# WHEN DOES RETRIEVAL HURT? CHARACTERIZING NEGATIVE RETRIEVAL IN RAG

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Retrieval can supply knowledge that a language model lacks, but it can also overturn answers that the model already gives reliably without context. We study the second effect, *answer preservation under retrieval*, using an evaluation protocol that separates closed-book screening from fresh, paired closed-book and retrieval evaluation. On questions screened as reliably answered, the difference between retrieval error and fresh closed-book error measures net preservation loss; marginal error rates also give sharp bounds on harmful and beneficial answer changes. The protocol reports overall utility and preservation on explicitly defined question populations. On 3,250 questions from five QA benchmarks, top-10 retrieval raises the primary reader's overall accuracy by 9.5 points, yet raises its error on the 1,878 screened questions by 12.8 points relative to the fresh control. A matched intervention on 600 screened questions keeps complete supporting evidence and replaces a single distractor: in-context RAG makes 25, 47 and 111 errors with an unrelated, topical near-miss and conflicting distractor, respectively, whereas Astute RAG makes 14, 19 and 34. Supporting evidence therefore does not prevent content-dependent errors; the conflict-induced increase is 14.3 percentage points for in-context RAG and 3.3 for Astute RAG. Elementary probability identities connect these measurements and bound the paired harm that marginal error rates can identify.

## 1 INTRODUCTION

Retrieval-augmented generation (RAG) gives language models access to evidence beyond their parameters (Lewis et al., 2020; Izacard et al., 2023; Ram et al., 2023). Its utility is usually summarized by the change in average answer accuracy. That change aggregates movement in two directions: a wrong closed-book answer can become correct, and a correct one can become wrong. The first direction extends knowledge coverage; the second matters whenever the reader already answers a question reliably. We study the second direction as *answer preservation under retrieval*.

Retrieval errors and knowledge conflicts are well documented. Irrelevant passages can distract readers (Shi et al., 2023; Yoran et al., 2024; Amiraz et al., 2025); additional passages can interact with otherwise useful evidence (Cuconasu et al., 2024); and retrieved claims can conflict with parametric knowledge (Longpre et al., 2021; Xie et al., 2024; Wu et al., 2024). Prior work has also stratified contexts by sufficiency (Joren et al., 2025) and analyzed the joint correctness of closed-book and retrieval answers (Wang et al., 2025). We ask a complementary measurement question: *on questions that independent evidence shows a model answers reliably without context, how much does retrieval change performance, and relative to what reference?*

The reference matters. Selecting a question because one closed-book answer was correct also selects favorable generation noise, so a later wrong answer need not reflect a retrieval loss: a fresh closed-book answer may be wrong as well. Requiring several successful screening calls selects more stable behavior but does not reduce future error to zero. We therefore use screening only to define the population and evaluate both arms on fresh calls. The difference in their error rates estimates net preservation loss on that population. We then use integer margins to bound harmful and beneficial changes under the paired evaluation. Applied to 3,250 questions from five benchmarks, the two views diverge sharply: top-10 retrieval raises the primary reader's accuracy by 9.5 points, yet its error on the 1,878 screened-stable questions exceeds the fresh control by 12.8 points.



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

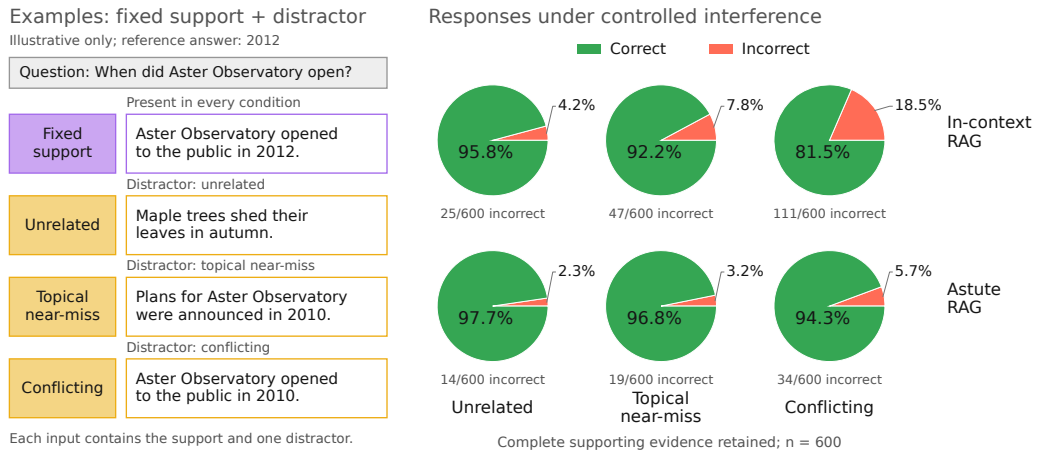


Figure 1: Controlled interference with complete supporting evidence. Left: an illustrative example, constructed for exposition rather than drawn from the study inputs, in which the supporting evidence is fixed and one distractor varies. Right: answer-match outcomes for in-context RAG and Astute RAG on the same 600 questions; sectors and percentages are computed from the error counts in Table 3.

A second question is what in the context drives the change. Natural retrieval bundles differ jointly in support, relevance, length and source composition. A contrast between unrelated and conflicting material is easier to interpret when the answer-supporting evidence stays present. Figure 1 illustrates such a design. On a 600-question controlled subset, replacing an unrelated distractor with a conflicting one raises the error count of in-context RAG from 25 to 111 and that of Astute RAG from 14 to 34. The matched comparison measures the sensitivity of each complete inference procedure to distractor content while keeping answer support available.

We make three contributions. First, we introduce an independently screened preservation protocol with a fresh closed-book reference (Section 2). Second, we connect aggregate utility and selected-set loss through exact accounting identities and sharp bounds on answer transitions. On a common evaluation population, these measurements expose a substantial preservation loss alongside a positive overall accuracy gain (Section 4.1). Third, a matched 600-question intervention tests interference with complete supporting evidence, and an expanded comparison evaluates five existing mitigation procedures using the same preservation target (Sections 4.2 and 4.3). Together, the protocol and intervention distinguish aggregate utility, selected-set preservation and sensitivity to conflicting evidence.

## 2 Measuring Negative Retrieval

### 2.1 Paired outcomes and aggregate utility

Let $q$ be a question with reference aliases $A_q$, and let $C_k(q)$ denote the first $k$ passages in a fixed ordering. We write $Y_0, Y_k \in \{0, 1\}$ for the scored closed-book and retrieval outcomes under a fixed generation and aggregation procedure. Probabilities are taken over questions and the declared sampling procedure. Unless stated otherwise, all quantities refer to the same population of questions with available context.

For a given pairing of the two outcomes, define the harmful and beneficial transition probabilities

$$H = \mathbb{P}(Y_0 = 1, Y_k = 0), \qquad G = \mathbb{P}(Y_0 = 0, Y_k = 1). \tag{1}$$

With $a_0 = \mathbb{E}Y_0$ and $a_k = \mathbb{E}Y_k$, the accuracy change satisfies

$$\Delta = a_k - a_0 = G - H. \tag{2}$$

Conditional destruction and rescue rates divide $H$ by $a_0$ and $G$ by $1 - a_0$, respectively, when these denominators are nonzero. They should not be combined with rates defined on a different screening population.

### 2.2 INDEPENDENT SCREENING AND A FRESH CONTROL

We draw three closed-book screening samples $S_1, S_2, S_3 \in \{0, 1\}$ and form

$$\mathcal{K} = \{q : S_1 = S_2 = S_3 = 1\}. \tag{3}$$

We call $\mathcal{K}$ the *screened-stable set*. Membership is an observed property of three calls, not a guarantee of latent knowledge or of future correctness. Questions outside $\mathcal{K}$ remain in the overall evaluation.

Evaluation uses fresh calls, disjoint from screening, in a closed-book arm and a retrieval arm. On $\mathcal{K}$, define

$$D_{\text{ret}} = \mathbb{P}(Y_k = 0 \mid \mathcal{K}), \quad D_0 = \mathbb{P}(Y_0 = 0 \mid \mathcal{K}), \quad E_{\mathcal{K}} = D_{\text{ret}} - D_0. \tag{4}$$

Here $D_0$ is the residual closed-book error after screening, and $E_{\mathcal{K}}$ is the *selected-set excess error*. With $m = |\mathcal{K}|$, its paired estimator is

$$\widehat{E_{\mathcal{K}}} = \frac{1}{m} \sum_{i:q_i \in \mathcal{K}} (Y_{0i} - Y_{ki}). \tag{5}$$

The fresh control supplies the baseline appropriate to the selected population. Subtracting it removes baseline error from the aggregate contrast; it does not label any individual retrieval failure as caused by context. Interpreting $E_{\mathcal{K}}$ as a context effect requires the same reader configuration, scoring and sampling conditions in both arms, with no systematic change in time or endpoint between them.

**Evaluation protocol (this work).** For each reader, we (i) use three closed-book calls to fix $\mathcal{K}$; (ii) issue separate closed-book and context-conditioned evaluation calls on all questions; and (iii) report overall accuracy, fresh closed-book error, selected-set excess error and margin-compatible transition bounds. Screening labels are used only by the evaluator. Inference policies receive neither screening labels nor reference answers. This separates population selection, answer generation and measurement.

**How screening changes the target.** Under a conditional-independence model, screening has a simple consequence. Write $\mathcal{K}_s$ for the set selected by $s$ successful screening calls, so that $\mathcal{K} = \mathcal{K}_3$. Suppose that question $q$ has single-call closed-book accuracy $p_q$, that the screening calls are conditionally independent given $q$, and that evaluation is independent of screening. Then $\mathbb{P}(q \in \mathcal{K}_s \mid q) = p_q^s$, and Bayes' rule gives

$$\mathbb{P}(q \in \mathcal{K}_s) = \mathbb{E}_P[p_q^s], \qquad E_{\mathcal{K}_s} = \frac{\mathbb{E}_P[p_q^s(b_q - r_q)]}{\mathbb{E}_P[p_q^s]}, \tag{6}$$

where $P$ is the question distribution and $b_q = \mathbb{E}[Y_0 \mid q]$ and $r_q = \mathbb{E}[Y_k \mid q]$ are the per-question accuracies of the two evaluation arms. Screening thus reweights questions toward larger $p_q$ but does not make the fresh control error-free, and changing the number of screening calls changes the target population. If screening calls are correlated, $p_q^s$ is replaced by the actual probability of the all-correct event. The empirical set definition and the accounting identities below do not require independent repeated calls.

### 2.3 WHAT THE MARGINAL CONTRAST IDENTIFIES

Let $H_{\mathcal{K}}$ and $G_{\mathcal{K}}$ denote the transition probabilities in Equation 1, conditioned on $\mathcal{K}$.

**Proposition 1** (Harm from marginal errors)**.** *For any joint distribution of the two scored outcomes on $\mathcal{K}$,*

$$E_{\mathcal{K}} = H_{\mathcal{K}} - G_{\mathcal{K}}, \qquad \max(0, E_{\mathcal{K}}) \le H_{\mathcal{K}} \le \min(D_{\text{ret}}, 1 - D_0). \tag{7}$$

*Both endpoints are attainable given only $D_{\text{ret}}$ and $D_0$.*

*Proof sketch.* Write $p_{ab} = \mathbb{P}(Y_0 = a, Y_k = b \mid \mathcal{K})$. Then $D_{\text{ret}} = p_{10} + p_{00}$ and $D_0 = p_{01} + p_{00}$, so $E_{\mathcal{K}} = p_{10} - p_{01} = H_{\mathcal{K}} - G_{\mathcal{K}}$. Since $H_{\mathcal{K}} = D_{\text{ret}} - p_{00}$, the bounds follow from nonnegativity of the four cells, and each endpoint is attained by setting $p_{00}$ to its largest or smallest feasible value. □

When $E_{\mathcal{K}} \ge 0$, the interval has width at most $D_0$, because $H_{\mathcal{K}} - E_{\mathcal{K}} = G_{\mathcal{K}} \le D_0$. The contrast is therefore an informative lower bound on paired harm whenever the fresh closed-book error is

small. Substituting estimated marginals yields an *identification interval from point estimates*; it is not a confidence interval and ignores sampling uncertainty. The same bounds hold within any subpopulation, and applying them separately within strata and summing the resulting counts never gives a wider interval than applying them to the pooled margins. Positive excess error is evidence of a preservation loss; it does not establish a passage-level causal mechanism.

### 2.4 Putting selected-set loss on the population scale

Let $w = \mathbb{P}(q \in \mathcal{K})$ and $L_{\mathcal{K}} = wE_{\mathcal{K}}$. Splitting the accuracy change over $\mathcal{K}$ and its complement gives

$$\Delta = -L_{\mathcal{K}} + (1-w)\Delta_{\mathcal{K}^c}, \qquad \Delta_{\mathcal{K}^c} = \mathbb{E}[Y_k - Y_0 \mid \mathcal{K}^c]. \tag{8}$$

$L_{\mathcal{K}}$ is the selected-set contribution to net population loss: its denominator is the full *paired* population, whereas $E_{\mathcal{K}}$ is normalized by $|\mathcal{K}|$. Neither equals the full-population harm $H$, and multiplying a selected-set rate by overall closed-book accuracy would use the wrong weight.

## 3 Evaluation Setting

**Question populations and contexts.** We evaluate subsets of HotpotQA (Yang et al., 2018), NQ-Open (Kwiatkowski et al., 2019), PopQA (Mallen et al., 2023), SituatedQA (Zhang & Choi, 2021) and StrategyQA (Geva et al., 2021). Restricting to questions with available context yields 1,000, 183, 96, 971 and 1,000 paired questions, respectively, for a total of 3,250. Results are conditional on this availability filter and are not full-benchmark scores; in particular, the small NQ-Open and PopQA subsets should not be read as benchmark rankings.

We keep the passage ordering supplied with each question and compare its top-1, top-5 and top-10 prefixes. This is a fixed-context reader evaluation: the ordering is held fixed and no method performs additional search. Top-10 is the primary condition, and the shorter prefixes serve as context-budget ablations.

**Comparison protocols.** We compare *closed-book QA* (Mallen et al., 2023) with *in-context RAG* (Ram et al., 2023), using the same reader configuration and scoring procedure. The retrieval condition supplies the top-$k$ passage prefix for $k \in \{1, 5, 10\}$; the closed-book condition supplies no passages.

The expanded comparison in Table 4 adds five methods. *NLI back-off*, adapted from Yoran et al. (2024), accepts the RAG answer only if an auxiliary NLI model (BART-large-MNLI) judges the question–answer pair to be entailed by some window of the context, and otherwise returns the closed-book answer. *SuRe* (Kim et al., 2024) generates answer candidates, summarizes the context conditioned on each, and validates and ranks the summaries. *InstructRAG-ICL* (Wei et al., 2025) uses denoising rationales as in-context demonstrations. *Astute RAG* (Wang et al., 2025) elicits internal knowledge and consolidates it with the retrieved passages over at most three iterations. *Sufficiency back-off*, adapted from Joren et al. (2025), returns the RAG answer when a separate R2 call judges the full top-10 context sufficient and the closed-book answer otherwise. Both back-off policies use a threshold of 0.5. They are full-coverage adaptations: the original designs rely, respectively, on a multi-step Self-Ask pipeline and on selective abstention.

All methods use the R2 reader, receive the same top-10 passages in their original order, and perform no additional search. Each complete inference procedure is run three times, with policy decisions made within each replicate before answer aggregation; rationales are excluded from answer matching. The fresh closed-book answers serve both as the common control and as the fallback candidates. No policy observes gold answers, screening labels or the correctness of either answer.

**Reader configurations.** We evaluate three API reader configurations: R1 is GPT-5.5 with low reasoning effort, R2 is GPT-5.6-Terra with medium reasoning effort, and R3 is GPT-5.6-Sol with medium reasoning effort. R2 is the primary reader throughout. These labels identify the API configurations used in this study. The generation settings and baseline adaptations are given in Appendix C.

Table 1: R2 with top-10 context. Accuracy columns use all $n$ paired questions; error columns use the $m$ screened-stable questions. $L_{\mathcal{K}} = (m/n)E_{\mathcal{K}}$ is their net loss contribution in percentage points of the paired population. All values are percentages or percentage points, rounded independently.

| | | | Closed-book QA (Mallen et al., 2023) | | In-context RAG (Ram et al., 2023) | | Paired contrast | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | $n$ | $m$ | $a_0$ | $D_0$ | $a_{10}$ | $D_{\text{ret}}$ | $\Delta$ | $E_{\mathcal{K}}$ | $L_{\mathcal{K}}$ |
| HotpotQA | 1000 | 579 | 57.5 | 0.9 | 85.0 | 5.7 | +27.5 | 4.8 | 2.8 |
| NQ-Open | 183 | 125 | 67.2 | 1.6 | 59.0 | 20.0 | −8.2 | 18.4 | 12.6 |
| PopQA | 96 | 62 | 64.6 | 0.0 | 53.1 | 22.6 | −11.5 | 22.6 | 14.6 |
| SituatedQA | 971 | 552 | 56.6 | 0.7 | 49.0 | 20.3 | −7.6 | 19.6 | 11.1 |
| StrategyQA | 1000 | 560 | 62.5 | 1.1 | 75.9 | 13.0 | +13.4 | 12.0 | 6.7 |
| All paired | 3250 | 1878 | 59.5 | 0.9 | 69.0 | 13.7 | +9.5 | 12.8 | 7.4 |

**Generation and scoring.** We extract the final short answer from each output and normalize it by lowercasing, removing punctuation and articles, and collapsing whitespace. The aggregated answer is the most frequent normalized answer across the three fresh evaluation calls; three-way ties go to the first output, without consulting reference answers. The aggregated answer is scored as correct if it matches a reference alias under normalized exact matching or whole-token substring matching. Empty or unparseable outputs are scored as incorrect.

The retrieval prompt permits the reader to use parametric knowledge when the evidence is incomplete or conflicting, so the contrast evaluates the full reader configuration with and without supplied context. Within-reader comparisons use that reader's own $\mathcal{K}$; differences across readers therefore combine response behavior with differences in the selected populations.

**Metrics and populations.** Let $Y_0$ and $Y_j$ be the scored aggregated answers of the fresh closed-book control and of method $j$. Overall accuracy $a_j = \mathbb{P}(Y_j = 1)$ uses all paired questions. On the screened-stable set $\mathcal{K}$, define

$$D_{\mathcal{K}}^{(j)} = \mathbb{P}(Y_j = 0 \mid \mathcal{K}), \qquad D_0 = \mathbb{P}(Y_0 = 0 \mid \mathcal{K}), \qquad E_{\mathcal{K}}^{(j)} = D_{\mathcal{K}}^{(j)} - D_0,$$

which compares the complete inference policy with the common closed-book control. For in-context RAG with top-10 context, $D_{\mathcal{K}}^{(j)} = D_{\text{ret}}$ and $E_{\mathcal{K}}^{(j)} = E_{\mathcal{K}}$, and its population loss contribution is $L_{\mathcal{K}} = (|\mathcal{K}|/n)E_{\mathcal{K}}$. Closed-book QA has zero excess error by definition. Because the control is common, $E_{\mathcal{K}}^{(j)} - E_{\mathcal{K}}^{(\ell)} = D_{\mathcal{K}}^{(j)} - D_{\mathcal{K}}^{(\ell)}$ for any two methods: subtracting the control sets the level and sign of the loss but leaves the within-set ranking unchanged. Methods can still rank differently when overall accuracy or paired harm is the objective.

**Statistical summaries.** We report point estimates for aggregate accuracy, selected-set error and between-method contrasts. The within-method confidence intervals in Table 3 are 95% percentile intervals from 2,000 question-level bootstrap replicates, stratified by dataset; each sampled question carries all context conditions for that method. The transition intervals in Table 2 instead describe the joint outcomes compatible with the integer margins. They are identification bounds on the observed sample, not confidence intervals. Appendix E distinguishes these two uses of intervals.

# 4 Results

## 4.1 Overall utility and screened-set preservation

Table 1 separates overall accuracy from preservation on the screened-stable set $\mathcal{K} = \mathcal{K}_3$, defined by correctness on all three screening calls. Retrieval raises R2's overall accuracy by 9.5 points, yet on $\mathcal{K}$ its error exceeds the fresh control by 12.8 points. Excess error is largest on NQ-Open, PopQA and SituatedQA (18.4–22.6 points), where retrieval also lowers overall accuracy, and it remains positive on HotpotQA (4.8 points), where overall accuracy rises by 27.5 points.

Table 2: Integer margins of the main R2 top-10 run and the harm $\mathbb{P}(Y_0 = 1, Y_{10} = 0)$ and benefit $\mathbb{P}(Y_0 = 0, Y_{10} = 1)$ they identify (Proposition 1), in percent of each row's $n_A$. Intervals are sharp given the listed margins; they are computed from point estimates and are not confidence intervals. The stratified row sums the $\mathcal{K}_3$ and $\mathcal{K}_3^c$ bounds in counts.

| Population | $n_A$ | Closed-book correct | RAG correct | Net (pp) | Harm (%) | Benefit (%) |
|---|---|---|---|---|---|---|
| All paired, pooled margins | 3,250 | 1,935 | 2,244 | +9.5 | [0.0, 31.0] | [9.5, 40.5] |
| $\mathcal{K}_3$ | 1,878 | 1,861 | 1,621 | −12.8 | [12.8, 13.7] | [0.0, 0.9] |
| $\mathcal{K}_3^c$ | 1,372 | 74 | 623 | +40.0 | [0.0, 5.4] | [40.0, 45.4] |
| All paired, stratified | 3,250 | 1,935 | 2,244 | +9.5 | [7.4, 10.2] | [16.9, 19.7] |

Table 3: Controlled interference on the same 600 questions. Counts are incorrect aggregated answers. Changes are relative to the unrelated condition and computed from unrounded counts; brackets give 95% paired bootstrap intervals.

| Method | Distractor | Errors/$n$ | Error (%) | Change (pp; 95% CI) |
|---|---|---|---|---|
| In-context RAG (Ram et al., 2023) | Unrelated | 25/600 | 4.2 | Reference |
| | Topical near-miss | 47/600 | 7.8 | +3.7 [2.1, 5.2] |
| | Conflicting | 111/600 | 18.5 | +14.3 [11.8, 17.0] |
| Astute RAG (Wang et al., 2025) | Unrelated | 14/600 | 2.3 | Reference |
| | Topical near-miss | 19/600 | 3.2 | +0.8 [−0.2, 2.1] |
| | Conflicting | 34/600 | 5.7 | +3.3 [1.8, 5.1] |

**Paired accounting.** For a population $A$, let $n_{ab,A}$ count questions with fresh closed-book correctness $a$ and retrieval correctness $b$. The harmful and beneficial counts are $n_{10,A}$ and $n_{01,A}$, and the net accuracy change is $(n_{01,A} - n_{10,A})/|A|$; the net change alone leaves the volume of answer turnover undetermined. Table 2 reports the integer margins of the main run together with the harm and benefit probabilities that Proposition 1 identifies from them.

On $\mathcal{K}_3$, the nearly error-free fresh control pins paired harm to [12.8, 13.7]%, so the marginal contrast recovers almost the full paired quantity. On the full paired population, the pooled margins alone constrain harm only to [0.0, 31.0]%. Applying the bounds separately to $\mathcal{K}_3$ and $\mathcal{K}_3^c$ and summing the counts narrows this interval to [7.4, 10.2]%. Although retrieval raises net accuracy by 9.5 points, the sample margins imply that at least 240 of the 3,250 questions (7.4%) change from a correct closed-book answer to an incorrect retrieval answer. This lower bound follows from count conservation and does not depend on choosing an unobserved four-cell table.

The two views answer different questions. Overall accuracy asks whether gains outweigh losses over the evaluation mixture; preservation asks whether performance falls relative to a fresh control on the screened population. The latter is a conditional comparison rather than a reweighting of the same accuracy score, and its interpretation depends on the screening rule.

### 4.2 Controlled interference with complete supporting evidence

**Experimental design.** The controlled subset contains 600 questions from $\mathcal{K}_3$: 200 from HotpotQA, 150 from StrategyQA, 150 from SituatedQA, 60 from NQ-Open and 40 from PopQA. Every condition retains a complete, audited answer-supporting bundle and varies a single distractor. The distractor is unrelated to the question, a topical near miss that mentions a plausible entity of the answer type without resolving the question, or a conflicting passage that supports an incorrect answer. Each bundle contains ten passages, and total context length and the position of the supporting bundle are held fixed across conditions. Fresh closed-book error on this subset is 3/600 (0.50%). All contrasts below are made within this subset.

**Within-method interference.** Relative to the unrelated condition, a near miss adds a net 22 errors for in-context RAG and a conflicting passage adds a net 86, increases of 3.7 and 14.3 percentage points. For Astute RAG, the net increases are 5 and 20 errors, or 0.8 and 3.3 points; its near-miss

Table 4: Inference procedures evaluated under the proposed preservation protocol, with R2 on 3,250 questions and the same 1,878-question screened set. $D_{\mathcal{K}}$ is selected-set error and $E_{\mathcal{K}} = D_{\mathcal{K}} - D_0$, with $D_0 = 17/1878 = 0.91\%$. Bold marks the best point estimate among retrieval procedures. †Fixed-context, full-coverage adaptation.

| Method | Acc. (%) ↑ | $D_{\mathcal{K}}$ (%) ↓ | $E_{\mathcal{K}}$ (pp) ↓ |
|---|---|---|---|
| Closed-book QA (Mallen et al., 2023) | 59.5 | 0.9 | 0.0 |
| In-context RAG (Ram et al., 2023) | 69.0 | 13.7 | 12.8 |
| NLI back-off† (Yoran et al., 2024) | 66.4 | 6.2 | 5.3 |
| SuRe (Kim et al., 2024) | 68.7 | 7.5 | 6.6 |
| InstructRAG-ICL (Wei et al., 2025) | 67.8 | 8.8 | 7.9 |
| Sufficiency back-off† (Joren et al., 2025) | 67.1 | 5.8 | 4.9 |
| Astute RAG (Wang et al., 2025) | **70.6** | **4.1** | **3.2** |

interval includes zero, whereas its conflict interval excludes it. Complete supporting evidence thus coexists with higher error rates when a single distractor is replaced. Each difference is a net quantity that can combine newly incorrect and newly correct answers.

**Direct comparison of interference.** Let $e_i^{j,c} = 1 - Y_i^{j,c}$ be the error indicator of method $j$ on question $i$ under context condition $c \in \{\mathrm{unr}, \mathrm{con}\}$ (unrelated or conflicting). The direct comparison uses all four outcomes of each question:

$$z_i = (e_i^{\mathrm{RAG,con}} - e_i^{\mathrm{RAG,unr}}) - (e_i^{\mathrm{Astute,con}} - e_i^{\mathrm{Astute,unr}}), \qquad \widehat{\delta} = \frac{1}{600}\sum_i z_i. \tag{9}$$

The conflict-induced error increase is 14.3 points for in-context RAG and 3.3 points for Astute RAG, giving an observed difference of $\widehat{\delta} = (86 - 20)/600 = 11.0$ percentage points. This descriptive contrast compares complete inference pipelines under the matched intervention, including their different prompting and inference computation. The 11.0-point contrast is a point estimate; the within-method intervals in Table 3 do not provide an interval for this difference.

**Directional changes and dataset composition.** For method $j$, let $h_j$ count questions answered correctly with the unrelated distractor and incorrectly with the conflicting one, and let $g_j$ count the reverse. The margins give $h_{\mathrm{RAG}} - g_{\mathrm{RAG}} = 86$ and $h_{\mathrm{Astute}} - g_{\mathrm{Astute}} = 20$. Because $g_j$ cannot exceed the error count in the unrelated condition, they also bound $h_{\mathrm{RAG}} \in [86, 111]$ and $h_{\mathrm{Astute}} \in [20, 34]$: at least 86 questions change from correct to incorrect between the unrelated and conflicting conditions for in-context RAG. The pooled estimates weight datasets by their share of the subset, with HotpotQA contributing one third. The target is this five-dataset mixture. Appendix D gives the corresponding directional bounds and explains the distinction between this pooled target and an equal-dataset average.

### 4.3 Preservation across inference procedures

Table 4 applies our preservation protocol to seven inference conditions on the common R2 population. Every method is evaluated against the same fresh closed-book answers. The common control determines the size and sign of excess error; it leaves the ordering by selected-set error unchanged.

All five mitigation procedures reduce excess error relative to in-context RAG, from 12.8 to 3.2–7.9 points. Astute RAG has both the highest overall accuracy (70.6%) and the lowest selected-set excess error (3.2 points) among the retrieval procedures in this run. The two back-off policies reduce excess error to 5.3 and 4.9 points, with overall accuracies of 66.4% and 67.1%, respectively. These are comparisons of observed point estimates; small gaps between methods are not interpreted as established population rankings.

The methods differ in prompting, intermediate calls and information processing. Their scores measure the performance of complete inference procedures under common external evidence. In particular, the Astute comparison includes internal-knowledge elicitation and consolidation, so it does not separate

Table 5: Reader and context-budget sensitivity. Accuracy columns use the 3,250 paired questions. Excess error uses each reader's own screened-stable set; the R2 budget rows share its 1,878-question set. Readers are defined in Section 3.

| Configuration | Closed-book QA (Mallen et al., 2023) $a_0$ (%) | In-context RAG (Ram et al., 2023) $a_k$ (%) | $\Delta$ (pp) | $E_{\mathcal{K}}$ (pp) |
|---|---|---|---|---|
| R1, top-10 | 56.6 | 51.6 | $-5.0$ | 24.1 |
| R2, top-10 | 59.5 | 69.0 | +9.5 | 12.8 |
| R3, top-10 | 49.7 | 56.6 | +6.9 | 9.9 |
| R2, top-1 | 59.5 | 32.6 | $-26.9$ | 52.1 |
| R2, top-5 | 59.5 | 58.1 | $-1.4$ | 23.7 |
| R2, top-10 | 59.5 | 69.0 | +9.5 | 12.8 |

Table 6: Closest evaluation settings and the focus of this protocol. The entries summarize the cited analyses and do not assert that those works exclude other evaluations.

| Prior work | Evidence already provided | Focus of this protocol |
|---|---|---|
| Yoran et al. (2024) | Closed-book and noisy-context comparisons; NLI back-off and robustness training | Independent screening and fresh closed-book re-evaluation |
| Joren et al. (2025) | Context-sufficiency strata, closed-book comparisons and selective generation | Preservation conditional on independently screened success |
| Wang et al. (2025) | Joint LLM/RAG correctness categories; conflict mitigation and token/API-call costs | Preservation, paired transitions and utility on a common screened population |
| Amiraz et al. (2025) | Fixed gold evidence with weak/hard distractors, compared across passage orders | Matched unrelated, near-miss and conflicting contexts with support held fixed |

the contribution of these stages from additional inference computation. A back-off policy also reads context before returning the closed-book answer; preservation and retrieval-cost savings are distinct objectives.

### 4.4 Reader and context-budget sensitivity

Table 5 varies the reader and the context budget. All three readers show substantial preservation loss at top-10 (9.9 to 24.1 points), including R3, whose overall accuracy rises by 6.9 points. Within R2, shorter prefixes are markedly worse: with top-1 context, overall accuracy falls by 26.9 points and excess error reaches 52.1 points. The budget comparison holds R2's selected population fixed, but a longer prefix changes both length and content, so it does not isolate a pure length effect. Cross-reader comparisons also change the screened population, so their conditional values are not a controlled capability ranking; a common-question analysis and a reader-specific analysis answer different questions.

## 5 Relation to Prior Work

Table 6 situates the protocol among the most closely related analyses. Existing work already includes closed-book comparisons, joint correctness categories, controlled distractors and inference-cost measurements. Our protocol organizes the preservation measurements around independently screened success and a fresh matched control, with exact accounting for what marginal results identify.

Relevance, sufficiency and knowledge conflict describe different properties of retrieved material (Joren et al., 2025; Longpre et al., 2021; Neeman et al., 2023; Xie et al., 2024; Wu et al., 2024). Our preservation contrast instead measures a change in scored behavior relative to a fresh control on an independently selected population. Context quality and reader preservation can therefore be assessed together without identifying one with the other.

Adaptive retrieval decides whether or when to retrieve using question properties, uncertainty or self-reflection (Mallen et al., 2023; Jiang et al., 2023; Jeong et al., 2024; Asai et al., 2024). Robust methods filter or reconcile evidence (Yoran et al., 2024; Yu et al., 2024; Kim et al., 2024; Wei et al., 2025; Wang et al., 2025). Our protocol evaluates the resulting complete policies. It complements the relevance, faithfulness and unsupported-generation measures of RAGAs, ARES and RAGTruth (Es et al., 2024; Saad-Falcon et al., 2024; Niu et al., 2024) by specifying the screening population, the fresh reference and the paired answer changes.

## 6 Implications and Limitations

**Evaluating a retrieval gate.** For a policy $g \in \{0, 1\}$ that returns the retrieval answer when $g = 1$ and the closed-book answer otherwise, the improvement over always retrieving is

$$\mathbb{E}[Y_g - Y_k] = \mathbb{P}(g = 0, Y_0 = 1, Y_k = 0) - \mathbb{P}(g = 0, Y_0 = 0, Y_k = 1), \tag{10}$$

since $Y_g - Y_k = (1 - g)(Y_0 - Y_k)$. The first term is harm avoided and the second is benefit forgone; both require a common population and a specified paired evaluation. Signals that use gold-answer presence or the correctness of both arms are audit tools, not deployable gates. A policy comparison should use held-out questions, a declared information budget and the full trade-off in Equation 10. This identity applies to routing between two final answer candidates. In the replicated back-off evaluations above, routing precedes answer aggregation, and each resulting policy is scored directly.

**Scope of the evidence.** The evaluation concerns fixed external contexts and reader behavior, rather than the quality of a newly trained retriever. Context availability and unequal dataset sizes determine the evaluation mixture; reader-specific screening also changes the conditional population. The controlled study covers a curated subset with audited support, and its pooled effect need not be uniform across datasets. The methods differ in prompting and computation, so the comparison does not identify a single component responsible for preservation. Finally, answer matching can reject valid paraphrases or accept spurious substring matches. The findings are expressed in terms of the declared answer-scoring procedure and the evaluated question populations.

## 7 Conclusion

Answer preservation is a conditional evaluation target: a screening rule selects the question population, and a fresh closed-book control supplies its reference performance. The resulting contrast separates selected-set loss from residual baseline error. On 3,250 questions, retrieval improves R2's overall accuracy by 9.5 points while raising its error on the screened-stable set by 12.8 points. Count conservation implies at least 240 correct-to-incorrect changes despite the positive net gain. The matched 600-question experiment shows larger error rates under conflicting content even when complete support remains present, with a smaller observed increase for the Astute RAG pipeline. Explicit populations, fresh controls and transition bounds make these distinct aspects of RAG performance visible.

## AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, consistency checks of reported quantities and statistical interpretations, and LaTeX preparation. No new model experiments were conducted as part of this manuscript revision. The authors are responsible for verifying the reported experiments, mathematical statements and references, and for the final content of the paper.

## REFERENCES

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258, 2025. doi: 10.18653/v1/2025.acl-long.892.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 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*, 2024. doi: 10.1145/3626772.3657834.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7036–7050, 2024.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. SuRe: Summarizing retrievals using answer candidates for open-domain QA of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2404.13081.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, 2023.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Scharli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 31210–31227, 2023.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, 2025. doi: 10.18653/v1/2025.acl-long.1476.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P1qhkp8gQT.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*, 2024.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14672–14685, 2024.

Michael Zhang and Eunsol Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, 2021. doi: 10.18653/v1/2021.emnlp-main.586.

# A Derivations

## A.1 Harm, benefit and the marginal contrast

Condition on any population, including $\mathcal{K}$, and write $p_{ab} = \mathbb{P}(Y_0 = a, Y_k = b)$ for $a, b \in \{0, 1\}$. The marginal errors satisfy $D_{\text{ret}} = p_{10} + p_{00}$ and $D_0 = p_{01} + p_{00}$. Hence $E_{\mathcal{K}} = p_{10} - p_{01} = H_{\mathcal{K}} - G_{\mathcal{K}}$. With $t = p_{00}$, nonnegativity of the four cells requires

$$\max(0, D_{\text{ret}} + D_0 - 1) \le t \le \min(D_{\text{ret}}, D_0).$$

Substituting $H_{\mathcal{K}} = D_{\text{ret}} - t$ gives

$$\max(0, D_{\text{ret}} - D_0) \le H_{\mathcal{K}} \le \min(D_{\text{ret}}, 1 - D_0).$$

For every $t$ in the displayed interval, the cells $(p_{00}, p_{10}, p_{01}, p_{11}) = (t, D_{\text{ret}} - t, D_0 - t, 1 - D_{\text{ret}} - D_0 + t)$ define a valid joint distribution, so the bounds are sharp given only the two pooled marginals. Question-level conditional marginals or a particular coupling can further restrict the joint distribution. If $E_{\mathcal{K}} \ge 0$, then $H_{\mathcal{K}} - E_{\mathcal{K}} = G_{\mathcal{K}} \le D_0$, which establishes the stated bound on the interval width.

**Stratified bounds.** Let the population be partitioned into strata $A_1, \ldots, A_S$ with harm counts $n_{10,A_s}$. Applying the bounds within each stratum and summing gives $\sum_s \max(0, x_s) \le n_{10} \le \sum_s \min(u_s, v_s)$, where $x_s$, $u_s$ and $v_s$ are the stratum-level count analogues of $D_{\text{ret}} - D_0$, $D_{\text{ret}}$ and $1 - D_0$. Because $\sum_s \max(0, x_s) \ge \max(0, \sum_s x_s)$ and $\sum_s \min(u_s, v_s) \le \min(\sum_s u_s, \sum_s v_s)$, the stratified interval is never wider than the pooled one. It is sharp given the stratum margins, since the within-stratum extremal tables can be chosen independently.

## A.2 The distribution induced by screening

**Proposition 2** (Screening weights)**.** *Let questions follow $P$. Conditional on $q$, suppose that $s$ screening calls are independent Bernoulli trials with success probability $p_q$ and that evaluation is independent of those calls. Let $b_q = \mathbb{E}[Y_0 \mid q]$ and $r_q = \mathbb{E}[Y_k \mid q]$ denote the accuracies of the declared evaluation procedure. If $\mathbb{E}_P[p_q^s] > 0$, then Equation 6 holds.*

*Proof.* Conditional independence gives $\mathbb{P}(q \in \mathcal{K}_s \mid q) = p_q^s$. Bayes' rule therefore reweights $P(dq)$ by $p_q^s / \mathbb{E}_P[p_q^s]$. Because evaluation is independent of screening, the conditional evaluation contrast remains $b_q - r_q$. Averaging under the reweighted distribution yields Equation 6. □

For a single fresh evaluation call, $b_q = p_q$. If the evaluation score instead declares success whenever at least two of three independent calls are gold-correct, then $b_q = 3p_q^2 - 2p_q^3$. This gold-based majority statistic differs from majority voting over generated answer strings, which must specify normalization, ties and answer selection without consulting the gold answer. Neither procedure should be silently substituted for the other.

An oracle that sees both scored outcomes could achieve $\mathbb{E}[\max(Y_0, Y_k)] = a_k + H$. Thus $H$ is the oracle improvement over always using the retrieval answer *for this specified pairing*. Under stochastic generation, this quantity depends on the joint sampling scheme. It is neither an intrinsic probability that a particular question will be harmed nor a gain available to a gate without access to correctness labels.

## A.3 Population accounting and gating

By the law of total expectation, $\Delta = w\mathbb{E}[Y_k - Y_0 \mid \mathcal{K}] + (1-w)\mathbb{E}[Y_k - Y_0 \mid \mathcal{K}^c]$. The first conditional expectation equals $-E_{\mathcal{K}}$, which proves Equation 8. For a gate, define $Y_g = gY_k + (1-g)Y_0$. Then $Y_g - Y_k = (1-g)(Y_0 - Y_k)$, which equals $+1$ on a declined harmful pair, $-1$ on a declined beneficial pair and zero otherwise. Taking expectations proves Equation 10. These identities hold regardless of how often the gate declines retrieval; deployment feasibility depends separately on the information used to compute $g$.

## B INTEGER ACCOUNTING

Table 7 gives the unique integer margins compatible with the reported sample sizes and one-decimal percentages in Table 1. These are marginal counts; reconstructing them does not reconstruct the question-level joint outcomes. All primary preservation results use $\mathcal{K} = \mathcal{K}_3$.

Table 7: Integer margins underlying Table 1. $C_0$ and $C_{10}$ are correct-answer counts on all $n$ questions. $F_0$ and $F_{10}$ are error counts on the $m$ screened-stable questions.

| Dataset | $n$ | $m$ | $C_0$ | $C_{10}$ | $F_0$ | $F_{10}$ |
|---|---|---|---|---|---|---|
| HotpotQA | 1000 | 579 | 575 | 850 | 5 | 33 |
| NQ-Open | 183 | 125 | 123 | 108 | 2 | 25 |
| PopQA | 96 | 62 | 62 | 51 | 0 | 14 |
| SituatedQA | 971 | 552 | 550 | 476 | 4 | 112 |
| StrategyQA | 1000 | 560 | 625 | 759 | 6 | 73 |
| Total | 3250 | 1878 | 1935 | 2244 | 17 | 257 |

The selected-set successes are $1878 - 17 = 1861$ for closed-book QA and $1878 - 257 = 1621$ for retrieval. The complement therefore has $1935 - 1861 = 74$ and $2244 - 1621 = 623$ successes, respectively. The principal rates follow as

$$\Delta = \frac{309}{3250}, \qquad E_{\mathcal{K}} = \frac{240}{1878}, \qquad L_{\mathcal{K}} = \frac{240}{3250}.$$

Multiplying by 100 gives 9.5077, 12.7796 and 7.3846 percentage points. All displayed percentages are rounded after the count arithmetic.

**Feasible paired tables.** Let $h = n_{10,\mathcal{K}}$ and $u = n_{10,\mathcal{K}^c}$. Every joint table compatible with these two sets of margins has

$$240 \le h \le 257, \qquad 0 \le u \le 74,$$

with integer $h, u$, and the following cells:

| Population | $(1,1)$ | $(1,0)$ | $(0,1)$ | $(0,0)$ |
|---|---|---|---|---|
| $\mathcal{K}$ | $1861 - h$ | $h$ | $h - 240$ | $257 - h$ |
| $\mathcal{K}^c$ | $74 - u$ | $u$ | $549 + u$ | $749 - u$ |
| All paired | $1935 - h - u$ | $h + u$ | $309 + h + u$ | $1006 - h - u$ |

Consequently, full-population harm lies between 240 and 331 questions, and benefit lies between 549 and 640 questions. These ranges produce the stratified row of Table 2. They specify feasible counts, not an estimated allocation of questions to the four cells.

## C PROTOCOL FOR THE INFERENCE COMPARISON

**Common conditions.** The seven conditions in Table 4 use the same R2 reader, 3,250 questions and 1,878-question screening set. All retrieval procedures receive the same ordered top-10 external passages and perform no additional search. Each complete procedure is repeated three times. Policy decisions are made within each replicate; the resulting short answers are then aggregated using the rule in Section 3. Fresh closed-book outputs provide both the shared control and the answer candidates used by the back-off policies. No policy receives reference answers, screening labels or answer-correctness labels.

**Reader and generation settings.** R1 uses GPT-5.5 with low reasoning effort; R2 uses GPT-5.6-Terra with medium reasoning effort; R3 uses GPT-5.6-Sol with medium reasoning effort. The specified generation settings are temperature 0.0, top-$p$ 1.0, inference seed 42 and a maximum output budget of 32,768 tokens including reasoning. Up to two additional attempts follow timeouts or parsing failures, with 1-second and 2-second waits. Empty or unparseable terminal outputs are scored as incorrect. The three calls are operational repeats under this configuration; the conditional-independence calculation in Proposition 2 states an additional model assumption.

**Answer aggregation.** The scoring unit is one question after answer aggregation. Each output contributes its final short answer, excluding any rationale. Normalization lowercases the answer, removes punctuation and articles, and collapses whitespace. The most frequent normalized string among the three evaluation answers is selected; a three-way tie selects the first output. The selected answer is then checked against reference aliases by normalized exact matching or whole-token substring matching. Voting occurs before gold-answer scoring. Repeated calls therefore do not increase the number of question-level observations.

**NLI back-off (adapted).** Following the entailment-based back-off principle of Yoran et al. (2024), BART-large-MNLI checks the question paired with the generated RAG answer against windows of the context. Ordered context is split into windows of 384 NLI tokens with a 64-token overlap, reserving at most 128 tokens for the question and answer. The rule accepts the RAG answer if any window has entailment probability at least 0.5, and otherwise returns the corresponding fresh closed-book answer. This is an answer-level, fixed-context adaptation of the original multi-step Self-Ask procedure.

**SuRe.** The procedure follows candidate generation, candidate-conditioned context summarization, validation and pairwise ranking from Kim et al. (2024), using two answer candidates. Each pair is ranked in both orders. Selection uses the highest sum of validity and ranking scores, with ties resolved by candidate order. R2 performs all stages on the common evidence pool.

**InstructRAG-ICL.** The in-context variant of Wei et al. (2025) uses R2-generated denoising rationales from labeled development examples. Two fixed question–rationale demonstrations are included at inference. Development answers enter rationale synthesis, while evaluation answers are withheld and reader weights remain fixed.

**Astute RAG.** The procedure follows internal-knowledge elicitation, source-aware consolidation and answer selection from Wang et al. (2025). It elicits at most one internal passage and uses at most three iterations. The ten external passages are unchanged. Internally generated text is a reader output rather than an additional retrieved passage. All stages use R2.

**Sufficiency back-off (adapted).** A separate R2 call applies the context-sufficiency criterion of Joren et al. (2025) to the question and complete top-10 bundle. A score of at least 0.5 routes to the in-context RAG answer; otherwise the policy returns the corresponding fresh closed-book answer. This is a full-coverage routing adaptation of a criterion originally used for selective abstention.

**Metric denominators.** Overall accuracy uses all 3,250 questions. Selected-set error uses the fixed 1,878-question set, and excess error subtracts its common fresh closed-book error without clipping negative values. The comparison concerns complete inference procedures. Intermediate calls, routing decisions and internal passages are parts of those procedures, rather than independent test cases.

## D Controlled Evidence and Directional Changes

The controlled subset consists of 200 HotpotQA, 150 StrategyQA, 150 SituatedQA, 60 NQ-Open and 40 PopQA questions from $\mathcal{K}_3$. The same 600 question identities are used for all conditions and both methods. Each input contains ten passages and retains a complete, audited supporting bundle. The context variants replace one distractor with unrelated, topical near-miss or conflicting material, while holding the position of the supporting evidence and total context length fixed. The contrast therefore concerns distractor content within these constructed bundles. Figure 1 provides an illustrative example; its passages are not examples sampled from the experiment.

**Directional bounds.** For each method, let $h$ count correct-to-incorrect changes from the unrelated to the conflicting condition, and let $g$ count the reverse. If the two conditions have error counts $e_{\mathrm{unr}}$ and $e_{\mathrm{con}}$, then $h - g = e_{\mathrm{con}} - e_{\mathrm{unr}}$ and $0 \le g \le e_{\mathrm{unr}}$. This yields Table 8.

The fresh closed-book error count on this subset is 3. Subtracting the common $3/600$ control from each raw error rate changes the level of each preservation contrast, but cancels in every within-method context increment and in the 11.0-point between-method increment difference. Thus the controlled

Table 8: Directional count bounds for unrelated-to-conflicting context changes. These are sharp ranges implied by the marginal error counts; $h - g$ is fixed within each row. They are not confidence intervals.

| Method | Correct → incorrect ($h$) | Incorrect → correct ($g$) | $h - g$ |
|---|---|---|---|
| In-context RAG | $[86, 111]$ | $[0, 25]$ | 86 |
| Astute RAG | $[20, 34]$ | $[0, 14]$ | 20 |

comparison and the preservation definition use compatible references without substituting the full-set control $17/1878$ for the subset control.

**Dataset weighting.** The pooled contrast uses dataset weights $(200, 150, 150, 60, 40)/600$ in the order above. An equal-dataset contrast would instead average the five dataset-specific increments with weight $1/5$. The pooled counts identify the former point estimate. They do not identify the latter or the allocation of the effect across datasets.

## E Statistical Interpretation

**Identification versus uncertainty.** The transition bounds in Tables 2 and 8 describe the observed sample: all paired tables with the stated margins lie inside the bounds, and their endpoints are attainable. Sampling uncertainty is a different question. In particular, a margin-compatible interval is not a 95% confidence interval for a population transition probability.

For $m > 1$ i.i.d. question-level records from a population induced by a fixed screening procedure, write $Z_i = Y_{0i} - Y_{ki} \in \{-1, 0, 1\}$. Then

$$\widehat{E_{\mathcal{K}}} = \bar{Z}, \qquad \widehat{\mathrm{SE}}(\bar{Z}) = \sqrt{\frac{1}{m(m-1)} \sum_{i=1}^{m} (Z_i - \bar{Z})^2}.$$

The covariance between arms is part of this standard error. Marginal totals identify the mean contrast but not that covariance. Resampling questions must carry all relevant outcomes together; three answers to one question do not constitute three independent question replicates.

**Within-method context intervals.** The 95% intervals in Table 3 concern each method's near-miss-minus-unrelated and conflicting-minus-unrelated error contrast. Question-level bootstrap resampling is stratified by dataset and retains each dataset's original sample size. All context outcomes of a sampled question travel together. Percentile endpoints are computed from the 2,000 resampled contrasts. This conditions on the observed screening set, constructed inputs and declared aggregation rule.

**Direct between-method contrast.** The point difference is

$$\widehat{\delta} = \frac{(111 - 25) - (34 - 14)}{600} = 0.11.$$

Its paired uncertainty depends on the joint distribution of all four error indicators in Equation 9. Subtracting endpoints of the two within-method confidence intervals would not produce a confidence interval for $\delta$. We report the direct difference as a point estimate.

**Arithmetic verification.** The accompanying aggregate-checking script recomputes the main rates, count-conservation identities and sharp transition bounds from the integer margins, and recomputes the controlled point contrasts from the six error counts. This verifies the reported arithmetic; it does not regenerate model outputs or independently recompute the reported bootstrap intervals.