

How Much Retrieval Depth Does Your Reader Need? Noise Ceilings and Cost-Effectiveness in Table QA

Jheng-Peng Huang^{1,2} Yao-Chung Fan³ Jiun-Long Huang¹

¹National Yang Ming Chiao Tung University, Hsinchu, Taiwan

²Digital Innovation Lab, Chunghwa Telecom Laboratories, Taoyuan, Taiwan

³National Chung Hsing University, Taichung, Taiwan

jphuang.cs13@nycu.edu.tw yfan@nchu.edu.tw jlhuang@nycu.edu.tw

Abstract

Retrieval and reading in Table QA are typically optimized in isolation, leaving no principled way to balance retrieval precision against reader capacity. We ask: *how much retrieval depth does a reader actually need?* Across three benchmarks (E2E-WTQ, OTT-QA, and FeTaQA) and four LLMs, we conduct a controlled single-table stress test that guarantees the inclusion of the gold table while increasing retrieval depth to $K = 50$ using soft or hard distractors. This design isolates reader noise tolerance from retrieval failure. We find that the Noise Ceiling is set jointly by task type, reader, and distractor condition, and is not predictable from reader class alone. With hard negatives, every reader has a Noise Ceiling of $K_c^{\text{hard}} \leq 5$ for long-form answer generation, favoring precision-oriented retrieval. By contrast, several readers remain robust through $K = 50$ on short-answer and multi-hop QA, depending on the distractor condition. A reader-side Cost per Success analysis further shows that the trade-off between robustness and inference cost varies across readers and distractor conditions. Overall, these results establish the reader’s condition-specific Noise Ceiling as a concrete operating target for balancing retrieval depth against precision in controlled single-table QA.

1 Introduction

Consider an analyst who asks “what was last quarter’s revenue in the APAC region?” The answer lives in one table among thousands in a corporate data lake. This is open-domain Table Question Answering (Table QA), a two-stage *retrieve-then-read* problem: a retriever narrows the corpus to a small set of candidate tables, and a Large Language Model (LLM) reader then reasons over them to produce the answer.

However, these two stages are traditionally optimized in isolation: retrieval studies emphasize

retrieving highly relevant evidence (Chen et al., 2021; Hsu et al., 2026; Chou et al., 2026), while many Table QA benchmarks assume that the relevant table is already available to the reader (Pasupat and Liang, 2015; Nan et al., 2022). Yet modern table retrievers, particularly those using synthetic query generation (Liang et al., 2025; Hsu et al., 2026; Chou et al., 2026), achieve high Recall@ K on established tasks. This progress exposes a critical but overlooked architectural dilemma:

- *How much retrieval precision does a downstream reader actually need—does loose, high-recall retrieval (a large K) suffice, or must the pipeline guarantee strict, top-ranked precision (a small K)?*

The answer depends not only on retrieval recall, but also on the reader’s noise tolerance. Returning K candidates sets the signal-to-noise ratio the reader faces: a robust reader may tolerate a large candidate set while still locating the correct evidence, whereas a more sensitive reader may fail even when the evidence is present. We call the maximum tolerable depth the reader’s *Noise Ceiling*, which provides a condition-specific target for the upstream retrieval precision needed to preserve reader performance.

This retriever–reader dependency remains underexplored because the two stages lack a shared evaluation protocol: retrieval metrics such as Hit@ K ignore downstream answer accuracy, while reader benchmarks assume idealized contexts. The result is a *specification gap*: architects lack a principled basis for deciding “how much precision is enough” for a given reader.

To bridge this gap, we run a controlled stress test that separates reader failure from retrieval failure. We keep the gold table in the reader’s context while progressively adding distractors; because the evidence is guaranteed to be present, performance

degradation isolates the reader’s sensitivity to retrieval noise. We focus on single-table retrieval and vary retrieval depth, gold-table position, and distractor similarity. This controlled setting provides an upper-bound analysis of reader robustness and lets us derive condition-specific Noise Ceilings as concrete retrieval specifications.

Specifically, we investigate two research questions:

RQ1 (Noise Ceiling): What determines a reader’s Noise Ceiling—the task type, the reader itself, or distractor hardness—and does gold-table position affect robustness beyond what the ceiling captures?

RQ2 (Cost-Effectiveness): How does reader noise tolerance affect the reader-side cost of obtaining a correct answer as retrieval depth grows? We quantify this trade-off with *Cost per Success (CpS)*, which amortizes reader inference cost over correct answers and excludes retriever/reranker and other deployment costs.

Key Findings Our stress tests yield three findings (summarized in Figure 1):

- **More retrieval can hurt (the Retrieval-Reasoning Paradox).** Beyond a reader’s Noise Ceiling, additional tables can overwhelm the benefit of deeper retrieval. The ceiling is condition-specific: several readers remain stable through $K = 50$ on short-answer and multi-hop tasks, whereas long-form answer generation is substantially more fragile. Gold-table placement can also materially affect performance on complex tasks.
- **Readers fail in distinct ways.** *Distractor Extraction* dominates noise-induced failures across question types and readers, particularly under hard negatives, while residual profiles differ by reader: GPT-OSS-20b shows substantially more *Reasoning Hallucination*, whereas Claude-Haiku-4.5 has the highest *Premature Refusal* share.
- **Cheap tokens are not always cheap correct answers.** Under our reader-side CpS metric, noise-induced failures can erode the nominal price advantage of cheaper readers, making robustness an important component of reader-side inference efficiency.

To support reproducibility, code and artifacts are publicly available.¹

¹<https://github.com/jphuang-cs13/NoiseCeil-TableQA>

2 Related Work

Our work connects two lines of research: retrieval-based Table QA and the robustness and cost of long-context LLM readers.

2.1 Evolution of Table QA: From Parsing to Retrieval

Research on structured-data QA has evolved from semantic parsing and text-to-SQL (Yu et al., 2018; Suhr et al., 2020) and tabular pre-training (Yin et al., 2020; Herzig et al., 2020) toward free-form answer generation over tables (Nan et al., 2022) and retrieval-augmented Table QA (Chen et al., 2021), alongside the broader RAG paradigm (Lewis et al., 2020). Retrieval has likewise progressed from sparse baselines to dense and late-interaction approaches (Karpukhin et al., 2020; Khattab and Zaharia, 2020), while recent systems incorporate external tool use (Schick et al., 2023), including multi-turn agentic retrieval for open-domain Table QA (Qiu, 2025). Complementarily, we study how the depth of retrieved candidate tables affects the downstream reader.

2.2 Reader Robustness & Cost-Effectiveness

Long contexts can make relevant evidence harder to use (Liu et al., 2024), while prior robustness work studies sensitivity to input perturbations (Gan et al., 2021; Pi et al., 2022). More directly, Cucinasu et al. (2024) show that retrieval noise is not uniformly harmful: semantically related distractors can degrade RAG accuracy, whereas random documents may sometimes help. Broader work likewise emphasizes that noise can be harmful or beneficial depending on its form and magnitude (Abdolazimi et al., 2026), and that performance can degrade or saturate as retrieval depth grows (Zou et al., 2026; Zhao et al., 2024).

Our focus differs in three respects. First, we characterize noise specifically in *tabular* reading through distractor-driven cell-level evidence-selection failures, question-type effects, and task-dependent positional sensitivity, rather than treating retrieved context as undifferentiated passages. Second, we operationalize reader tolerance as a condition-specific Noise Ceiling and translate it into an upstream retrieval-depth specification. Third, we introduce Cost per Success (CpS) as a reader-side diagnostic that connects retrieval depth, reasoning robustness, and reader-side inference cost under retrieval noise.

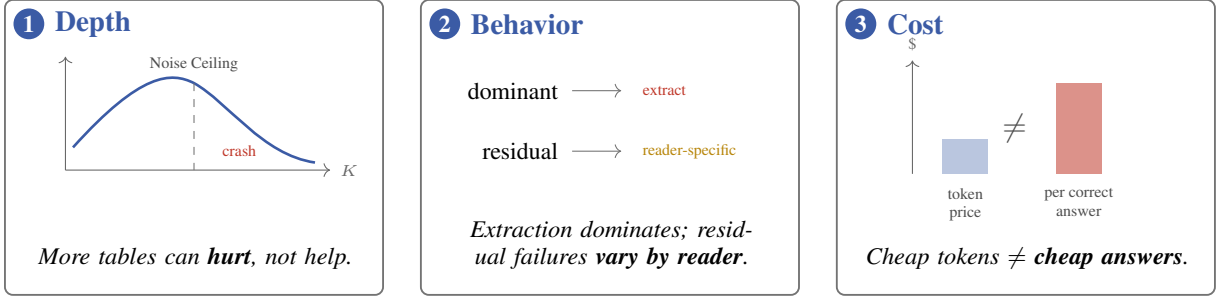


Figure 1: **What to take away.** (1) Beyond a condition-specific *Noise Ceiling*, additional tables can reduce reader accuracy even when the gold table is present; (2) *Distractor Extraction* dominates noise-induced failures, while residual profiles vary by reader; (3) cheap tokens do not necessarily yield cheap correct answers, because noise robustness affects reader-side CpS.

3 Methodology

3.1 Problem Formulation

Given a query and a repository of tables, a single-table Table QA system retrieves an ordered context of K candidate tables, $C = [D_1, \dots, D_K]$, and a reader generates the answer. Our controlled protocol guarantees that the gold table is present ($\text{Hit}@K = 1$), isolating reader noise tolerance from retrieval failure as a function of depth K .

3.2 Experimental Setup

3.2.1 Datasets

We evaluate three benchmarks, each representing a distinct task type: **E2E-WTQ** (Pan et al., 2021) (a short-answer lookup and cell-selection subset of WikiTableQuestions (Pasupat and Liang, 2015)), **OTT-QA** (Chen et al., 2021) (multi-hop open-domain QA over tables and text), and **Fe-TaQA** (Nan et al., 2022) (long-form, multi-fact synthesis). Here, *task type* denotes the dominant answering demand of each benchmark, whereas *question type* denotes the finer-grained query-level categories used in our error analysis, such as *Lookup*.

3.2.2 Models

We evaluate two open-weight models—**Qwen-3-32b** (Qwen Team, 2025) and **GPT-OSS-20b** (OpenAI, 2025)—and two proprietary models—**Claude-Haiku-4.5** (Anthropic, 2025) and **GPT-4o** (OpenAI, 2024).

3.3 Evaluation Metrics

We evaluate reader accuracy ($\text{Score}@K$), noise robustness (NRR; RQ1), and reader-side Cost per Success (CpS; RQ2). Scores are averaged over three independent runs per query.

Score@K. Retrieval-side $\text{Hit}@K$ indicates whether the gold table appears within the top- K results and is fixed to 1 in our controlled setting. Reader-side $\text{Score}@K$ is the mean answer correctness given K contexts. Because strict string matching is brittle to formatting and free-form responses, we use GPT-OSS-20b as a semantic judge (Li et al., 2024). Given the question, reader output, and gold answer, it performs a binary factual-equivalence check and returns only YES or NO. The judge is blind to retrieved tables and answer provenance, runs at temperature 0.0 with a 10-token output cap, and neither ranks candidates nor assigns open-ended quality scores (see Appendix A). We validate these judgments against benchmark-specific metrics and blinded human annotations in Appendix C.

Noise Robustness Ratio (NRR). We define $\text{NRR}@K = \text{Score}@K / \text{Score}@1$ and the *Noise Ceiling* as $K_c = \max\{K \in \mathcal{K} : \text{NRR}@K \geq \tau\}$, where $\mathcal{K} = \{1, 5, 10, 20, 30, 40, 50\}$ and $\tau = 0.9$. We report separate K_c^{hard} and K_c^{soft} values and use $K_c \geq 50$ when the criterion holds at the maximum tested depth. We do not assume monotonic degradation: K_c is the maximum tested depth satisfying the retention criterion.

Cost per Success (CpS). We define $\text{CpS} = \frac{\text{Avg. Input Cost} + \text{Avg. Output Cost}}{\text{Score}@K}$, which amortizes reader-inference cost over successful answers. CpS charges reader-token cost only and excludes retriever/reranker cost, latency, indexing, batching, cache reuse, hardware utilization, and other deployment overheads; it is therefore a reader-side diagnostic rather than an end-to-end cost estimate. For a fixed reader, uniformly scaling both input and output token prices by the same factor rescales CpS without changing the shape of its CpS- K trajectory; absolute cross-reader comparisons remain price-dependent.

3.4 Experimental Design & Injection Protocols

We vary three controlled factors: retrieval depth, gold-table position, and distractor hardness.

3.4.1 Representation & Distractors

Tables are serialized in Markdown (Min et al., 2024) with dynamic row chunking. In the *Controlled* environment, the gold table is combined with either *soft negatives* (random tables) or *hard negatives* (top-ranked distractors from the dense-retrieval mode of BGE-M3 (Chen et al., 2024)). The *Uncontrolled* environment uses the natural Top- K BGE-M3 ranking without gold-table injection.

3.4.2 Controlled Injection Dimensions

For each query, the controlled stress test contains 37 configurations: six depths with $K > 1$ crossed with three gold positions and two noise types ($6 \times 3 \times 2 = 36$), plus one gold-only $K = 1$ baseline. Thus, hard and soft conditions are identical at $K = 1$ by construction.

Dimension 1: Retrieval Depth. We evaluate $K \in \{1, 5, 10, 20, 30, 40, 50\}$ with a single gold table and $K - 1$ distractors. This controlled setting isolates reader-side noise tolerance; multi-table joins, linking, and evidence integration remain outside our scope.

Dimension 2: Gold Position. We place the gold table at $P_{\text{gold}} \in \{\text{Start}, \text{Middle}, \text{End}\}$ to measure positional sensitivity (Liu et al., 2024).

Dimension 3: Noise Intensity. We compare soft and hard negative conditions to vary semantic interference.

4 Experimental Results & Analysis

4.1 The Noise Ceiling

We report separate K_c^{soft} and K_c^{hard} values to capture how the Noise Ceiling varies with distractor hardness, task type, and reader.

The Retrieval–Reasoning Paradox. Figure 2 contrasts reader-side NRR under natural BGE-M3 retrieval (*Uncontrolled*), where retrieval failures and reader robustness jointly determine answer performance, with Controlled hard/soft NRR, where the gold table is guaranteed ($\text{Hit}@K = 1$), allowing performance declines to isolate sensitivity

to added distractors. Under natural retrieval, performance can improve at shallow K ; under controlled retrieval, NRR may instead fall as distractors accumulate. On FeTaQA with hard negatives, Qwen-3-32b drops from $\text{Score}@1 = 0.867 \pm 0.012$ to $\text{Score}@50 = 0.604 \pm 0.003$ ($\text{NRR}@50 = 0.697$), while GPT-4o drops from 0.775 ± 0.006 to 0.662 ± 0.002 ($\text{NRR}@50 = 0.854$). These changes are substantially larger than the run-to-run variation, indicating that the degradation trend is not dominated by generation stochasticity.

We call this contrast the *Retrieval–Reasoning Paradox*: deeper natural retrieval yields diminishing reader-side gains, while controlled robustness can deteriorate as distractors accumulate. Full per-depth scores and trial standard deviations are reported in Appendix D.

Noise tolerance also remains condition-specific. On E2E-WTQ with hard negatives, Qwen-3-32b reaches its Noise Ceiling earlier than GPT-OSS-20b and GPT-4o, but this ordering is not universal across tasks (Table 1).

As an auxiliary trajectory-level diagnostic, we define *Recall Exploitation Efficiency* (REE) to summarize how well a reader preserves robustness over the pre-plateau range of natural retrieval. For each dataset–reader pair, let K^* denote the tested depth where the mean *Uncontrolled* BGE-M3 NRR curve reaches its empirically identified plateau, and let $\mathcal{K}_{\leq K^*}$ denote the evaluated depths up to K^* . For each controlled noise condition, $\text{REE} = \frac{1}{|\mathcal{K}_{\leq K^*}|} \sum_{K \in \mathcal{K}_{\leq K^*}} \text{NRR}@K$. Higher REE indicates better preservation of reader performance across this range. REE is used only as a complementary summary; the retrieval-depth specifications in this paper are determined by the condition-specific Noise Ceiling K_c .

Accordingly, the condition-specific K_c can serve as an operating target: higher ceilings permit more recall-oriented retrieval at larger K , whereas lower ceilings favor precision-oriented retrieval or reranking within the reader’s demonstrated tolerance.

4.2 Evaluating Semantic Distraction

The hard–soft gap varies substantially across readers, task types, and depths. Across many settings, GPT-4o, GPT-OSS-20b, and Claude-Haiku-4.5 show narrower gaps than Qwen-3-32b, but neither the magnitude nor the direction of the gap is universal, reinforcing that noise tolerance is condition-specific.

On OTT-QA, some reader–depth combinations

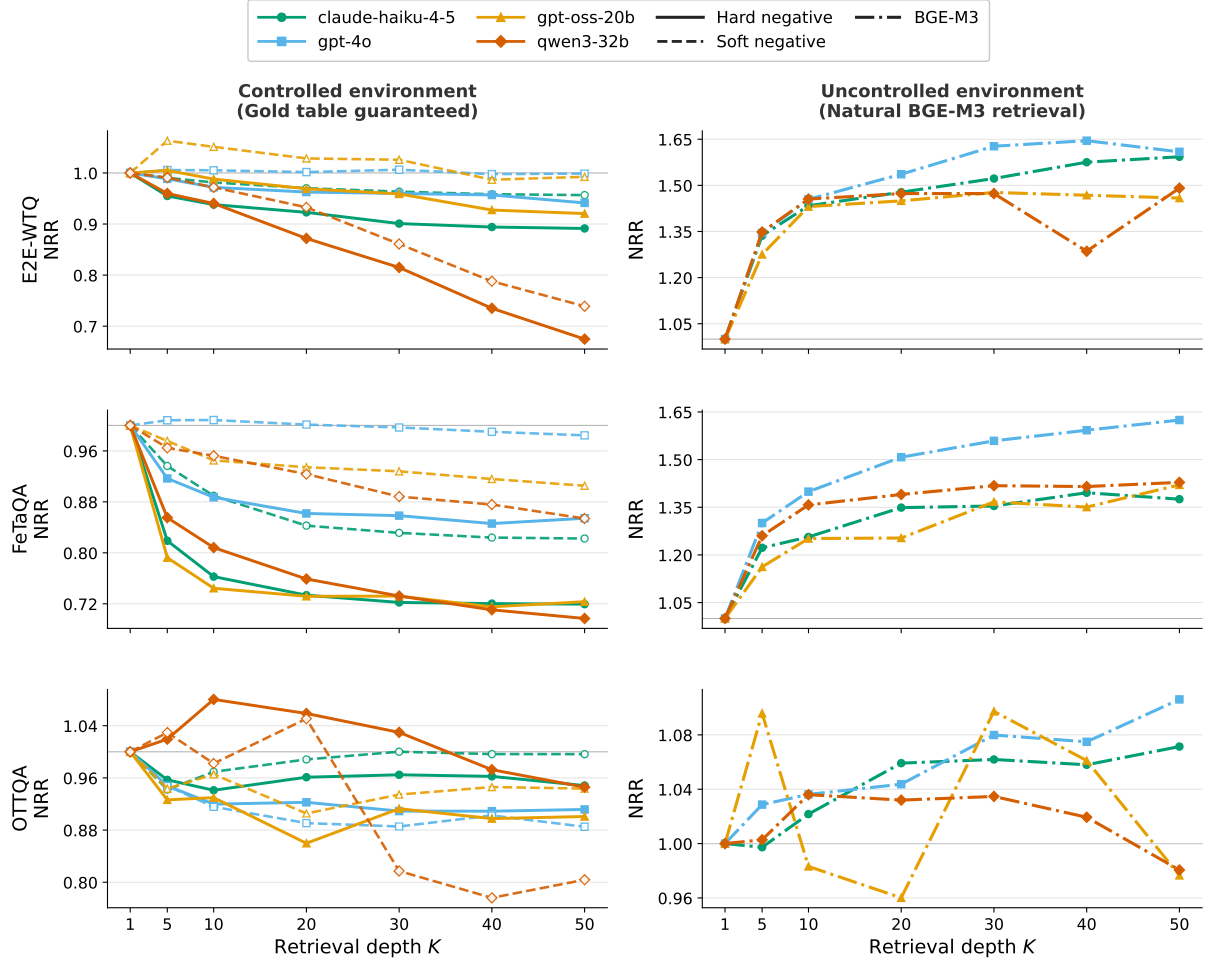


Figure 2: **The Retrieval-Reasoning Paradox.** Reader-side NRR under natural BGE-M3 retrieval (*Uncontrolled*) can initially improve with retrieval depth, whereas Controlled NRR isolates sensitivity to additional distractors once the gold table is guaranteed. Dash-dotted curves denote Uncontrolled NRR; solid and dashed curves denote Controlled hard- and soft-negative NRR, respectively.

(e.g., Qwen-3-32b and GPT-4o) even perform slightly better with hard than soft negatives. One possible explanation is that, for multi-hop questions, semantically related distractors can sometimes provide useful contextual cues rather than pure interference.

4.3 Task-Dependent Positional Sensitivity

Rather than a universal “Lost-in-the-Middle” pattern, positional sensitivity is strongly task-dependent. We quantify it using *Absolute Positional Variance*, the mean absolute deviation of accuracy across Start, Middle, and End placements. All estimated input lengths at $K = 50$ remain below the available context limits of the evaluated model endpoints, making truncation unlikely to explain these positional effects (Appendix B). To expose the direction hidden by variance alone, Table 10 reports Start/Middle/End accuracies separately at each high-depth setting $K \in \{30, 40, 50\}$.

Figure 3 shows substantially greater positional sensitivity on FeTaQA than on E2E-WTQ, particularly under hard negatives. The disaggregated results show that this effect is driven primarily by an advantage for Start placement rather than a universal Middle-specific deficit. The pattern extends to GPT-4o and Claude-Haiku-4.5, indicating that even stronger readers remain position-sensitive on long-form answer generation under dense distractors. Thus, when reasoning is complex, guaranteeing the gold table’s presence does not eliminate sensitivity to where that evidence appears.

4.4 Reader-Side Cost-Effectiveness

To address RQ2, we compare reader-side Cost per Success (CpS) across retrieval configurations. CpS amortizes reader-inference cost over successful answers, capturing how both context growth and failures affect the cost of a correct answer; it excludes retrieval and deployment overheads. Figure 4 sum-

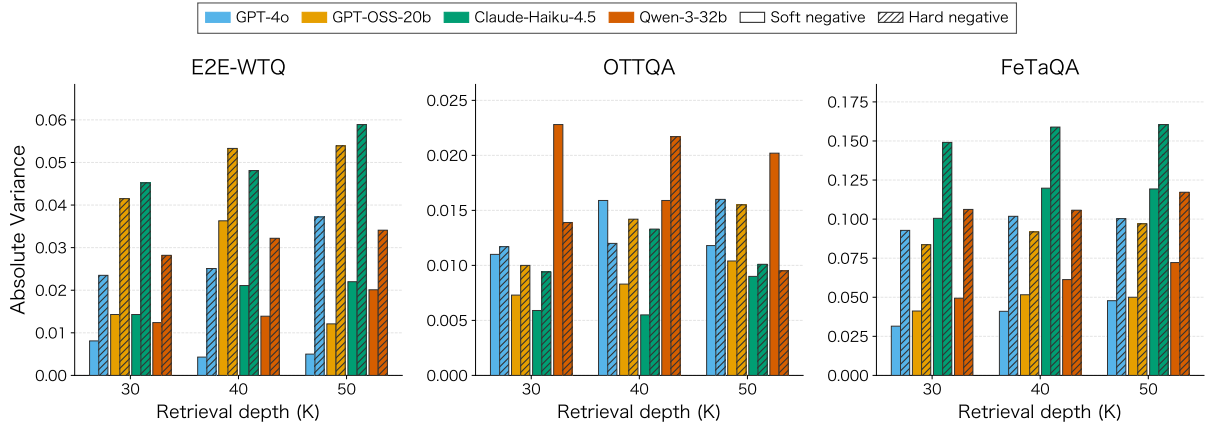


Figure 3: **Positional Sensitivity.** At $K \geq 30$, short-answer tasks (E2E-WTQ) remain relatively stable, whereas long-form answer generation (FeTaQA) exhibits substantially greater positional sensitivity under hard negatives. The corresponding Start/Middle/End accuracies are reported in Table 10.

marizes these trajectories.

Reader-Side Cost–Robustness Trade-offs

Across the evaluated settings, CpS generally increases with retrieval depth, particularly for the proprietary readers, while GPT-OSS-20b maintains a comparatively flatter trajectory. The ordering of hard, soft, and Uncontrolled conditions remains dataset- and reader-dependent. Claude-Haiku-4.5 incurs substantially lower reader-side CpS than GPT-4o across the evaluated datasets, although their relative noise robustness is task-dependent; among open-weight readers, GPT-OSS-20b maintains low reader-side CpS under deep retrieval, while its robustness still varies by task type.

These results make CpS a reader-side diagnostic rather than an end-to-end cost estimate: a reader’s demonstrated noise tolerance indicates how far retrieval depth can increase before additional context yields unfavorable reader-side cost–robustness trade-offs. Deployment decisions must separately account for retrieval/reranking cost, latency, indexing, batching, caching, and hardware utilization.

4.5 Qualitative Error Analysis

We first track each query from the clean baseline ($K = 1$) to the maximum tested depth ($K = 50$) under hard negatives, yielding four behavioral quadrants: *Absolute Robustness* (success $\rightarrow$ success), *Inherent Capability Deficit* (failure $\rightarrow$ failure), *Serendipitous Hallucination* (failure $\rightarrow$ success), and *Noise-Induced Failure* (success $\rightarrow$ failure). Figure 5 visualizes these transitions. Because our goal is to explain failures introduced by retrieval noise, the subsequent error-type analysis fo-

cuses on *Noise-Induced Failures*, aggregated across the full Controlled hard- and soft-negative grid.

Automatic Failure Labeling We assign mutually exclusive labels using deterministic structural heuristics in a fixed order, with an LLM pass used only as auxiliary diagnostic evidence. First, *Distractor Extraction* is assigned when the incorrect answer contains an exact normalized match to a distractor data-cell value absent from the gold table. Among remaining cases, outputs with explicit refusal or insufficient-information phrases are labeled *Premature Refusal*; all others are labeled *Reasoning Hallucination*, covering residual arithmetic, aggregation, or synthesis errors. Provider-filtered cases with no reader output are excluded.

Impact of Distractor Type *Distractor Extraction* dominates under both conditions (84.9% hard; 78.2% soft), but hard negatives shift the distribution further toward direct extraction (Figure 6a). Soft negatives instead show higher *Premature Refusal* (5.6% vs. 3.3%) and *Reasoning Hallucination* (16.2% vs. 11.8%). This pattern appears across all three benchmarks, with dataset-dependent magnitude.

Failure Modes by Question Type *Distractor Extraction* is also cross-cutting rather than Lookup-specific: under hard negatives it accounts for 84.2% of Lookup, 86.3% of Comparison, and 90.4% of Aggregation failures. Although Lookup questions constitute 73.1% of all hard-negative extraction errors, they occupy a similar share of the analyzed failures. We therefore interpret extraction as a general evidence-selection failure and use Lookup only as the most prevalent question type for the reader-

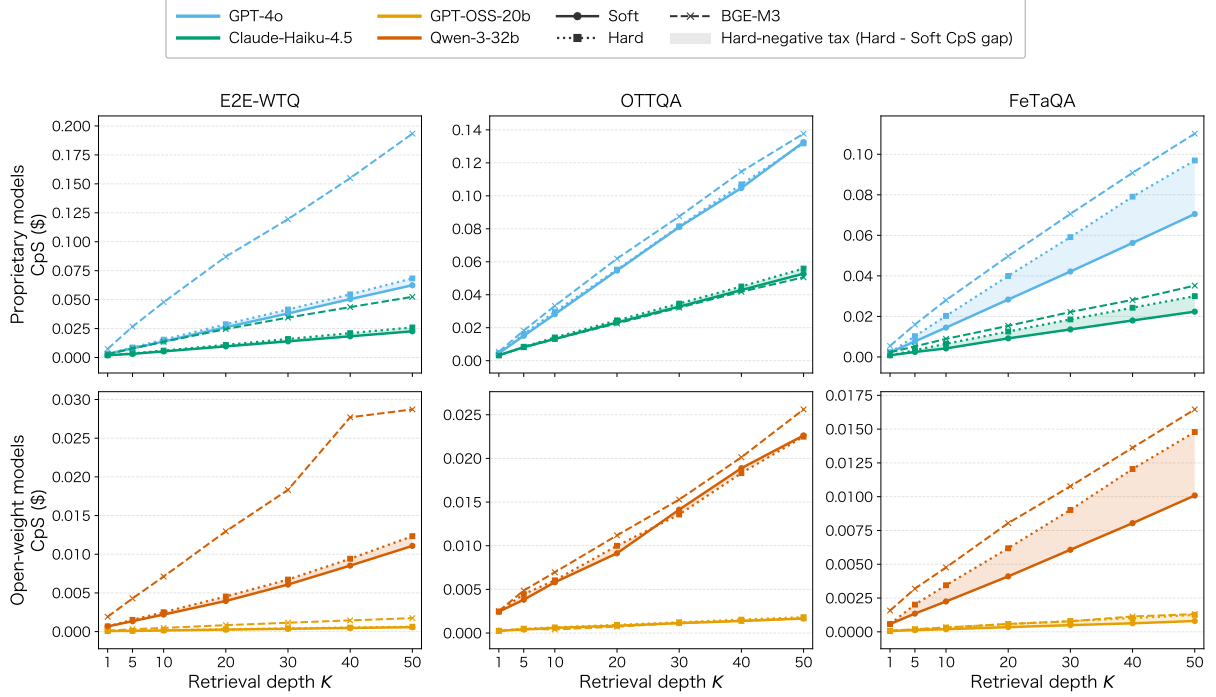


Figure 4: **Reader-Side Cost per Success (CpS) Dynamics.** CpS generally increases with retrieval depth for the proprietary readers under the evaluated conditions, while GPT-OSS-20b exhibits a comparatively flatter reader-side trajectory. Shaded regions visualize the CpS difference between hard and soft negatives.

level comparison below.

Model Behavioral Profiles Within Lookup questions under hard negatives, Figure 6(b) shows distinct residual profiles despite Distractor Extraction remaining dominant for every reader. GPT-OSS-20b has the lowest extraction share (70.8%) and the highest *Reasoning Hallucination* share (28.2%), whereas extraction accounts for roughly 88–91% of failures for the other readers. Claude-Haiku-4.5 has the highest *Premature Refusal* share (6.0%), although refusal remains a minority mode. Thus, readers differ not only in robustness but also in how their noise-induced failures manifest.

Conditional Retrieval Specifications Finally, Table 1 translates these condition-specific robustness patterns into retrieval specifications. Under hard negatives, FeTaQA constrains every reader to $K_c^{\text{hard}} \leq 5$, whereas E2E-WTQ and OTT-QA permit substantially deeper retrieval for several readers. Retrieval depth should therefore be specified by task type, reader, and distractor hardness rather than by reader class: on E2E-WTQ under hard negatives, the open-weight GPT-OSS-20b tolerates $K \geq 50$ while the proprietary Claude-Haiku-4.5 reaches its ceiling at $K = 30$. These aggregate specifications should be interpreted alongside the

Table 1: **Conditional Retrieval Specification Matrix.** Noise Ceilings over the full grid $K \in \{1, 5, 10, 20, 30, 40, 50\}$ at the $\text{NRR}@K \geq 0.9$ retention threshold. ≥ 50 denotes robustness through the maximum evaluated depth.

Task Type	$K_c^{\text{hard}} / K_c^{\text{soft}}$			
	Qwen-3-32b	GPT-OSS-20b	Claude-Haiku-4.5	GPT-4o
E2E-WTQ (Short-answer)	10/20	$\geq 50 / \geq 50$	30/ ≥ 50	$\geq 50 / \geq 50$
OTT-QA (Multi-hop)	$\geq 50 / 20$	$\geq 50 / \geq 50$	$\geq 50 / \geq 50$	$\geq 50 / 40$
FeTaQA (Long-form)	1/20	1/ ≥ 50	1/5	5/ ≥ 50

position-specific accuracies in Table 10.

5 Conclusion

Current practices typically treat retrieval and reading in Table QA as separate tasks. Within our controlled single-table setting, we connect the two by showing that the useful retrieval depth is constrained by the reader’s condition-specific noise tolerance. The resulting *Retrieval Specification Matrix* (Table 1) provides a practical guide to reader-retriever configuration.

The Reader–Retriever Trade-off (RQ1). Retrieval depth should not be maximized blindly, nor inferred from reader class; it should be chosen per task type, reader, and distractor hardness. Under

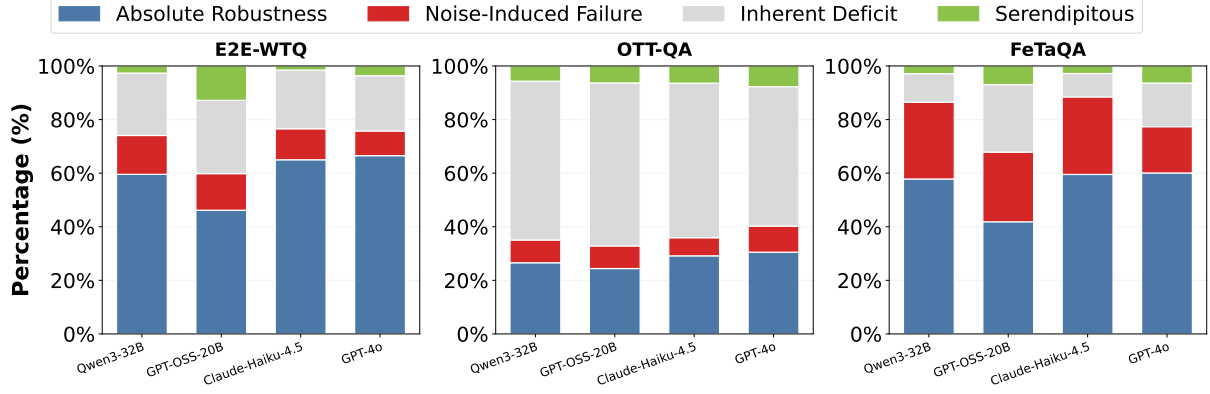
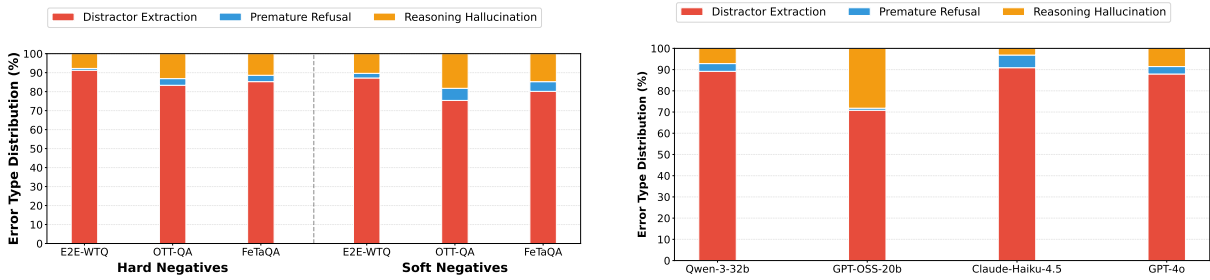


Figure 5: **Behavioral Shifts under Dense Noise.** Query-level transitions from $K = 1$ to $K = 50$ under hard negatives. *Noise-Induced Failure* denotes queries answered correctly at $K = 1$ but incorrectly after dense distractors are introduced.



(a) Error Modes by Distractor Type
Figure 6: **Anatomy of Noise-Induced Failures.** (a) *Distractor Extraction* dominates under both noise conditions, with a larger share under hard negatives; soft negatives shift a modest fraction of failures toward *Premature Refusal* and *Reasoning Hallucination*. (b) Within Lookup questions under hard negatives, extraction remains the dominant failure mode for every reader, while GPT-OSS-20b shows a substantially larger share of residual *Reasoning Hallucination*.

hard negatives, long-form answer generation on FeTaQA imposes the strictest limits, with all evaluated readers reaching $K_c^{\text{hard}} \leq 5$. Soft negatives often permit deeper contexts, although this ordering is not universal. Several readers remain robust through $K = 50$ on short-answer and multi-hop tasks, permitting more recall-oriented retrieval within those conditions. Gold-table position can further affect robustness on complex tasks, so aggregate Noise Ceilings should be interpreted alongside the position-specific results. Across noise-induced failures, Distractor Extraction is the dominant mechanism across question types, while residual refusal and reasoning profiles vary by reader.

Reader-Side Cost-Effectiveness (RQ2). Nominal token price alone is insufficient to characterize reader-side cost-effectiveness under retrieval noise. Because failures are amortized into CpS, retrieval depth and noise tolerance jointly shape the cost of a correct answer. For example, Claude-Haiku-4.5 achieves substantially lower reader-side CpS than GPT-4o across the evaluated datasets, although their relative noise robust-

ness remains task-dependent. CpS excludes retrieval/reranking and other deployment costs and should therefore be interpreted as a reader-side diagnostic, not an end-to-end cost estimate.

Future Work. Future work should extend this controlled analysis to multi-table retrieval and evidence integration, vary distractor ordering and serialization, and evaluate the resulting retrieval-depth recommendations under end-to-end retrieval/reranking behavior and cost.

Limitations

We note several limitations of our study.

Restriction to Single-Table Retrieval: Our controlled design assumes that all required evidence resides in a single gold table, allowing us to isolate reader noise tolerance from cross-table linking and join errors. The reported Noise Ceilings should therefore be interpreted within this favorable single-table setting and should not be directly extrapolated to multi-table QA. Multi-table settings introduce additional evidence-linking and integration challenges that may further reduce effective retrieval

tolerance. Within each benchmark, question-level heterogeneity may introduce additional variation not captured by the dataset-level Noise Ceilings.

Context Serialization Format: Our study exclusively evaluates a Markdown serialization strategy with dynamic row-chunking. Alternative formats (e.g., HTML, JSON, or a programmatic REPL) may alter attention allocation and therefore shift both the magnitude of noise-induced degradation and the exact location of a model’s Noise Ceiling. Our conclusions about these quantities should therefore be interpreted within the evaluated serialization setting; testing whether the observed distractor-driven failure patterns persist across alternative formats remains an important direction for future work.

Scope of Evaluated Models: We evaluate four readers: two open-weight and two proprietary models (Qwen-3-32b, GPT-OSS-20b, Claude-Haiku-4.5, and GPT-4o). Smaller or aggressively fine-tuned readers, or jointly trained retrievers, may exhibit different noise-tolerance profiles and Noise Ceilings.

Semantic Judge Bias: Although the semantic judge is restricted to a constrained YES/NO equivalence decision, is blind to retrieved context and answer provenance, and achieves 86.29% agreement with adjudicated human judgments, residual evaluator bias cannot be ruled out. Agreement is also dataset-dependent, with lower human–judge agreement on FeTaQA than on E2E-WTQ and OTT-QA. In addition, because GPT-OSS-20b is itself one of the evaluated readers, shared model tendencies may still influence equivalence judgments despite provenance blinding.

Fixed Distractor Ordering: Within each controlled condition, we hold the relative order among distractor tables fixed while varying the gold-table position. This isolates the effect of gold placement but does not capture sensitivity to permutations among distractors themselves. Future work should randomize distractor order to quantify this additional source of positional variability.

Scope of Cost per Success (CpS): CpS is a reader-side diagnostic based on nominal reader-inference cost and should not be interpreted as an end-to-end deployment-cost estimate. It excludes retriever/reranker cost, latency, indexing, batching, cache reuse, and other deployment overheads. Nominal API prices are also volatile, while the realized cost of self-hosted open-weight models depends on hardware utilization, batch size, and

KV-cache optimizations. CpS therefore supports standardized reader-side comparison rather than full deployment budgeting.

Acknowledgments

This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under Grants NSTC 115-2221-E-005-095-MY3, NSTC 113-2221-E-A49-165-MY3, and NSTC 115-2218-E-005-002, and in part by the Delta Research Center, Delta Electronics, Inc. This work was also supported by the Smart Sustainable New Agriculture Research Center (SMARTer), funded by NSTC under Grant NSTC 114-2634-F-005-002.

During the preparation of this manuscript, the authors used Google Gemini and OpenAI ChatGPT/Codex for text polishing, language consistency checks, and assistance with data-processing workflows. All AI-generated outputs were thoroughly reviewed and verified by the authors, who take full responsibility for the contents and findings of this work.

References

- Reyhaneh Abdolazimi, Shengmin Jin, Pramod Varshney, and Reza Zafarani. 2026. [Harnessing the power of noise: A survey of techniques and applications](#). *ACM Comput. Surv.*, 58(13).
- Anthropic. 2025. [Claude Haiku 4.5 system card](#). Technical report, Anthropic.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W. Cohen. 2021. Open question answering over tables and text. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Tsung-Hsiang Chou, Chen-Jui Yu, Shui-Hsiang Hsu, and Yao-Chung Fan. 2026. [Cgpt: Cluster-guided partial tables with llm-generated supervision for table retrieval](#). In *Proceedings of the ACM Web Conference 2026*, WWW ’26, page 8585–8588, New York, NY, USA. Association for Computing Machinery.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for

- rag systems. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 719–729.
- Yujian Gan, Xinyun Chen, Qiuping Huang, Matthew Purver, John R Woodward, Jinxia Xie, and Pengsheng Huang. 2021. Towards robustness of text-to-sql models against synonym substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2505–2515.
- Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin 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.
- Shui-Hsiang Hsu, Tsung-Hsiang Chou, Chen-Jui Yu, and Yao-Chung Fan. 2026. **STAR: semantic table representation with header-aware clustering and adaptive weighted fusion**. In *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pages 8773–8776. ACM.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.
- Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. **Retrieval-augmented generation for knowledge-intensive nlp tasks**. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.
- Hsing-Ping Liang, Che-Wei Chang, and Yao-Chung Fan. 2025. Improving table retrieval with question generation from partial tables. In *Proceedings of the 4th Table Representation Learning Workshop*, pages 217–228.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12:157–173.
- Dehai Min, Nan Hu, Rihui Jin, Nuo Lin, Jiaoyan Chen, Yongrui Chen, Yu Li, Guilin Qi, Yun Li, Nijun Li, and Qianren Wang. 2024. **Exploring the impact of table-to-text methods on augmenting LLM-based question answering with domain hybrid data**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 464–482, Mexico City, Mexico. Association for Computational Linguistics.
- Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir Radev. 2022. **FeTaQA: Free-form table question answering**. *Transactions of the Association for Computational Linguistics*, 10:35–49.
- OpenAI. 2024. **GPT-4o system card**.
- OpenAI. 2025. **GPT-OSS-120b & GPT-OSS-20b model card**. *Preprint*, arXiv:2508.10925.
- Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and Peter Fox. 2021. **CLTR: An end-to-end, transformer-based system for cell-level table retrieval and table question answering**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 202–209, Online. Association for Computational Linguistics.
- Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1470–1480. Association for Computational Linguistics.
- Xinyu Pi, Bing Wang, Yan Gao, Jiaqi Guo, Zhoujun Li, and Jian-Guang Lou. 2022. **Towards robustness of text-to-SQL models against natural and realistic adversarial table perturbation**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2022, Dublin, Ireland. Association for Computational Linguistics.
- Matt Post. 2018. **A call for clarity in reporting BLEU scores**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Zipeng Qiu. 2025. **OpenTable-R1: A reinforcement learning augmented tool agent for open-domain table question answering**. *Preprint*, arXiv:2507.03018.
- Qwen Team. 2025. **Qwen3 technical report**. *Preprint*, arXiv:2505.09388.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

Alane Suhr, Ming-Wei Chang, Peter Shaw, and Kenton Lee. 2020. Exploring unexplored generalization challenges for cross-database semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8372–8388.

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.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. [Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Shengming Zhao, Yuheng Huang, Jiayang Song, Zhijie Wang, Chengcheng Wan, and Lei Ma. 2024. Towards understanding retrieval accuracy and prompt quality in RAG systems. *arXiv preprint arXiv:2411.19463*.

Jiaru Zou, Dongqi Fu, Sirui Chen, Xinrui He, Zihao Li, Yada Zhu, Jiawei Han, and Jingrui He. 2026. [RAG over tables: Hierarchical memory index, multi-stage retrieval, and benchmarking](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 38133–38152, San Diego, California, United States. Association for Computational Linguistics.

A Experimental Reproducibility Details

To ensure full reproducibility of our experiments, we detail the hyperparameter configurations and the exact prompt templates utilized for both the reader models and the semantic judge.

A.1 Hyperparameter Settings

To accurately reflect our experimental setup and accommodate the varying requirements of generation and evaluation, we utilized distinct configurations for the reader models and the semantic judge.

Reader Models (Qwen-3-32b², GPT-OSS-20b, Claude-Haiku-4.5, GPT-4o):

- **Temperature:** 0.7 (This parameter setting induces generation stochasticity, fundamentally necessitating our three independent trials to robustly amortize variance, as detailed in [section 3](#)).
- **Top-p:** 1.0
- **Max New Tokens:** Provider-specific generation limits, set sufficiently high to accommodate free-form explanatory generation for complex tasks such as FeTaQA.

Semantic Judge (GPT-OSS-20b):

- **Temperature:** 0.0
- **Top-p:** 1.0
- **Max New Tokens:** 10 (Strictly constrained to prevent verbosity and restrict the output to binary YES/NO tokens).

A.2 Semantic Judge Prompt Template

As introduced in [subsection 3.3](#), strict string matching is inadequate for evaluating complex Table QA. We utilize GPT-OSS-20b as a semantic judge. The evaluation uses the following constrained prompts to keep the decision focused on factual equivalence:

System Prompt:

You are an expert evaluator. Decide if the user answer correctly answers the question and aligns with the gold answer. Treat paraphrases, extra context, different order, and formatting (markdown/HTML/punctuation/case) as equivalent. If the gold answer text is clearly present or entailed in the user answer, and it answers the question without contradiction, respond YES. Respond with ONLY 'YES' or 'NO', nothing else.

User Prompt:

*{question_part}
Gold Answer: {gold_answer}
User Answer: {user_answer}*

Are these semantically equivalent? Respond with only YES or NO.

A.3 Reader Generation Prompt Template

For the reader models, the retrieved tabular context and the user query are formatted using a standardized, zero-shot template across all test environments. This minimal design keeps prompting consistent across models and test environments, reducing prompt-engineering variation as a potential confound in the comparison.

System Prompt:

You are an assistant that answers questions about tables.

²Qwen-3-32b was accessed through OpenRouter, where the available endpoint supports up to 131K context tokens.

User Prompt:

Answer the following question about the given table(s).

Question: {question}

Tables:
{tables_str}

Answer:

B Dataset Token Statistics and Context Limits

A critical requirement for our stress test (especially the positional sensitivity analysis in [subsection 4.3](#)) is ensuring that no retrieved tables are truncated due to context window limitations.

[Table 2](#) summarizes the average and maximum input token lengths at our maximum retrieval depth ($K = 50$). Even under the most extreme conditions (i.e., E2E-WTQ at $K = 50$), the maximum prompt length is approximately 118,000 tokens. These estimated prompt lengths remain below the available context limits of all evaluated model endpoints, indicating that input truncation is unlikely to explain the observed performance degradation or positional variance under the evaluated settings.

Table 2: Input token length statistics at retrieval depth $K = 50$ (estimated using tiktoken, c1100k_base).

Dataset	Avg. Tokens	Max Tokens
E2E-WTQ	57,619	117,914
OTT-QA	21,621	33,151
FeTaQA	25,279	53,413

C Semantic Judge Validation

The semantic judge performs a constrained factual-equivalence decision using only the question, reference answer, and reader output. Because benchmark-aligned automatic metrics differ substantially in granularity and sensitivity to free-form output, we use them as complementary external checks rather than treating any single automatic metric as human ground truth.

Benchmark-specific agreement. We first compare the semantic judgments against benchmark-aligned automatic evaluation signals. For E2E-WTQ, we use denotation correctness from the official WikiTableQuestions evaluator ([Pasupat and Liang, 2015](#)); OTT-QA uses its official exact-match (EM) and token-F1 metrics ([Chen et al., 2021](#)); FeTaQA uses corpus-level sacreBLEU ([Nan et al.,](#)

Table 3: **Automatic comparison** We compare the semantic judge against the corresponding benchmark-aligned automatic evaluation signal. E2E-WTQ and OTT-QA permit item-level comparison, whereas FeTaQA provides a corpus-level sacreBLEU score.

Dataset	Automatic comparison	Result
E2E-WTQ	Denotation agreement	26.67%
OTT-QA	EM agreement	66.52%
FeTaQA	Corpus BLEU association	$r = 0.448, \rho = 0.583$

Table 4: **Deterministic answer-extraction sensitivity.** RAW applies the corresponding benchmark-aligned metric directly to the full free-form reader output. EXTRACTION applies the same metric after the frozen, candidate-only short-answer extractor.

Dataset / metric	RAW	EXTRACTION	Change
E2E-WTQ agreement	26.67%	27.86%	+1.19 pp
E2E-WTQ κ	0.012	0.020	+0.008
OTT-QA EM agreement	66.52%	71.76%	+5.24 pp
OTT-QA κ	0.035	0.223	+0.188

[2022; Post, 2018](#)), which does not provide an item-level correctness label.

Interpreting automatic agreement.

Benchmark-aligned automatic metrics and semantic factual-equivalence judgments measure different aspects of free-form QA outputs. In particular, exact or denotation-based metrics may penalize a correct answer embedded in explanatory text. We therefore conducted a pre-specified sensitivity analysis using a deterministic, candidate-only answer extractor that was frozen before rescoring and had no access to references or judge labels. As shown in [Table 4](#), extraction had a larger effect on OTT-QA than on E2E-WTQ, indicating substantial surface-form sensitivity for the former while leaving much of the metric-judge disagreement in E2E-WTQ unexplained. At the group level, OTT-QA EM association changed from $r = -0.448, \rho = -0.656$ under RAW scoring to $r = 0.110, \rho = 0.169$ after extraction.

Human validation. We further validate the semantic judge against blinded human factual-equivalence annotations on a primary sample of 200 judgments per dataset (600 total). Two independent annotators evaluate only the question, reference answer, and original reader response, without access to the judge decision, benchmark metric, reader identity, retrieval condition, or answer

Table 5: **Blinded human validation of the semantic judge.** Two final UNCERTAIN judgments are excluded from binary judge–human agreement.

Dataset	Eligible N	Agreement	κ
E2E-WTQ	199	90.95%	0.785
FeTaQA	200	80.50%	0.561
OTT-QA	199	87.44%	0.713
Overall	598	86.29%	0.723

provenance. The two annotators were members of the authors’ research team; no external annotators were recruited or separately compensated for this validation. Disagreements and uncertain cases are independently re-evaluated by a blinded adjudicator. Before adjudication, annotators agree on 89.67% of the 600 judgments ($\kappa = 0.796$).

After adjudication, two judgments remain UNCERTAIN and are excluded from binary judge–human comparison. Across the remaining 598 judgments, the semantic judge agrees with the final human labels on 86.29% ($\kappa = 0.723$). Agreement is highest on E2E-WTQ (90.95%, $\kappa = 0.785$), followed by OTT-QA (87.44%, $\kappa = 0.713$) and FeTaQA (80.50%, $\kappa = 0.561$), showing dataset-dependent agreement, with the lowest agreement on FeTaQA’s long-form answer generation setting.

Human Annotation Instructions. Annotators were instructed to judge whether the reader response was factually equivalent to the reference answer using one of three labels: CORRECT, INCORRECT, or UNCERTAIN. CORRECT denotes factual equivalence despite harmless differences in wording or formatting; INCORRECT denotes a factual mismatch, missing required information, or an error that changes the answer; UNCERTAIN is reserved for cases where equivalence cannot be determined confidently from the provided information alone. Annotators were instructed not to consult the other annotator or any model/judge metadata during independent annotation.

D Full Quantitative Results

To support the visual trends presented in the main text and facilitate future reproducibility, we provide detailed quantitative results across all datasets and models.

Table 6, Table 7, and Table 8 detail the Score@ K and Noise Robustness Ratio (NRR) across the full retrieval-depth grid used to plot Figure 2. All values represent the mean across three independent

trials (Temperature=0.7), with variance reported as the standard deviation. Table 9 reports the corresponding token consumption and Cost per Success (CpS) at representative depths $K \in \{1, 10, 30, 50\}$ for the trends summarized in Figure 4.

Magnitude Relative to Trial Variability. To assess whether the observed degradation is dominated by generation stochasticity, we compare changes in Score@ K with the run-to-run standard deviations across the three trials. For example, on FeTaQA under hard negatives, Qwen-3-32b decreases from 0.867 ± 0.012 at $K = 1$ to 0.604 ± 0.003 at $K = 50$. The corresponding drop of 0.263 is substantially larger than the observed trial-to-trial variation. Similar patterns occur for the other readers under this condition. These comparisons indicate that the degradation trend is not dominated by run-to-run variability; we do not treat this standard-deviation comparison as a formal significance test.

Per-Position Accuracy. To complement the positional variance analysis in subsection 4.3, Table 10 reports the underlying Start, Middle, and End accuracies separately at each high-depth setting $K \in \{30, 40, 50\}$ used in Figure 3.

Table 6: Comprehensive evaluation results (Score@ K and NRR) for **E2E-WTQ**. Retrieval depths are evaluated across the full experimental grid ($K \in \{1, 5, 10, 20, 30, 40, 50\}$).

Dataset	Model	K	Hard negatives		Soft negatives		BGE-M3 (Uncontrolled)	
			Score	NRR	Score	NRR	Score	NRR
E2E-WTQ	GPT-4o	1	0.809 $\pm$ 0.100	1.000	0.809 $\pm$ 0.100	1.000	0.460 $\pm$ 0.057	1.000
		5	0.800 $\pm$ 0.114	0.989	0.814 $\pm$ 0.113	1.006	0.619 $\pm$ 0.113	1.345
		10	0.786 $\pm$ 0.110	0.972	0.813 $\pm$ 0.110	1.005	0.669 $\pm$ 0.094	1.454
		20	0.779 $\pm$ 0.107	0.963	0.810 $\pm$ 0.098	1.002	0.707 $\pm$ 0.108	1.536
		30	0.776 $\pm$ 0.097	0.960	0.814 $\pm$ 0.095	1.006	0.749 $\pm$ 0.094	1.627
		40	0.774 $\pm$ 0.100	0.957	0.807 $\pm$ 0.103	0.998	0.757 $\pm$ 0.101	1.645
		50	0.762 $\pm$ 0.103	0.941	0.808 $\pm$ 0.109	0.999	0.741 $\pm$ 0.100	1.609
	Claude-Haiku-4.5	1	0.826 $\pm$ 0.111	1.000	0.826 $\pm$ 0.111	1.000	0.473 $\pm$ 0.064	1.000
		5	0.788 $\pm$ 0.131	0.955	0.818 $\pm$ 0.122	0.990	0.632 $\pm$ 0.130	1.336
		10	0.775 $\pm$ 0.142	0.938	0.810 $\pm$ 0.120	0.981	0.678 $\pm$ 0.124	1.434
		20	0.762 $\pm$ 0.132	0.923	0.801 $\pm$ 0.119	0.970	0.699 $\pm$ 0.132	1.478
		30	0.744 $\pm$ 0.135	0.901	0.795 $\pm$ 0.123	0.963	0.720 $\pm$ 0.131	1.522
		40	0.738 $\pm$ 0.135	0.894	0.791 $\pm$ 0.125	0.958	0.745 $\pm$ 0.136	1.575
		50	0.736 $\pm$ 0.136	0.891	0.790 $\pm$ 0.128	0.957	0.753 $\pm$ 0.139	1.593
	GPT-OSS-20b	1	0.736 $\pm$ 0.166	1.000	0.736 $\pm$ 0.166	1.000	0.456 $\pm$ 0.103	1.000
		5	0.740 $\pm$ 0.141	1.005	0.783 $\pm$ 0.114	1.063	0.582 $\pm$ 0.140	1.275
		10	0.728 $\pm$ 0.131	0.988	0.774 $\pm$ 0.112	1.051	0.653 $\pm$ 0.117	1.431
		20	0.714 $\pm$ 0.146	0.969	0.757 $\pm$ 0.118	1.028	0.661 $\pm$ 0.147	1.450
		30	0.706 $\pm$ 0.139	0.959	0.755 $\pm$ 0.122	1.026	0.674 $\pm$ 0.132	1.477
		40	0.683 $\pm$ 0.153	0.927	0.727 $\pm$ 0.139	0.987	0.669 $\pm$ 0.152	1.468
		50	0.678 $\pm$ 0.161	0.920	0.731 $\pm$ 0.146	0.992	0.665 $\pm$ 0.158	1.459
	Qwen-3-32b	1	0.809 $\pm$ 0.122	1.000	0.809 $\pm$ 0.122	1.000	0.469 $\pm$ 0.071	1.000
		5	0.776 $\pm$ 0.108	0.960	0.801 $\pm$ 0.112	0.991	0.632 $\pm$ 0.109	1.348
		10	0.761 $\pm$ 0.088	0.940	0.786 $\pm$ 0.086	0.972	0.682 $\pm$ 0.079	1.455
		20	0.705 $\pm$ 0.042	0.872	0.755 $\pm$ 0.036	0.933	0.690 $\pm$ 0.043	1.473
		30	0.659 $\pm$ 0.019	0.815	0.696 $\pm$ 0.031	0.861	0.690 $\pm$ 0.020	1.473
		40	0.595 $\pm$ 0.084	0.735	0.637 $\pm$ 0.118	0.788	0.603 $\pm$ 0.083	1.286
		50	0.546 $\pm$ 0.151	0.675	0.598 $\pm$ 0.160	0.739	0.699 $\pm$ 0.193	1.491

Table 7: Comprehensive evaluation results (Score@ K and NRR) for **OTT-QA**. Retrieval depths are evaluated across the full experimental grid ($K \in \{1, 5, 10, 20, 30, 40, 50\}$).

Dataset	Model	K	Hard negatives		Soft negatives		BGE-M3 (Uncontrolled)	
			Score	NRR	Score	NRR	Score	NRR
OTT-QA	GPT-4o	1	0.425 $\pm$ 0.035	1.000	0.425 $\pm$ 0.035	1.000	0.362 $\pm$ 0.030	1.000
		5	0.403 $\pm$ 0.030	0.948	0.403 $\pm$ 0.046	0.948	0.372 $\pm$ 0.029	1.029
		10	0.391 $\pm$ 0.028	0.920	0.389 $\pm$ 0.047	0.916	0.375 $\pm$ 0.027	1.036
		20	0.392 $\pm$ 0.021	0.923	0.379 $\pm$ 0.039	0.891	0.378 $\pm$ 0.020	1.044
		30	0.387 $\pm$ 0.017	0.909	0.377 $\pm$ 0.036	0.885	0.391 $\pm$ 0.017	1.080
		40	0.386 $\pm$ 0.010	0.909	0.384 $\pm$ 0.029	0.902	0.389 $\pm$ 0.009	1.075
		50	0.388 $\pm$ 0.008	0.912	0.376 $\pm$ 0.031	0.885	0.400 $\pm$ 0.009	1.106
	Claude-Haiku-4.5	1	0.363 $\pm$ 0.004	1.000	0.363 $\pm$ 0.004	1.000	0.336 $\pm$ 0.003	1.000
		5	0.347 $\pm$ 0.007	0.957	0.342 $\pm$ 0.006	0.942	0.335 $\pm$ 0.006	0.997
		10	0.341 $\pm$ 0.009	0.941	0.352 $\pm$ 0.004	0.969	0.343 $\pm$ 0.009	1.022
		20	0.349 $\pm$ 0.009	0.961	0.359 $\pm$ 0.002	0.988	0.356 $\pm$ 0.010	1.059
		30	0.350 $\pm$ 0.012	0.965	0.363 $\pm$ 0.004	1.000	0.357 $\pm$ 0.012	1.062
		40	0.349 $\pm$ 0.017	0.962	0.361 $\pm$ 0.002	0.996	0.355 $\pm$ 0.016	1.058
		50	0.344 $\pm$ 0.021	0.948	0.361 $\pm$ 0.005	0.996	0.360 $\pm$ 0.022	1.071
	GPT-OSS-20b	1	0.324 $\pm$ 0.005	1.000	0.324 $\pm$ 0.005	1.000	0.273 $\pm$ 0.004	1.000
		5	0.300 $\pm$ 0.019	0.926	0.306 $\pm$ 0.002	0.943	0.299 $\pm$ 0.020	1.096
		10	0.301 $\pm$ 0.022	0.930	0.313 $\pm$ 0.016	0.965	0.269 $\pm$ 0.020	0.983
		20	0.278 $\pm$ 0.010	0.860	0.293 $\pm$ 0.009	0.905	0.262 $\pm$ 0.011	0.960
		30	0.296 $\pm$ 0.026	0.913	0.303 $\pm$ 0.016	0.934	0.300 $\pm$ 0.026	1.097
		40	0.291 $\pm$ 0.025	0.898	0.306 $\pm$ 0.018	0.946	0.290 $\pm$ 0.024	1.061
		50	0.292 $\pm$ 0.028	0.901	0.306 $\pm$ 0.016	0.944	0.267 $\pm$ 0.026	0.977
	Qwen-3-32b	1	0.349 $\pm$ 0.006	1.000	0.349 $\pm$ 0.006	1.000	0.325 $\pm$ 0.005	1.000
		5	0.356 $\pm$ 0.039	1.019	0.360 $\pm$ 0.045	1.030	0.326 $\pm$ 0.041	1.003
		10	0.377 $\pm$ 0.071	1.080	0.343 $\pm$ 0.007	0.982	0.337 $\pm$ 0.049	1.036
		20	0.370 $\pm$ 0.072	1.059	0.367 $\pm$ 0.073	1.051	0.336 $\pm$ 0.054	1.032
		30	0.360 $\pm$ 0.055	1.030	0.285 $\pm$ 0.079	0.817	0.337 $\pm$ 0.051	1.035
		40	0.340 $\pm$ 0.023	0.973	0.271 $\pm$ 0.100	0.776	0.332 $\pm$ 0.024	1.019
		50	0.330 $\pm$ 0.013	0.946	0.281 $\pm$ 0.076	0.804	0.319 $\pm$ 0.013	0.981

Table 8: Comprehensive evaluation results (Score@ K and NRR) for **FeTaQA**. Retrieval depths are evaluated across the full experimental grid ($K \in \{1, 5, 10, 20, 30, 40, 50\}$).

Dataset	Model	K	Hard negatives		Soft negatives		BGE-M3 (Uncontrolled)	
			Score	NRR	Score	NRR	Score	NRR
FeTaQA	GPT-4o	1	0.775 $\pm$ 0.006	1.000	0.775 $\pm$ 0.006	1.000	0.359 $\pm$ 0.003	1.000
		5	0.711 $\pm$ 0.010	0.917	0.781 $\pm$ 0.004	1.008	0.468 $\pm$ 0.009	1.300
		10	0.688 $\pm$ 0.004	0.887	0.782 $\pm$ 0.006	1.008	0.503 $\pm$ 0.003	1.399
		20	0.668 $\pm$ 0.009	0.862	0.776 $\pm$ 0.006	1.001	0.542 $\pm$ 0.010	1.508
		30	0.665 $\pm$ 0.008	0.858	0.773 $\pm$ 0.005	0.997	0.560 $\pm$ 0.007	1.559
		40	0.656 $\pm$ 0.004	0.846	0.767 $\pm$ 0.006	0.990	0.573 $\pm$ 0.003	1.593
		50	0.662 $\pm$ 0.002	0.854	0.763 $\pm$ 0.002	0.984	0.584 $\pm$ 0.002	1.625
	Claude-Haiku-4.5	1	0.886 $\pm$ 0.008	1.000	0.886 $\pm$ 0.008	1.000	0.396 $\pm$ 0.004	1.000
		5	0.725 $\pm$ 0.009	0.819	0.830 $\pm$ 0.007	0.936	0.483 $\pm$ 0.004	1.222
		10	0.676 $\pm$ 0.005	0.763	0.788 $\pm$ 0.002	0.889	0.497 $\pm$ 0.004	1.257
		20	0.650 $\pm$ 0.004	0.734	0.747 $\pm$ 0.007	0.843	0.533 $\pm$ 0.003	1.349
		30	0.640 $\pm$ 0.016	0.722	0.737 $\pm$ 0.014	0.831	0.535 $\pm$ 0.013	1.354
		40	0.638 $\pm$ 0.026	0.720	0.730 $\pm$ 0.025	0.824	0.552 $\pm$ 0.027	1.396
		50	0.637 $\pm$ 0.024	0.719	0.729 $\pm$ 0.024	0.822	0.544 $\pm$ 0.020	1.375
	GPT-OSS-20b	1	0.677 $\pm$ 0.006	1.000	0.677 $\pm$ 0.006	1.000	0.308 $\pm$ 0.003	1.000
		5	0.537 $\pm$ 0.005	0.792	0.661 $\pm$ 0.005	0.976	0.358 $\pm$ 0.004	1.162
		10	0.504 $\pm$ 0.005	0.744	0.640 $\pm$ 0.010	0.945	0.386 $\pm$ 0.004	1.252
		20	0.496 $\pm$ 0.001	0.732	0.633 $\pm$ 0.004	0.934	0.386 $\pm$ 0.002	1.253
		30	0.496 $\pm$ 0.003	0.732	0.629 $\pm$ 0.003	0.928	0.421 $\pm$ 0.003	1.367
		40	0.485 $\pm$ 0.002	0.715	0.621 $\pm$ 0.006	0.916	0.416 $\pm$ 0.003	1.351
		50	0.490 $\pm$ 0.002	0.723	0.613 $\pm$ 0.004	0.905	0.438 $\pm$ 0.002	1.421
	Qwen-3-32b	1	0.867 $\pm$ 0.012	1.000	0.867 $\pm$ 0.012	1.000	0.381 $\pm$ 0.005	1.000
		5	0.741 $\pm$ 0.002	0.855	0.836 $\pm$ 0.002	0.965	0.479 $\pm$ 0.002	1.260
		10	0.700 $\pm$ 0.004	0.808	0.825 $\pm$ 0.005	0.952	0.516 $\pm$ 0.003	1.357
		20	0.658 $\pm$ 0.004	0.759	0.800 $\pm$ 0.005	0.924	0.529 $\pm$ 0.005	1.390
		30	0.635 $\pm$ 0.004	0.732	0.770 $\pm$ 0.007	0.888	0.539 $\pm$ 0.004	1.418
		40	0.616 $\pm$ 0.004	0.711	0.759 $\pm$ 0.002	0.876	0.538 $\pm$ 0.004	1.415
		50	0.604 $\pm$ 0.003	0.697	0.740 $\pm$ 0.004	0.854	0.543 $\pm$ 0.002	1.428

Table 9: **Reader token consumption and reader-side Cost per Success (CpS)** corresponding to Figure 4. CpS amortizes nominal reader inference cost over correct answers and excludes retrieval/reranking and other deployment overheads.

Dataset	Model	K	Soft negatives			Hard negatives			BGE-M3 (Uncontrolled)		
			In.	Out.	CpS (\$)	In.	Out.	CpS (\$)	In.	Out.	CpS (\$)
E2E-WTQ	GPT-4o	1	696	84	\$0.0032	696	84	\$0.0032	1,166	49	\$0.0074
		10	4,118	95	\$0.0138	4,287	142	\$0.0154	12,363	106	\$0.0478
		30	11,931	116	\$0.0381	12,177	177	\$0.0415	35,243	137	\$0.1195
		50	19,693	120	\$0.0624	20,095	181	\$0.0683	56,677	149	\$0.1933
	Claude-Haiku-4.5	1	689	181	\$0.0019	689	181	\$0.0019	830	102	\$0.0028
		10	3,694	119	\$0.0053	4,003	125	\$0.0060	8,598	121	\$0.0136
		30	10,446	132	\$0.0140	11,176	140	\$0.0160	24,231	120	\$0.0345
		50	17,186	139	\$0.0226	18,382	149	\$0.0260	38,865	121	\$0.0524
	GPT-OSS-20b	1	655	439	\$0.0001	655	439	\$0.0001	1,240	284	\$0.0001
		10	4,218	306	\$0.0001	4,349	396	\$0.0002	12,495	584	\$0.0005
		30	11,996	371	\$0.0004	12,242	508	\$0.0004	35,484	628	\$0.0011
		50	19,749	213	\$0.0006	20,165	253	\$0.0006	56,958	190	\$0.0017
	Qwen-3-32b	1	759	543	\$0.0007	759	543	\$0.0007	1,342	865	\$0.0019
		10	4,581	685	\$0.0022	4,662	938	\$0.0025	14,462	1,118	\$0.0071
		30	13,082	753	\$0.0061	13,142	1,058	\$0.0067	41,289	1,141	\$0.0183
		50	21,365	725	\$0.0111	21,465	859	\$0.0123	65,708	1,716	\$0.0287
OTT-QA	GPT-4o	1	479	72	\$0.0045	479	72	\$0.0045	485	73	\$0.0054
		10	3,970	105	\$0.0282	4,116	134	\$0.0297	4,403	146	\$0.0332
		30	11,713	122	\$0.0810	11,968	159	\$0.0815	12,917	189	\$0.0875
		50	19,444	129	\$0.1326	19,746	171	\$0.1318	21,247	199	\$0.1377
	Claude-Haiku-4.5	1	434	151	\$0.0033	434	151	\$0.0033	372	149	\$0.0033
		10	3,652	195	\$0.0132	3,785	206	\$0.0141	3,568	189	\$0.0132
		30	10,828	213	\$0.0328	10,998	225	\$0.0346	10,459	205	\$0.0322
		50	17,993	216	\$0.0528	18,087	227	\$0.0559	17,181	206	\$0.0506
	GPT-OSS-20b	1	530	732	\$0.0003	530	732	\$0.0003	549	523	\$0.0002
		10	4,042	1,039	\$0.0006	4,190	1,103	\$0.0006	4,485	201	\$0.0004
		30	11,807	1,075	\$0.0011	12,061	1,228	\$0.0012	13,046	813	\$0.0011
		50	19,573	1,151	\$0.0017	19,859	1,322	\$0.0018	21,431	198	\$0.0017
	Qwen-3-32b	1	534	1,185	\$0.0024	534	1,185	\$0.0024	542	1,121	\$0.0025
		10	4,409	1,213	\$0.0058	4,633	1,582	\$0.0060	4,983	1,525	\$0.0070
		30	11,528	1,173	\$0.0141	13,510	1,638	\$0.0136	14,642	1,522	\$0.0153
		50	19,444	1,214	\$0.0226	22,290	1,638	\$0.0225	24,095	2,002	\$0.0256
FeTaQA	GPT-4o	1	476	52	\$0.0022	476	52	\$0.0022	556	56	\$0.0054
		10	4,265	65	\$0.0145	5,203	92	\$0.0203	5,261	94	\$0.0280
		30	12,724	76	\$0.0421	15,274	115	\$0.0591	15,338	121	\$0.0706
		50	21,197	83	\$0.0705	25,164	128	\$0.0969	25,206	132	\$0.1102
	Claude-Haiku-4.5	1	332	79	\$0.0008	332	79	\$0.0008	387	101	\$0.0023
		10	2,867	88	\$0.0042	3,820	111	\$0.0065	3,843	118	\$0.0089
		30	9,440	111	\$0.0136	11,236	120	\$0.0185	11,253	124	\$0.0222
		50	15,727	111	\$0.0223	18,508	121	\$0.0300	18,520	125	\$0.0352
	GPT-OSS-20b	1	544	304	\$0.0001	544	304	\$0.0001	617	131	\$0.0001
		10	4,357	371	\$0.0002	5,304	558	\$0.0003	5,333	161	\$0.0003
		30	12,840	546	\$0.0005	15,401	874	\$0.0008	15,424	195	\$0.0008
		50	21,323	654	\$0.0008	25,283	1,091	\$0.0013	25,303	661	\$0.0013
	Qwen-3-32b	1	540	582	\$0.0006	540	582	\$0.0006	631	704	\$0.0016
		10	4,865	753	\$0.0022	5,947	1,169	\$0.0034	5,984	1,228	\$0.0048
		30	14,506	794	\$0.0061	17,398	1,150	\$0.0090	17,426	1,282	\$0.0108
		50	24,142	786	\$0.0101	28,597	1,077	\$0.0148	28,618	1,095	\$0.0165

Table 10: **Per-Position Reader Accuracy.** Start, Middle, and End denote the placement of the gold table within the retrieved context. Each position is reported separately at $K \in \{30, 40, 50\}$, corresponding to the high-depth settings used in the positional analysis in Figure 3.

Dataset	Reader	K	Hard negatives			Soft negatives		
			Start	Middle	End	Start	Middle	End
E2E-WTQ	Qwen-3-32b	30	0.702	0.621	0.656	0.703	0.678	0.709
		40	0.643	0.550	0.591	0.644	0.617	0.651
		50	0.597	0.502	0.538	0.612	0.568	0.614
	GPT-OSS-20b	30	0.769	0.672	0.678	0.777	0.739	0.750
		40	0.763	0.636	0.650	0.781	0.689	0.710
		50	0.759	0.644	0.630	0.749	0.725	0.718
	Claude-Haiku-4.5	30	0.812	0.721	0.699	0.816	0.796	0.774
		40	0.810	0.717	0.688	0.823	0.787	0.764
		50	0.824	0.709	0.675	0.823	0.785	0.762
	GPT-4o	30	0.812	0.760	0.757	0.823	0.817	0.802
		40	0.812	0.753	0.757	0.813	0.808	0.801
		50	0.817	0.736	0.731	0.812	0.812	0.801
OTT-QA	Qwen-3-32b	30	0.381	0.341	0.358	0.266	0.320	0.271
		40	0.372	0.328	0.319	0.272	0.247	0.294
		50	0.345	0.316	0.331	0.304	0.251	0.288
	GPT-OSS-20b	30	0.311	0.289	0.288	0.311	0.305	0.292
		40	0.312	0.277	0.283	0.319	0.299	0.301
		50	0.315	0.276	0.284	0.321	0.295	0.301
	Claude-Haiku-4.5	30	0.364	0.345	0.341	0.372	0.358	0.359
		40	0.369	0.342	0.337	0.370	0.358	0.357
		50	0.359	0.336	0.337	0.375	0.358	0.351
	GPT-4o	30	0.404	0.382	0.373	0.393	0.369	0.367
		40	0.404	0.378	0.377	0.408	0.373	0.371
		50	0.412	0.374	0.377	0.394	0.368	0.367
FeTaQA	Qwen-3-32b	30	0.794	0.543	0.567	0.844	0.745	0.720
		40	0.774	0.538	0.535	0.851	0.721	0.705
		50	0.780	0.505	0.528	0.848	0.695	0.677
	GPT-OSS-20b	30	0.621	0.435	0.432	0.690	0.605	0.591
		40	0.622	0.419	0.413	0.698	0.596	0.567
		50	0.636	0.415	0.420	0.688	0.592	0.559
	Claude-Haiku-4.5	30	0.863	0.534	0.522	0.887	0.654	0.668
		40	0.877	0.512	0.526	0.910	0.617	0.663
		50	0.878	0.512	0.523	0.908	0.606	0.672
	GPT-4o	30	0.805	0.602	0.590	0.820	0.762	0.736
		40	0.809	0.587	0.572	0.829	0.755	0.718
		50	0.813	0.596	0.578	0.835	0.741	0.713