THE VALUE OF EVIDENCE SELECTION DEPENDS ON REASONING EFFORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence selection reduces input length but also changes the information available for reasoning. Its accuracy and cost benefits may therefore depend on the reader's reasoning effort. We test this dependence in multi-hop retrieval-augmented question answering with the model, prompt, retrieved passages, and selected contexts fixed. On held-out 2Wiki questions, a learned selector improves exact-match accuracy by 4.4 percentage points at minimal effort but reduces it by 3.0 points at medium effort. Despite the same 67.3% prompt-token reduction, reader-cost savings fall from 60.0% to 8.7%. To examine the role of evidence content, we restore omitted support in an annotation-guided intervention across three benchmarks. Relative to adding a similar-length non-support passage, restoration improves medium-effort accuracy by 9.3 points and reduces mean reported reasoning tokens by 42.2%. We extend this investigation to information computed within a source-only retrieval pipeline. Grounded execution reuses intermediate answers already extracted to guide search; exposing rather than withholding these answers, with acquired passages fixed, improves minimal-effort accuracy by 5.3 points on a three-benchmark development panel without another model call. Reuse also lowers medium-effort total API cost, including planning and extraction, while its accuracy effect is unresolved. These effort-dependent benefits motivate joint configuration selection. We evaluate joint calibration of existing evidence policies and reader effort on held-out questions.

1 INTRODUCTION

Retrieval-augmented question answering requires a reader to combine information from external passages. Selection and compression can reduce input cost and distraction (Xu et al., 2024; Zhang et al., 2025b; Gu et al., 2026), but also change the available information. Passage value depends on the reader (Zhang et al., 2025a); reasoning models introduce an additional choice within that reader, its inference-effort setting. Since useful test-time computation depends on the task and inference strategy (Snell et al., 2025; Yue et al., 2025), we ask whether evidence selected for one effort remains beneficial when that setting changes.

This dependence affects both accuracy and efficiency. On held-out 2Wiki questions, a released selector improves exact-match accuracy by 4.4 percentage points at minimal effort but reduces it by 3.0 points at medium effort. With input reduction unchanged, reader-cost savings fall from 60.0% to 8.7%, and selected contexts induce more reported reasoning tokens at medium effort. Three-benchmark comparisons and an abstractive compressor characterize the scope of this interaction.

To investigate how information contributes to this tradeoff, we intervene on omitted support. Restoring an annotated supporting passage, compared with adding a similar-length non-support passage, improves accuracy and reduces medium-effort generation. This tests a particular information loss and connects the supplied evidence to computational outcomes beyond input length.

Restoration uses annotations unavailable at inference. We therefore examine information already computed within a source-only retrieval pipeline. In grounded execution, question decomposition and source-quoted answers guide retrieval for dependent subquestions. Reusing these answers in the final reader adds no model call. With acquired passages fixed, reuse improves minimal-effort accuracy by 5.3 points on a 360-question, three-benchmark development panel. At medium effort it reduces reported reasoning tokens and total API cost, including planning and extraction, while the accuracy change is unresolved.

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.

These effort-dependent effects make evidence and effort a joint decision. We therefore evaluate calibration over the existing policy–effort menu.

The paper makes three contributions:

- A controlled demonstration that reasoning effort can reverse fixed evidence-policy rankings within one reader and alter their realized reader-cost savings.
- Controlled support-restoration and source-only answer-reuse interventions that connect supplied information to accuracy and generation. Reusing answers already computed for retrieval improves minimal-effort accuracy without another call and reduces medium-effort API cost on a three-benchmark development panel.
- An empirical evaluation of joint accuracy–cost calibration over the existing policy–effort menu, compared with selecting the evidence policy at a single effort.

2 Related Work

Evidence selection and compression. Evidence policies reduce retrieval-augmented contexts through selection or synthesis. RECOMP trains extractive and abstractive compressors against downstream language-model performance and includes an abstention mechanism for unhelpful retrievals (Xu et al., 2024). Utility-based selection distills pseudo-answers and passage-utility judgments into a smaller selector (Zhang et al., 2025b); related results show that passage utility can depend on the downstream reader (Zhang et al., 2025a). BRIEF-Pro instead synthesizes retrieved evidence into a controllable-length summary while measuring both reader performance and compressor overhead (Gu et al., 2026). We examine a complementary deployment variable: how the value of a fixed selector or compressor output changes with the same reader's reasoning-effort setting.

Evidence passages can also have conditional rather than independent value. ParaSet and SetCE learn compatibility scores for complete evidence sets (Song & Lee, 2026). RepoShapley combines a coalition surrogate with exact Shapley values (Huo et al., 2026), while document-attribution work characterizes redundancy, complementarity, and synergy in retrieval-augmented generation (Nematov et al., 2025). Set-level deletion risk has also been modeled for agent histories (Wang et al., 2026). Our paired support-restoration intervention follows this interaction-aware perspective: it compares an omitted support passage with a same-count, similar-length addition from the same acquired pool.

Reasoning effort and joint efficiency. Reader behavior depends on both context structure and inference allocation. Long-context models are sensitive to the position of relevant evidence (Liu et al., 2024). Adaptive-RAG routes questions among no-retrieval, single-step, and iterative strategies (Jeong et al., 2024), while Self-Ask uses explicit follow-up questions and search (Press et al., 2023). IRCoT interleaves retrieval with reasoning so that intermediate conclusions can guide subsequent evidence acquisition (Trivedi et al., 2023). Our grounded-acquisition experiment builds on this use of intermediate answers. It separately varies whether answers already used for retrieval are exposed to the final reader, holding the acquired passages fixed and measuring the effect at different reasoning efforts. Test-time scaling methods allocate inference compute according to problem difficulty (Snell et al., 2025). Yue et al. (2025) allocate retrieved documents, demonstrations, and iterative prompting steps for long-context RAG; their short-answer accounting omits output tokens. ThinkDial trains discrete low-, medium-, and high-effort modes (He et al., 2025). Recent systems connect reasoning and compression more directly. ReCo coordinates key–value-cache compression, reflection penalties, and stopping, and finds that input compression can induce additional generation (Zhu et al., 2026). GRC jointly trains generation, representation, and compression objectives (Miao et al., 2026).

Together, this literature motivates treating evidence policy and reader effort as coupled deployment controls, with resource accounting that includes generated output. We study this coupling by crossing identical questions and fixed policy outputs with three documented effort settings of one dated reader, while including both input and generated tokens in the resource analysis. We then select joint evidence–effort actions on calibration data, freeze those choices, and evaluate them on held-out questions under several declared price schedules.

3 PROBLEM FORMULATION

An evidence policy can change answer accuracy and the amount of generated output at a given effort. We therefore model evidence policy and reasoning effort jointly, so that their interaction enters both answer quality and computational cost.

3.1 EVIDENCE POLICIES, ACCURACY, AND COST

For a question q, let $D(q)$ denote the retrieved passage pool. An evidence policy $s \in \mathcal{S}$ produces a context $s(q, D(q))$ by selecting or rewriting passages. The reader generates an answer at effort setting $e \in \mathcal{E}$. Let $Y(q, s, e; \xi)$ be its evaluation score, where ξ represents generation randomness. Expected accuracy is

$$A(s, e) = \mathbb{E}_q \mathbb{E}_\xi \big[Y(q, s, e; \xi) \big]. \tag{1}$$

The fixed-pool policy comparisons use the same questions and retrieved passages. Our primary experiments use $\mathcal{E} = \{\text{minimal}, \text{low}, \text{medium}\}$, the reader's explicit API effort controls. These settings regulate inference behavior; they do not impose equal token or floating-point-operation budgets. Generated length is measured as an outcome. Section 4.3 extends the evaluation to additional acquisition and intermediate-answer inputs.

Let C_{acq} and $C_{\text{sel}}(s)$ denote acquisition and selection cost. Let T_s and $G_{s,e}$ be the numbers of billed input and output tokens, with respective unit prices p_{in} and p_{out}. Expected deployment cost is

$$C(s, e) = C_{\text{acq}} + C_{\text{sel}}(s) + p_{\text{in}} \mathbb{E}[T_s] + p_{\text{out}} \mathbb{E}[G_{s,e}]. \tag{2}$$

The output term $G_{s,e}$ depends on both policy and effort: selecting evidence can change generation as well as input length. Latency is measured separately from dollar cost.

Because $G_{s,e}$ responds to the supplied evidence, a shorter context need not lower the bill. Let ΔT and ΔG denote the changes in mean billed input and output tokens when a policy replaces a reference context at the same effort. The reader-cost change is $p_{\text{in}} \Delta T + p_{\text{out}} \Delta G$, so an input reduction ($\Delta T < 0$) lowers reader cost exactly when

$$\Delta G < -\frac{p_{\text{in}}}{p_{\text{out}}} \Delta T, \tag{3}$$

and lowers total cost only if the saving also covers selection, $\Delta G < -(p_{\text{in}} \Delta T + C_{\text{sel}}(s))/p_{\text{out}}$. For the reader studied here, $p_{\text{in}}/p_{\text{out}} = 1/8$ at list prices, so each additional output token offsets the saving from eight removed input tokens. With input caching, the input term separates into cached and uncached changes at their respective prices.

For a cost weight $\lambda \geq 0$, net utility is $U_\lambda(s, e) = A(s, e) - \lambda C(s, e)$. Larger λ places greater weight on savings. A budget-constrained alternative maximizes accuracy subject to mean cost at most B. Either objective can select a single configuration or a question-independent mixture of configurations.

3.2 POLICY–EFFORT INTERACTIONS

Let $M(s, e)$ denote either accuracy or net utility under a fixed cost definition. For two policies s_0, s_1 and two efforts e_L, e_H, define the relative policy effect as $\delta_e = M(s_1, e) - M(s_0, e)$. Their interaction is

$$I_M = \delta_{e_L} - \delta_{e_H}. \tag{4}$$

A nonzero I_M means that a policy's relative value depends on effort. When δ_{e_L} and δ_{e_H} have opposite signs, the ranking reverses: each effort favors a different policy, so no single policy choice is best at both. Joint calibration (Section 3.3) therefore estimates the value of each policy–effort pair directly.

3.3 JOINT CALIBRATION

Because a policy's value can depend on effort, we select the evidence policy and the reader effort together for a given deployment objective. An action $a = (s, e)$ pairs an evidence policy s with an effort e; the menu contains the four extractive policies of Section 4 at all three efforts, giving 12 actions. For a labeled calibration sample Q, let $\overline{\text{EM}}_Q(a)$ be mean exact match on the zero-to-one scale

and $\overline{C}_{\mathrm{reader},Q}(a)$ mean billed reader cost in US dollars per question. Calibration then maximizes a calibration-set estimate of the net utility U_λ defined in Section 3.1:

$$\widehat{U}_\lambda(a;Q) = \overline{\mathrm{EM}}_Q(a) - \lambda\,\overline{C}_{\mathrm{reader},Q}(a) - \gamma\mathbf{1}\{s = \textsc{Utility}\}, \qquad \gamma = 0.01. \tag{5}$$

Equation 5 approximates U_λ: acquisition cost is common to all actions and cancels, and the selection cost is replaced by γ, a fixed one-point penalty for invoking the learned selector that does not scale with λ; selector compute is measured separately (Appendix F). The weight λ has units of utility per US dollar, and net-utility points are 100 times utility differences.

For each $\lambda \in \{0, 100, 250, 500, 1000\}$, the procedure is:

1. Compute unrounded calibration means for all 12 actions.
2. *Joint calibration* selects the action maximizing Equation 5.
3. *Medium-first* selects the policy with the highest medium-effort EM, breaking ties by lower reader cost, and then the effort that maximizes Equation 5 within that policy.
4. *Minimal-first* is defined analogously, using minimal-effort EM in its first stage.
5. Ties within 10^{-12} are broken by lower reader cost, then a fixed action order

All choices are fitted on the calibration sample and applied unchanged to held-out questions. Under a mean-cost budget instead of a cost weight, calibration also considers randomized mixtures of actions; an optimal mixture uses at most two actions (Appendix A).

4 Methodology

Fixed-context comparisons estimate policy–effort interactions; restoration and answer-reuse interventions test specified changes to the reader's information.

4.1 Evidence Policies

We consider four policies over five retrieved passages. Full5 retains all five; Top2 and Top3 retain the first two or three in retrieval order; Utility applies the released UtilityQwen1.7B learned selector (Zhang et al., 2025b). The rank-based subsets control for reducing context without learned selection. A separate experiment evaluates the released abstractive compressor Brief-Pro (Gu et al., 2026).

Every policy receives the same passage pool, and its output is reused across effort settings. Reader version, prompt, answer extraction, scoring, and output-token limit are fixed.

The primary contrast is the minimal-minus-medium accuracy interaction for Utility versus Full5 in Equation 4. Evidence for a ranking reversal additionally requires opposite signs for the two within-effort contrasts, with both paired confidence intervals excluding zero. Comparisons at low effort and against the other policies characterize the interaction in greater detail.

4.2 Support-Restoration Interventions

We examine an information loss caused by selection: omission of a passage annotated by the benchmark as supporting the answer. We construct interventions on questions for which the retrieved pool contains all annotated support but the Utility context, denoted U, omits exactly one supporting passage g. Eligibility depends on the evidence annotations, independently of answer correctness. Among the unselected passages not annotated as support, let d be the closest in length to g. The two intervention contexts are

$$R = U \cup \{g\}, \qquad C = U \cup \{d\}. \tag{6}$$

Here, R restores support and C is the non-support control. Each adds one complete passage from the same pool. Both preserve the original retrieval order and text representation. Passage counts are matched exactly and token lengths approximately; insertion positions can differ.

For effort e, the paired effect is $\rho_e = \mathbb{E}[Y(R,e) - Y(C,e)]$. We test whether this effect changes between medium and minimal effort and repeat the analysis on cases whose request lengths differ

by at most 32 tokens. A follow-up intervention appends either passage to the same selected prefix, matching the insertion position as well. Since these constructions use gold annotations, they measure the effect of specified evidence changes rather than the performance of a deployable selector.

4.3 Grounded Acquisition and Answer Reuse

A retrieval pipeline can also supply information it computes without annotations. Grounded execution uses question decomposition and source-quoted intermediate answers to guide retrieval, then tests whether those answers should be passed to the final reader. A question-only plan specifies subquestions and dependencies. A *root* subquestion has no dependencies; its answer is extracted from the initial five passages. For a dependent subquestion whose parents are resolved roots, a BM25 query combines that subquestion with the parent answers. Each query adds at most one novel passage, with at most three additions per question. Original passages remain unchanged. Exact source quotations verify provenance, not correctness. No reference answers or support annotations enter the procedure.

We fix the acquired passages and plan, then compare final-reader inputs that withhold or expose the extracted root answers. A third condition exposes root and dependent answers, using one batched call to resolve the dependent subquestions from the augmented evidence. Invalid or unresolved answers remain in the evaluation. The root-answer comparison varies only whether the reader sees answers used for retrieval.

With or without root-answer reuse, the complete pipeline requires three model calls: planning, extraction, and final reading. Reuse adds input but no new call; dependent resolution requires a fourth call. All helper calls use fixed low effort; only final-reader effort varies. API costs include all calls, with local retrieval reported separately. The ablation uses 360 development questions across three benchmarks. Appendix H.4 details the controls and retrieval-transfer evaluation.

5 Experiments

We evaluate whether policy value changes with effort, whether controlled information changes affect the reader's response, and whether joint calibration improves configuration selection.

5.1 Experimental Setup

Data and retrieval. We use MuSiQue, 2WikiMultiHopQA (2Wiki), and HotpotQA (Trivedi et al., 2022; Ho et al., 2020; Yang et al., 2018). The main evaluation contains 500 held-out 2Wiki development questions, excluding question identities, normalized questions, and connected groups of annotated support that appeared in earlier experiments. Its retrieval corpus contains 6,350 deduplicated documents drawn from these questions and 500 additional questions. BM25 retrieves 50 candidates, and `bge-reranker-v2-m3` reranks them to five complete passages (Robertson & Zaragoza, 2009; Beijing Academy of Artificial Intelligence, 2024). Every evidence policy operates on this same retrieved pool. Additional experiments and calibration use separate three-benchmark samples. Their composition and overlap are specified in Appendix B.

Reader and metrics. The primary reader is `gpt-5-mini-2025-08-07`, accessed through AIMLAPI with explicit minimal, low, or medium reasoning effort. All conditions share a two-message question-answering prompt and a 4,000-token completion limit. We report normalized exact match (EM), which tests whether the answer matches a reference after normalization and alias handling. Accuracy differences are in percentage points. Token-overlap F1 and complete request settings appear in the appendix. Resource measurements distinguish input tokens, reported reasoning tokens, visible answer tokens, and billed reader cost.

Statistical analysis. All contrasts are paired by question. Confidence intervals use 4,000 bootstrap resamples of connected question groups defined by shared support or source membership, retaining each question's conditions together. The held-out sample has 490 such groups. Its primary interaction and reversal criteria were specified before generation. Secondary intervals are unadjusted. The resampling estimates question-sampling uncertainty conditional on the observed generations; exact duplicate requests share a response, as detailed in Appendix C.

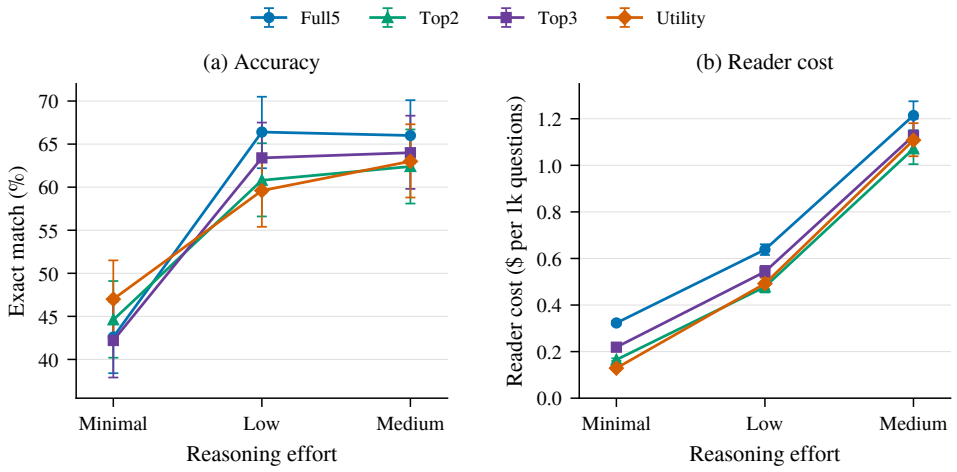

Figure 1: Accuracy and billed reader cost on 500 held-out 2Wiki questions. Each policy supplies identical contexts at every effort setting. Error bars show 95% bootstrap intervals; cost excludes retrieval and selection.

Table 1: Held-out 2Wiki accuracy. EM is in percent; paired differences are in percentage points with 95% confidence intervals. The minimal − medium interaction is the pre-specified primary contrast; the other rows are secondary.

Effort	Full5	Top2	Top3	Utility	Utility − Full5
Minimal	42.6	44.6	42.2	47.0	+4.4 [1.2, 7.7]
Low	66.4	60.8	63.4	59.6	-6.8 [-9.6, -4.1]
Medium	66.0	62.4	64.0	63.0	-3.0 [-5.6, -0.4]
Min. − med. interaction	–	–	–	–	+7.4 [3.4, 11.4]

5.2 Effect of Reasoning Effort

The same learned selector helps at minimal effort and harms at higher effort. On the held-out sample, Utility improves EM over Full5 by 4.4 points at minimal effort, but reduces it by 6.8 points at low effort and 3.0 points at medium effort (Table 1). The minimal-to-medium interaction is 7.4 points, and both endpoint contrasts satisfy the ranking-reversal criterion. Because the selected contexts are identical across efforts, the result shows that a policy's accuracy benefit does not transfer automatically between settings of the same reader. The same interaction appears under a different retriever: on a separate 1,000-question 2Wiki development sample whose passages come from HippoRAG2 (Appendix C.3), the Utility–Full5 minimal-minus-medium interaction is +9.0 points [5.7, 12.3], although the minimal-effort gain alone is not resolved (+2.4 [−0.3, 5.2]).

The interaction is nonmonotonic: the accuracy deficit is larger at low than at medium effort (Figure 1). Against Top3, Utility also changes rank between minimal and low effort (+4.8 [1.6, 8.1] versus −3.8 [−6.8, −0.8]); against Top2, no effort-specific difference is resolved (Table 9). Learned selection is therefore not consistently preferable to retrieval-rank subsets.

On the three-benchmark sample, the pooled selector advantage decreases between minimal and low effort, primarily on MuSiQue and 2Wiki; the HotpotQA contrast is unresolved. Brief-Pro also has a positive minimal-minus-medium interaction, with an uncertain minimal-effort gain and a resolved medium-effort loss. These results extend effort-dependent compression beyond extraction without establishing a reversal for every operator or dataset (Appendix C).

5.3 Computational Cost

The accuracy interaction is accompanied by a change in efficiency (Table 2). Utility removes the same 67.3% of prompt tokens at every effort, yet its reader-cost saving falls from 60.0% at minimal

Table 2: Resource response of UTILITY relative to FULL5 on the held-out 2Wiki sample. Input reduction is identical at every effort, but additional reasoning erodes the reader-cost saving. Extra reasoning tokens are per question, with 95% intervals (Appendix F).

Effort	Input-token reduction	Extra reasoning tokens	Reader-cost saving
Minimal	67.3%	0.0	60.0%
Low	67.3%	+18.2 [9.0, 27.6]	22.8%
Medium	67.3%	+33.4 [10.7, 57.8]	8.7%

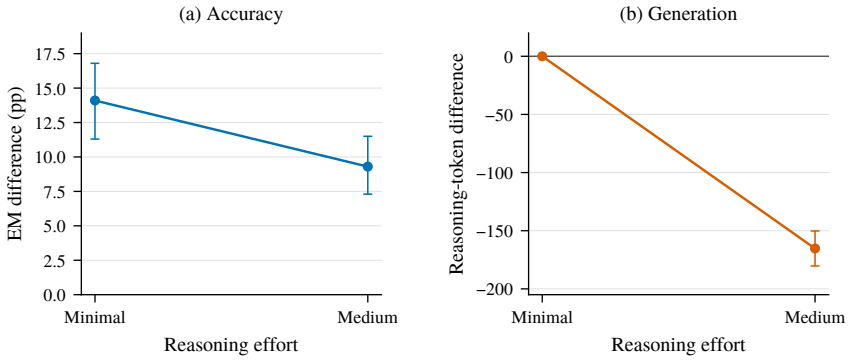

Figure 2: Support restoration relative to the non-support control ($R - C$, Equation 6) on 1,147 eligible questions; both conditions add one passage. (a) Exact-match difference in percentage points. (b) Difference in reported reasoning tokens per question. At minimal effort, both conditions report zero reasoning tokens; at medium effort, restoration reduces mean reported reasoning tokens by 165 (42.2%). Error bars show paired 95% confidence intervals.

to 8.7% at medium effort, because selected contexts induce additional reasoning at low and medium effort. Since one output token costs as much as eight input tokens (Equation 3), input reduction overstates the reader-cost saving.

The response differs by compressor: BRIEF-PRO retains positive reader-cost savings at both efforts, with an unresolved change in medium-effort reasoning tokens. The fixed input reduction therefore needs to be evaluated together with each reader's generation response.

Selection overhead can further offset reader savings. Warm single-query profiling gives a UTILITY break-even price of approximately $0.48 per occupied GPU-hour at minimal effort. Appendix F reports runtime, hardware, and billed versus standardized token costs.

5.4 EFFECTS OF SUPPORT RESTORATION

The factorial compares complete policy outputs. To examine a specified evidence difference, we compare restoring an omitted supporting passage with a non-support addition, matching passage count and approximate length. Across 1,147 eligible questions, restoration improves EM by 14.1 points at minimal effort and 9.3 points at medium effort (Figure 2). The restoration advantage is smaller at medium effort, with the same conclusion when request lengths differ by at most 32 tokens. This effort dependence is resolved on 2Wiki and HotpotQA; the MuSiQue interaction remains uncertain.

At medium effort, restoration reduces mean reported reasoning tokens by 42.2%, with no resolved change in visible answer length. The intervention therefore improves accuracy while reducing generation relative to the control, despite adding the same number of passages. At minimal effort, the reader reports no reasoning tokens and the restoration condition costs slightly more. The computational effect of the evidence change depends on the effort setting.

A position-matched follow-up on 135 eligible held-out questions again finds a smaller restoration advantage at medium effort; its primary test of effort-dependent non-support harm is inconclusive.

Table 3: Intermediate-answer reuse on 360 development questions. Acquired-context rows share passages, plan, and root extraction. Only final-reader effort varies. EM is in percent; API costs include every model call. Paired intervals and shared local retrieval costs are reported in Appendix H.4.

Configuration	Total API	Minimal final effort		Medium final effort	
	calls	EM	Billed $/1k	EM	Billed $/1k
Direct (original five)	1	34.4	0.280	62.2	1.294
Acquired; root answers withheld	3	44.4	2.291	64.7	3.211
Root answers exposed	3	49.7	2.303	65.6	3.152
Root + dependent answers exposed	4	52.2	3.531	64.2	4.335

Table 4: Configuration selected at each cost weight λ under three-dataset calibration, with its held-out 2Wiki accuracy and billed reader cost, shown as (EM %, $ per 1,000 questions).

λ	Joint calibration	Medium-first	Minimal-first
0	Medium FULL5 (66.0, 1.21)	Medium FULL5 (66.0, 1.21)	Medium UTILITY (63.0, 1.11)
100	Low FULL5 (66.4, 0.64)	Low FULL5 (66.4, 0.64)	Low UTILITY (59.6, 0.49)
250	Low UTILITY (59.6, 0.49)	Low FULL5 (66.4, 0.64)	Low UTILITY (59.6, 0.49)
500	Low UTILITY (59.6, 0.49)	Low FULL5 (66.4, 0.64)	Low UTILITY (59.6, 0.49)
1000	Minimal UTILITY (47.0, 0.13)	Minimal FULL5 (42.6, 0.32)	Minimal UTILITY (47.0, 0.13)

Restoration is also positive under three reader configurations, including Qwen2.5-3B and Phi-3.5-mini (Appendix D). These effects concern eligible omission cases. Attributing the complete policy reversal would require separating multiple deletions and changes in context composition.

5.5 INTERMEDIATE-ANSWER REUSE

The annotation-guided intervention motivates examining information available without support labels. Grounded execution tests whether the final reader benefits from answers already computed for retrieval, with acquired passages fixed. Exposing rather than withholding root answers improves minimal-effort EM by 5.3 points $[1.69, 14.36]$ in the secondary comparison (Table 3). The observed API-bill increase is 0.5%, with an interval including zero. The primary medium-effort accuracy difference is 0.8 points $[-1.87, 2.41]$. At medium effort, reuse reduces reported reasoning by 29.7 tokens per question and API cost by 1.8%. These paired resource reductions are resolved with planning and extraction included; the accuracy interval does not establish equivalence. The same reuse decision therefore has different benefits at the two final-reader efforts.

The distinction between reuse and additional computation matters for deployment. Dependent resolution adds cost without an established accuracy gain; all staged paths cost more than direct medium-effort reading. Acquisition improves on an equal-count BGE rank-continuation control in development, but does not establish a gain over the original five passages in a separate 450-question Qwen study with a different starting retriever. Appendix H reports further analyses.

5.6 JOINT CALIBRATION

Effort-dependent accuracy and cost make the deployment objective essential to configuration selection. We evaluate joint calibration on the original twelve pairs of fixed evidence policies and reader efforts; answer reuse is not part of this menu. We estimate their performance on 750 calibration questions, with 250 per benchmark, and evaluate the selected configurations on the held-out 2Wiki sample used above. A second calibration excludes 2Wiki and uses only MuSiQue and HotpotQA. Neither procedure uses held-out outcomes for selection.

The comparison depends on the cost regime (Figure 3; Table 4). With little or no cost penalty, joint calibration chooses FULL5 at medium or low effort and improves over minimal-first calibration. At the largest tested cost weight, it chooses minimal-effort UTILITY and improves over medium-first calibration. These outcomes show that selecting a policy at one effort can restrict the configurations available when the deployment objective changes.

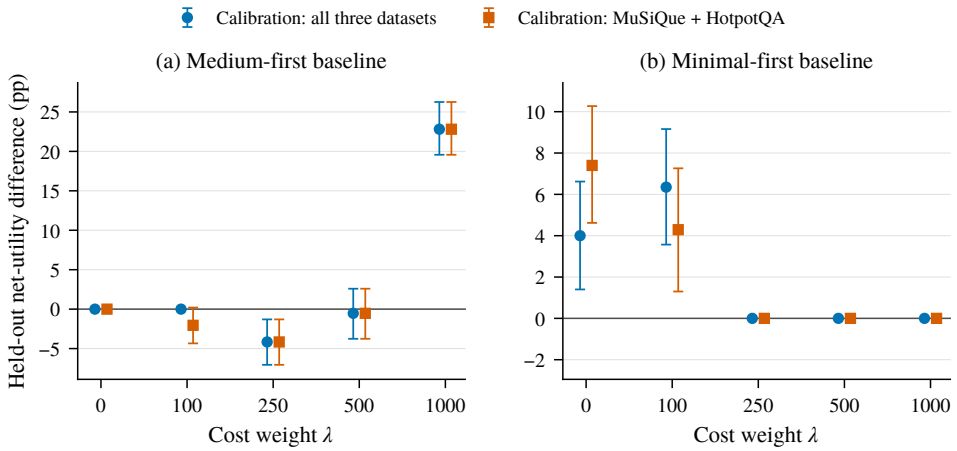

Figure 3: Held-out net-utility difference (Equation 5) between joint calibration and medium-first (a) or minimal-first (b) selection, in percentage points; positive values favor joint calibration. Circles use the three-dataset calibration set, squares use MuSiQue and HotpotQA only; both are evaluated on the same held-out 2Wiki sample. Error bars show paired 95% confidence intervals.

Joint calibration also makes an intermediate-price error. At $\lambda = 250$, it chooses low-effort UTILITY, which loses 4.2 net-utility points to the low-effort FULL5 configuration selected by medium-first calibration. The calibration without 2Wiki exhibits the same high-price gain and intermediate-price loss. Joint selection expands the available choices, but still requires calibration data representative of deployment. Appendix E reports fixed-action comparisons and budget-constrained mixtures.

6 DISCUSSION

Accuracy and cost results make inference configuration part of evidence-policy evaluation. A selected context should be assessed at its deployment effort, for both answer quality and generated output.

Restoration and answer reuse connect this dependence to supplied information. Outcomes change under matched passage counts or fixed acquired passages, respectively. These controls support conclusions about the reader's response to specified inputs, without identifying its internal mechanism.

The findings suggest two deployment decisions: whether to reuse already-computed answers at a given final-reader effort, and which policy–effort pair meets the intended accuracy–cost objective. The first weighs added input against its benefit. The second needs representative calibration data.

Limitations. The primary reversal uses one dated reader and held-out 2Wiki questions. Grounded execution uses development data; the Qwen acquisition study changes the starting retriever and does not replicate answer reuse. Support annotations and exact match can miss alternative evidence or valid answers, motivating the reported controls and F1. Broader readers, corpora, and provider/hardware settings require evaluation.

7 CONCLUSION

Evidence selection changes both available information and the reader's computational response. The resulting effort dependence appears in accuracy, cost savings, and intermediate-answer reuse. Evidence and effort should be evaluated jointly, using answer quality and measured reader cost while accounting separately for acquisition and processing overhead.

Reproducibility Statement

The appendix specifies the datasets, retrieved evidence, model checkpoints, request settings, prompts, scoring, and statistical analysis. It also describes sample overlap and response reuse. The accompanying implementation produces the reported tables and figures from the experimental results. The held-out evaluation criteria and calibration choices were specified before its responses were collected. Distribution of datasets, model weights, and provider responses is subject to their respective licenses and terms.

Ethics Statement

The experiments use public question-answering benchmarks and Wikipedia-derived passages. They measure answer accuracy and resource use in research settings and do not establish suitability for safety-critical applications. Benchmark support annotations are used for offline analysis, training controls, and explicitly identified oracle interventions; they are not inference inputs to the source-only policies.

AI-Use Statement

AI assistants were used to polish writing and draft parts of the manuscript text, figures, and tables; to search and discover related literature; to support research execution, including developing experimental protocols, implementing and reviewing code, coordinating experiments, and analyzing results. The reader models and evidence policies evaluated in the study are described in the experimental setup. The authors reviewed all AI-assisted work and remain responsible for the accuracy, attribution, and reproducibility of the work.

REFERENCES

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Beijing Academy of Artificial Intelligence. BAAI/bge-reranker-v2-m3 model card. Hugging Face model card, 2024. URL https://huggingface.co/BAAI/bge-reranker-v2-m3. Accessed 2026-09-24.

Jia-Chen Gu, Junyi Zhang, Di Wu, Yuankai Li, Kai-Wei Chang, and Nanyun Peng. BRIEF-pro: Universal context compression with short-to-long synthesis for fast and accurate multi-hop reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14221–14241. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.findings-acl.696. URL https://aclanthology.org/2026.findings-acl.696/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Qianyu He, Siyu Yuan, Xuefeng Li, Mingxuan Wang, and Jiangjie Chen. ThinkDial: An open recipe for controlling reasoning effort in large language models. *arXiv preprint arXiv:2508.18773*, 2025. URL https://arxiv.org/abs/2508.18773.

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 the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/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. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10390–10412. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.findings-acl.505. URL https://aclanthology.org/2026.findings-acl.505/.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7036–7050. Association for Computational Linguistics, June 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

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. URL https://aclanthology.org/2024.tacl-1.9/.

Zhongtao Miao, Qiyu Wu, and Yoshimasa Tsuruoka. GRC: Unifying reasoning-driven generation, retrieval and compression. *arXiv preprint arXiv:2605.09100*, 2026. URL https://arxiv.org/abs/2605.09100.

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.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Mooho Song and Jay-Yoon Lee. Retrieving a set, not independent passages: Set-level compatibility learning for efficient set exploration. *arXiv preprint arXiv:2607.05712*, 2026. URL https://arxiv.org/abs/2607.05712.

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. URL https://aclanthology.org/2022.tacl-1.31/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Mingxuan Wang, Bo Wang, Fei Luo, Guorun Yao, Chao Ning, Yinglong Guo, Hongyue Chen, Yanbiao Ma, and Jungong Han. DRSR: Learning set-level deletion risk for efficient long-horizon agents. *arXiv preprint arXiv:2609.27276*, 2026. URL https://arxiv.org/abs/2609.27276.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b5747175878345040 8f5c9b70257d15b-Abstract-Conference.html.

Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. LLM-specific utility for retrieval-augmented generation. *arXiv preprint arXiv:2510.11358*, 2025a. URL https://arxiv.org/abs/2510.11358.

Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. Distilling a small utility-based passage selector to enhance retrieval-augmented generation. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 22–30. Association for Computing Machinery, 2025b. doi: 10.1145/3767695.3769496. URL https://doi.org/10.1145/3767695.3769496.

Qiyuan Zhu, Dezhi Li, Pengyu Cheng, Tianle Chen, Jiacheng Wang, Ruijie Shen, Hao Gu, Sida Lin, Zirui Liu, Jiacheng Liu, and Sirui Han. Fewer tokens, smaller cache: Reward-coordinated efficient reasoning. *arXiv preprint arXiv:2608.04771*, 2026. URL https://arxiv.org/abs/2608.04771.

Appendix guide. Appendix A contains the proofs, and Appendix B provides experimental details. Additional results appear in Appendices C and D. Appendices E and F describe calibration and computational cost. Appendices G and H evaluate additional methods, evidence repair, and grounded execution.

A Mathematical Details

This section gives the formal result for budget-constrained mixtures used in Section 3.3.

A.1 Mixtures under one mean-cost constraint

Consider K actions. Action k has population mean accuracy A_k and mean cost C_k, and w_k is the probability assigned to it. For a feasible mean-cost budget B, consider

$$\max_{w \in \mathbb{R}^K} \sum_{k=1}^{K} w_k A_k \quad \text{subject to} \quad \sum_{k=1}^{K} w_k C_k \le B, \quad \sum_{k=1}^{K} w_k = 1, \quad w_k \ge 0. \tag{7}$$

Proposition 1 (Two-action sufficiency)**.** *If Equation 7 is feasible, it has an optimal solution supported on at most two actions.*

Proof. The feasible set is a nonempty compact polytope, so a linear objective has an optimal extreme point. Suppose an extreme point has at least three positive coordinates. On those coordinates, there is a nonzero vector v satisfying $\sum_k v_k = 0$ and $\sum_k C_k v_k = 0$, because these are at most two independent linear constraints on at least three unknowns. For sufficiently small $\epsilon > 0$, both $w + \epsilon v$ and $w - \epsilon v$ remain nonnegative and normalized, and they have the same cost as w. Both are feasible and their midpoint is w, contradicting extremality. Thus an optimal extreme point has at most two positive coordinates. □

The optimizer can be found by enumerating feasible single actions and action pairs whose mean costs bracket B. If $C_i < B < C_j$, the cost-saturating weight on action i is $(C_j - B)/(C_j - C_i)$. Comparing the corresponding expected accuracies identifies the best candidate; for equal-cost pairs, the higher-accuracy endpoint suffices. This construction yields a question-independent mixture. In empirical calibration, estimates replace the population means, and held-out evaluation measures the resulting accuracy and budget transfer.

B Experimental Details

This section specifies the reader, evidence, sampling, retrieval, and inference procedures. The treatment is an evidence-policy and reasoning-effort pair. Within each natural-context comparison, the question, acquired evidence pool, policy output, reader snapshot, prompt, output limit, and scorer remain fixed across effort settings.

B.1 Reader and evaluation

B.1.1 Reader configuration

The factorial experiments requested `openai/gpt-5-mini-2025-08-07` through AIMLAPI, and the returned model identifier was `gpt-5-mini-2025-08-07`. Each request specified the model, the two messages below, `reasoning_effort`, and `max_completion_tokens=4000`. The effort value was `minimal`, `low`, or `medium`. No temperature parameter was supplied.

The system message was:

```
Answer the question using ONLY the passages provided.  Reply
with the shortest exact answer span -- a name, date, number
or noun phrase (or yes/no if the question is a yes/no
question).  No sentence, no explanation, no punctuation at
the end.  If the passages do not contain the answer, reply
with your best guess anyway.
```

For a context containing m passages, passage i was rendered as the literal prefix `[i]`, one space, the title, a newline, and the complete passage text. Consecutive passages were separated by two newlines. The user message was:

```
Passages:
[1] ⟨title⟩
⟨text⟩
...
Question: ⟨question⟩
Answer:
```

Gold answers and aliases were stored separately and were unavailable when reader messages were constructed.

B.1.2 Response processing and resource fields

The prediction is the returned message content after stripping outer whitespace and trailing periods. An empty completion at the length limit is scored as incorrect; a nonempty length-limited completion is scored as returned. Byte-identical requests within a study share one collected response, with the evaluation convention described in Section B.5.

Let T_{in} denote provider-reported prompt tokens, T_{out} total completion tokens, and T_{reason} the reported field `completion_tokens_details.reasoning_tokens`. Visible answer tokens are $T_{\text{vis}} = T_{\text{out}} - T_{\text{reason}}$. Reported reasoning tokens are generation telemetry rather than a controlled computational dose. The primary reader-cost measure is the provider-reported value `meta.usage.usd_spent`. The standardized uncached-price sensitivity is

$$C_{\text{list}} = 0.25\,T_{\text{in}}/10^6 + 2.00\,T_{\text{out}}/10^6 \tag{8}$$

in US dollars. Retrieval and local evidence selection are excluded from both reader-cost measures and analyzed separately in Section F.

B.1.3 Answer metrics

Exact match (EM) and token F1 use the SQuAD normalization: lowercasing, removal of ASCII punctuation and the articles *a*, *an*, and *the*, followed by whitespace normalization. EM equals one when the normalized prediction matches any normalized official answer or alias. Token F1 is the maximum bag-of-token overlap score over aliases. Percentage results multiply question-level scores by 100.

B.2 Evidence policies

For a five-document acquisition pool, FULL5 retains all five documents in acquisition order. TOP2 and TOP3 retain the first two or three documents in that order. The UTILITY policy applies the released UtilityQwen1.7B selector of Zhang et al. (2025b); it may return zero through five documents and may reorder them. The reader receives the complete, untruncated source passages in the selector's output order. We use F and U to denote the FULL5 and UTILITY contexts in subsequent contrasts.

B.2.1 Utility selector implementation

We used `hengranZhang/UtilityQwen1.7B`, revision `986a2ccbef147749fb29c3728a25ea1e4a1101df`, with the released inference code at commit `28f1f993c0369d03d3b572d153fadf32df93ea5d`. Each candidate view is the whitespace-collapsed string `Title: <title> Content: <body>`, truncated to 300 whitespace-delimited words. The five candidates are numbered 1 through 5. Decoding is greedy with temperature zero, thinking disabled, and a 512-token output limit.

The released conversation begins with:

> **System:** `You are the utility selector, an intelligent assistant that can select the passages that have utility in answering the question.`

Assistant: `Yes, i am the utility selector.`
User: `I will provide you with 5 passages, each indicated by number identifier [].`
`Select the passages that have utility in answering the question: ⟨question⟩.`
Assistant: `Okay, please provide the passages.`

For each $i = 1, \ldots, 5$, a user turn contains `[i] <selector view>` and the next assistant turn is `Received passage [i].` The final user turn is:

```
Question: ⟨question⟩
Utility Requirement: A passage has **utility** only if it
is **both relevant to the question AND useful for generating
a correct and reasonable answer.**
Firstly, **provide the answer** to the question based on the
provided 5 passages or your own knowledge without sources,
which can help judge passages utility.
Then, **select the passages that have utility in answering
the question**.
**Output Format STRICTLY:**
Answer: [Your generated answer here]
My selection: [[i],[j],...]
Only respond to the answer and the selected list, do not say
any word or explain.
```

The parser searches for the first case-sensitive occurrence of `selection:`, extracts bracketed decimal identifiers from the remaining text, converts one-based identifiers to zero-based indices, removes repetitions after their first occurrence, and preserves model order. Indices outside 1 through 5 are discarded. A missing marker or empty list yields an empty context. In the held-out 500-question sample, the selector returned 42 empty, 278 one-document, 160 two-document, 17 three-document, 2 four-document, and 1 five-document contexts.

B.3 Datasets, sampling, and retrieval

The experiments use 2WikiMultiHopQA, HotpotQA, and MuSiQue (Ho et al., 2020; Yang et al., 2018; Trivedi et al., 2022). Development samples support method comparison and calibration; the separately sampled 2Wiki population provides held-out evaluation.

B.3.1 2Wiki comparison sample

The comparison factorial uses 1,000 questions from a 2Wiki sample used during method development and its fixed `raw_hippo` top-five packets. These packets were produced by the HippoRAG2 acquisition path (Gutiérrez et al., 2025). The comparison crosses FULL5 and UTILITY with minimal and medium effort. The sample contains 886 canonical-document dependency components.

B.3.2 Three-dataset sample

For each benchmark, a fixed 500-question roster was sorted by the SHA-256 digest of a prespecified salt, benchmark name, and question identifier, with the identifier used to break ties. The first 250 questions were retained, yielding 750 questions in total; reader outcomes did not enter the sampling rule. Before subsampling, retrieval was run independently for each benchmark over the text-deduplicated contexts of its 500 evaluation and 500 filler questions. The resulting corpora contain 6,319 2Wiki documents, 9,786 HotpotQA documents, and 8,913 MuSiQue documents. BM25 retained 50 candidates and the BGE reranker described below retained five complete documents. All four evidence policies were evaluated at minimal, low, and medium effort on the same questions and policy outputs. Exact-support-page components were formed on each complete 500-question roster before restriction to 250 questions, and benchmark identifiers prefix components in pooled analyses.

B.3.3 Held-out 2Wiki sample

The sampling frame is the 12,576-question official 2Wiki development split. We excluded question identifiers used in preceding development samples and required normalized question strings to be

unique in the full split. Normalization lowercases, replaces characters outside `[a-z0-9 ]` with spaces, and collapses whitespace.

We formed a graph over the full split by connecting questions that share an exact annotated-support page, represented by its title and concatenated official sentences. Every component containing a previously used question was excluded. The remaining candidates were ordered by the SHA-256 digest of a prespecified salt, dataset name, and question identifier, with the identifier as the tie break. The first 500 questions form the evaluation sample and the next 500 provide retrieval-corpus fillers. Answers, aliases, model outputs, question type, retrieval success, and cost were absent from the selection rule. This construction separates the held-out sample from development data in question identifiers and annotated-support components; non-support pages, entities, and pretraining exposure can still overlap.

B.3.4 HELD-OUT RETRIEVAL

Text deduplication over the official context pages of the 500 evaluation and 500 filler questions yields 6,350 documents. BM25 (Robertson & Zaragoza, 2009) tokenizes by Unicode casefold followed by the regular expression `[^\W_]+`. Each indexed document is its canonical title, one ASCII space, and exact text. Let $V(q)$ be the set of distinct tokens in query q, qtf_t the query frequency of token t, N the corpus size, df_t the document frequency of t, $tf_{t,d}$ its frequency in document d, $|d|$ the document length, and $\overline{|d|}$ the mean document length. The BM25 score is

$$\sum_{t \in V(q)} qtf_t \log\left(1 + \frac{N - df_t + 0.5}{df_t + 0.5}\right) \frac{tf_{t,d}(k_1 + 1)}{tf_{t,d} + k_1(1 - b + b|d|/\overline{|d|})}, \tag{9}$$

with $k_1 = 1.5$ and $b = 0.75$. Documents are ranked by decreasing score, with corpus row as the tie break, and the top 50 proceed to reranking.

The reranker is `BAAI/bge-reranker-v2-m3`, revision `953dc6f6f85a1b2dbfca4c34a2796e7dde08d41e` (Beijing Academy of Artificial Intelligence, 2024). Sequence A is the original question; sequence B is the exact title, a newline, and exact text. Scoring uses half-precision floating point (FP16), a maximum sequence length of 512, and pair microbatches of 16. Decreasing logit determines rank, with BM25 rank as the tie break. The reader receives the five complete documents, including when a shorter view was used for reranker scoring.

B.3.5 ABSTRACTIVE COMPRESSION SAMPLE

The compression study uses 450 development questions, 150 from each benchmark. The raw condition is a fixed, interleaved union of evidence from an interpretation-search plan and the HippoRAG2 pipeline, and the compressed condition is a cached BRIEF-Pro summary of that same parent context (Gu et al., 2026). We used `uclanlp/brief-pro`, revision `287329d91a06cb8158a94254e15030bea4f0ce19`, and the authors' Auto prompt at code commit `077943328a555f49925ac1d459e00ed229bc63b9`. The authors' vLLM 0.6.4.post1 inference contract was implemented with Transformers 5.17.0, PyTorch 2.11.0, scaled dot-product attention, and FP16, including the tied-embedding mapping. Generation used temperature 0.01, top-p 0.95, seed 20260923, a 1,024-token limit, and the tokenizer's end-of-sequence and end-of-turn stop tokens. The generated summaries were used without a post-generation length cutoff. Raw and compressed evidence were each evaluated at minimal and medium reader effort.

B.4 GOLD-ANNOTATION EVIDENCE INTERVENTIONS

These interventions use official support annotations offline to construct explanatory input contrasts. They do not define a deployable evidence policy. In addition to U and F, let R denote the context obtained by adding omitted annotated support and C the context obtained by adding a passage-count-matched annotated non-support control.

B.4.1 SUPPORT-RESTORATION SAMPLE

A question is eligible when its development five-document pool contains all official support, U omits exactly one support passage, and an unselected non-support passage is available. The sample contains

419 2Wiki, 455 HotpotQA, and 273 MuSiQue questions, for 1,147 questions total. Context R adds the omitted support passage. Context C adds the unselected non-support passage whose standalone `o200k_base` token length is nearest to that support passage; a fixed salted order resolves ties. Both contexts follow the original acquisition order. Since the two additions can have different acquisition ranks, their insertion positions need not coincide. Both contexts are evaluated at minimal and medium effort. A prespecified length sensitivity retains the 700 questions for which the two exact reader messages differ by at most 32 tokens.

B.4.2 Overlapping 2Wiki intervention subset

For the 419 2Wiki questions in the support-restoration sample, the 2Wiki comparison factorial supplies the U and F responses, while the intervention study supplies R and C. All four contexts share the acquired pool, prompt, reader snapshot, and effort settings. The subset therefore supports paired explanatory contrasts among the four contexts while contributing no additional questions beyond its parent samples.

B.4.3 Held-out 2Wiki intervention subset

Eligibility was determined within the held-out 500-question sample before reader outcomes were analyzed. The top-five pool must contain all official support, U must be an exact subset that omits one support passage, and at least one omitted non-support passage must be available. The resulting 135 questions belong to 134 inherited support components. Contexts R and C append the omitted support or nearest-length non-support passage to the unchanged U prefix at the same final position. Both are evaluated at all three effort settings; the U and F responses come from the parent factorial. Because R uses append order while F uses acquisition order, the $R - F$ contrast includes both evidence selection and presentation order.

B.5 Statistical inference and response reuse

All confidence intervals use 4,000 paired, two-sided percentile bootstrap replicates with seed 20260923. Each sample uses a precomputed dependency map based on exact annotated support or shared canonical source documents. A bootstrap replicate samples whole components with replacement, retains every question in a sampled component, and applies the same replicate to all arms of a contrast. The held-out 500-question sample has 490 support components; the 1,000-question comparison sample has 886 canonical-document components; the support-restoration sample has 1,046 components; and its 419-question 2Wiki subset has 404. The held-out intervention subset inherits 134 components from its parent. Components for the three-dataset sample are formed on each complete 500-question source roster before subsampling. Dataset-pooled analyses prefix component identifiers by benchmark and weight questions equally.

For effort settings e_1 and e_2, the natural-context interaction is

$$[\mathrm{EM}(U, e_1) - \mathrm{EM}(F, e_1)] - [\mathrm{EM}(U, e_2) - \mathrm{EM}(F, e_2)]. \tag{10}$$

Secondary cell, adjacent-effort, resource, and intervention intervals are descriptive and unadjusted unless identified as primary. A fix is a question on which the first policy is correct and the reference is incorrect; a harm is the converse.

Each unique request contributes one collected reader generation. A byte-identical request alias reuses that response and remains a logical observation in its destination cell, inheriting the measured prediction, usage, and cost. Component-bootstrap intervals condition on these realized generations. They quantify question and component sampling uncertainty, excluding variation that would arise from repeating reader generation or provider serving.

C Additional Results

C.1 Study populations and analytic reuse

Table 5 summarizes the experimental populations and their roles. Several analyses intentionally share observations. The 419-question 2Wiki intervention subset belongs to both the 2Wiki comparison

factorial and the 1,147-question support-restoration sample. The 135-question held-out intervention subset belongs to the held-out 2Wiki factorial and reuses its U and F responses. The three-effort analysis and calibration study use the same 750 questions, and all calibration rules are evaluated on the same held-out 2Wiki sample. These relationships support paired comparisons but do not constitute additional independent samples.

Table 5: Summary of evaluation samples. The support-restoration and overlapping intervention analyses share questions; calibration and held-out evaluation samples are reused across rows, so sample sizes are not additive.

Study	Population	n	Evidence actions	Effort	Role
Policy–effort factorial	2Wiki comparison sample	1000	Full5, Utility	Minimal, medium	Policy–effort interaction
BRIEF-Pro factorial	2Wiki, HotpotQA, MuSiQue	450	Raw, BRIEF-Pro	Minimal, medium	Compression comparison
Four-policy factorial	Held-out 2Wiki; support-component-disjoint	500	Full5, Top2, Top3, Utility	Minimal, low, medium	Held-out policy comparison
Oracle restoration	Three-dataset one-omission sample	1147	Support, matched non-support	Minimal, medium	Support intervention
Intervention subset	Held-out 2Wiki omission subset	135	4 fixed contexts	Minimal, low, medium	Held-out intervention comparison
Complete-menu calibration	Three-dataset calibration → held-out 2Wiki	750/500	12 policy–effort actions	Held-out evaluation	Price-dependent deployment choice

C.2 HELD-OUT 2WIKI FACTORIAL

The held-out factorial establishes a strict ranking reversal between UTILITY and FULL5. The paired EM difference, $U - F$, is $+4.40$ points at minimal effort, -6.80 at low effort, and -3.00 at medium effort. The primary minimal-minus-medium interaction is $+7.40$ points with a 95% interval of $[3.39, 11.42]$. The corresponding interaction is $+5.80$ $[2.01, 9.44]$ against TOP3 and $+1.80$ $[-1.20, 4.82]$ against TOP2.

Table 6: Held-out 2Wiki sample ($n = 500$), with four policies and three effort levels. EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Full5	500	42.6 [38.4, 46.8]	48.8 [44.7, 52.9]
Minimal	Top2	500	44.6 [40.2, 49.1]	50.2 [46.0, 54.3]
Minimal	Top3	500	42.2 [37.9, 46.7]	47.9 [43.8, 52.1]
Minimal	Utility	500	47.0 [42.5, 51.5]	53.2 [49.1, 57.4]
Low	Full5	500	66.4 [62.2, 70.5]	74.9 [71.4, 78.3]
Low	Top2	500	60.8 [56.6, 65.1]	69.2 [65.5, 73.0]
Low	Top3	500	63.4 [59.3, 67.5]	71.8 [68.1, 75.2]
Low	Utility	500	59.6 [55.4, 63.8]	68.7 [64.9, 72.4]
Medium	Full5	500	66.0 [61.9, 70.1]	74.8 [71.3, 78.2]
Medium	Top2	500	62.4 [58.1, 66.7]	70.2 [66.5, 73.8]
Medium	Top3	500	64.0 [59.8, 68.3]	72.1 [68.4, 75.6]
Medium	Utility	500	63.0 [58.8, 67.3]	71.0 [67.4, 74.7]

Relative to FULL5, UTILITY changes incorrect answers to correct ones and correct answers to incorrect ones on 45 and 23 questions at minimal effort, 8 and 42 at low effort, and 14 and 29 at medium effort. The aggregate reversal therefore reflects bidirectional question-level changes, and its magnitude depends on the compact-policy reference.

Table 7: Held-out 2Wiki sample ($n = 500$), with four policies and three effort levels. Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Full5	500	893.5 [865.0, 921.1]	0.0 [0.0, 0.0]
Minimal	Top2	500	403.1 [385.7, 421.3]	0.0 [0.0, 0.0]
Minimal	Top3	500	571.6 [549.9, 593.5]	0.0 [0.0, 0.0]
Minimal	Utility	500	292.4 [275.6, 308.7]	0.0 [0.0, 0.0]
Low	Full5	500	893.5 [865.0, 921.1]	121.6 [114.1, 129.9]
Low	Top2	500	403.1 [385.7, 421.3]	120.6 [112.0, 129.0]
Low	Top3	500	571.6 [549.9, 593.5]	124.5 [115.6, 133.6]
Low	Utility	500	292.4 [275.6, 308.7]	139.8 [130.2, 150.1]
Medium	Full5	500	893.5 [865.0, 921.1]	343.4 [320.1, 366.6]
Medium	Top2	500	403.1 [385.7, 421.3]	348.7 [323.3, 375.5]
Medium	Top3	500	571.6 [549.9, 593.5]	349.1 [323.4, 376.7]
Medium	Utility	500	292.4 [275.6, 308.7]	376.8 [350.2, 405.3]

Table 8: Held-out 2Wiki sample ($n = 500$), with four policies and three effort levels. Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Visible tok.	Reader $/1k
Minimal	Full5	500	13.24 [12.96, 13.54]	0.323 [0.314, 0.332]
Minimal	Top2	500	13.00 [12.77, 13.23]	0.165 [0.160, 0.171]
Minimal	Top3	500	13.10 [12.86, 13.34]	0.219 [0.212, 0.226]
Minimal	Utility	500	12.97 [12.73, 13.20]	0.129 [0.124, 0.135]
Low	Full5	500	12.70 [12.50, 12.91]	0.638 [0.615, 0.661]
Low	Top2	500	12.57 [12.38, 12.77]	0.478 [0.454, 0.501]
Low	Top3	500	12.74 [12.53, 12.95]	0.543 [0.518, 0.568]
Low	Utility	500	12.74 [12.54, 12.96]	0.492 [0.468, 0.519]
Medium	Full5	500	12.73 [12.53, 12.93]	1.214 [1.151, 1.275]
Medium	Top2	500	12.61 [12.42, 12.81]	1.071 [1.005, 1.141]
Medium	Top3	500	12.72 [12.52, 12.93]	1.127 [1.058, 1.200]
Medium	Utility	500	12.60 [12.42, 12.79]	1.108 [1.039, 1.181]

Table 9: Paired Utility contrasts on the held-out 2Wiki sample ($n = 500$), in exact-match percentage points. Interaction rows compare the indicated effort-specific Utility effects (Utility minus the fixed reference); intervals use the component bootstrap.

Reference	Estimand	Estimate [95% CI]	n	Components
Full5	Minimal	+4.4 [1.2, 7.7]	500	490
Full5	Low	-6.8 [-9.6, -4.1]	500	490
Full5	Medium	-3.0 [-5.6, -0.4]	500	490
Full5	Interaction: min. − med.	+7.4 [3.4, 11.4]	500	490
Full5	Interaction: min. − low	+11.2 [7.4, 15.1]	500	490
Full5	Interaction: low − med.	-3.8 [-6.6, -1.2]	500	490
Top2	Minimal	+2.4 [-0.4, 5.2]	500	490
Top2	Low	-1.2 [-4.0, 1.6]	500	490
Top2	Medium	+0.6 [-2.2, 3.2]	500	490
Top2	Interaction: min. − med.	+1.8 [-1.2, 4.8]	500	490
Top2	Interaction: min. − low	+3.6 [0.4, 6.7]	500	490
Top2	Interaction: low − med.	-1.8 [-4.4, 1.0]	500	490
Top3	Minimal	+4.8 [1.6, 8.1]	500	490
Top3	Low	-3.8 [-6.8, -0.8]	500	490
Top3	Medium	-1.0 [-3.6, 1.6]	500	490
Top3	Interaction: min. − med.	+5.8 [2.0, 9.4]	500	490
Top3	Interaction: min. − low	+8.6 [5.0, 12.3]	500	490
Top3	Interaction: low − med.	-2.8 [-5.6, 0.0]	500	490

C.3 2WIKI COMPARISON FACTORIAL

The 1,000-question comparison sample shows the same policy-by-effort interaction without resolving a positive minimal-effort endpoint. The sample was used during method development and is reported separately from held-out evaluation. The $U - F$ effect is $+2.40$ EM points $[-0.30, 5.18]$ at minimal effort and -6.60 $[-8.68, -4.59]$ at medium effort, giving an interaction of $+9.00$ $[5.74, 12.33]$. UTILITY reduces prompt tokens by 61.3%; its reader-bill reduction is 54.5% at minimal effort and an unresolved 2.1% at medium effort. This sample corroborates the interaction while reserving the strict-reversal conclusion for the held-out evaluation.

Table 10: 2Wiki comparison sample ($n = 1{,}000$), with two policies and two effort levels. EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Full5	1000	47.7 [44.5, 50.8]	56.3 [53.4, 59.1]
Minimal	Utility	1000	50.1 [46.8, 53.4]	57.8 [54.7, 60.7]
Medium	Full5	1000	72.8 [69.8, 75.6]	81.2 [78.9, 83.4]
Medium	Utility	1000	66.2 [63.2, 69.2]	75.2 [72.8, 77.6]

Table 11: 2Wiki comparison sample ($n = 1{,}000$), with two policies and two effort levels. Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Full5	1000	882.9 [858.0, 908.8]	0.0 [0.0, 0.0]
Minimal	Utility	1000	341.3 [328.8, 353.6]	0.0 [0.0, 0.0]
Medium	Full5	1000	882.9 [858.0, 908.8]	304.5 [291.5, 317.8]
Medium	Utility	1000	341.3 [328.8, 353.6]	363.5 [347.0, 381.2]

Table 12: 2Wiki comparison sample ($n = 1{,}000$), with two policies and two effort levels. Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Completion tok.	Reader $/1k
Minimal	Full5	1000	13.38 [13.21, 13.57]	0.320 [0.312, 0.328]
Minimal	Utility	1000	13.23 [13.06, 13.40]	0.146 [0.142, 0.150]
Medium	Full5	1000	317.63 [304.56, 330.89]	1.113 [1.078, 1.149]
Medium	Utility	1000	376.54 [360.16, 394.26]	1.091 [1.048, 1.136]

Table 13: Paired EM contrasts on the 2Wiki comparison sample ($n = 1{,}000$), in percentage points.

Contrast	Estimate [95% CI]	n	Components
utility minus full medium	-6.6 [-8.7, -4.6]	1000	886
utility minus full minimal	+2.4 [-0.3, 5.2]	1000	886
minimal minus medium full5	-25.1 [-28.1, -22.1]	1000	886
minimal minus medium utility	-16.1 [-18.8, -13.3]	1000	886
effort by policy interaction	+9.0 [5.7, 12.3]	1000	886

Table 14: Paired F1 contrasts on the 2Wiki comparison sample ($n = 1{,}000$), in percentage points.

Contrast	Estimate [95% CI]	n	Components
utility minus full medium	-5.97 [-7.78, -4.16]	1000	886
utility minus full minimal	+1.50 [-1.09, 4.21]	1000	886
minimal minus medium full5	-24.90 [-27.73, -22.08]	1000	886
minimal minus medium utility	-17.43 [-20.16, -14.71]	1000	886
effort by policy interaction	+7.47 [4.47, 10.59]	1000	886

C.4 Effort profile across three datasets

The three-dataset sample locates most of the change in the transition from minimal to low effort. Pooled $U - F$ is $+3.73$ EM points $[0.57, 6.90]$ at minimal effort, -1.73 $[-4.19, 0.77]$ at low effort, and -1.87 $[-4.46, 0.75]$ at medium effort. The minimal-minus-low change is $+5.47$ $[1.72, 9.24]$, the low-minus-medium change is $+0.13$ $[-2.37, 2.68]$, and their second difference is $+5.33$ $[0.43, 10.12]$.

Table 15: Evaluation sample: All datasets ($n = 750$). EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Full5	750	35.2 [30.7, 39.7]	47.5 [43.6, 51.3]
Minimal	Top2	750	33.5 [29.4, 37.8]	45.0 [41.0, 48.9]
Minimal	Top3	750	34.8 [30.5, 39.3]	47.2 [43.3, 51.2]
Minimal	Utility	750	38.9 [34.5, 43.4]	50.2 [46.2, 54.3]
Low	Full5	750	58.9 [53.9, 63.7]	71.6 [67.5, 75.5]
Low	Top2	750	56.9 [51.8, 61.8]	68.3 [64.2, 72.3]
Low	Top3	750	57.9 [52.8, 62.7]	69.9 [65.7, 73.8]
Low	Utility	750	57.2 [52.2, 62.0]	68.2 [64.2, 72.2]
Medium	Full5	750	59.2 [54.1, 64.0]	71.3 [67.3, 75.3]
Medium	Top2	750	56.9 [51.9, 61.7]	68.3 [64.2, 72.3]
Medium	Top3	750	58.3 [53.0, 63.2]	70.2 [66.1, 74.2]
Medium	Utility	750	57.3 [52.3, 62.2]	68.1 [63.9, 72.2]

Table 16: Evaluation sample: All datasets ($n = 750$). Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Full5	750	770.4 [741.0, 798.8]	0.0 [0.0, 0.0]
Minimal	Top2	750	329.6 [318.4, 341.7]	0.0 [0.0, 0.0]
Minimal	Top3	750	478.0 [460.2, 495.5]	0.0 [0.0, 0.0]
Minimal	Utility	750	252.8 [241.9, 265.0]	0.0 [0.0, 0.0]
Low	Full5	750	770.4 [741.0, 798.8]	149.6 [136.0, 164.5]
Low	Top2	750	329.6 [318.4, 341.7]	138.0 [127.3, 149.3]
Low	Top3	750	478.0 [460.2, 495.5]	147.4 [135.8, 159.6]
Low	Utility	750	252.8 [241.9, 265.0]	132.9 [124.1, 142.0]
Medium	Full5	750	770.4 [741.0, 798.8]	410.8 [373.9, 454.9]
Medium	Top2	750	329.6 [318.4, 341.7]	374.8 [344.2, 407.0]
Medium	Top3	750	478.0 [460.2, 495.5]	392.0 [363.1, 425.3]
Medium	Utility	750	252.8 [241.9, 265.0]	354.5 [330.0, 378.5]

The minimal-to-low change is positive on 2Wiki and MuSiQue and unresolved with the opposite sign on HotpotQA. The pooled interaction is therefore concentrated at the first effort transition and varies across datasets.

Table 17: Evaluation sample: All datasets ($n = 750$). Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Completion tok.	Reader $/1k
Minimal	Full5	750	13.19 [12.94, 13.46]	0.285 [0.275, 0.294]
Minimal	Top2	750	13.51 [12.89, 14.54]	0.143 [0.138, 0.148]
Minimal	Top3	750	13.10 [12.89, 13.32]	0.190 [0.184, 0.196]
Minimal	Utility	750	12.98 [12.76, 13.22]	0.116 [0.113, 0.120]
Low	Full5	750	162.30 [148.64, 177.16]	0.673 [0.640, 0.708]
Low	Top2	750	150.58 [139.85, 161.91]	0.499 [0.472, 0.529]
Low	Top3	750	159.98 [148.49, 172.11]	0.572 [0.543, 0.603]
Low	Utility	750	145.64 [136.75, 154.69]	0.461 [0.439, 0.485]
Medium	Full5	750	423.41 [386.47, 467.49]	1.352 [1.259, 1.463]
Medium	Top2	750	387.32 [356.79, 419.40]	1.114 [1.035, 1.197]
Medium	Top3	750	404.60 [375.69, 437.83]	1.208 [1.133, 1.295]
Medium	Utility	750	367.08 [342.41, 391.03]	1.037 [0.974, 1.101]

Table 18: Evaluation sample: 2Wiki ($n = 250$). EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Full5	250	35.2 [29.4, 41.1]	41.5 [35.7, 47.2]
Minimal	Top2	250	33.6 [28.0, 39.6]	39.4 [34.1, 45.2]
Minimal	Top3	250	32.4 [26.8, 38.6]	38.8 [33.4, 44.6]
Minimal	Utility	250	40.0 [34.1, 46.0]	45.1 [39.3, 50.9]
Low	Full5	250	65.6 [59.8, 71.4]	74.0 [69.0, 78.9]
Low	Top2	250	64.4 [58.4, 70.3]	73.1 [68.1, 77.9]
Low	Top3	250	63.2 [57.1, 69.2]	71.8 [66.7, 77.0]
Low	Utility	250	62.0 [56.0, 67.9]	70.3 [65.0, 75.5]
Medium	Full5	250	65.2 [59.4, 71.0]	73.5 [68.5, 78.3]
Medium	Top2	250	64.4 [58.6, 70.4]	73.2 [68.1, 78.2]
Medium	Top3	250	63.6 [57.7, 69.5]	71.9 [66.8, 77.0]
Medium	Utility	250	63.2 [57.3, 69.1]	70.9 [65.6, 76.1]

Table 19: Evaluation sample: 2Wiki ($n = 250$). Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Full5	250	915.2 [877.8, 953.7]	0.0 [0.0, 0.0]
Minimal	Top2	250	353.2 [335.1, 372.3]	0.0 [0.0, 0.0]
Minimal	Top3	250	551.8 [524.8, 579.1]	0.0 [0.0, 0.0]
Minimal	Utility	250	264.9 [248.1, 283.5]	0.0 [0.0, 0.0]
Low	Full5	250	915.2 [877.8, 953.7]	151.0 [139.6, 162.6]
Low	Top2	250	353.2 [335.1, 372.3]	145.7 [133.4, 158.8]
Low	Top3	250	551.8 [524.8, 579.1]	154.9 [141.4, 168.8]
Low	Utility	250	264.9 [248.1, 283.5]	154.9 [143.6, 166.5]
Medium	Full5	250	915.2 [877.8, 953.7]	412.9 [380.4, 447.2]
Medium	Top2	250	353.2 [335.1, 372.3]	394.0 [361.2, 430.2]
Medium	Top3	250	551.8 [524.8, 579.1]	420.4 [384.8, 458.0]
Medium	Utility	250	264.9 [248.1, 283.5]	409.1 [379.1, 440.2]

Table 20: Evaluation sample: 2Wiki ($n = 250$). Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Completion tok.	Reader $/1k
Minimal	Full5	250	13.34 [13.01, 13.71]	0.331 [0.319, 0.344]
Minimal	Top2	250	13.39 [13.07, 13.76]	0.150 [0.144, 0.156]
Minimal	Top3	250	13.42 [13.10, 13.78]	0.215 [0.206, 0.224]
Minimal	Utility	250	13.38 [12.99, 13.78]	0.121 [0.116, 0.128]
Low	Full5	250	164.36 [152.93, 175.90]	0.725 [0.693, 0.758]
Low	Top2	250	158.93 [146.45, 172.16]	0.529 [0.495, 0.564]
Low	Top3	250	168.15 [154.68, 182.07]	0.617 [0.580, 0.654]
Low	Utility	250	168.22 [156.91, 179.96]	0.524 [0.495, 0.555]
Medium	Full5	250	426.23 [393.61, 460.61]	1.406 [1.321, 1.496]
Medium	Top2	250	407.11 [374.09, 443.24]	1.173 [1.086, 1.268]
Medium	Top3	250	433.62 [398.12, 471.19]	1.307 [1.214, 1.407]
Medium	Utility	250	422.42 [392.32, 453.50]	1.185 [1.107, 1.267]

Table 21: Evaluation sample: HotpotQA ($n = 250$). EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Full5	250	52.0 [45.8, 58.2]	66.0 [60.7, 71.1]
Minimal	Top2	250	49.6 [43.4, 56.0]	61.1 [55.6, 66.6]
Minimal	Top3	250	52.4 [46.2, 58.6]	65.9 [60.7, 71.1]
Minimal	Utility	250	51.6 [45.2, 57.8]	65.9 [60.5, 71.1]
Low	Full5	250	61.2 [55.2, 67.5]	77.8 [73.4, 82.1]
Low	Top2	250	60.4 [54.4, 66.4]	73.5 [68.7, 78.2]
Low	Top3	250	61.2 [55.2, 67.5]	75.6 [70.9, 80.1]
Low	Utility	250	62.4 [56.4, 68.4]	74.7 [69.7, 79.4]
Medium	Full5	250	63.2 [57.2, 69.2]	79.2 [74.9, 83.3]
Medium	Top2	250	60.0 [53.8, 66.3]	73.6 [68.7, 78.3]
Medium	Top3	250	62.0 [56.2, 68.3]	76.6 [72.1, 80.9]
Medium	Utility	250	62.0 [56.0, 67.9]	75.6 [70.7, 80.2]

Table 22: Evaluation sample: HotpotQA ($n = 250$). Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Full5	250	730.6 [711.9, 749.6]	0.0 [0.0, 0.0]
Minimal	Top2	250	320.2 [310.4, 330.5]	0.0 [0.0, 0.0]
Minimal	Top3	250	453.8 [441.3, 467.4]	0.0 [0.0, 0.0]
Minimal	Utility	250	277.5 [264.7, 292.2]	0.0 [0.0, 0.0]
Low	Full5	250	730.6 [711.9, 749.6]	105.5 [93.8, 118.0]
Low	Top2	250	320.2 [310.4, 330.5]	104.7 [92.4, 118.0]
Low	Top3	250	453.8 [441.3, 467.4]	101.4 [90.5, 113.3]
Low	Utility	250	277.5 [264.7, 292.2]	95.5 [85.8, 105.6]
Medium	Full5	250	730.6 [711.9, 749.6]	286.5 [257.0, 320.0]
Medium	Top2	250	320.2 [310.4, 330.5]	269.6 [242.6, 299.2]
Medium	Top3	250	453.8 [441.3, 467.4]	281.9 [251.9, 314.1]
Medium	Utility	250	277.5 [264.7, 292.2]	252.7 [227.8, 280.1]

Table 23: Evaluation sample: HotpotQA ($n = 250$). Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Completion tok.	Reader $/1k
Minimal	Full5	250	13.08 [12.68, 13.51]	0.272 [0.266, 0.278]
Minimal	Top2	250	12.83 [12.44, 13.30]	0.138 [0.134, 0.142]
Minimal	Top3	250	12.93 [12.58, 13.30]	0.182 [0.177, 0.186]
Minimal	Utility	250	12.84 [12.49, 13.22]	0.124 [0.120, 0.129]
Low	Full5	250	118.13 [106.50, 130.64]	0.545 [0.513, 0.579]
Low	Top2	250	117.12 [104.76, 130.51]	0.409 [0.377, 0.445]
Low	Top3	250	113.85 [102.96, 125.88]	0.444 [0.415, 0.476]
Low	Utility	250	108.02 [98.39, 118.13]	0.372 [0.346, 0.399]
Medium	Full5	250	298.95 [269.30, 332.61]	1.015 [0.937, 1.103]
Medium	Top2	250	281.95 [254.67, 311.61]	0.838 [0.767, 0.915]
Medium	Top3	250	294.30 [264.37, 326.61]	0.913 [0.835, 0.998]
Medium	Utility	250	265.11 [240.27, 292.66]	0.780 [0.715, 0.852]

Table 24: Evaluation sample: MuSiQue holdout ($n = 250$). EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Full5	250	18.4 [11.3, 26.8]	35.0 [27.8, 41.8]
Minimal	Top2	250	17.2 [10.9, 24.1]	34.4 [26.3, 42.4]
Minimal	Top3	250	19.6 [12.5, 27.9]	36.9 [29.0, 44.9]
Minimal	Utility	250	25.2 [17.5, 33.2]	39.7 [31.5, 47.4]
Low	Full5	250	50.0 [38.8, 61.5]	62.9 [53.2, 72.1]
Low	Top2	250	46.0 [34.7, 57.1]	58.2 [48.7, 67.5]
Low	Top3	250	49.2 [37.7, 60.7]	62.1 [52.5, 71.5]
Low	Utility	250	47.2 [36.5, 57.6]	59.5 [50.4, 68.4]
Medium	Full5	250	49.2 [37.6, 60.5]	61.3 [51.6, 70.5]
Medium	Top2	250	46.4 [35.7, 56.9]	58.1 [48.9, 67.3]
Medium	Top3	250	49.2 [37.1, 61.0]	62.0 [52.2, 71.2]
Medium	Utility	250	46.8 [35.4, 58.3]	57.8 [48.5, 67.3]

Table 25: Evaluation sample: MuSiQue holdout ($n = 250$). Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Full5	250	665.4 [611.9, 722.8]	0.0 [0.0, 0.0]
Minimal	Top2	250	315.5 [290.9, 343.8]	0.0 [0.0, 0.0]
Minimal	Top3	250	428.4 [394.5, 465.4]	0.0 [0.0, 0.0]
Minimal	Utility	250	216.0 [196.8, 238.3]	0.0 [0.0, 0.0]
Low	Full5	250	665.4 [611.9, 722.8]	192.3 [161.8, 225.6]
Low	Top2	250	315.5 [290.9, 343.8]	163.6 [140.3, 188.5]
Low	Top3	250	428.4 [394.5, 465.4]	185.9 [160.8, 212.3]
Low	Utility	250	216.0 [196.8, 238.3]	148.5 [128.3, 167.9]
Medium	Full5	250	665.4 [611.9, 722.8]	533.0 [445.5, 636.7]
Medium	Top2	250	315.5 [290.9, 343.8]	460.8 [390.0, 534.5]
Medium	Top3	250	428.4 [394.5, 465.4]	473.9 [407.0, 551.9]
Medium	Utility	250	216.0 [196.8, 238.3]	401.7 [347.7, 457.3]

Table 26: Evaluation sample: MuSiQue holdout ($n = 250$). Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Completion tok.	Reader $/1k
Minimal	Full5	250	13.16 [12.66, 13.84]	0.251 [0.233, 0.270]
Minimal	Top2	250	14.30 [12.66, 17.24]	0.140 [0.129, 0.156]
Minimal	Top3	250	12.96 [12.56, 13.39]	0.173 [0.162, 0.186]
Minimal	Utility	250	12.74 [12.33, 13.16]	0.104 [0.098, 0.111]
Low	Full5	250	204.42 [174.08, 237.75]	0.748 [0.674, 0.830]
Low	Top2	250	175.70 [152.69, 200.59]	0.560 [0.500, 0.625]
Low	Top3	250	197.94 [172.94, 224.27]	0.655 [0.590, 0.722]
Low	Utility	250	160.69 [140.30, 180.24]	0.489 [0.434, 0.543]
Medium	Full5	250	545.06 [457.77, 648.82]	1.634 [1.410, 1.895]
Medium	Top2	250	472.89 [402.25, 546.65]	1.333 [1.148, 1.523]
Medium	Top3	250	485.88 [419.06, 563.83]	1.403 [1.232, 1.605]
Medium	Utility	250	413.70 [359.64, 469.31]	1.147 [1.005, 1.295]

Table 27: Utility-minus-reference EM contrasts for All datasets ($n = 750$). Endpoint, adjacent-effort, and second-difference estimands are shown.

Reference	Estimand	EM pp [95% CI]	n	Comp.
Full5	Minimal	+3.7 [0.6, 6.9]	750	572
Full5	Low	-1.7 [-4.2, 0.8]	750	572
Full5	Medium	-1.9 [-4.5, 0.8]	750	572
Full5	Overall: min. − med.	+5.6 [1.3, 9.8]	750	572
Full5	Adjacent: min. − low	+5.5 [1.7, 9.2]	750	572
Full5	Adjacent: low − med.	+0.1 [-2.4, 2.7]	750	572
Full5	Adjacent second difference	+5.3 [0.4, 10.1]	750	572
Top2	Minimal	+5.5 [2.7, 8.3]	750	572
Top2	Low	+0.3 [-2.3, 2.8]	750	572
Top2	Medium	+0.4 [-2.3, 3.3]	750	572
Top2	Overall: min. − med.	+5.1 [1.8, 8.3]	750	572
Top2	Adjacent: min. − low	+5.2 [1.9, 8.4]	750	572
Top2	Adjacent: low − med.	-0.1 [-3.3, 2.7]	750	572
Top2	Adjacent second difference	+5.3 [-0.1, 10.6]	750	572
Top3	Minimal	+4.1 [0.9, 7.2]	750	572
Top3	Low	-0.7 [-3.1, 1.9]	750	572
Top3	Medium	-0.9 [-3.4, 1.6]	750	572
Top3	Overall: min. − med.	+5.1 [1.6, 8.4]	750	572
Top3	Adjacent: min. − low	+4.8 [1.2, 8.3]	750	572
Top3	Adjacent: low − med.	+0.3 [-2.4, 2.8]	750	572
Top3	Adjacent second difference	+4.5 [-0.7, 9.6]	750	572

Table 28: Utility-minus-reference EM contrasts for 2Wiki ($n = 250$). Endpoint, adjacent-effort, and second-difference estimands are shown.

Reference	Estimand	EM pp [95% CI]	n	Comp.
Full5	Minimal	+4.8 [-0.8, 10.1]	250	246
Full5	Low	-3.6 [-8.0, 1.2]	250	246
Full5	Medium	-2.0 [-6.0, 2.0]	250	246
Full5	Overall: min. − med.	+6.8 [0.0, 13.4]	250	246
Full5	Adjacent: min. − low	+8.4 [2.0, 14.6]	250	246
Full5	Adjacent: low − med.	-1.6 [-5.6, 2.8]	250	246
Full5	Adjacent second difference	+10.0 [1.6, 18.3]	250	246
Top2	Minimal	+6.4 [2.0, 11.2]	250	246
Top2	Low	-2.4 [-7.2, 2.8]	250	246
Top2	Medium	-1.2 [-5.2, 2.8]	250	246
Top2	Overall: min. − med.	+7.6 [2.0, 13.0]	250	246
Top2	Adjacent: min. − low	+8.8 [2.4, 15.0]	250	246
Top2	Adjacent: low − med.	-1.2 [-5.3, 3.2]	250	246
Top2	Adjacent second difference	+10.0 [0.8, 18.9]	250	246
Top3	Minimal	+7.6 [2.0, 12.9]	250	246
Top3	Low	-1.2 [-6.0, 3.6]	250	246
Top3	Medium	-0.4 [-4.4, 3.6]	250	246
Top3	Overall: min. − med.	+8.0 [1.6, 13.9]	250	246
Top3	Adjacent: min. − low	+8.8 [2.4, 15.1]	250	246
Top3	Adjacent: low − med.	-0.8 [-5.5, 4.0]	250	246
Top3	Adjacent second difference	+9.6 [0.0, 18.9]	250	246

Table 29: Utility-minus-reference EM contrasts for HotpotQA ($n = 250$). Endpoint, adjacent-effort, and second-difference estimands are shown.

Reference	Estimand	EM pp [95% CI]	n	Comp.
Full5	Minimal	-0.4 [-4.0, 3.2]	250	249
Full5	Low	+1.2 [-2.4, 4.8]	250	249
Full5	Medium	-1.2 [-4.4, 2.0]	250	249
Full5	Overall: min. − med.	+0.8 [-4.4, 6.0]	250	249
Full5	Adjacent: min. − low	-1.6 [-6.4, 3.2]	250	249
Full5	Adjacent: low − med.	+2.4 [-1.6, 6.4]	250	249
Full5	Adjacent second difference	-4.0 [-11.2, 3.2]	250	249
Top2	Minimal	+2.0 [-2.4, 6.4]	250	249
Top2	Low	+2.0 [-1.2, 5.2]	250	249
Top2	Medium	+2.0 [-1.6, 5.6]	250	249
Top2	Overall: min. − med.	+0.0 [-4.0, 4.0]	250	249
Top2	Adjacent: min. − low	+0.0 [-4.4, 4.4]	250	249
Top2	Adjacent: low − med.	+0.0 [-2.8, 3.2]	250	249
Top2	Adjacent second difference	+0.0 [-6.4, 6.3]	250	249
Top3	Minimal	-0.8 [-5.2, 3.2]	250	249
Top3	Low	+1.2 [-2.8, 5.2]	250	249
Top3	Medium	+0.0 [-4.0, 4.0]	250	249
Top3	Overall: min. − med.	-0.8 [-5.6, 4.0]	250	249
Top3	Adjacent: min. − low	-2.0 [-6.8, 2.8]	250	249
Top3	Adjacent: low − med.	+1.2 [-2.4, 4.4]	250	249
Top3	Adjacent second difference	-3.2 [-10.0, 3.6]	250	249

Table 30: Utility-minus-reference EM contrasts for MuSiQue holdout ($n = 250$). Endpoint, adjacent-effort, and second-difference estimands are shown.

Reference	Estimand	EM pp [95% CI]	n	Comp.
Full5	Minimal	+6.8 [-0.3, 13.6]	250	77
Full5	Low	-2.8 [-7.7, 2.0]	250	77
Full5	Medium	-2.4 [-8.2, 3.6]	250	77
Full5	Overall: min. − med.	+9.2 [-0.4, 18.4]	250	77
Full5	Adjacent: min. − low	+9.6 [1.8, 17.3]	250	77
Full5	Adjacent: low − med.	-0.4 [-5.2, 4.6]	250	77
Full5	Adjacent second difference	+10.0 [0.7, 19.0]	250	77
Top2	Minimal	+8.0 [2.9, 13.6]	250	77
Top2	Low	+1.2 [-3.5, 6.2]	250	77
Top2	Medium	+0.4 [-5.8, 7.6]	250	77
Top2	Overall: min. − med.	+7.6 [0.7, 14.2]	250	77
Top2	Adjacent: min. − low	+6.8 [0.4, 13.0]	250	77
Top2	Adjacent: low − med.	+0.8 [-7.7, 7.1]	250	77
Top2	Adjacent second difference	+6.0 [-4.9, 18.8]	250	77
Top3	Minimal	+5.6 [-1.3, 11.8]	250	77
Top3	Low	-2.0 [-6.4, 2.4]	250	77
Top3	Medium	-2.4 [-7.5, 2.7]	250	77
Top3	Overall: min. − med.	+8.0 [0.8, 15.1]	250	77
Top3	Adjacent: min. − low	+7.6 [0.4, 14.3]	250	77
Top3	Adjacent: low − med.	+0.4 [-4.9, 5.5]	250	77
Top3	Adjacent second difference	+7.2 [-3.0, 17.0]	250	77

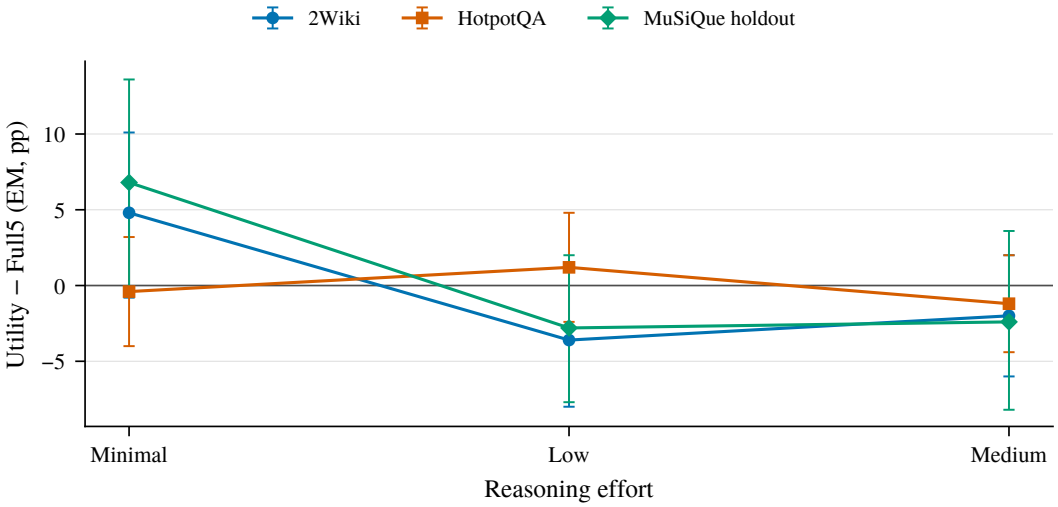

Figure 4: Paired UTILITY−FULL5 exact-match differences across effort on the three evaluation datasets. Bars are 95% component-bootstrap intervals; lines connect the three specified effort settings.

C.5 ABSTRACTIVE COMPRESSION

BRIEF-Pro exhibits an effort-dependent accuracy effect across the three datasets. The pooled minimal-minus-medium interaction is $+8.89$ EM points $[4.26, 13.48]$, with an F1 interaction of $+7.29$ $[3.24, 11.36]$. The within-effort EM effects are $+2.22$ $[-1.94, 6.28]$ at minimal effort and -6.67 $[-10.05, -3.43]$ at medium effort.

Table 31: Three-dataset BRIEF-Pro sample ($n = 450$), pooled and by dataset. EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Pooled	Minimal/Raw	450	49.6 [44.8, 54.8]	59.4 [55.0, 64.2]
Pooled	Minimal/BRIEF-Pro	450	51.8 [46.8, 56.7]	62.0 [57.4, 66.6]
Pooled	Medium/Raw	450	70.2 [65.5, 74.7]	79.9 [75.9, 83.6]
Pooled	Medium/BRIEF-Pro	450	63.6 [58.6, 68.4]	75.2 [71.2, 79.3]
2Wiki	Minimal/Raw	150	46.7 [38.4, 55.0]	56.2 [48.7, 63.6]
2Wiki	Minimal/BRIEF-Pro	150	48.7 [40.5, 56.7]	57.3 [49.8, 64.5]
2Wiki	Medium/Raw	150	75.3 [68.0, 82.2]	83.7 [78.3, 88.8]
2Wiki	Medium/BRIEF-Pro	150	67.3 [59.5, 75.0]	78.5 [72.7, 84.2]
HotpotQA	Minimal/Raw	150	66.0 [58.0, 73.3]	77.8 [71.8, 83.3]
HotpotQA	Minimal/BRIEF-Pro	150	69.3 [62.0, 76.5]	82.7 [77.5, 87.6]
HotpotQA	Medium/Raw	150	73.3 [66.2, 80.0]	85.2 [80.4, 89.6]
HotpotQA	Medium/BRIEF-Pro	150	71.3 [63.8, 78.7]	85.8 [80.9, 90.1]
MuSiQue	Minimal/Raw	150	36.0 [26.7, 45.7]	44.1 [35.1, 53.3]
MuSiQue	Minimal/BRIEF-Pro	150	37.3 [28.6, 46.7]	46.0 [37.3, 54.9]
MuSiQue	Medium/Raw	150	62.0 [52.6, 71.4]	70.8 [62.4, 79.1]
MuSiQue	Medium/BRIEF-Pro	150	52.0 [42.2, 61.6]	61.4 [52.4, 69.9]

Table 32: Three-dataset BRIEF-Pro sample ($n = 450$), pooled and by dataset. Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Pooled	Minimal/Raw	450	1054.4 [1014.3, 1094.8]	0.0 [0.0, 0.0]
Pooled	Minimal/BRIEF-Pro	450	344.0 [332.0, 355.2]	0.0 [0.0, 0.0]
Pooled	Medium/Raw	450	1054.4 [1014.3, 1094.8]	362.0 [325.3, 404.9]
Pooled	Medium/BRIEF-Pro	450	344.0 [332.0, 355.2]	350.3 [319.3, 382.3]
2Wiki	Minimal/Raw	150	1210.1 [1127.1, 1295.7]	0.0 [0.0, 0.0]
2Wiki	Minimal/BRIEF-Pro	150	367.7 [343.4, 392.9]	0.0 [0.0, 0.0]
2Wiki	Medium/Raw	150	1210.1 [1126.9, 1294.2]	286.3 [256.0, 322.1]
2Wiki	Medium/BRIEF-Pro	150	367.7 [344.0, 391.6]	318.7 [279.8, 361.3]
HotpotQA	Minimal/Raw	150	870.8 [834.5, 908.9]	0.0 [0.0, 0.0]
HotpotQA	Minimal/BRIEF-Pro	150	298.0 [287.4, 308.9]	0.0 [0.0, 0.0]
HotpotQA	Medium/Raw	150	870.8 [834.8, 908.3]	228.3 [207.4, 249.2]
HotpotQA	Medium/BRIEF-Pro	150	298.0 [287.2, 308.8]	222.7 [198.3, 252.1]
MuSiQue	Minimal/Raw	150	1082.2 [1017.3, 1144.1]	0.0 [0.0, 0.0]
MuSiQue	Minimal/BRIEF-Pro	150	366.3 [346.5, 385.7]	0.0 [0.0, 0.0]
MuSiQue	Medium/Raw	150	1082.2 [1019.9, 1146.2]	571.3 [488.2, 674.4]
MuSiQue	Medium/BRIEF-Pro	150	366.3 [346.7, 385.2]	509.4 [445.1, 575.1]

The pooled medium-effort difference in reported reasoning tokens is -11.66 $[-39.88, 15.13]$. Thus the compression result is an accuracy interaction with an unresolved minimal endpoint and no resolved increase in reported reasoning tokens.

Table 33: Three-dataset BRIEF-Pro sample ($n = 450$), pooled and by dataset. Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Visible tok.	Reader $/1k
Pooled	Minimal/Raw	450	14.02 [13.23, 15.28]	0.380 [0.366, 0.393]
Pooled	Minimal/BRIEF-Pro	450	14.08 [13.31, 15.12]	0.149 [0.144, 0.154]
Pooled	Medium/Raw	450	12.97 [12.74, 13.22]	1.317 [1.218, 1.434]
Pooled	Medium/BRIEF-Pro	450	13.03 [12.78, 13.30]	1.057 [0.975, 1.142]
2Wiki	Minimal/Raw	150	13.73 [13.21, 14.29]	0.429 [0.403, 0.458]
2Wiki	Minimal/BRIEF-Pro	150	13.88 [13.37, 14.43]	0.156 [0.148, 0.165]
2Wiki	Medium/Raw	150	13.46 [12.97, 13.95]	1.171 [1.080, 1.271]
2Wiki	Medium/BRIEF-Pro	150	13.53 [13.04, 14.04]	0.984 [0.880, 1.095]
HotpotQA	Minimal/Raw	150	12.87 [12.44, 13.36]	0.317 [0.305, 0.329]
HotpotQA	Minimal/BRIEF-Pro	150	13.05 [12.58, 13.55]	0.131 [0.127, 0.135]
HotpotQA	Medium/Raw	150	12.54 [12.15, 12.96]	0.910 [0.855, 0.967]
HotpotQA	Medium/BRIEF-Pro	150	12.67 [12.25, 13.15]	0.710 [0.645, 0.786]
MuSiQue	Minimal/Raw	150	15.47 [13.22, 18.93]	0.392 [0.370, 0.415]
MuSiQue	Minimal/BRIEF-Pro	150	15.32 [13.16, 18.20]	0.159 [0.149, 0.170]
MuSiQue	Medium/Raw	150	12.91 [12.58, 13.24]	1.871 [1.645, 2.147]
MuSiQue	Medium/BRIEF-Pro	150	12.90 [12.52, 13.29]	1.478 [1.307, 1.649]

Table 34: Paired EM contrasts for BRIEF-Pro versus raw context on the three-dataset sample ($n = 450$), in percentage points.

Panel	Contrast	Estimate [95% CI]	n	Components
Pooled	brief minus raw minimal	+2.2 [-1.9, 6.3]	450	403
Pooled	brief minus raw medium	-6.7 [-10.0, -3.4]	450	403
Pooled	interaction	+8.9 [4.3, 13.5]	450	403
2Wiki	brief minus raw minimal	+2.0 [-5.9, 9.5]	150	146
2Wiki	brief minus raw medium	-8.0 [-12.9, -3.4]	150	146
2Wiki	interaction	+10.0 [1.3, 19.2]	150	146
HotpotQA	brief minus raw minimal	+3.3 [-2.6, 9.3]	150	149
HotpotQA	brief minus raw medium	-2.0 [-6.0, 2.0]	150	149
HotpotQA	interaction	+5.3 [0.0, 10.7]	150	149
MuSiQue	brief minus raw minimal	+1.3 [-5.7, 8.2]	150	108
MuSiQue	brief minus raw medium	-10.0 [-17.4, -2.7]	150	108
MuSiQue	interaction	+11.3 [2.5, 20.5]	150	108

Table 35: Paired F1 contrasts for BRIEF-Pro versus raw context on the three-dataset sample ($n = 450$), in percentage points.

Panel	Contrast	Estimate [95% CI]	n	Components
Pooled	brief minus raw minimal	+2.63 [-1.09, 6.26]	450	403
Pooled	brief minus raw medium	-4.66 [-7.41, -2.05]	450	403
Pooled	interaction	+7.29 [3.24, 11.36]	450	403
2Wiki	brief minus raw minimal	+1.07 [-6.52, 8.53]	150	146
2Wiki	brief minus raw medium	-5.16 [-8.75, -1.73]	150	146
2Wiki	interaction	+6.23 [-2.20, 14.68]	150	146
HotpotQA	brief minus raw minimal	+4.91 [0.83, 9.16]	150	149
HotpotQA	brief minus raw medium	+0.56 [-2.05, 3.27]	150	149
HotpotQA	interaction	+4.34 [1.18, 7.90]	150	149
MuSiQue	brief minus raw minimal	+1.91 [-4.44, 8.32]	150	108
MuSiQue	brief minus raw medium	-9.38 [-16.04, -3.03]	150	108
MuSiQue	interaction	+11.29 [3.08, 19.69]	150	108

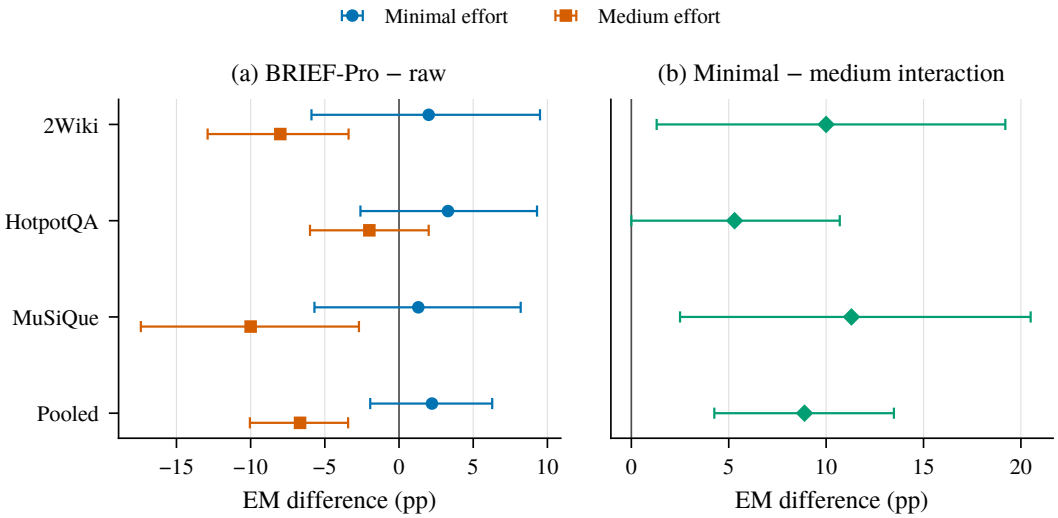

Figure 5: Paired BRIEF-Pro-minus-raw exact-match differences by dataset. Panel (a) shows effects within effort; panel (b) shows minimal-minus-medium interactions. Bars are 95% component-bootstrap intervals. The pooled row weights questions equally across the three datasets.

D Evidence Interventions

D.1 Support restoration across three datasets

Adding omitted annotated support improves EM relative to adding a passage-count-matched annotated non-support passage at both effort settings. The effects are $+14.12$ points $[11.33, 16.82]$ at minimal effort and $+9.33$ $[7.25, 11.49]$ at medium effort. The primary effort interaction is

$$\begin{aligned} I_{R,C} &= [\mathrm{EM}(R, \text{medium}) - \mathrm{EM}(C, \text{medium})] \\ &\quad - [\mathrm{EM}(R, \text{minimal}) - \mathrm{EM}(C, \text{minimal})] \\ &= -4.80\ [-7.93, -1.76]. \end{aligned} \tag{11}$$

Table 36: Support-restoration cells for the pooled three-dataset sample ($n = 1147$). The comparator adds a length-matched annotated non-support passage from the same five-document pool.

Effort	Context	n	EM [95% CI]	F1 [95% CI]
Minimal	Restore support	1147	60.2 [57.2, 63.1]	74.2 [71.9, 76.5]
Minimal	Matched non-support	1147	46.1 [43.2, 49.1]	58.7 [56.0, 61.4]
Medium	Restore support	1147	71.8 [68.9, 74.5]	83.7 [81.8, 85.6]
Medium	Matched non-support	1147	62.4 [59.5, 65.4]	75.3 [73.0, 77.5]

Table 37: Support-restoration cells for the 2Wiki sample ($n = 419$). The comparator adds a length-matched annotated non-support passage from the same five-document pool.

Effort	Context	n	EM [95% CI]	F1 [95% CI]
Minimal	Restore support	419	55.8 [50.9, 60.7]	70.1 [66.2, 74.0]
Minimal	Matched non-support	419	39.4 [34.5, 44.4]	50.4 [46.1, 54.8]
Medium	Restore support	419	69.9 [65.1, 74.3]	81.8 [78.5, 85.1]
Medium	Matched non-support	419	60.6 [55.8, 65.5]	74.3 [70.5, 78.0]

On the same 1,147 questions at medium effort, mean reported reasoning falls from 392.1 tokens in the non-support control to 226.8 after support restoration. The 42.2% reduction is computed from these pooled means as $(\overline{G}_C - \overline{G}_R)/\overline{G}_C$, where $\overline{G}_C$ and $\overline{G}_R$ denote mean reported reasoning tokens

Table 38: Support-restoration cells for the HotpotQA sample ($n = 455$). The comparator adds a length-matched annotated non-support passage from the same five-document pool.

Effort	Context	n	EM [95% CI]	F1 [95% CI]
Minimal	Restore support	455	66.8 [62.4, 71.1]	81.3 [78.2, 84.1]
Minimal	Matched non-support	455	54.9 [50.3, 59.6]	69.6 [65.9, 73.2]
Medium	Restore support	455	72.7 [68.6, 76.8]	85.6 [83.1, 88.1]
Medium	Matched non-support	455	67.3 [62.8, 71.5]	80.5 [77.4, 83.5]

Table 39: Support-restoration cells for the MuSiQue sample ($n = 273$). The comparator adds a length-matched annotated non-support passage from the same five-document pool.

Effort	Context	n	EM [95% CI]	F1 [95% CI]
Minimal	Restore support	273	56.0 [49.1, 62.8]	68.7 [63.1, 74.1]
Minimal	Matched non-support	273	41.8 [34.6, 48.7]	53.2 [47.1, 59.4]
Medium	Restore support	273	72.9 [66.4, 78.9]	83.4 [78.6, 87.8]
Medium	Matched non-support	273	57.1 [49.6, 64.0]	68.0 [62.2, 73.5]

Table 40: Support-restoration cells for the pooled length-gap subset ($n = 700$). The comparator adds a length-matched annotated non-support passage from the same five-document pool.

Effort	Context	n	EM [95% CI]	F1 [95% CI]
Minimal	Restore support	700	63.3 [59.6, 67.0]	76.0 [73.2, 78.8]
Minimal	Matched non-support	700	49.1 [45.1, 53.0]	60.7 [57.2, 64.1]
Medium	Restore support	700	74.9 [71.3, 78.2]	85.2 [82.8, 87.5]
Medium	Matched non-support	700	65.3 [61.6, 69.0]	76.7 [73.9, 79.5]

Table 41: Restoration-minus-matched-non-support effects for the pooled three-dataset sample ($n = 1147$).

Metric	Estimand	Estimate [95% CI]	n	Comp.
EM	Minimal	+14.1 [11.3, 16.8]	1147	1046
EM	Medium	+9.3 [7.3, 11.5]	1147	1046
EM	Med. − min. interaction	-4.8 [-7.9, -1.8]	1147	1046
F1	Minimal	+15.5 [13.1, 17.9]	1147	1046
F1	Medium	+8.4 [6.7, 10.4]	1147	1046
F1	Med. − min. interaction	-7.1 [-9.7, -4.6]	1147	1046

Table 42: Restoration-minus-matched-non-support effects for the 2Wiki sample ($n = 419$).

Metric	Estimand	Estimate [95% CI]	n	Comp.
EM	Minimal	+16.5 [11.8, 21.2]	419	404
EM	Medium	+9.3 [5.7, 13.0]	419	404
EM	Med. − min. interaction	-7.2 [-12.4, -1.9]	419	404
F1	Minimal	+19.7 [15.4, 23.9]	419	404
F1	Medium	+7.5 [4.3, 10.6]	419	404
F1	Med. − min. interaction	-12.2 [-16.8, -7.7]	419	404

Table 43: Restoration-minus-matched-non-support effects for the HotpotQA sample ($n = 455$).

Metric	Estimand	Estimate [95% CI]	n	Comp.
EM	Minimal	+11.9 [7.9, 15.8]	455	453
EM	Medium	+5.5 [2.6, 8.4]	455	453
EM	Med. − min. interaction	-6.4 [-10.8, -2.0]	455	453
F1	Minimal	+11.6 [8.4, 15.1]	455	453
F1	Medium	+5.1 [2.9, 7.4]	455	453
F1	Med. − min. interaction	-6.6 [-9.9, -3.2]	455	453

Table 44: Restoration-minus-matched-non-support effects for the MuSiQue sample ($n = 273$).

Metric	Estimand	Estimate [95% CI]	n	Comp.
EM	Minimal	+14.3 [8.5, 20.3]	273	189
EM	Medium	+15.8 [10.5, 21.0]	273	189
EM	Med. − min. interaction	+1.5 [-5.6, 8.3]	273	189
F1	Minimal	+15.5 [10.1, 21.0]	273	189
F1	Medium	+15.4 [11.0, 19.9]	273	189
F1	Med. − min. interaction	-0.1 [-5.9, 5.8]	273	189

Table 45: Restoration-minus-matched-non-support effects for the pooled length-gap subset ($n = 700$).

Metric	Estimand	Estimate [95% CI]	n	Comp.
EM	Minimal	+14.1 [10.6, 17.7]	700	653
EM	Medium	+9.6 [6.9, 12.2]	700	653
EM	Med. − min. interaction	-4.6 [-8.4, -0.9]	700	653
F1	Minimal	+15.4 [12.2, 18.4]	700	653
F1	Medium	+8.5 [6.3, 10.8]	700	653
F1	Med. − min. interaction	-6.8 [-10.1, -3.5]	700	653

in the control and restoration conditions. Table 46 reports the paired absolute differences and their intervals.

Table 46: Resource differences for restoration minus matched non-support on the pooled three-dataset sample ($n = 1147$).

Metric	Estimand	Estimate [CI]	n	Comp.
Input tokens	Minimal	+7.3 [2.7, 12.1]	1147	1046
Input tokens	Medium	+7.3 [2.4, 12.2]	1147	1046
Input tokens	Med. − min. interaction	+0.0 [0.0, 0.0]	1147	1046
Reasoning tokens	Minimal	+0.0 [0.0, 0.0]	1147	1046
Reasoning tokens	Medium	-165.3 [-180.3, -150.2]	1147	1046
Reasoning tokens	Med. − min. interaction	-165.3 [-180.9, -150.2]	1147	1046
Visible tokens	Minimal	+0.04 [-0.12, 0.18]	1147	1046
Visible tokens	Medium	+0.06 [-0.01, 0.14]	1147	1046
Visible tokens	Med. − min. interaction	+0.02 [-0.13, 0.19]	1147	1046
Reader $/1k	Minimal	+0.002 [0.001, 0.004]	1147	1046
Reader $/1k	Medium	-0.427 [-0.467, -0.389]	1147	1046
Reader $/1k	Med. − min. interaction	-0.430 [-0.470, -0.390]	1147	1046

Table 47: Resource differences for restoration minus matched non-support on the pooled length-gap subset ($n = 700$).

Metric	Estimand	Estimate [CI]	n	Comp.
Input tokens	Minimal	+0.1 [-1.1, 1.3]	700	653
Input tokens	Medium	+0.1 [-1.1, 1.2]	700	653
Input tokens	Med. − min. interaction	+0.0 [0.0, 0.0]	700	653
Reasoning tokens	Minimal	+0.0 [0.0, 0.0]	700	653
Reasoning tokens	Medium	-166.9 [-186.9, -148.4]	700	653
Reasoning tokens	Med. − min. interaction	-166.9 [-186.1, -148.9]	700	653
Visible tokens	Minimal	+0.06 [-0.14, 0.25]	700	653
Visible tokens	Medium	+0.08 [-0.01, 0.18]	700	653
Visible tokens	Med. − min. interaction	+0.02 [-0.18, 0.23]	700	653
Reader $/1k	Minimal	+0.000 [-0.000, 0.001]	700	653
Reader $/1k	Medium	-0.434 [-0.483, -0.388]	700	653
Reader $/1k	Med. − min. interaction	-0.434 [-0.484, -0.385]	700	653

Restricting the analysis to the 700 questions whose two reader messages differ by at most 32 tokens gives an interaction of -4.57 $[-8.37, -0.86]$. At medium effort, the resource difference is primarily in reported reasoning tokens rather than visible answer length. Because the added passages follow their acquisition ranks, the $R - C$ estimand compares the complete oracle input constructions, including any rank difference. It characterizes the value of annotated support within the eligible population; deployment would require a support detector unavailable in this experiment.

D.2 Overlapping 2Wiki intervention subset

The overlapping 419-question 2Wiki subset shows both restoration gains and effort-dependent attenuation of the matched non-support penalty. At minimal effort, $R-U$, $C-U$, and $R-C$ are $+5.73$ $[1.42, 9.79]$, -10.74 $[-14.98, -6.68]$, and $+16.47$ $[11.85, 21.01]$ EM points. At medium effort, the corresponding effects are $+7.40$ $[4.27, 10.66]$, -1.91 $[-4.56, 0.72]$, and $+9.31$ $[5.78, 12.95]$.

Table 48: Overlapping 2Wiki intervention subset ($n = 419$). EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Utility	419	50.1 [45.2, 55.2]	62.2 [58.0, 66.5]
Minimal	Restore support	419	55.8 [50.9, 60.8]	70.1 [66.3, 73.9]
Minimal	Matched non-support	419	39.4 [34.6, 44.0]	50.4 [46.0, 54.7]
Minimal	Full5	419	50.4 [45.7, 55.0]	64.2 [60.2, 68.2]
Medium	Utility	419	62.5 [57.7, 67.3]	75.4 [71.8, 78.9]
Medium	Restore support	419	69.9 [65.2, 74.4]	81.8 [78.5, 84.9]
Medium	Matched non-support	419	60.6 [55.8, 65.5]	74.3 [70.6, 78.0]
Medium	Full5	419	69.9 [65.2, 74.5]	81.9 [78.8, 85.1]

Table 49: Overlapping 2Wiki intervention subset ($n = 419$). Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Utility	419	321.7 [304.8, 339.1]	0.0 [0.0, 0.0]
Minimal	Restore support	419	480.3 [457.9, 504.1]	0.0 [0.0, 0.0]
Minimal	Matched non-support	419	451.3 [431.0, 472.0]	0.0 [0.0, 0.0]
Minimal	Full5	419	943.1 [902.8, 983.8]	0.0 [0.0, 0.0]
Medium	Utility	419	321.7 [304.9, 339.3]	360.6 [336.9, 386.0]
Medium	Restore support	419	480.3 [457.3, 504.0]	221.8 [208.2, 236.7]
Medium	Matched non-support	419	451.3 [430.2, 472.6]	470.1 [442.4, 499.2]
Medium	Full5	419	943.1 [902.9, 983.7]	254.6 [239.3, 271.1]

Table 50: Overlapping 2Wiki intervention subset ($n = 419$). Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Visible tok.	Reader $/1k
Minimal	Utility	419	13.74 [13.49, 14.01]	0.141 [0.135, 0.147]
Minimal	Restore support	419	13.82 [13.58, 14.07]	0.192 [0.185, 0.200]
Minimal	Matched non-support	419	13.89 [13.60, 14.21]	0.183 [0.176, 0.190]
Minimal	Full5	419	14.01 [13.73, 14.29]	0.341 [0.329, 0.355]
Medium	Utility	419	13.40 [13.17, 13.64]	1.078 [1.016, 1.145]
Medium	Restore support	419	13.54 [13.30, 13.77]	0.769 [0.729, 0.808]
Medium	Matched non-support	419	13.51 [13.27, 13.75]	1.405 [1.333, 1.482]
Medium	Full5	419	13.58 [13.34, 13.83]	1.005 [0.961, 1.053]

The medium-minus-minimal interaction for $C - U$ is $+8.83$ $[4.03, 13.40]$. These paired comparisons explain behavior within the shared parent population; the subset reuses questions and responses from the comparison factorial and support-restoration study.

Table 51: EM contrasts on the overlapping 2Wiki intervention subset ($n = 419$).

Effort	Contrast	EM pp [95% CI]	n	Comp.
Minimal	R − U	+5.7 [1.4, 9.8]	419	404
Minimal	C − U	-10.7 [-15.0, -6.7]	419	404
Minimal	R − C	+16.5 [11.8, 21.0]	419	404
Minimal	R − F	+5.5 [2.2, 9.1]	419	404
Minimal	F − U	+0.2 [-4.1, 4.3]	419	404
Minimal	U − F	-0.2 [-4.3, 3.8]	419	404
Medium	R − U	+7.4 [4.3, 10.7]	419	404
Medium	C − U	-1.9 [-4.6, 0.7]	419	404
Medium	R − C	+9.3 [5.8, 12.9]	419	404
Medium	R − F	+0.0 [-2.2, 2.3]	419	404
Medium	F − U	+7.4 [3.9, 10.7]	419	404
Medium	U − F	-7.4 [-10.8, -4.0]	419	404
Med. − min.	R − U	+1.7 [-2.6, 5.9]	419	404
Med. − min.	C − U	+8.8 [4.0, 13.4]	419	404
Med. − min.	R − C	-7.2 [-12.4, -1.7]	419	404
Med. − min.	R − F	-5.5 [-9.7, -1.2]	419	404
Med. − min.	F − U	+7.2 [2.3, 12.3]	419	404
Med. − min.	U − F	-7.2 [-12.0, -2.1]	419	404

D.3 Held-out position-matched intervention

The held-out 135-question subset does not resolve the primary interaction for harm from adding matched non-support evidence. The effect $(C - U)_{\text{medium}} - (C - U)_{\text{minimal}}$ is $+3.70$ points $[-3.62, 11.03]$. Its minimal-to-low component is positive, followed by a nonmonotonic low-to-medium change.

Restoration exceeds the matched control by $+12.59$ $[5.22, 20.15]$ at minimal effort and $+6.67$ $[1.48, 11.94]$ at low effort; the medium estimate is $+4.44$ $[-1.48, 10.37]$. The medium-minus-minimal $R - C$ interaction is -8.15 $[-15.67, -0.74]$.

Table 52: Held-out 2Wiki intervention subset ($n = 135$). EM and F1 cell means are percentages with 95% component-bootstrap intervals.

Panel	Action	n	EM	F1
Minimal	Utility	135	63.7 [55.5, 71.9]	73.5 [66.7, 80.0]
Minimal	Restore support	135	71.1 [63.2, 78.7]	81.5 [75.7, 87.0]
Minimal	Matched non-support	135	58.5 [50.4, 66.9]	67.9 [60.8, 75.0]
Minimal	Full5	135	61.5 [53.0, 70.1]	72.2 [65.4, 79.0]
Low	Utility	135	69.6 [61.5, 77.6]	80.2 [74.3, 86.1]
Low	Restore support	135	79.3 [72.1, 85.9]	89.3 [84.9, 93.3]
Low	Matched non-support	135	72.6 [64.7, 80.0]	82.0 [76.3, 87.4]
Low	Full5	135	79.3 [72.1, 85.8]	88.5 [83.8, 92.7]
Medium	Utility	135	75.6 [68.1, 82.8]	84.6 [79.2, 89.7]
Medium	Restore support	135	78.5 [71.1, 85.8]	88.5 [83.9, 92.8]
Medium	Matched non-support	135	74.1 [66.7, 81.3]	84.6 [79.3, 89.6]
Medium	Full5	135	80.7 [73.5, 87.3]	89.9 [85.4, 93.9]

The position-matched construction supports an effort-dependent reduction in the relative benefit of restored support. The unresolved primary $C - U$ interaction indicates that matched non-support addition alone does not account for the natural-context reversal.

Table 53: Held-out 2Wiki intervention subset ($n = 135$). Reader-input and reported-reasoning cell means with 95% component-bootstrap intervals; token counts are per question.

Panel	Action	n	Input tok.	Reasoning tok.
Minimal	Utility	135	279.7 [252.1, 309.4]	0.0 [0.0, 0.0]
Minimal	Restore support	135	431.9 [393.9, 472.3]	0.0 [0.0, 0.0]
Minimal	Matched non-support	135	418.2 [383.0, 456.2]	0.0 [0.0, 0.0]
Minimal	Full5	135	922.8 [869.1, 977.3]	0.0 [0.0, 0.0]
Low	Utility	135	279.7 [252.1, 309.4]	155.0 [138.0, 172.9]
Low	Restore support	135	431.9 [393.9, 472.3]	77.3 [66.8, 88.8]
Low	Matched non-support	135	418.2 [383.0, 456.2]	156.0 [141.4, 171.1]
Low	Full5	135	922.8 [869.1, 977.3]	87.2 [78.7, 96.2]
Medium	Utility	135	279.7 [252.1, 309.4]	362.7 [319.5, 409.9]
Medium	Restore support	135	431.9 [393.9, 472.3]	202.0 [179.6, 232.0]
Medium	Matched non-support	135	418.2 [383.0, 456.2]	413.4 [366.9, 459.6]
Medium	Full5	135	922.8 [869.1, 977.3]	239.4 [212.9, 271.8]

Table 54: Held-out 2Wiki intervention subset ($n = 135$). Reader-output and billed-cost cell means with 95% component-bootstrap intervals. Reader cost is dollars per 1,000 questions and excludes retrieval and selection.

Panel	Action	n	Visible tok.	Reader $/1k
Minimal	Utility	135	13.89 [13.39, 14.41]	0.127 [0.118, 0.137]
Minimal	Restore support	135	13.70 [13.28, 14.13]	0.176 [0.164, 0.190]
Minimal	Matched non-support	135	13.90 [13.34, 14.52]	0.173 [0.161, 0.185]
Minimal	Full5	135	14.07 [13.47, 14.74]	0.337 [0.319, 0.355]
Low	Utility	135	13.70 [13.24, 14.19]	0.530 [0.485, 0.579]
Low	Restore support	135	13.60 [13.17, 14.06]	0.377 [0.345, 0.413]
Low	Matched non-support	135	13.54 [13.11, 13.99]	0.577 [0.536, 0.620]
Low	Full5	135	13.66 [13.23, 14.11]	0.560 [0.530, 0.591]
Medium	Utility	135	13.56 [13.15, 13.98]	1.070 [0.957, 1.194]
Medium	Restore support	135	13.58 [13.16, 14.02]	0.701 [0.638, 0.782]
Medium	Matched non-support	135	13.59 [13.16, 14.03]	1.247 [1.123, 1.369]
Medium	Full5	135	13.59 [13.17, 14.03]	0.956 [0.882, 1.046]

Table 55: Paired EM contrasts on the held-out 2Wiki intervention subset ($n = 135$).

Contrast	Estimand	EM pp [95% CI]	n	Comp.
C − U	Minimal	-5.2 [-11.9, 0.7]	135	134
C − U	Low	+3.0 [-2.2, 8.1]	135	134
C − U	Medium	-1.5 [-5.2, 2.2]	135	134
C − U	Interaction: med. − min.	+3.7 [-3.6, 11.0]	135	134
C − U	Interaction: low − min.	+8.1 [0.7, 15.7]	135	134
C − U	Interaction: med. − low	-4.4 [-9.0, 0.0]	135	134
R − U	Minimal	+7.4 [0.7, 14.2]	135	134
R − U	Low	+9.6 [4.4, 15.6]	135	134
R − U	Medium	+3.0 [-2.2, 8.8]	135	134
R − U	Interaction: med. − min.	-4.4 [-11.9, 3.0]	135	134
R − U	Interaction: low − min.	+2.2 [-4.4, 8.9]	135	134
R − U	Interaction: med. − low	-6.7 [-11.9, -1.5]	135	134
R − F	Minimal	+9.6 [3.0, 16.2]	135	134
R − F	Low	+0.0 [-3.0, 3.0]	135	134
R − F	Medium	-2.2 [-5.9, 0.7]	135	134
R − F	Interaction: med. − min.	-11.9 [-18.7, -4.4]	135	134
R − F	Interaction: low − min.	-9.6 [-16.4, -2.9]	135	134
R − F	Interaction: med. − low	-2.2 [-6.0, 1.5]	135	134
R − C	Minimal	+12.6 [5.2, 20.1]	135	134
R − C	Low	+6.7 [1.5, 11.9]	135	134
R − C	Medium	+4.4 [-1.5, 10.4]	135	134
R − C	Interaction: med. − min.	-8.1 [-15.7, -0.7]	135	134
R − C	Interaction: low − min.	-5.9 [-13.3, 1.5]	135	134
R − C	Interaction: med. − low	-2.2 [-6.0, 2.2]	135	134

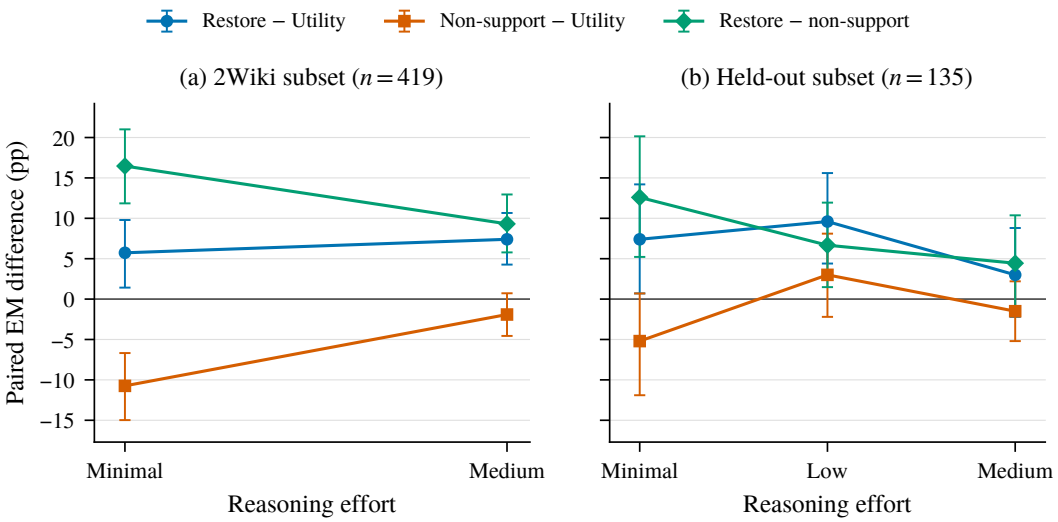

Figure 6: Paired EM differences for support restoration (R), matched annotated non-support addition (C), and Utility evidence (U). The 2Wiki comparison subset and held-out position-matched subset are displayed separately. Bars are 95% component-bootstrap intervals.

D.4 Cross-reader restoration

The restoration advantage is positive under three reader regimens on a shared 1,286-question development population. This population comprises the 1,147 one-omission questions above and 139 questions with two omitted support documents, grouped into 1,139 dependency components. Table 56 reports the pooled $R - C$ effects; the dataset-specific contrasts were the primary analyses.

Qwen2.5-3B-Instruct (Qwen Team, 2024), revision `aa8e72537993ba99e69dfaafa59ed015b17504d1`, and Phi-3.5-mini-instruct (Abdin et al., 2024), revision `2fe192450127e6a83f7441aef6e3ca586c338b77`, use a common short-answer prompt, native chat templates, FP16 greedy decoding, and a 64-token output limit. The GPT-5 mini condition uses the `openai/gpt-5-mini` API alias, omits an explicit reasoning-effort field, and sets a 4,000-token output limit. This regimen is distinct from the dated reader used in the factorial experiments.

Table 56: Cross-reader restoration-minus-size-control EM effects on the pooled three-dataset sample ($n = 1,286$; 1,139 bootstrap components). The sample comprises 1,147 one-omission and 139 two-omission questions.

Reader	Regimen	EM effect [95% CI]	n
Qwen2.5-3B-Instruct	Greedy; 64-token cap	+17.26 [14.53, 20.08]	1286
GPT-5 mini	Default-effort alias; 4,000-token cap	+11.43 [9.19, 13.73]	1286
Phi-3.5-mini-instruct	Greedy; 64-token cap	+15.63 [13.22, 18.17]	1286

The shared questions and contexts make this a comparison of reader regimens on one population. The consistent direction across readers strengthens the input-intervention result while retaining the pooled analysis's descriptive status.

D.5 Illustrative response patterns

Table 57 presents two descriptive cases from the held-out intervention subset. A fixed post-analysis rule selected the lexicographically first identifier for two opposing patterns: U correct and C incorrect only at minimal effort (12 matches), and the corresponding pattern only at medium effort (3 matches). Parenthetical values are official alias-aware normalized EM. The examples have no inferential weight and include only final answers and short passage prefixes.

Table 57: Illustrative response patterns under Utility evidence (U), support restoration (R), and matched annotated non-support addition (C).

Case and question	Context titles	Minimal final answer (EM)	Medium final answer (EM)
A. Who died first, Shaikh Ayaz or Horton Foote? Gold: Shaikh Ayaz	U: Horton Foote C adds: Mary Hunter Wolf	U: Shaikh Ayaz (1) R: Shaikh Ayaz (1) C: Horton Foote (0)	U: Shaikh Ayaz (1) R: Shaikh Ayaz (1) C: Shaikh Ayaz (1)
Added-passage opening: "Mary Hunter Wolf(1905? – November 3, 2000), born Mary Hunter, was an American theater director and producer. She made her Broadway directorial debut on 4 April 1944 at the Bijou Theater with Only the Heart, the first play written by Horton Foote."			
B. Which film was released more recently, 3 Nafas Likas or The Bravest Way? Gold: 3 Nafas Likas	U: The Bravest Way C adds: 3 (2012 Tamil film)	U: The Bravest Way (0) R: 3 Nafas Likas (1) C: 3 Nafas Likas (1)	U: 3 Nafas Likas (1) R: 3 Nafas Likas (1) C: 3 (0)
Added-passage opening: "3 is a 2012 Indian Tamil-language romantic psychological thriller film written and directed by Aishwarya R. Dhanush in her directorial debut."			

The opposite transitions illustrate within-question heterogeneity across effort and do not estimate the frequency of either response pattern.

E CALIBRATION DETAILS

Joint calibration changes the selected policy-effort action as reader cost receives greater weight, and its held-out advantage over sequential calibration changes sign across price settings. The calibration population is the 750-question three-dataset sample. We consider a three-dataset scope and a nested 500-question HotpotQA–MuSiQue scope; both are evaluated on the held-out 2Wiki sample.

E.1 PRICE-BASED ACTION SELECTION

Section 3.3 defines the calibration objective (Equation 5) and the joint, medium-first, and minimal-first selection rules. For the final tie-break, the 12 actions are ordered first by effort (minimal, low, medium) and then by policy (FULL5, TOP2, TOP3, UTILITY). The tables below report the held-out accuracy and cost of every action and the actions selected at each λ.

Table 58: Accuracy of all twelve policy–effort actions on the held-out 2Wiki sample ($n = 500$).

Action	EM [95% CI]	F1 [95% CI]	n	Comp.
Minimal/Full5	42.6 [38.4, 46.8]	48.8 [44.7, 52.9]	500	490
Minimal/Top2	44.6 [40.2, 49.1]	50.2 [46.0, 54.3]	500	490
Minimal/Top3	42.2 [37.9, 46.7]	47.9 [43.8, 52.1]	500	490
Minimal/Utility	47.0 [42.5, 51.5]	53.2 [49.1, 57.4]	500	490
Low/Full5	66.4 [62.2, 70.5]	74.9 [71.4, 78.3]	500	490
Low/Top2	60.8 [56.6, 65.1]	69.2 [65.5, 73.0]	500	490
Low/Top3	63.4 [59.3, 67.5]	71.8 [68.1, 75.2]	500	490
Low/Utility	59.6 [55.4, 63.8]	68.7 [64.9, 72.4]	500	490
Medium/Full5	66.0 [61.9, 70.1]	74.8 [71.3, 78.2]	500	490
Medium/Top2	62.4 [58.1, 66.7]	70.2 [66.5, 73.8]	500	490
Medium/Top3	64.0 [59.8, 68.3]	72.1 [68.4, 75.6]	500	490
Medium/Utility	63.0 [58.8, 67.3]	71.0 [67.4, 74.7]	500	490

Table 59: Reader cost for all twelve policy–effort actions on the held-out 2Wiki sample ($n = 500$), in dollars per 1,000 questions.

Action	Reader $/1k [95% CI]	n	Comp.
Minimal/Full5	0.323 [0.314, 0.332]	500	490
Minimal/Top2	0.165 [0.160, 0.171]	500	490
Minimal/Top3	0.219 [0.212, 0.226]	500	490
Minimal/Utility	0.129 [0.124, 0.135]	500	490
Low/Full5	0.638 [0.615, 0.661]	500	490
Low/Top2	0.478 [0.454, 0.501]	500	490
Low/Top3	0.543 [0.518, 0.568]	500	490
Low/Utility	0.492 [0.468, 0.519]	500	490
Medium/Full5	1.214 [1.151, 1.275]	500	490
Medium/Top2	1.071 [1.005, 1.141]	500	490
Medium/Top3	1.127 [1.058, 1.200]	500	490
Medium/Utility	1.108 [1.039, 1.181]	500	490

Table 60: Complete-menu calibration on the MuSiQue + HotpotQA calibration sample ($n = 500$), evaluated on the held-out 2Wiki sample ($n = 500$) relative to the medium-first baseline. Entries are net-utility differences in percentage points.

λ	Joint action	Medium-first action	Δ net utility [95% CI]
0	Medium/Full5	Medium/Full5	+0.00 [0.00, 0.00]
100	Low/Top3	Low/Full5	-2.05 [-4.34, 0.20]
250	Low/Utility	Low/Full5	-4.16 [-7.06, -1.30]
500	Low/Utility	Low/Full5	-0.53 [-3.76, 2.59]
1000	Minimal/Utility	Minimal/Full5	+22.81 [19.57, 26.26]

Table 61: Complete-menu calibration on the MuSiQue + HotpotQA calibration sample ($n = 500$), evaluated on the held-out 2Wiki sample ($n = 500$) relative to the minimal-first baseline. Entries are net-utility differences in percentage points.

λ	Joint action	Minimal-first action	Δ net utility [95% CI]
0	Medium/Full5	Low/Utility	+7.40 [4.62, 10.27]
100	Low/Top3	Low/Utility	+4.29 [1.30, 7.26]
250	Low/Utility	Low/Utility	+0.00 [0.00, 0.00]
500	Low/Utility	Low/Utility	+0.00 [0.00, 0.00]
1000	Minimal/Utility	Minimal/Utility	+0.00 [0.00, 0.00]

Table 62: Complete-menu calibration on the three-dataset calibration sample ($n = 750$), evaluated on the held-out 2Wiki sample ($n = 500$) relative to the medium-first baseline. Entries are net-utility differences in percentage points.

λ	Joint action	Medium-first action	Δ net utility [95% CI]
0	Medium/Full5	Medium/Full5	+0.00 [0.00, 0.00]
100	Low/Full5	Low/Full5	+0.00 [0.00, 0.00]
250	Low/Utility	Low/Full5	-4.16 [-7.06, -1.30]
500	Low/Utility	Low/Full5	-0.53 [-3.76, 2.59]
1000	Minimal/Utility	Minimal/Full5	+22.81 [19.57, 26.26]

Table 63: Complete-menu calibration on the three-dataset calibration sample ($n = 750$), evaluated on the held-out 2Wiki sample ($n = 500$) relative to the minimal-first baseline. Entries are net-utility differences in percentage points.

λ	Joint action	Minimal-first action	Δ net utility [95% CI]
0	Medium/Full5	Medium/Utility	+4.00 [1.40, 6.62]
100	Low/Full5	Low/Utility	+6.35 [3.57, 9.16]
250	Low/Utility	Low/Utility	+0.00 [0.00, 0.00]
500	Low/Utility	Low/Utility	+0.00 [0.00, 0.00]
1000	Minimal/Utility	Minimal/Utility	+0.00 [0.00, 0.00]

Under three-dataset calibration, the joint actions across increasing λ are medium-FULL5, low-FULL5, low-UTILITY, low-UTILITY, and minimal-UTILITY. The medium-first actions are medium-FULL5, low-FULL5, low-FULL5, low-FULL5, and minimal-FULL5. On held-out 2Wiki, joint minus medium-first net utility is -4.164 points $[-7.055, -1.303]$ at $\lambda = 250$, -0.527 $[-3.756, 2.591]$ at $\lambda = 500$, and $+22.805$ $[19.569, 26.257]$ at $\lambda = 1000$. Joint search expands the action set, but its transferred value remains sensitive to the calibration distribution and cost regime.

E.2 QUESTION-INDEPENDENT BUDGET MIXTURES

With one mean-cost constraint, an optimal extreme-point mixture uses at most two actions. For calibration sample Q and budget $B \in \{0.000125, 0.00025, 0.0005, 0.001, 0.0015\}$ dollars per question, the mixture weights w_a solve

$$\max_{w \geq 0} \quad \sum_a w_a \left(\overline{\mathrm{EM}}_Q(a) - 0.01 \mathbf{1}\{s_a = \textsc{Utility}\} \right), \tag{12}$$

$$\text{s.t.} \quad \sum_a w_a = 1, \qquad \sum_a w_a \overline{C}_Q(a) \leq B. \tag{13}$$

The implementation enumerates feasible pure actions and every action pair, including each cost-saturating pairwise mixture and its endpoints. Ties are resolved by lower calibration cost and then lexicographically lower weights in the fixed action order.

Mixture weights are fitted on calibration data and applied unchanged to held-out 2Wiki. Reported metrics are per-question expectations over randomized assignment, equivalent to Rao–Blackwellization over the assignment draw.

Table 64: Budget-constrained action mixtures calibrated on the MuSiQue + HotpotQA calibration sample ($n = 500$) and evaluated on the held-out 2Wiki sample ($n = 500$).

Calibration budget	Action mixture	Held-out EM [95% CI]	Held-out reader $/q [95% CI]
0.000125	0.035 Low/Utility + 0.965 Minimal/Utility	47.44 [43.02, 51.91]	0.0001 [0.0001, 0.0001]
0.000250	0.430 Low/Utility + 0.570 Minimal/Utility	52.42 [48.44, 56.51]	0.0003 [0.0003, 0.0003]
0.000500	0.586 Low/Top3 + 0.414 Low/Utility	61.83 [57.88, 65.79]	0.0005 [0.0005, 0.0005]
0.001000	0.479 Low/Full5 + 0.521 Medium/Full5	66.19 [62.12, 70.13]	0.0009 [0.0009, 0.0010]
0.001500	1.000 Medium/Full5	66.00 [61.89, 70.12]	0.0012 [0.0012, 0.0013]

Table 65: Budget-constrained action mixtures calibrated on the three-dataset calibration sample ($n = 750$) and evaluated on the held-out 2Wiki sample ($n = 500$).

Calibration budget	Action mixture	Held-out EM [95% CI]	Held-out reader $/q [95% CI]
0.000125	0.025 Low/Utility + 0.975 Minimal/Utility	47.31 [42.88, 51.79]	0.0001 [0.0001, 0.0001]
0.000250	0.387 Low/Utility + 0.613 Minimal/Utility	51.88 [47.87, 55.98]	0.0003 [0.0003, 0.0003]
0.000500	0.990 Low/Top2 + 0.010 Low/Top3	60.83 [56.66, 65.16]	0.0005 [0.0005, 0.0005]
0.001000	0.518 Low/Full5 + 0.482 Medium/Full5	66.21 [62.14, 70.15]	0.0009 [0.0009, 0.0010]
0.001500	1.000 Medium/Full5	66.00 [61.89, 70.12]	0.0012 [0.0012, 0.0013]

The budget constrains mean cost on the calibration population. Held-out expected cost is reported directly and can differ under distribution shift, so the table characterizes transferred mixtures rather than rematching them to the test distribution.

F COMPUTATIONAL COST

We separate total deployment cost into acquisition, evidence selection, and reader generation:

$$C_{\text{total}} = C_{\text{acq}} + C_{\text{select}} + C_{\text{reader}}. \tag{14}$$

Here C_{acq} covers retrieval and reranking, C_{select} covers the Utility model, and C_{reader} covers answer generation. Provider-billed reader dollars are the primary reader measure; Equation 8 supplies an uncached list-price sensitivity. Both include all reported completion tokens and exclude local acquisition and selection. Acquisition cost cancels between policies only when both deployments construct the same top-five pool.

Table 66: Reader input tokens on the held-out 2Wiki sample ($n = 500$).

Effort	Full5 input tokens	Utility input tokens	Reduction (%)
Minimal	893.5	292.4	67.3
Low	893.5	292.4	67.3
Medium	893.5	292.4	67.3

Table 67: Utility-minus-Full5 output-token differences on the held-out 2Wiki sample ($n = 500$).

Effort	Δ reported reasoning [95% CI]	Δ visible output [95% CI]
Minimal	+0.0 [0.0, 0.0]	-0.27 [-0.54, -0.03]
Low	+18.2 [9.0, 27.6]	+0.04 [-0.07, 0.15]
Medium	+33.4 [10.7, 57.8]	-0.13 [-0.23, -0.03]

Table 68: Reader cost on the held-out 2Wiki sample ($n = 500$), in dollars per 1,000 questions.

Effort	Full5 [95% CI]	Utility [95% CI]	Cost reduction (%)
Minimal	0.323 [0.314, 0.332]	0.129 [0.124, 0.135]	+60.0
Low	0.638 [0.615, 0.661]	0.492 [0.468, 0.519]	+22.8
Medium	1.214 [1.151, 1.275]	1.108 [1.039, 1.181]	+8.7

Table 69: Reader-cost sensitivity to billed and uncached list prices on the held-out 2Wiki sample ($n = 500$), in dollars per 1,000 questions.

Effort	Pricing	Full5 [95% CI]	Utility [95% CI]	Utility − Full5 [95% CI]
Minimal	Actual bill	0.323 [0.314, 0.332]	0.129 [0.124, 0.135]	-0.194 [-0.203, -0.185]
Minimal	Uncached list price	0.250 [0.243, 0.257]	0.099 [0.095, 0.103]	-0.151 [-0.157, -0.144]
Low	Actual bill	0.638 [0.615, 0.661]	0.492 [0.468, 0.519]	-0.145 [-0.169, -0.120]
Low	Uncached list price	0.492 [0.474, 0.511]	0.378 [0.359, 0.398]	-0.114 [-0.132, -0.094]
Medium	Actual bill	1.214 [1.151, 1.275]	1.108 [1.039, 1.181]	-0.106 [-0.166, -0.041]
Medium	Uncached list price	0.936 [0.887, 0.983]	0.852 [0.799, 0.908]	-0.084 [-0.130, -0.035]

Table 70: Warm batch-one latency by policy ($n = 90$); retrieval, acquisition, and model loading are excluded.

Arm	Mean total s	Median	P95	Selector s	Reader s
Full5	0.224	0.214	0.368	0.000	0.224
Top2	0.158	0.147	0.288	0.000	0.158
Top3	0.175	0.168	0.296	0.000	0.175
Utility	1.404	1.252	2.412	1.255	0.149

On held-out 2Wiki, UTILITY reduces mean prompt length from 893.488 to 292.392 tokens at every effort. Relative reader-bill savings are 60.0% at minimal, 22.8% at low, and 8.7% at medium

Table 71: Warm batch-one prompt length and peak allocated memory by policy ($n = 90$).

Arm	Mean prompt tok.	Peak GiB
Full5	818.9	6.12
Top2	386.1	6.02
Top3	525.4	6.04
Utility	322.9	9.39

effort. At low and medium effort, reported reasoning increases by 18.18 $[9.00, 27.59]$ and 33.41 $[10.68, 57.82]$ tokens, respectively. Input reduction therefore does not translate into a constant reduction in generation cost.

On the three-dataset sample at minimal effort, the reader-bill saving for UTILITY relative to FULL5 is 0.0001683587 dollars per question, with interval $[0.0001604691, 0.0001762750]$. Let g_w and g_b denote dollars per occupied GPU-hour in the warm batch-one and bulk-generation regimes. The warm profile uses 1.254850 selector seconds per question, giving selector cost $0.000348569 g_w$ and break-even condition $g_w < 0.4830$. Bulk selection uses 454.171 seconds for 500 questions, giving $0.000252317 g_b$ per question and $g_b < 0.6673$; whole-process wall time gives $0.000257911 g_b$ and $g_b < 0.6528$.

These thresholds are symbolic because no market GPU price is imposed. In the co-resident warm profile, selection reduces local reader time by 0.074881 seconds per question but adds 1.254850 selector seconds, for a net increase of 1.179968 seconds and a 6.258-fold total-time ratio relative to FULL5. The warm profile measures local co-resident execution, whereas the bulk profile measures throughput; neither is an API tail-latency estimate.

G ADDITIONAL METHODS

These comparisons evaluate question-dependent alternatives to the fixed action menu on development data. The following controller studies reuse response outcomes from the corresponding fixed-menu experiments; the subsequent evidence-editing and execution studies collect their own specified responses.

G.1 SEMANTIC JOINT CONTROLLER

The coupled controller does not resolve an improvement over a factorized main-effects model. The source-only representation encodes the question paired with each of five ordered passages using the BGE reranker backbone. It concatenates five normalized 1,024-dimensional first-token representations with classifier logits and 26 length and truncation features. After excluding overlapping source components, 554 questions from the three-dataset sample train a multioutput ridge model for the 12 action values; evaluation uses 500 2Wiki development questions. The coupled version adds policy-by-effort modes to the factorized model and optimizes $\mathrm{EM} - 100 C_{\mathrm{reader}} - 0.01 \mathbf{1}\{s = \mathrm{UTILITY}\}$.

The coupled controller obtains 62.20% EM, \$0.4060 per 1,000 questions, and 57.75 net-utility points. The factorized model obtains 61.80%, \$0.4046, and 57.34 points. Their paired utility difference is $+0.41$ $[-4.18, 4.06]$. A calibration-selected low-FULL5 action reaches 66.40% EM and 61.48 points, while a held-out cost-matched TOP2/TOP3 mixture exceeds the coupled controller by 1.00 point $[-4.22, 6.24]$. Because one source component contains 408 of the 500 evaluation questions, uncertainty is dominated by that component. The results favor the fixed-action calibration baseline for this representation and linear value model.

G.2 ANSWER-INFORMED ESCALATION

The answer-informed gate performs below both always-medium FULL5 and a source-only router, while its difference from a fixed agreement cascade is unresolved. The gate observes the minimal-FULL5 and minimal-UTILITY final answers plus fixed runtime features, then chooses minimal- or medium-FULL5. The source-only comparator routes before reader generation. Both use a `HistGradientBoostingRegressor` with squared-error loss, learning rate 0.05, 100 iterations,

7 leaf nodes, minimum leaf size 50, L_2 penalty 1, no early stopping, and seed 20260924. Training excludes every source component represented in the 750-question evaluation sample. The agreement cascade escalates exactly when the two minimal-effort answers differ.

On the three-dataset sample, the answer gate reaches 52.67% EM, costs \$1.2527 per 1,000 questions, escalates 57.07% of questions, and has net utility 0.3914, where utility is EM minus 100 times reader dollars per question. The source-only router reaches 55.60% EM, \$1.0913 per 1,000, a 73.07% escalation rate, and utility 0.4469. Always-medium FULL5 reaches 59.20% EM, \$1.3519 per 1,000, and utility 0.4568. The answer gate is lower than these comparators by 0.0555 $[0.0399, 0.0722]$ and 0.0654 $[0.0467, 0.0863]$, respectively. The agreement cascade reaches 52.27% EM and utility 0.4001; the gate-minus-cascade difference is -0.0087 $[-0.0253, 0.0087]$. These results provide no evidence that the additional answer probe improves net utility for this escalation rule.

G.3 SET-VALUE LEARNERS

None of the tested set-value learners improves tuning utility relative to FULL5. A learner using the final four BGE blocks applies the same capacity to three targets: joint downstream value, additive singleton value, and annotated-support coverage. On a fixed 100-question MuSiQue tuning sample, mean target utility relative to FULL5 is -0.1883, -0.3833, and -0.0153, respectively. A listwise UtilityQwen adaptation selects FULL5 or a fixed pair for all 100 questions under the three targets and yields no positive tuning utility over FULL5. A separate 32-question training-set fit recovers nearly all available headroom, indicating interpolation capacity without evidence of transfer. A cross-fitted source-policy gate also satisfies none of nine reader-by-dataset evaluation criteria. Across these architectures, the development evidence supports reporting the fixed-action menu rather than a learned per-question selector.

H LOCAL EVIDENCE EDITING AND MULTISTAGE EXECUTION

We next study two extensions to the fixed policy–effort menu on a balanced panel of 360 questions, with 120 questions from each benchmark. These analyses reuse an exposed development panel and its 149 exact-source components. They test whether additional evidence actions or multistage inference provide useful headroom; they are not independent confirmation experiments.

H.1 ANSWER-PRIVILEGED EVIDENCE REPAIR

The evidence-repair screen expands the original menu of Full5, Top2, Top3, and Utility contexts. For each question it adds five contexts that remove one ranked Full5 passage and contexts that restore one omitted Full5 passage to Utility. Exact duplicate contexts share a response within a draw. At each effort separately, a discovery draw uses the reference answer to select the best expanded-menu action and the best original-menu action under the fixed net-utility objective. A fresh repeat draw then evaluates both selected contexts, Full5, a fixed Top4 context, and a prespecified random local edit. Selection never moves across effort levels. Confidence intervals use 4,000 dataset-stratified source-component bootstrap samples and average the three efforts within each question.

The expanded menu selected a context absent from the original menu in 305 of 1,080 question–effort cells (28.2%). On the fresh repeat, its selected context exceeded the original-menu selection by 2.50 EM points $[0.76, 5.42]$ and 2.43 net-utility points $[0.77, 5.34]$. The corresponding discovery-draw EM difference was 5.09 points, so the fresh result also shows the expected reduction from selection on a single draw. Table 72 gives the repeat controls. Because reference-answer outcomes choose both actions, this is a repeat-validated oracle opportunity estimate rather than a deployable repair policy.

H.2 SOURCE-ONLY EDIT POLICIES

We test whether the local-edit opportunity can be recovered without answers. Each physical context is represented by its retained, removed, and added sources; source rank and length; cached BGE relevance features; question-token coverage; and lexical redundancy. The models predict Full5-relative accuracy and reader-cost changes and score all available contexts with a Full5 zero-change option. Dataset identity, question identifiers, reference answers, support annotations, and reader

Table 72: Fresh-repeat results for answer-privileged evidence-menu selection. The first five rows are Draw-B means; the paired row reports 95% source-component-bootstrap intervals. Reader cost is standardized list-price dollars per 1,000 question–effort evaluations, and net utility is reported in percentage points for $\mathrm{EM}_{\mathrm{fraction}} - 100C_{\mathrm{reader}}$. Discovery and selection costs are excluded.

Draw-B action	EM	List $/1k	Net utility
Expanded-menu winner	56.8	0.397	52.8
Old-menu winner	54.3	0.390	50.4
Full5	52.2	0.575	46.5
Fixed Top4	50.5	0.545	45.0
Random local edit	46.7	0.545	41.2
Expanded − old (paired)	+2.5 [0.8, 5.4]	+0.007 [-0.017, 0.016]	+2.4 [0.8, 5.3]

outputs are excluded from the features. Five-fold cross-fitting keeps each exact-source component in one fold and balances question–effort cells and datasets in training.

The linear screen uses a fixed-penalty multioutput ridge regressor. Its effort-aware version adds fixed effort contrasts and their interactions with every base feature; the otherwise identical effort-blind version omits those terms. The nonlinear sensitivity uses prespecified Extra Trees regressors with 200 trees, depth four, and minimum leaf size 20 under the same folds, targets, and controls. Neither model class improves on both an identical effort-blind model and the training-fold best fixed action for each effort. For ridge, the effort-aware minus effort-blind net-utility difference is -0.91 $[-2.94, 1.00]$, and the difference from the fixed-by-effort control is -1.77 $[-5.03, -0.50]$. For Extra Trees, the corresponding differences are $+0.40$ $[-1.50, 3.17]$ and $+0.20$ $[-1.71, 2.93]$. Table 73 reports absolute out-of-fold results. Thus the available source-only features and these two fixed learners do not convert the oracle edit opportunity into a policy gain.

Table 73: Out-of-fold source-only evidence-edit policies on the Stage-A development outcomes. Every prediction is held out by exact-source component. Costs are standardized list-price reader dollars per 1,000 question–effort evaluations; net utility uses the same definition as Table 72.

Model	Policy	EM	List $/1k	Net utility
Ridge	Effort-aware	47.9	0.506	42.8
Ridge	Effort-blind	48.8	0.508	43.7
Ridge	Fold-best by effort	49.7	0.514	44.6
Extra Trees	Effort-aware	50.3	0.550	44.8
Extra Trees	Effort-blind	49.8	0.543	44.4
Extra Trees	Fold-best by effort	49.7	0.514	44.6
Fixed	Utility	50.2	0.414	46.0
Fixed	Full5	51.0	0.574	45.3

H.3 Fixed multistage evidence operations

The execution pilot holds the ordered Full5 evidence constant and compares one-call direct reading with three four-call strategies. Sequential execution first produces a question-only dependency plan, resolves root facts, and then resolves dependent facts with the root state available. Flat decomposition shares the same plan but produces two independent complete fact maps. Generic repeated inference produces three source-grounded maps without a supplied plan. Each staged strategy then makes one final answer call at minimal or medium effort; direct reading is evaluated at minimal, low, and medium effort. Intermediate citations must quote an input passage exactly, which verifies provenance but not entailment. Standalone costs assign every deployment its own planning or mapping calls.

At minimal final effort, sequential execution exceeds generic repeated inference by 3.33 EM points $[-3.27, 5.77]$ and flat decomposition by 1.94 points $[-3.39, 6.67]$; neither interval resolves a gain. It exceeds direct minimal by 9.72 points $[2.51, 14.02]$, but falls below direct low by 16.11 points $[-21.37, -11.22]$ and direct medium by 18.06 points $[-24.77, -11.94]$. The observed direct-medium point has higher EM and lower list and billed cost than every staged arm, so none of these multistage strategies adds an accuracy–cost frontier point on this panel.

The prespecified adaptive policy chooses sequential execution only for nontrivial chain plans and otherwise chooses flat decomposition. It uses minimal final effort only when the selected operator returns valid, fully resolved state, with agreement also required between flat maps; all other cases use medium effort. The policy selects flat-medium for 88 questions, flat-minimal for 40, sequential-medium for 149, and sequential-minimal for 83. It reaches 59.17% EM at \$4.470 billed per 1,000 answers. Relative to the identical operator rule with medium effort throughout, it saves \$0.169 per 1,000 but loses 2.22 EM points $[-3.49, -0.82]$. Its difference from direct low is -1.11 points $[-4.22, 1.68]$ while costing \$3.811 more per 1,000. The fixed adaptive rule therefore does not motivate independent confirmation.

Table 74: Fixed evidence operation pilot and prespecified adaptive policy on the 360-question development panel. EM includes 95% source-component-bootstrap intervals. Costs are standalone deployment dollars per 1,000 answers and include every planning, fact, and final-answer call; local retrieval and parsing are excluded.

Method	Final effort	Calls	EM [95% CI]	List \$/1k	Billed \$/1k
Direct	Minimal	1	34.4 [30.3, 48.9]	0.221	0.280
Direct	Low	1	60.3 [56.4, 71.5]	0.510	0.658
Direct	Medium	1	62.2 [58.4, 73.0]	1.004	1.294
Flat decomposition	Minimal	4	42.2 [37.3, 55.3]	3.353	4.256
Flat decomposition	Medium	4	60.3 [56.5, 71.2]	4.006	5.100
Sequential execution	Minimal	4	44.2 [40.2, 56.2]	2.613	3.388
Sequential execution	Medium	4	61.4 [57.2, 72.1]	3.284	4.256
Generic repeated maps	Minimal	4	40.8 [36.7, 56.4]	6.474	8.133
Generic repeated maps	Medium	4	60.3 [56.2, 72.3]	7.135	8.987
Adaptive flat/sequential	Adaptive	4	59.2 [54.9, 70.3]	3.477	4.470

H.4 State-conditioned evidence acquisition

State-conditioned queries improve annotated-support availability and medium-effort answer accuracy relative to fixed size controls on the 360-question development panel. Their accuracy advantage over the strongest direct Full5 reader is smaller and unresolved, so this experiment establishes an acquisition effect without establishing a new accuracy–cost frontier.

The acquisition procedure reuses the question-only dependency graph and root extraction defined above. It constructs each context as follows:

1. A low-effort planning call maps the question to at most four factual nodes and their directed dependencies. A second low-effort call extracts every root fact from the original ordered Full5 context. Each extracted fact must cite an exact source substring; this check establishes provenance, not entailment.
2. A dependent node is eligible when its parent set is nonempty and all its parents are graph roots with resolved extracted answers. Its retrieval query concatenates the dependent-node question with those parent answers in graph order.
3. The query searches the fixed dataset corpus using BM25 with $k_1 = 1.5$, $b = 0.75$, and a deterministic top-50 ranking. The first document that is novel relative to Full5 and earlier additions is selected. At most one document is added per eligible node and at most three per question. Invalid plans and unresolved roots remain in the evaluation with no addition.
4. Selected documents are appended to Full5 in node order. The count-matched rank control instead appends the same number of novel documents from BGE ranks six onward; Full10 uses the first ten BGE-ranked documents. All final readers receive only the passages and original question.

Planning and root extraction use low effort with a 2,000-token cap. API final readers use `openai/gpt-5-mini-2025-08-07`, a 4,000-token cap, and minimal, low, or medium effort.

The grounded and rank-matched paths are charged for planning, root extraction, and final reading. Full10 and direct Full5 require one final-reader call. The local BM25 acquisition took 7.489 seconds for the panel and used neither an API nor a GPU.

Support annotations were opened only after acquisition contexts were fixed. The grounded queries added 283 documents across 275 questions. They increased the number of questions containing all annotated support pages from 166 to 247; adding the same number of BGE-ranked documents increased it by zero. Child-only and original-question-plus-child queries provide controls for the query content while preserving the admitted dependency nodes and per-question addition counts.

Table 75: Annotated-support availability under state-conditioned acquisition and fixed query controls. Entries in the five rightmost columns count questions containing every annotated support page. All acquisition variants append the same number of documents for each question; the rank control takes the next documents in the fixed BGE order.

Dataset	n	Added docs	Full5	Rank control	Child	Question + child	Grounded
2Wiki	120	100	44	44	44	45	88
HotpotQA	120	80	101	101	101	101	112
MuSiQue holdout	120	103	21	21	22	22	47
Pooled	360	283	166	166	167	168	247

The primary API comparison is grounded acquisition at medium final effort minus Full10 at medium effort. It yields $+5.00$ EM points $[1.00, 10.31]$ and $+4.34$ F1 points $[0.89, 8.30]$. The grounded-minus-Full10 EM differences are $+9.17$ $[2.14, 16.47]$ at minimal effort and $+1.11$ $[-2.59, 6.84]$ at low effort. Thus the benefit of the acquired evidence depends on reader effort even within this acquisition family.

Table 76: API reader accuracy and standalone cost on the 360-question three-dataset panel. The grounded and rank-matched acquisition paths charge the shared planning and root-extraction calls plus the final reader; Full10 and direct Full5 charge only the final reader. Local retrieval is excluded. EM intervals use the 149 source components.

Method	Final effort	Calls	EM [95% CI]	List $/1k	Billed $/1k
Grounded acquisition	Minimal	3	44.4 [40.0, 56.7]	1.738	2.257
Grounded acquisition	Low	3	62.5 [58.6, 73.1]	1.989	2.577
Grounded acquisition	Medium	3	64.7 [60.7, 74.6]	2.439	3.164
Rank-matched acquisition	Minimal	3	33.6 [29.9, 48.0]	1.744	2.267
Rank-matched acquisition	Low	3	60.0 [55.7, 68.3]	2.021	2.619
Rank-matched acquisition	Medium	3	59.4 [55.3, 68.3]	2.493	3.231
Full10	Minimal	1	35.3 [31.5, 47.0]	0.395	0.494
Full10	Low	1	61.4 [57.2, 70.7]	0.678	0.857
Full10	Medium	1	59.7 [56.3, 70.2]	1.179	1.511
Direct Full5	Minimal	1	34.4 [30.3, 48.9]	0.221	0.280
Direct Full5	Low	1	60.3 [56.4, 71.5]	0.510	0.658
Direct Full5	Medium	1	62.2 [58.4, 73.0]	1.004	1.294

The medium-effort grounded context reaches 64.72% EM at a provider-billed cost of $3.164 per 1,000 answers. Direct Full5 at medium effort reaches 62.22% at $1.294 per 1,000; their paired difference is $+2.50$ points $[-2.06, 5.87]$. The comparison supports an acquisition benefit over Full10, while the evidence does not resolve broader superiority over the direct Full5 effort points.

An exploratory acquisition ablation tests whether scanning past a Full5 duplicate to force a novel document explains the remaining errors. The rule was fixed before scoring: retain a grounded addition only when its unfiltered BM25 rank is one, with no alternative rank cutoff. It retains 121 additions across 116 questions; 244 questions receive Full5 only. Of the 360 resulting medium-effort requests, 359 exactly match an existing Direct or grounded request and one previously unevaluated context was read after the complete context roster was frozen. The rank-1-only context reaches 63.61% EM. Its paired differences are $+1.39$ points $[-2.22, 3.42]$ relative to Direct Full5, -1.11 $[-4.24, 1.71]$ relative to the parent grounded context, and $+3.89$ $[-1.04, 8.12]$ relative to Full10. Thus removing forced lower-ranked novelty does not yield a resolved improvement over any of these controls.

Table 77: Paired API-reader contrasts for state-conditioned acquisition. Accuracy is in percentage points and cost is provider-billed dollars per 1,000 answers. Intervals are paired source-component-bootstrap intervals.

Effort	Contrast	ΔEM [95% CI]	ΔF1 [95% CI]	Δ billed $/1k
Minimal	Grounded − Full10	+9.17 [2.14, 16.47]	+10.78 [3.54, 16.36]	+1.763 [1.722, 1.893]
Minimal	Grounded − rank control	+10.83 [2.47, 17.27]	+13.07 [3.51, 19.12]	-0.010 [-0.019, -0.005]
Low	Grounded − Full10	+1.11 [-2.59, 6.84]	+1.67 [-1.29, 4.63]	+1.720 [1.674, 1.841]
Low	Grounded − rank control	+2.50 [0.04, 8.77]	+3.42 [0.71, 6.96]	-0.042 [-0.094, -0.025]
Medium	Grounded − Full10 (primary)	+5.00 [1.00, 10.31]	+4.34 [0.89, 8.30]	+1.653 [1.592, 1.790]
Medium	Grounded − rank control	+5.28 [2.97, 9.93]	+4.89 [2.01, 7.27]	-0.067 [-0.130, 0.022]
Medium	Grounded − direct Full5/low	+4.44 [-0.62, 8.39]	+3.67 [-1.02, 5.88]	+2.506 [2.414, 2.656]
Medium	Grounded − direct Full5/medium	+2.50 [-2.06, 5.87]	+2.65 [-0.01, 4.23]	+1.870 [1.806, 2.010]

Table 78: Exploratory no-forced-novelty acquisition ablation at medium effort. The rank-1-only and parent grounded paths charge planning, root extraction, and final reading for every question; Direct Full5 and Full10 charge one final read. Accuracy intervals use the 149 original source components on the exposed 360-question development panel.

Context	Calls	EM [95% CI]	F1 [95% CI]	List $/1k	Billed $/1k
Rank-1-only acquisition	3	63.61 [59.99, 73.09]	76.18 [73.27, 84.22]	2.461	3.188
Parent grounded acquisition	3	64.72 [60.93, 74.98]	77.30 [74.31, 85.28]	2.439	3.164
Direct Full5	1	62.22 [58.44, 73.02]	74.65 [71.45, 83.42]	1.004	1.294
Full10	1	59.72 [56.20, 70.03]	72.96 [70.28, 81.25]	1.179	1.511

To test the role of the grounded parent values, a separate medium-effort ablation holds the dependent nodes, addition count, corpus, retrieval rule, and planning and extraction charges fixed. The primary control queries each child using the original question plus the child question; a secondary control uses the child question alone. Relative to the primary question-aware control, adding the resolved parent values improves pooled EM by $+5.28$ points $[0.02, 7.85]$ and F1 by $+4.98$ $[0.74, 6.23]$. This contrast identifies the effect of these specified query constructions and does not establish a general entity-binding mechanism.

Table 79: Medium-effort grounded-query ablation. Every path uses the same admitted dependency nodes, number of added documents, corpus, and planning and root-extraction charges. Dataset rows are descriptive; the pooled grounded-minus-question-plus-child contrast is primary.

Scope	Query contrast	ΔEM [95% CI]	ΔF1 [95% CI]	Δ billed $/1k
Pooled	Grounded − question + child	+5.28 [0.02, 7.85]	+4.98 [0.74, 6.23]	-0.094 [-0.148, -0.007]
2Wiki	Grounded − question + child	+8.33 [-3.85, 13.80]	+9.32 [-0.37, 10.88]	-0.210 [-0.252, 0.038]
HotpotQA	Grounded − question + child	+3.33 [0.81, 6.61]	+3.47 [1.00, 6.46]	-0.027 [-0.104, 0.040]
MuSiQue holdout	Grounded − question + child	+4.17 [-3.03, 9.80]	+2.16 [-5.10, 3.73]	-0.045 [-0.197, -0.008]
Pooled	Grounded − child only	+5.56 [1.96, 8.59]	+4.85 [1.68, 6.13]	-0.091 [-0.155, -0.009]

Auxiliary-state factorial. To distinguish evidence acquisition from passing extracted state to the final reader, a fixed 2×2 factorial crosses original versus grounded evidence with an empty versus populated root-state object. Every final prompt includes the same question-only plan, and final-reader effort is minimal or medium; planning and root extraction use fixed low effort. The primary bridge-context contrast at medium effort is $+0.83$ EM points $[-1.87, 2.41]$ for populated versus empty state. The state-by-acquisition interaction is -1.67 points $[-7.77, 0.29]$, and the root-plus-bridge cell does not satisfy the prespecified comparisons with both the existing grounded reader and direct medium reader. The medium-effort contrast therefore does not establish an additional accuracy gain from inserting root state into the final prompt.

Table 80: State-by-acquisition factorial. Costs follow each cell's natural standalone path: original evidence without root state uses planning and final reading; every other cell also charges root extraction.

Effort	Evidence	Final state	Calls	EM [95% CI]	List $/1k	Billed $/1k
Minimal	Original Full5	No state	2	35.00 [31.00, 43.17]	0.750	0.974
Minimal	Original Full5	Root state	3	42.78 [39.04, 52.71]	1.768	2.287
Minimal	Grounded acquisition	No state	3	44.44 [38.87, 52.01]	1.770	2.291
Minimal	Grounded acquisition	Root state	3	49.72 [46.06, 59.50]	1.791	2.303
Medium	Original Full5	No state	2	59.17 [55.50, 69.61]	1.529	1.982
Medium	Original Full5	Root state	3	61.67 [57.71, 73.58]	2.479	3.206
Medium	Grounded acquisition	No state	3	64.72 [60.36, 76.35]	2.483	3.211
Medium	Grounded acquisition	Root state	3	65.56 [61.56, 75.72]	2.447	3.152

At minimal effort, the secondary grounded-context state contrast is $+5.28$ EM points $[1.69, 14.36]$. A post-completion descriptive comparison finds larger root-state effects at minimal than at medium effort: this effort interaction is $+4.44$ points $[1.15, 14.04]$ with grounded evidence and $+5.28$ $[0.71, 11.29]$ with original evidence. Because the pattern appears with both evidence contexts, it suggests an effort-dependent state effect rather than an acquisition-specific interaction. These comparisons were not part of the primary confirmation criteria.

The acquired-context comparison charges the same planning and root-extraction calls whether the final reader receives or does not receive the extracted answers. At minimal effort, the mean billed cost changes from \$2.2905 to \$2.3031 per 1,000 questions; the paired cost difference includes zero. At medium effort, reuse changes the corresponding cost from \$3.2109 to \$3.1525, a 1.8% reduction, and reduces reported reasoning by 29.7 tokens per question. The paired reductions hold for both provider-billed and standardized uncached-list cost. Table 81 gives the resource contrasts and intervals. Since helper outputs and charges are identical within each pair, the absolute token and API-cost differences arise in the final read. These comparisons do not establish accuracy noninferiority or a cost advantage over one-call direct reading.

Table 81: Paired root-answer minus withheld-answer resource contrasts for the grounded acquired context. Cells are mean differences with 95% component-bootstrap intervals. Every value covers the complete three-call API path: planning and root extraction are common to both conditions and cancel questionwise, leaving the final reader's response to the already computed root answers. Uncached-list cost uses standardized list pricing; billed cost uses provider charges. Local retrieval is excluded.

Effort	Input tok.	Reported reasoning tok.	Uncached-list API \$/1k	Billed API \$/1k
Minimal	+92.0 [87.5, 97.9]	+0.0 [0.0, 0.0]	+0.022 [0.004, 0.025]	+0.013 [-0.013, 0.031]
Medium	+92.0 [87.5, 97.9]	-29.7 [-67.7, -18.9]	-0.036 [-0.113, -0.015]	-0.058 [-0.156, -0.025]

Post-acquisition dependent resolution. A further fixed study on the same development panel adds one low-effort call that resolves every non-root plan node in order from the grounded 5–7-passage context and validated root bindings. Resolved values require exact source quotations; malformed and unresolved outputs remain in the paired analysis. The final reader receives either root state alone or root plus dependent state at minimal and medium effort. At minimal effort, dependent state changes EM by $+2.50$ points $[-0.10, 11.33]$, which does not establish the prespecified improvement; at medium effort, the difference is -1.39 $[-4.24, 1.69]$. The secondary minimal-minus-medium interaction is $+3.89$ points $[1.02, 11.79]$, indicating that the observed state effect is larger at minimal effort. The four-call minimal path costs \$3.531 billed per 1,000 answers, versus \$2.303 for the three-call root-state path, and does not add an accuracy–cost frontier point relative to the completed bridge-direct medium reader.

In a separate descriptive analysis, augmentation increases the number of questions with at least one provenance-valid resolved dependent binding from 129 to 201; 71 questions cite an added source in such a binding. Plan depths 1, 2, and 3 or greater contain 52, 285, and 23 questions, respectively, so moderation by plan depth remains exploratory. These source-defined records do not establish the truth of a binding or causal mediation.

Table 82: Post-acquisition dependent-state experiment on the 360-question development panel. Root-state paths charge planning, root extraction, and final reading; dependent-state paths also charge one dependent-resolution call. Exact historical replays retain their measured standalone costs. Local retrieval is excluded.

Effort	Final state	Calls	EM [95% CI]	F1 [95% CI]	List \$/1k	Billed \$/1k
Minimal	Root state	3	49.72 [46.06, 59.50]	63.61 [60.47, 72.61]	1.791	2.303
Minimal	Root + dependent state	4	52.22 [47.69, 66.89]	66.11 [62.61, 78.18]	2.748	3.531
Medium	Root state	3	65.56 [61.56, 75.72]	78.26 [75.18, 86.71]	2.447	3.152
Medium	Root + dependent state	4	64.17 [60.15, 75.59]	77.12 [74.04, 86.00]	3.374	4.335

Qwen reader transfer and confirmation. We transfer the four fixed context families to Qwen2.5-3B-Instruct (Qwen Team, 2024), revision `aa8e72537993ba99e69dfaafa59ed015`

b17504d1. The local reader uses FP16, greedy decoding, a 64-token output cap, and batch size four. Qwen does not construct the plan, root bindings, or acquired documents. On the 360-question development panel, grounded acquisition exceeds Full10 by $+8.06$ EM points $[3.53, 13.47]$ and direct Full5 by $+6.11$ $[3.48, 9.64]$. The three dataset-specific grounded-minus-Full10 estimates are positive, although the MuSiQue interval includes zero.

Table 83: Fixed-context transfer to Qwen2.5-3B-Instruct. Timings divide each batch wall time among its responses and are descriptive rather than causal latency estimates. Upstream planning and acquisition are not included.

Context	EM [95% CI]	F1 [95% CI]	Batched GPU s/response
Grounded acquisition	32.50 [26.06, 38.79]	42.70 [36.07, 47.02]	0.337 [0.325, 0.353]
Rank-matched control	26.39 [19.65, 31.46]	36.08 [29.08, 39.92]	0.337 [0.327, 0.354]
Full10	24.44 [17.46, 30.58]	33.15 [26.66, 37.92]	0.345 [0.334, 0.361]
Direct Full5	26.39 [20.33, 31.31]	35.48 [29.21, 39.40]	0.339 [0.326, 0.351]

Table 84: Paired Qwen fixed-context contrasts. The pooled grounded-minus-Full10 comparison is primary; dataset-specific rows are descriptive. Intervals resample source components within dataset.

Scope	Contrast	ΔEM [95% CI]	ΔF1 [95% CI]
Pooled	Grounded − Full10	+8.06 [3.53, 13.47]	+9.55 [4.14, 14.52]
Pooled	Grounded − rank control	+6.11 [3.43, 10.60]	+6.62 [3.80, 10.50]
Pooled	Grounded − direct Full5	+6.11 [3.48, 9.64]	+7.21 [4.37, 10.67]
2Wiki	Grounded − Full10	+7.50 [3.57, 18.75]	+9.34 [2.96, 20.94]
HotpotQA	Grounded − Full10	+8.33 [1.65, 14.96]	+9.91 [2.88, 16.89]
MuSiQue	Grounded − Full10	+8.33 [-3.70, 18.18]	+9.39 [-4.67, 16.20]

The development Qwen reader required 425.411 seconds of inference wall time summed over sequential batches after a 5.543-second model load. Per-response times in Table 83 divide each batch time among its responses and should not be interpreted as causal latency differences.

We then evaluate the four acquisition strategies on a separate 450-question panel with no initial-source component shared with development. This transfer also changes the starting retriever: the initial five passages come from cached HippoRAG 2 retrieval, and the widening and count-matched controls use original-question BM25 rankings, rather than the BGE rankings used in development. Direct Full5 reaches 34.89% EM, grounded acquisition 33.56%, the count-matched control 32.22%, and Full10 31.56%. The primary grounded-minus-Full5 difference is -1.33 points $[-3.99, 1.12]$. Grounded acquisition differs from Full10 by $+2.00$ points $[-1.12, 5.10]$ and from the count-matched control by $+1.33$ $[-1.33, 3.83]$. Neither prespecified accuracy criterion is met. A sensitivity that joins questions sharing any source across all evaluated contexts also leaves the lower endpoints below zero. The development gain therefore does not establish improvement over direct Full5 under this joint question and retrieval shift.

Table 85: Component-disjoint Qwen confirmation on 450 questions. GPU times divide batch wall time among responses and exclude upstream planning and acquisition.

Context	EM [95% CI]	F1 [95% CI]	Batched GPU s/response
Direct Full5	34.89 [30.48, 39.43]	44.07 [39.89, 48.24]	0.309 [0.300, 0.317]
Grounded acquisition	33.56 [29.25, 37.95]	44.64 [40.56, 48.74]	0.305 [0.297, 0.315]
Count-matched control	32.22 [28.04, 36.64]	41.68 [37.66, 45.83]	0.310 [0.301, 0.319]
Full10	31.56 [27.26, 35.96]	40.65 [36.59, 44.87]	0.310 [0.302, 0.319]

As a secondary transfer analysis, we apply the single rank-1-only acquisition rule fixed on the development panel to both Qwen panels. Exact request matching reuses 808 of 810 responses; only two strict-subset contexts require a new generation. On development, the rule differs from Full5 by $+5.56$ EM points $[3.31, 9.42]$ and from the parent grounded context by -0.56 $[-1.51, 0.83]$. On confirmation, the corresponding differences are $+0.44$ $[-1.08, 2.00]$ and $+1.78$ $[-0.40, 3.97]$. Its

Table 86: Paired Qwen confirmation contrasts. The primary analysis uses 421 global original-source components, represented by 422 blocks after dataset stratification. The sensitivity joins questions sharing any source across the four evaluated contexts, yielding 244 components. The grounded-minus-Full5 contrast is primary.

Component grouping	Contrast	ΔEM [95% CI]	ΔF1 [95% CI]
Original	Grounded − Full5	-1.33 [-3.99, 1.12]	+0.57 [-2.09, 3.16]
Original	Grounded − Full10	+2.00 [-1.12, 5.10]	+3.99 [0.56, 7.31]
Original	Grounded − count control	+1.33 [-1.33, 3.83]	+2.96 [0.35, 5.45]
Acquired union	Grounded − Full5	-1.33 [-6.86, 0.91]	+0.57 [-4.72, 2.94]
Acquired union	Grounded − Full10	+2.00 [-4.02, 4.60]	+3.99 [-1.85, 6.83]
Acquired union	Grounded − count control	+1.33 [-4.36, 3.47]	+2.96 [-2.12, 5.05]

advantage over Full10 is positive on both panels, but it does not consistently improve on Full5 or the parent grounded context. The original four-arm confirmation remains the primary evaluation.

Table 87: Secondary transfer of the fixed rank-1-only acquisition rule to the Qwen reader. The development panel is exposed; the confirmation panel is component-disjoint. Panels are analyzed separately, and the original four-arm confirmation remains primary.

Panel	Contrast	ΔEM [95% CI]	ΔF1 [95% CI]
Development	Rank-1 only − Full5	+5.56 [3.31, 9.42]	+6.29 [3.85, 9.78]
Development	Rank-1 only − Parent grounded	-0.56 [-1.51, 0.83]	-0.93 [-2.01, 0.23]
Development	Rank-1 only − Full10	+7.50 [3.37, 13.22]	+8.62 [3.54, 13.64]
Confirmation	Rank-1 only − Full5	+0.44 [-1.08, 2.00]	+1.07 [-0.41, 2.67]
Confirmation	Rank-1 only − Parent grounded	+1.78 [-0.40, 3.97]	+0.49 [-1.61, 2.63]
Confirmation	Rank-1 only − Full10	+3.78 [0.86, 6.71]	+4.49 [1.32, 7.74]

Across these studies, state-conditioned acquisition improves annotated-support coverage and beats fixed size controls on the exposed development panel. Its advantage over direct Full5 does not reproduce on the component-disjoint Qwen confirmation, and neither auxiliary state nor the fixed rank-1 rule establishes a broader accuracy–cost improvement.