From Evidence Scores to RAG Answers: Budget, Search, and Presentation

Anonymous Authors

Anonymous Institution

Abstract. Evidence selection in retrieval-augmented generation determines how much context a model receives, which passages it uses, and their presentation. Endpoint comparisons can conceal large, opposing answer changes: reversing a selected set can change a correct answer, while reselection restores it. We expose these changes through two controlled paths over fixed six-passage pools. The first improves combination search at three passages and then permits smaller sets; the second reverses a selected set and then selects again under reversed reference-answer likelihood. Complete answer tables supply every feasible intermediate response and cross each selected set with both presentation orders. Across 539 multi-hop questions and three generators, budget flexibility accounts for 4.79 of Mistral's 5.91-point token-F1 gain over three-step greedy selection. On 30 matched questions, Gemma loses 20.56 points under fixed-set reversal and recovers 16.54 through exact-three reselection; the same membership change lowers F1 by 23.35 under the original order. A shared cross-encoder selects identical greedy and exact sets in 75 of 89 pools. These findings reveal model-specific budget effects and order-dependent reselection gains, showing how intermediate answers explain the behavior of evidence selectors.

Keywords: Retrieval-augmented generation · Evidence selection · Context construction · Empirical evaluation

1 Introduction

Retrieval-augmented generation (RAG) turns retrieved passages into a context for answering a question [13]. Building that context involves three decisions: how many passages to include, which passages to combine, and how to present them. A final answer score summarizes their combined effect. Intermediate answers reveal which decision changed the response, helping locate whether a gain comes from budget choice, combination search, or presentation.

Consider a question about the band founded by Devilment's stable vocalist. Qwen answers *Cradle of Filth* from a selected two-passage set. Reversing those passages changes the answer to *Falling in Reverse*; selecting another set under reversed-order likelihood restores *Cradle of Filth*. The endpoints agree, while the intermediate answer changes completely. This example captures our central observation: large, opposing answer changes can be hidden inside a small endpoint difference.

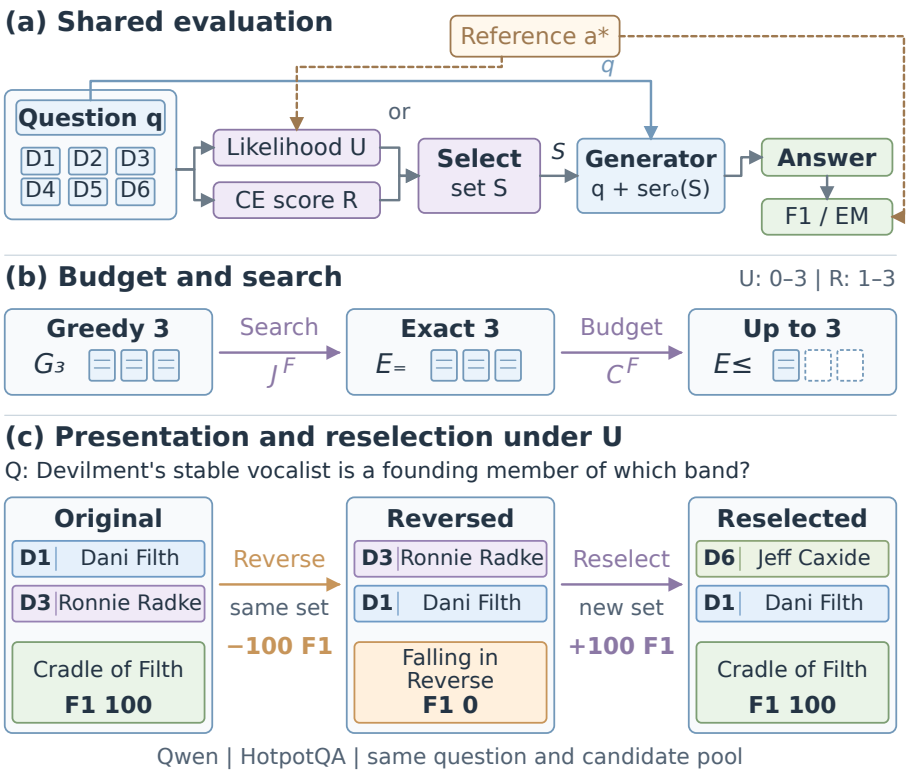


Fig. 1. Evidence-selection paths reveal answer changes hidden by their endpoints. (a) A question and a fixed passage pool pass through scoring, selection, and generation; reference answers support likelihood scoring and answer evaluation. (b) The budget-and-search path compares greedy and exact three-passage selections, then permits smaller sets under the same score. (c) In the Qwen example, reversing the selected passages changes *Cradle of Filth* to *Falling in Reverse*; reselection restores *Cradle of Filth*. The three responses score 100, 0, and 100 token F1.

Existing research has established that retrieved relevance, evidence utility, and information position each influence generation [18,14,9]. Set-level selectors increasingly address the information supplied by passages together [12,11]. These developments motivate a specific question: *where along an evidence-selection decision does the generated answer change?* A higher final score can come from a better combination, a different set size, or a more favorable presentation. An intermediate comparison isolates each change by retaining the same question, candidate pool, and scoring objective.

We study these objects in fixed pools of six passages. Building on complete evidence-utility evaluation [5], we first compare three-step greedy selection with the exact three-passage optimum under reference-answer likelihood. We then select the best set of up to three passages. A second path holds the selected

set fixed while reversing its presentation, then replaces it with the set selected under the reversed-order score. Complete generation on matched question subsets supplies every feasible intermediate answer. It also lets us present both the original and reselected sets in both orders, revealing whether a membership change helps under each presentation. Figure 1 shows the comparison paths. A public cross-encoder extends the budget-and-search path to a scorer that reads the question and candidate evidence.

The experiments reveal three complementary patterns. Across 539 questions, budget flexibility accounts for 4.79 of Mistral's 5.91-point token-F1 advantage over three-step greedy selection, concentrated in one- and two-passage contexts. On 30 matched questions, Gemma loses 20.56 points when its exact-three selected set is reversed and recovers 16.54 through reselection. Presenting that reselected set in the original order instead lowers F1 by 23.35 points, exposing a strong dependence on presentation. Finally, greedy and exact cross-encoder search choose the same set on 75 of 89 retrieval-constructed pools, concentrating answer changes in the remaining cases.

Our study makes three contributions. First, two matched comparison paths connect evidence decisions to their intermediate answers, locating model-specific budget effects and opposing presentation and reselection effects. Second, complete answer tables expose higher-scoring contexts beyond the selected-set union and enable a crossed comparison of membership and presentation. Third, a shared cross-encoder analysis links additional score optimization to set and answer changes, showing where greedy and exact search converge and where their responses diverge.

2 Related Work

Evidence selection for generation. RAG systems combine retrieval with generative inference [13,8,17]. The resulting context depends on both individual relevance and the information supplied jointly. Maximal marginal relevance (MMR) balances query relevance with novelty relative to selected passages [2]. SetR reasons about a query's information needs to select a supporting set [12], while OptiSet learns set preferences and jointly optimizes selection and set-level ranking [11]. CARROT optimizes chunk combinations and ordering under a cost constraint, using utility to select contexts that may leave part of the budget unused [20]. We examine the answers produced as a shared score is optimized first at a fixed cardinality and then over a flexible cardinality budget.

Utility and generated answers. Tian et al. [18] find a small positive relationship between retrieved relevance and generation utility, with weaker correlation at larger context sizes. CUE-R studies individual evidence items through removal, replacement, and duplication, tracking changes in answers and evidence-use signals [9]. Seo and Jang examine how similarity and evidence signals weight document-conditioned answers during aggregation [16]. We follow adjacent selection states under a shared objective: greedy to exact search at the same size,

then exact search with a flexible budget. Their generated answers locate which decision changes the response.

Target feedback and complete observations. Complex Utility, Simple Calibration [5] uses exhaustive likelihood tables to study the target probes needed for an effective evidence decision. Its singleton and adaptive-pair methods access selected entries, and its answer evaluation follows the sets chosen by those methods. Building on this foundation, we study the answers along fixed-size search, budget release, and presentation/reselection paths. Complete-generation cohorts extend the observed answer space to every feasible subset, including contexts outside the selected-set union and both presentations of each reselected set.

Passage presentation. The position of relevant information affects how language models use their input context [14]. We examine presentation together with the selection rule: a set chosen under one order can be evaluated in another order, and the selector can then choose again under that presentation. Crossing both selected sets with both orders connects presentation sensitivity to the evidence chosen under each scoring condition.

3 Tracing Evidence Decisions

3.1 Scores and Answers on the Same Candidate Pool

Fix a question q, a generator g, and six saved passages $D_q = (d_1, \ldots, d_6)$. A selected set $S \subseteq D_q$ is serialized in a specified order o. The canonical order c follows the saved candidate positions; the reversed order r traverses those positions backwards. Passage identifiers and text stay attached to their documents in both orders.

We pair each context with two observations. The first is the generator's average log-likelihood of the reference answer $y_{1:T}$:

$$L_o(S) = \frac{1}{T} \sum_{t=1}^{T} \log p_g(y_t \mid q, \mathrm{ser}_o(S), y_{<t}), \qquad U_o(S) = L_o(S) - L_o(\emptyset). \quad (1)$$

The second is the answer obtained by greedy decoding from the question and the same serialized evidence. We write $F_o(S)$ for its normalized token F1 against the accepted answer dictionary and also record exact match (EM). The common subset identity joins the score, selected passages, and generated response.

With six passages, all 64 likelihood values can be evaluated directly. A maximum budget of three passages contains 42 feasible subsets, including the empty set. On the complete-generation cohorts, evaluating every feasible answer makes it possible to inspect the response at any selected intermediate state.

3.2 Combination Search and Budget Choice

Suppose a selector takes three passages. It can change the combination while keeping that size, or choose a smaller set under the same maximum budget. We evaluate these two changes in sequence.

Let G_3 be the result of three greedy additions under canonical utility U_c. At each step, greedy selects the available passage with the largest marginal gain

$$\Delta_i(S) = U_c(S \cup \{d_i\}) - U_c(S).$$

Define the exact selections

$$E_= = \operatorname*{arg\,max}_{S:\,|S|=3} U_c(S), \qquad E_\leq = \operatorname*{arg\,max}_{S:\,|S|\leq 3} U_c(S). \tag{2}$$

Exact ties use the smallest subset bitmask, and greedy ties use the lowest candidate index. The path $G_3 \to E_= \to E_\leq$ first improves the combination at size three, then allows the selected size to vary. The corresponding score differences are $U_c(E_=) - U_c(G_3)$ and $U_c(E_\leq) - U_c(E_=)$.

We follow the generated answers along this path:

$$J^F = F_c(E_=) - F_c(G_3), \qquad \text{same-size search,} \tag{3}$$

$$C^F = F_c(E_\leq) - F_c(E_=), \qquad \text{budget release.} \tag{4}$$

Their sum is the endpoint difference $F_c(E_\leq) - F_c(G_3)$. The intermediate three-passage optimum makes the two comparisons observable separately. In the results, F1 and EM are expressed on a 0–100 scale, so the contrasts are reported in score points.

Complete answer generation also gives the best feasible answer score,

$$F^* = \max_{S:\,|S|\leq 3} F_c(S), \qquad R^F(\pi) = F^* - F_c(S_\pi). \tag{5}$$

Here S_π is the set chosen by a selection rule π. The gap R^F locates its answer within the observed feasible answer space. For a family of selection rules Π, we also compare F^* with $\max_{\pi\in\Pi} F_c(S_\pi)$. Their difference measures the additional best-answer score exposed by complete generation beyond that family's selected sets.

3.3 Presentation and Reselection

The same evidence can produce different answers when its order changes. A selector also sees a different utility table when the generator scores that reversed context. We therefore evaluate the original set under reversal before selecting again.

For a selection rule π, let S_c^π and S_r^π be its choices under canonical and reversed utility, respectively. We evaluate

$$F_c(S_c^\pi), \qquad F_r(S_c^\pi), \qquad F_r(S_r^\pi).$$

The adjacent changes are

$$\Delta^F_{\text{pres}}(\pi) = F_r(S^\pi_c) - F_c(S^\pi_c), \tag{6}$$

$$\Delta^F_{\text{resel}}(\pi) = F_r(S^\pi_r) - F_r(S^\pi_c). \tag{7}$$

Their sum is the difference between the two selected endpoints. Presentation holds membership fixed. Reselection holds the reversed presentation fixed and changes the chosen membership. We apply this path to $E_{\leq}$, $E_{=}$, and G_3, preserving the same rule across the two orders.

We complete the comparison with $F_c(S^\pi_r)$: the reverse-selected set presented in canonical order. The same membership change can now be evaluated under both presentations. Its crossed answer difference is

$$I^F(\pi) = \big[F_r(S^\pi_r) - F_r(S^\pi_c)\big] - \big[F_c(S^\pi_r) - F_c(S^\pi_c)\big]. \tag{8}$$

A positive value means that the answer-score change from replacing S^π_c with S^π_r is larger under reversed presentation. All four readouts share the same question, passage pool, generator, and selection rule.

3.4 A Shared Cross-Encoder

To apply the search and budget comparisons to a public relevance scorer, we use BGE reranker v2-m3 [1]. It takes the question and serialized evidence as its two inputs. Let $R(q, S)$ be its raw logit on a canonically ordered set. We score all 41 nonempty sets of sizes one through three and compare five rules:

- **Singleton top-3** selects the three highest singleton logits.
- **MMR-3** combines singleton relevance with the novelty of each addition [2].
- **Joint greedy-3** starts with the best singleton and adds the passage producing the highest joint score at each step.
- **Joint exact-3** maximizes R over all three-passage sets.
- **Joint best-1-to-3** maximizes R over all nonempty feasible sets.

The last three rules use the same joint score. Their selected answers therefore support the same sequence of comparisons: combination search at size three, followed by release of the size constraint. We also record whether each comparison changes the selected set and whether the generator then changes its answer.

4 Experimental Design

Questions and candidate pools. We study 539 development questions from three multi-hop benchmarks: 180 from 2WikiMultiHopQA [7], 180 from HotpotQA distractor [22], and 179 from answerable MuSiQue [19]. Each question has two six-passage constructions from its benchmark-local candidates. Construction A starts from the annotated supporting passages and fills the remaining positions with high-scoring distractors. Construction B uses reciprocal-rank fusion of BM25 and BGE ranks, with constant 60 [15,4,3]. Both use a fixed hash-based candidate order. Passage text is capped once at 256 tokens with the Qwen tokenizer, and the saved text is shared across models and subsets.

Table 1. Matched question groups support progressively richer answer observations. Each group uses both candidate constructions. The order and cross-encoder studies use Qwen and Gemma.

Study	Questions	Available answer observations
Budget and search	539	G_3, $E_=$, and $E_\leq$ for three generators
Complete canonical	89	All 64 subsets for Qwen and Gemma
Order and reselection	30	All 42 feasible answers in both orders; crossed selected sets
Shared cross-encoder	89	Answers for every set chosen by the five rules

Generators and prompts. The models are Qwen2.5-7B-Instruct, Gemma-3-4B-it, and Mistral-7B-Instruct-v0.3 [21,6,10]. The prompt asks for the shortest correct answer span from the supplied evidence. Each model uses its chat template, BF16 weights, and greedy decoding with a 64-token output limit. Likelihood uses float32 log-softmax and the reference-answer tokens. Model revisions, full prompts, decoding settings, and the saved passage text accompany the research materials.

Complete and matched observations. Table 1 summarizes the question groups. The complete canonical cohort contains 30 2Wiki, 30 HotpotQA, and 29 MuSiQue questions from the larger evaluation. The other 450 questions contribute 150 per dataset. The order cohort contains ten questions per dataset, selected by question-hash order from the complete cohort. Reversal retains each passage's original identifier, title, and saved content. Each order condition has all 64 utility values and all 42 budget-feasible answers.

Cross-encoder configuration. The cross-encoder uses raw logits and an 8,192-token context limit; the longest evaluated pair has 918 tokens. MMR uses $\lambda = 0.5$, sigmoid-transformed singleton logits for relevance, and $(1 + \cos)/2$ for similarity between normalized BGE-M3 passage embeddings [3]. The principal application results use retrieval-constructed pools B. The five rules access 6, 6, 15, 20, and 41 scores, respectively, when evaluated individually.

Answer scoring and uncertainty. Token F1 and EM use the same frozen answer dictionary across models, selection rules, and presentation orders. Answer normalization lowercases text, removes punctuation and English articles, and collapses whitespace. For results combining constructions, each question contributes its mean paired contrast across A and B. We use 20,000 dataset-stratified bootstrap draws over questions, resampling all matched observations together. Intervals are 95% percentile intervals.

Table 2. Budget flexibility and same-size search have different answer effects. Mean paired token-F1 changes on 539 questions, with 95% question-bootstrap intervals.

Model	Budget C^F	Search J^F	Total
Qwen	-1.46 $[-3.99, 0.97]$	-0.35 $[-2.13, 1.46]$	-1.81 $[-4.60, 0.95]$
Gemma	$+0.95$ $[-1.32, 3.23]$	$+0.65$ $[-0.79, 2.12]$	$+1.60$ $[-0.90, 4.15]$
Mistral	$+4.79$ $[3.12, 6.53]$	$+1.11$ $[-0.002, 2.23]$	$+5.91$ $[4.06, 7.82]$

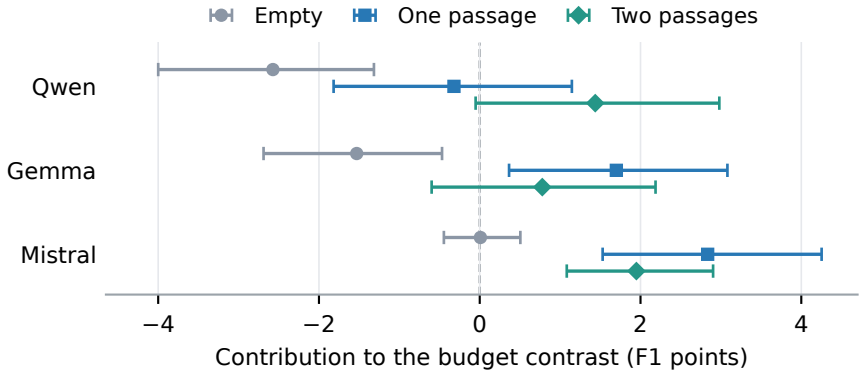


Fig. 2. One- and two-passage choices account for Mistral's budget gain. Points show the contributions of empty, one-passage, and two-passage selections to the mean paired budget contrast. Lines show 95% question-bootstrap intervals.

5 Budget Flexibility Has a Model-Specific Answer Effect

5.1 Mistral's Main Gain Comes from the Chosen Set Size

Releasing the three-passage constraint accounts for most of Mistral's advantage over fixed-step greedy selection (Table 2). Across 539 questions, the budget contrast is +4.79 F1 points, with a 95% interval of $[3.12, 6.53]$. The same-size search contrast is +1.11 $[-0.002, 2.23]$, yielding a total of +5.91 $[4.06, 7.82]$. The comparison assigns the larger observed gain to choosing how much evidence to retain.

The corresponding budget means for Qwen and Gemma are -1.46 and $+0.95$. Direct paired model comparisons place Mistral's budget effect 6.25 points $[3.29, 9.23]$ above Qwen's and 3.85 points $[0.97, 6.83]$ above Gemma's. Thus the budget choice produces different answer effects even when the candidate constructions and selection definitions are shared.

Table 3. Likelihood selection leaves a substantial feasible-answer gap. Complete answer tables give the highest feasible token F1 at budget three and the values reached by likelihood-optimal and fixed-greedy selection. The paired gap from $E_{\leq}$ to the F1 optimum has a 95% question-bootstrap interval.

Model / pool	Best F1	$F(E_{\leq})$	$F(G_3)$	F1 gap
Qwen / A	86.96	65.42	77.83	21.54 [13.84, 29.69]
Qwen / B	71.96	51.16	60.21	20.80 [13.38, 28.63]
Gemma / A	86.10	71.79	69.71	14.31 [8.41, 20.80]
Gemma / B	72.94	59.14	55.17	13.80 [8.30, 19.94]

The selected size further locates the change (Figure 2). For Mistral, one-passage choices contribute 2.84 points and two-passage choices 1.95 points to the overall budget contrast. Empty selections contribute 0.01 points. The combined contribution of nonempty smaller sets is 4.78 $[3.18, 6.46]$. The positive effect is therefore concentrated in selections that retain a compact evidence context.

This view also resolves the other models' aggregate means. Gemma's nonempty smaller selections contribute $+2.48$ points, while empty selections contribute -1.53. Qwen has contributions of $+1.12$ and -2.57, respectively. The selected-size view exposes different mixtures behind the three model averages and identifies the concrete decisions responsible for the budget contrast.

5.2 Question Coverage Changes the Qwen Comparison

Qwen's endpoint difference varies substantially across the two question groups. On the 89-question cohort, the F1 difference between $E_{\leq}$ and G_3 is -10.73 points, with interval $[-17.96, -3.85]$. On the remaining 450, the estimate is -0.04 $[-3.04, 2.93]$. Over all 539, it is -1.81 $[-4.60, 0.95]$. The direct descriptive difference between the two groups is -10.68 $[-18.50, -3.08]$.

The intermediate selection locates part of this variation in combination search. Qwen's same-size contrast is -4.84 points on the nested 89 and $+0.54$ on the remaining 450. Mistral's total contrasts are positive in both groups: $+8.62$ $[4.12, 13.72]$ and $+5.37$ $[3.37, 7.46]$, respectively.

5.3 Complete Answers Expose Both Objective and Coverage Gaps

The complete answer tables locate each likelihood decision within the feasible answer space (Table 3). Qwen's likelihood-optimal selection trails the highest feasible F1 by 21.54 points in construction A and 20.80 in B; Gemma's gaps are 14.31 and 13.80. The table also shows the model-dependent effect of score optimization: fixed likelihood-greedy obtains higher mean F1 than exact likelihood selection for Qwen and lower mean F1 for Gemma.

Table 4. Complete generation reveals higher-scoring answers beyond the selected-set union. On 89 questions, "Best of six" and "Best of twelve" give the highest observed token F1 among sets chosen by each rule family; "Complete" maximizes over all 42 feasible sets. The excess over twelve rules has a 95% question-bootstrap interval.

Model	Best of six	Best of twelve	Complete	Excess over twelve
Qwen	76.59	77.04	79.46	2.42 [0.30, 5.45]
Gemma	73.92	74.86	79.52	4.67 [1.88, 8.15]

Complete generation also reveals answers beyond those reached by a family of selection rules. We first take the best F1 among sets chosen by six rules: BM25 top-3, BGE top-3, singleton-likelihood top-3, likelihood-greedy with positive-gain stopping, fixed-step likelihood-greedy, and $E_{\leq}$. The stopped rule adds passages while the largest utility gain is positive, up to three passages. We then enlarge this family to twelve rules by adding $E_{=}$ and the five cross-encoder rules from Section 3.4.

Table 4 shows what the full answer table adds. Relative to the six-rule union, the highest feasible F1 is 2.87 points higher for Qwen and 5.60 for Gemma. With all twelve rules, the remaining excess is 2.42 points $[0.30, 5.45]$ and 4.67 points $[1.88, 8.15]$, respectively. Thus, even the enlarged union leaves higher-scoring contexts outside its selections. Complete answers provide both the intermediate readouts needed for the paired comparisons and a broader view of the answers available within each candidate pool.

6 Reselection Can Offset a Presentation Loss

6.1 Large Intermediate Changes Appear Within Smaller Endpoint Changes

Gemma exhibits opposing presentation and reselection effects across all three selection rules (Table 5). For the exactly-three optimum, reversing the selected set lowers token F1 by 20.56 points $[-33.64, -8.53]$. Choosing the best three-passage set under reversed utility recovers 16.54 points $[8.07, 26.20]$. The endpoint change is -4.02 $[-11.00, 2.48]$. Holding three passages throughout the path makes the large, oppositely directed intermediate changes especially clear.

The other Gemma rules show the same directional pattern. The at-most-three optimum loses 17.85 points under fixed-set reversal and gains 10.28 through reselection. Fixed-step greedy loses 23.79 and gains 18.72. Their endpoint changes are -7.57 and -5.07. In each case, the final selected-context comparison combines a substantial presentation loss with a subsequent gain from changing the selected evidence.

Qwen produces a more varied pattern. Its fixed-greedy path has a presentation change of -10.40 points $[-20.95, -1.11]$ and a reselection change of $+6.98$

Table 5. Following the same selected passages through reversal separates presentation from reselection. Paired token-F1 changes on 30 matched questions; brackets give 95% question-bootstrap intervals.

Model / rule	Presentation	Reselection	Total
Qwen $E_{\leq}$	-7.73 $[-21.75, 6.00]$	$+11.03$ $[-1.83, 25.00]$	$+3.30$ $[-6.19, 13.97]$
Qwen $E_{=}$	-2.04 $[-11.11, 7.67]$	-0.91 $[-8.69, 6.94]$	-2.95 $[-9.98, 3.98]$
Qwen G_3	-10.40 $[-20.95, -1.11]$	$+6.98$ $[0.24, 14.84]$	-3.41 $[-14.05, 7.14]$
Gemma $E_{\leq}$	-17.85 $[-31.27, -5.62]$	$+10.28$ $[1.85, 19.00]$	-7.57 $[-18.57, 3.00]$
Gemma $E_{=}$	-20.56 $[-33.64, -8.53]$	$+16.54$ $[8.07, 26.20]$	-4.02 $[-11.00, 2.48]$
Gemma G_3	-23.79 $[-37.34, -11.08]$	$+18.72$ $[6.68, 31.79]$	-5.07 $[-14.57, 3.14]$

Table 6. The same membership change has different answer effects across presentation orders. Four token-F1 readouts for exact-three likelihood selection on 30 questions. S_c and S_r are selected under canonical and reversed likelihood. The final column is the paired crossed difference from Equation 8, with a 95% question-bootstrap interval.

	S_c		S_r		Crossed
Model	F_c	F_r	F_c	F_r	difference I^F
Qwen	69.30	67.26	55.73	66.34	$+12.65$ $[0.31, 26.02]$
Gemma	68.99	48.43	45.64	64.97	$+39.89$ $[24.72, 56.96]$

$[0.24, 14.84]$. The at-most-three means also point in opposite directions, with wider intervals. Its exactly-three path has two small negative point estimates. Observing each rule separately retains this variation alongside Gemma's consistent directional pattern.

6.2 Reselection Gains Depend on the Presentation Order

The four-cell comparison shows how the same membership change behaves under each order (Table 6). For Gemma, the canonical-selected three-passage set scores 68.99 F1 in canonical order and 48.43 when reversed. The reverse-selected set scores 45.64 in canonical order and 64.97 when reversed. Replacing the former set with the latter therefore lowers F1 by 23.35 points $[-35.58, -11.83]$ in canonical order and raises it by 16.54 points $[8.07, 26.20]$ in reversed order. The crossed difference is 39.89 points $[24.72, 56.96]$. The same membership change has opposite answer effects across the two presentations.

Table 7. The same cross-encoder selections produce model-dependent answer scores. Five selection rules on 89 retrieval-constructed pools. Answer scores use a 0–100 scale; scoring accesses follow each selection rule.

		Qwen		Gemma	
Selection rule	Scores	F1	EM	F1	EM
Singleton top-3	6	52.76	43.82	43.25	34.83
MMR-3	6	46.74	39.33	41.58	33.71
Joint greedy-3	15	51.32	41.57	45.81	38.20
Joint exact-3	20	48.41	39.33	44.39	34.83
Joint best-1-to-3	41	44.63	35.96	41.39	31.46

Qwen has a smaller crossed difference of 12.65 points $[0.31, 26.02]$. Its reverse-selected set has lower mean F1 under both presentations, with a larger mean loss in canonical order (-13.57) than in reversed order (-0.91). Reading the four cells together distinguishes this pattern from Gemma's opposing signs and shows precisely how each model's response depends on the selected set and its presentation.

6.3 The Running Example Shows What the Intermediate State Adds

The Devilment example in Figure 1 makes the three states tangible. In Qwen's construction A, the canonical likelihood-optimal set contains d_1 and d_3. These passages concern Dani Filth and Ronnie Radke; d_6 concerns Jeff Caxide. The generated answer is *Cradle of Filth*. Presenting the same blocks as (d_3, d_1) produces *Falling in Reverse*. Under the reversed-order utility table, selection moves to $\{d_1, d_6\}$, serialized as (d_6, d_1), and the answer returns to *Cradle of Filth*.

The three responses have token F1 values of 100, 0, and 100. The endpoint change is zero, while the two adjacent changes are -100 and $+100$ points. The zero endpoint difference therefore contains two complete answer flips, exactly the variation that the intermediate comparison is designed to reveal.

7 A Shared Scorer Localizes the Remaining Answer Changes

7.1 Greedy and Exact Search Often Reach the Same Set

Table 7 compares the five cross-encoder rules on 89 retrieval-constructed pools. Joint greedy-3 and joint exact-3 select identical sets on 75 questions. Across the 89 pools, the mean cross-encoder logit gain is 0.03185, with interval $[0.01209, 0.05556]$. The selected set changes on 14 questions; Qwen changes seven answers and Gemma changes nine. The mean F1 contrasts are -2.91 points $[-7.54, 0.98]$ and -1.42 $[-6.29, 3.07]$, respectively.

Following the changed sets gives a more detailed account of the answer effects. Qwen's F1 increases on two questions and decreases on four; Gemma's increases on three and decreases on four. The other questions retain the same F1. Thus the score improvements are accompanied by a small set of answer changes whose directions can be inspected individually.

The singleton and diversity rules provide useful context for the joint comparisons. Singleton top-3 attains mean F1 of 52.76 for Qwen and 43.25 for Gemma. Joint greedy-3 attains 51.32 and 45.81. The relative ordering of these rules changes with the generator, even though they use the same candidate pools and cross-encoder. These results connect the scoring decision to the particular model that consumes the selected context.

7.2 Budget Release Changes More Sets Under the Cross-Encoder

Allowing the cross-encoder to choose one, two, or three passages changes 41 of the 89 sets relative to joint exact-3. The mean logit gain is 0.1670 $[0.1064, 0.2347]$. The corresponding F1 contrasts are -3.78 points $[-8.13, 0.00]$ for Qwen and -3.00 $[-8.02, 1.57]$ for Gemma. The budget comparison therefore produces more membership changes than same-size search in this setting, with the generated answers supplying their downstream readout.

The five rules also expose the scoring work associated with each decision. Singleton top-3 and MMR use six cross-encoder scores, joint greedy uses 15, joint exact-3 uses 20, and joint best-1-to-3 uses 41. These schedules pair additional scoring with the precise set and answer changes it produces.

8 Implications for Evidence-Selection Evaluation

The results motivate retaining intermediate contexts when evaluating evidence selectors. At a chosen size, compare the combinations reached by the search rule; then vary the size under the same score. Mistral's results place the larger answer gain in smaller nonempty contexts, while the shared cross-encoder often reaches the same set through greedy and exact search. The selected sets and their answers make these distinct outcomes visible.

Presentation calls for a crossed comparison. Gemma's large reversal loss is partly offset by reselection, and the reselected membership helps specifically under the reversed presentation. Evaluating both sets in both orders exposes this dependence. Qwen's four-cell pattern differs, showing why the set change and presentation change should be observed together for each generator.

Complete answer tables broaden this account beyond the chosen paths. Their best feasible F1 remains above the best observed answer among twelve selection rules. Recording the shared candidate pool, selected membership, presentation, score, and response therefore supports two complementary questions: which decision changed the answer, and which higher-scoring answers remain outside the selector's observed choices?

9 Conclusion

Endpoint comparisons can combine large, opposing answer changes. Two matched evidence-selection paths locate Mistral's main gain in smaller nonempty contexts, expose Gemma's opposing presentation and reselection effects, and show frequent agreement between greedy and exact cross-encoder search. Crossing selected sets with presentation orders further reveals that the same membership change can help under one order and hurt under the other. Complete answer tables supply these intermediate responses and expose higher-scoring contexts beyond a broad selected-set union, connecting evidence decisions to the answers they produce.

10 Limitations

Scope and selection information. The experiments cover short-answer multi-hop QA, benchmark-local pools of six passages, and three specified generators. Reference-answer likelihood provides a diagnostic selection objective; using its choices on a new question would require that reference. The feasible-answer optimum and best-of-rule-family scores use evaluated responses and the reference dictionary. They quantify answer coverage rather than an executable answer-blind selection policy; the result depends on the included rule family. Complete canonical generation covers 89 questions, and complete reversed budget-feasible generation covers 30. Larger open-corpus pools, other prompts, and long-form answers remain to be studied.

Presentation, uncertainty, and cost. The order study evaluates one global reversal. The 30-question resource tier was chosen after pilots and the budget/cross-encoder runs, before formal order evaluation and effect aggregation. The cohorts are nested, and bootstrap intervals are pointwise. The signed and crossed differences organize matched response comparisons; they do not identify a unique content mechanism. The coverage and crossed-order analyses were added during revision using the saved observations. Their intervals describe question-sampling uncertainty and do not account for analysis selection. A confidence interval crossing zero leaves the effect's direction unresolved. Cross-encoder selection is evaluated in canonical order with a fixed pretrained scorer applied to concatenated passages. The scoring counts measure accesses to that scorer; per-policy online latency and answer-blind reselection under reversal have not been evaluated.

Answer references and semantic assessment. Token F1 measures lexical overlap and can give partial credit to an incorrect entity. The answer dictionary retains benchmark references and a limited set of reviewed answer forms. An attempted dual-model semantic assessment produced unresolved judgments and confirmed shared judge errors, so its automatic labels are excluded from the reported correctness evidence. We also identified two reference conflicts and three questions with insufficient supporting packets. Excluding the two conflicts leaves Mistral's budget contrast at $+4.78$ points $[3.11, 6.54]$; excluding all five gives

+4.73 [3.04, 6.55]. These questions are outside the order and cross-encoder cohorts. Full factual reannotation remains open.

Research materials and assistance. The accompanying materials retain the saved passage text, subset identities, model and prompt configurations, generated responses, and scripts for the reported comparisons. ChatGPT and Codex assisted with implementation, analysis, literature checking, figure preparation, and writing.

References

1. Beijing Academy of Artificial Intelligence: BGE-Reranker-v2-M3: Model card. https://huggingface.co/BAAI/bge-reranker-v2-m3 (2024), accessed 9 November 2026
2. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 335–336. ACM (1998). https://doi.org/10.1145/290941.291025
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 2318–2335 (2024). https://doi.org/10.18653/v1/2024.findings-acl.137
4. Cormack, G.V., Clarke, C.L.A., Büttcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 758–759 (2009). https://doi.org/10.1145/1571941.1572114
5. Du, W.: Complex utility, simple calibration: Decision sufficiency in rag evidence selection (2027), accepted to WSDM 2027 Findings
6. Gemma Team, et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025), https://arxiv.org/abs/2503.19786
7. Ho, X., Nguyen, A.K.D., Sugawara, S., Aizawa, A.: Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In: Proceedings of COLING. pp. 6609–6625 (2020). https://doi.org/10.18653/v1/2020.coling-main.580
8. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 874–880. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.eacl-main.74
9. Jain, S., Vedam, V.N.: CUE-R: Beyond the final answer in retrieval-augmented generation. arXiv preprint arXiv:2604.05467 (2026). https://doi.org/10.48550/arXiv.2604.05467
10. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., El Sayed, W.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023), https://arxiv.org/abs/2310.06825
11. Jiang, Y., Zhao, S., Li, J., Wang, H., Hu, B., Du, Y., Qin, B.: OptiSet: Unified optimizing set selection and ranking for retrieval-augmented generation. In: Findings of the Association for Computational Linguistics: EMNLP 2026. pp. 7125–7139. Association for Computational Linguistics (2026). https://doi.org/10.18653/2026.findings-emnlp.339

12. Lee, D., Jo, Y., Park, H., Lee, M.: Shifting from ranking to set selection for retrieval augmented generation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 17606–17619. Association for Computational Linguistics (2025). https://doi.org/10.18653/v1/2025.acl-long.861
13. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
14. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024). https://doi.org/10.1162/tacl_a_00638
15. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval (2009). https://doi.org/10.1561/1500000019
16. Seo, D., Jang, B.: When retrieval similarity hurts: A diagnostic framework for document weighting in black-box RAG. In: Findings of the Association for Computational Linguistics: EMNLP 2026. pp. 12534–12558. Association for Computational Linguistics (2026). https://doi.org/10.18653/2026.findings-emnlp.594
17. Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.t.: REPLUG: Retrieval-augmented black-box language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 8371–8384. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.naacl-long.463
18. Tian, F., Ganguly, D., Macdonald, C.: Is relevance propagated from retriever to generator in RAG? In: Advances in Information Retrieval: ECIR 2025, Part I. Lecture Notes in Computer Science, vol. 15572, pp. 32–48. Springer (2025). https://doi.org/10.1007/978-3-031-88708-6_3
19. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: MuSiQue: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics **10**, 539–554 (2022). https://doi.org/10.1162/tacl_a_00475
20. Wang, Z., Yuan, H., Dong, W., Cong, G., Li, F.: CARROT: A learned cost-constrained retrieval optimization system for RAG. arXiv preprint arXiv:2411.00744 (2026), https://arxiv.org/abs/2411.00744, accepted to ICDE 2026; earlier title used CORAG
21. Yang, A., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024), https://arxiv.org/abs/2412.15115
22. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: 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. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1259