MAPPING THE SET-FUNCTION GEOMETRY OF EVIDENCE UTILITY IN RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence selection depends on the structure of generator utility, the search strategy, and the evaluation target. We introduce Evidence Utility Geometry, an exhaustive subset-intervention audit that measures all three within a fixed evidence interface. Across three multi-hop question-answering datasets, three generator families, and controlled versus support-blind six-passage pools, all 18 dataset–model–construction cells exhibit material non-monotonicity and increasing returns. Negative marginals occur in 27.3–37.9% of tests and positive discrete Hessians in 35.5–42.9%. Exact regret decomposition separates the price of filling a budget from positive same-cardinality search loss, including in an eight-passage extension. A complete generation audit evaluates all 64 subsets for 90 existing questions, two models, and both constructions. Generated F1 also exhibits negative marginals and positive Hessians. Likelihood–F1 rank correlations of 0.55 and 0.49 coexist with different optima: at budget three, exact likelihood selection has F1 regret of 0.184 and 0.175 against the exact F1 optimum, and its paired F1 contrasts with fixed likelihood-greedy span zero. Cardinality-standardized comparisons reveal how weighting subset sizes changes cross-size measurements. EUG connects exhaustive structural measurement to exact selection losses and target alignment, identifying where evidence-selection losses arise.

1 INTRODUCTION

Evidence selection combines a search problem with a stopping decision. In retrieval-augmented generation (RAG), these decisions are commonly guided by relevance, coverage, diversity, or estimated generator utility (Lewis et al., 2020; Lee et al., 2025; Wang et al., 2025). A designed selection objective can support efficient optimization; evaluating its relationship to generator behavior determines the relevance of that guarantee.

Three properties determine the quality of this decision. Evidence utility can decrease after an addition or increase through combinations with weak individual contributions. A greedy selector can incur a cost from filling its budget and a further cost from choosing the wrong combination at that size. Finally, likelihood and generated-answer quality can favor different evidence sets. Separating these properties identifies whether the relevant intervention concerns stopping, search, or the optimization target.

Complete subset evaluation makes each distinction measurable. For each question and generator, Evidence Utility Geometry (EUG) measures centered teacher-forced answer log-likelihood $U(S)$ for every subset S. The passage text, prompt, and serialization rule are held fixed. Complete enumeration directly yields marginal changes, discrete Hessians, and exact cardinality-constrained optima. Figure 1 separates this structural measurement from its decision and generation tests.

The primary experiment covers 540 questions from three multi-hop datasets, three instruction-tuned models, and two six-passage constructions. One construction includes annotated supports; the other uses support-blind local retrieval. Material violations recur in every dataset–model–construction cell and persist in an eight-passage extension. Prior work models set utility, harmful additions, and non-additive evidence interventions (Jiang et al., 2026; Wang et al., 2026; Jain & Vedam, 2026). Our

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.

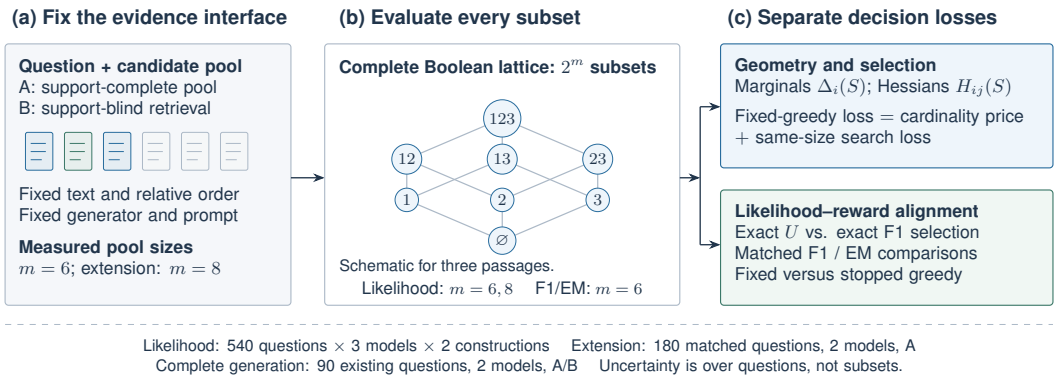

Figure 1: **Measure structure, selection loss, and target alignment.** Complete likelihood lattices identify violations and exact selection losses. Complete generated-reward lattices on a fixed subcohort test whether optimizing likelihood also optimizes answer quality. All comparisons are conditional on the declared evidence interface. The generation audit uses a 90-question subset of the primary roster.

contribution is a complete, matched measurement of *objective structure and its consequences for evidence selection.*

Exact selection and complete generation connect geometry to decision quality. Exact regret decomposition separates the cost of requiring exactly b passages from loss relative to the best set of that same size. Both terms are positive on average, and stopped greedy retains a gap to the exact optimum. A complete generated-answer audit then scores all 64 subsets for 90 of the existing questions. F1 exhibits non-monotonicity and increasing returns. The exact U optimum has substantial regret against the F1 optimum, and its paired F1 contrasts with fixed U-greedy span zero. The result separates optimization under likelihood from optimization of generated-answer quality.

We contribute an exhaustive diagnostic with three linked readouts: interface-conditioned set-function geometry, exact cardinality and search losses, and complete generated-reward landscapes. Matched comparisons reveal when stopping and combination search address different losses, and how likelihood can track answer quality while favoring a different optimum. Cardinality-standardized comparisons also identify a weighting effect in cross-size measurements.

2 Related work

Set selection and generator alignment. SetR selects evidence jointly using query information requirements (Lee et al., 2025); InfoGain-RAG estimates document information gain for reranking and filtering (Wang et al., 2025); OptiSet learns set selection and ordering from generator-conditional utility changes (Jiang et al., 2026). GeoRAG constructs a demand-coverage objective with monotone-submodular structure (Zhang et al., 2026). The classical greedy approximation theorem applies when its structural hypotheses hold (Nemhauser et al., 1978). Those guarantees concern the specified selection objective. We measure the interface-conditioned generator target on every subset of a fixed pool, allowing its structure to be checked separately.

CARROT explicitly addresses non-monotone utility and stopping before budget exhaustion (Wang et al., 2026). Stochastic RAG relaxes document independence through expected-utility optimization (Zamani & Bendersky, 2024). CoRAG jointly optimizes reranker and generator decisions, while BAR-RAG uses generator feedback to train an evidence selector and then adapts the generator (Song et al., 2026; Sun et al., 2026). These learning-based approaches motivate a complementary diagnosis of search error and target alignment. Our answer-aware oracles give exact references for both quantities within the evaluated evidence pools.

Coalitions, interventions, and evaluation targets. Source attribution studies Shapley-style document contributions and their approximations (Nematov et al., 2025). RepoShapley uses teacher-forced probing, an interaction-aware surrogate game, and bounded generator verification (Huo et al., 2026).

CUE-R evaluates removal, replacement, and duplication, including a two-support non-additivity test (Jain & Vedam, 2026). EUG complements these analyses with full-lattice evaluation and exact budget-constrained regret. Relevance–utility studies already distinguish retrieved relevance from generation outcomes (Tian et al., 2025); our generated-reward audit makes that distinction within the same fully observed subset space. Position sensitivity (Liu et al., 2024) further motivates treating serialization as part of the estimand.

3 EVIDENCE UTILITY, GEOMETRY, AND EXACT LOSSES

3.1 A CONDITIONAL SET FUNCTION

Fix question q, generator g, candidate universe $D_q = \{d_1, \dots, d_m\}$, and canonical answer tokens $y_{1:T}$. For a fixed serializer ser, define

$$L_{q,g}(S) = \frac{1}{T}\sum_{t=1}^{T} \log p_g(y_t \mid q, \mathrm{ser}(S), y_{<t}), \qquad U_{q,g}(S) = L_{q,g}(S) - L_{q,g}(\emptyset). \tag{1}$$

Centering sets $U(\emptyset) = 0$ and preserves every finite difference. The resulting U is a signed likelihood utility. Candidate text is prepared once, and every subset preserves canonical relative order. Removing a passage changes absolute positions and subsequent exposure; these interface effects are part of U.

Identifying passages with their indices, for $i, j \notin S$, $i \neq j$, the marginal and discrete Hessian are

$$\Delta_i(S) = U(S \cup \{i\}) - U(S), \tag{2}$$

$$H_{ij}(S) = U(S \cup \{i,j\}) - U(S \cup \{i\}) - U(S \cup \{j\}) + U(S). \tag{3}$$

Monotonicity requires every marginal to be nonnegative. Submodularity is equivalent to every such Hessian being nonpositive. Positive H identifies increasing returns under the declared interface. Complete evaluation requires 2^m contexts and yields $m2^{m-1}$ edges and $\binom{m}{2}2^{m-2}$ Hessian tests: 192 and 240, respectively, at $m = 6$.

Magnitude and uncertainty. For each dataset–model–construction cell, s is the interquartile range of centered U values over all subsets of its calibration questions. We use $\epsilon = \max(10^{-3}, 0.05s)$ for material signs, counting $\Delta < -\epsilon$ and $H > \epsilon$. Normalized positive-Hessian mass is $\sum_{S,i<j} [H_{ij}(S)]_+ / s$; within a fixed m its sum and mean differ only by a constant. Rates and masses are first computed per question–model cube. Question-level resampling retains the dependence among finite differences. Threshold-free quantities and fixed threshold sensitivity are retained. Appendix J separates calibration scales, sign tolerances, and effect-size thresholds in a compact reference table.

3.2 SEPARATE STOPPING AND COMBINATION SEARCH

Let $O_{\le b} = \max_{|S| \le b} U(S)$ and $O_{=b} = \max_{|S|=b} U(S)$. Fixed true-marginal greedy starts empty and adds the largest available marginal until it has b passages, giving G_b. Stopped greedy instead terminates when its best marginal is nonpositive. Exact optima use the smallest bitmask to resolve ties; greedy ties use the smallest candidate index.

For fixed greedy, the following identity separates two nonnegative quantities:

$$\underbrace{O_{\le b} - U(G_b)}_{\text{total loss}} = \underbrace{O_{\le b} - O_{=b}}_{\text{cardinality price}} + \underbrace{O_{=b} - U(G_b)}_{\text{same-cardinality search loss}}. \tag{4}$$

The first is the best possible price of requiring exactly b items. The second compares two sets of identical size. Dividing each term by the same s preserves the identity. The decomposition separates two exact optimization gaps. The fixed-versus-stopped comparison evaluates a separate policy choice, because early stopping can also miss useful combinations.

We therefore inspect stopped states with at least two remaining budget slots. A *pair escape* occurs when every available singleton gain is nonpositive but some pair has joint gain greater than ϵ. The pair provides an explicit improving move beyond a singleton stopping point. Singleton-U, BM25, BGE, MMR, and semantic-deduplication selectors provide additional comparisons (Appendix B). All U-based oracles are reference-answer-aware diagnostics.

Table 1: **Questions are the independent units; cubes are repeated measurements.** The eight-passage and complete-generation studies reuse subsets of the primary final roster. Calibration is separate.

Study	Questions	Models	Tracks	Subsets/cube	Cubes
Calibration, $m = 6$	180	3	2	64	1,080
Primary, $m = 6$	540	3	2	64	3,240
Extension, $m = 8$	180	2	1	256	360
Complete generation, $m = 6$	90	2	2	64	360

Table 2: Study chronology. Extension and generation cohorts are nested within the primary roster.

Study	Timing relative to primary outcomes
Primary likelihood	Protocol and calibration fixed before final scoring.
Eight-passage extension	Design fixed beforehand; execution chosen after primary results.
Complete generation	Designed after primary results; membership fixed before generation.

3.3 A DIFFERENT LANDSCAPE FOR GENERATED ANSWERS

Let $F(S)$ and $E(S)$ be normalized token F1 and exact match (EM) of the greedily decoded answer, maximized over official aliases. Their finite differences use the same lattice as U. We separately compute the exact F1 optimum $F^*_{\leq b} = \max_{|S|\leq b} F(S)$ and

$$R^F_b(\pi) = F^*_{\leq b} - F(S_\pi), \qquad D^F_b(\pi) = F(S^*_U) - F(S_\pi), \tag{5}$$

where S^*_U maximizes U under the same budget. R^F_b is nonnegative; D^F_b can have either sign. Rank association between U and F characterizes the landscape as a whole; $R^F_b(S^*_U) = 0$ and $D^F_b(\pi) > 0$ respectively characterize optimal-value alignment and improvement over a comparator. When optima tie, value regret and optimum-value hits are more informative than bitmask equality alone.

4 EXPERIMENTAL DESIGN

Data and candidate constructions. We use the official development distributions of 2WikiMultiHopQA, HotpotQA distractor, and MuSiQue answerable (Ho et al., 2020; Yang et al., 2018; Trivedi et al., 2022). Each contributes 60 calibration and 180 disjoint final questions, selected by a fixed ordering. Track A includes every annotated support and fills the remaining slots with hard non-support distractors selected without generator outcomes. Track B uses support-blind reciprocal-rank fusion (Cormack et al., 2009) of BM25 (Robertson & Zaragoza, 2009) and BGE within the supplied local candidates. Both constructions use benchmark-supplied local candidates. Each passage retains the same representation, capped at 256 reference-tokenizer tokens, across all subsets.

Generators and likelihood evaluation. The primary generators are Qwen2.5-7B-Instruct, Gemma-3-4B-it, and Mistral-7B-Instruct-v0.3 (Yang et al., 2024; Gemma Team et al., 2025; Jiang et al., 2023). The extension and complete-generation audit use Qwen and Gemma. Model revisions, prompts, targets, and candidate memberships are fixed. Inference uses BF16 weights, batch size one, and float32 log-softmax for likelihood; the numerical configuration is specified in Appendix A. The primary and extension evaluate 207,360 and 92,160 likelihood contexts, respectively.

Sequential studies and analysis scope. Table 2 separates protocol specification from the decision to execute each study. The complete-generation audit uses a fixed ordering of 30 existing final questions per dataset, chosen independently of measured outcomes, across two models and both tracks. Its 23,040 generations are repeated measurements of those questions, using greedy decoding with a 64-token limit. Appendix F.1 reports the separate Gemma end-of-turn sensitivity.

We use 10,000 percentile-bootstrap draws, stratifying by dataset and retaining all compared model outputs in each resampled question cluster. Datasets receive equal weight. Undefined correlations for constant reward cubes remain undefined; sign rates include those cubes. Each interval describes

Table 3: **Violations persist beyond annotated support-complete pools.** Ranges span dataset–model cells and use calibration-based materiality. All 18 six-passage and all six eight-passage cells pass the predeclared material-geometry rule. These rates concern likelihood utility, not generated F1.

Track, size	Cells	$\Delta < -\epsilon$ (%)	$H > \epsilon$ (%)
A, $m = 6$	9	27.8–37.1	35.5–41.7
B, $m = 6$	9	27.3–37.9	35.7–42.9
A, $m = 8$	6	27.2–36.7	34.0–39.8

Table 4: **Stopping and combination search are different problems.** Mean losses at $b = 3$, divided by the calibration scale; brackets are 95% question-cluster intervals. Fixed loss equals cardinality price plus search loss pointwise. Stopped regret is measured against the at-most-budget optimum and is not a term in that decomposition.

Track, size	Fixed regret	Cardinality price	Search loss	Stopped regret
A, $m = 6$	0.164 [0.138, 0.193]	0.097 [0.081, 0.115]	0.067 [0.051, 0.085]	0.050 [0.038, 0.063]
B, $m = 6$	0.223 [0.195, 0.254]	0.144 [0.123, 0.165]	0.080 [0.065, 0.095]	0.051 [0.039, 0.066]
A, $m = 8$	0.171 [0.125, 0.221]	0.071 [0.046, 0.102]	0.100 [0.068, 0.138]	0.075 [0.048, 0.109]

its specified comparison. Primary association tests, supplemental decompositions, and post hoc sensitivities retain their distinct roles. Details and all cell estimates appear in the appendix.

5 Results

5.1 Material violations recur in both constructions

All 18 six-passage cells exhibit material non-monotonicity and positive Hessians (Table 3). The recurrence spans models, datasets, and evidence constructions. At budget three, the predeclared joint rule linking positive-Hessian mass to fixed-greedy regret holds in seven of nine controlled cells and eight of nine support-blind cells. Cell correlations range from 0.019 to 0.359 in Track A and from 0.117 to 0.420 in Track B; all cell estimates and intervals are reported in Appendix C. These comparisons measure how geometry and regret co-vary within the observed landscapes.

The eight-passage extension exhibits material violations in every cell. The following analyses resolve its selection losses and the cardinality composition of its finite differences.

5.2 Filling the budget explains only part of greedy loss

Budget enforcement and combination search both contribute to greedy loss. At $b = 3$, the cardinality price and same-cardinality search loss are positive in both six-passage tracks and the eight-passage extension (Table 4). The latter term directly measures better combinations at equal set size.

Early stopping substantially reduces mean U regret, while pair escapes expose remaining opportunities. Among eligible stopped states, material pair escapes occur in 15.9% of Track A and 16.2% of Track B cases at $b = 3$; the corresponding eight-passage rate is 26.9%. Both individual additions can be non-improving while their joint addition improves utility. An existing $b = 4$ witness has stopped state $S = \{d_6\}$, singleton gains -0.6556 and -2.5739 for d_2, d_3, and joint gain $+0.4277 > \epsilon = 0.1774$. These values exhibit complementarity under the fixed interface. Continuing fixed greedy eventually reaches the optimum in this example (Appendix D).

Full budget curves distinguish the consequences further. At $b = m$, same-cardinality search loss is necessarily zero because there is only one set of size m. Total fixed-greedy loss can nevertheless remain positive relative to a smaller optimum. Thus the budget separately governs the attainable cardinality price and the search space (Appendix D.1).

Table 5: **Answer quality also has non-monotone, non-submodular local structure.** Complete generation on 90 questions, two models, and two tracks. Rates include constant cubes; correlations exclude undefined cubes before question-level aggregation. Intervals are 95% question-cluster intervals.

Track	Reward	Negative marginal (%)	Positive Hessian (%)	$\rho(U, \text{reward})$
A	F1	7.19 [6.24, 8.22]	9.78 [8.56, 11.07]	0.553 [0.495, 0.608]
A	EM	4.38 [3.64, 5.13]	6.09 [5.08, 7.13]	0.551 [0.497, 0.604]
B	F1	6.62 [5.54, 7.76]	8.36 [7.10, 9.67]	0.489 [0.414, 0.563]
B	EM	3.93 [3.09, 4.80]	4.84 [3.86, 5.88]	0.555 [0.488, 0.620]

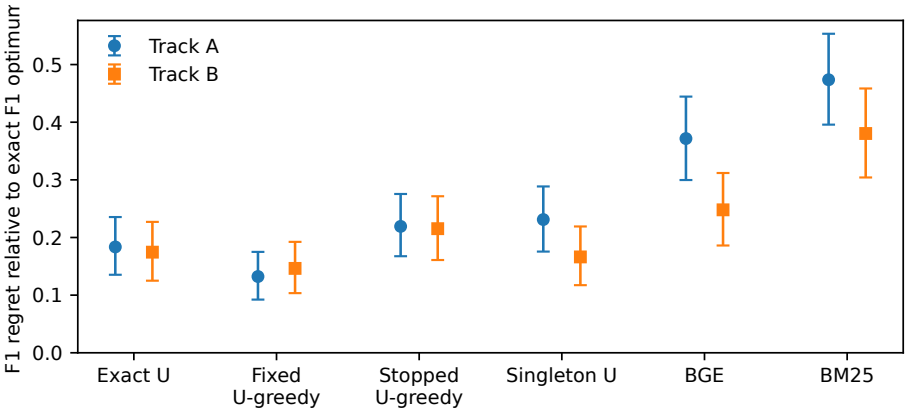

Figure 2: **Exact likelihood optimization does not remove generated-answer regret.** At $b = 3$, every policy is scored against the exact F1 optimum on the same complete generated cube. Error bars are 95% question-cluster intervals. U-based methods are answer-aware diagnostic oracles; BM25 and BGE are answer-blind retrieval references.

Retrieval and diversity comparisons. The local BM25-to-BGE membership comparison at budget four also exposes destructive changes: support recall can increase while U decreases by more than ϵ. Across six-passage dataset–model cells, their frequencies range from 2.2% to 10.6%. This retrospective comparison measures support–utility disagreement in the local pools. Eight-passage MMR and semantic-deduplication comparisons use the same similarity representations and selection definitions as the six-passage analysis. Their mean normalized $b = 3$ regrets are 1.148 and 1.060, respectively, versus 1.063 for BGE, quantifying the residual utility-selection gap for each policy. Full definitions and numerical comparisons appear in Appendix B.

5.3 Generated reward has structure, but a different optimum

Complete generated-reward lattices also exhibit negative marginals and increasing returns (Table 5). Their sign rates use a numerical tolerance of 10^{-10}; likelihood materiality uses its own calibrated scale. Requiring an F1 difference greater than 0.10 still leaves positive-Hessian rates of 9.39% and 7.99% in A and B. A post hoc cardinality breakdown on the same 90 questions gives F1 positive-Hessian rates of 14.48/10.64/9.31/8.44/9.70% in A and 13.63/9.45/7.76/7.26/6.78% in B for $k = 0, \ldots, 4$. The low-k point rates and Track A's rebound at the largest base cardinality reveal a heterogeneous cardinality profile (Appendix J).

Likelihood tracks reward across subsets while selecting a different optimum. Rank association is 0.553 in Track A and 0.489 in Track B. At $b = 3$, selecting the exact U optimum leaves mean F1 regret of 0.184 and 0.175 against the exact F1 optimum (Figure 2). The answer-aware F1 oracle supplies the diagnostic upper bound.

The paired contrasts quantify this alignment (Table 6). Exact-U minus fixed-U-greedy F1 is -0.051 in A and -0.028 in B, with paired intervals $[-0.108, 0.003]$ and $[-0.084, 0.026]$. The paired intervals do not establish a generated-F1 advantage for exact likelihood selection over fixed likelihood-greedy.

Table 6: **Generated-answer quality under matched selection policies.** Complete generation at $b = 3$. Positive differences favor exact likelihood selection; negative differences favor the named comparator. Brackets are paired 95% question-cluster intervals. Exact F1 is a reward oracle. Additional selectors and EM appear in Appendix E.

Selector	A: F1	A: $F_1(S_U^*) - F_1(S_\pi)$	B: F1	B: $F_1(S_U^*) - F_1(S_\pi)$
Exact F1	0.868	-0.184 [-0.235, -0.135]	0.723	-0.175 [-0.227, -0.125]
Exact U	0.684	0.000 [0.000, 0.000]	0.548	0.000 [0.000, 0.000]
Fixed U-greedy	0.735	-0.051 [-0.108, 0.003]	0.576	-0.028 [-0.084, 0.026]
Stopped U-greedy	0.648	0.036 [-0.003, 0.076]	0.507	0.041 [0.006, 0.081]
Singleton U	0.637	0.047 [-0.022, 0.117]	0.556	-0.008 [-0.074, 0.057]
BGE	0.496	0.188 [0.112, 0.269]	0.475	0.073 [-0.001, 0.145]
BM25	0.394	0.290 [0.197, 0.380]	0.342	0.206 [0.119, 0.294]

Against stopped greedy, the corresponding interval is positive in B and includes zero in A. The paired results quantify the distinction between landscape association and target-optimal selection.

Direct reward-greedy diagnostics isolate a further distinction. At $b = 3$, fixed F1-greedy has F1 regret 0.087 $[0.054, 0.122]$ in A and 0.045 $[0.020, 0.074]$ in B; stopped F1-greedy has regret 0.100 $[0.062, 0.142]$ and 0.036 $[0.013, 0.062]$ (Appendix E). These diagnostics use F1 marginals directly. Their residual regret includes both selected-size and combination effects, extending the selection diagnosis to the reward target itself.

The earlier selected-set evaluation over all 540 questions and three models gave positive mean exact-U minus fixed-greedy F1 differences, 0.021 in A and 0.016 in B. Restricting that same parent evaluation to Qwen and Gemma yields -0.0032 and $+0.0005$; the new subcohort gives the values above. This post hoc decomposition identifies the model and cohort composition associated with the different aggregate contrasts (Appendix F).

5.4 Cross-size attenuation depends on cardinality weighting

The matched extension preserves the six original candidates, their text and relative order, and the calibration scale. Even so, a uniform average over all lattice tests changes the distribution of base-set size $k = |S|$ when m changes. At size k, there are $\binom{m}{k}\binom{m-k}{2}$ Hessian tests. A raw full-lattice comparison therefore mixes changes within a cardinality stratum with changes in its weight.

On matched questions and models, the positive-Hessian material rate falls by 2.85 percentage points from $m = 6$ to $m = 8$. Weighting both sizes by the six-passage reference distribution on common cardinalities instead gives $+0.52$ points, with a 95% interval $[+0.07, +0.99]$. The corresponding standardized difference in mean positive-Hessian magnitude, normalized by the calibration scale, is -0.0025, with interval $[-0.0086, +0.0037]$. The materiality threshold is held fixed throughout. The contrast between these estimands identifies sensitivity to cardinality weighting (Appendix G). Within common cardinalities, the point estimates of material positive-Hessian rate differences are positive at each $k = 0, \ldots, 4$, whereas mean positive magnitude has mixed signs. At $k = 3$, the normalized magnitude difference is -0.0100 $[-0.0197, -0.0011]$; the other four magnitude intervals contain zero. These pointwise paired intervals resolve the cardinality-specific magnitude changes (Appendix J).

6 Discussion and limitations

The diagnostic separates three actionable questions: how utility responds to evidence, where a selector loses value, and whether its target agrees with generated-answer quality. Exact losses and complete reward landscapes turn these questions into distinct measurements. A surrogate's guarantee applies to its own objective; failure of a sufficient greedy condition does not, by itself, establish that a practical selector is ineffective.

Interface and interpretation. The estimand conditions on short-form multi-hop answers, six or eight candidates, fixed passage representations, serialization, and evaluated generators. Centered

U can have either sign and has a calibration distinct from reward calibration. The algebraic loss decomposition separates optimization gaps. Positive Hessians and pair escapes describe the joint effect of content, position, and exposure under that interface, rather than identifying neural or semantic causes. Pair escapes supply local improving moves, not a global optimization guarantee. Likelihood uses a canonical answer, with official-alias sensitivity reported separately.

Cohorts and uncertainty. The complete-generation study reuses 90 primary questions with Qwen and Gemma; its 23,040 generations are repeated measurements. Question-cluster intervals are pointwise unless specified otherwise. Cardinality and cohort analyses are post hoc, and their intervals do not establish simultaneous trends or isolate the cause of cross-size changes. Permutation controls identify local finite differences and only partly cover complete permutation-specific regret. For Gemma, first-turn truncation changes 13 answer scores while preserving all main $b = 3$ policy contrasts (Appendix F.1).

Oracle information and transfer. The answer-aware oracles diagnose target landscapes; they are not information-matched deployment baselines. Long-form quality, faithfulness, open-corpus retrieval, and new prompt or generator distributions require corresponding evaluations.

7 Conclusion

Complete evidence lattices reveal material non-monotonicity and increasing returns across controlled and support-blind local RAG pools. Exact decomposition distinguishes the price of filling a budget from loss due to selecting the wrong combination. Complete generated-answer evaluation confirms structural effects in reward while revealing that likelihood and answer-quality optima can diverge. EUG provides a reproducible diagnostic for these distinctions; a selector's geometry, stopping rule, and evaluation target should be validated separately.

Reproducibility statement

The numerical supplement supplies likelihood cubes, complete-generation scores and outputs, calibration scales, and table/figure inputs. CPU scripts reconstruct finite differences, exact losses, selected paired intervals, answer scores, and cardinality-resolved analyses. A 36-cube subset provides a small example. The appendices and configuration files specify model revisions, prompts, selection, and decoding. Model inference and historical controls require their original configurations; datasets and weights retain their providers' access conditions and licenses.

Ethics statement

The study evaluates public research benchmarks and released models. Its empirical objects are benchmark examples and model outputs. Benchmark answers and model outputs can reflect source errors and biases. High-stakes RAG applications require task-specific factuality and safety evaluation beyond the likelihood and token-overlap targets studied here.

AI use disclosure

ChatGPT and Codex assisted with hypothesis refinement, conceptual and mathematical formalization, experimental design and implementation, data processing, statistical analysis, numerical verification, literature checking, figure construction, and manuscript drafting and editing. Final workflow diagrams are editable vector drawings checked against the measurement design. The authors take responsibility for the scientific claims, data, citations, and submitted text.

References

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pp. 335–336, 1998. doi: 10.1145/290941.291025.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 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. doi: 10.1145/1571941.1572114.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of COLING*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580.

Yu Huo, Kun Zeng, Siyu Zhang, Yuquan Lu, Cheng Yang, Yifu Guo, and Xiaoying Tang. RepoShapley: Shapley-enhanced context filtering for repository-level code completion. *arXiv preprint arXiv:2601.03378*, 2026. URL https://arxiv.org/abs/2601.03378. ACL 2026 Findings.

Siddharth Jain and Venkat Narayan Vedam. CUE-R: Beyond the final answer in retrieval-augmented generation. *arXiv preprint arXiv:2604.05467*, 2026. URL https://arxiv.org/abs/2604.05467.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Yi Jiang, Sendong Zhao, Jianbo Li, Bairui Hu, Yanrui Du, Haochun Wang, and Bing Qin. OptiSet: Unified optimizing set selection and ranking for retrieval-augmented generation. *arXiv preprint arXiv:2601.05027*, 2026. URL https://arxiv.org/abs/2601.05027.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 17606–17619, 2025. doi: 10.18653/v1/2025.acl-long.861.

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.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Ikhtiyor Nematov, Tarik Kalai, Elizaveta Kuzmenko, Gabriele Fugagnoli, Dimitris Sacharidis, Katja Hose, and Tomer Sagi. Source attribution in retrieval-augmented generation. *arXiv preprint arXiv:2507.04480*, 2025. URL https://arxiv.org/abs/2507.04480.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009. doi: 10.1561/1500000019.

Lichang Song, Ting Long, and Yi Chang. Rethinking retrieval-augmented generation as a cooperative decision-making problem. *arXiv preprint arXiv:2602.18734*, 2026. URL https://arxiv.org/abs/2602.18734.

Jiashuo Sun, Pengcheng Jiang, Saizhuo Wang, Jiajun Fan, Heng Wang, Siru Ouyang, Ming Zhong, Yizhu Jiao, Chengsong Huang, Xueqiang Xu, Pengrui Han, Peiran Li, Jiaxin Huang, Ge Liu, Heng Ji, and Jiawei Han. Rethinking the reranker: Boundary-aware evidence selection for robust retrieval-augmented generation. *arXiv preprint arXiv:2602.03689*, 2026. URL https://arxiv.org/abs/2602.03689.

Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. Is relevance propagated from retriever to generator in RAG? *arXiv preprint arXiv:2502.15025*, 2025. URL https://arxiv.org/abs/2502.15025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. InfoGain-RAG: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of EMNLP*, 2025. doi: 10.18653/v1/2025.emnlp-main.365.

Ziting Wang, Haitao Yuan, Wei Dong, Gao Cong, and Feifei Li. CARROT: A learned cost-constrained retrieval optimization system for RAG. *arXiv preprint arXiv:2411.00744*, 2026. URL https://arxiv.org/abs/2411.00744. Accepted to ICDE 2026; earlier title used CORAG.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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

Hamed Zamani and Michael Bendersky. Stochastic RAG: End-to-end retrieval-augmented generation through expected utility maximization. In *Proceedings of SIGIR*, 2024. URL https://arxiv.org/abs/2405.02816.

Bingxue Zhang, Jianying Jia, and Feida Zhu. Covering the unseen: Information demand coverage optimization for retrieval-augmented generation. *arXiv preprint arXiv:2606.29328*, 2026. URL https://arxiv.org/abs/2606.29328.

A MEASUREMENT PROTOCOL AND STUDY POPULATIONS

A.1 QUESTIONS, POOLS, AND FIXED REPRESENTATIONS

The primary roster contains 60 calibration and 180 final questions from each of 2WikiMultiHopQA, HotpotQA distractor, and MuSiQue answerable. The two splits are disjoint. Eligibility requires a nonempty canonical answer, at least one annotated support passage, and at least six distinct normalized candidates. The controlled construction includes all annotated supports and fills remaining slots with non-supports ranked by BGE, with deterministic BM25 and candidate-identifier tie breaks. The support-blind construction combines BM25 and BGE rankings by reciprocal-rank fusion without consulting support labels or answer correctness. Both operate within benchmark-supplied local candidate pools.

A passage is rendered once as a title and text and capped at 256 tokens using the fixed Qwen reference tokenizer. The representation is fixed across subsets. A deterministic global order assigns each candidate a permanent position. Serialization retains that relative order and wraps each selected passage in an evidence block whose label remains its original position. Subset membership therefore varies at fixed passage labels and relative order. Removing an earlier block can shift later token positions, making serialization an explicit part of the set function.

The eight-passage roster contains 60 primary final questions per dataset, evaluated with Qwen and Gemma in Track A. The original six candidates, their representations, and their relative order are preserved. The complete-generation roster contains 30 primary final questions per dataset, selected by a fixed hash ordering before supplementary generation, and evaluates both tracks with those two models. Selection is outcome-independent within the primary population. All 360 generation cubes contain exactly 64 scored subsets.

A.2 PROMPTS, LIKELIHOOD, AND DECODING

The instruction is: "You are a question-answering system. Use the supplied evidence when it is available. Return only the shortest correct answer span and do not explain." The user text has the following structure:

```
Question:
{question}
Evidence:
{serialized evidence, or [NO EVIDENCE]}
Return only the answer.
```

The prefix is constructed from one system message and one user message by the pinned tokenizer's chat template, with tokenization and the generation prompt enabled. Teacher forcing appends one ASCII space followed by the stripped canonical answer, with no added target special tokens and no scored EOS. The mean log probability includes every target token. The empty-evidence score is computed separately for each query and model. Subtracting it leaves every marginal and Hessian unchanged.

The three checkpoints are Qwen2.5-7B-Instruct, Gemma-3-4B-it, and Mistral-7B-Instruct-v0.3. Configuration files specify the fixed checkpoint revisions. Inference uses BF16 weights, no weight quantization or CPU offload, batch size one, and float32 log-softmax for likelihood. The experiments ran on one RTX 5090 workstation. Calibration revealed padding-shape sensitivity in small likelihood differences, motivating batch size one.

Generated answers use deterministic greedy decoding, one beam, and at most 64 new tokens. The implementation explicitly supplies the tokenizer EOS token and its fixed padding rule. Decoded special tokens are removed before scoring. Appendix F.1 gives the Gemma end-of-turn sensitivity. EM and token F1 use lowercasing, ASCII punctuation removal, removal of the articles *a*, *an*, and *the*, and whitespace normalization. F1 uses multiset token overlap; the score is the maximum over the frozen official aliases. The reward evaluation consists of these exact-match and token-overlap measures.

A.3 CALIBRATION AND AGGREGATION

For a dataset–model–track cell, the calibration scale is the interquartile range of *all centered query–subset utilities* from its 60 complete calibration cubes, with a floor of 10^{-6}. The calibration utilities, rather than final marginal effects, determine the scale. The materiality threshold is $\epsilon = \max(10^{-3}, 0.05s)$. The matching six-passage scale is reused at eight passages. A complete m-candidate cube has $m2^{m-1}$ directed addition edges and $\binom{m}{2}2^{m-2}$ unordered-pair Hessian tests: 192 and 240 for $m = 6$, and 1,024 and 1,792 for $m = 8$.

Each cube is first reduced to query–model summaries. Pooled statistics average models within a query, queries within a dataset, and then datasets equally. Percentile bootstrap resampling is performed at the question level within each dataset, keeping paired model and method outputs together, with 10,000 draws. Cellwise estimates resample questions for that cell. Undefined rank correlations are omitted only from the correlation aggregation, with their counts reported; constant cubes still contribute zero sign rates. Intervals are marginal to each specified comparison. The supplement provides the randomization needed for the reported numerical reconstruction.

B EXACT SELECTION COMPARISONS

B.1 ALGORITHMS AND TIES

The at-most-budget oracle enumerates every bitmask with population count at most b, including the empty set. The exact-size oracle restricts the count to b. Both break utility ties by the smallest numeric bitmask. Fixed true-marginal greedy starts empty and takes the largest currently available marginal, breaking ties by the original candidate index, for exactly b additions. The stopped version uses the same choices but terminates when the best marginal is nonpositive. Singleton-U ranks candidates by their empty-context utilities and takes the top b. Reward oracles and reward-greedy methods apply these rules to the recorded F1 or EM cube rather than to U.

BM25 (Robertson & Zaragoza, 2009) and BGE select the top b local candidates by their fixed scores. MMR uses equal weights on min–max-normalized BGE relevance and maximal cosine similarity to an already selected passage; the redundancy term is zero for the first selection (Carbonell & Goldstein, 1998). Ties use the fixed candidate identifier. Semantic deduplication scans the BGE ranking and skips a passage if its cosine similarity to any selected passage is at least 0.90; it can return fewer than b passages. These comparison policies use fixed, untuned hyperparameters.

Candidate-pair similarities use normalized BGE-M3 CLS representations, with the checkpoint revision held fixed (Chen et al., 2024). The eight-passage comparison uses the same BGE representation and precision for every candidate pair. The reconstruction inventory records 5,040 required pairs; overlapping entries agree across the two pool sizes. Table 7 reports the resulting comparison.

Table 7: **Diversity heuristics do not close the measured utility gap.** Eight-passage Track A mean normalized regret, over 180 questions and two models. Budgets two through four are shown for comparability with the primary analysis. Intervals are question-cluster 95% intervals.

Selector	$b = 2$	$b = 3$	$b = 4$
Fixed U-greedy	0.106 [0.069, 0.152]	0.171 [0.125, 0.221]	0.211 [0.157, 0.272]
Stopped U-greedy	0.039 [0.022, 0.060]	0.075 [0.048, 0.109]	0.093 [0.061, 0.130]
Singleton U	0.349 [0.281, 0.425]	0.509 [0.408, 0.617]	0.596 [0.492, 0.705]
BM25	1.287 [1.117, 1.463]	1.268 [1.098, 1.446]	1.100 [0.942, 1.275]
BGE	1.077 [0.910, 1.249]	1.063 [0.894, 1.237]	0.929 [0.784, 1.083]
MMR	1.228 [1.056, 1.403]	1.148 [0.989, 1.313]	1.053 [0.897, 1.211]
Semantic dedup	1.078 [0.915, 1.247]	1.060 [0.900, 1.232]	0.920 [0.778, 1.079]

B.2 WHY THE LOSS DECOMPOSITION HOLDS

Because every set of size b is feasible for the at-most-b problem, $O_{\le b} \ge O_{=b}$. Because fixed greedy returns a set of size b, $O_{=b} \ge U(G_b)$. Adding and subtracting $O_{=b}$ gives Equation 4, with both terms nonnegative. The identity follows solely from feasible-set inclusion and the value of the selected set.

Dividing by the same positive scale preserves the equality. At $b = m$, $O_{=m} = U(D)$ and $G_m = D$, so same-cardinality search loss is exactly zero. The cardinality price can remain nonzero if a proper subset has higher utility.

Stopped greedy follows a trajectory-dependent policy and can select a different number of elements. An initially unfavorable addition can lead to a favorable combination, so either stopping or continuing can produce the lower regret on an individual query. Its regret is evaluated separately from the same-cardinality term.

C COMPLETE LIKELIHOOD GEOMETRY AND ASSOCIATION ESTIMATES

The predeclared material-geometry criterion accepts a cell when either the material negative-marginal rate or the material positive-Hessian rate is at least 5%, with its corresponding lower confidence limit above 1%. In the observed complete cubes, both types of violation substantially exceed these references. Table 8 reports the rates and intervals for every cell.

The primary geometry–regret test relates normalized positive-Hessian mass to fixed-greedy regret. Within a fixed cube size, the mass is the sum of positive Hessians divided by s; replacing the sum by its mean merely multiplies it by a constant and leaves its Spearman association unchanged. The joint rule at a budget requires correlation at least 0.15 with a positive lower limit, a top-minus-bottom-quartile regret gap at least 0.05 calibration units with a positive lower limit, and exact-optimum set mismatch of at least 10%. Table 9 retains every correlation and quartile gap at $b = 3$, including those whose intervals contain zero. Eight-passage associations are displayed for comparison, separate from the extension's original material-geometry criterion.

Table 8: **All likelihood-geometry cells.** Percentages of material finite differences after query-level aggregation, with 95% intervals. There are 180 questions per six-passage cell and 60 per eight-passage cell. The matching original scale is used at both sizes.

Track, m	Dataset	Model	Negative marginal (%)	Positive Hessian (%)
A, 6	2Wiki	Gemma	36.7 [35.1, 38.3]	40.1 [38.8, 41.3]
A, 6	2Wiki	Mistral	37.1 [35.2, 39.0]	41.6 [40.1, 43.1]
A, 6	2Wiki	Qwen	35.2 [33.7, 36.8]	41.7 [40.4, 42.9]
A, 6	HotpotQA	Gemma	29.6 [28.2, 31.1]	37.4 [36.2, 38.6]
A, 6	HotpotQA	Mistral	27.8 [25.8, 29.7]	35.5 [33.7, 37.2]
A, 6	HotpotQA	Qwen	30.2 [28.7, 31.8]	38.1 [36.7, 39.6]
A, 6	MuSiQue	Gemma	34.4 [32.9, 36.0]	40.6 [39.5, 41.8]
A, 6	MuSiQue	Mistral	32.2 [30.8, 33.6]	38.5 [37.3, 39.7]
A, 6	MuSiQue	Qwen	29.4 [27.9, 30.8]	38.1 [36.9, 39.3]
A, 8	2Wiki	Gemma	36.7 [34.0, 39.2]	38.8 [36.8, 40.7]
A, 8	2Wiki	Qwen	33.3 [30.8, 35.7]	39.8 [37.3, 42.1]
A, 8	HotpotQA	Gemma	31.3 [28.5, 34.1]	35.5 [33.3, 37.5]
A, 8	HotpotQA	Qwen	30.2 [27.4, 32.9]	35.5 [33.0, 37.9]
A, 8	MuSiQue	Gemma	34.9 [32.4, 37.3]	38.6 [37.0, 40.1]
A, 8	MuSiQue	Qwen	27.2 [25.0, 29.3]	34.0 [32.2, 35.8]
B, 6	2Wiki	Gemma	37.9 [36.4, 39.5]	40.4 [39.2, 41.6]
B, 6	2Wiki	Mistral	37.2 [35.3, 39.1]	42.7 [41.2, 44.2]
B, 6	2Wiki	Qwen	36.6 [35.0, 38.2]	42.9 [41.7, 44.1]
B, 6	HotpotQA	Gemma	30.4 [28.9, 31.9]	37.9 [36.6, 39.1]
B, 6	HotpotQA	Mistral	27.3 [25.4, 29.2]	35.7 [33.9, 37.5]
B, 6	HotpotQA	Qwen	30.2 [28.6, 31.8]	38.1 [36.7, 39.5]
B, 6	MuSiQue	Gemma	37.2 [35.6, 38.7]	41.9 [40.7, 43.0]
B, 6	MuSiQue	Mistral	36.1 [34.5, 37.6]	41.7 [40.5, 42.8]
B, 6	MuSiQue	Qwen	33.2 [31.8, 34.6]	40.6 [39.5, 41.8]

Table 9: **Geometry–regret association across datasets and generators.** Positive-Hessian mass versus fixed-greedy regret at $b = 3$. The quartile gap compares the upper and lower geometry quartiles; regret is divided by s. Brackets are 95% query-bootstrap intervals. These are observational relationships within the measured cubes.

Track, m	Dataset	Model	Spearman ρ	Quartile gap / s
A, 6	2Wiki	Gemma	0.254 [0.097, 0.402]	0.411 [0.221, 0.673]
A, 6	2Wiki	Mistral	0.264 [0.110, 0.401]	0.681 [0.316, 1.096]
A, 6	2Wiki	Qwen	0.359 [0.207, 0.499]	1.065 [0.618, 1.589]
A, 6	HotpotQA	Gemma	0.256 [0.100, 0.398]	0.333 [0.158, 0.553]
A, 6	HotpotQA	Mistral	0.084 [-0.059, 0.224]	0.195 [0.017, 0.423]
A, 6	HotpotQA	Qwen	0.303 [0.161, 0.436]	0.228 [0.118, 0.364]
A, 6	MuSiQue	Gemma	0.347 [0.202, 0.474]	0.252 [0.144, 0.372]
A, 6	MuSiQue	Mistral	0.019 [-0.135, 0.170]	0.097 [0.005, 0.209]
A, 6	MuSiQue	Qwen	0.270 [0.124, 0.407]	0.120 [0.054, 0.204]
A, 8	2Wiki	Gemma	0.406 [0.136, 0.629]	0.527 [0.204, 0.835]
A, 8	2Wiki	Qwen	0.380 [0.113, 0.612]	0.935 [0.398, 1.644]
A, 8	HotpotQA	Gemma	0.387 [0.097, 0.629]	0.422 [0.158, 0.704]
A, 8	HotpotQA	Qwen	0.254 [0.011, 0.481]	0.194 [0.026, 0.407]
A, 8	MuSiQue	Gemma	0.373 [0.144, 0.561]	0.192 [0.048, 0.390]
A, 8	MuSiQue	Qwen	0.123 [-0.125, 0.362]	0.015 [-0.024, 0.184]
B, 6	2Wiki	Gemma	0.322 [0.171, 0.468]	0.538 [0.314, 0.783]
B, 6	2Wiki	Mistral	0.310 [0.155, 0.453]	0.761 [0.416, 1.114]
B, 6	2Wiki	Qwen	0.359 [0.208, 0.493]	1.093 [0.674, 1.470]
B, 6	HotpotQA	Gemma	0.272 [0.119, 0.412]	0.335 [0.153, 0.586]
B, 6	HotpotQA	Mistral	0.117 [-0.033, 0.259]	0.229 [0.052, 0.469]
B, 6	HotpotQA	Qwen	0.323 [0.184, 0.451]	0.287 [0.139, 0.422]
B, 6	MuSiQue	Gemma	0.420 [0.281, 0.547]	0.632 [0.400, 0.861]
B, 6	MuSiQue	Mistral	0.313 [0.158, 0.455]	0.402 [0.213, 0.607]
B, 6	MuSiQue	Qwen	0.358 [0.223, 0.482]	0.282 [0.161, 0.441]

C.1 MATERIALITY SENSITIVITY

The primary multiplier remains 0.05. Supplemental recomputation uses $\epsilon_\alpha = \max(10^{-3}, \alpha s)$ at the fixed multipliers in Table 10. The fixed multipliers vary the reporting scale while retaining the data, selectors, and exact optima. Selection regret is consequently unchanged.

Table 10: **Material violations persist across the declared reporting thresholds.** Pooled question-cluster estimates and 95% intervals. Percentages use $\Delta < -\epsilon_\alpha$ and $H > \epsilon_\alpha$.

Track, m	α	Negative (%)	Positive H (%)
A, 6	0.01	40.2 [39.6, 40.8]	46.5 [46.1, 46.9]
A, 6	0.025	36.8 [36.2, 37.5]	43.5 [43.0, 44.0]
A, 6	0.05	32.5 [31.8, 33.2]	39.1 [38.5, 39.7]
A, 6	0.1	26.3 [25.5, 27.0]	32.4 [31.7, 33.1]
A, 8	0.01	41.5 [40.6, 42.3]	45.9 [45.4, 46.4]
A, 8	0.025	37.4 [36.4, 38.4]	42.2 [41.5, 43.0]
A, 8	0.05	32.2 [30.9, 33.5]	37.0 [35.9, 38.0]
A, 8	0.1	25.2 [23.8, 26.7]	29.5 [28.1, 30.9]
B, 6	0.01	41.3 [40.7, 41.8]	47.0 [46.6, 47.5]
B, 6	0.025	38.1 [37.4, 38.8]	44.2 [43.7, 44.7]
B, 6	0.05	34.0 [33.3, 34.7]	40.2 [39.6, 40.8]
B, 6	0.1	27.8 [27.1, 28.6]	33.8 [33.1, 34.5]

D STOPPING, PAIR ESCAPES, AND BUDGET CURVES

At a stopped state S, an escape diagnostic is eligible only if at least two passages remain available and at least two budget slots remain. It records whether every available singleton gain is nonpositive (up to numerical tolerance) while some pair satisfies $U(S \cup \{i, j\}) - U(S) > \epsilon$. The conditional rates at $b = 3$ are 116/728 in six-passage Track A, 129/794 in six-passage Track B, and 36/134 at eight passages. Their query-cluster 95% intervals are [13.3, 18.6]%, [13.6, 19.0]%, and [19.0, 35.1]%, respectively. The denominator in each rate is the set of eligible query–model states.

A concrete local witness. A HotpotQA Track-B Qwen cube illustrates the distinction at $b = 4$. The stopped state is $S = \{d_6\}$ with utility 4.9913. Adding d_2 alone changes utility by -0.6556; adding d_3 alone changes it by -2.5739. Joint addition changes utility by $+0.4277$, above $\epsilon = 0.1774$, yielding a positive Hessian of 3.6572. The annotations mark d_2 as support and d_3 as a distractor; these labels describe the benchmark annotations. Continuing fixed greedy eventually reaches the exact optimum $\{d_2, d_4, d_5, d_6\}$ with utility 5.8238. The case is a deterministic illustration selected from cubes satisfying the stated pattern.

D.1 BUDGET DEPENDENCE

Figures 3–5 extend the exact accounting to every possible budget using the same complete cubes. Fixed regret equals cardinality price plus search loss; the stopped curve evaluates a separate policy. At the final budget, the sole full-size set makes the search-loss term exactly zero.



Figure 3: **Track A, six passages: larger budgets leave a cardinality price.** Exact mean normalized losses over the primary question population. Full query-cluster intervals are in the numeric table accompanying the figure.

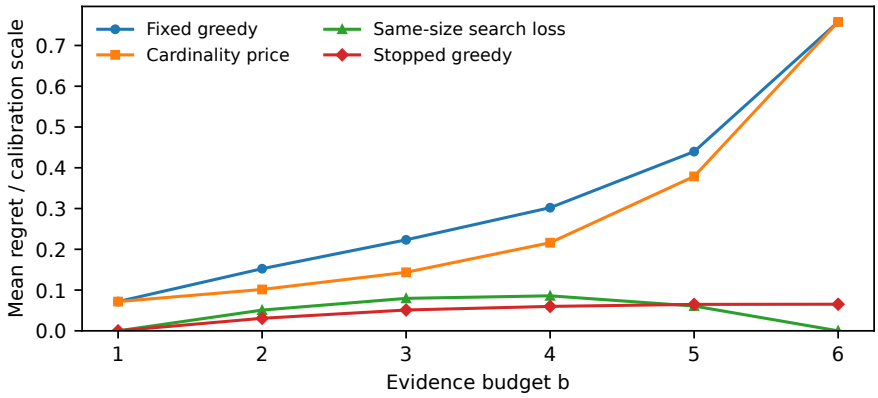

Figure 4: **Track B, six passages: stopping and same-size search remain distinct.** The same accounting in the support-blind local pools. Means weight questions and datasets as in the primary analysis.



Figure 5: **Track A, eight passages: complete enumeration covers all budgets.** This cohort contains 180 matched questions and two models, not the entire six-passage population.

E COMPLETE GENERATED-REWARD EVALUATION

E.1 CELLS, CONSTANTS, AND REWARD MAGNITUDE

Table 11 shows every generated-F1 cell. Complete enumeration avoids selecting only the subsets exposed by named selectors. The independent units are the 90 questions shared by two models; subsets and models are repeated measurements within those units. A complete reward cube can be constant even when its likelihood cube is not. Such a cube has zero reward finite differences but no defined rank correlation. Table 12 records these cases explicitly, preserving undefined correlations.

Table 11: **Generated-F1 structure and alignment in every cell.** Each row contains 30 complete cubes. Rates use a numerical sign tolerance; intervals resample questions. Correlation estimates use only nonconstant reward cubes.

Track	Dataset	Model	F1 negative (%)	F1 positive H (%)	$\rho(U, F_1)$
A	2Wiki	Gemma	11.0 [8.4, 13.7]	12.2 [9.4, 15.2]	0.521 [0.390, 0.637]
A	2Wiki	Qwen	7.6 [5.3, 10.3]	11.7 [8.4, 15.1]	0.580 [0.484, 0.666]
A	HotpotQA	Gemma	6.0 [4.4, 7.7]	8.4 [6.2, 10.7]	0.605 [0.483, 0.711]
A	HotpotQA	Qwen	6.1 [4.1, 8.1]	10.0 [7.0, 13.1]	0.548 [0.395, 0.676]
A	MuSiQue	Gemma	7.0 [5.2, 9.0]	9.0 [6.6, 11.8]	0.496 [0.368, 0.614]
A	MuSiQue	Qwen	5.5 [3.6, 7.6]	7.3 [4.8, 10.0]	0.628 [0.506, 0.725]
B	2Wiki	Gemma	11.7 [9.0, 14.5]	12.7 [9.7, 15.9]	0.483 [0.338, 0.618]
B	2Wiki	Qwen	7.7 [5.0, 10.8]	11.0 [7.8, 14.3]	0.565 [0.444, 0.675]
B	HotpotQA	Gemma	5.2 [3.5, 7.0]	7.7 [5.3, 10.2]	0.593 [0.459, 0.711]
B	HotpotQA	Qwen	6.0 [3.9, 8.4]	8.2 [5.5, 11.2]	0.529 [0.387, 0.653]
B	MuSiQue	Gemma	5.4 [3.4, 7.4]	6.1 [3.9, 8.2]	0.408 [0.222, 0.583]
B	MuSiQue	Qwen	3.8 [1.9, 6.0]	4.6 [2.5, 7.0]	0.446 [0.211, 0.642]

Table 12: **Constant reward cubes remain in the rate denominator.** Each track has 180 cubes from 90 questions and two models. All-zero cubes are a subset of constant cubes. Pooled correlations average finite model results within questions and datasets, rather than treating all finite cubes as independent.

Track	Reward	Constant / 180	All zero / 180	Finite correlations	ρ on finite cubes
A	F1	14	9	166	0.553 [0.495, 0.608]
A	EM	48	44	132	0.551 [0.497, 0.604]
B	F1	37	32	143	0.489 [0.414, 0.563]
B	EM	73	69	107	0.555 [0.488, 0.620]

Table 13: **The F1 sign pattern is not confined to changes near numerical zero.** Negative marginals and positive Hessians exceeding each listed absolute F1 magnitude, with question-cluster 95% intervals. These thresholds are on F1, not on the likelihood calibration scale.

Track	Magnitude threshold	F1 negative (%)	F1 positive H (%)
A	0.01	7.19 [6.23, 8.21]	9.76 [8.55, 11.05]
A	0.05	7.10 [6.18, 8.11]	9.63 [8.45, 10.88]
A	0.1	6.93 [6.04, 7.87]	9.39 [8.26, 10.59]
B	0.01	6.61 [5.54, 7.76]	8.35 [7.09, 9.66]
B	0.05	6.45 [5.43, 7.52]	8.18 [6.97, 9.42]
B	0.1	6.31 [5.33, 7.34]	7.99 [6.83, 9.18]

E.2 ALL BUDGET-THREE POLICY COMPARISONS

Table 14 includes the original likelihood-based and retrieval policies as well as F1- and EM-based diagnostic oracles. Every reported set is scored on its exact recorded generation condition. Reward ties are common, so two distinct maximizing bitmasks can both have zero reward regret. Reward regret therefore measures answer quality even when maximizing sets differ.

For F1 and EM, the greedy rules recompute the exact marginal after each addition. These methods have access to all required generated rewards and reference answers. They are answer-aware landscape diagnostics. The U and reward oracles provide exact targets, while BM25 and BGE supply retrieval-based selection references.

Table 14: **All generated-reward policies at $b = 3$.** Absolute scores and regret against the exact at-most-budget F1 optimum. Brackets are question-cluster 95% intervals. "Fixed" fills the budget; "stopped" terminates on a nonpositive marginal of its own target.

Track	Selector	F1	EM	F1 regret
A	Exact U	0.684 [0.616, 0.747]	0.578 [0.500, 0.656]	0.184 [0.135, 0.235]
A	Fixed U-greedy	0.735 [0.671, 0.797]	0.628 [0.544, 0.706]	0.132 [0.092, 0.175]
A	Stopped U-greedy	0.648 [0.583, 0.713]	0.533 [0.456, 0.611]	0.219 [0.168, 0.275]
A	Singleton U	0.637 [0.562, 0.709]	0.528 [0.444, 0.611]	0.231 [0.176, 0.289]
A	BM25	0.394 [0.310, 0.479]	0.317 [0.239, 0.400]	0.474 [0.396, 0.553]
A	BGE	0.496 [0.414, 0.576]	0.394 [0.306, 0.478]	0.372 [0.300, 0.444]
A	MMR	0.458 [0.374, 0.542]	0.367 [0.278, 0.456]	0.410 [0.335, 0.487]
A	Semantic dedup	0.490 [0.408, 0.571]	0.389 [0.306, 0.478]	0.377 [0.306, 0.450]
A	Exact F1	0.868 [0.817, 0.914]	0.756 [0.678, 0.828]	0.000 [0.000, 0.000]
A	Fixed F1-greedy	0.781 [0.719, 0.838]	0.672 [0.594, 0.750]	0.087 [0.054, 0.122]
A	Stopped F1-greedy	0.767 [0.709, 0.822]	0.644 [0.567, 0.722]	0.100 [0.062, 0.142]
A	Exact EM	0.769 [0.696, 0.838]	0.756 [0.678, 0.828]	0.099 [0.061, 0.141]
B	Exact U	0.548 [0.474, 0.624]	0.450 [0.367, 0.533]	0.175 [0.125, 0.227]
B	Fixed U-greedy	0.576 [0.497, 0.657]	0.461 [0.372, 0.550]	0.146 [0.103, 0.192]
B	Stopped U-greedy	0.507 [0.435, 0.583]	0.406 [0.322, 0.489]	0.215 [0.161, 0.272]
B	Singleton U	0.556 [0.477, 0.635]	0.461 [0.372, 0.550]	0.166 [0.117, 0.219]
B	BM25	0.342 [0.260, 0.428]	0.283 [0.200, 0.367]	0.380 [0.304, 0.458]
B	BGE	0.475 [0.394, 0.558]	0.383 [0.300, 0.472]	0.248 [0.186, 0.312]
B	MMR	0.424 [0.341, 0.506]	0.344 [0.261, 0.433]	0.298 [0.227, 0.373]
B	Semantic dedup	0.475 [0.394, 0.557]	0.383 [0.300, 0.472]	0.248 [0.185, 0.314]
B	Exact F1	0.723 [0.656, 0.789]	0.606 [0.522, 0.689]	0.000 [0.000, 0.000]
B	Fixed F1-greedy	0.677 [0.601, 0.751]	0.567 [0.478, 0.650]	0.045 [0.020, 0.074]
B	Stopped F1-greedy	0.686 [0.621, 0.751]	0.567 [0.483, 0.650]	0.036 [0.013, 0.062]
B	Exact EM	0.621 [0.540, 0.698]	0.606 [0.522, 0.683]	0.102 [0.063, 0.145]

F Parent evaluation and supplemental cohort

The parent selected-set evaluation and the complete-generation subcohort estimate different finite averages. Table 15 holds the contrast direction fixed while separating model composition and query selection. The 90 questions are a subset of the 540; the remaining 450-question mean provides a retrospective decomposition. Subcohort membership and selection rules were fixed before this decomposition. The three-model parent mean and the two-model supplemental mean consequently summarize different finite populations.

Table 15: **Model and cohort composition change the aggregate contrast.** Post hoc descriptive F1 differences, exact likelihood oracle minus fixed likelihood-greedy, at $b = 3$. These nested rows are not independent replications.

Population	Questions	Track A	Track B
Parent, three models	540	+0.0211	+0.0163
Parent, Qwen and Gemma	540	-0.0032	+0.0005
Complete-generation subcohort, Qwen and Gemma	90	-0.0513	-0.0282
Remaining parent, Qwen and Gemma	450	+0.0065	+0.0063

F.1 End-of-turn sensitivity

Generation uses the tokenizer EOS token as its stopping criterion. Gemma's separate `<end_of_turn>` marker can occur before that endpoint, allowing subsequent content in a recorded

trajectory. Removing special tokens often removes only repeated markers or whitespace, but in 13 of the 11,520 Gemma conditions, spanning four questions, post-turn content affects F1 or EM.

The post hoc first-turn sensitivity applies the same scorer to each recorded trajectory truncated at its first end-of-turn marker. The original decoding remains primary; model, prompt, subset membership, and reference answers remain fixed. Table 16 reports the affected geometry summaries. Every main $b = 3$ policy F1/EM value and paired contrast, including its interval, remains unchanged. The budget-three contrasts are stable under this rescoring of the recorded trajectories.

Table 16: **End-of-turn truncation changes a small number of scores, not the main policy contrast.** Percentages of F1 finite-difference signs over the complete generation cubes. The truncation column is a post-hoc sensitivity; frozen decoding remains primary.

Track	Quantity	Frozen decoding	First-turn truncation
A	Negative marginal	7.193	7.167
A	Positive Hessian	9.775	9.731
B	Negative marginal	6.617	6.586
B	Positive Hessian	8.363	8.289

G MATCHED CROSS-SIZE COMPARISONS

The paired comparison uses only questions and models available at both sizes. All 360 pairs preserve the six original candidates, their fixed text and order, and their calibration scale. New candidates can nevertheless change the distribution of the base set on which an addition or Hessian is measured. At base cardinality k, the numbers of edges and Hessians are

$$N^{\Delta}_{m,k} = \binom{m}{k}(m-k), \qquad N^{H}_{m,k} = \binom{m}{k}\binom{m-k}{2}. \tag{6}$$

An unstratified full-lattice rate weights each cardinality by these counts. The supplemental common-cardinality estimand instead uses

$$w^{\Delta}_{k} = \frac{\binom{6}{k}(6-k)}{192}, \quad k = 0, \ldots, 5; \qquad w^{H}_{k} = \frac{\binom{6}{k}\binom{6-k}{2}}{240}, \quad k = 0, \ldots, 4. \tag{7}$$

For either universe size, we compute the rate or mean normalized positive magnitude separately within each k, then apply these common six-passage weights. These reference-lattice weights define a common distribution over k restricted to shared cardinalities. Paired intervals preserve each query's two sizes and all its models.

Table 17: **A full-lattice average and a common-cardinality average answer different questions.** Paired $m = 8$ minus $m = 6$ material-rate differences on the same 180-question cohort, with question-cluster 95% intervals. The materiality threshold is unchanged.

Quantity	Weighting	$m = 8$ minus $m = 6$ (pp)
Negative marginals	Full lattice	-0.11 [-0.65, 0.44]
Negative marginals	Common cardinalities	1.46 [0.98, 1.94]
Positive Hessians	Full lattice	-2.85 [-3.37, -2.31]
Positive Hessians	Common cardinalities	0.52 [0.07, 0.99]

For normalized positive-Hessian magnitude, the common-cardinality difference is -0.00248, with interval $[-0.00858, +0.00374]$. The interval spans zero, whereas the standardized rate comparison reveals a dependence on cardinality weighting.

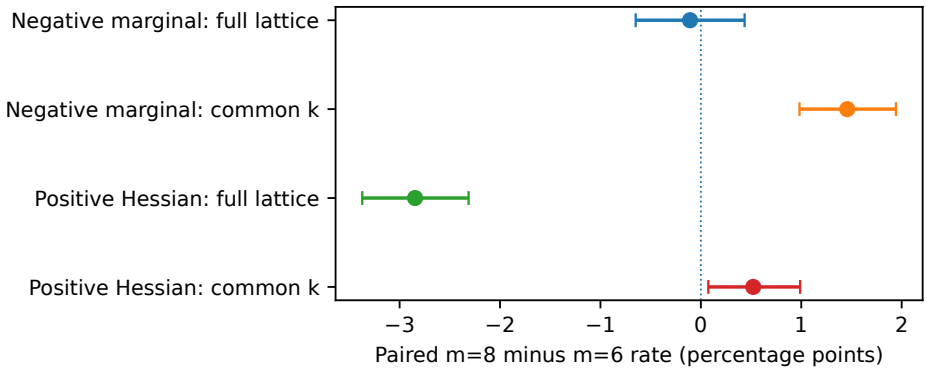

Figure 6: **Hessian-rate attenuation is sensitive to base-cardinality composition.** The common-cardinality comparison fixes the weights in Equation 7. It fixes cardinality weights while retaining the different candidate universes.

H Order, aliases, and measurement controls

H.1 What the permutation data identify

The historical order study uses 40 Track-A questions per dataset, all three models, and three fixed global permutations. Evidence labels, text, and membership remain fixed while block order changes. Its 1,080 noncanonical question–model–permutation conditions each contain 52–63 of the 64 subsets, covering a prescribed local set of finite differences, including 32 Hessian tests and their required subsets.

Table 18 compares the observed local finite differences with the same triples and edges in the canonical order. The specified subsets identify exact at-most-budget-three optima in 27 conditions. The remaining 1,053 identify the planned local finite differences. The table therefore compares matched local effects across orders.

Table 18: **Matched local sensitivity to evidence order.** Change relative to the canonical order on the same observed finite differences, in percentage points; brackets are query-bootstrap 95% intervals. Each row uses 40 questions. The three permutations were fixed before this control.

Dataset	Model	Permutation	Δ negative (pp)	Δ positive H (pp)
2Wiki	Gemma	A	0.42 [-1.81, 2.68]	1.72 [-1.33, 4.69]
2Wiki	Gemma	B	2.41 [-0.01, 4.77]	2.11 [-1.41, 5.70]
2Wiki	Gemma	C	-0.42 [-2.58, 1.72]	1.41 [-1.80, 4.92]
2Wiki	Mistral	A	-0.25 [-4.05, 3.29]	0.70 [-2.50, 3.98]
2Wiki	Mistral	B	-1.38 [-4.59, 1.73]	-1.33 [-5.31, 2.66]
2Wiki	Mistral	C	0.29 [-2.98, 3.61]	0.39 [-3.28, 4.38]
2Wiki	Qwen	A	-1.28 [-4.19, 1.56]	0.55 [-2.34, 3.52]
2Wiki	Qwen	B	-0.29 [-3.74, 2.85]	-1.41 [-4.53, 1.88]
2Wiki	Qwen	C	-2.94 [-5.66, -0.22]	1.80 [-0.62, 4.30]
HotpotQA	Gemma	A	1.57 [-1.14, 4.35]	0.94 [-2.73, 4.69]
HotpotQA	Gemma	B	1.90 [-0.57, 4.52]	0.00 [-3.59, 3.36]
HotpotQA	Gemma	C	1.68 [-0.73, 4.09]	0.55 [-1.95, 3.20]
HotpotQA	Mistral	A	0.77 [-2.86, 4.24]	1.48 [-2.34, 5.55]
HotpotQA	Mistral	B	2.41 [-1.03, 5.91]	1.09 [-2.66, 4.84]
HotpotQA	Mistral	C	1.40 [-2.47, 4.93]	1.02 [-2.27, 4.30]
HotpotQA	Qwen	A	0.47 [-2.23, 3.09]	-0.16 [-3.67, 3.44]
HotpotQA	Qwen	B	1.01 [-1.81, 3.68]	0.62 [-2.73, 4.22]
HotpotQA	Qwen	C	0.63 [-1.68, 2.91]	2.81 [-0.31, 5.86]
MuSiQue	Gemma	A	0.60 [-1.60, 2.74]	-1.80 [-5.00, 1.33]
MuSiQue	Gemma	B	2.78 [0.43, 5.34]	-0.70 [-3.59, 2.42]
MuSiQue	Gemma	C	1.91 [-0.56, 4.47]	-2.03 [-5.23, 1.33]
MuSiQue	Mistral	A	0.01 [-2.63, 2.64]	-0.62 [-3.67, 2.50]
MuSiQue	Mistral	B	0.78 [-2.29, 3.84]	-0.62 [-4.45, 3.21]
MuSiQue	Mistral	C	2.06 [-0.27, 4.31]	0.47 [-2.97, 3.91]
MuSiQue	Qwen	A	0.19 [-2.23, 2.57]	-1.56 [-4.61, 1.56]
MuSiQue	Qwen	B	-0.82 [-2.54, 0.79]	0.47 [-2.34, 3.12]
MuSiQue	Qwen	C	-0.59 [-2.70, 1.40]	-2.97 [-6.02, -0.08]

H.2 ALIAS SENSITIVITY AND OTHER CHECKS

The alias sensitivity replaces the canonical mean answer log probability by the maximum mean log probability over the official aliases, centers it by its own empty-evidence value, and retains the original calibration scale. Distinct alternative aliases are available for 38 questions in each of 2Wiki and MuSiQue in this control; the examined HotpotQA records do not supply distinct alternatives. Table 19 evaluates this two-dataset subset. Variable alias lengths and maximization define the alternative utility being compared.

Table 19: **Alternative official answers can change utility geometry and regret.** Alias-max minus canonical-answer results. Each n is the number of questions per model and track. Changes are paired; brackets are 95% intervals. Fixed-greedy regret is measured at $b = 3$ with the corresponding target's own optimum.

Dataset	Model	Track	n	Δ positive H (pp)	Δ regret / s
2Wiki	Gemma	A	38	-2.14 [-3.49, -0.89]	-0.090 [-0.186, -0.008]
2Wiki	Gemma	B	38	-2.35 [-3.82, -0.95]	-0.121 [-0.230, -0.027]
2Wiki	Mistral	A	38	-3.82 [-5.91, -1.96]	-0.115 [-0.203, -0.043]
2Wiki	Mistral	B	38	-3.99 [-6.11, -2.20]	-0.177 [-0.290, -0.081]
2Wiki	Qwen	A	38	-3.07 [-4.53, -1.78]	-0.129 [-0.324, 0.030]
2Wiki	Qwen	B	38	-2.85 [-4.48, -1.38]	-0.197 [-0.411, -0.024]
MuSiQue	Gemma	A	38	0.24 [-0.47, 0.92]	0.002 [-0.031, 0.034]
MuSiQue	Gemma	B	38	0.25 [-0.43, 0.95]	-0.002 [-0.093, 0.078]
MuSiQue	Mistral	A	38	-0.10 [-1.26, 1.10]	-0.039 [-0.078, -0.006]
MuSiQue	Mistral	B	38	0.04 [-1.04, 1.10]	-0.040 [-0.120, 0.028]
MuSiQue	Qwen	A	38	0.58 [-0.90, 2.04]	-0.001 [-0.033, 0.033]
MuSiQue	Qwen	B	38	0.79 [-0.57, 2.24]	-0.043 [-0.136, 0.022]

The measurement protocol verifies empty centering, complete finite canonical cubes, and fixed passage representations within each cube. Support-removal comparisons on calibration questions assess evidence sensitivity before final scoring. Prompts contain the question and selected passages; support annotations and canonical answers remain evaluation inputs.

I DIAGNOSTIC RECONSTRUCTION AND EVIDENCE AVAILABILITY

The lightweight diagnostic contains 36 complete cached likelihood cubes from nine questions, two models, and both six-passage tracks. Its independent CPU script reconstructs finite differences and exact selector comparisons from those fixed values. This subset provides a compact worked example of the reconstruction. The companion package supplies figure inputs and CPU scripts for checking cube completeness, rebuilding numerical tables, recomputing selected question-cluster intervals, and rescoring recorded answers. It also retains a mapping from manuscript claims to source tables.

The reconstruction scripts operate on recorded model values. End-to-end inference uses the original licensed benchmarks, fixed tokenizer and model revisions, candidate construction, and execution configuration. The supplement also provides the recorded local permutation and alias statistics with their comparison coverage, alongside the numerical reconstruction materials. Distributed benchmark-derived records retain their source licenses; model weights and additional benchmark assets are obtained separately from their original providers.

J CARDINALITY-RESOLVED GEOMETRY

The following post hoc analyses use the existing complete lattices. Table 20 separates quantities that determine effect size from numerical sign conventions and sampling uncertainty. The original evidence selectors are retained throughout these reporting views.

Table 20: **Calibration, numerical signs, and uncertainty serve different roles.**

Quantity	Definition	Role
Utility scale s	Calibration IQR of centered U, floored at 10^{-6}	One scale per dataset–model–track cell; reused at $m = 8$.
Material utility effect	$\epsilon = \max(10^{-3}, 0.05s)$	Count $\Delta < -\epsilon$ or $H > \epsilon$; .05 is a scale multiplier.
Unthresholded utility sign	Strict comparison with zero	Same-cohort sign description, separate from material rates.
Numerical F1 sign	$\Delta < -10^{-10}$ or $H > 10^{-10}$	Numerical tolerance, not an effect-size requirement.
F1 magnitude sensitivity	$\Delta < -0.10$ or $H > 0.10$	Previously specified absolute reward-difference threshold.
Sampling uncertainty	10,000 stratified question-cluster draws	Pointwise percentile 95% intervals; no simultaneous coverage across k.

J.1 Comparison units and denominators

Each cardinality layer has two oriented statistics. Negative-edge rows report the frequency of $\Delta_i(S) < 0$ (or $\Delta_i(S) < -\epsilon$) and the mean loss magnitude $\mathbb{E}[(-\Delta_i(S))_+]$. Positive-Hessian rows report the frequency of positive differences and the mean $\mathbb{E}[(H_{ij}(S))_+]$. Here $[x]_+ = \max(x, 0)$, and each mean averages over *all* tests of its type at $|S| = k$, including zero contributions. Thus the negative-edge amplitude measures lost utility, while the Hessian amplitude measures increasing returns. Likelihood amplitudes are divided by the calibration scale s; F1 amplitudes retain reward units. These are all-test means, not means conditional on a violation.

The U–F1 comparison uses exactly the same 90 questions, two models, and A/B constructions. Their units and sign tolerances remain different. Rates describe each target on the matched roster. The cross-size comparison uses the original 180 paired questions, two models, and Track A, with matching candidate identities, text, order, and calibration scales. The full 540-question U summary is reported separately. F1 is available only for the complete six-passage generation cubes.

New intervals resample questions within each dataset, retaining every model, track, size, and cardinality layer for a selected question together; datasets receive equal weight. The fixed thresholds and pointwise intervals describe each comparison separately. Utility summaries use the original question-cluster, dataset-stratified resampling. Repeated estimates share the canonical interval source documented in Appendix K; source-specific resampling records remain available in the supplement.

J.2 Generated-reward layers

Tables 21–22 retain all six edge layers and five Hessian layers. Positive-Hessian point rates are largest at $k = 0$ in both tracks, with a final-layer rebound in A. The magnitudes and intervals resolve the resulting nonuniform layer profile.

Table 21: Post hoc F1 cardinality description, Track A: 90 questions, two models, 180 complete cubes. Signs use 10^{-10}; 0.10 is the previously specified magnitude sensitivity. Mean $[H]_+$ includes zero contributions. Brackets are pointwise 95% question-cluster intervals; models and all cardinality layers stay paired.

k	Negative Δ (%)	Positive H (%)	Mean $[H]_+$	$H > 0.10$ (%)
0	9.26 [5.74, 13.06]	14.48 [10.19, 19.11]	0.107 [0.068, 0.152]	14.04 [9.81, 18.56]
1	7.57 [6.04, 9.15]	10.64 [8.83, 12.53]	0.073 [0.059, 0.089]	10.04 [8.33, 11.86]
2	7.15 [6.01, 8.36]	9.31 [7.87, 10.80]	0.068 [0.057, 0.080]	8.96 [7.59, 10.36]
3	6.93 [5.75, 8.18]	8.44 [7.07, 9.91]	0.062 [0.051, 0.074]	8.18 [6.85, 9.56]
4	7.11 [5.74, 8.50]	9.70 [7.26, 12.33]	0.081 [0.058, 0.107]	9.63 [7.19, 12.26]
5	6.76 [4.72, 8.98]	–	–	–

Table 22: Post hoc F1 cardinality description, Track B: 90 questions, two models, 180 complete cubes. Signs use 10^{-10}; 0.10 is the previously specified magnitude sensitivity. Mean $[H]_+$ includes zero contributions. Brackets are pointwise 95% question-cluster intervals; models and all cardinality layers stay paired.

k	Negative Δ (%)	Positive H (%)	Mean $[H]_+$	$H > 0.10$ (%)
0	9.07 [5.56, 12.87]	13.63 [9.26, 18.37]	0.102 [0.062, 0.149]	13.22 [8.89, 17.89]
1	7.02 [5.48, 8.67]	9.45 [7.65, 11.37]	0.062 [0.049, 0.076]	8.96 [7.24, 10.77]
2	6.25 [5.05, 7.49]	7.76 [6.27, 9.34]	0.053 [0.042, 0.064]	7.35 [5.98, 8.78]
3	6.64 [5.29, 8.03]	7.26 [5.72, 8.86]	0.052 [0.040, 0.065]	7.01 [5.50, 8.57]
4	6.44 [4.94, 8.09]	6.78 [4.85, 8.81]	0.057 [0.039, 0.077]	6.67 [4.74, 8.70]
5	6.48 [4.35, 8.80]	–	–	–

J.3 Utility and reward on the same questions

Tables 23–24 compare sign rates within the reward cohort. A positive-Hessian rate under U and the corresponding F1 rate use identical evidence memberships but different response functions. The accompanying numerical materials also retain cell-level estimates and the paired sign-rate differences; raw magnitudes remain in their respective units.

Table 23: Paired target descriptions on exactly the same 90 questions and two models, Track A. U uses its original strict-zero sign definition; F1 uses numerical tolerance 10^{-10}. Differences concern measured sign prevalence on a shared cohort, not equivalent material effect sizes or a causal target intervention. F1 and U amplitudes retain their own units. Intervals are pointwise paired question-cluster 95% intervals.

Test	k	U sign rate (%)	F1 sign rate (%)	F1 minus U (pp)
edge	0	64.81 [60.74, 68.80]	9.26 [5.74, 13.06]	-55.56 [-60.74, -50.19]
edge	1	43.24 [40.70, 45.76]	7.57 [6.04, 9.15]	-35.67 [-38.20, -33.09]
edge	2	41.64 [39.79, 43.50]	7.15 [6.01, 8.36]	-34.49 [-36.49, -32.57]
edge	3	41.75 [39.94, 43.58]	6.93 [5.75, 8.18]	-34.82 [-36.96, -32.69]
edge	4	43.33 [41.28, 45.39]	7.11 [5.74, 8.50]	-36.22 [-38.74, -33.70]
edge	5	46.39 [42.59, 50.19]	6.76 [4.72, 8.98]	-39.63 [-43.89, -35.46]
hessian	0	70.07 [65.37, 74.59]	14.48 [10.19, 19.11]	-55.59 [-61.93, -49.04]
hessian	1	49.11 [46.55, 51.57]	10.64 [8.83, 12.53]	-38.47 [-41.43, -35.38]
hessian	2	48.06 [46.04, 50.09]	9.31 [7.87, 10.80]	-38.74 [-41.12, -36.33]
hessian	3	47.29 [45.28, 49.34]	8.44 [7.07, 9.91]	-38.84 [-41.25, -36.39]
hessian	4	47.33 [43.85, 50.89]	9.70 [7.26, 12.33]	-37.63 [-41.89, -33.30]

Table 24: Paired target descriptions on exactly the same 90 questions and two models, Track B. U uses its original strict-zero sign definition; F1 uses numerical tolerance 10^{-10}. Differences concern measured sign prevalence on a shared cohort, not equivalent material effect sizes or a causal target intervention. F1 and U amplitudes retain their own units. Intervals are pointwise paired question-cluster 95% intervals.

Test	k	U sign rate (%)	F1 sign rate (%)	F1 minus U (pp)
edge	0	67.87 [63.89, 71.76]	9.07 [5.56, 12.87]	-58.80 [-63.80, -53.70]
edge	1	43.87 [41.50, 46.33]	7.02 [5.48, 8.67]	-36.85 [-39.52, -34.24]
edge	2	42.73 [40.94, 44.56]	6.25 [5.05, 7.49]	-36.48 [-38.51, -34.45]
edge	3	43.41 [41.71, 45.11]	6.64 [5.29, 8.03]	-36.77 [-38.91, -34.59]
edge	4	45.35 [43.46, 47.28]	6.44 [4.94, 8.09]	-38.91 [-41.26, -36.56]
edge	5	44.91 [40.93, 48.89]	6.48 [4.35, 8.80]	-38.43 [-42.59, -34.26]
hessian	0	72.52 [68.30, 76.63]	13.63 [9.26, 18.37]	-58.89 [-65.00, -52.59]
hessian	1	48.80 [46.46, 51.15]	9.45 [7.65, 11.37]	-39.34 [-42.09, -36.65]
hessian	2	47.98 [46.19, 49.79]	7.76 [6.27, 9.34]	-40.22 [-42.67, -37.78]
hessian	3	47.50 [45.59, 49.42]	7.26 [5.72, 8.86]	-40.24 [-42.44, -37.96]
hessian	4	49.63 [46.07, 53.22]	6.78 [4.85, 8.81]	-42.85 [-47.15, -38.74]

J.4 Within-cardinality changes between six and eight passages

Table 25 separates rate changes from magnitude changes on the paired extension. The five material positive-Hessian rate differences have positive point estimates; only the $k = 0$ rate interval excludes zero. For normalized positive magnitude, the estimates have mixed signs and only the $k = 3$ pointwise interval is wholly negative. This layer-specific pattern refines the weighting-sensitive overall comparison by locating the rate and magnitude differences within the fixed interface.

Table 25: Post hoc within-k differences, m=8 minus m=6, for the same 180 questions and two models in Track A. Calibration scales and candidate mappings are preserved. Edges use k=0 through 5; Hessians use k=0 through 4. Positive magnitudes average all tests within a layer. Intervals are pointwise paired question-cluster 95% intervals.

Test	k	Material-rate difference (pp)	$\mathbb{E}[(-\Delta)_+]/s$ or $\mathbb{E}[H_+]/s$ difference
Negative edge	0	+2.87 [+1.86, +3.85]	+0.0279 [+0.0117, +0.0443]
Negative edge	1	+0.79 [-0.10, +1.65]	+0.0065 [-0.0052, +0.0183]
Negative edge	2	+1.04 [+0.32, +1.80]	+0.0010 [-0.0053, +0.0070]
Negative edge	3	+1.74 [+0.98, +2.50]	-0.0015 [-0.0083, +0.0051]
Negative edge	4	+2.22 [+1.11, +3.36]	+0.0053 [-0.0026, +0.0129]
Negative edge	5	+0.92 [-1.24, +3.03]	+0.0078 [-0.0073, +0.0224]
Positive Hessian	0	+1.45 [+0.43, +2.48]	+0.0080 [-0.0134, +0.0297]
Positive Hessian	1	+0.39 [-0.60, +1.40]	+0.0012 [-0.0145, +0.0174]
Positive Hessian	2	+0.11 [-0.79, +1.00]	-0.0024 [-0.0100, +0.0053]
Positive Hessian	3	+0.74 [-0.56, +2.01]	-0.0100 [-0.0197, -0.0011]
Positive Hessian	4	+1.69 [-0.37, +3.77]	+0.0014 [-0.0140, +0.0165]

Table 26: Larger-only layers of the eight-passage Track-A lattice, 180 questions and two models. These layers are excluded from common-k standardization. Brackets are pointwise 95% question-cluster intervals.

Test	k	Material rate (%)	$\mathbb{E}[(-\Delta)_+]/s$ or $\mathbb{E}[H_+]/s$
Negative edge	6	31.25 [29.22, 33.38]	0.124 [0.104, 0.146]
Negative edge	7	29.51 [26.46, 32.67]	0.116 [0.094, 0.142]
Positive Hessian	5	33.56 [31.86, 35.22]	0.119 [0.104, 0.134]
Positive Hessian	6	34.17 [31.60, 36.82]	0.126 [0.107, 0.147]

J.5 Likelihood geometry by base cardinality

The following tables report likelihood geometry by base cardinality using the original pooled summaries. Six-passage summaries use the full 540-question roster and three models per track; the eight-passage summary uses 180 questions and two models in A. These rows describe their respective cohorts. Tables 25 and 17 provide the matched comparisons.

Table 27: Likelihood geometry by base cardinality, Track A, m=6: 540 questions and 3 models. Existing point estimates and pointwise 95% question-cluster intervals are reused. Edges report negative marginals; Hessians report positive differences. U sign threshold is zero; material threshold is max(0.001,0.05s). Positive parts are averaged over all tests, including zeros.

Test	k	Sign rate (%)	Material rate (%)	$\mathbb{E}[(-\Delta)_+]/s$ or $\mathbb{E}[H_+]/s$
Negative edge	0	64.18 [62.67, 65.69]	59.32 [57.74, 60.91]	0.581 [0.533, 0.634]
Negative edge	1	45.10 [44.26, 45.94]	36.76 [35.90, 37.63]	0.217 [0.201, 0.235]
Negative edge	2	42.55 [41.90, 43.22]	31.70 [30.94, 32.46]	0.148 [0.138, 0.159]
Negative edge	3	43.33 [42.64, 44.02]	30.12 [29.25, 30.99]	0.128 [0.119, 0.139]
Negative edge	4	45.61 [44.83, 46.40]	30.00 [29.01, 30.98]	0.118 [0.108, 0.128]
Negative edge	5	47.21 [45.92, 48.51]	29.08 [27.71, 30.46]	0.109 [0.097, 0.123]
Positive Hessian	0	69.83 [68.18, 71.48]	65.49 [63.70, 67.26]	0.676 [0.616, 0.740]
Positive Hessian	1	50.80 [50.00, 51.59]	42.56 [41.66, 43.44]	0.249 [0.229, 0.270]
Positive Hessian	2	47.96 [47.29, 48.64]	36.90 [36.08, 37.69]	0.163 [0.152, 0.176]
Positive Hessian	3	47.03 [46.33, 47.72]	33.57 [32.64, 34.48]	0.135 [0.125, 0.146]
Positive Hessian	4	48.72 [47.57, 49.86]	33.77 [32.51, 35.05]	0.125 [0.114, 0.137]

Table 28: Likelihood geometry by base cardinality, Track B, m=6: 540 questions and 3 models. Existing point estimates and pointwise 95% question-cluster intervals are reused. Edges report negative marginals; Hessians report positive differences. U sign threshold is zero; material threshold is max(0.001,0.05s). Positive parts are averaged over all tests, including zeros.

Test	k	Sign rate (%)	Material rate (%)	$\mathbb{E}[(-\Delta)_+]/s$ or $\mathbb{E}[H_+]/s$
Negative edge	0	66.73 [65.22, 68.29]	61.82 [60.23, 63.49]	0.647 [0.595, 0.705]
Negative edge	1	45.64 [44.79, 46.46]	37.91 [37.02, 38.78]	0.234 [0.216, 0.252]
Negative edge	2	43.05 [42.38, 43.72]	33.13 [32.34, 33.91]	0.160 [0.148, 0.171]
Negative edge	3	43.87 [43.21, 44.56]	31.60 [30.75, 32.46]	0.139 [0.129, 0.149]
Negative edge	4	46.20 [45.39, 47.03]	31.61 [30.60, 32.64]	0.128 [0.118, 0.140]
Negative edge	5	47.28 [45.94, 48.63]	31.47 [30.05, 32.90]	0.120 [0.108, 0.134]
Positive Hessian	0	71.05 [69.46, 72.65]	67.28 [65.56, 69.01]	0.736 [0.674, 0.803]
Positive Hessian	1	50.97 [50.20, 51.74]	43.36 [42.49, 44.22]	0.264 [0.244, 0.285]
Positive Hessian	2	48.16 [47.53, 48.79]	38.01 [37.23, 38.78]	0.173 [0.161, 0.185]
Positive Hessian	3	47.17 [46.47, 47.86]	35.06 [34.14, 35.95]	0.144 [0.133, 0.155]
Positive Hessian	4	48.47 [47.30, 49.65]	34.52 [33.24, 35.84]	0.129 [0.118, 0.141]

Table 29: Likelihood geometry by base cardinality, Track A, m=8: 180 questions and 2 models. Existing point estimates and pointwise 95% question-cluster intervals are reused. Edges report negative marginals; Hessians report positive differences. U sign threshold is zero; material threshold is max(0.001,0.05s). Positive parts are averaged over all tests, including zeros.

Test	k	Sign rate (%)	Material rate (%)	$\mathbb{E}[(-\Delta)_+]/s$ or $\mathbb{E}[H_+]/s$
Negative edge	0	67.33 [64.20, 70.31]	63.19 [59.97, 66.32]	0.668 [0.570, 0.776]
Negative edge	1	44.69 [43.07, 46.33]	37.73 [36.00, 39.46]	0.241 [0.207, 0.278]
Negative edge	2	41.50 [40.43, 42.57]	32.18 [30.93, 33.42]	0.159 [0.139, 0.179]
Negative edge	3	43.37 [42.40, 44.34]	31.85 [30.46, 33.22]	0.139 [0.122, 0.157]
Negative edge	4	45.10 [44.17, 46.05]	31.76 [30.36, 33.18]	0.128 [0.112, 0.146]
Negative edge	5	45.98 [44.67, 47.28]	30.92 [29.25, 32.60]	0.121 [0.104, 0.138]
Negative edge	6	47.09 [45.40, 48.79]	31.25 [29.22, 33.38]	0.124 [0.104, 0.146]
Negative edge	7	47.15 [44.38, 50.03]	29.51 [26.46, 32.67]	0.116 [0.094, 0.142]
Positive Hessian	0	72.03 [68.68, 75.23]	68.28 [64.81, 71.67]	0.759 [0.645, 0.886]
Positive Hessian	1	51.05 [49.40, 52.69]	44.05 [42.29, 45.78]	0.278 [0.239, 0.319]
Positive Hessian	2	47.18 [46.24, 48.12]	37.40 [36.20, 38.59]	0.173 [0.152, 0.196]
Positive Hessian	3	47.29 [46.43, 48.16]	35.43 [34.09, 36.79]	0.146 [0.128, 0.165]
Positive Hessian	4	48.82 [47.76, 49.89]	35.37 [33.92, 36.83]	0.135 [0.117, 0.152]
Positive Hessian	5	48.43 [47.12, 49.68]	33.56 [31.86, 35.22]	0.119 [0.104, 0.134]
Positive Hessian	6	50.14 [47.67, 52.63]	34.17 [31.60, 36.82]	0.126 [0.107, 0.147]

K CONSISTENT REPORTING OF REPEATED INTERVALS

Repeated estimates share a fixed question-bootstrap summary. Eight-passage fixed and stopped regrets use the decomposition summary. Exact-U F1 regret uses the negative of the paired exact-U-minus-F1-oracle contrast, with interval endpoints negated and reversed; Figure 2 and the corresponding tables use that same interval. The larger-only layers in Table 26 use the original cardinality summary reported in Table 29. The supplement records the source keys and fixed resampling seeds and provides deterministic scripts that reconstruct every repeated display from its designated summary.

Table 30: Canonical intervals reused for repeated statistics. All entries are for budget three and pooled question-cluster summaries.

Statistic	Cohort	Estimate [95% interval]
Fixed U regret / s	A, eight passages	0.171 [0.125, 0.221]
Stopped U regret / s	A, eight passages	0.075 [0.048, 0.109]
Exact-U F1 regret	A, six passages	0.184 [0.135, 0.235]
Exact-U F1 regret	B, six passages	0.175 [0.125, 0.227]