al., 2024; Wang et al., 2024). A trustworthy table reasoning system should recognize when an answer is reliably supported by the underlying table structure and abstain rather than produce a hallucinated answer when critical evidence remains insufficient or structurally inconsistent (Zhang et al., 2025a).

Most uncertainty estimation methods for language models are answer-distribution based. Confidence methods ask whether the model assigns high likelihood to an answer or verbally reports high confidence (Guo et al., 2017). Sampling methods ask whether multiple generations agree (Wang et al., 2023). Semantic uncertainty methods cluster sampled responses by meaning and measure dispersion over semantic alternatives. These signals are valuable when uncertainty appears as variability in the model’s answer distribution. However, table reasoning exposes a failure mode that these signals can miss: uncertainty can be structural rather than distributional (Zhang et al., 2024). A model may consistently choose the wrong entity, bind a question to a plausible but incorrect column, aggregate the wrong rows, or hallucinate missing cross-table evidence. The final answer may be stable across samples, semantically consistent, and accompanied by high self-reported confidence (Kadavath et al., 2022; Tian et al., 2023). In these cases, answer consistency and self-reported confidence can become a mirage: they appear reliable even when the underlying table support is structurally incorrect.

We believe that reliability estimation for table reasoning should ask not only whether the answer is likely or stable, but whether the reasoning process has satisfied the structural requirements needed for the answer to be trusted. We refer to the

[†]Corresponding authors.

[‡]Project Leader.

remaining risk as residual structural uncertainty: uncertainty about whether an answer is correctly grounded, properly computed, and sufficiently supported by the table evidence. This paper introduces MIRAGE: **M**odel **I**nference **R**eliability **A**ssessment via **G**rounded **E**vidence. Given a question, table context, candidate answer, and observable reasoning signals, MIRAGE estimates residual structural uncertainty and makes a selective decision: answer or abstain. The framework also produces lightweight reliability certificates that summarize whether the answer appears stable, grounded, computed, and supported. MIRAGE is complementary to table question-answering systems. Prior table reasoning work primarily improves answer generation through better semantic parsers, table encoders, reasoning chains, tools, or retrieval mechanisms (Herzig et al., 2020; Liu et al., 2022; Wu et al., 2025; Zhang et al., 2024; Tian et al., 2026b). In contrast, MIRAGE operates after an answer has been produced and asks whether that answer is reliable enough to return. This reliability layer is motivated by a common table-reasoning failure mode: a reasoner may produce a stable answer even when its entity grounding, schema binding, computation, or evidence support remains uncertain.

We evaluate MIRAGE primarily on controlled diagnostic stress tests that isolate structural table-reasoning failures, and use natural-table and cross-backbone experiments as transfer evidence (Zhang et al., 2025b; Liu et al., 2025). MIRAGE substantially reduces selective risk in the matched diagnostic setting, while transfer gains are smaller and metric-dependent.

Our contributions are as follows:

1. We introduce residual structural uncertainty as an abstraction for structure-grounded reliability in table reasoning, organized around whether an answer is stable, grounded, computed, and supported.
2. We propose MIRAGE, a reliability framework that estimates answer trustworthiness from test-time trace signals and produces calibrated answer-or-abstain decisions with lightweight reliability certificates.
3. Extensive experiments show that structure-grounded reliability signals substantially reduce selective risk on controlled structural stress tests.

2 Related Work

Table question answering and table reasoning. Table question answering has been studied through semantic parsing, neural table encoders, and pretrained table-text models (Tian et al., 2026a). WikiTableQuestions introduced compositional question answering over semi-structured tables (Pasupat and Liang, 2015). Later systems improved table understanding through table-specific pretraining and neural execution, including TaBERT (Yin et al., 2020), TAPAS (Herzig et al., 2020), and TAPEX (Liu et al., 2022). This line of work primarily targets answer accuracy. Our work instead asks whether a produced answer is reliable enough to return, especially when structural ambiguities remain unresolved.

Uncertainty estimation for language models. Neural confidence scores are often miscalibrated even when predictive accuracy is high (Guo et al., 2017). For language models, reliability estimation has been studied through model probabilities, self-evaluation, and elicited confidence (Jiang et al., 2021; Kadavath et al., 2022; Tian et al., 2023). These methods are general-purpose and largely operate over answer likelihoods or verbalized beliefs. We study a setting where the answer distribution can be stable even when the table reasoning process is structurally unsupported.

Semantic uncertainty and hallucination detection. Sampling-based methods estimate reliability from agreement among generations. Self-consistency improves reasoning by marginalizing over multiple reasoning paths (Wang et al., 2023). Semantic Entropy clusters generations by semantic equivalence and measures uncertainty over meanings (Kuhn et al., 2023). SelfCheckGPT detects hallucinations by comparing sampled passages for consistency (Manakul et al., 2023). These methods are strong generic baselines, but they are not designed to verify whether a table answer has satisfied entity, schema, operation, evidence, and dependency obligations. Our comparison asks whether answer-distribution signals are sufficient under structural table failures, not whether these methods are universally weak.

Selective prediction and abstention calibration. Selective prediction studies when a model should abstain rather than answer (Geifman and El-Yaniv, 2019). Conformal prediction motivates calibration procedures based on held-out nonconformity

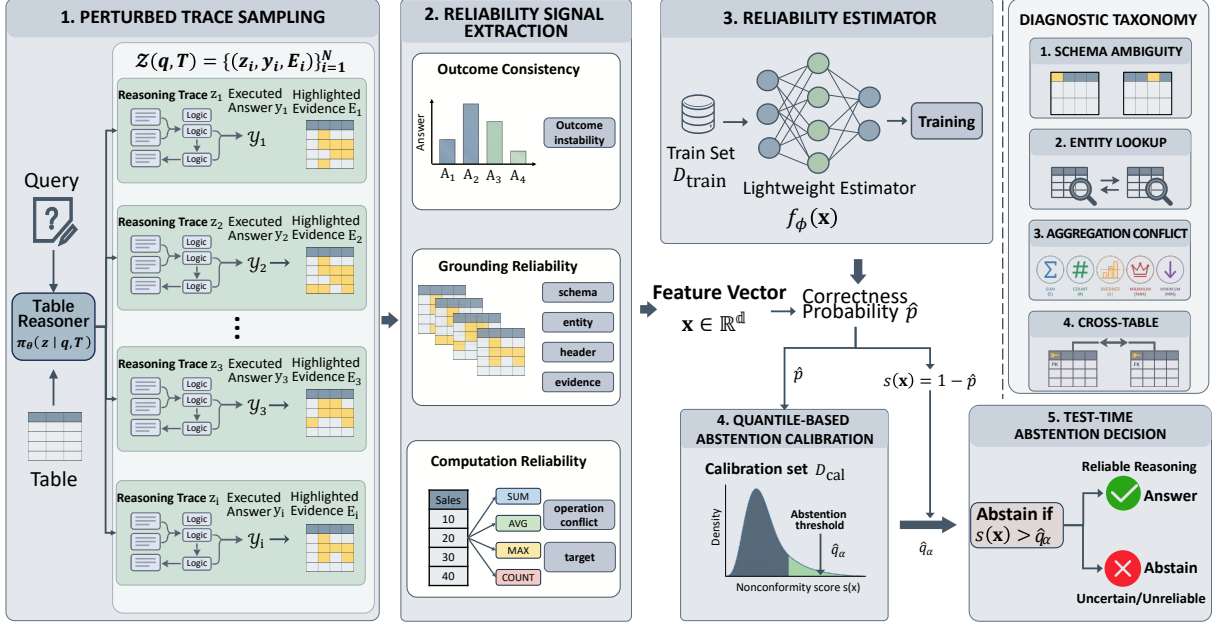


Figure 1: Overview of MIRAGE. A base table reasoner produces sampled or perturbed traces for a question–table pair. MIRAGE extracts reliability signals from trace behavior, including outcome consistency, grounding uncertainty, and computation uncertainty. These signals are used to estimate correctness, apply quantile-based abstention calibration, and make a selective answer-or-abstain decision with a lightweight reliability certificate.

scores (Vovk et al., 2005; Angelopoulos and Bates, 2023). We use a calibration-set quantile over a learned structural reliability score to choose a data-dependent abstention threshold. This step selects an operating point; it does not control the conditional error rate among accepted answers.

3 Method

MIRAGE is a reliability layer for table reasoning systems that operate over structured evidence. Given a question, a table context, and a candidate answer produced by a base reasoner, MIRAGE estimates whether the answer is reliable enough to return. It assumes access to observable test-time signals such as sampled or perturbed reasoning traces, candidate answers, evidence regions, schema bindings, entity candidates, and operation choices. From these signals, MIRAGE estimates *residual structural uncertainty*: uncertainty that remains after an answer has been produced because the answer may still be incorrectly grounded or computed. As shown in Figure 1, a base table reasoner first produces trace-level observations for a question–table pair. MIRAGE then converts these observations into reliability signals over outcome consistency, grounding, and computation (Section 3.3). A lightweight reliability head estimates the probability that the selected answer is correct, and

quantile-based abstention calibration converts this score into a selective answer-or-abstain decision with a lightweight reliability certificate.

3.1 Problem Formulation

Let q be a natural-language question and T a table context. A table reasoning system induces a distribution over executable or semi-executable reasoning traces:

$$z \sim \pi_\theta(z \mid q, T). \quad (1)$$

Each trace z yields an answer y , an evidence region E , and optional metadata m such as selected columns, linked entities, and operation choices:

$$(y, E, m) = \text{Exec}(z, T). \quad (2)$$

At test time, we observe a set of sampled or perturbed trajectories:

$$\mathcal{Z}(q, T) = \{(z_i, y_i, E_i, m_i)\}_{i=1}^N. \quad (3)$$

The goal is not to generate a better answer directly. Instead, given the answer selected by the underlying reasoner, MIRAGE estimates whether the answer is reliable enough to return. Formally, it learns a selective decision function:

$$g(q, T, \mathcal{Z}) \in \{\text{ANSWER}, \text{ABSTAIN}\}. \quad (4)$$

3.2 Perturbed Trace Sampling

The first stage collects test-time observations from the base table reasoner. Given (q, T) , we run the reasoner multiple times under sampling, perturbation, or serialization variation to obtain $\mathcal{Z}(q, T)$. These traces expose how the reasoner arrives at candidate answers: which answers it produces, which table regions it highlights as evidence, which entities or columns it binds to, and which operations it applies.

This stage does not require model parameters, hidden states, or unrestricted chain-of-thought. Full MIRAGE requires compact structured records that are either exposed natively by the reasoner or elicited in a fixed output format. These records include candidate answers, evidence regions, schema bindings, entity candidates, and operation choices. Answer-only interfaces can use only the outcome-level signals.

3.3 Reliability Signal Extraction

From $\mathcal{Z}(q, T)$, MIRAGE constructs a reliability feature vector $\mathbf{x} \in \mathbb{R}^d$. These features do not use gold labels, oracle answerability, oracle evidence masks, or hidden generator state. They are computed from observable trace behavior.

Outcome consistency. MIRAGE does not discard answer-distribution uncertainty. It includes outcome-level signals that measure whether sampled traces agree on the final answer. Let $P(y)$ be the empirical distribution over executed answers. We define execution entropy:

$$H_{\text{exec}} = - \sum_y P(y) \log P(y). \quad (5)$$

High entropy indicates disagreement over executed outcomes. We also compute an answer vote-margin feature A_{margin} from the gap between the top two answer masses. These features capture whether the final outcome is stable, but they do not by themselves determine whether the stable answer is structurally reliable.

Grounding reliability. Grounding signals measure whether traces consistently identify the relevant rows, entities, columns, fields, and supporting cells. Under schema or serialization perturbations $\{\tilde{T}_k\}_{k=1}^K$, we compute schema-induced an-

swer volatility:

$$V_{\text{schema}} = 1 - \frac{1}{K} \sum_{k=1}^K \mathbb{I}(\hat{y}(q, \tilde{T}_k) = \hat{y}(q, T)). \quad (6)$$

High volatility indicates that the selected answer is sensitive to schema presentation. We also include header-collision and schema-target disagreement features, denoted C_{header} and D_{ambig} , based on competing column bindings. Entity-resolution signals measure lookup quality, including a status score E_{status} and a top-1/top-2 entity-candidate gap E_{gap} .

We further measure evidence alignment across traces. Represent each evidence region E_i as a binary mask $\mathbf{m}_i$ over table cells. Evidence stability is:

$$S_{\text{evid}} = \frac{2}{N(N-1)} \sum_{i < j} \frac{|\mathbf{m}_i \cap \mathbf{m}_j|}{|\mathbf{m}_i \cup \mathbf{m}_j|}. \quad (7)$$

Low overlap indicates that traces rely on inconsistent table regions, even when their final answers agree.

Computation reliability. Computation signals measure whether traces agree on the operation required to answer the question. For aggregation, comparison, filtering, and ranking questions, we compute operation conflict features. These include disagreement over candidate operation types and disagreement over operation targets, denoted C_{op} and D_{op} . High disagreement indicates that traces may have grounded similar table regions but applied different reasoning operations.

The exact thresholds used for header similarity, operation consensus, and entity alignment are implementation hyperparameters. The framework does not depend on any single threshold; it depends on whether observable trace behavior exposes residual structural uncertainty.

3.4 Reliability Estimator

The third stage maps the reliability feature vector to a correctness estimate. We train a lightweight estimator:

$$f_{\phi}(\mathbf{x}) \rightarrow \hat{p}, \quad (8)$$

where $\hat{p}$ approximates:

$$\Pr(\text{answer is correct} \mid \mathbf{x}). \quad (9)$$

The estimator is intentionally simple: the contribution is not a new neural architecture, but a reliability interface that combines outcome consistency with structure-grounded signals.

The same signals also produce a lightweight reliability certificate:

$$c = \{c_{\text{outcome}}, c_{\text{grounding}}, c_{\text{computation}}\}. \quad (10)$$

Each component summarizes whether the corresponding aspect appears stable, resolved, or uncertain. For more detailed error analysis, the grounding component can be decomposed into entity, schema, header, and evidence-alignment summaries, while the computation component can be decomposed into operation and target summaries. These certificates are diagnostic summaries for selective prediction and error analysis, not formal proofs of correctness.

3.5 Quantile-Based Abstention Calibration

The fourth stage converts the reliability score into a calibrated threshold for selective prediction. Let the nonconformity score be:

$$s(\mathbf{x}) = 1 - \hat{p}. \quad (11)$$

On a calibration set $\mathcal{D}_{\text{cal}}$, we compute:

$$\hat{q}_\alpha = \text{Quantile}_{1-\alpha} \left(\{s(\mathbf{x}_j)\}_{(\mathbf{x}_j, y_j) \in \mathcal{D}_{\text{cal}}} \right). \quad (12)$$

Under exchangeability between calibration and test data, this quantile rule calibrates the marginal acceptance event induced by the score distribution. It does not guarantee semantic correctness or bound the conditional error (selective risk) among accepted answers; direct risk control requires a different calibration objective.

3.6 Test-Time Abstention Decision

At test time, MIRAGE applies the calibrated threshold to the selected answer. It abstains when:

$$s(\mathbf{x}) > \hat{q}_\alpha. \quad (13)$$

Equivalently, it returns the selected answer only when:

$$s(\mathbf{x}) \leq \hat{q}_\alpha. \quad (14)$$

The output is therefore either an accepted answer $\hat{y}$ or an abstention, together with the lightweight reliability certificate described above. The certificate is intended to indicate which aspect of the trace behavior appears stable, resolved, or uncertain; it should not be interpreted as a formal proof of correctness.

4 Diagnostic Benchmark

We use a controlled diagnostic benchmark to test whether uncertainty estimators respond to structural table-reasoning failures. The benchmark is not intended as a replacement for natural table QA datasets. Its purpose is causal diagnosis: it isolates failure regimes in which answer-distribution confidence may be misleading.

The benchmark taxonomies instantiate common ways in which the broader structural obligations from Section 3 can fail. They are used for evaluation and stress testing; they are not part of the MIRAGE framework itself, nor are they intended to exhaust the space of possible structural failures.

Failure families. Each instance belongs to one of four diagnostic failure families:

- schema ambiguity;
- entity lookup ambiguity;
- aggregation conflict;
- missing cross-table evidence.

These families correspond to different manifestations of residual structural uncertainty. Schema ambiguity and entity lookup ambiguity primarily test grounding uncertainty. Aggregation conflict primarily tests computation uncertainty, often with additional schema ambiguity. Missing cross-table evidence primarily tests support incompleteness through absent dependencies or unresolved joins. For every example, we retain multi-path executed outcomes, evidence masks, perturbation responses, and trace metadata. This makes it possible to analyze reliability at the level of both final-answer behavior and structural reasoning behavior.

Label balance and anti-shortcut design. A central design requirement is to avoid trivial separability. The benchmark contains 2,000 examples, balanced across the four failure families and ANSWER/ABSTAIN labels, with 250 examples in each family-label cell. The generator is hardened so that no single failure family directly maps to answer or abstain decisions. The benchmark therefore tests whether a reliability method detects unresolved structural uncertainty, not whether it memorizes taxonomy labels.

Construction and label validation. Latent table structures, executable queries, answerability

states, and gold answers are generated deterministically in Python; Qwen3-8B is used only to lexicalize surface forms. In a blinded audit of 200 examples across the four failure families, two annotators achieved 96% agreement ($\kappa = 0.92$); after adjudication, consensus matched the programmatic labels on all 200 examples. Additional validation is reported in Appendix F.

5 Experiments

5.1 Experimental Setup

Evaluation sets. Our main experiments use the controlled diagnostic benchmark described in Section 4. This benchmark is designed to isolate structural table-reasoning failures and contains four diagnostic taxonomies: schema ambiguity, entity lookup ambiguity, aggregation conflict, and missing cross-table evidence. Unless otherwise stated, results are reported on the harder diagnostic setting with $N_{\text{trace}} = 7$ sampled traces per example. We also evaluate transfer on WikiTableQuestions (Pasupat and Liang, 2015) and FinQA (Chen et al., 2021), and instantiate the controlled diagnostic pipeline with Llama-3.1-8B-Instruct as a second reasoning backbone. Qwen3-8B is the default backbone; within each setting, all compared reliability methods use the same backbone to isolate reliability estimation from answer-generation capability.

Protocol and metrics. Primary controlled results are reported as mean $\pm$ standard deviation over five group-stratified rotations of group-disjoint folds. Groups sharing latent-table or generative-template provenance are kept within the same fold; each rotation uses three folds for training, one for calibration, and one for held-out testing, preserving the 1,200/400/400 scale. All compared methods use identical folds. WikiTableQuestions and FinQA statistics are likewise computed over five disjoint table groups. Risk@50% is recomputed independently within each held-out group. We also report AUROC, AUPRC, ECE, selective accuracy, coverage, and unsafe answer rate.

Baselines. We compare MIRAGE with representative uncertainty, confidence, conformal, and hallucination-detection baselines. Confidence baselines include MaxProb, predictive entropy, and temperature-scaled MaxProb, following standard calibration practice (Guo et al., 2017). Sampling-based baselines include self-

consistency, which uses agreement among sampled reasoning paths (Wang et al., 2023). Self-assessment baselines include $P(\text{True})$ and verbalized confidence, which ask the model to assess whether its answer is likely to be correct (Kadavath et al., 2022; Tian et al., 2023). Semantic uncertainty baselines include Semantic Entropy and Evidential Semantic Entropy, which estimate uncertainty over semantic answer clusters rather than surface forms (Kuhn et al., 2023; Farquhar et al., 2024; Kunitomo-Jacquin et al., 2026). Hallucination-detection baselines include SelfCheckGPT, which scores consistency across sampled generations (Manakul et al., 2023). Conformal baselines include Conformal-only, ConU, and SConU, which calibrate generic uncertainty scores or selective decisions using conformal prediction (Wang et al., 2024, 2025). These baselines primarily rely on answer likelihood, answer agreement, semantic dispersion, or sample consistency, rather than table-specific grounding or computation signals.

We organize the experiments as a diagnostic evidence chain. First, we test whether answer-distribution uncertainty detects structural failures. Second, we test whether trace-level reliability signals reduce selective risk. Third, we analyze which signal groups matter. Finally, we examine robustness, taxonomy transfer, and real-table transfer.

5.2 RQ1: Do Answer-Distribution Signals Detect Structural Failures?

Table 1 compares seven uncertainty and hallucination-detection baselines on the harder diagnostic benchmark. All methods are evaluated on the same five group-disjoint rotations, with identical score orientation and calibration protocols where applicable. Self-consistency obtains AUROC 0.55 ± 0.04 , Semantic Entropy 0.58 ± 0.03 , and SelfCheckGPT 0.63 ± 0.03 . Other confidence, self-assessment, and semantic-uncertainty baselines show similarly weak-to-moderate discrimination. These methods primarily estimate reliability from confidence or answer-distribution behavior, which can remain stable even when the underlying table reasoning is structurally unsupported. Consistent with this failure mode, trace agreement differs by only 0.01–0.06 between answerable and abstain cases in three of the four diagnostic families (Appendix E). Thus, answer agreement alone does not reliably distinguish

Method	AUROC $\uparrow$	ECE $\downarrow$	Sel. Acc $\uparrow$	Risk@50% $\downarrow$
Self-consistency	0.55 $\pm$ 0.04	0.19 $\pm$ 0.03	0.54 $\pm$ 0.04	0.46 $\pm$ 0.03
$P(\text{True})$	0.57 $\pm$ 0.03	0.23 $\pm$ 0.04	0.52 $\pm$ 0.03	0.45 $\pm$ 0.03
Verbalized confidence	0.62 $\pm$ 0.04	0.16 $\pm$ 0.03	0.51 $\pm$ 0.04	0.43 $\pm$ 0.03
SConU	0.55 $\pm$ 0.05	0.20 $\pm$ 0.03	0.54 $\pm$ 0.04	0.46 $\pm$ 0.04
Evidential SE	0.50 $\pm$ 0.04	0.33 $\pm$ 0.05	0.52 $\pm$ 0.03	0.48 $\pm$ 0.04
Semantic Entropy	0.58 $\pm$ 0.03	0.33 $\pm$ 0.04	0.51 $\pm$ 0.04	0.43 $\pm$ 0.03
SelfCheckGPT	0.63 $\pm$ 0.03	0.17 $\pm$ 0.03	0.52 $\pm$ 0.03	0.43 $\pm$ 0.03
MIRAGE	0.94$\pm$0.02	0.08$\pm$0.02	0.90$\pm$0.03	0.16$\pm$0.02

Table 1: Performance on the controlled diagnostic benchmark. Values are mean $\pm$ standard deviation over five group-disjoint evaluation rotations.

Method	AUROC $\uparrow$	ECE $\downarrow$	Sel. Acc $\uparrow$	Risk@50% $\downarrow$
MaxProb	0.38 $\pm$ 0.03	0.30 $\pm$ 0.04	0.50 $\pm$ 0.02	0.50 $\pm$ 0.03
Entropy	0.38 $\pm$ 0.04	0.30 $\pm$ 0.03	0.50 $\pm$ 0.02	0.50 $\pm$ 0.03
TS+MaxProb	0.38 $\pm$ 0.03	0.16 $\pm$ 0.03	0.50 $\pm$ 0.02	0.50 $\pm$ 0.03
Conformal-only	0.38 $\pm$ 0.04	0.30 $\pm$ 0.04	0.50 $\pm$ 0.02	0.50 $\pm$ 0.04
MIRAGE	0.94$\pm$0.02	0.08$\pm$0.02	0.90$\pm$0.03	0.16$\pm$0.02

Table 2: Comparison with simple confidence baselines on the harder diagnostic benchmark. Values are mean $\pm$ standard deviation over five group-disjoint rotations.

structurally supported from unsupported answers in the represented diagnostic settings.

5.3 RQ2: Do Trace-Level Reliability Signals Reduce Selective Risk?

Across the five group-disjoint rotations, MIRAGE achieves AUROC 0.94 ± 0.02 , ECE 0.08 ± 0.02 , and selective accuracy 0.90 ± 0.03 at its calibrated operating point (approximately 49% coverage). At the same fixed 50% coverage for every method, MIRAGE achieves Risk@50% 0.16 ± 0.02 , compared with 0.43–0.48 for the evaluated answer-distribution baselines. Figure 2 shows that MIRAGE maintains lower risk across coverage levels, while answer-distribution baselines cannot effectively trade coverage for reliability.

The strong in-taxonomy diagnostic score should not be interpreted as evidence that table reliability is solved. Rather, it shows that trace-level reliability signals are highly diagnostic when calibration and test examples share the same controlled failure families. Later transfer experiments show that the same signals do not fully generalize to unseen taxonomies or natural tables.

5.4 RQ3: Which Reliability Signals Explain the Gains?

Table 3 examines which parts of the MIRAGE reliability interface drive the selective-prediction gains. The largest degradation comes from removing schema-related grounding signals, suggesting that many diagnostic failures arise from unresolved schema or column binding rather than

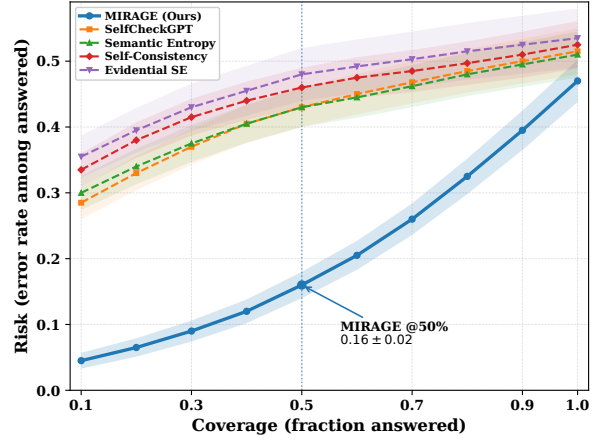


Figure 2: Mean risk–coverage curves over five group-disjoint diagnostic folds. Shaded bands indicate $\pm$ one standard deviation across folds.

from answer instability alone. Entity and operation signals also contribute substantially, showing that reliable table reasoning depends on checking both whether the model has identified the right table entity and whether it has applied the intended computation. Dependency and evidence-stability signals have smaller marginal effects in this matched setting, but this should be interpreted as a property of the benchmark mixture rather than evidence that support signals are generally unnecessary. The answer-distribution-only variant further shows that outcome consistency is useful but insufficient without structure-grounded signals.

5.5 RQ4: Are the Reliability Signals Robust or Taxonomy-Specific?

Taxonomy-held-out transfer. When the target taxonomy is excluded from both training and calibration, Table 5 shows that transfer is uneven across failure families. Schema ambiguity and missing cross-table evidence retain partial transfer, likely because they share observable symptoms with other structural failures, such as unstable bindings or incomplete support. Entity lookup and aggregation transfer less reliably, indicating that some cues are more failure-family-specific. The higher coverage and selective risk under the transferred threshold further suggest that calibration becomes over-permissive when the target failure family is absent from calibration.

Shortcut audits. We additionally test whether the retained features encode trivial label or split information. Shuffled-label training collapses to AUROC 0.54, while a split-leakage classifier ob-

Variant	AUROC $\uparrow$	ECE $\downarrow$	Risk@50% $\downarrow$	Sel. Acc $\uparrow$	Coverage
MIRAGE full	0.94	0.08	0.15	0.91	0.49
w/o schema signals	0.79 $\downarrow 16.0\%$	0.14 $\uparrow 75.0\%$	0.31 $\uparrow 106.7\%$	0.78 $\downarrow 14.3\%$	0.65 $\uparrow 32.7\%$
w/o entity signals	0.87 $\downarrow 7.4\%$	0.11 $\uparrow 37.5\%$	0.22 $\uparrow 46.7\%$	0.84 $\downarrow 7.7\%$	0.58 $\uparrow 18.4\%$
w/o operation signals	0.90 $\downarrow 4.3\%$	0.10 $\uparrow 25.0\%$	0.19 $\uparrow 26.7\%$	0.87 $\downarrow 4.4\%$	0.55 $\uparrow 12.2\%$
w/o dependency signals	0.93 $\downarrow 1.1\%$	0.09 $\uparrow 12.5\%$	0.16 $\uparrow 6.7\%$	0.90 $\downarrow 1.1\%$	0.48 $\downarrow 2.0\%$
w/o evidence stability	0.93 $\downarrow 1.1\%$	0.08 $\pm 0.0\%$	0.15 $\pm 0.0\%$	0.90 $\downarrow 1.1\%$	0.49 $\pm 0.0\%$
answer-distribution only	0.55 $\downarrow 41.5\%$	0.19 $\uparrow 137.5\%$	0.46 $\uparrow 206.7\%$	0.54 $\downarrow 40.7\%$	0.91 $\uparrow 85.7\%$

Table 3: Ablation by reliability-signal group. Coverage is the answered fraction under the corresponding decision rule. Risk@50% is computed by ranking examples by reliability and answering the top 50%.

Taxonomy	n	AUROC $\uparrow$	ECE $\downarrow$	Sel. Acc $\uparrow$
Aggregation conflict	100	0.96	0.06	0.93
Missing cross-table evidence	100	0.95	0.07	0.92
Schema ambiguity	100	0.93	0.09	0.89
Entity lookup	100	0.94	0.08	0.90

Table 4: Per-taxonomy diagnostic performance. These in-taxonomy results are strong, but they do not imply transfer to unseen failure families.

Held-out taxonomy	AUROC $\uparrow$	Coverage	Sel. Risk $\downarrow$	n
Schema ambiguity	0.72	0.74	0.36	500
Entity lookup	0.53	0.66	0.45	500
Aggregation conflict	0.62	0.79	0.41	500
Missing cross-table evidence	0.78	0.68	0.31	500

Table 5: Taxonomy-held-out transfer. Performance drops when the target failure family is excluded from training and calibration, revealing limited structural transfer across diagnostic taxonomies.

tains AUROC 0.50. We also removed two candidate features that directly revealed answerability before running the reported experiments. These checks suggest that the diagnostic gains are not explained by direct label leakage or split artifacts (Appendix H).

Lexicalization robustness. To test whether the diagnostic gains depend on Qwen-specific surface-generation artifacts, we independently lexicalize 500 diagnostic examples using Llama-3.1-8B-Instruct while keeping the latent table specifications and executable programs unchanged. The original MIRAGE estimator is applied without retraining or recalibration, and results are evaluated over five latent-specification-disjoint groups. As shown in Table 6, MIRAGE retains higher AUROC (0.76 ± 0.03 vs. 0.62 ± 0.04) and lower Risk@50% (0.38 ± 0.03 vs. 0.45 ± 0.04) than SelfCheckGPT. This experiment does not remove the intentional alignment between the diagnostic benchmark and MIRAGE’s structural features, but indicates that the observed advantage is not ex-

Method	AUROC $\uparrow$	Risk@50% $\downarrow$
SelfCheckGPT	0.62 ± 0.04	0.45 ± 0.04
MIRAGE	0.76 ± 0.03	0.38 ± 0.03

Table 6: Independent lexicalization of 500 diagnostic examples using Llama-3.1-8B-Instruct. Values are mean $\pm$ standard deviation over five latent-specification-disjoint groups; MIRAGE is applied without retraining or recalibration.

plained solely by Qwen-specific lexicalization artifacts.

5.6 RQ5: How Does MIRAGE Transfer Beyond the Matched Setting?

We first evaluate transfer to WikiTableQuestions, where correctness is measured against gold answers and no explicit structural-failure annotations are available. As shown in Table 7, the results are metric-dependent. MIRAGE does not outperform SelfCheckGPT in global AUROC (0.61 ± 0.02 vs. 0.63 ± 0.03), indicating that its overall ranking of correct versus incorrect answers is not better in this heterogeneous real-table setting. However, at the same fixed 50% coverage, MIRAGE achieves a lower Risk@50% (0.34 ± 0.03 vs. 0.42 ± 0.04). Because the comparison fixes the answered fraction, this reduction cannot be attributed simply to MIRAGE abstaining on more examples. We therefore interpret the WikiTableQuestions result as evidence that structure-grounded signals can identify a safer high-reliability subset, rather than as evidence of superior global discrimination on natural tables.

We further evaluate transfer beyond the original dataset and reasoning backbone. On FinQA with Qwen3-8B, MIRAGE obtains AUROC 0.69 ± 0.03 and Risk@50% 0.32 ± 0.03 , compared with 0.65 ± 0.04 and 0.38 ± 0.04 for SelfCheckGPT. On the controlled diagnostic benchmark with Llama-3.1-8B-Instruct as the reasoning back-

Setting	Method	AUROC $\uparrow$	Risk@50% $\downarrow$
WTQ / Qwen3-8B	SelfCheckGPT	0.63$\pm$0.03	0.42 $\pm$ 0.04
	MIRAGE	0.61 $\pm$ 0.02	0.34$\pm$0.03
FinQA / Qwen3-8B	SelfCheckGPT	0.65 $\pm$ 0.04	0.38 $\pm$ 0.04
	MIRAGE	0.69$\pm$0.03	0.32$\pm$0.03
Diagnostic / Llama-3.1-8B	SelfCheckGPT	0.74 $\pm$ 0.03	0.34 $\pm$ 0.04
	MIRAGE	0.77$\pm$0.02	0.29$\pm$0.03

Table 7: Transfer beyond the matched Qwen diagnostic setting. WikiTableQuestions and FinQA results are reported as mean $\pm$ standard deviation over five disjoint table groups; the Llama diagnostic results use five group-disjoint train/calibration/test rotations.

bone, MIRAGE obtains AUROC 0.77 ± 0.02 and Risk@50% 0.29 ± 0.03 , compared with 0.74 ± 0.03 and 0.34 ± 0.04 for SelfCheckGPT. These gains are substantially smaller than those in the matched Qwen diagnostic setting and should therefore be viewed as transfer evidence within the evaluated datasets and backbones, rather than as establishing universal AUROC superiority.

6 Conclusion

Table reasoning exposes a reliability failure mode that answer-distribution uncertainty alone cannot fully capture. A model may produce answers that appear stable across samples while still relying on incorrect entity grounding, ambiguous schema bindings, unsupported evidence, or inconsistent computations. We introduced MIRAGE, a structure-grounded reliability framework that estimates residual structural uncertainty from observable trace-level signals and produces calibrated answer-or-abstain decisions with lightweight reliability certificates. Experiments show that structure-grounded reliability signals substantially reduce selective risk under the controlled structural failure families studied here. Transfer gains on natural tables and a second reasoning backbone are smaller and metric-dependent, providing evidence beyond the matched setting without establishing universal superiority.

Future work should investigate more transferable structural reliability signals, including signals that generalize across heterogeneous table schemas, unseen failure families, and real-world reasoning environments. Another important direction is integrating reliability estimation more tightly with table reasoners, enabling systems not only to abstain under unresolved structural uncertainty, but also to repair or verify unsupported reasoning steps before producing a final answer.

Limitations

The diagnostic benchmark is deliberately aligned with the structural failure families targeted by MIRAGE. It is designed to test whether grounding, computation, and support signals detect stable but structurally unsupported answers under controlled conditions, rather than to represent all table-reasoning failures. Human auditing and independent lexicalization reduce labeling and surface-generation concerns, but do not remove this intentional alignment.

Transfer remains substantially weaker than performance in the matched diagnostic setting. On WikiTableQuestions, MIRAGE improves Risk@50% but does not outperform SelfCheckGPT in AUROC; gains on FinQA and with a second reasoning backbone are also smaller. The current results therefore support transfer only to the evaluated datasets and models, not universal table-reliability claims.

Full MIRAGE requires native or elicitable structured traces, including schema fields, entity candidates, operation or tool choices, and supporting rows or cells. Such information is available from many executable or tool-using reasoners and can be elicited in fixed-format outputs, but final-answer-only interfaces can use only the outcome-level subset of the method.

Finally, the approximately 49% coverage reported for the diagnostic setting is one data-dependent operating point rather than an intrinsic coverage limit. Different thresholds trade coverage against selective risk. Our quantile-based calibration selects such an operating point but does not guarantee semantic correctness or bound the conditional error rate among accepted answers. Performance and calibration may also degrade under distribution shift.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 72601425 and 62302503, Innovation Research Foundation of NUDT No. ZK23-15, the Open Research Fund from State Key Laboratory of High Performance Computing of China Grant No. 202401-09, and Young Elite Scientists Sponsorship Program by CAST No. YESS20240767.

References

- Anastasios N Angelopoulos and Stephen Bates. 2023. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4320–4333.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. volume 9, pages 962–977.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.
- Lucie Kunitomo-Jacquín, Edison Marrese-Taylor, Ken Fukuda, and Masahiro Hamasaki. 2026. Evidential semantic entropy for llm uncertainty quantification. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7107–7122.
- Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. Tapex: Table pre-training via learning a neural sql executor. In *International Conference on Learning Representations*.
- Shixuan Liu, Changjun Fan, Kewei Cheng, Yunfei Wang, Peng Cui, Yizhou Sun, and Zhong Liu. 2024a. Inductive meta-path learning for schema-complex heterogeneous information networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10196–10209.
- Shixuan Liu, Yue He, Haotian Wang, Wenjing Yang, Yunfei Wang, Peng Cui, and Zhong Liu. 2025. Environment inference for learning generalizable dynamical system. *Advances in Neural Information Processing Systems*, 38:164076–164105.
- Tianyang Liu, Fei Wang, and Muhao Chen. 2024b. Rethinking tabular data understanding with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 450–482.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.
- Panupong Pasupat and Percy Liang. 2015. [Compositional semantic parsing on semi-structured tables](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1470–1480. Association for Computational Linguistics.
- Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. 2024. Selective selective prediction: Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12935–12948.
- Jinlong Tian, Shixuan Liu, Ruochun Jin, Mengmeng Li, Yanfang Zhou, Xinhai Xu, and Yuhua Tang. 2026a. Efficient table embeddings via self-supervised structural-semantic graph autoencoder. *Information Processing & Management*, 63(1):104298.
- Jinlong Tian, Jiang Yu, Kewei Cheng, Fengxiang Cheng, Yue He, Yunfei Wang, Haotian Wang, Haoxuan Li, Wenjing Yang, and Shixuan Liu. 2026b. Logicsage: Neuro-symbolic reasoning with socratic-guided enhancement. In *Forty-third International Conference on Machine Learning*.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. pages 5433–5442.

Vladimir Vovk, Alexander Gammernan, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.

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

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6886–6898.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. Sconu: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19052–19075.

Jian Wu, Linyi Yang, Dongyuan Li, Yulian Ji, Manabu Okumura, and Yue Zhang. 2025. Mmq: Evaluating llms with multi-table multi-hop complex questions. In *The thirteenth international conference on learning representations*.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8413–8426.

Xuanliang Zhang, Dingzirui Wang, Keyan Xu, Qingfu Zhu, and Wanxiang Che. 2025a. Rot: Enhancing table reasoning with iterative row-wise traversals. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 559–579.

Yue Zhang, Seiji Maekawa, and Nikita Bhutani. 2025b. Same content, different representations: A controlled study for table qa. *arXiv preprint arXiv:2509.22983*.

Yunjia Zhang, Jordan Henkel, Avriella Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M. Patel. 2024. [ReAcTable: Enhancing ReAct for table question answering](#). *Proceedings of the VLDB Endowment*, 17(8):1981–1994.

A Implementation Details

Evaluation protocol. Primary controlled comparisons use five group-stratified rotations over a fixed set of group-disjoint folds. Groups are defined by shared latent-table and generative-template provenance, and all examples from the

same group remain in one fold. Each rotation uses three folds for training, one for calibration, and one for held-out testing. Metrics are computed independently on each held-out fold and reported as mean $\pm$ standard deviation across the five rotations. All applicable methods are re-fitted or re-calibrated on identical folds. The full MIRAGE estimator uses all retained trace-level reliability signals. Ablations remove individual signal groups to assess their contribution. The default configuration disables an optional partial-join rescue rule so that all main tables use the same inference pipeline; the rule is enabled only in controlled ablations.

Base reasoner and reliability estimator.

Qwen3-8B is the default table-reasoning backbone. We additionally instantiate the complete pipeline with Llama-3.1-8B-Instruct for the cross-backbone transfer evaluation. Within each setting, MIRAGE and all applicable baselines use traces generated by the same backbone. The backbone itself is not fine-tuned as part of MIRAGE; only the lightweight reliability estimator is trained.

The reliability estimator is intentionally lightweight. In the reported experiments, we use a shallow two-layer MLP over the extracted structural reliability features to estimate the probability that a candidate answer is correct. The goal is not to introduce a large reasoning architecture, but to evaluate whether observable trace-level signals contain sufficient information for selective prediction.

Compute and infrastructure. All experiments were conducted on a single NVIDIA A100 GPU. Training the lightweight reliability estimator requires less than 2 GPU hours in the default configuration.

Diagnostic benchmark and trace setting. The primary diagnostic benchmark contains 2000 examples balanced across four failure taxonomies and answer/abstain labels. We use disjoint train, calibration, and test splits of sizes 1200/400/400, stratified jointly by taxonomy and label.

For perturbed trace sampling, the base table reasoner is executed multiple times under controlled perturbations, including stochastic decoding and schema or serialization variation. Unless otherwise stated, the harder diagnostic setting uses $N_{\text{trace}} = 7$ sampled traces per example, perturbation strength 0.55, and overlap rate 0.45. These

Certificate component	Status	Diagnostic interpretation
Schema grounding	Unstable	Different traces bind the question to different columns or schema targets.
Entity resolution	Ambiguous	Top candidate entities receive similar scores, producing low lookup confidence.
Operation agreement	Stable	Most traces agree on the intended aggregation or comparison operation.
Evidence support	Missing	Required join evidence or supporting rows are absent or unresolved.
Final decision	Abstain	Structural obligations remain unresolved despite answer agreement.

Table 8: Example lightweight reliability certificate produced by MIRAGE. The certificate summarizes observable structural reliability signals used for selective prediction.

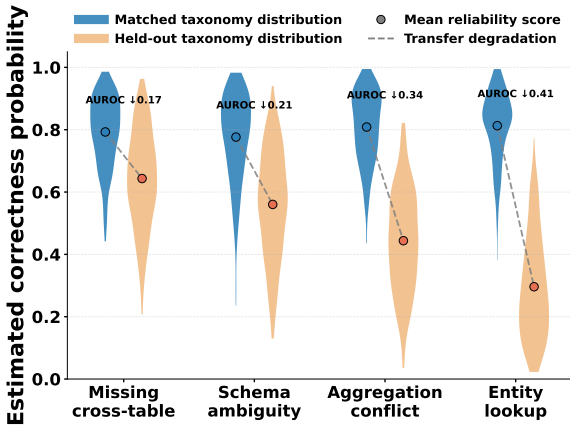


Figure 3: Distributional transfer under taxonomy-held-out evaluation. Dashed connectors show transfer degradation between matched and held-out taxonomies.

values define the diagnostic stress regime used in the experiments and are not required by the general framework.

Although MIRAGE is designed as a reliability layer for safer table reasoning, incorrect reliability estimates may still lead to over-confident acceptance or unnecessary abstention under unseen structural failures or distribution shifts.

B Distributional Transfer Analysis

Figure 3 visualizes the reliability-score distributions under matched-taxonomy and held-out-taxonomy evaluation. Each violin shows the distribution of estimated correctness probabilities produced by the MIRAGE reliability estimator. The matched setting uses train, calibration, and test examples drawn from the same diagnostic failure family, whereas the held-out setting evaluates on a taxonomy excluded from both training and calibration.

Several patterns are visible. First, held-out

transfer consistently shifts the reliability distribution downward, indicating that the estimator becomes less confident under unseen structural failure families. Second, the magnitude of this collapse differs substantially across taxonomies. Missing cross-table evidence retains partial transfer, while entity lookup exhibits the largest distributional collapse and AUROC degradation. This supports the interpretation that some structural reliability cues generalize across failure families, while others remain strongly taxonomy-specific.

The figure should not be interpreted as showing calibration failure alone. Rather, it visualizes a broader structural transfer gap: reliability signals learned from one family of table failures do not always preserve their ranking behavior under unseen structural perturbations.

C Reliability Certificate Examples

Table 8 illustrates simplified reliability certificates produced by MIRAGE. Each certificate summarizes observable trace-level structural signals rather than hidden model states. The goal is not to explain the full reasoning chain, but to expose whether key structural obligations—such as schema grounding, entity resolution, operation agreement, and evidence support—remain unresolved at decision time.

D Qualitative Case Studies

Table 9 summarizes representative diagnostic failure patterns using simplified pseudo-examples. These cases illustrate why answer-distribution agreement alone is insufficient for trustworthy table reasoning: the final answer can appear stable even when the underlying grounding, aggregation, or support obligations remain unresolved.

Failure type	Example question	Baseline signal	Structural issue detected by MIRAGE	Decision
Entity lookup	“Which Paris entry has the highest population?” where multiple rows correspond to Paris, France and Paris, Texas.	Samples repeatedly return the same answer with high agreement.	Low top-1/top-2 entity gap and unresolved lookup ambiguity indicate unstable entity grounding.	Abstain.
Aggregation conflict	“What is the average sales in Europe?” where the table contains both Sales and Revenue columns.	Low semantic entropy and consistent final answers across traces.	High schema-target disagreement and aggregation conflict reveal inconsistent operation grounding.	Abstain.
Missing cross-table evidence	“Which player scored the most goals after the transfer?” where the transfer table is missing.	Samples hallucinate a consistent answer despite agreement.	Join conflict and missing-evidence mentions indicate incomplete support structure.	Abstain.

Table 9: Representative pseudo-cases illustrating structural uncertainty in table reasoning. MIRAGE abstains when grounding or support obligations remain unresolved, even when answer-distribution baselines report stable outputs.

Feature	Signal type	Description	Test-time observable
H_{exec}	Outcome consistency	Entropy of final answers across sampled traces.	Yes
S_{evid}	Evidence stability	Consistency of retrieved evidence across traces.	Yes
V_{schema}	Schema volatility	Variation in schema bindings across perturbations.	Yes
A_{margin}	Answer margin	Confidence gap between top answer candidates.	Yes
C_{header}	Schema grounding	Header collision or schema-target conflict.	Yes
D_{ambig}	Schema grounding	Degree of schema-target ambiguity across traces.	Yes
C_{agg}	Operation consistency	Aggregation conflict across sampled traces.	Yes
D_{agg}	Operation consistency	Disagreement in selected operation targets.	Yes
D_{cross}	Dependency support	Cross-table dependency instability.	Yes
C_{join}	Dependency support	Missing or conflicting join evidence.	Yes
R_{entity}	Dependency support	Unresolved entity references in dependencies.	Yes
$M_{missing}$	Dependency support	Mentions of missing or unsupported evidence.	Yes
C_{lookup}	Entity resolution	Ambiguity among entity lookup candidates.	Yes
Q_{lookup}	Entity resolution	Quality of lookup resolution across traces.	Yes
E_{status}	Entity resolution	Resolution status of retrieved entities.	Yes
E_{gap}	Entity resolution	Score gap between top entity candidates.	Yes

Table 10: Summary of retained MIRAGE reliability features. All features are computed from observable trace behavior at inference time.

E Trace Agreement Analysis

Table 11 reports average trace agreement, defined as the fraction of traces producing the most common answer. For three of four taxonomies, the answerable-versus-abstain gap is less than 0.06, showing why answer-distribution agreement alone is unreliable for structural failures.

Taxonomy	Answerable	Abstain	Gap
Entity lookup	0.70	0.64	0.06
Schema ambiguity	0.68	0.65	0.03
Aggregation conflict	0.64	0.65	0.01
Missing cross-table evidence	0.31	0.76	0.45

Table 11: Average trace agreement for answerable versus abstain cases. Small gaps in the top three rows show that agreement does not reliably identify structural support.

F Benchmark Construction and Additional Validation

The benchmark generation pipeline separates latent specification from surface realization. Latent table structure, executable queries, answerability states, and gold answers are generated deterministically in Python. Qwen3-8B is used only for surface lexicalization, including entity names, headers, legends, and question wording; model judgments do not determine the benchmark labels.

We blindly audited 200 examples sampled across the four diagnostic families. Two annotators independently assessed answerability and the gold answer, achieving 96% agreement with Cohen’s $\kappa = 0.92$. After adjudication, the consensus labels matched the programmatic labels for all 200 examples.

To test sensitivity to the lexicalization model, we additionally generated 500 evaluation examples with Llama-3.1-8B-Instruct while keeping the latent table specifications and executable programs unchanged. The original MIRAGE estimator was applied without retraining or recalibration. The corresponding results are reported in Table 6 in the main paper.

G Feature Summary

Table 10 summarizes the retained reliability features used by the MIRAGE estimator. All features are observable at inference time and are computed from sampled reasoning traces, schema bindings, operation targets, or dependency states. No feature uses gold labels, oracle answerability, or hidden model activations.

H Feature Legitimacy Audit

We audit all retained features to ensure that none directly encodes gold labels, oracle answerability, oracle evidence masks, or generator hidden states. Two candidate features were removed during audit because they trivially revealed answerability status: $E_{\text{answerable}}$ and $E_{\text{invariant}}$. Each showed nearly perfect label correlation and was therefore excluded from all reported experiments.

As an additional sanity check, shuffled-label training collapses to near-chance performance with AUROC 0.54. A split-leakage classifier achieves AUROC 0.50, suggesting that train, calibration, and test examples are not distinguishable from retained features alone.

I Artifact Usage and Licensing

WikiTableQuestions and FinQA are used under their respective research and distribution terms. All evaluated baselines are used only for research purposes. The controlled diagnostic benchmark introduced in this work contains synthetic diagnostic examples and does not include personally identifying information. The benchmark is intended for research on structural reliability estimation and diagnostic evaluation for table reasoning systems.

J AI Usage Disclosure

AI assistants were used only for language polishing of the manuscript. Qwen3-8B and Llama-3.1-8B-Instruct were additionally used as experimental language models as described in the paper. All

scientific claims, methodological design, experimental validation, and final manuscript decisions were made and verified by the authors.