COVERAGE BEFORE COMPRESSION: ONLINE CALL SCHEDULING FOR REPEATED LONG-CONTEXT REVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated long-context review amortizes candidate proposal while incurring target-model verification cost on every case, making call scheduling a first-order online design variable. We introduce policy-isolated verification, a paired estimand that holds the output, span partition, candidate order, target model, and call cap fixed, together with EvidenceCall, a coalition-state policy organized by coverage before compression. EvidenceCall tests necessity and direction, continues from verified coalitions, and delays redundancy removal until evidence coverage is secured. At an identical $B{=}50$ cap over 3,600 RULER-128K, MuSiQue, and QuALITY examples, it improves decisive-span P@5 by $+9.3$ over Greedy Minimal-Set, with 80.6% of the margin retained under a literal ranking freeze. Across three frozen proposers, it also leads outside-family MCTS-Subset by $+1.2$ to $+1.8$ P@5, and every paired-bootstrap lower bound is positive. The identified gain transfers to decisions: reviewer yield rises by 31.7% in a 32-reviewer mechanism-isolation crossover and by 28.1% in a 24-developer code deployment, while a disjoint 36-reviewer citation deployment improves accepted findings from 3.88 to 4.28 per hour over MCTS-Subset. Blind full-context adjudication, external proposer supervision, three held-out benchmarks, and two additional target families preserve the effect. The results establish coverage-before-compression scheduling as a consequential systems principle for repeated review.

1 INTRODUCTION

A reviewer-facing long-context system makes two coupled decisions: which spans to propose and how to verify them. The costs separate under repeated use. A proposer can be trained once and cached across thousands of frozen outputs, whereas every intervention consumes target-model capacity and adds latency to every audit. The consequential online question is therefore a policy question: given a fixed answer, span partition, candidate order, target model, and call cap, which adaptive schedule exposes the most decisive evidence?

We make that policy effect an identifiable experimental object. Our paired policy-isolation estimand evaluates competing schedules on the same frozen audit units and charges every target-model query to the same budget. Its companion measure, policy-isolation retention, reports how much of an end-to-end margin survives when proposal order is held literally constant. This design separates the value of ranking candidates from the value of acting on observed intervention outcomes, a distinction hidden by aggregate attribution accuracy.

EvidenceCall realizes the controlled object through a coverage-before-compression policy. It first establishes whether a candidate is necessary, resolves whether its influence is supportive or misleading, and then reconditions the remaining search on the verified coalition. Only after this discovery phase does it spend calls removing redundant spans. The state transition is the technical distinction: subsequent calls depend on both the content and direction of earlier flips, while Greedy Minimal-Set follows a deletion sweep and MCTS-Subset optimizes a generic flip-and-size rollout reward.

The evaluation pairs this mechanism with a broad identification design. The controlled bed contains 3,600 QA examples at $B{=}50$, and the code evaluation contains all 500 SWE-bench Verified bugs. Ten-seed paired tests, a 692-example blind full-context audit, three frozen proposer families, and transfer from Llama-3.1-70B-Instruct to Qwen2.5-72B-Instruct and Claude 3.5 Sonnet isolate ranking, label, operator, policy-family, and target effects (Grattafiori et al., 2024; Qwen et al., 2024; Anthropic,

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.

2024). Three disjoint reviewer studies then connect recovered evidence to decisions: a 32-reviewer mechanism-isolation crossover, a 24-developer external code deployment, and a 36-reviewer external citation deployment.

The evidence identifies a stable and practically important policy effect. EvidenceCall gains $+9.3$ pooled-QA P@5 over Greedy, retains a 7.5-point margin under an exact ranking freeze, and leads MCTS-Subset across original, externally supervised, and MonoT5-3B candidate orders. The controlled and external code studies raise reviewer throughput by 31.7% and 28.1%; the citation study adds $+0.40$ accepted findings per hour over MCTS-Subset. Thus reviewer systems can gain throughput by changing the online policy while preserving both the proposer and the target-call cap.

The contributions are:

1. a paired policy-isolation estimand and retention measure that identify post-proposal scheduling value under exact resource and ranking control;
2. EvidenceCall's coverage-before-compression state machine, which converts observed necessity, direction, and coalition outcomes into the next charged intervention;
3. a cross-family identification design spanning literal ranking freezes, independent full-context labels, three proposer families, and a matched-call MCTS policy; and
4. decision-level validation in controlled QA/code review and external code/citation workflows, with transfer across benchmarks, supervision sources, and target models.

The evidence path moves from matched-call recovery (Table 2) through blind labels (Table 3) and fixed-ranking identification (Tables 4–5) to reviewer decisions (Tables 6–7).

2 Related Work

Removal, Shapley, and perturbation methods assign importance to features or input regions (Shapley, 1953; Lundberg & Lee, 2017; Fong & Vedaldi, 2017; Covert et al., 2021), while infidelity and faithfulness criteria evaluate the resulting explanations (Yeh et al., 2019; Jacovi & Goldberg, 2020). Rationale, attributed-QA, citation, and context-attribution methods select or predict supporting text (Lei et al., 2016; Chen et al., 2018; DeYoung et al., 2020; Bohnet et al., 2022; Gao et al., 2023; Radevski et al., 2025; Wang et al., 2025). These methods assess or produce rankings, rationales, or attributed answers; policy isolation begins at that interface, freezes the ranking, and measures the paired effect of the subsequent intervention schedule. SynQA synthesizes training data for support selection, while TracLLM uses informed search to locate contributive texts. These works motivate the candidate interface; our frozen-proposer comparisons use the three rankings specified in Appendix C.

Long-context benchmarks measure answer quality, retrieval, and context utilization as input length grows (Liu et al., 2024; Bai et al., 2024; Hsieh et al., 2024; Yen et al., 2025). Test-time compute allocation improves answer generation by deciding where additional inference is valuable (Snell et al., 2024); our setting allocates a capped sequence of verification calls to improve decisive-evidence recovery from an already frozen answer. The policy-isolation estimand connects this allocation decision to reviewer-time utility while controlling the candidate interface. Contemporaneous budget-aware systems learn when to compress an agent trajectory (Wu et al., 2026), while evidence-grounded review assistants allocate retrieval and execution to claim verification (Yue et al., 2026). Policy-isolated verification complements these systems by identifying the value of an outcome-conditioned target-call schedule after both the answer and candidate interface have been fixed.

Bug localization and reviewer aids rank suspicious regions for inspection (Zhou et al., 2012; Wong et al., 2016). EvidenceCall adds a coalition-state transition after ranking, and the fixed-order design attributes its reviewer-facing effect to that transition. MCTS-Subset supplies an outside-family UCT control (Kocsis & Szepesvári, 2006): it searches subsets under the same charged-call cap, allowing the comparison to distinguish coverage-before-compression scheduling from generic adaptive subset search.

3 POLICY-ISOLATED VERIFICATION

3.1 CONTROLLED POLICY EFFECT

Given context $\mathbf{c}$, query $\mathbf{q}$, and frozen answer $\mathbf{a} = M(\mathbf{c}, \mathbf{q})$, the audit seeks a subset-minimal decisive coalition $\mathcal{S}^* \subseteq \{s_1, \ldots, s_K\}$. Every policy receives the same answer, span partition, candidate ranking, and hard cap B. For span s_k, the operational intervention indicator is

$$\tau_\Delta(s_k; \mathbf{c}, \mathbf{q}, \mathbf{a}) = \mathbf{1}[\Delta(\mathbf{a}, M(\mathbf{c} \setminus \{s_k\}, \mathbf{q})) = 1], \tag{1}$$

where Δ is the task-specific answer-change comparator. A separate correctness judge labels a decisive span as supportive when removal changes a correct answer to an incorrect one and misleading when the direction reverses.

For audit unit i, paired seed u, frozen ranking r_i, and policy π, let $Y_{iu}(\pi; r_i, B)$ denote the evaluation endpoint. The endpoint is selected for each comparison: decisive-span P@5 or coalition F1 for model audits, and accepted findings per hour for reviewer studies. The empirical policy-isolation effect is

$$\widehat{\Delta}_{\mathrm{PI}}(\pi_1, \pi_0; r, B) = \frac{1}{nU} \sum_{i=1}^{n} \sum_{u=1}^{U} [Y_{iu}(\pi_1; r_i, B) - Y_{iu}(\pi_0; r_i, B)]. \tag{2}$$

Because every paired term shares $(\mathbf{a}_i, \{s_k\}_i, r_i, M, B, u)$, Equation 2 attributes the remaining model-audit difference to the online policy. For P@5, Y is decisive-span precision; coalition F1 and reviewer yield use the corresponding paired unit and endpoint.

Let policy π map state $x_t = (Q_t, \widehat{\mathcal{S}}_t, F_t, D_t, B - t)$ to its next charged intervention, where Q_t is the residual ranking and F_t, D_t record observed flips and directions. The shared budgeted target remains

$$\max_{\pi \in \Pi_B} \mathbb{E}\Big[\mathrm{Cov}_5(\widehat{\mathcal{S}}_\pi, \mathcal{S}^*) - \lambda_{\mathrm{mis}} \,\mathrm{MisDir}(\pi)\Big] \quad \text{s.t.} \quad C(\pi) \le B. \tag{3}$$

Equation 3 specifies the policy family and charged resource. Equation 2 is the controlled estimand used to compare members of that family.

We quantify separation from proposal quality with *policy-isolation retention*

$$\rho_{\mathrm{PI}} = \frac{\widehat{\Delta}_{\mathrm{PI}}^{\mathrm{fixed}}}{\widehat{\Delta}_{\text{end-to-end}}}, \tag{4}$$

the fixed-ranking EvidenceCall–Greedy P@5 effect divided by their full-pipeline P@5 difference. The numerator uses one literal shared ranking, whereas the denominator uses the full-pipeline rows with their own proposer ranking. The observed ratios are $7.5/9.3 = 80.6\%$ on pooled QA and $(64.1 - 59.2)/(64.1 - 58.4) = 4.9/5.7 = 86.0\%$ on SWE-bench Verified.

3.2 COVERAGE-BEFORE-COMPRESSION TRANSITIONS

Stage 1 trains a two-layer, 256-hidden-unit proposer once on $5{,}000$ restricted-pool exact-search examples using attention, semantic, and lexical features. It then caches a ranked span list for repeated use. Stage 2 is a deterministic state machine over x_t. For the next candidate $s = \mathrm{head}(Q_t)$, EvidenceCall first queries the target after removing $\widehat{\mathcal{S}}_t \cup \{s\}$. A flip enables a task-safe direction query; supportive evidence enters $\widehat{\mathcal{S}}_t$ and changes the coalition used by subsequent calls. Discovery continues until the queue is exhausted or the reserved tail is reached, after which leave-one-member-in queries remove redundant spans.

This coverage-before-compression order resolves one uncertainty per transition. Necessity tests causal sensitivity, direction separates support from misinformation, coalition continuation reveals evidence conditional on what is already verified, and delayed minimality compresses the returned set after coverage. The policy therefore continues beyond the first sufficient coalition and records a charged, outcome-linked reason for every call. In the cumulative fixed-ranking ablation, adding direction, coalition continuation, and delayed minimality is associated with successive gains of $+2.7$, $+2.5$, and $+2.3$ pooled-QA P@5, respectively.

Table 1: EvidenceCall transition contract. Each executed intervention consumes one target-model call; queue, outcome, and coalition updates are bookkeeping.

Transition	Eligibility	Charged intervention	State consequence
Necessity	$s = \mathrm{head}(Q_t)$	Delete $\widehat{\mathcal{S}}_t \cup \{s\}$ and compare with $\mathbf{a}$	Record the flip in F_t; advance Q_t
Direction	Necessity produces a flip	Replace s with a task-safe neutral alternative	Record supportive or misleading direction in D_t
Coalition continuation	s is supportive	Test the next candidate jointly with $\widehat{\mathcal{S}}_t \cup \{s\}$	Add s and condition later tests on the enlarged coalition
Delayed minimality	Discovery ends or reaches the reserved tail	Restore one coalition member while deleting the others	Remove the member when the coalition flip persists

4 Evaluation Design

Benchmarks and metrics. RULER-128K, MuSiQue, and QuALITY contribute $1{,}200$ test examples each (Hsieh et al., 2024; Trivedi et al., 2022; Pang et al., 2022). The external-label settings are HELMET-LongQA ($N{=}750$), SWE-bench Verified ($N{=}500$), and CiteBench-Long ($N{=}680$). HELMET and SWE-bench are described by Yen et al. (2025) and Jimenez et al. (2024); the Verified subset contains 500 human-validated problems (OpenAI, 2024b). Attributed QA and citation-generation evaluation provide the related citation-evaluation context (Bohnet et al., 2022; Gao et al., 2023). For QA, sentence spans shorter than 12 tokens are merged; for code, executable lines are combined into logical units containing no more than five lines. Model-audit metrics are decisive-span P@5 and coalition F1; reviewer studies measure accepted true findings per hour and time to first accepted finding.

Models, budgets, and compute. Llama-3.1-70B-Instruct is the primary target (Grattafiori et al., 2024). Qwen2.5-72B-Instruct and Claude 3.5 Sonnet test target transfer (Qwen et al., 2024; Anthropic, 2024). All matched comparisons use $B{=}50$ charged target-model calls. On the primary QA serving configuration with $4{\times}$H100 80GB, vLLM 0.5.4, FlashAttention-2, PyTorch 2.3.1, and CUDA 12.4, EvidenceCall averages 46.8 calls and $41.2{\pm}3.8$ seconds at 92K mean context length. All target calls use greedy decoding with 256 generated tokens for QA/citation and 512 for code.

Comparators and statistics. LOO Erasure, Budgeted Shapley, and Greedy Minimal-Set are matched-budget controls. AttrExtractor combines GPT-4o evidence extraction (OpenAI, 2024a) with a MonoT5-3B reranker (Nogueira et al., 2020). The model benchmark evaluates this path at $1.6{\times}$ online compute, while the citation deployment uses a cost-matched configuration. MCTS-Subset uses UCT over evidence subsets, charges every expansion/rollout query, and stops at the same $B{=}50$ cap. Headline model results average ten fixed seeds; uncertainty shown as "$\pm$" is seed standard deviation. Comparative intervals use $10{,}000$ paired example-level bootstrap resamples. Model metrics are example-level means within benchmark; pooled QA is a micro-average over three equally sized benchmarks, whereas Table 8 uses an equal-weight benchmark macro. Documents, repositories, and commits are partitioned before proposer supervision, and target-recalibration slices are disjoint from test rows.

Reviewer protocols. The preregistered 32-reviewer crossover contains $1{,}600$ decisions under a fixed interface. Each reviewer completes ten sessions, each containing five 30-minute graded cases (150 minutes per session; 25 graded hours per reviewer); warm-up, breaks, and surveys are excluded. The 24-developer external code study contains $1{,}920$ decisions in reviewers' own VS Code environments. The 36-reviewer citation study contains $2{,}304$ decisions on 480 held-out CiteBench-Long cases. Both external studies measure paid wall-clock from case open to final submission and include inference latency.

Table 2: Decisive-span P@5 at matched B=50 (mean ± seed standard deviation; ten seeds).

Method	RULER-128K	MuSiQue	QuALITY
LOO Erasure	64.7±1.2	59.2±1.4	57.8±1.5
Budgeted Shapley	61.3±1.6	56.4±1.7	54.2±1.8
Greedy Minimal-Set	68.9±1.1	64.7±1.2	62.1±1.3
EvidenceCall	78.3±0.8	73.6±1.0	71.8±1.1

Table 3: Blind full-context agreement by benchmark. The final column is the lower endpoint of the two-sided 95% exact Clopper–Pearson interval for the displayed agreement count.

Benchmark	Agreement	Rate (%)	95% exact lower bound (%)
RULER-128K	170/180	94.4	90.0
MuSiQue	207/220	94.1	90.1
QuALITY	201/220	91.4	86.8
LongBugBench	64/72	88.9	79.3

5 Results

5.1 Matched call caps reveal policy-sensitive evidence

Does allocation matter when answers, partitions, and call caps are identical? Table 2 shows that EvidenceCall leads all matched-budget controls on every QA benchmark. The displayed rows imply a $+9.3$ example-level pooled P@5 margin over Greedy, while semantic re-adjudication on 400 intervention flips retains $+9.1$ P@5 with 95% CI $[+7.2, +11.0]$.

Coalition F1 rises from 63.2 to 72.6 on RULER-128K, 59.8 to 68.4 on MuSiQue, and 57.4 to 66.2 on QuALITY; the corresponding paired gains are $+9.4$ $[7.5, 11.2]$, $+8.6$ $[6.7, 10.4]$, and $+8.8$ $[6.9, 10.6]$. The mean reflects broad gains rather than uniform movement. Pooled-QA per-example ΔP@5 has median $+7.8$, IQR $[+2.4, +14.6]$, p95 $+24.3$, and a 9.5% loss fraction. SWE-bench Verified has median $+5.2$ and IQR $[+0.8, +11.4]$. Thus allocation changes which evidence reaches the first five positions while retaining a visible minority of ties and losses.

5.2 Blind labels validate the recovered evidence

Table 3 evaluates the restricted-pool labels against expanded search and blind full-context adjudication. Agreement is $642/692$ (92.8%): $578/620$ (93.2%) for QA and $64/72$ (88.9%) for code. The 50 disagreements comprise 20 semantic-equivalence cases, 16 pool misses, 8 coalition ambiguities, 2 accepted-fix proxies, and 4 post-hoc annotation errors. The independently adjudicated EvidenceCall-minus-Greedy gap remains positive on every benchmark with lower bound at least $+5.4$ P@5.

5.3 Exact ranking freezes identify post-proposal value

Table 4 freezes the original proposer order and changes only the budget-spending policy. The paired policy effects are 7.5 pooled-QA and 4.9 SWE-bench P@5, yielding policy-isolation retention of 80.6% and 86.0%, respectively. Each cumulative policy extension has a positive paired lower bound. These contrasts compare successively expanded policies; they do not separately identify the effect of action order.

Cross-family isolation asks a harder question: does the scheduling advantage survive both materially different rankings and a strong policy outside EvidenceCall's four-action family? Table 5 answers yes. MCTS-Subset's UCT search materially improves over Greedy on all six proposer–task combinations, showing that generic adaptive subset search is competitive. EvidenceCall remains ahead with positive confidence-interval lower bounds and substantial win, tie, and loss counts, supporting the complete coverage-before-compression policy across the three candidate interfaces.

Across the six blocks, realized calls are $48.4/50/50$ for Greedy, $46.7/50/50$ for EvidenceCall, and $50/50/50$ for MCTS-Subset (mean/p95/maximum). Every maximum equals the common cap. The

Table 4: Policy isolation with the original proposer ranking fixed exactly, $B{=}50$. Brackets are 95% paired-bootstrap intervals for successive cumulative policy extensions.

Policy	QA P@5	Increment	SWE-V P@5	Increment
Greedy deletion-only	$67.1{\pm}1.0$	—	$59.2{\pm}1.3$	—
+ direction test	$69.8{\pm}1.0$	$+2.7\ [1.6, 3.8]$	$61.0{\pm}1.2$	$+1.8\ [0.6, 3.0]$
+ coalition continuation	$72.3{\pm}0.9$	$+2.5\ [1.4, 3.6]$	$62.8{\pm}1.2$	$+1.8\ [0.7, 2.9]$
EvidenceCall	$74.6{\pm}0.9$	$+2.3\ [1.3, 3.3]$	$64.1{\pm}1.2$	$+1.3\ [0.3, 2.3]$

Table 5: Cross-family policy isolation under three frozen proposers. P@5 uses identical rankings within each block and $B{=}50$ for every policy; EC denotes EvidenceCall.

Frozen proposer	Task (N)	Greedy	MCTS	EC	EC−MCTS [95% CI]
Original	QA (3,600)	67.1	72.8	74.6	+1.8 [+0.7,+2.9]
Original	SWE-V (500)	59.2	62.5	64.1	+1.6 [+0.3,+2.9]
External-only	QA (3,600)	66.5	72.0	73.4	+1.4 [+0.3,+2.5]
External-only	SWE-V (500)	58.8	61.8	63.0	+1.2 [+0.1,+2.3]
MonoT5-3B	QA (3,600)	65.9	70.8	72.1	+1.3 [+0.2,+2.4]
MonoT5-3B	SWE-V (500)	58.5	61.4	62.6	+1.2 [+0.1,+2.3]

heterogeneous W/T/L distributions reported in Appendix C contain non-zero losses in every block. The advantage therefore reflects a positive average policy effect across external supervision, an independently constructed MonoT5 order, and competitive subset search without extra target calls.

5.4 The policy effect improves reviewer throughput

The controlled crossover uses graded audit hours; the external code workflow uses paid wall-clock from case open through final submission and includes model latency. Each controlled task–aid cell aggregates 200 decisions over 100 graded hours. Each external-code aid aggregates 480 decisions. Table 6 preserves these denominators and reports uncertainty on the comparative effect.

In the controlled study, pooled yield rises from 4.05 to 5.33 true issues per graded hour ($+31.7\%$ from unrounded cell means), with 29 of 32 reviewer-level effects positive. In the external code workflow, EvidenceCall raises accepted findings from 3.02 to 3.87 per paid hour ($+28.1\%$): the corresponding raw accounting is 371 findings over 122.8 hours for Greedy and 457 over 118.1 hours for EvidenceCall. Preference partitions exactly as 21/24 EvidenceCall, 2/24 Greedy, and 1/24 no preference.

The external citation workflow tests the same mechanism when labels, proposer supervision, and reviewer population all come from outside the original QA/code path. 36 professional research, fact-checking, and technical-document reviewers inspect 480 held-out CiteBench-Long cases under a locked Latin square. Each reviewer completes 64 cases across EvidenceCall-Ext, MCTS-Subset, cost-matched AttrExtractor, and Greedy, yielding $2{,}304$ graded decisions.

EvidenceCall-Ext exceeds AttrExtractor by $+0.68$ findings/hour (95% CI $[+0.43, +0.93]$). Reviewer-level deltas are positive for $27/36$ participants, and the preference partition includes $4/36$ with no preference. The strong comparators, paid-time endpoint, and visible heterogeneity establish that conditional call allocation improves non-code review throughput as well as evidence recovery.

5.5 The identified effect persists across shifts

Table 8 moves benchmark labels, target family, proposer supervision, and comparator family outward. The benchmark-macro EvidenceCall-over-Greedy gap is $+8.7$ P@5 on Llama, $+7.8$ on Qwen, $+7.4$ on Claude, and $+8.0$ with external-only supervision. AttrExtractor improves over Greedy by $+3.7$ benchmark-macro P@5; on the $1{,}930$-example N-weighted pool, EvidenceCall remains $+5.2$ ahead of AttrExtractor at lower online cost.

Table 6: Reviewer-time outcomes. TIY denotes true issues per hour. Controlled columns are from 32 reviewers and 1,600 decisions; external code is from 24 developers and 1,920 decisions under paid wall-clock.

Aid	MuSiQue TIY/hr	LongBugBench TIY/hr	External code findings/hr
No aid	2.19 ± 0.12	1.92 ± 0.14	—
BM25 highlight	2.92 ± 0.12	2.57 ± 0.13	—
Greedy Minimal-Set	4.30 ± 0.11	3.79 ± 0.12	3.02
EvidenceCall	5.69 ± 0.10	4.96 ± 0.12	3.87
EC−Greedy [95% CI]	$+1.39$ $[1.12, 1.65]$	$+1.17$ $[0.94, 1.41]$	$+0.85$ $[0.63, 1.07]$

Table 7: External citation review under paid wall-clock. Yield is accepted true findings/hour; parenthesized values are raw accepted findings/paid hours. Preference is reviewer-level after exposure to all aids.

Aid	Yield (accepted/paid h)	Time to first accepted	Preference
Greedy	3.14 (502/159.9)	—	—
AttrExtractor	3.60 (563/156.4)	—	—
MCTS-Subset	3.88 (606/156.2)	8.2 min	8/36
EvidenceCall-Ext	4.28 (663/154.9)	7.4 min	24/36
EC-Ext−MCTS [95% CI]	+0.40 [+0.17,+0.63]	—	—

For column c, macro retention is $\mathrm{Ret}_{\mathrm{macro}}(c) = \Delta_{\mathrm{macro}}(c)/\Delta_{\mathrm{macro}}(\text{EvidenceCall})$, computed from unrounded benchmark-level seed means. This ratio of macro gaps preserves the aggregation unit; row-wise ratios provide benchmark-specific transfer detail. The policy effect remains large across independently supplied labels, externally supervised candidate generation, open-weight target transfer, API target transfer, and an outside-family practical extractor.

6 Robustness of Policy Isolation

Three robustness views identify the same policy effect. Blind full-context labels preserve a positive gap on all four audited benchmarks. Alternate operators retain 81.9% of the pooled-QA gap with NLI-Neutral-Mask, 78.5% with Sentence-Shuffle, and 84.2% of the SWE-bench gap with AST-Neutralize. NLI-Neutral-Mask accepts a paraphrase only when DeBERTa-v3-large-MNLI finds non-entailment in both directions. Cross-family and target-transfer tests preserve positive lower bounds under different rankings and target models.

EvidenceCall targets recurring, high-value audits where proposer reuse makes the online cap the latency-controlling resource. Its gains are governed by candidate coverage, intervention validity, and unambiguous direction labels: expanded pools, semantic adjudication, compile-safe replacements, and external supervision directly stress these conditions. The observed loss fraction and non-unanimous reviewer preferences show that the average gain coexists with realistic case and reviewer heterogeneity.

7 Conclusion

Repeated long-context review has two separable system levers: candidate proposal and online verification scheduling. The policy-isolation effect measures the second lever under exact control, while policy-isolation retention quantifies its share of the end-to-end gain. EvidenceCall shows that coverage-before-compression transitions turn the same candidate order and call cap into better evidence recovery than deletion sweeps and generic MCTS subset search. The effect survives independent labels, three proposer families, target transfer, and external supervision, then converts into higher controlled, code, and citation reviewer throughput. Reporting the policy-isolation effect and retention separates scheduling value from proposer quality and reveals when throughput can improve without retraining or additional target calls. Online verification scheduling is therefore an independently measurable and practically consequential design layer for repeated review systems.

Table 8: Externalization at B=50. EC is EvidenceCall; Qwen, Claude, and EC-Ext cells report ΔP@5/retention (%) relative to the row-wise Greedy baseline. The macro row reports 95% lower bounds in brackets.

Benchmark (N)	Greedy	EC Δ	Qwen Δ/Ret.	Claude Δ/Ret.	EC-Ext Δ/Ret.	AttrExtr. Δ
HELMET (750)	63.2±1.1	+8.4	+7.6/90.5	+7.2/85.7	+7.8/92.9	+3.5
SWE-V (500)	58.4±1.4	+5.7	+5.1/89.5	+5.0/87.7	+5.1/89.5	+2.3
CiteBench (680)	54.3±1.4	+12.1	+10.6/87.6	+10.0/82.6	+11.2/92.6	+5.3
Macro (95% LB)	58.6±1.1	+8.7 [7.2]	+7.8 [6.3]/88.9	+7.4 [5.9]/84.7	+8.0 [6.4]/92.0	+3.7 [2.3]

Ethics Statement

All reviewer studies received institutional approval. Participants provided informed consent, were compensated for their time, and could stop without penalty. Analyses used anonymized records, and session audio/video was discarded after transcription.

AI Use Statement

Generative AI was used only for manuscript writing, editing, and presentation; the authors reviewed the resulting edits and remain responsible for the final content.

References

Anthropic. Claude 3.5 Sonnet model card addendum. *Technical Report*, 2024. URL https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022. doi: 10.48550/arXiv.2212.08037. URL https://arxiv.org/abs/2212.08037.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 883–892, 2018. URL https://proceedings.mlr.press/v80/chen18j.html.

Ian Covert, Scott M Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://jmlr.org/papers/v22/20-1316.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017. doi: 10.1109/ICCV.2017.371.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati (eds.), *Human Mental Workload*, volume 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988. doi: 10.1016/S0166-4115(08)62386-9.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. doi: 10.48550/arXiv.2404.06654. URL https://arxiv.org/abs/2404.06654.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European Conference on Machine Learning*, pp. 282–293, 2006. doi: 10.1007/11871842_29.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1011. URL https://aclanthology.org/D16-1011/.

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/.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 708–718. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.63. URL https://aclanthology.org/2020.findings-emnlp.63/.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024a. doi: 10.48550/arXiv.2410.21276. URL https://arxiv.org/abs/2410.21276.

OpenAI. Introducing SWE-bench verified, 2024b. URL https://openai.com/index/introducing-swe-bench-verified/.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.391. URL https://aclanthology.org/2022.naacl-main.391/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

Gorjan Radevski, Kiril Gashteovski, Shahbaz Syed, Christopher Malon, Sebastien Nicolas, Chia-Chien Hung, Timo Sztyler, Verena Heußer, Wiem Ben Rim, Masafumi Enomoto, Kunihiro Takeoka, Masafumi Oyamada, Goran Glavaš, and Carolin Lawrence. On synthesizing data for context attribution in question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16929–16950. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.828. URL https://aclanthology.org/2025.acl-long.828/.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. doi: 10.48550/arXiv.2408.03314. URL https://arxiv.org/abs/2408.03314.

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/.

Yanting Wang, Wei Zou, Runpeng Geng, and Jinyuan Jia. TracLLM: A generic framework for attributing long context LLMs. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3845–3864. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/wang-yanting.

W. Eric Wong, Ruizhi Gao, Yihao Li, Rui Abreu, and Franz Wotawa. A survey on software fault localization. *IEEE Transactions on Software Engineering*, 42(8):707–740, 2016. doi: 10.1109/TSE.2016.2521368.

Yong Wu, Yanzhao Zheng, Tianze Xu, Zhentao Zhang, Yuanqiang Yu, Jihuai Zhu, Chao Ma, Binbin Lin, Baohua Dong, Hangcheng Zhu, Ruohui Huang, and Gang Yu. ContextBudget: Budget-aware context management for long-horizon search agents. *arXiv preprint arXiv:2604.01664*, 2026. doi: 10.48550/arXiv.2604.01664.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (in)fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f5332c8273d02729730a9c24dec2135e-Abstract-Conference.html.

Ling Yue, Chaoqian Ouyang, Hang Xu, Ruijun Huang, Yuchen Liu, Libin Zheng, Wei Liu, Shaowu Pan, Shimin Di, and Min-Ling Zhang. FactReview: Evidence-grounded peer review with execution-based claim verification. *arXiv preprint arXiv:2604.04074*, 2026. doi: 10.48550/arXiv.2604.04074.

Jian Zhou, Hongyu Zhang, and David Lo. Where should the bugs be fixed? — more accurate information retrieval-based bug localization based on bug reports. *Proceedings of the 34th International Conference on Software Engineering (ICSE)*, pp. 14–24, 2012. doi: 10.1109/ICSE.2012.6227210.

A Protocol Details

Proposal labels and split hygiene. Before training the proposer, a deterministic lexical, semantic, and attention ranker forms a top-8 pool. Offline exact search enumerates all non-empty subsets, records answer-changing subsets, and retains subset-minimal coalitions. The proposer uses these labels only during one-time training. QA source documents, book chapters, repositories, and commits are partitioned before label construction; no evaluation source enters proposer training. The ten seeds are $\{13, 29, 47, 61, 83, 97, 113, 131, 149, 167\}$.

Comparators and judges. QA uses exact-answer change during large-scale search and a separate correctness judge for direction. The semantic audit re-adjudicates 400 stratified flips with an NLI-backed judge and human review. NLI-Neutral-Mask generates ten paraphrases per span with Llama-3.1-8B-Instruct at temperature 0.7 and top-p 0.95, then accepts only candidates classified as non-entailing in both directions by DeBERTa-v3-large-MNLI. Code interventions use AST-aware neutralization and require successful parsing before charging the result as valid.

Serving. The primary target runs in bf16 with tensor parallelism four, batch size one, and prefix caching. Mean context is 92K tokens (p95 124K). The 46.8 EvidenceCall calls average 881 ms each, producing 41.2 ± 3.8 seconds per audit on $4\times$H100 80GB. A cheaper Llama-3.1-8B-Instruct configuration on one A100 80GB averages 12.4 ± 1.3 seconds.

B EvidenceCall Transition Contract

EvidenceCall consumes a cached ranking Q, initializes the verified coalition $\widehat{\mathcal{S}} = \varnothing$, and charges every target-model query against B. For the next candidate s, it evaluates deletion of $\widehat{\mathcal{S}} \cup \{s\}$. An answer flip triggers a task-safe replacement query that assigns supportive or misleading direction. A supportive span joins $\widehat{\mathcal{S}}$, after which the residual queue is reconditioned on the updated coalition. Once discovery has consumed the ranked candidates or reached the reserved tail budget, EvidenceCall tests members of $\widehat{\mathcal{S}}$ for redundancy and removes any member whose exclusion preserves the flip. The policy terminates when the queue and eligible minimality tests are exhausted or when $B = 50$ charged calls have been used. This contract makes proposal order an input, observed outcomes the transition signal, and coalition coverage the priority before set compression.

C MCTS-Subset and Multi-Proposer Isolation

Frozen rankings. The original proposer is the restricted-pool two-layer MLP. EvidenceCall-Ext uses a proposer trained on 3,800 externally supervised HELMET-aligned, SWE-bench-style, and citation-sentence examples. The MonoT5-3B proposer caches the sentence/line order produced by the reranking stage of AttrExtractor, without EvidenceCall proposal features. Within each proposer–task block, all three policies consume the identical cached order.

UCT policy. MCTS-Subset follows UCT (Kocsis & Szepesvári, 2006). A state contains the tested subset, observed answer change, current best flipping subset, remaining ranked candidates, and residual call budget. An action adds one untested candidate span. Each expansion performs exactly one target-model query on the resulting subset; selection and back-propagation are free. The tree score is

$$\mathrm{UCT}(v) = \overline{R}(v) + \sqrt{2}\sqrt{\frac{\log N(\mathrm{parent}(v))}{N(v)}}, \tag{5}$$

with rollout reward equal to the observed flip indicator minus 0.02 times subset size. Rollouts break ties by proposer order, continue until a terminal flip or no candidate remains, and consume all remaining calls. Every rollout query is charged, hidden rollouts are disallowed, and search stops at exactly 50 charged calls. Its scalar reward supplies a competitive adaptive-search control, while its absence of direction labels and coverage-before-compression transitions creates the policy-family contrast tested in Table 5.

Statistics and call ledger. Table 5 uses paired differences over the same examples and ten fixed seeds. Its 95% intervals use 10,000 paired bootstrap resamples. EvidenceCall-versus-MCTS W/T/L counts are $1128/1614/858$ and $156/238/106$ for the original proposer, $1062/1687/851$ and $148/249/103$ for external-only supervision, and $1034/1714/852$ and $145/254/101$ for MonoT5-3B (QA/SWE-V). Mean/p95/maximum calls are reported in the main text; every policy has maximum 50.

D Reviewer Study Details

Recruitment and assignment. The controlled crossover recruited 22 university-CS reviewers and 10 professional code-review-marketplace reviewers from 41 screened candidates; all 32 passed a held-out qualification task and were compensated at $50 USD/hour. The 24 external-code reviewers were recruited separately, had median seven years of industry experience, and were disjoint from the controlled cohort. Aid order in both studies was fixed before data collection, and the controlled interface changed only the evidence pane.

Controlled crossover. The 32-reviewer mechanism study uses a locked assignment schedule across four aids. Each session contains a 15-minute warm-up, two 30-minute cases, a 15-minute break, three 30-minute cases, and a 15-minute survey. Thus each session contains 150 graded minutes, and ten sessions yield exactly 25 graded hours per reviewer. Warm-up, break, and survey time are excluded from the preregistered graded-hour endpoint. The fixed interface changes only the evidence pane.

External code workflow. Twenty-four industry developers with median seven years of experience review 80 LongBugBench cases each in their own VS Code environments. A 4×4 Latin square assigns EvidenceCall, Greedy, BM25, and no aid, giving 20 cases per aid per reviewer. Paid wall-clock includes navigation, reading, inference, and final submission. Time to first true issue falls from 9.1 minutes for Greedy to 6.2 for EvidenceCall.

External citation workflow. The 36 citation reviewers are disjoint from every prior cohort and are compensated at $60 USD/hour. They complete 64 cases each from a 480-case CiteBench-Long holdout disjoint from the earlier pilot. A preregistered Latin square balances EvidenceCall-Ext, MCTS-Subset, cost-matched AttrExtractor, and Greedy across aid order and case source, with $9/9/9/9$ reviewers assigned to its four sequences. The evidence pane is controlled; normal browser search and note-taking remain available. The analysis uses reviewer-clustered and case-clustered bootstrap intervals plus a mixed-effects model with reviewer and case random intercepts.

Table 9: Secondary outcomes in the 36-reviewer external citation workflow. Raw false-finding counts are shown in parentheses; each condition has 576 decisions. NASA-TLX (Hart & Staveland, 1988) is lower-is-better.

Aid	Accepted/case	False findings/hr (count)	NASA-TLX
Greedy	0.87	0.87 (139)	55.1
AttrExtractor	0.98	0.76 (119)	49.6
MCTS-Subset	1.05	0.71 (111)	47.8
EvidenceCall-Ext	1.15	0.62 (96)	44.3

Consistent with the $+0.40$ clustered-bootstrap contrast, the mixed-effects coefficient for EvidenceCall-Ext over MCTS-Subset is +0.39 findings/hour with 95% CI [+0.15,+0.63]. The reviewer-level delta distribution has median +0.43, IQR [+0.11,+0.71], p10 -0.12, and p90 +0.96 findings/hour. Reviewer-experience interaction is +0.05 with 95% CI [-0.18,+0.28]; the largest case-source interaction is +0.07 with 95% CI [-0.16,+0.30]. All 36 reviewers complete all assigned cases; no decision is excluded.

Earlier citation scope check. The primary 36-reviewer deployment supersedes this disjoint 16-reviewer design pilot, which contains 320 decisions under paid wall-clock. EvidenceCall reaches 4.37 true citation findings/hour versus 3.68 for AttrExtractor. Its 22/25 preference result is session-level; reviewer-level aggregation is 14/16 EvidenceCall, 1/16 AttrExtractor, and 1/16 no preference.

E SYSTEMS EVIDENCE MATRIX

Table 10: Claim-to-evidence matrix. Each row identifies the controlled object, strongest stress test, and decision-bearing endpoint for one evaluation domain.

Domain	Controlled object	Strongest stress test	Decision-bearing evidence
QA recovery	Frozen answer, span partition, ranking, and B=50 calls	692 blind full-context labels, alternate operators, three proposers, MCTS-Subset	Positive P@5/CoalF1 gaps; Tables 2–5 and 3
Code review	Frozen patch context, evidence pane, and paid-time clock	SWE-bench Verified, AST-neutralization validity, 24 external developers	Accepted findings/hour, latency, and preference; Tables 6 and 8
Citation review	Held-out labels, external proposer, and paid-time clock	Cost-matched AttrExtractor, MCTS-Subset, 480 cases, 36 reviewers	Paid yield, false rate, TLX, and preference; Tables 7 and 9
Target transfer	Cached ranking and fixed Stage-2 policy	HELMET, SWE-V, CiteBench, Qwen, and Claude	Positive gaps, macro retention, and online cost; Table 8

F ADDITIONAL UNCERTAINTY AND ROBUSTNESS

Paired-bootstrap 95% intervals for EvidenceCall minus Greedy at B=50 are $+9.4$ $[7.6, 11.1]$ on RULER-128K, $+8.9$ $[7.0, 10.7]$ on MuSiQue, $+9.7$ $[7.8, 11.5]$ on QuALITY, and $+5.7$ $[3.9, 7.5]$ on SWE-bench Verified. For the fixed-ranking comparison, intervals for successive cumulative policy extensions appear in Table 4. Seed variation and example-bootstrap uncertainty are reported separately throughout.

The independent-label disagreement taxonomy sums exactly to 50: $20 + 16 + 8 + 2 + 4 = 50$. Removing all disagreement cases changes no benchmark gap by more than 0.5 P@5. Under NLI-Neutral-Mask plus AST-Neutralize on a 200-example stratified slice, blind-label agreement is 186/200 and the EvidenceCall-minus-Greedy gap is $+7.1$ P@5 with 95% lower bound $+5.2$.