When Memory Is Enough to Decide but Not Enough to Reconsider: Memory Selection for Argument Survival

Anonymous authors
Paper under double-blind review

Abstract

Language models answer from selected or compressed memories, commonly evaluated by whether they preserve the current answer. Evidence redundant for that answer can become essential after a source is withdrawn. We study this gap across five evidence and reasoning benchmarks using budgeted memories and several open-weight language models. Memories that preserve a correct, supported initial answer can lose every evaluated *revision basis*, a set of records sufficient to justify the answer after withdrawal, even while the full history remains sufficient; this loss reaches 42.2% of such answers. Two histories can share a memory yet warrant different answers after the same withdrawal, so memory alone cannot identify the warranted answer. Paired interventions link model responses to argument completeness and dependence on withdrawn evidence. Replacing one necessary premise reduces revised-answer accuracy by 66.0 percentage points at matched record counts; abstention increases when alternative arguments share the withdrawn premise. These findings motivate argument survival as a selection objective across a prespecified family of withdrawals. Two records can each have zero marginal gain under this objective yet positive joint gain by completing a surviving argument. Pair-rescue implements simple joint admission to capture this value. Executable reasoning tasks and held-out GO-Plus ontology queries with surviving alternatives in the candidate pool show gains in retention and grounded review (correct revised answers with sufficient valid citations) at equal budgets, with current accuracy within a two-point non-inferiority margin. In a separate experiment, identifying and showing the model a verified basis already in memory adds 28.5 points of grounded review. A memory can be enough to decide yet not enough to reconsider.

1 Introduction

Language-model systems answer from selected or compressed memories (Fang et al., 2026; Xu et al., 2025; Jiang et al., 2024). Evidence discarded as redundant can become necessary after a source is withdrawn. Over 10,000 scientific papers were retracted in 2023 (Van Noorden, 2023); Crossref distributes machine-readable retraction records (Rittman, 2025). Retained evidence determines which answers remain justified.

Consider the claim *"The bridge may remain open"* (Figure 1). A damage report (a_1) and a rule requiring damaged bridges to close (a_2) form a *current proof* A that refutes the claim. A dangerous-water-level report (b_1) and its closure rule (b_2) form another argument B. After the damage report is withdrawn as unreliable, the full history still warrants closure. A memory that kept both water-level records can justify that answer; one that dropped the rule loses this argument. The answer stays the same while its justification changes. We call a set of records sufficient to justify an answer after withdrawal a *revision basis*.

Preserving the current proof can thus conceal the loss of every surviving alternative. Two histories can even yield the same memory and supported current answer yet warrant different answers after the same withdrawal (Section 2). A model reading that memory cannot tell them apart. Revision-basis retention complements tests of recall, updating, and conflict resolution (Wu et al., 2025; Hu et al., 2026).

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

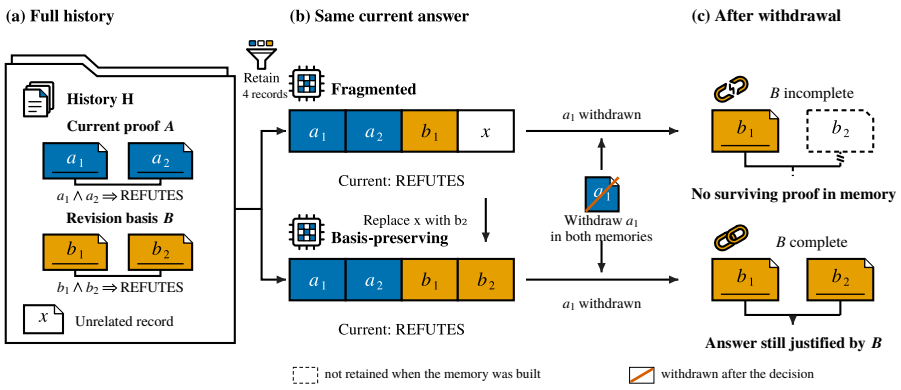

Figure 1: **The same current answer can hide a missing revision basis.** Both memories justify bridge closure using the damage argument A. Only the lower retains the complete water-level argument B, which still justifies closure after a_1 is withdrawn.

Among correct answers with complete current proofs, loss reaches 42.2% on eQASC, a science question-answering dataset with annotated reasoning chains (Jhamtani & Clark, 2020). Two common safeguards also leave alternatives incomplete: LongLLMLingua keeps 92.9% of the records in the most complete revision basis on average, yet retains a whole basis in only 74.3% of memories; protecting the current proof retains one in 57.6% of ProofWriter memories at $K = 180$ (Table 1A). Our principal experiments hold the full-history answer fixed while withdrawing its current evidence, so recovery requires an alternative justification.

Three paired interventions test how evidence structure affects model responses (Section 3.2). Removing the written explanation mainly affects current grounding; removing revision records mainly affects revised-answer accuracy. Replacing one necessary revision fact at matched record counts reduces revised-answer accuracy by 66.0 percentage points; targeted restoration outperforms unrelated additions. With initial argument counts matched, sharing the withdrawn premise increases abstention. A complete surviving argument is the unit needed for reconsideration.

These findings lead to *survival scoring*: select records by whether complete arguments survive a prespecified family of withdrawals. Under this objective, two records can each have zero marginal gain but positive joint gain when both are needed to complete a surviving argument. *Pair-rescue* implements *joint admission*, adding such pairs when single-record search stalls (Section 4). Controlled reasoning tasks separately test survival scoring and joint admission. On held-out GO-Plus ontology queries (The Gene Ontology Consortium, 2026), Pair-rescue raises *grounded review*, correct revised answers with sufficient valid citations, from 20.3% to 62.8% over current-proof-preserving relevance selection at the same budget.

We then trace retained arguments into grounded answers. In a separate experiment, showing the model a verified basis already in memory adds 28.5 points of grounded review. Natural-text experiments examine the discovery and repair needed to realize this principle (Section 5).

Our contributions are:

- **Hidden revision-basis loss.** We define revision-basis retention, measure hidden loss across five sources, and show that identical memories can require different answers after the same withdrawal (Sections 2–3.1).
- **A structural account.** Paired interventions connect revised-answer accuracy and abstention to argument completeness and withdrawal dependence (Section 3.2).
- **A selection principle and its tests.** We formulate an argument-survival objective, test its value and joint admission, and distinguish failures of argument discovery, complete retention, and evidence use (Sections 4–5).

2 REVISABILITY UNDER EVIDENCE WITHDRAWAL

A history H contains *records*: facts, rules, axioms, or evidence sentences. For question q and token budget K, selector f retains memory $S \subseteq H$ before withdrawal. A fixed language-model *reader* then answers from S and a notice listing withdrawn records D. Writing $y^*_X(D)$ for the *warranted answer* from $X \setminus D$, the task is to recover $y^*_H(D)$ from the retained state, with omitted history unavailable.

The current proof A justifies the initial answer. Let $\mathcal{B}_q(H, D)$ contain the sets of records sufficient to justify $y^*_H(D)$ under executable checks or released annotations. These revision bases may share unwithdrawn records with A. Define

$$r(S, D) = \mathbf{1}[\exists B \in \mathcal{B}_q(H, D) : B \subseteq S \setminus D], \qquad R = \Pr(r = 1). \tag{1}$$

Revision-basis retention R is the fraction of questions whose memory still contains a complete revision basis after withdrawal.

Identical memories, different revision targets. Figure 1 varies the memory for one history. We now fix the memory and vary the history:

$$\begin{gathered} f(H^+, q) = f(H^-, q) = S, \qquad y^*_S(\varnothing) = y^*_{H^+}(\varnothing) = y^*_{H^-}(\varnothing), \\ y^*_{H^+}(D) \neq y^*_{H^-}(D). \end{gathered} \tag{2}$$

Under fixed rules $a \Rightarrow q$ and $b_1 \wedge b_2 \Rightarrow q$, let $H^+ = \{a, b_1, b_2\}$, $H^- = \{a, b_1\}$, and retain $S = \{a, b_1\}$ from both. Withdrawing a leaves a proof only in H^+. Identical reader inputs require different full-history answers: any deterministic or randomized reader has mean accuracy at most one half under equal history weights (Appendix A.1).

From retention to an answer. *Current accuracy* checks the initial answer; *grounded review* requires a correct revised answer with sufficient valid citations. G_{ref} checks the citations against an annotated reference basis, while G_{exec} runs an executable *checker* on the cited records. For an *answerable* target, justified by $H \setminus D$, let g indicate a grounded answer under the same evidence check used for retention. Grounded success implies a retained basis, $g = 1 \Rightarrow r = 1$, giving

$$G = \mathbb{E}[g] = \Pr(r = 1) \Pr(g = 1 \mid r = 1). \tag{3}$$

The factors separate argument availability from the reader's success given a retained basis. The second factor is computed over questions for which each selector retains a basis.

Experimental setup. RuleTaker (Clark et al., 2020), ProofWriter (Tafjord et al., 2021), CLUTRR-derived reasoning (Sinha et al., 2019), and GO-Plus support executable checks; FEVEROUS, SciFact, and eQASC provide evidence annotations (Aly et al., 2021; Wadden et al., 2020; Jhamtani & Clark, 2020). Qwen3.8-27B and Gemma-4-31B-it are the main readers, supplemented by Mistral-Small-24B on ProofWriter and GPT-OSS-20B for repair. The main ProofWriter experiments use greedy decoding; ProofWriter groups questions by *theory*, a shared set of facts and rules. We report paired 95% cluster-bootstrap confidence intervals (CIs), prespecified Holm test families, and a two-percentage-point (pp) current-accuracy non-inferiority margin (Appendix A).

3 HIDDEN REVISION-BASIS LOSS AND ITS CAUSES

3.1 CURRENT SUCCESS CAN HIDE INCOMPLETE ALTERNATIVES

We measure loss against two current checks: record coverage and correct, supported answers, conditioning on full histories that remain answerable after withdrawal.

Coverage can hide a missing argument. For questions with at least one revision basis, *best-basis coverage* measures how much of the most complete alternative remains:

$$C(S, D) = \max_{B \in \mathcal{B}_q(H,D)} \frac{|B \cap (S \setminus D)|}{|B|}. \tag{4}$$

A revision basis is retained when $C = 1$; $0 < C < 1$ means that every evaluated alternative is incomplete. LongLLMLingua reaches 92.9% mean coverage but keeps a complete basis in only 74.3% of memories (Figure 2; Appendix B.1). High average coverage can therefore conceal the missing record that prevents an entire argument from being used.

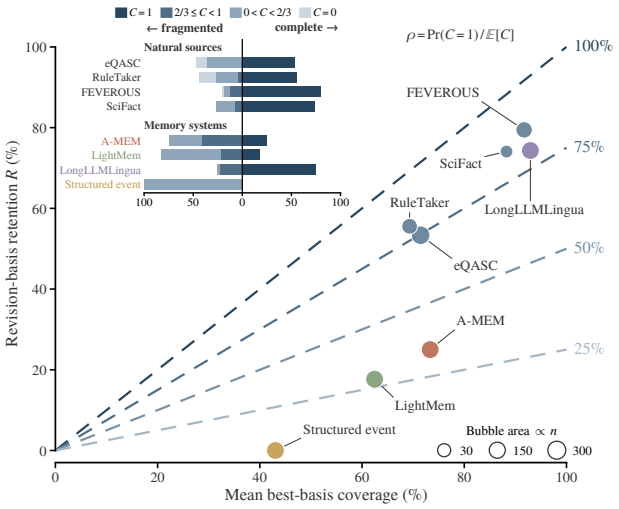

Figure 2: **High coverage can hide an incomplete revision basis.** Mean best-basis coverage versus retention R. Rays hold $R/\mathbb{E}[C]$ fixed; bubble area indicates sample size. The inset partitions complete, partial, and zero coverage.

Correct answers conceal the loss. Among correct answers with complete current proofs, basis loss occurs in 269/638 eQASC observations (question–reader pairs; 42.2%), 66/249 RuleTaker observations (26.5%), and 111/658 ProofWriter observations (16.9%). FEVEROUS and SciFact yield 3.1% and 3.8%, respectively. After withdrawal, 73.2% of responses on these 269 eQASC observations still label the claim supported, although memory no longer holds any annotated argument for it. Appendix B.2 gives the complete counts and evidence checks. **A correct, supported answer and high record coverage can both conceal the loss of every revision basis.**

3.2 The evidence requirements of reconsideration

The first intervention distinguishes information roles; the next two vary completeness and withdrawal dependence. Endpoints are *current grounding* (initial answer and exact proof citations), *revised-answer accuracy* (full-history answer recovery), and *abstention* (INSUFFICIENT).

Written explanations and revision records serve different roles. On 242 FEVEROUS, Climate-FEVER (Diggelmann et al., 2020), and FactNet (Shen et al., 2026) cases, memory also holds a model-written explanation citing the current proof, first checked by an independent model. We remove either this explanation or the revision-basis records. Revised-answer accuracy scores the final label after the model is told to reconsider its answer.

Removing the explanation changes current grounding from 85.95% to 29.34% and revised-answer accuracy from 76.03% to 80.99%. Removing revision records instead gives 88.84% and 21.07%. Comparing the two removals within each outcome gives differences of $59.50\,[52.89, 65.70]$ and $59.92\,[53.31, 66.53]$ pp, respectively (Figure 3A). The explanation mainly supports the current answer, and the revision records the revised answer (Appendix C.1).

Relevant components must form a complete argument. In 300 controlled cases, we replace a necessary revision fact f with a fact from another argument, retaining the rule that requires f. Current-proof records, distractors, and fact/rule counts match; lengths differ by at most four tokens and 2%. The complete memory exceeds the fragmented one in revised-answer accuracy by $66.00\,[61.83, 70.17]$ pp (Figure 3B).

Targeted repair restores f. Sham repair adds another fact, leaving the rule's required premise unsatisfied. Targeted repair exceeds sham by $64.67\,[60.50, 68.83]$ pp, linking recovery to the missing premise. Current-answer tests use the shared current-proof records (Appendix C.2.1).

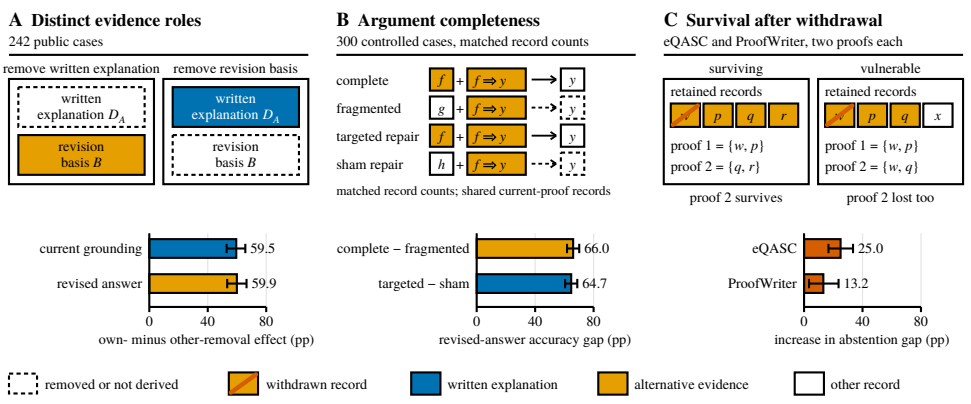

Figure 3: **Paired evidence interventions.** Effects and 95% CIs are pp. **A:** Removing the written explanation versus the revision records. **B:** Replacing and restoring a necessary revision fact. **C:** Increase in the vulnerable-minus-surviving abstention gap after withdrawal. Absolute outcomes: Appendix C.

Complete arguments can share the same point of failure. We match current proofs, record counts, and initial argument counts, varying whether the second argument avoids the withdrawn premise (*surviving*) or depends on it (*vulnerable*). Records that support no argument balance the counts; text lengths are similar. eQASC uses annotated chains; ProofWriter checks remaining derivability exhaustively (Appendix C.3).

On 176 eQASC questions, abstention rises from 6.5% to 33.2% in surviving memories and from 6.3% to 58.0% in vulnerable memories. Withdrawal widens the vulnerable-minus-surviving gap by $25.0\,[16.7, 33.5]$ pp: the reader becomes more likely to abstain when both arguments fail. The corresponding interaction is $13.2\,[3.4, 23.6]$ pp on 76 matched ProofWriter questions (Figure 3C). **Reconsideration requires a complete argument that avoids the withdrawn evidence; it may share other records with the current proof.**

The paired memories differ in completeness or withdrawal dependence. These controlled changes affect answer recovery and abstention, linking evidence structure to reader behavior. Together, the results motivate scoring surviving arguments and admitting complementary records jointly.

4 Selecting Records for Argument Survival

An objective that values surviving arguments. The interventions suggest a selection principle: preserve the complete arguments that remain usable when evidence is withdrawn. For a prespecified withdrawal family $\mathcal{D}$, memory token count c, and executable checker Derives, we score

$$J(S) = \frac{1}{|\mathcal{D}|} \sum_{D' \in \mathcal{D}} \mathbf{1}[\text{Derives}(q; S \setminus D')], \qquad c(S) \le K. \tag{5}$$

Each withdrawal contributes to J when a complete argument remains. Under the rules in Section 2, consider withdrawing a, b_1, or b_2 with equal weight. Memory $\{a\}$ scores $2/3$: it loses its proof when a is withdrawn. Adding either b_1 or b_2 leaves J unchanged; adding both raises it to one. Complete-argument scoring therefore creates value that a single-record marginal gain can miss.

ProofWriter permits argument checks independently of the reader. Here $\mathcal{D}$ contains all single-fact withdrawals and all pairs of current-proof facts, with equal weights. ProofWriter current proofs contain at most four facts, yielding at most six pairs. Selection precedes the actual withdrawal (Appendix D.4).

Atomic selects by relevance; *Verified* also protects the current proof when it fits. *Robust* scores single-record additions by J; *Pair-rescue* also tests pairs when no single addition improves J. All share candidate records, Qwen3-8B relevance scores, memory format, and budget. Robust–Verified tests selection by argument survival; Pair-rescue–Robust tests joint admission under the same objective. Oracle-basis receives a surviving argument (Appendix D).

Table 1: **Selection for argument survival and the value of joint additions.** Current accuracy (a), revision-basis retention (R), and grounded review (G), in percent. Each block compares methods on the same questions and evidence.

A. ProofWriter: survival scoring
443 questions / 300 theories; $K = 180$

Method	a	R	G_{ref}
Atomic	57.1	55.8	24.7
Verified	62.5	57.6	25.1
Robust	64.7	**82.6**	**29.3**

B. ProofWriter: joint admission
Same questions; separate generation

Method	a	R	G_{exec}
Robust	64.6	82.6	33.1
Pair-rescue	64.4	**87.6**	**33.6**

C. GO-Plus: held-out ontology
300 queries / 296 classes; $K = 240$

Method	a	R	G_{exec}
Atomic	79.0	25.0	17.7
Verified	97.7	32.0	20.3
Robust	97.5	32.7	21.3
Pair-rescue	98.2	**82.7**	**62.8**

D. CLUTRR-derived: relations
300 problems; $K = 920$

Method	a	R	G_{exec}
Atomic	40.3	67.7	12.0
Verified	42.2	67.7	12.0
Robust	42.2	67.7	12.0
Pair-rescue	42.7	**79.0**	**13.2**

Paired contrast	Retention Δ [95% CI]	Grounded review Δ [95% CI]
A. Robust–Verified	$+25.1\ [19.8, 30.4]$	$+4.3\ [2.7, 6.2]$
B. Pair-rescue–Robust	$+5.0\ [2.3, 8.0]$	$+0.6\ [0.1, 1.2]$
C. Pair-rescue–Verified	$+50.7\ [44.9, 56.3]$	$+42.5\ [37.3, 47.8]$
D. Pair-rescue–Robust	$+11.3\ [7.7, 15.0]$	$+1.2\ [0.3, 2.2]$

Differences are pp with paired 95% CIs. A and B come from separate model runs; G_{ref} checks a reference basis, whereas G_{exec} checks the actual citations. Qwen and Gemma read every block. Current-accuracy contrasts meet the -2 pp margin. Complete outcomes, including Oracle-basis, are in Appendices E.1, E.2, and E.3.

Survival scoring adds value beyond the current proof. At $K = 180$, Robust raises ProofWriter retention from Verified's 57.6% to 82.6% and G_{ref} from 25.1% to 29.3% (Table 1A). Verified gains $0.3\ [-1.3, 1.9]$ pp over Atomic in grounded review. Verified and Robust share current-proof initialization and the same relevance-based fallback. Robust first adds records that increase argument survival.

Joint admission completes otherwise invisible alternatives. Pair-rescue checks feasible pairs when single-record search stalls, adds the best positive-gain pair, and resumes single-record search. The objective, information, and budget stay fixed (Algorithm 2).

On ProofWriter, Pair-rescue restores a basis in 22 additional memories: $+4.97\ [2.33, 8.04]$ retention points and $+0.56\ [0.11, 1.22]$ in G_{exec}. For the actual withdrawal, *completion distance* is the minimum number of records outside the current proof needed to complete a surviving argument. All 22 gains occur among the 61 distance-two questions; 380 have distance one and two have distance three (Appendix E.2).

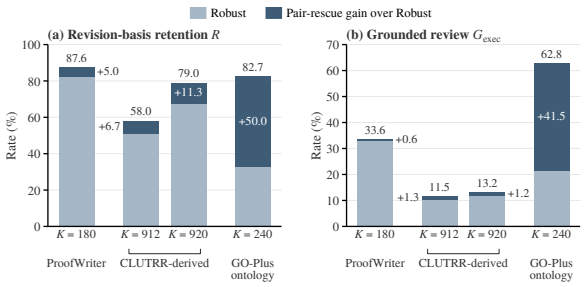

Figure 4: **Joint admission gains.** Light bars show Robust; dark segments add Pair-rescue's gains at fixed objective and budget. Panels report **(a)** retention R and **(b)** grounded review G_{exec}. Top labels give rates (%); + labels give gains (pp).

The CLUTRR-derived kinship task concentrates this requirement: 293/300 generated kinship problems have distance two. At $K = 920$, Pair-rescue retains a revision basis in 34 additional memories, gaining $11.3\ [7.7, 15.0]$ retention points and $1.2\ [0.3, 2.2]$ grounded-review points (Table 1D). Figure 4 shows the joint-admission gains at both tested budgets.

The same joint value appears in a public ontology. GO-Plus tests argument survival on a public biological ontology (The Gene Ontology Consortium, 2026). Records are logical axioms; questions ask whether one class is contained in another. OWL 2 EL permits exact entailment checks (W3C OWL Working Group, 2012). We evaluate 300 held-out queries from 296 classes at a 240-token budget fixed on development queries. Each candidate pool contains a complete justification that survives withdrawal; we measure whether selection retains an available alternative under the budget (Appendix E.3.1). Pair-rescue raises retention from 32.0% to 82.7% and G_{exec} from 20.3% to 62.8% over Verified (Table 1C). Both gains pass Holm-adjusted tests, and current accuracy meets the two-pp non-inferiority margin. The ELK reasoner independently confirms support and citation sufficiency (Kazakov et al., 2014). SUPPORTS responses without sufficient citations fall from 202/600 to 124/600 reader responses (Appendix E.3.1).

On GO-Plus, all 204 questions where Verified loses every revision basis have an alternative that fits alongside the current proof. Pair-rescue recovers 152 and preserves every previously retained basis (Table 1C).

Withdrawal sensitivity and computational cost. On 144 ProofWriter questions, a withdrawal in $\mathcal{D}$ removes the same arguments as the actual withdrawal. Robust gains 25.7 $[16.9, 34.9]$ retention points over Verified. Fixed Pair-rescue memories exceed Verified by 5.18 and 14.69 points under random and mixed current-proof/other-record withdrawals (Appendices E.5–E.6). Atomic, Verified, and Robust mean lengths differ by less than 0.24 token at each ProofWriter budget. Median CPU selection takes 31.45 ms for Robust at $K = 180$ and 0.343 s for GO-Plus Pair-rescue, excluding inference (Appendices D.5–D.10).

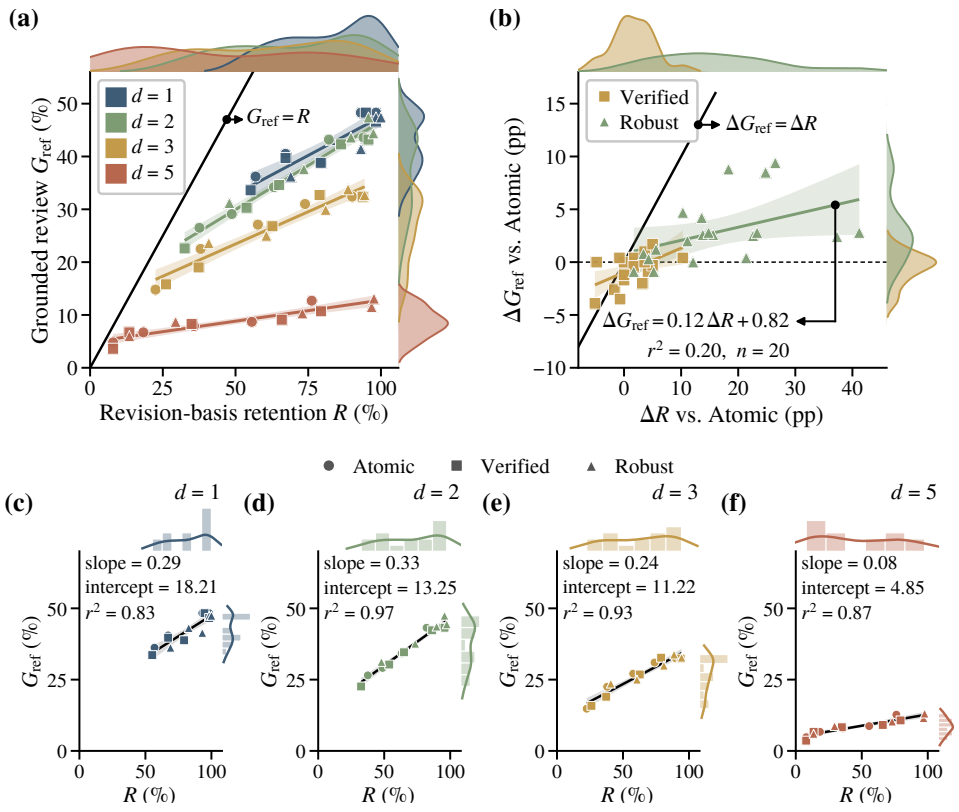

Figure 5: **Retention and evidence use across ProofWriter operating conditions. (a)** Retention R and grounded review G_{ref} by proof depth d and selector. **(b)** Gains over Atomic. **(c–f)** Within-depth points and linear fits. Markers identify selectors; marginal plots show point distributions. Diagonals mark equal rates or gains. Fits summarize aggregate associations.

5 From Preserved Arguments to Grounded Answers

We examine argument discovery, complete retention, and evidence use separately. Figure 5 relates retention to grounded review within proof-depth groups. In the two-reader budget sweep, Robust retains a basis on 96.8% of questions at $K = 240$ yet grounds 32.2% of answers. Equation 3 separates availability from use.

Retained arguments still have to be used. Figure 6 reports the sweep over 443 questions from 300 theories. Both gains over Verified have positive paired 95% CIs at every budget. At $K = 180$, retention gains $25.06\,[19.77, 30.54]$ pp and grounded review $4.29\,[2.67, 6.13]$ pp. At $K = 240$, these narrow to 6.09 and 1.24 pp as Verified retains more alternatives. Retention gains exceed grounded-review gains throughout. All questions remain in the sweep, including 112 whose rules alone exceed $K = 120$ (Appendix E.4).

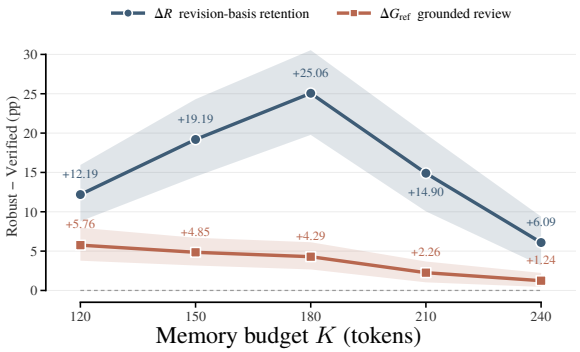

Figure 6: **Paired gains from survival scoring.** Robust minus Verified in R and G_{ref} (pp) across token budgets. Bands: paired 95% CIs clustered by theory.

In a separate three-reader comparison at $K = 180$, both selectors achieve 38.6% grounded review on 762 question–reader pairs where both retain a basis. Robust alone retains one on another 336 pairs: 42 answers (12.5%) are grounded and 73.8% abstain. All net additional grounded answers arise here (Appendix F.1). The groups contain different questions; a separate Atomic–Robust comparison finds deeper proofs and larger revision bases among newly retained cases (Appendix F.2). On GO-Plus, 377/496 responses to Pair-rescue memories with a retained basis are grounded (76.0%).

On 100 independent ProofWriter theories, an executable checker finds a surviving basis within memory and presents its record IDs to the model. This adds $28.5\,[19.5, 37.5]$ pp of grounded review over ordinary reading (Appendix F.3). **Selection preserves an argument; identifying verified support helps the reader use it.**

Identifying arguments enables selection and basis repair. We next examine discovery and complete retention in natural text. With annotated alternatives, exact selection gains 30.5 retention points on 82 eQASC questions where a better memory fits the budget. With model-proposed groups at the primary per-case budgets, revised-answer accuracy improves by $8.9\,[3.0, 15.2]$ pp on RuleTaker, with small nonpositive changes on FEVEROUS and SciFact. Relevance already ranks basis members highly on the latter two (areas under the ROC curve 0.896 and 0.851; Appendix F.5 also reports fixed-budget results).

On 300 cases with the current answer and alternative size (one record) given, we propose and check alternatives. At exactly matched reader-prompt cost, grounded review gains $15.67\,[10.67, 20.67]$ pp over review-aware selection for reconsideration, and $9.67\,[5.00, 14.33]$ over neutral selection for broadly useful information. Natural-length reruns support this pattern. Discovery and checking require additional model calls (Appendix F.6). On QASPER research-paper questions (Dasigi et al., 2021), with support annotations withheld from selection, estimated survival yields lower grounded-review rates than review-aware selection. Diagnostics identify gaps in both discovering alternatives and retaining proposed supports completely (Appendix F.7).

Basis repair replaces records outside the current proof to complete a model-checked alternative. We apply it to A-MEM (Xu et al., 2025), LongLLMLingua, LightMem (Fang et al., 2026), and our structured event-memory baseline, which stores records with event and source identifiers. Across 600 cases, A-MEM and LongLLMLingua gain 7.44 and 6.83 source-match points; A-MEM's gain comes from FactNet (Table 2).

Table 2: **Basis repair: frequency, benefits, and current-answer preservation.** All 600 cases (200 per source) remain in the means, averaging three readers within case.

	Changed	Source-match Δ (pp)			Pooled source-match		Current accuracy
Interface	/600	FV	CF	FN	Base → Repair	Δ [95% CI]	Δ [95% CI]
A-MEM	53	0.0	0.0	+22.3	37.9 → 45.4	+7.44 [5.56, 9.44]	+0.50 [0.17, 0.89]
LongLLMLingua	52	+1.0	+6.5	+13.0	33.7 → 40.5	+6.83 [4.89, 8.89]	+2.50 [1.39, 3.72]
Structured event	5	0.0	+0.5	+0.8	48.7 → 49.1	+0.44 [0.06, 0.94]	0.00 [−0.17, 0.17]
LightMem	4	0.0	+0.3	+0.5	54.8 → 55.1	+0.28 [−0.11, 0.83]	+0.06 [0.00, 0.17]
Raw (reference)	184	+6.3	+16.7	+56.8	16.7 → 43.3	+26.61 [23.61, 29.61]	+22.11 [18.94, 25.39]

FV: FEVEROUS; CF: Climate-FEVER; FN: FactNet. Source-match checks the revised answer and reference source IDs. Base/repair scores are percentages; changes and paired case-bootstrap CIs are pp. All current-accuracy contrasts meet the −2 pp non-inferiority criterion. Raw records are budget matched. Appendix F.8 reports content sufficiency and source-specific CIs.

6 Related Work

Memory construction and selection. A-MEM organizes linked notes (Xu et al., 2025); LightMem stages filtering, consolidation, and offline updates (Fang et al., 2026). LongLLMLingua compresses prompts (Jiang et al., 2024), while OAS chooses retention or consolidation under budget pressure (Kang et al., 2026). We evaluate the complete alternatives these operations leave available after withdrawal and derive a selection criterion that captures the joint value of records.

Knowledge updates and forgetting. LongMemEval tests knowledge updates and temporal reasoning (Wu et al., 2025); MemoryAgentBench includes selective forgetting in incremental interactions (Hu et al., 2026). We withdraw current evidence while holding the full-history answer fixed, measuring loss of justification despite a correct answer.

Truth maintenance and belief revision. Assumption-based truth maintenance tracks justification dependencies (de Kleer, 1986); OWL explanation methods enumerate minimal justifications (Kalyanpur et al., 2007). Resource-bounded revision studies efficient belief change (Alechina et al., 2006). Kumiho stores versioned beliefs (Park, 2026); NeuSymMS retracts conflicting facts (Sultan et al., 2026). We measure how budgeted selection before withdrawal loses complete alternatives despite current-answer success, and how this affects grounded reconsideration (Appendix G.4).

Evidence sufficiency and use. Sufficient Context separates missing evidence from failures to use sufficient context (Joren et al., 2025); +VeriRel uses verification feedback for evidence retrieval (Deng et al., 2025). We trace this distinction through withdrawal: selection preserves complete alternatives, verified support guides their use, and basis repair tests the effect of restoring missing records.

7 Discussion and Conclusion

Current-answer success can hide revision-basis loss. Paired interventions connect these structural losses to changes in model responses. Argument survival turns these findings into a selection objective; joint-admission experiments test the consequences for retention and grounded review.

Exact checks establish sufficiency under the stated semantics. Natural annotations may omit alternatives: if 25% of measured eQASC losses hid an unannotated argument, loss would still be 31.6% (Appendix B.3). Natural-text applications require identifying usable alternatives and preserving them completely (Table 2; Appendices F.5–F.8). In our LoCoMo experiment without withdrawal (Maharana et al., 2024), every candidate pool fits the budget, so all selectors retain identical records and yield identical outcomes (Appendix G.3). Larger missing arguments require larger joint additions; Appendix D.10 reports costs as candidate pools grow. Extending the single-claim, single-withdrawal experiments to repeated changes requires tracking dependencies among surviving arguments.

Like memory selection, training-data selection precedes future questions and evidence changes; extending argument survival to evidence encoded in model parameters remains open.

AI USE STATEMENT

Generative AI tools were used to assist us with writing and editing, retrieving and discovering related work, research ideation, and implementing, debugging, and checking the consistency of experimental code. LLMs were also used within our experimental pipeline to generate memory summaries, retained-state representations, candidate support groups, and reviewer outputs under fixed configurations. The underlying public datasets were not generated or relabeled by LLMs; controlled logical datasets were generated programmatically and verified using executable rules.

ETHICS STATEMENT

The study uses public benchmark evidence and generated diagnostic instances, with no new human annotation or human-subject study. Evaluation relies on released annotations, executable checks, and model outputs. Consequential applications also require reliable provenance, appeal procedures, and human oversight.

REPRODUCIBILITY STATEMENT

Appendix A describes data versions, sampling, metrics and representative prompts; the cited dataset and software publications link to their public sources. Appendix D specifies the algorithms, token accounting and timing. Intervention constructions appear in Appendix C; Appendices B–G provide complete results, statistical definitions, resource accounting and sensitivity analyses.

REFERENCES

Natasha Alechina, Mark Jago, and Brian Logan. Resource-bounded belief revision and contraction. In Matteo Baldoni, Ulle Endriss, Andrea Omicini, and Paolo Torroni (eds.), *Declarative Agent Languages and Technologies III*, volume 3904 of *Lecture Notes in Computer Science*, pp. 141–154. Springer, 2006. doi: 10.1007/11691792_9. URL https://link.springer.com/chapter/10.1007/11691792_9.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/68d30a9594728bc39aa24be94b319d21-Abstract-round1.html.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10: 746–763, 2022. doi: 10.1162/tacl_a_00486. URL https://aclanthology.org/2022.tacl-1.43/.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336. Association for Computing Machinery, 1998. doi: 10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. In *International Conference on Learning Representations*, pp. 4413–4446, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0856bc553d3e3b9827e5140d0ad3bf8d-Abstract-Conference.html.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025: 28th European Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3882–3890, 2020. doi: 10.24963/ijcai.2020/537. URL https://www.ijcai.org/Proceedings/2020/537.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986. doi: 10.1016/0004-3702(86)90080-9. URL https://doi.org/10.1016/0004-3702(86)90080-9.

Xingyu Deng, Xi Wang, and Mark Stevenson. +VeriRel: Verification feedback to enhance document retrieval for scientific fact checking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4706–4711. Association for Computing Machinery, 2025. doi: 10.1145/3746252.3760822. URL https://doi.org/10.1145/3746252.3760822.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. CLIMATE-FEVER: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020. URL https://www.climatechange.ai/papers/neurips2020/67.

Xinyi Fan, Miri Liu, Ruozhen Yang, Siru Ouyang, and Jiawei Han. Can agent memory systems track evolving state? *arXiv preprint arXiv:2608.19652*, 2026. doi: 10.48550/arXiv.2608.19652. URL https://doi.org/10.48550/arXiv.2608.19652.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, pp. 98706–98729, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a05b72653ec5b473732129829ae04195-Abstract-Conference.html.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, Jiaxin Pei, Julian McAuley, Yejin Choi, and Alex Pentland. Benchmarking agent memory in interdependent multi-session agentic tasks. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JHYmxqS9Jv.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.890. URL https://aclanthology.org/2024.findings-acl.890/.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, pp. 156259–156291, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fd1eff9dd295df50a41f2521942fa31d-Abstract-Conference.html.

Harsh Jhamtani and Peter Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 137–150. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.10. URL https://aclanthology.org/2020.emnlp-main.10/.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=6ox8XZGOqP.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, pp. 20310–20334, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Aditya Kalyanpur, Bijan Parsia, Matthew Horridge, and Evren Sirin. Finding all justifications of OWL DL entailments. In *The Semantic Web: 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, volume 4825 of *Lecture Notes in Computer Science*, pp. 267–280. Springer, 2007. doi: 10.1007/978-3-540-76298-0_20. URL https://doi.org/10.1007/978-3-540-76298-0_20.

Qingcan Kang, Mingyang Liu, Shixiong Kai, Kaichao Liang, Zhentao Tang, Yuqi Cui, Tao Zhong, and Mingxuan Yuan. Retain or consolidate? budget-dependent operator selection for language agent memory. *arXiv preprint arXiv:2607.17545*, 2026. doi: 10.48550/arXiv.2607.17545. URL https://doi.org/10.48550/arXiv.2607.17545.

Yevgeny Kazakov, Markus Krötzsch, and František Simančík. The incredible ELK: From polynomial procedures to efficient reasoning with EL ontologies. *Journal of Automated Reasoning*, 53 (1):1–61, 2014. doi: 10.1007/s10817-013-9296-3. URL https://doi.org/10.1007/s10817-013-9296-3.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, pp. 54738–54778, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59f6421e64707225fdf5b28840679a07-Abstract-Conference.html.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability. *arXiv preprint arXiv:2605.05583*, 2026. doi: 10.48550/arXiv.2605.05583. URL https://doi.org/10.48550/arXiv.2605.05583.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. doi: 10.48550/arXiv.2310.08560. URL https://doi.org/10.48550/arXiv.2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.57. URL https://aclanthology.org/2024.findings-acl.57/.

Young Bin Park. Graph-native cognitive memory for AI agents: Formal belief revision semantics for versioned memory architectures. *arXiv preprint arXiv:2603.17244*, 2026. doi: 10.48550/arXiv.2603.17244. URL https://doi.org/10.48550/arXiv.2603.17244.

Vedant Patel. Supersede: Diagnosing and training the memory-update gap in LLM agents. *arXiv preprint arXiv:2606.27472*, 2026. doi: 10.48550/arXiv.2606.27472. URL https://doi.org/10.48550/arXiv.2606.27472.

Om Phadke, Jeffrey Guo, Nolan Koch, Andy Xu, Shanduojiao Jiang, and Kevin Zhu. Truth-maintained memory agent: Proactive quality control for reliable long-context dialogue. In *NeurIPS Workshop on Socially Responsible and Trustworthy Foundation Models*, 2025. URL https://openreview.net/forum?id=n2oOEU1rf9.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. doi: 10.48550/arXiv.2501.13956. URL https://doi.org/10.48550/arXiv.2501.13956.

Martyn Rittman. Retraction watch retractions now in the crossref API. Crossref, January 2025. URL https://www.crossref.org/blog/retraction-watch-retractions-now-in-the-crossref-api/.

Yingli Shen, Wen Lai, Jie Zhou, Xueren Zhang, Yudong Wang, Kangyang Luo, Shuo Wang, Ge Gao, Alexander Fraser, and Maosong Sun. FactNet: A billion-scale knowledge graph for multilingual factual grounding. *arXiv preprint arXiv:2602.03417*, 2026. doi: 10.48550/arXiv.2602.03417. URL https://doi.org/10.48550/arXiv.2602.03417.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1458. URL https://aclanthology.org/D19-1458/.

Mujahid Sultan, Sri Thuraisamy, and Daya Rajaratnam. NeuSymMS: A hybrid neuro-symbolic memory system for persistent, self-curating LLM agents. *arXiv preprint arXiv:2605.17596*, 2026. doi: 10.48550/arXiv.2605.17596. URL https://doi.org/10.48550/arXiv.2605.17596.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 2026. doi: 10.1093/nar/gkaf1292. URL https://doi.org/10.1093/nar/gkaf1292.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26814–26841. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1337. URL https://aclanthology.org/2026.findings-acl.1337/.

Richard Van Noorden. More than 10,000 research papers were retracted in 2023—a new record. *Nature*, 624(7992):479–481, 2023. doi: 10.1038/d41586-023-03974-8. URL https://www.nature.com/articles/d41586-023-03974-8.

vLLM Project. Reproducibility. vLLM documentation, 2026. URL https://docs.vllm.ai/en/stable/usage/reproducibility/. Accessed 25 September 2026.

W3C OWL Working Group. OWL 2 Web Ontology Language Profiles (second edition). W3C Recommendation, 11 December 2012, 2012. URL https://www.w3.org/TR/2012/REC-owl2-profiles-20121211/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Kuan Wang. MemDelta: Controlled baselines and hidden confounds in agent memory evaluation. *arXiv preprint arXiv:2606.29914*, 2026. doi: 10.48550/arXiv.2606.29914. URL https://doi.org/10.48550/arXiv.2606.29914.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, pp. 86809–86836, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Wujiang Xu, Yu Wang, Kai Mei, Kaiqu Liang, Zhenting Wang, Mingyu Jin, Han Zhang, Shi-Xiong Zhang, Wenyue Hua, Sambit Sahu, and Dimitris N. Metaxas. MemGym: a long-horizon memory environment for LLM agents. *arXiv preprint arXiv:2605.20833*, 2026. doi: 10.48550/arXiv.2605.20833. URL https://doi.org/10.48550/arXiv.2605.20833.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710. Curran Associates, Inc., 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

Jianing Zhu, Yeonju Ro, John T. Robertson, Kevin Wang, Junbo Li, Haris Vikalo, Aditya Akella, and Zhangyang Wang. Your agents are aging too: Agent lifespan engineering for deployed systems. *arXiv preprint arXiv:2605.26302*, 2026. doi: 10.48550/arXiv.2605.26302. URL https://doi.org/10.48550/arXiv.2605.26302.

APPENDIX GUIDE

The appendix follows the questions in the main text. Each section gives the corresponding definitions, complete comparisons, or sensitivity analyses.

Main-text link	Supplementary material	Page
Section 2	A. Task, data, and evaluation	15
Section 3.1; Figure 2	B. Occurrence, coverage, and annotation sensitivity	23
Section 3.2; Figure 3	C. Evidence roles, completeness, and withdrawal dependence	25
Section 4	D. Selection algorithms and computational cost	30
Table 1; Figures 4, 6	E. Selection results and withdrawal robustness	34
Section 5; Figure 5; Table 2	F. Evidence use, argument discovery, and memory repair	44
Section 7	G. Representation and temporal scope	53

A TASK, DATA, AND EVALUATION

This section specifies the information available to selection and review, the evidence semantics, and the units used for inference. In the appendix, active basis and active support denote the current proof A; a citation-bound active derivation is the written explanation of that proof. Reviewer denotes the reader. Reproduced prompts retain their original wording. The formal construction appears in Appendix A.1; Appendices A.2, A.3, A.3.3, A.4 and A.5 provide the data, model configurations, scoring rules, and source-level candidate construction.

A.1 INFORMATION LOSS UNDER EVIDENCE WITHDRAWAL

A.1.1 INITIALLY SUFFICIENT STATES WITH DIFFERENT REVISION TARGETS

Initial structural sufficiency means that the retained records establish the current answer under the task semantics. We check this property independently of the reviewer's prediction.

Consider fixed rules $a \Rightarrow q$ and $b_1 \wedge b_2 \Rightarrow q$, histories $H^+ = \{a, b_1, b_2\}$ and $H^- = \{a, b_1\}$, and a selector that retains $S = \{a, b_1\}$ from both histories. The retained a establishes the current answer. Withdrawing $D = \{a\}$ leaves a complete proof in H^+ and no proof in H^-. For any reviewer architecture restricted to the claim, retained state, and withdrawal, these inputs induce the same output distribution. Since the full-history targets differ, the two success probabilities sum to at most one. Under equal weight on the histories, mean accuracy is at most one half.

This pair instantiates Eq. (2): selection preserves the same current proof while discarding the fact that distinguishes the two post-withdrawal targets.

A deterministic construction check instantiates 10,000 matched pairs across four topologies and four application domains, using source invalidation, fact correction, expiry, and rule-exception updates. Each pair has a common initial label, an active proof, a selective update, and no hidden alternative. The executor checks that active-only selection preserves the initial outcome and that the full history and oracle revision state preserve the revised target. These checks validate instances of Eq. (2); behavioral effects are measured in the reviewer experiments.

A.1.2 RETENTION AND EVIDENCE USE

Omission removes every component of a needed alternative. *Fragmentation* leaves relevant components while breaking all complete arguments. *Use failure* occurs when a complete basis is retained and the reviewer fails to produce a correct, grounded review. Revision-basis retention separates the first two from the third; the component-retention example distinguishes completeness from atomic coverage.

For a basis of m necessary records, retaining $m-1$ gives atomic recall $(m-1)/m$ and zero complete-basis retention. The deterministic size sweep checks this construction for $m \in \{2, 3, 4, 5\}$. A required rule or provenance relation is part of the basis when represented as a record. These executor checks establish the intended structural difference; Table C.5 measures its effect on model responses.

A.2 Data, evidence semantics, and sampling

A.2.1 Evidence sources

FEVEROUS. FEVEROUS pairs claims with Wikipedia sentences and table evidence (Aly et al., 2021). We withdraw active evidence selectively so that a designated alternative still supports the claim in the full candidate pool. The pool contains both supports and up to 15 neighboring page or table records ranked by claim-token Jaccard overlap. Annotations identify the evidence and source pages for task construction; the selector then receives the resulting records and constructor scores.

We require usable text, a valid withdrawal, a multi-record alternative with a closure-critical cut, and a budget-feasible state. A single-record screen rejects alternatives containing a record that covers at least 90% of the claim's content tokens. Certified cases have a table/number cut or mixed-unit dependency; exploratory cases rely on annotated set sufficiency. The evaluation includes 180 cases in 179 source clusters, comprising 141 certified and 39 exploratory cases. The 33 development cases belong to 30 separate clusters.

SciFact. SciFact provides scientific claims, abstracts, and annotated rationale sentences (Wadden et al., 2020). We require two rationales from different abstracts with the same verdict, including at least two sentences in the surviving rationale. Candidate pools contain all sentences of the selected abstracts. The 27 evaluation claims comprise 18 supported and nine refuted claims in 21 source clusters. Their support guarantee is annotated set containment, and all are exploratory. Ordinary-selection results use all 27 claims. The alternative-path screen retains 11 for its corresponding diagnostic.

eQASC. eQASC provides annotated two-fact explanations for question answering (Jhamtani & Clark, 2020). We express a question–answer pair as a claim, withdraw part of its active explanation, and check whether a complete annotated alternative remains. The ordinary-selection evaluation contains 375 questions in 180 connected components of shared basis facts. Selection development uses 150 questions; the matched support-survival experiment uses 69 development and 176 confirmation questions. Shared facts define the resampling clusters, and each experiment keeps its own question denominator.

RuleTaker and ProofWriter. RuleTaker and ProofWriter share the reasoning-data lineage introduced by Clark et al. (2020) and extended by Tafjord et al. (2021). RuleTaker cases require a positively derivable query, multiple annotated proofs, and an active-fact withdrawal that leaves another proof with at least two facts. Forward chaining verifies the initial support and the theory after withdrawal. Occurrence counts use any surviving proof. The group-selection experiment additionally reports containment of its designated reference basis, a narrower check distinguished in Appendix B.2.

ProofWriter serves as an executable testbed for isolating support survival. We use forward chaining with open-world semantics, deriving positive and explicitly negative literals separately. Rules remain valid and are charged to the state budget; facts are selectable and withdrawable. Eligible questions have a multi-fact active proof and an alternative that survives the withdrawal. The verifier agrees with released labels on all 4,000 checked instances with formal representations. The evaluation follows these forward semantics, including their treatment of negation. Development and confirmation are disjoint by theory, including exclusion of RuleTaker theories used elsewhere in the study.

Climate-FEVER. Climate-FEVER contains climate claims with evidence and verdict annotations (Diggelmann et al., 2020). The public evidence extension uses 483 label-consistent, source-disjoint cases selected from 1,535 claims. The memory-interface experiment uses 200 cases, and the citation-bound role experiment includes 54 qualified cases. These samples serve different comparisons and retain their respective denominators.

FactNet. FactNet provides multilingual factual grounding data derived from a knowledge graph (Shen et al., 2026). We use verbalized evidence records, with 500 supported cases selected from 2,110 eligible synsets for the public extension. The memory-interface experiment uses 200 cases and the qualified role experiment includes 121. Alternatives in the broad extension are predominantly single-record; dedicated reasoning and multi-record panels test joint completeness.

Data use and attribution. We use the cited public releases and their evidence identifiers. SciFact licenses its claims and evidence annotations under CC BY 4.0 and its abstract corpus under ODC-By 1.0.

Generated reasoning instances. The additional generated panel contains 300 rule-based instances with controlled chain depth. Each instance is an independent unit; the comparison matches initial support counts and tests survival under the same withdrawal.

Table A.1: Principal cohorts and resampling units. All reviewers and conditions from a source unit are resampled as one cluster.

Source	Use	Questions	Units	Resampling unit
FEVEROUS	Selection development	33	30	Source cluster
	Selection evaluation	180	179	Source cluster
SciFact	Selection development	9	8	Source cluster
	Selection evaluation	27	21	Source cluster
RuleTaker	Selection development	50	50	Theory
	Selection evaluation	135	135	Theory
eQASC	Ordinary selection	375	180	Shared-fact component
	Matched development	69	57	Shared-fact component
	Matched confirmation	176	99	Shared-fact component
ProofWriter	Selection development	83	60	Theory
	Selection confirmation	443	300	Theory
	Matched confirmation	76	48	Theory
Generated	Matched structure	300	300	Instance

A.2.2 Screening and sampling

The FEVEROUS construction starts with 1,766 candidates. Of these, 1,585 have usable text and a valid withdrawal; 924 pass the single-record screen; 918 have a closure-critical cut; and 724 are budget feasible. This eligible pool contains 573 certified and 151 exploratory cases. Removing development cases leaves 691 evaluation candidates. The 180-case evaluation sample is stratified by certification, label, and basis size.

We use two probes to identify cases for the natural-evidence intervention analysis. An active-only input tests answerability without the alternative; an active-plus-distractor input tests other routes after removing the designated alternative. When both reviewers answer a probe correctly, the corresponding case enters the descriptive analysis for that screen. Ordinary-selection results retain all cases, while intervention results report the smaller screened population. The intervention results describe cases that pass these reviewer-based screens.

Table A.2: ProofWriter screening under forward-chaining semantics. Counts are questions before panel sampling. The two support-overlap categories partition eligible cases.

Outcome	Validation	Test	Train
Conclusion not derivable	37,013	73,157	253,705
Single-record support	11,953	23,623	82,950
No surviving alternative	1,617	3,109	10,380
Eligible	261	561	1,761
Disjoint supports	11	24	87
Shared supports	250	537	1,674
Unparsed or NatLang	—	5,270	—

Dashes mark unreported categories. The method-confirmation sample contains 25 disjoint-support and 418 shared-support questions.

A.2.3 STATE BUDGETS

The FEVEROUS, RuleTaker, and SciFact occurrence rows in Table B.1 use the primary per-case construction budget

$$K_i = c_i^{\text{basis}} + \alpha\big(c_i^{\text{pool}} - c_i^{\text{basis}}\big), \qquad \alpha = 0.35. \tag{6}$$

Here c_i^{basis} and c_i^{pool} are the reference-basis and candidate-pool costs used in construction. This design controls capacity relative to a known support. The eQASC ordinary-selection panel uses its primary record-selection setting, with mean Atomic state cost 111.2 tokens. ProofWriter uses a fixed 180-token state budget.

To assess selection under a budget fixed independently of each evaluation case, we choose one tight and one loose budget per source using development data (Table A.3). At the tight budget, a complete reference basis fits in 162/180 FEVEROUS, 115/135 RuleTaker, and 23/27 SciFact cases. All cases remain in the main means; a paired sensitivity analysis uses the feasible subset. Appendix F.5 distinguishes this experiment from the per-case budget design.

Table A.3: Source-level budgets fixed from development data. Costs include the charged rule context and state formatting.

Source	Tight K	Loose K	Development cases	Tight-budget feasible
FEVEROUS	230	600	33	162/180
RuleTaker	130	190	50	115/135
SciFact	480	1,400	9	23/27
ProofWriter method	180	—	83	—

Appendices F.6, F.8 and G.1 specify the cohorts used for interface repair and the public evidence extensions. Their evidence units, budgets, and selector initializations are defined within each comparison.

A.3 OUTCOME DEFINITIONS AND STATISTICAL INFERENCE

Evidence checks for the reported outcomes. Figures 2 and 5 use R for revision-basis retention; Figure 5 uses G_{ref} for reference-checked grounded review. Figure 6 plots paired differences in these two outcomes. Table 2 reports the separate source-match score, which checks the correct revised answer and reference source IDs at compressed-memory interfaces. Grounded review instead requires sufficient evidence in the cited records. For G_{exec}, the executable checker tests whether the cited valid records derive the target. For G_{ref}, the cited records must contain a designated reference basis; in natural-evidence tasks, citation checks use the released evidence annotations. The ProofWriter pair comparison reports both checks on the same answers. We use *grounded review* for these answer-and-argument checks and specify the applicable definition with each experiment. The separate *source-match score* at compressed memory interfaces checks the answer and source identifiers. Content sufficiency of rewritten notes is evaluated separately. Native conversational outcomes use the original dataset rubrics.

A.3.1 LABELS AND GROUNDING

In the principal selector comparison, every full history remains answerable after withdrawal, and every selector is evaluated against that same answer. An appropriate abstention on an incomplete selected state still represents a lost answer for the combined selection-and-review system. The intervention experiments answer different questions: component replacement measures recovery of the full-history label, matched survival measures abstention, and role-specific removal measures current grounding and post-withdrawal final-decision accuracy.

Current accuracy checks the initial label; *current grounding* additionally requires the exact current-proof citations. In the main text, *revised-answer accuracy* checks recovery of the answer warranted by the full history after withdrawal. In the role experiment this is the final-decision endpoint, which also requires the instructed `REOPEN` field; all post-withdrawal outputs satisfy that requirement. In the component experiment it is full-history label recovery. *Grounded review* additionally requires

a complete valid citation basis. Input-relative review accuracy, where reported separately, uses the label warranted by the shown evidence. *Abstention* records an INSUFFICIENT response regardless of the reference label.

Answers are normalized to SUPPORTS, REFUTES, or INSUFFICIENT. Ordinary-selection current tests use the retained state; the Complete–Fragmented experiment uses a common active-only input. Review labels are compared with the reference specified for each experiment. A direct deletion experiment can change that reference, whereas the matched survival analysis measures the same abstention event in both conditions and phases.

Let C be the cited IDs, V the visible IDs, and $\mathcal{B}$ the evaluated valid-basis family. For an answerable target, grounding requires

$$\exists B \in \mathcal{B} : \; B \subseteq C \subseteq V, \qquad C \cap D = \varnothing. \tag{7}$$

V includes records still shown in the prompt even when they are marked withdrawn; $C \cap D = \varnothing$ excludes their use as evidence. Additional visible and valid citations are permitted. Fabricated or withdrawn citations fail the check. Grounded review requires both a correct label and this certificate. Correct INSUFFICIENT is evaluated separately on insufficient-reference cases and requires an empty citation set under the relevant protocol. An assertion of SUPPORTS or REFUTES without a valid citation basis is counted as an unsupported assertion relative to the evaluated bases.

For the ProofWriter selector comparison, all 443 full-history targets are SUPPORTS both before and after withdrawal. Current label accuracy therefore measures answers in this selected positive-query population. Grounded review additionally checks that the cited records contain a complete surviving proof. Every selected state remains in the all-case denominator.

The citation-bound role-specific post-withdrawal endpoint is

$$\mathbf{1}[\text{action} = \text{REOPEN} \;\wedge\; \text{normalized final label} = \text{reference label}]. \tag{8}$$

The output fields are `action`, `final_label`, and `evidence_ids`. The instruction supplies `REOPEN`, so this endpoint measures final-label recomputation under instructed reopening. All 726 post-withdrawal outputs contain that action; omitting the action requirement changes no scored final decision. Thus the increase after active-derivation removal is present in the label outcome itself. Citation completeness is scored by grounded review, a separate endpoint. The pre-withdrawal endpoint additionally requires the exact active citations. Appendix C.1 gives the removal conditions and the corresponding outcome contrasts.

A.3.2 UNITS AND ESTIMATORS

Point estimates average question–reviewer observations. In the ProofWriter comparison, 443 questions belong to 300 theories; bootstrap draws resample theories, preserving all questions, reviewers, and selector conditions within each sampled theory. Questions receive equal weight, with equal reviewer weight within question. Retention is checked once per state and repeated in reviewer tables only to display the shared structural property.

For eQASC, connected components of shared basis facts define the resampling units. Other source panels resample source clusters or theories, and the generated panel resamples instances. Interface repair averages three reviewers within each of 600 cases before paired case-level inference. The exact-prompt-cost experiment assigns one constructor to each of 300 source cases. In the role experiment, all intervention outcomes remain paired when resampling the 242 qualified cases.

For the behavioral evaluations and budget sweep, we calculate 95% percentile cluster-bootstrap intervals from 20,000 draws and compute differences before display rounding. Seeds are 20260909 for the ProofWriter/eQASC mechanism analyses, 20260908 for the token-accounted multi-source and fixed-budget analyses, and 20260906 for controlled complete-basis analyses. ProofWriter/eQASC two-sided bootstrap tail estimates have a reporting floor of 0.0001; a four-test Holm adjustment gives 0.0004 when all raw estimates reach that floor. Arm means and paired differences have separately calculated intervals.

Holm adjustment is applied within the stated test families. The primary ProofWriter grounded-review family contains Robust–Verified, Robust–Atomic, Verified–Atomic, and Oracle–Atomic. The additional-reviewer panel has its own four contrasts. The multi-source family contains three

basis-group–Atomic and three destroy–sham tests. The fixed-budget selected-method family contains guarded basis-group–Atomic on each source's feasible subset; the unguarded comparisons form a secondary family. Current non-inferiority requires the lower paired 95% bound to exceed -0.02. For the matched pre-withdrawal differences, inference uses estimation intervals.

The post-hoc budget sweep uses 20,000 paired theory-cluster bootstrap draws with seed 20260914. Point estimates pool question–reviewer observations, so theories determine dependence without receiving equal weight in the mean. At each budget, the six-test Holm family comprises Robust against Atomic and Verified on retention, current accuracy, and grounded review. Holm correction is applied within each budget, using a separate family from the primary confirmation. Two-sided bootstrap sign-tail estimates have a 0.0001 reporting floor. Conditional-use intervals resample theories within each selector's own retained subset.

The completion-distance analysis uses one structural observation per question on the original 443-question cohort. Its Pair-rescue contrasts use paired 95% theory-cluster bootstrap intervals within each distance stratum (Appendix E.2). This secondary analysis reuses the primary question cohort.

The proxy-coverage analysis in Appendix E.5 reuses the original states and three-reviewer outputs. It uses 20,000 paired theory-cluster bootstrap draws with seed 20260919, retaining all questions and reviewer outputs from a sampled theory. Within this secondary analysis, a four-test Holm family covers the Robust–Verified retention and grounded-review contrasts on structurally uncovered questions and the two corresponding uncovered-minus-covered interactions. Other group-specific estimates have pointwise intervals and unadjusted p-values. The coverage-stratified budget analysis uses the separate two-reviewer sweep.

A.3.3 Reviewers and decoding

Table A.4: Model roles, decoding, and execution resources. Configuration rows distinguish the primary two-reader comparison from additional-reader panels.

Role or panel	Models or hardware	Recorded setting
Primary ProofWriter	Qwen3.8-27B; Gemma-4-31B-it	Temperature 0; at most 256 new tokens; seed 20260906
Additional ProofWriter	Mistral-Small-24B-Instruct-2501	Same frozen states; development qualification in Table A.5
GO-Plus reading	Qwen3.8-27B; Gemma-4-31B-it	At most 256 new tokens; two NVIDIA A100 services
Interface repair	Qwen3.8-27B; Gemma-4-31B-it; GPT-OSS-20B	Three readers averaged within each source case
Shared relevance	Qwen3-8B	One cached record-scoring response per question
ProofWriter selection	Dual-socket AMD EPYC 7713	One CPU process; three repeats after warm-up
ProofWriter budget sweep	Four NVIDIA A100-SXM4-80GB GPUs	Two GPUs per reader; serving concurrency eight

Principal ProofWriter, its sweep, and the multi-source group-selection protocol share the stated greedy configuration. Role-task generation and independent qualification use the templates in Appendices A.4 and C.1. Hardware timings and request counts remain separate in Appendices D.6, D.10 and E.3.1.

Qwen3.8-27B and Gemma-4-31B-it form the primary reviewer pair, using the public repositories `https://huggingface.co/Qwen/Qwen3.8-27B` and `https://huggingface.co/google/gemma-4-31B-it`. The smaller `https://huggingface.co/Qwen/Qwen3-8B` checkpoint supplies shared relevance scores. The reviewer pair was specified before constructing and evaluating confirmation states. An additional Mistral-Small-24B evaluation uses the same states after development-only qualification: at least 0.95 schema validity in each phase and 0.50 Atomic current accuracy. Table A.5 reports this task-specific screen.

Additional model IDs are `mistralai/Mistral-Small-24B-Instruct-2501` and `openai/gpt-oss-20b`. Supplementary reasoning-style checkpoints use a 1,024-token output cap and are reported separately from the 256-token primary runs in Table A.4.

Table A.5: ProofWriter development qualification. Values are percentages. Eligibility requires at least 95% schema validity in each phase and 50% Atomic current accuracy.

Checkpoint	Current schema	Review schema	Atomic current	Eligible
Qwen3.8-27B	100.0	97.0	66.3	Yes
Gemma-4-31B-it	98.8	100.0	60.2	Yes
Mistral-Small-24B	98.8	100.0	68.7	Yes
Llama-3.3-70B-AWQ	40.1	99.1	10.8	No
Llama-3.2-3B-Instruct	97.6	98.8	18.1	No
DeepSeek-R1-Distill-Llama-70B	66.9	74.7	49.4	No

A.4 Representative prompts

The templates below reproduce the core instructions, with dynamic fields enclosed in braces and line breaks added for readability. The evidence-review template is used in the component experiment and evaluations with that output format. Other versions of that template substitute task-specific fields. The citation-bound role task separately uses the post-withdrawal action-and-label format defined in Eq. (8).

A.4.1 Evidence reviewer

```
Claim: {claim}
Reconsider the claim after the following records or provenances
become invalid and may not be used: {invalid_record_ids}.
Cite a complete sufficient set of remaining shown record IDs.
If the supplied evidence is insufficient, answer INSUFFICIENT
with an empty evidence list.

Retained state:
{retained_state}

Return exactly one JSON object and no other text:
{"label":"SUPPORTS|REFUTES|INSUFFICIENT","evidence_ids":["ID"]}
```

For the current phase, the reconsideration instruction is replaced by:

```
Decide the claim from the retained state as it currently stands.
Cite a complete sufficient set of shown record IDs.
```

The Complete–Fragmented current phase inserts only the common active records into this template. Its review phase inserts the corresponding full retained state. The review instruction permits abstention on insufficient supplied evidence; the target-label-recovery metric compares the answer with the original full-history label.

A.4.2 Record relevance and candidate-basis construction

The Qwen3-8B constructor receives this system message:

```
Infer evidence structure from the supplied pool only.
```

Its user prompt is:

```
You are given a claim and a pool of retained evidence records.
Score every record 0-3 for how directly it bears on the exact claim,
then propose up to five record groups that could EACH independently
justify a verdict on the claim if the currently cited support
(records {active_ids}) becomes unusable. A group must contain every
premise it needs and must not rely on those cited records.
Return exactly one JSON object:
{"record_scores":{"F00":3},"candidate_bases":[["F00","F01"]]}.
```

```
Claim: {claim}
Rules (always valid):
{shared_context}

Pool:
{record_blocks}
```

All three ProofWriter selectors use the same returned record scores. Verified and Robust evaluate selected facts with the solver, as specified in Appendix D; the group-based comparison instead admits the constructor's proposed bases as units.

A.4.3 PRE-WITHDRAWAL ACTIVE-DERIVATION WRITER

```
Before any future challenge is known, write a short executable
derivation for the recorded current decision. Use every listed
evidence record, cite only its opaque ID, and explain why the
evidence entails or contradicts the claim. Do not discuss alternative
evidence, future review, invalidation, or revision. Return exactly
one JSON object with keys decision, evidence_ids, and rationale,
with no surrounding commentary.

Claim: {claim}
Recorded current decision: {gold_label}
Recorded current evidence:
{active_records}
```

Generation checks require the correct decision, in-state citations, the exact active-citation set, and a nonempty rationale. Independent qualification then tests whether a reviewer can recover the decision and exact citations from the derivation.

A.4.4 INDEPENDENT DERIVATION QUALIFICATION

```
Independently recover the current claim label from the frozen
pre-challenge derivation. The decision field is intentionally hidden.
Use only the citation-bound rationale and cited records. Return
the recovered label and the exact evidence IDs needed.

Claim: {claim}
Cited evidence records:
{cited_records}
Frozen rationale: {rationale}
```

Qualification requires the correct recovered label and exact active citations. The role experiment's post-withdrawal output instead contains `action`, `final_label`, and `evidence_ids`; its instruction gives `REOPEN`. The final-decision endpoint in Eq. (8) checks label recomputation and the required action field independently of evidence citations.

A.5 CANDIDATE POOLS AND INFORMATION ACCESS

Table A.6: Candidate construction and evidence definitions by source. Each row specifies the pool, record unit, support check, and budget.

Source	Pool and record unit	Support check and budget
FEVEROUS	Selected sentence/table supports and up to 15 neighboring records ranked by claim overlap	Annotation containment; source-level or per-case budgets stated in Appendix A.2
SciFact	Sentences from selected abstracts	Annotation containment; 27 exploratory claims
eQASC	Released explanation facts; shared facts define clusters	Annotated chains; matched states use a 50% character-length caliper
RuleTaker	Theory facts with fixed rules	Forward verification; any proof for occurrence, designated basis in the group-selection endpoint
ProofWriter	Selectable facts plus fixed, charged rules	Exact two-tokenizer state cost; disjoint theory development/evaluation
CLUTRR-derived	Generated families rendered by fixed relation templates	Executable mappings; 300 independent generated evaluation families
GO-Plus	At most 30 logical axioms from the pinned supported fragment	Formal subsumption; exact $K = 240$, eligibility-accrual panel of 300 targets
QASPER	Native full-text paragraphs, with common candidate construction	Released evidence sets; no gold answer or support-size input to constructor
PersonaMem	Released conversational prefix for each question	Model-estimated supports; native answer score, binding 2K reading
LoCoMo	Chronologically maintained native Mem0 memories	Model-estimated supports; nonbinding $K = 4{,}096$ read cap; native F1/adversarial-detection score; no withdrawal

B Occurrence, coverage, and annotation sensitivity

These analyses support Section 3.1: Appendix B.1 distinguishes partial coverage from complete retention, Appendix B.2 conditions on supported current answers, and Appendix B.3 quantifies sensitivity to unannotated alternatives.

B.1 Best-basis coverage and complete retention

For each retained state and evaluated revision family, best-basis coverage is the largest fraction of one basis present after invalidated records are removed. Complete retention is the event that this fraction equals one. Figure 2 plots mean coverage against the complete-state rate, with separate source and memory-system populations. Its radial guides hold the ratio $\Pr(C = 1)/\mathbb{E}[C]$ fixed. These guides compare the complete-state rate with mean coverage. The sources retain their own sample sizes and budgets. The inset counts the realizations $c = 1$, $2/3 \leq c < 1$, $0 < c < 2/3$, and $c = 0$.

The conditional occurrence analysis below additionally restricts question–reviewer observations to a correct current answer and a retained complete initial support. Its denominator is consequently different from the state-level coverage display. Each rate retains its own evidence family and sampling unit.

B.2 Current sufficiency and revision-basis loss

B.2.1 Initially sufficient natural-source states

On the four sources, record-wise selection preserves the active support, so current-correct observations are initially sufficient by construction. Within these observations, post-withdrawal support is retained versus lost in 285 versus 9 FEVEROUS cases, 369 versus 269 eQASC cases, 183 versus 66 RuleTaker cases, and 50 versus 2 SciFact cases, counting question–reviewer observations. The corresponding loss rates are 3.06%, 42.16%, 26.51%, and 3.8%. The initially insufficient cells are empty under these active-preserving selection rules.

Table B.1: **Current correctness hides revision-basis loss within initially sufficient states.** The last column counts observations with a supported, correct current answer and no retained revision basis after withdrawal. Its denominator is the preceding column.

Source	Questions	Current accuracy (%)	Supported current answers	Revision basis lost: count (%)
eQASC	375	85.1	638	269 (42.2)
FEVEROUS	180	81.7	294	9 (3.1)
SciFact	27	96.3	52	2 (3.8)
RuleTaker	135	92.2	249	66 (26.5)
ProofWriter	443	59.4	658	111 (16.9)

Counts are question–reviewer observations: two reviewers per question in the first four rows and three in ProofWriter. Natural-source checks use annotations; SciFact is an exploratory 27-question panel. Rule-task checks search for any proof. Budgets differ by source. RuleTaker and ProofWriter share a data-generation lineage.

The RuleTaker occurrence check executes the entire retained theory after withdrawal. Ten questions retain another proof even though the designated alternative is incomplete. Their 20 reviewer observations therefore belong to the surviving-support group, giving 183 retained and 66 lost observations among 249 correct initial answers. The certificate columns in Appendix F.5 check the designated reference bases and retain that narrower interpretation.

Natural-language checks count complete annotated explanations in eQASC and complete evidence sets in FEVEROUS and SciFact. Losing such a certificate establishes the loss of an evaluated justification, with unannotated alternatives left unresolved. Appendix A.2 specifies the support guarantee for each source.

B.2.2 Correct answers without an initially sufficient state

Without active-support initialization, Atomic can answer correctly on a state that already lacks a proof. Of 790 correct answers in 1,329 ProofWriter observations, 658 are initially sufficient and 132 are insufficient. The sufficient group contains 547 observations with surviving support and 111 without, yielding $111/658 = 16.9\%$ revision-basis loss. Conditioning only on correctness gives $243/790 = 30.8\%$, because that count also includes the 132 initially unsupported answers.

Table B.2: Initial sufficiency among current-correct ProofWriter observations. The final two columns are percentages within the correct-and-sufficient subset.

Selector	Correct and sufficient	Correct and insufficient	Basis retained within supported-correct (%)	Grounded review within supported-correct (%)
Atomic	658	132	83.1	41.9
Verified	855	1	66.1	32.9
Robust	882	1	86.2	36.4

These conditional subsets describe the structure of correct answers. Selector effects in Table 1 use the complete common cohort.

Verified and Robust each retain an initially sufficient state on 442 of 443 questions. Among their current-correct observations, only one lacks initial support in each arm. Their active-support initialization makes this property nearly universal, with the remaining question invoking the budget fallback described in Appendix D.

B.2.3 Behavior after support loss

Table B.3 crosses current correctness with support loss before filtering on initial sufficiency. Requiring initial sufficiency gives 111 ProofWriter observations, with 24.3% SUPPORTS and 74.8% INSUFFICIENT, and 66 RuleTaker observations, with 10.6% SUPPORTS and 87.9% INSUFFICIENT. These conditional subsets have the denominators used in Table B.1; the ProofWriter cell in Table B.3 has the broader denominator. Table B.3 shows different response patterns among support-lost states

Table B.3: Joint current correctness and post-withdrawal support under Atomic. Counts are question–reviewer observations; behavioral columns are percentages within each row.

Source	Current	Basis	Count	All (%)	Post SUPPORTS (%)	Grounded review (%)
eQASC	Correct	Retained	369	49.2	90.5	68.8
	Correct	Lost	269	35.9	73.2	0.0
	Wrong	Retained	31	4.1	6.5	3.2
	Wrong	Lost	81	10.8	4.9	0.0
ProofWriter	Correct	Retained	547	41.2	65.3	50.5
	Correct	Lost	243	18.3	32.5	0.0
	Wrong	Retained	194	14.6	13.4	6.7
	Wrong	Lost	345	26.0	4.6	0.0

eQASC: 375 questions and 750 observations. ProofWriter: 443 questions and 1,329 observations. Grounded review requires a retained certificate.

across tasks. In the eQASC current-correct, support-lost cell, 73.2% of responses assert SUPPORTS and 26.4% abstain. In the corresponding ProofWriter cell, 32.5% assert SUPPORTS and 66.7% abstain. Grounded review is zero in both cells because the required support is absent, even though their label behavior differs markedly.

B.3 ANNOTATION INCOMPLETENESS

Let L_{ann} be the rate of states with no complete evaluated support, and let α be the unknown fraction of these states that in fact contain an unannotated valid support. Assuming accepted annotated supports are semantically sound, the corresponding semantic loss is

$$L_{\text{semantic}} = L_{\text{ann}}(1 - \alpha). \tag{9}$$

This identity gives a sensitivity analysis indexed by an assumed false-loss fraction α. For the eQASC conditional rate $269/638$, hypothetical α values of 0, 0.10, 0.25, and 0.50 imply semantic-loss rates of 42.16%, 37.95%, 31.62%, and 21.08%. Method contrasts require selector-specific omission rates. Executable entailment checks determine support directly under the stated rule and ontology semantics.

C EVIDENCE ROLES, COMPLETENESS, AND WITHDRAWAL DEPENDENCE

The following constructions test the three questions in Section 3.2. We describe role-specific removal (Appendix C.1), component replacement and repair (Appendices C.2 and C.2.1), and matched withdrawal dependence (Appendix C.3), including complete outcomes and concrete examples.

C.1 EVIDENCE REMOVAL AND BEHAVIORAL ROLES

C.1.1 CITATION-BOUND ACTIVE DERIVATIONS

Before any withdrawal, a writer produces a current derivation with exact opaque active-evidence citations and a rationale. Generation checks require a correct decision, in-state citations, the exact active-citation set, and a nonempty rationale. The independent qualification reviewer receives the cited records and rationale with the decision field hidden, and must recover the correct label and exact active citations (Appendix A.4).

The confirmation cohort contains 270 attempted public cases and excludes the 18 development cases. All 270 pass the generation checks; 242 pass independent qualification, exceeding the predefined minimum of 150. These qualified cases contain 168 supported and 74 refuted claims, with 67 FEVEROUS, 54 Climate-FEVER, and 121 FactNet cases. Writers contribute 101 Mistral, 117 Qwen, and 24 Llama cases. Three paired conditions and two endpoints give 1,452 reviewer outputs.

Active removal deletes the generated, citation-bound active derivation. Revision removal deletes the dormant revision basis. For endpoint $e \in \{A, R\}$, let $Y_e(I)$ be intact accuracy and $Y_e(-A)$ and

Table C.1: **Current derivations and revision bases affect different behaviors.** The same 242 qualified public cases receive both removals. Upper panel: accuracy (%). Lower panel: own-role minus other-role removal loss (pp), with paired 95% confidence intervals.

Condition	Current grounding	Final-decision accuracy
Intact	85.95	76.03
Remove active derivation	29.34	80.99
Remove revision basis	88.84	21.07
Role specificity (pp)		
Active specificity		**59.50** $[52.89, 65.70]$
Review specificity		**59.92** $[53.31, 66.53]$

Final-decision accuracy tests label recomputation under instructed `REOPEN`; grounded review evaluates citation completeness separately. Each specificity compares the two removal effects within one endpoint.

$Y_e(-R)$ be the accuracies under the two removals. Define $\text{Loss}_e(-X) = Y_e(I) - Y_e(-X)$. The selective effects are

$$\text{Specificity}_A = \text{Loss}_A(-A) - \text{Loss}_A(-R) = Y_A(-R) - Y_A(-A), \tag{10}$$

$$\text{Specificity}_R = \text{Loss}_R(-R) - \text{Loss}_R(-A) = Y_R(-A) - Y_R(-R). \tag{11}$$

The intact condition cancels in each specificity contrast. Current grounding evaluates both the current label and exact active citations. Final-decision accuracy checks the final label and required action field under instructed `REOPEN`. Grounded review separately evaluates citation completeness.

Table C.2 separates intact-to-removal losses from the contrasts between removals.

Table C.2: Role-specific effects on the 242 qualified cases. Loss is intact accuracy minus intervened accuracy; specificity compares the two removals within the same endpoint. Effects and confidence intervals are pp.

Contrast	Effect [95% CI]
Current-grounding loss from active removal	$+56.61$ $[+50.00, +62.81]$
Current-grounding loss from revision removal	-2.89 $[-6.20, 0.00]$
Active specificity	$+59.50$ $[+52.89, +65.70]$
Final-decision loss from revision removal	$+54.96$ $[+48.35, +61.57]$
Final-decision loss from active removal	-4.96 $[-9.09, -0.83]$
Review specificity	$+59.92$ $[+53.31, +66.53]$

Both specificities have positive intervals within each source. Active specificity is numerically larger on FEVEROUS and Climate-FEVER, while review specificity is larger on FactNet. The source-specific intervals evaluate each removal contrast within the qualified cohort.

C.2 Component replacement and repair

C.2.1 Complete and fragmented revision bases

The controlled component-replacement experiment contains 300 cases. Complete retains an active basis, a complete revision basis, and distractors. Fragmented substitutes a necessary fact in that revision basis with a fact from a different support chain, leaving the original rule in place. The active records and distractors are identical. Fact and rule counts are matched, while the substituted record's identity and content change.

Both states fit within 256 tokens. For each of the Qwen and Gemma tokenizers, the absolute length difference is at most four tokens and the relative difference is at most 2%. The current phase shows only the common active records. The review phase shows the corresponding Complete or Fragmented state and the same withdrawal.

The sole Complete–Fragmented substitution is EVID00 $\rightarrow$ EVID12. EVID12 belongs to another revision chain and cannot satisfy the premise of EVID14. The shared withdrawal invalidates EVID10,

Table C.3: The double dissociation holds within all three sources of the same qualified cohort. Upper panel: endpoint accuracies (%). Lower panel: selective effects and paired 95% confidence intervals (pp).

Source	Endpoint	Intact	$-A$	$-R$
FEVEROUS	Current grounding	92.54	26.87	95.52
	Final decision	64.18	73.13	28.36
Climate-FEVER	Current grounding	79.63	12.96	81.48
	Final decision	57.41	68.52	16.67
FactNet	Current grounding	85.12	38.02	88.43
	Final decision	90.91	90.91	19.01

Source	Cases	Active specificity	Review specificity
FEVEROUS	67	$+68.66$ $[+56.72, +80.60]$	$+44.78$ $[+31.34, +58.21]$
Climate-FEVER	54	$+68.52$ $[+55.56, +79.63]$	$+51.85$ $[+37.04, +64.81]$
FactNet	121	$+50.41$ $[+41.32, +59.50]$	$+71.90$ $[+63.64, +80.17]$
Pooled	242	$+59.50$ $[+52.89, +65.70]$	$+59.92$ $[+53.31, +66.53]$

All estimates use the same reviewer outputs. Each interval uses 20,000 paired percentile bootstrap draws over original cases, within the stated source for source-specific results. $-A$ removes the active derivation; $-R$ removes the revision basis.

Table C.4: A complete record pair from the controlled replacement experiment. The claim is "The proposal is approved." A check mark denotes a retained record. Withdrawing EVID10 leaves a complete revision basis only in Complete.

Record	Content	Complete	Fragmented
EVID16	Condition A1 is satisfied.	✓	✓
EVID10	Condition A2 is satisfied.	✓	✓
EVID02	If Conditions A1 and A2 are satisfied, the proposal is approved.	✓	✓
EVID00	Condition B1 is satisfied.	✓	—
EVID12	Condition C1 is satisfied.	—	✓
EVID14	If Condition B1 is satisfied, the proposal is approved.	✓	✓
EVID01	N0 satisfied.	✓	✓
EVID03	N4 satisfied.	✓	✓
EVID04	N5 satisfied.	✓	✓

The pair contains eight records per state. Complete replaces EVID12 with EVID00; all other records are shared. Distractor descriptions are abbreviated. Each original state uses 126 tokens under Qwen and 118 under Gemma in this example.

breaking the active rule's A2 premise. Complete can still derive approval from EVID00 and EVID14; Fragmented has no complete surviving support. Targeted repair reverses the substitution and produces the Complete state exactly. Sham repair substitutes another fact that fails to complete the retained rule; in this case it uses EVID21, "Condition C2 is satisfied."

Table C.5: Controlled component replacement and repair, 300 cases. The outcome is recovery of the original full-history label in every condition. Effects and 95% confidence intervals are pp.

Contrast	Effect	95% CI
Complete − Fragmented	+66.00	[+61.83, +70.17]
Targeted repair − Sham repair	+64.67	[+60.50, +68.83]

Both conditions use the original full-history label, SUPPORTS, as their recovery target. This label-only endpoint assigns zero credit to INSUFFICIENT; grounded review separately requires complete citations. Current responses use the common active-only input. Associated identifier, layout, and order controls change accuracy by at most two points.

A separate deterministic component sweep varies basis size from two to five. Partial retention, including $m - 1$ of m necessary components, gives zero revision-basis availability, while retaining the complete basis permits successful execution. This threshold describes the executor's dependence on complete support.

C.2.2 PROMPT WORDING SENSITIVITY

On 100 paired cases, the complete–fragmented review gap is $66.5\,[59.5, 73.5]$ points with the original prompt and $69.0\,[62.0, 75.5]$ with a paraphrase. Their difference is $2.5\,[-2.0, 7.0]$ points. The current phase uses the same active-only records under either wording.

C.3 MATCHED SUPPORT-SURVIVAL EXPERIMENTS

C.3.1 CONSTRUCTION

Each pair contains an active support A, a second support, and inert records. We construct

$$S_{\text{surviving}} = A \cup B \cup I_a, \qquad S_{\text{vulnerable}} = A \cup C \cup I_d, \tag{12}$$

where B survives the realized withdrawal D and C depends on a withdrawn premise. Both states contain complete supports before withdrawal. The inert sets I_a and I_d adjust record counts without contributing a proof. The paired state check requires equal initial support counts, a surviving alternative in the surviving state, and no surviving evaluated support in the vulnerable state.

For eQASC, the second supports are released valid explanation chains with equal numbers of facts. Candidate pairs satisfy a 50% character-length caliper, and inert records balance state record counts. Annotation checks then count the complete chains before and after withdrawal. All 176 confirmation pairs and 69 development pairs satisfy the recorded structural checks. The length caliper controls textual size approximately; exact tokenizer costs can differ across a pair.

The ProofWriter construction enumerates minimal supports of at most four facts, builds the two states, and verifies them by forward chaining. Of 91 fresh pairs, 76 also have equal record counts; of 38 additional pairs, 37 do. The equal-record analyses use those subsets. Table C.6 additionally reports the full panels, retaining the pairs with count differences as sensitivity analyses. The support-count and post-withdrawal derivability requirements apply to both.

C.3.2 AN EXECUTABLE EXAMPLE

Consider the query "Fiona is young." The active support is `F08`: "Fiona is kind" and `F09`: "Fiona is smart." The rules include

$$\text{smart}(x) \Rightarrow \text{cold}(x), \qquad \text{cold}(x) \wedge \text{furry}(x) \Rightarrow \text{kind}(x),$$
$$\text{kind}(x) \wedge \text{cold}(x) \Rightarrow \text{young}(x), \qquad \text{big}(x) \Rightarrow \text{cold}(x).$$

The surviving state adds F07: "Fiona is furry." Its second support is {F07, F09}: smart derives cold, cold and furry derive kind, and kind and cold derive young. The vulnerable state replaces F07 with F05: "Fiona is big." Its second support is {F05, F08}, which shares the kind premise with the active route.

Withdrawing F08 leaves the surviving route available and breaks both routes in the vulnerable state. Both states also contain the same five inert records, F00–F04. Each has eight records and two initial minimal supports. The post-withdrawal support counts are one and zero. This particular pair has exactly equal Qwen/Gemma state lengths of 139/131 tokens.

C.3.3 AN ANNOTATED-CHAIN EXAMPLE

An eQASC item asks whether arthropods can destroy crops and spread disease. Its active chain uses "Insects spread disease and destroy crops" and "All insects are arthropods." Both states also contain "Scorpions are non-insect arthropods." The surviving state adds "Insects can spread disease and destroy crops"; the vulnerable state instead adds "Most arthropods are insects." Withdrawing the first active fact leaves an annotated chain in the surviving state and breaks both annotated chains in the vulnerable state.

Both states contain four records and two initial annotated chains. Their Qwen/Gemma lengths are 87/78 tokens in this example. Chain validity follows the dataset annotations. The experiment measures survival of annotated chains, as specified in Appendix A.2.

C.3.4 PAIRED EFFECTS AND ABSOLUTE OUTCOMES

Let $a_{z,t}$ be the probability of INSUFFICIENT in condition z and phase t. The matched interaction is

$$\tau = (a_{\text{vulnerable,post}} - a_{\text{surviving,post}}) - (a_{\text{vulnerable,pre}} - a_{\text{surviving,pre}}). \tag{13}$$

Table C.6: Support-survival interactions and full-panel sensitivity estimates. Effects are percentage points with paired 95% intervals. The full ProofWriter panels include pairs with unequal record counts.

Panel	Questions	Units	Interaction	p
eQASC confirmation	176	99	$25.0\ [16.7, 33.5]$	< 0.0001
eQASC development	69	57	$19.6\ [11.1, 28.8]$	< 0.0001
ProofWriter fresh, matched	76	48	$13.2\ [3.4, 23.6]$	0.0091
ProofWriter separate panel, matched	37	27	$6.8\ [-7.1, 21.2]$	0.3914
Generated controlled extension	300	300	$56.5\ [52.5, 60.5]$	< 0.0001
ProofWriter fresh, full pool	91	57	$13.2\ [4.7, 22.3]$	—
ProofWriter separate panel, full pool	38	28	$6.6\ [-6.9, 20.8]$	—

Table C.7: Absolute abstention before and after withdrawal. eQASC counts pool two reviewers; the generated-panel rates average the same reviewers. Percentages.

Panel	State	Before withdrawal	After withdrawal
eQASC	Surviving	$23/352 = 6.5$	$117/352 = 33.2$
	Vulnerable	$22/352 = 6.3$	$204/352 = 58.0$
Generated	Surviving	0.0	0.3
	Vulnerable	0.0	56.8

For eQASC confirmation, the absolute abstention rates rise from 6.5% to 33.2% in the surviving arm and from 6.3% to 58.0% in the vulnerable arm. The interaction is 25.0 percentage points. Current-accuracy vulnerable–surviving differences are $+0.9\ [-2.6, 4.3]$ points on eQASC and $+4.0\ [-2.3, 10.1]$ on the 76-question ProofWriter subset. These intervals characterize pre-withdrawal differences at the achieved precision.

The fresh ProofWriter interaction is $+13.2\ [3.4, 23.6]$ points. Its post-withdrawal contrast alone is $+9.9\ [-0.7, 21.0]$. The interaction uses the paired change across phases and therefore has its

own uncertainty. The additional 37-question panel gives $+6.8\,[-7.1, 21.2]$, extending the range of observed effect sizes.

C.3.5 GENERATED STRUCTURAL EXTENSION

The generated panel contains 300 independent rule-based instances with varying rule depth. Both reviewers have zero pre-withdrawal abstention. The pooled post-withdrawal interaction is $+56.5\,[52.5, 60.5]$ points. Qwen's interaction is $+52.0\,[46.3, 57.7]$ and Gemma's is $+61.0\,[55.7, 66.3]$. Surviving-arm review accuracy is 100.0% for Qwen and 99.3% for Gemma, or 99.7% pooled after rounding. These results characterize the generated instance population and its realized rule depths, alongside the released-source panels.

D SELECTION ALGORITHMS AND COMPUTATIONAL COST

This section specifies the selectors in Section 4, their shared information, and exact budget accounting. Algorithms 1 and 2 distinguish singleton from pair admission; Appendices D.5, D.6 and D.10 connect these operations to support evaluations, measured latency, and candidate-pool scaling.

D.1 RELEVANCE SCORES AND RECORD ADMISSION

A Qwen3-8B constructor scores each record for its direct relevance to the claim on an instructed 0–3 scale. Missing or nonnumeric scores default to zero; numeric values are used as returned. The resulting order is descending score, then lexicographic record ID. All record-wise filling follows this order. A record that exceeds the remaining budget is skipped, and the scan continues. Admission always keeps an entire record.

Atomic starts with no facts in the ProofWriter experiment. The record-wise comparators in FEVEROUS, eQASC, RuleTaker, and SciFact begin with the active support. This difference determines whether initial sufficiency is guaranteed by admission or must be checked after selection.

D.2 CURRENT VERIFICATION AND SURVIVAL SCORING

In the selector pseudocode, R denotes the fixed rule set and S the selected facts. Let $U(S) = \mathbf{1}[R \cup S \vdash q]$ be current derivability. Verified and Robust use the same forward-greedy procedure in Algorithm 1, with objective $F = U$ for Verified and $F = J$ for Robust. Both initialize with A and fall back to Atomic if that state exceeds K. The active support already gives $U(A) = 1$, so Verified immediately reaches its objective maximum and fills by relevance. This is exactly its behavior on 442 confirmation questions. On the remaining question, the active-support state costs 181 tokens and triggers the Atomic fallback.

Robust scores each feasible one-record addition by the number of proxy withdrawals under which the query remains derivable, proportional to the uniform mean in Eq. (5). It continues only while an addition strictly improves this score. Otherwise, it fills by relevance using the same scores as Verified. Multi-record completions whose individual additions have zero gain can enter during this final fill. The example in Section 4 illustrates the zero-gain case; Appendix E.2 reports retention by completion distance. The measured solver-call counts and selection times appear in Section 4 and Table D.1.

D.3 COMPLEMENTARY-PAIR EXTENSION

Pair cost is computed from the jointly serialized state. Selection uses the current candidate pool, active support, and prespecified withdrawal family; the realized withdrawal and surviving reference basis are reserved for evaluation. Pair admission completes two-record complements under the original score and filling rule. Larger complements may require more than two records to be admitted in one step.

Algorithm 1 Forward-greedy evidence selection

Require: Fact pool H, active support A, fixed rules R, budget K, scores s, objective F

$L \leftarrow$ records sorted by $(-s_r, \mathrm{ID}(r))$
if $\mathrm{cost}(R, A) > K$ **then**
    return FILL$(\varnothing, L, K)$
end if
$S \leftarrow A$
loop
    $E \leftarrow \{r \in H \setminus S : \mathrm{cost}(R, S \cup \{r\}) \leq K\}$
    if $E = \varnothing$ **then**
        return S
    end if
    Choose $r^* \in \arg\max_{r \in E}[F(S \cup \{r\}) - F(S)]$, ties by L
    if $F(S \cup \{r^*\}) - F(S) \leq 0$ **then**
        return FILL(S, L, K)
    end if
    $S \leftarrow S \cup \{r^*\}$
end loop
function FILL(S, L, K)
    for r in L with $r \notin S$ **do**
        if $\mathrm{cost}(R, S \cup \{r\}) \leq K$ **then**
            $S \leftarrow S \cup \{r\}$
        end if
    end for
    return S
end function

Algorithm 2 Pair-rescue with the original survival objective

Require: Candidate records H, active support A, fixed context, scores s, budget K

$L \leftarrow$ the frozen relevance order; $S \leftarrow A$
if $c(S) > K$ **then**
    return FILL$(\varnothing, L, K)$
end if
loop
    $E_1 \leftarrow \{\{x\} : x \in H \setminus S,\ c(S \cup \{x\}) \leq K\}$
    Choose $U_1 \in E_1$ of maximal gain $J(S \cup U_1) - J(S)$, if nonempty
    if $E_1 \neq \varnothing$ and its maximal gain is positive **then**
        $S \leftarrow S \cup U_1$; **continue**
    end if
    $E_2 \leftarrow \{\{x, y\} \subseteq H \setminus S : c(S \cup \{x, y\}) \leq K\}$
    Choose $U_2 \in E_2$ of maximal gain, if nonempty; use the frozen tie-break
    if $E_2 = \varnothing$ or its maximal gain is nonpositive **then**
        return FILL(S, L, K)
    end if
    $S \leftarrow S \cup U_2$
end loop

D.4 THE PROXY WITHDRAWAL FAMILY

Let H_F contain the candidate fact IDs and let $P_{12}(A)$ denote the first at most 12 pairs from the lexicographically ordered combinations of active IDs. The family is

$$\mathcal{D}(H, A) = \{\{f\} : f \in H_F\} \cup P_{12}(A), \qquad Q(D') = |\mathcal{D}|^{-1}. \tag{14}$$

Every scenario withdraws facts, while rules remain valid. The family is generated from the current input before the realized withdrawal is revealed to the evaluator.

Realized withdrawals contain one, two, three, and four facts on 284, 141, 17, and one questions. Exact set matching covers the 284 single-fact withdrawals. A further 144 contain a minimal effective core that matches a proxy scenario, and one has the same proof-survival signature as a scenario despite matching neither its set nor an effective core. These comparisons leave 14 structurally uncovered questions. Appendix E.5 defines the categories and reports their retention, behavioral outcomes, and evidence requirements. The classification is computed after selection; the selector and its proxy family are unchanged.

D.5 NUMBER OF DERIVABILITY EVALUATIONS

Let n be the number of candidate facts, $a = |A|$, and $M = n + \min\{12, \binom{a}{2}\}$ the number of proxy scenarios. With T successful additions, evaluating every feasible single-record candidate at each round requires $O((T+1)nM)$ derivability checks before accounting for caching. Since $T \le n$, this gives an $O(n^3)$ worst-case count. Runtime also includes the cost of forward chaining and token-budget checks. The measured call counts in Table D.1 concern the finite pools in this experiment.

No cap binds in the 83 development and 443 evaluation questions: at most four active facts yield at most six pairs. The cap could affect pair coverage only when $a \ge 6$; active supports of at most five facts generate at most ten pairs. For larger supports, record-ID order determines which pairs are retained. Candidate filtering, scenario sampling, and cached proof evaluations are possible extensions for controlling search size.

D.6 SELECTION LATENCY

We measure selection and forward verification on a shared dual-socket AMD EPYC 7713 host, using one process and sequential states with no batching and `TOKENIZERS_PARALLELISM=false`. After ten warm-up cases per selector, each state is timed three times. We take its median and summarize these per-state values across 443 questions. The timing excludes tokenizer loading, theory preprocessing, relevance-score loading, and constructor inference. One precomputed constructor response per question is shared across all selectors and budgets.

At $K = 180$, summing the per-question timings gives 2.45 seconds for Atomic, 4.68 seconds for Verified, and 17.20 seconds for Robust across 443 states. These totals share the exclusions of the per-state measurements.

The reviewer sweep uses four NVIDIA A100-SXM4-80GB GPUs, two per reviewer, with a serving concurrency limit of eight sequences. Software versions are vLLM 0.27.1, PyTorch 2.13.0+cu130, and Transformers 5.5.3. The latency measurements above describe CPU selection under the shared-host conditions and exclude reviewer generation.

D.7 STATE SERIALIZATION AND TOKEN COST

The ProofWriter state consists of fixed rule context followed by selected record blocks:

```
Rules (always valid):  <fixed rule text>
RECORD F00:  <record text>
RECORD F01:  <record text>
```

Let T_Q and T_G be the tokenizers loaded with the saved Qwen3.8-27B and Gemma-4-31B-it reviewer checkpoints identified in Appendix A.3. We charge

$$\mathrm{cost}(R, S) = \max_{T \in \{T_Q, T_G\}} \left| T\big(\mathrm{serialize}(R, S);\ \texttt{add_special_tokens=False}\big) \right|. \tag{15}$$

Table D.1: Selection latency and solver calls across state budgets. Latencies are milliseconds; each cell summarizes 443 states.

		Selection latency (ms)			Solver calls		
K	Selector	Median	Mean	P90	Median	Mean	P90
120	Atomic	4.94	4.70	6.57	0	0.00	0
	Verified	6.75	6.83	10.69	1	4.63	14
	Robust	7.16	16.26	44.81	4	106.98	416
150	Atomic	5.47	5.15	7.35	0	0.00	0
	Verified	8.51	8.75	13.78	9	6.94	15
	Robust	16.60	25.22	58.90	143	166.64	476
180	Atomic	5.87	5.52	7.93	0	0.00	0
	Verified	10.78	10.57	15.98	11	10.15	15
	Robust	31.45	38.83	77.31	260	249.81	476
210	Atomic	6.07	5.80	8.60	0	0.00	0
	Verified	11.00	10.96	16.61	11	10.91	15
	Robust	39.35	44.60	85.25	260	283.78	476
240	Atomic	6.20	6.02	9.07	0	0.00	0
	Verified	11.23	11.21	17.06	11	10.91	15
	Robust	40.02	45.09	85.85	260	284.34	476

Selection latency excludes shared relevance scoring, tokenizer initialization, and theory preprocessing. Each state is timed three times after warm-up; its median timing is used in the cross-question summaries. Solver calls give a separate measure of computation.

The serialized rules, record IDs, prefixes, separators, and newlines all count toward $K = 180$. The claim, withdrawal instructions, system and reviewer prompts, chat template, and generated output lie outside this retained-state budget. All four selection arms use the same serialization and cost function. Their mean state costs round to 168 tokens.

When the serialized rules alone exceed K, no fact can be admitted. The implementation returns the rules-only state at its actual serialized cost, which can exceed the nominal limit. The sweep marks these questions as infeasible and retains them in every selector's denominator; all have zero revision-basis retention. This occurs only at $K = 120$ in the reported grid. The constraint in Eq. (5) applies to feasible states.

D.8 Oracle support admission

Oracle-basis receives a designated gold support that survives the realized withdrawal. It prioritizes those records, skips any record that would exceed the budget, and then fills remaining capacity in relevance order. It has no requirement to retain A.

Oracle retains the entire supplied support on 437 of 443 questions. On the other six, that support remains logically valid but costs 181–184 tokens after adding fixed rules and formatting. Three failures share a theory. Removing record-label overhead would make all six supports fit; with the prescribed rendering, each resulting 171- or 180-token state omits one required record. All six questions remain in the evaluation.

D.9 Constructor-based grouping and repair procedures

Basis-group admits the supports proposed by a pre-withdrawal constructor as groups, then fills with individual records in relevance order. *Source grouping* instead uses page, abstract, or subject-entity bundles. Its feasibility-adjusted comparison is *feasible grouping*. *Guarded basis-group* keeps the Atomic state when it already contains a complete proposed support and otherwise uses basis-group selection. The fixed-block control has identical outcomes to Atomic in the reported natural-source and eQASC panels.

The eQASC *basis-priority* selector adds a bonus of $\beta = 3$ to records belonging to proposed supports, then admits records individually. Basis-group admission instead treats a proposed support as a unit.

At a fixed budget, group membership therefore changes both which useful records enter and how much capacity their companions consume.

The *predicted-basis selector* (Appendix F.6) discovers and qualifies alternatives, selects qualified evidence, and reconstructs the state. *Basis repair* (Appendix F.8) applies qualification, budgeted replacement, and self-verification or retention of the original state to an existing memory state. These procedures test estimated support and interface-level repair under their respective evidence and cost definitions.

D.10 Candidate-pool scaling and accounting conventions

The candidate-pool scaling experiment uses 20 ProofWriter theories and three timing repeats per theory in each cell. We take the within-theory median and summarize these values across theories. Added irrelevant facts preserve the original proof structure. The implementation's exact caches reproduce the selected states of both methods on all 443 original questions.

Table D.2: Median/P90 selection time in seconds under larger candidate pools. Each entry summarizes 20 theories, with three repeats per theory. No displayed cell times out.

Candidates	K	Robust median / P90	Pair-rescue median / P90
16	180	0.0287 / 0.0433	0.0942 / 0.1610
32	180	0.0639 / 0.0971	0.3774 / 0.6642
64	180	0.1369 / 0.2072	1.3976 / 2.9584
128	180	0.2936 / 0.4361	5.3354 / 12.9103
256	180	0.5933 / 0.8957	21.6516 / 54.8011
16	512	0.0372 / 0.0611	0.1390 / 0.1925
32	512	0.0909 / 0.1470	0.6794 / 0.9392
64	512	0.2088 / 0.3255	3.0622 / 4.1733
128	512	0.4439 / 0.6881	13.2766 / 18.0659
256	512	0.9246 / 1.4320	55.8455 / 77.0576

A separate cached native-support timing test contains 20 candidate pools, three repeats, and 4,096/8,192-token caps. All pools fit both caps. Median times are approximately 0.0675 seconds for singleton search and 0.0668 seconds for Pair-rescue; P90 is approximately 0.082 and 0.390 seconds. These measurements characterize the estimated-group evaluator on nonbinding inputs; Table D.2 characterizes executable search as the candidate pool expands.

For all methods, state tokens, upstream model input/output tokens, support evaluations, and CPU latency are separate resources. Shared cached constructor output is counted once physically but remains a dependency of each method that uses it. Identical final reader prompts can share a generated response under the same model configuration. Stand-alone cost includes writing, retrieval, support estimation, and reading. Historical writer costs are reported only where logs are available; amortized comparisons require a measured per-read saving.

E Selection results and withdrawal robustness

These results support Table 1 and Section 4. We give the primary ProofWriter comparison (Appendix E.1), the completion-distance analysis (Appendix E.2), and external executable evaluations (Appendix E.3). Budget, proxy-coverage, and withdrawal-stress analyses follow in Appendices E.4 to E.6.

E.1 ProofWriter selector results

E.1.1 Counts and reviewer-specific outcomes

Among the 886 primary-pair observations per selector, Atomic yields 219 grounded reviews, Verified 222, Robust 260, and Oracle 308. Robust adds 38 grounded answers over Verified, or 4.29 percentage points. Across the three-reviewer panel, the counts are 289, 294, 336, and 407 out of 1,329, giving a 3.16-point Robust–Verified gain. These aggregates weight the indicated reviewers equally.

Table E.1: ProofWriter counts by reviewer and selector. Every row has 443 questions from 300 theories. Basis retention is a shared state property across reviewers.

Reviewer	Selector	Current correct	Grounded correct	Basis retained
Qwen	Atomic	282	110	247
	Verified	306	112	255
	Robust	307	127	366
	Oracle-basis	294	160	437
Gemma	Atomic	224	109	247
	Verified	248	110	255
	Robust	266	133	366
	Oracle-basis	252	148	437
Mistral	Atomic	284	70	247
	Verified	302	72	255
	Robust	310	76	366
	Oracle-basis	309	99	437

Table E.2: Paired grounded-review effects with 95% intervals, in percentage points. QG denotes Qwen–Gemma, M denotes Mistral, and All gives an equal-weight three-reviewer average. Each panel has a four-contrast Holm family.

Panel	Contrast	Difference	Holm p
QG	Robust–Verified	$4.3\ [2.7, 6.2]$	0.0004
	Robust–Atomic	$4.6\ [2.8, 6.6]$	0.0004
	Verified–Atomic	$0.3\ [-1.3, 1.9]$	0.7220
	Oracle–Atomic	$10.0\ [7.7, 12.6]$	0.0004
M	Robust–Verified	$0.9\ [0.2, 1.9]$	0.0936
	Robust–Atomic	$1.4\ [0.0, 2.8]$	0.1466
	Verified–Atomic	$0.5\ [-0.9, 1.8]$	0.6127
	Oracle–Atomic	$6.5\ [4.3, 9.1]$	0.0004
All	Robust–Verified	$3.2\ [2.0, 4.5]$	0.0004
	Robust–Atomic	$3.5\ [2.1, 5.1]$	0.0004
	Verified–Atomic	$0.4\ [-0.9, 1.7]$	0.6076
	Oracle–Atomic	$8.9\ [6.8, 11.1]$	0.0004

Point differences are calculated from the unrounded counts in Table E.1; intervals use paired theory-level resampling.

Per-reviewer Robust–Verified grounded-review improvements are $5.2\,[3.1, 7.6]$ points for Gemma, $3.4\,[1.7, 5.3]$ for Qwen, and $0.9\,[0.2, 1.9]$ for Mistral. Current differences are $4.1\,[1.6, 6.7]$, $0.2\,[-2.3, 2.7]$, and $1.8\,[-0.2, 4.0]$ points, respectively. The primary-panel current gain is $2.1\,[0.2, 4.1]$ points. Its lower bound satisfies the two-point non-inferiority criterion; Qwen's individual interval extends slightly below that margin.

Providing a surviving basis as a privileged reference improves primary-pair grounded review over Atomic by 10.0 percentage points. Of 443 questions, 437 retain that supplied basis, and 308/886 reviewer observations give a correct, grounded answer. This response rate applies to the specified reference-state construction, budget, and reviewers.

E.2 Withdrawal-conditioned completion distance and Pair-rescue

For a realized withdrawal D, define completion distance as the minimum number of records outside A that must be added to $A \setminus D$ to derive the target using no invalid records:

$$d(A, D) = \min_{B \subseteq H \setminus D:\, R \cup B \vdash q} |B \setminus A|. \tag{16}$$

The realized withdrawal is used only in this diagnostic. Every candidate support must avoid D. Exhaustive checking by added-record count verifies these withdrawal-conditioned distances on all 443 questions.

Table E.3: Pair-rescue–Robust retention by withdrawal-conditioned completion distance. Effects and 95% intervals are percentage points.

Distance	Questions	Theories	Retention difference [95% CI]
1	380	264	$0.00\,[0.00, 0.00]$
2	61	38	$36.07\,[20.29, 52.24]$
3	2	2	$0.00\,[0.00, 0.00]$

Original Robust retains a post-withdrawal proof on 366/443 questions and Pair-rescue on 388/443, a $4.97\,[2.33, 8.04]$-point increment. Both preserve initial support on 442/443. Mean state tokens are 168.480813 and 168.408578. The distance-two minus distance-one retention interaction is $36.07\,[20.00, 52.31]$ points. The two distance-three questions remain a sparse descriptive group.

In the same-run behavioral comparison, the original designated-basis grounding counts are 266/886 for Robust and 273/886 for Pair-rescue, a $0.79\,[0.12, 1.61]$-point difference (Holm $p = 0.02835$ in the two-test family with current non-inferiority). Current label accuracy is 572/886 versus 571/886, a $-0.11\,[-0.68, 0.33]$-point difference meeting the -2 pp boundary. Input-relative review accuracy changes by $-0.79\,[-2.99, 1.42]$ points.

The additional any-proof citation check runs the verifier on the cited records themselves. It yields 293/886 versus 298/886 grounded answers, a $0.56\,[0.11, 1.22]$-point difference. The original 25.1%/29.3% confirmation comparison uses its own generations and designated-reference scoring; the executable comparison evaluates the new paired generations. The pair comparison is secondary: it reuses the 443-question cohort of the primary comparison.

E.3 External executable validation

E.3.1 GO-Plus: source, semantics, and sampling

Formal task. The source is the pinned GO-Plus dated archive of 5 August 2026, whose embedded version IRI is dated 26 July 2026 (The Gene Ontology Consortium, 2026). We retain logical axioms in the supported OWL 2 EL fragment of the merged distribution (W3C OWL Working Group, 2012). We load the merged GO-Plus distribution and its available imports closure; the merged file contains no literal OWL import statement. One logical axiom is one selectable record. Targets are inferred, non-asserted named-class subsumptions. The proof relation is relative to this permitted formal input. ELK 0.6.0 (Kazakov et al., 2014), OWLAPI 5.1.20, and PULi 0.1.0 provide entailment and justifications. A justification is a subset-minimal set of axioms entailing the target (Kalyanpur et al., 2007). The source SHA-256 is

```
52f51e70a94667f887f89205daeb442a
be5f6f8b6cd8dda51ff5ad7fe7524bf5.
```

The renderer, local record identifiers, and separators count toward the retained-state budget. Question and answer instructions are charged to reader input separately. Logical identity remains attached to the original axiom; its local identifier is only a presentation key.

Sampling. We sample inferred subsumptions from held-out subsumee classes in depth-stratified, deterministic hash order. A target is eligible when the candidate pool contains a complete justification surviving the withdrawal. The budget $K = 240$ is fixed on a disjoint development split. The original sample supplies 78 eligible targets. Before computing any held-out method outcomes, we extend the sample by 222 eligible targets in the same fixed order, reaching 300 targets from 296 classes, with at most two per class. The eligibility criteria, budget, and selection algorithms remain fixed. Development and engineering classes are excluded throughout. This is a held-out evaluation conditional on surviving-support eligibility.

For source selection, the accompanying FoodOn screen yields 99/510 development targets with surviving alternatives (19.41%), below the prespecified 20% threshold. FoodOn is therefore excluded from the treatment comparison. GO-Plus supplies 223 eligible targets from 5,700 additional candidates; the first 222 complete the panel. These qualification counts record the yield of the fixed candidate-sampling procedure.

Shared information and construction. All four methods receive the same pool H of at most 30 axioms, active support A, frozen relevance scores, exact serializer, budget, and executable verifier. Atomic starts from no selected axioms. Verified preserves A and fills by relevance. Robust uses the original singleton-greedy survival search; Pair-rescue adds the pair check in Algorithm 2. The same prespecified proxy is used within the comparison. Each state is frozen before the evaluator-only singleton withdrawal is supplied. The realized withdrawal, surviving reference justification, and reader outputs enter the subsequent evaluation stage.

Every test target has a complete H justification after withdrawal. Thus the reported rates condition on structural eligibility. Joint feasibility of the active proof and a surviving alternative is checked by $c(A \cup B) \leq 240$.

E.3.2 GO-PLUS ABSOLUTE OUTCOMES AND PAIRED EFFECTS

Table E.4: GO-Plus absolute counts on 300 targets. Structural counts have denominator 300; answer and grounded-answer counts have denominator 600, comprising two reviewers per target.

	Complete support		Correct answer		Grounded
Method	Current	After D	Current	After D	After D
Atomic	214	75	474	250	106
Verified	300	96	586	324	122
Robust	300	98	585	327	128
Pair-rescue	300	248	589	501	377

Table E.5: Primary GO-Plus Pair-rescue–Verified comparisons. Effects and intervals are percentage points. Wins, losses, and ties refer to target-level paired outcomes after averaging the two reviewers.

Outcome	Difference [95% CI]	Wins/losses/ties	Holm p
Post-withdrawal support	$50.67\ [44.93, 56.33]$	152/0/148	0.00015
Post-withdrawal grounding	$42.50\ [37.25, 47.83]$	144/0/156	0.00015
Current answer, NI at -2	$0.50\ [-0.50, 1.67]$	5/3/292	0.00015

Each raw directional centered-bootstrap estimate is 0.00005. The current-accuracy test is a non-inferiority test against the -2 pp boundary.

The three-test primary family uses 20,000 paired subsumee-cluster bootstrap draws over 296 groups. Both reviewers and all methods remain paired within each target. Point estimates give equal target

weight; clusters account for repeated subsumee classes. Intervals are two-sided percentile 95% intervals. Directional centered-bootstrap tests use the declared superiority and non-inferiority boundaries, followed by Holm correction across the three tests. Shared ontology content can induce dependencies beyond these target groups, so inference describes this pinned ontology and sampling design.

Secondary post-withdrawal retention differences are $+0.67\,[0, 1.67]$ points for Robust–Verified and $+50.00\,[44.26, 55.70]$ for Pair-rescue–Robust. Pair-rescue retains support on 64/78 original targets and 184/222 accrued targets. These counts show retention in the original and accrued parts of the common held-out panel.

Opportunity and evidence use. All 204 Verified losses have a surviving minimal justification B with exact cost $c(A \cup B) \leq 240$. Pair-rescue recovers 152, misses 52, and never destroys a Verified-retained post-withdrawal support. It admits a pair on 208 targets, with 224 pair additions and 51 subsequent singleton additions across the panel. These counts describe the pair admissions and subsequent single-record additions used to construct the selected states.

Of Pair-rescue's 496 supported reviewer inputs, 377 give grounded answers, or 76.0%; 119 remain unsuccessful. Verified gives 122/192, or 63.5%, on its own retained inputs. Each percentage measures evidence use within that selector's retained population. Post-withdrawal abstentions total 238/600 for Verified and 86/600 for Pair-rescue. Unsupported SUPPORTS assertions total 202/600 and 124/600, respectively.

Independent support and citation checks. Direct ELK evaluates exactly the selected or cited axioms, with no omitted pool axioms supplied. It rechecks 1,104 distinct current/post-withdrawal selected sets representing 2,400 logical state-phase checks, finding zero support-status mismatches or verifier failures. It also rechecks 906 distinct actual citation sets spanning 4,302 citation-valid logical rows, finding zero sufficiency mismatches. The remaining 498 rows have invalid output or withdrawn/invalid citations and remain nongrounded. No grounded answer occurs on a structurally unsupported post-withdrawal state.

Qwen has 180 invalid JSON/label parses among its 2,400 logical method–phase rows; Gemma has none. Exact one-prefix normalization of `RECORD` changes no citation-validity row. Valid wrong or abstaining outputs are never retried. Output-cap terminations occur on 81 unique Qwen requests and 29 unique Gemma requests; reaching a cap is recorded separately from final-answer validity.

Compute. The logical reader matrix contains 4,800 uses. Exact model-plus-full-prompt reuse yields 2,208 physical requests, 1,104 per reviewer. Qwen consumes 485,082 input and 52,459 output tokens; Gemma consumes 452,606 and 40,073. The combined total is 937,688 input and 92,532 output tokens. Logged request spans are approximately 592 seconds per model on separate A100 services, with the two spans overlapping.

Original Robust takes 9.22 summed CPU seconds over 300 states and 312,694 exact support evaluations. Pair-rescue takes 97.95 summed seconds, evaluates 183,472 pair candidates and 6,046,657 support checks, and performs 207,449 exact serialized-cost checks. Its median/P90 state-selection time is 0.343/0.453 seconds. Evaluations of the available exact support relation are distinguished from the independent direct-reasoner validation above. All four methods retain approximately 231 tokens on average, and every state satisfies the 240-token limit.

E.3.3 CLUTRR-DERIVED RELATION REASONING

We use a controlled-language adaptation of the public CLUTRR relation system (Sinha et al., 2019). The task uses fixed relation templates with explicit story-to-fact mappings. Of 600 generated families, 557 satisfy the structural eligibility checks; 50 groups are reserved for development and 300 independent families for evaluation. Both budgets were fixed from structural statistics on the 50 development groups, before any test-set reviewer evaluation. A coarse grid from 256 to 2,048 tokens left no budget at which the active support fit while selection pressure remained, so we refined the grid between 768 and 1,024 tokens. The primary budget, $K = 920$, is the only refined-grid value at which at least 70% of development groups can accommodate the active support together with a surviving alternative while at least 50% still require deleting candidate records. The adjacent tighter value, $K = 912$, was fixed at the same time as a budget-sensitivity condition. Test completion distances

are one, two, and three in six, 293, and one families, respectively. These counts characterize the completion requirements in the generated evaluation families.

Table E.6: CLUTRR-derived evaluation. Retention counts have denominator 300, and executable grounded-answer counts have denominator 600. Current accuracy averages two reviewers.

K	Method	Retained	Grounded	Current (%)
912	Atomic	154	61	39.17
	Active-preserving relevance	154	61	42.17
	Robust	154	61	42.17
	Pair-rescue	174	69	42.83
920	Atomic	203	72	40.33
	Active-preserving relevance	203	72	42.17
	Robust	203	72	42.17
	Pair-rescue	237	79	42.67

Pair-rescue–Robust retention differences are $6.67\,[4.00, 9.67]$ and $11.33\,[7.67, 15.00]$ points at the two budgets; executable grounded-answer differences are $1.33\,[0.50, 2.50]$ and $1.17\,[0.33, 2.17]$. The current-accuracy differences, $0.67\,[0.17, 1.33]$ and $0.50\,[-0.67, 1.83]$, meet the declared -2 pp boundary. The six-test family includes these three outcomes at both budgets. Paired family-cluster bootstrap uses 20,000 draws. Holm-adjusted grounded-review p-values are 0.0299 at both budgets; structural p-values are 0.0003.

Verified and Robust have equal structural and behavioral aggregate scores at each budget. Atomic has the same post-withdrawal aggregates but different current accuracy. All net grounded gains arise in the 20 and 34 questions that transition from missing to retained support. The common-retention group's net contribution is zero and there are no support-destroying transitions. In the new-retention groups, grounded use is 20.0% and 10.3%, while abstention is 25.0% and 22.1%. The remaining failures include both abstentions and answers without an executable citation certificate.

A separate controlled-length evaluation uses 120 of these families. Native, approximately 4K, and approximately 16K histories all pass through the same 32-record candidate stage at $K = 920$. Full history and retrieved candidates preserve support on all 120 cases; Robust retains 86 and Pair-rescue 98 at every length. This isolates a downstream selection loss after successful candidate acquisition. Of 480 case–method combinations, 312 have identical selected text across all lengths. Seven repeated prompt/model groups change grounding across executions. Their variation is recorded separately from input-history effects.

E.4 Five-budget comparison

The secondary sweep evaluates 443 questions from 300 theories with the same Qwen–Gemma pair at $K \in \{120, 150, 180, 210, 240\}$. Each cell has 886 observations. At $K = 180$, state text and withdrawal match the primary comparison, while the separate generation gives 222, 223, and 261 grounded answers for Atomic, Verified, and Robust, versus 219, 222, and 260 in the primary run. The Robust–Verified difference is 38/886 in both runs. All budget calculations consistently use the sweep outputs.

Temperature zero fixes greedy token choice, but numerical and serving configurations can still affect exact outputs (vLLM Project, 2026). The recorded inputs establish the state match; intervals condition on the saved generations. Tables E.7 and E.8 give all arm means and paired contrasts. Figure 6 displays the Robust–Verified retention and grounded-review contrasts from the latter table, using differences calculated before display rounding. Holm correction applies separately to the six tests at each budget, as specified in Appendix A.3.

Retention gains occur at every budget. Grounded-review gains over Verified pass the within-budget correction throughout; those over Atomic pass through $K = 210$, with the $K = 240$ interval including zero.

Selection pressure decreases as more of the pool fits in memory (Table E.9). The share of pools not fully retained falls from 94.58% at $K = 120$ to 55.53% at $K = 240$. At the lowest budget, 112 questions have rules longer than the nominal cap; their actual rules-only costs remain in the

Table E.7: Secondary budget sweep on 443 questions from 300 theories, with two reviewers and 886 observations per row. Accuracy and retention are percentages; actual cost is the mean serialized state length in tokens.

K	Selector	Basis retention	Current accuracy	Grounded review	Actual tokens
120	Atomic	26.86	31.83	17.83	117.76
	Verified	26.41	38.15	16.48	117.81
	Robust	38.60	39.73	22.23	117.85
150	Atomic	39.05	42.78	22.12	141.82
	Verified	38.83	52.03	21.22	141.81
	Robust	58.01	54.40	26.07	141.91
180	Atomic	55.76	57.00	25.06	168.33
	Verified	57.56	62.53	25.17	168.48
	Robust	82.62	65.01	29.46	168.48
210	Atomic	73.36	63.43	30.14	193.00
	Verified	79.23	65.91	30.59	193.24
	Robust	94.13	67.61	32.84	193.19
240	Atomic	88.26	65.69	31.83	213.12
	Verified	90.74	66.48	30.93	213.16
	Robust	96.84	66.93	32.17	213.05

The primary confirmation and secondary sweep use the same questions and reviewer pair but separate generations. Budget curves and equivalent-budget estimates consistently use the sweep outputs. Actual-token means include the rules-only floor on infeasible questions.

Table E.8: Complete paired contrasts in the secondary budget sweep. Each cell gives Robust minus the named baseline in percentage points, a paired 95% interval, and the Holm-adjusted p-value.

K	Baseline	Retention	Current accuracy	Grounded review
120	Atomic	$+11.74$ $[7.53, 16.16]$ $p_{\mathrm{H}} = 0.0006$	$+7.90$ $[5.23, 10.90]$ $p_{\mathrm{H}} = 0.0006$	$+4.40$ $[1.84, 7.09]$ $p_{\mathrm{H}} = 0.0016$
	Verified	$+12.19$ $[8.79, 15.95]$ $p_{\mathrm{H}} = 0.0006$	$+1.58$ $[0.46, 2.80]$ $p_{\mathrm{H}} = 0.0050$	$+5.76$ $[3.78, 7.97]$ $p_{\mathrm{H}} = 0.0006$
150	Atomic	$+18.96$ $[14.00, 24.24]$ $p_{\mathrm{H}} = 0.0006$	$+11.63$ $[8.14, 15.27]$ $p_{\mathrm{H}} = 0.0006$	$+3.95$ $[1.79, 6.19]$ $p_{\mathrm{H}} = 0.0008$
	Verified	$+19.19$ $[14.47, 24.31]$ $p_{\mathrm{H}} = 0.0006$	$+2.37$ $[1.01, 3.86]$ $p_{\mathrm{H}} = 0.0008$	$+4.85$ $[3.17, 6.70]$ $p_{\mathrm{H}} = 0.0006$
180	Atomic	$+26.86$ $[20.95, 32.87]$ $p_{\mathrm{H}} = 0.0006$	$+8.01$ $[5.19, 10.97]$ $p_{\mathrm{H}} = 0.0006$	$+4.40$ $[2.52, 6.44]$ $p_{\mathrm{H}} = 0.0006$
	Verified	$+25.06$ $[19.77, 30.54]$ $p_{\mathrm{H}} = 0.0006$	$+2.48$ $[0.58, 4.39]$ $p_{\mathrm{H}} = 0.0106$	$+4.29$ $[2.67, 6.13]$ $p_{\mathrm{H}} = 0.0006$
210	Atomic	$+20.77$ $[15.37, 26.29]$ $p_{\mathrm{H}} = 0.0006$	$+4.18$ $[2.05, 6.36]$ $p_{\mathrm{H}} = 0.0006$	$+2.71$ $[1.15, 4.34]$ $p_{\mathrm{H}} = 0.0014$
	Verified	$+14.90$ $[10.05, 19.87]$ $p_{\mathrm{H}} = 0.0006$	$+1.69$ $[0.33, 3.11]$ $p_{\mathrm{H}} = 0.0229$	$+2.26$ $[1.03, 3.68]$ $p_{\mathrm{H}} = 0.0006$
240	Atomic	$+8.58$ $[5.30, 12.12]$ $p_{\mathrm{H}} = 0.0006$	$+1.24$ $[-0.23, 2.74]$ $p_{\mathrm{H}} = 0.3342$	$+0.34$ $[-0.89, 1.65]$ $p_{\mathrm{H}} = 0.6701$
	Verified	$+6.09$ $[3.17, 9.39]$ $p_{\mathrm{H}} = 0.0006$	$+0.45$ $[-0.12, 1.10]$ $p_{\mathrm{H}} = 0.4100$	$+1.24$ $[0.45, 2.23]$ $p_{\mathrm{H}} = 0.0020$

Twenty thousand paired theory-cluster draws preserve both readers and all methods for each sampled theory. At each K, Holm correction covers the six-test family of two baselines and three outcomes.

token means and their retention is zero under every selector. The 331 rules-fit questions give a Robust–Atomic retention difference of $15.71\,[10.24, 21.43]$ points (Table E.10). This sensitivity subset complements the common 443-question denominator.

Table E.9: Pool coverage and rules-only infeasibility in the budget sweep. Percentages have the common denominator of 443 questions and coincide across the three selectors.

K	Pool not fully retained (%)	Rules-only infeasible (%)
120	94.58	25.28
150	90.29	0.00
180	85.55	0.00
210	71.78	0.00
240	55.53	0.00

Pool not fully retained denotes omission of at least one candidate record or rules-only infeasibility. It measures pool coverage under the admission procedure.

Table E.10: Sensitivity analysis at $K = 120$ on the rules-fit subset: 331 questions, 662 observations, and 240 theories. Values are percentages.

Selector	Basis retention	Current accuracy	Grounded review
Atomic	35.95	42.60	23.87
Verified	35.35	51.06	22.05
Robust	51.66	53.17	29.76

The remaining 112 questions have zero retention under all selectors. The main sweep keeps all 443 questions. Complete-support feasibility additionally includes the required fact records and their formatting.

For an outcome y, equivalent budgets interpolate between adjacent baseline points that bracket the treatment value:

$$K_{\text{eq}} = K_0 + (K_1 - K_0)\frac{y_{\text{Robust},180} - y_{\text{Verified},K_0}}{y_{\text{Verified},K_1} - y_{\text{Verified},K_0}}. \tag{17}$$

Using unrounded sweep means gives 218.82 tokens for retention, interpolating between $K_0 = 210$ and $K_1 = 240$, and 203.75 for grounded review, interpolating between 180 and 210. These are outcome-specific interpolations between nominal budget limits.

Let c denote mean actual retained-state tokens, a current accuracy, and r and G the retention and grounded-review means. For either outcome pair (r, G) or (a, G), a baseline dominates a Robust operating point when its c is no larger and neither outcome is lower, with at least one strict improvement across cost and outcomes. No Atomic or Verified point dominates a Robust point under either comparison in the measured grid. The comparison covers the 15 measured operating points. Robust at $K = 210$ has higher current accuracy and grounded review than Robust at $K = 240$, whereas the larger state has higher retention.

E.5 Structural coverage of proxy withdrawals

We examine how the realized withdrawal relates to the scenarios used for survival scoring. The analysis reuses the 443 ProofWriter questions and the original retained states at $K = 180$. Qwen, Gemma, and Mistral provide 1,329 question–reviewer observations. The selector, relevance scores, and proxy family remain fixed.

E.5.1 Exact sets and effective withdrawal cores

A withdrawal's *survival signature* records which evaluated minimal supports for the claim remain intact. A *minimal effective core* is an inclusion-minimal subset of the withdrawal with the same signature. Removed records outside such a core leave the survival of these supports unchanged. We assign each question to the first applicable category: *exact* when the realized set equals a proxy scenario; *effective core* when a minimal core equals a scenario; *signature only* when a scenario has the same signature without either preceding match; and *uncovered* otherwise.

Every singleton withdrawal is in the proxy family, so the 284 exact matches follow from its construction. The other 159 realized sets contain multiple records. Effective-core matching adds 144 questions, giving $428/443 = 96.6\%$ exact-or-core coverage; one further question is signature-equivalent and 14 are uncovered. Coverage is measured relative to the evaluated minimal supports after the original selections are fixed.

E.5.2 Retention and grounded review by coverage

Robust improves retention over Verified by 25.4 points under exact matching and 25.7 points under effective-core matching (Table E.11). Grounded-review gains are 3.5 and 2.8 points, respectively, with positive intervals in both groups. Each contrast pairs selectors within its coverage group.

Table E.11: **Retention and grounded-review gains occur under both exact and effective-core coverage.** Upper panel: Robust minus Verified on the original three-reviewer evaluation at $K = 180$. Lower panel: the four-test coverage family. Effects and 95% intervals are percentage points.

Coverage	Questions	Theories	Retention gain [95% CI]	Grounded gain [95% CI]
Exact	284	196	$+25.4$ $[+18.6, +32.2]$	$+3.5$ $[+1.9, +5.5]$
Effective core	144	99	$+25.7$ $[+16.9, +34.9]$	$+2.8$ $[+1.2, +4.7]$
Signature only	1	1	0.0 $[0.0, 0.0]$	0.0 $[0.0, 0.0]$
Uncovered	14	11	$+14.3$ $[0.0, +37.5]$	0.0 $[0.0, 0.0]$

Contrast	Difference [95% CI]	Raw p	Holm p
Uncovered: retention	$+14.3$ $[0.0, +37.5]$	0.2284	0.6852
Uncovered: grounded review	0.0 $[0.0, 0.0]$	1.0000	1.0000
Coverage interaction: retention	-11.12 $[-27.32, +12.58]$	0.2976	0.6852
Coverage interaction: grounded	-3.26 $[-4.66, -2.03]$	< 0.0001	< 0.0001

Each question contributes three reviewer observations; retention is shared across reviewers. Theories can contribute questions to more than one group. Exact and effective-core contrasts have raw $p < 0.0001$ for both outcomes; signature-only contrasts have $p = 1$. The coverage interaction subtracts the Robust–Verified effect in all 429 represented questions from the effect in the 14 uncovered questions. Intervals use 20,000 paired theory-cluster bootstrap draws. A $[0, 0]$ interval reflects no variation in the observed paired differences.

On the 14 uncovered questions, Verified retains a basis on 28.6% and Robust on 42.9%. Neither yields a grounded answer. The retention contrast is $14.3\,[0.0, 37.5]$ points, with Holm $p = 0.6852$; the uncovered-minus-covered retention interaction is $-11.12\,[-27.32, 12.58]$. The grounded-review interaction is $-3.26\,[-4.66, -2.03]$ points: the represented group gains grounded answers while the uncovered group has none. The evidence requirements differ across these groups, as detailed next.

E.5.3 Evidence requirements in uncovered questions

The uncovered questions have larger effective withdrawal cores and more initial proofs (Table E.12). Mean core size is 2.07, compared with 1.00 in both exact and effective-core groups; mean proof count is 5.14, compared with 2.26 and 2.28. Their active supports, alternative supports, and candidate pools also cost more. These measures describe several related aspects of the same difference in problem structure.

One uncovered example is `PW|RelNoneg-OWA-D2-645|82ff3bd1`. The claim has three independent proofs, and the realized withdrawal removes three facts. Its effective core is $\{$`F01`, `F05`$\}$; `F02` is inert padding. The core combines an active fact with a non-active fact. The proxy family draws pairs only from the active basis and excludes this mixed pair. No proxy scenario has the same survival signature for this case. This example locates a specific coverage boundary in the permitted combinations of withdrawn facts.

E.5.4 Coverage across state budgets

The secondary sweep uses Qwen and Gemma on the same 443 questions, with 28 reviewer observations from the 14 uncovered questions and 858 from the 429 represented questions at every budget. It uses separate generations from the original three-reviewer comparison. Table E.13 reports paired retention and grounded-review differences within each group.

Table E.12: **Structurally uncovered questions require larger withdrawals and more evidence.** Group characteristics on the same 443 questions. Sizes and proof counts are means; costs are mean token counts for the indicated evidence sets.

Coverage	Questions	Raw withdrawal size	Effective-core size	Initial proofs
Exact	284	1.00	1.00	2.26
Effective core	144	2.06	1.00	2.28
Signature only	1	3.00	2.00	3.00
Uncovered	14	2.71	2.07	5.14

Coverage	Active-support cost	Alternative-support cost	Candidate-pool cost
Exact	118.4	131.8	234.7
Effective core	117.8	132.4	248.1
Signature only	158.0	172.0	290.0
Uncovered	133.6	147.8	275.9

Costs measure the indicated supports and full candidate pools. Selected-state lengths are reported separately.

Table E.13: **The represented group retains positive gains across five budgets.** Robust minus Verified in the secondary Qwen–Gemma sweep. Represented combines exact, effective-core, and signature-only coverage. Effects and paired 95% intervals are percentage points.

K	Coverage	Retention gain [95% CI]	Grounded gain [95% CI]
120	Represented	$+12.6$ $[+9.0, +16.5]$	$+5.9$ $[+3.9, +8.2]$
	Uncovered	0.0 $[0.0, 0.0]$	0.0 $[0.0, 0.0]$
150	Represented	$+19.8$ $[+15.0, +24.9]$	$+5.0$ $[+3.3, +7.0]$
	Uncovered	0.0 $[0.0, 0.0]$	0.0 $[0.0, 0.0]$
180	Represented	$+25.4$ $[+20.1, +30.9]$	$+4.4$ $[+2.8, +6.4]$
	Uncovered	$+14.3$ $[0.0, +37.5]$	0.0 $[0.0, 0.0]$
210	Represented	$+15.2$ $[+10.3, +20.3]$	$+2.2$ $[+1.0, +3.7]$
	Uncovered	$+7.1$ $[0.0, +25.0]$	$+3.6$ $[0.0, +12.5]$
240	Represented	$+6.3$ $[+3.3, +9.7]$	$+1.3$ $[+0.5, +2.3]$
	Uncovered	0.0 $[0.0, 0.0]$	0.0 $[0.0, 0.0]$

The intervals are coverage-specific; the full-cohort six-test Holm families are reported separately in Appendix E.4. The 180-token retention differences agree across the two evaluations because they use the same retained states. Their grounded-review comparisons use the reviewer panel and generations stated for each evaluation.

The represented group has positive retention and grounded-review intervals at every budget. In the uncovered group, grounded-review gains are zero at four budgets and $3.6\,[0.0, 12.5]$ points at 210 tokens. These estimates describe 14 questions with consistently greater evidence requirements. Their evidence requirements and the complete group-specific intervals are reported alongside the covered group.

E.6 Frozen-state withdrawal stress

The withdrawal stress analysis reuses frozen executable states and changes the evaluator's withdrawal conditions. In the reported aggregate comparisons, Pair-rescue exceeds Verified by 5.18 percentage points under random withdrawals and 14.69 points under mixed active/non-active withdrawals. These are structural retention effects. The evaluator applies the stress withdrawals after selection is complete.

A mixed multi-record withdrawal can be absent as a literal set from the construction proxy while still sharing an effective core or support-survival signature with a proxy member. The reported aggregates measure retention under these random and mixed withdrawal conditions. Appendix E.5 separately reports exact, effective-core, and structural coverage for the original realized withdrawals. A withdrawal that also destroys every full-history support has a different interpretation from a memory-induced loss in an otherwise answerable history. The revision-specific subset is $r(H, D) = 1$ and $r(A, D) = 0$. The table below reports aggregate point differences; Appendix E.5 gives coverage-stratified effects with paired inference.

Table E.14: Descriptive support-retention differences under alternative withdrawals applied to frozen states. Values are Pair-rescue minus Verified in percentage points.

Withdrawal condition	Retention difference
Random withdrawal	$+5.18$
Mixed active/non-active withdrawal	$+14.69$

F Evidence use, argument discovery, and memory repair

This section follows Section 5. We first trace retained arguments into answers (Appendix F.1), distinguish the compared question populations (Appendix F.2), and test showing the model a verified basis (Appendix F.3). Appendix F.4 gives descriptive aggregate relationships. The remaining comparisons examine argument discovery and basis repair in text (Appendices F.5–F.8).

F.1 Converting retained support into grounded answers

Table F.1: Paired retention and evidence use across three reviewers. Counts are observations out of 1,329; the grounded columns give within-cell success percentages.

Retention pattern	Count	Share (%)	Robust grounded review	Verified grounded review
Both retain	762	57.3	38.6	38.6
Robust only	336	25.3	12.5	0.0
Verified only	3	0.2	0.0	0.0
Both lose	228	17.2	0.0	0.0

The paired retention groups contain 254 questions where both selectors retain support, 112 where only Robust does, one where only Verified does, and 76 where neither does. Repeating each state across three reviewers gives the observation counts in Table F.1. Both selectors attain 38.6% grounded review in the shared group; this equality is at the group-mean level and permits different answers on individual observations.

Of the 336 Robust-only observations, 42 yield a correct, grounded review, 11.3% assert a label without an accepted complete citation basis, and 73.8% return INSUFFICIENT. On these same questions, Verified abstains in 87.8% of observations and asserts SUPPORTS in 10.4%. The additional support recovers 42 grounded answers, with abstention remaining the most common response.

The paired comparison identifies where the added grounded answers arise: 42 occur in the Robust-only group, and the shared group's mean is unchanged. Conditional use is 12.5% in the newly retained group and 38.6% in the shared group. These groups contain different questions. The difficulty analysis in Appendix F.2 uses a separate Atomic–Robust comparison.

F.2 Case difficulty and matched retained states

We compare Robust's retained cases at $K = 180$ according to whether Atomic also preserves a basis. The newly retained group has deeper proofs, larger active and alternative supports, and larger evidence pools (Table F.2). Its lower current accuracy and grounded review accompany these greater evidence requirements.

Table F.2: Additional retention reaches larger and deeper problems at $K = 180$. Shared denotes the baseline-retained group; Robust-only denotes cases additionally retained by Robust. Values are descriptive summaries of the two groups.

Property	Shared	Robust-only
Proof depth	2.61	3.40
Active-basis cost (tokens)	98.90	133.53
Alternative-basis cost (tokens)	99.16	133.55
Candidate-pool cost (tokens)	207.83	272.43
Current accuracy (%)	72.6	50.4
Grounded review given retention (%)	46.0	17.8

Both groups retain a complete basis under Robust, but they contain different questions. The Atomic–Robust groups contain 237 shared and 129 newly retained questions. Active and alternative costs are computed independently, including rules and formatting; integer rounding gives 99/134 for both rows. The three-reviewer comparison in Table F.1 uses Verified–Robust groups and distinct conditional summaries.

The budget-wise use rates average over these different retained populations. To compare representations on common questions, we restrict attention to cases where both selectors retain support. Verified and Robust usually produce identical states in that subset (Table F.3). Robust and Atomic provide more distinct states, though the number of such pairs remains small.

Table F.3: State overlap in conditional comparisons. Verified–Robust identity rates condition on both selectors retaining a basis. The final column counts distinct-state pairs in the Robust–Atomic comparison.

K	Verified–Robust identical states (%)	Robust–Atomic distinct-state pairs
120	99.1	11
150	100.0	20
180	99.2	43
210	99.1	65
240	100.0	57

At $K = 180$, jointly retained Verified–Robust states differ in 0.8% of cases; Robust–Atomic supplies 43 distinct-state pairs. The newly retained Atomic–Robust group has deeper proofs and larger supports. The Verified–Robust state-identity rates concern its separately defined shared-retention group. These comparisons distinguish expanded answerability from changes in how the same question is presented.

F.3 Verifier-assisted presentation and a fixed check instruction

On a separate 100-theory panel, the system uses its executable verifier after withdrawal to identify a complete support within the retained state, then explicitly presents the corresponding record IDs to the reader. Compared with ordinary reading, this basis-ID interface improves grounded answering by $28.5\,[19.5, 37.5]$ percentage points. The presented IDs identify a support computed entirely from the available evidence. Finding the basis requires verifier computation in addition to memory selection; the comparison tests how explicitly identifying retained support affects the model's answer.

The contrast estimates the effect of the complete verifier-assisted presentation procedure on a common panel. Support is computed using the retained state and the revealed withdrawal, after selection has finished. The verifier has already established entailment when it returns a basis. The intervention supplies the entailment result and the supporting record identities, measuring the benefit of this verified information at reading time.

A different, fixed same-state comparison uses 100 ProofWriter states: 50 newly retained, 25 jointly retained, and 25 unsupported controls. Both conditions use JSON-only structured output, the same evidence, and no visible reasoning requirement. Among the 50 newly retained states, direct reading grounds 8/50 answers and a check-first instruction grounds 11/50, a $6.0\,[0.0, 14.0]$-point difference with two-sided $p = 0.1209$. Both conditions parse successfully on all 100 inputs. These experiments distinguish a normal-format instruction change from the additional verified information supplied by basis presentation.

F.4 Depth-stratified retention and review

Figure 5 groups the retention–review operating points by the displayed proof-depth strata $d \in \{1, 2, 3, 5\}$. Marker shapes distinguish Atomic, Verified and Robust. The upper-right panel subtracts Atomic within the matched setting; the lower panels show each depth group separately. The fitted slopes, intercepts and R^2 values summarize the plotted aggregate points. They are descriptive associations across operating conditions, while the paired theory-level tests use the saved question outcomes. Proof depth is a property of the underlying derivation; the withdrawal-conditioned completion distance in Appendix E.2 counts records needed beyond the current proof.

F.4.1 Conditional use across budgets

All grounded-review successes occur among basis-retaining observations. Table F.4 divides those successes by the retained count to obtain conditional use. Since different selectors retain different questions, these percentages describe different populations. Appendix F.2 examines their difficulty and the remaining state variation on jointly retained questions.

Table F.4: Conditional use in the secondary budget sweep, calculated from raw observation counts. Use and 95% intervals are percentages.

K	Selector	Basis retained	Grounded successes	Conditional use	95% CI
120	Atomic	238	158	66.39	$[58.70, 73.55]$
	Verified	234	146	62.39	$[54.09, 70.24]$
	Robust	342	197	57.60	$[50.92, 63.99]$
150	Atomic	346	196	56.65	$[49.42, 63.76]$
	Verified	344	188	54.65	$[47.21, 62.08]$
	Robust	514	231	44.94	$[38.91, 50.99]$
180	Atomic	494	222	44.94	$[38.70, 51.22]$
	Verified	510	223	43.73	$[37.70, 50.00]$
	Robust	732	261	35.66	$[30.59, 40.97]$
210	Atomic	650	267	41.08	$[35.80, 46.59]$
	Verified	702	271	38.60	$[33.43, 44.13]$
	Robust	834	291	34.89	$[30.14, 39.95]$
240	Atomic	782	282	36.06	$[31.11, 41.15]$
	Verified	804	274	34.08	$[29.21, 39.06]$
	Robust	858	285	33.22	$[28.57, 38.04]$

Conditional use is the number of grounded-review successes among basis-retaining observations divided by the number of basis-retaining observations. Intervals use 20,000 theory-cluster bootstrap draws. Each selector is conditioned on its own retained subset.

F.5 Group selection and budget effects across sources

F.5.1 Primary per-case budgets

Table F.5 reports the per-case-budget comparison; Tables F.6 and F.7 report source-level fixed-budget comparisons.

Table F.5: Constructor-based selection under per-case budgets, $\alpha = 0.35$ in Eq. (6). Percentages. Certificate retention uses each panel's designated reference bases.

Source	Selector	Current	Review	Grounded	Certificate
FEVEROUS	Atomic	81.67	75.00	52.22	96.67
	Basis-group	82.22	73.33	52.50	91.67
	Full-History	86.11	75.56	53.89	100.00
RuleTaker	Atomic	92.22	55.93	33.70	64.44
	Basis-group	94.44	64.81	40.74	82.96
	Full-History	95.93	75.56	50.74	100.00
SciFact	Atomic	96.30	90.74	59.26	96.30
	Basis-group	94.44	88.89	61.11	96.30
	Full-History	94.44	92.59	64.81	100.00

The group comparison asks whether admitting a proposed support jointly improves on individual record admission. Its certificate column requires complete containment of one designated reference basis, and grounding uses the same family. In RuleTaker, another executable proof may survive when this designated certificate is incomplete. The any-proof occurrence results in Appendix B.2 therefore answer a broader structural question.

At the primary per-case budget, basis-group changes review over Atomic by $+8.89\,[2.96, 15.19]$ points on RuleTaker, $-1.67\,[-3.61, 0.00]$ on FEVEROUS, and $-1.85\,[-5.77, 0.00]$ on SciFact. Holm-adjusted p-values in the six-contrast family are 0.0128, 0.3015, and 1.0, respectively. Full history indicates the information available before selection and carries a higher state cost.

Oversized source bundles account for much of the low retention under page or abstract grouping. Although such bundles contain useful records, they can fail the budget check as a unit: utilization is 21% on FEVEROUS and 72% on SciFact. Feasible grouping uses 96–99% of the budget and raises reference-basis retention to 93.9% and 88.9%, respectively. These comparisons show that the feasibility of the admission unit matters alongside its evidence content.

F.5.2 Fixed source budgets

Table F.6: Constructor-based selection under fixed source budgets, evaluated on all cases. Percentages. Certificate retention uses the designated reference bases.

Source	Selector	Current	Review	Grounded	Certificate
FEVEROUS	Atomic	81.67	70.00	46.11	79.44
	Basis-group	79.72	65.56	43.06	67.22
	Guarded groups	80.83	66.39	44.17	76.11
RuleTaker	Atomic	90.00	50.74	30.37	55.56
	Basis-group	90.00	58.15	35.93	71.11
	Guarded groups	90.00	56.30	32.96	67.41
SciFact	Atomic	94.44	90.74	44.44	74.07
	Basis-group	94.44	68.52	44.44	66.67
	Guarded groups	96.30	72.22	46.30	74.07

Tables F.6 and F.7 use the budgets chosen on development data in Table A.3. They include every evaluation case and separately report the subset with a budget-feasible reference basis. Guarded groups use the basis-group procedure and retain the Atomic state when it already contains a complete proposed support. In the feasible-subset family, its review differences are $-2.47\,[-4.94, -0.61]$

Table F.7: Review differences from Atomic under fixed source budgets, in percentage points. Holm adjustment uses the three feasible-subset comparisons separately for basis-group and guarded basis-group.

Source	Selector	All cases	Feasible cases	Feasible Holm p
FEVEROUS	Basis-group	$-4.44\ [-8.06, -1.11]$	$-3.70\ [-7.14, -0.61]$	0.0656
	Guarded groups	$-3.61\ [-6.39, -1.38]$	$-2.47\ [-4.94, -0.61]$	0.0408
RuleTaker	Basis-group	$7.41\ [1.11, 13.70]$	$9.13\ [1.74, 16.52]$	0.0471
	Guarded groups	$5.56\ [0.37, 10.74]$	$6.96\ [1.30, 13.04]$	0.0458
SciFact	Basis-group	$-22.22\ [-43.33, -4.17]$	$-8.70\ [-20.00, 0.00]$	0.0727
	Guarded groups	$-18.52\ [-40.62, 0.00]$	$-4.35\ [-14.29, 0.00]$	0.7083

points on FEVEROUS, $+6.96\,[1.30, 13.04]$ on RuleTaker, and $-4.35\,[-14.29, 0.00]$ on SciFact, with Holm $p = 0.0408$, 0.0458, and 0.7083.

F.5.3 Relevance and proposed groups

Relevance scores rank reference-basis membership more accurately on FEVEROUS and SciFact than on RuleTaker, with AUCs of 0.896, 0.851, and 0.720. Proposed groups contain the reference basis in 63%, 85%, and 48% of cases; median per-record costs occupy 17.4%, 14.6%, and 12.8% of the tight budgets. These correlated source properties are consistent with stronger individual relevance selection on the two natural-evidence tasks.

A structural sweep on 200 held-out cases varies the available budget over eight fractions. Atomic and basis-group differ by at most one retained reference basis at each fraction, including 142/200 for both at the primary fraction 0.35. State-level disagreement can therefore coexist with nearly equal revision-basis retention.

F.5.4 eQASC relevance bonuses and survival transfer

On the 375-question eQASC ordinary-selection panel, basis-priority raises grounded review over Atomic by $3.33\,[0.26, 6.40]$ points and annotated support retention by $5.87\,[1.65, 9.95]$. Review-label accuracy changes by $-1.60\,[-4.28, 1.07]$ and current accuracy by $-1.20\,[-2.85, 0.44]$. The grounded-review contrast has raw $p = 0.0387$. This procedure changes record scores through proposed-basis membership.

Table F.8: Structural transfer of survival scoring to estimated eQASC supports. The 150-question development panel compares Robust with Verified. Retention effects are percentage points.

Budget	Changed states	Budget utilization	Full history fits	ΔBasis retention (points)
120	$6/150$	0.93	0.000	$-1.3\ [-4.0, 1.3]$
170	$14/150$	0.95	0.007	$1.3\ [0.0, 3.3]$

The survival-objective transfer uses estimated candidate groups on 150 development questions. Preserving overlap between the active and alternative supports is important: 260 of 534 annotated surviving supports share a record with the active support. The candidate procedure retains these shared members while evaluating withdrawal-specific survival.

At 120 and 170 tokens, Verified and Robust produce different states for six and 14 questions, respectively. Their paired annotated-retention differences are $-1.3\,[-4.0, 1.3]$ and $+1.3\,[0.0, 3.3]$ points. The candidate withdrawal family covers all single-record withdrawals and pairs from the active support; the realized single-record withdrawal belongs to that family by construction.

Accurate support descriptions expose additional opportunities for selection. With estimates, the comparison state is optimal on 116/150 cases, and exact selection on the remainder changes annotated retention by $0.0\,[-14.7, 17.6]$ points. With annotated supports, only 68/150 states are optimal; exact selection on the other 82 gains $30.5\,[20.7, 40.2]$ points. The latter effect is conditional on those 82 questions. This eQASC diagnostic evaluates annotated structural retention.

F.6 PREDICTED SUPPORTS AND PROMPT-COST COMPARISONS

F.6.1 CANDIDATE DISCOVERY AND QUALIFICATION

The predicted-basis procedure uses the supplied current verdict and support-size condition to discover and qualify alternative support, then selects qualified candidates to rebuild the state. We evaluate Qwen3-8B, Mistral-Small-24B, and Llama-3.3-70B-AWQ constructors on 323 FEVEROUS, 483 Climate-FEVER, and 500 FactNet cases. The three constructors each process all 1,306 source cases, producing 3,918 constructor–case observations.

Table F.9 reports the larger candidate-discovery evaluation; the matched-cost comparison uses its own independent source-case denominator.

Table F.9: Predicted-basis selection on public evidence. Each source case is processed by three constructors. Support retention and grounded review are percentages.

Source	Cases	Constructor outputs	Basis retention	Grounded review
FEVEROUS	323	969	97.01	73.17 $[70.38, 75.95]$
Climate-FEVER	483	1,449	78.05	64.32 $[61.84, 66.80]$
FactNet	500	1,500	99.60	92.33 $[91.00, 93.67]$

Pooled revision-basis retention is 90.99% and grounded review is 77.23% ($[75.91, 78.53]$). Climate-FEVER has the lowest retention and grounded review among the three sources. Six-constructor structural checks extend coverage of candidate discovery and qualification. The behavioral results here use the three named constructors.

The 300-case matched-input experiment supplies the recorded outcome and an annotation-derived required alternative size, which equals one on every case. It ranks up to five candidates, applies a topic-usability qualification, selects the required-size prefix of the accepted list, and combines it with the active record. Qualification uses the model's estimate of topical usability; the grounding outcome uses the released evidence annotations. The unknown-size QASPER experiment in Appendix F.7 removes these supplied answer/size conditions.

F.6.2 NATURAL-LENGTH INPUTS AND PADDING SENSITIVITY

Table F.10: Paired padding comparison on the same 300 source cases. Each comparison retains its original pair-specific padding length. The values are annotation-bound grounded-answer counts.

Control	Method padded	Method natural	Control padded	Control natural
Neutral generic	227	240	198	208
Review-aware	225	240	180	192

Natural-length method differences: $+10.67\,[5.67, 15.67]$ pp versus generic and $+16.00\,[10.67, 21.33]$ versus review-aware. Within this rerun, padded differences are $+9.67\,[5.00, 14.33]$ and $+15.00\,[10.00, 20.00]$. The five-control comparison below uses its original generations.

Removing padding leaves the chosen evidence unchanged and removes the whole non-citable suffix. Both arms improve in absolute performance. Padded-minus-natural method/control effects are $-4.33\,[-7.33, -1.33]$ and $-3.33\,[-6.00, -0.67]$ points for the generic comparison, and $-5.00\,[-8.00, -2.00]$ and $-4.00\,[-6.67, -1.33]$ for review-aware. The differences in method effects across padding conditions are $-1.00\,[-3.33, 1.33]$ and $-1.00\,[-3.67, 1.67]$. Natural inputs have unequal lengths: the method averages 440.4133 tokens, generic 430.9633, and review-aware 423.8333. This comparison measures gains at natural input lengths; the actual costs differ across methods. All four cells per comparison are generated within the same run; the two unpadded contrasts have Holm $p = 0.0002$ in the declared four-test family.

F.6.3 AN EXACT-PROMPT-COST COMPARISON

Table F.11 gives all five controls, including the review-aware and neutral-generic comparisons reported in Section 5. MMR provides the relevance–diversity control (Carbonell & Goldstein, 1998).

Table F.11: **Qualified predicted bases improve grounded review at exactly matched prompt cost.** Each contrast uses 300 independent public cases. Effects and paired 95% intervals are percentage points.

Matched control	Grounded-review gain	95% CI
Neutral generic	$\mathbf{+9.67}$	$[\mathbf{5.00}, \mathbf{14.33}]$
Current-focused	$+14.67$	$[9.67, 20.00]$
Review-aware	$+15.67$	$[10.67, 20.67]$
Atomic relevance	$+16.67$	$[11.67, 21.67]$
MMR	$+35.67$	$[30.00, 41.67]$

Non-citable neutral padding equalizes standardized prompt cost within each pair while keeping its selected evidence fixed. This experiment evaluates the predicted-basis procedure; Robust uses the executable selection task.

To compare evidence choices at the same prompt cost, we use 300 independent public cases balanced at 100 per source and 100 per constructor, assigning one constructor to each case. For each of five controls, semantically empty, non-citable padding equalizes standardized reviewer-prompt tokens while keeping the chosen evidence IDs fixed. Outcomes are paired within case; bootstrap inference resamples the 300 source cases. Five contrasts yield 600 outputs each and 3,000 in total.

Predicted-basis selection improves grounded review over the five controls in Table F.11. Its answer is unchanged across the five pair-specific padding lengths on 289/300 cases (96.33%). This stability check tracks the predicted-basis arm across pair-specific padding lengths; the preceding padding rerun compares both arms with their natural-length inputs. The intervals describe evidence-selection contrasts under the stated padding. Cost means the complete standardized reviewer input, with upstream candidate discovery and qualification outside that match.

F.6.4 Selection pressure

A separate equal-record-capacity comparison finds a larger predicted-basis advantage where selection is constrained. Against current-focused selection, grounded review improves by $7.99\,[0.91, 15.30]$ points in the binding stratum and by $2.67\,[-6.00, 11.19]$ under low pressure. Realized lengths can differ when a method stops filling early. These results characterize available capacity; the padding experiment controls realized prompt-token cost directly.

F.7 Discovering and retaining paragraph-level supports in QASPER

QASPER supplies full research articles, questions, answers, and evidence annotations (Dasigi et al., 2021). Candidate construction starts from all native full-text paragraphs. Qualification yields 138 questions from 118 papers; a fixed paper-ID split assigns 30 papers to development and 105 questions from 88 papers to evaluation. Evaluation contains 63 yes/no questions from 57 papers and 42 extractive questions from 39 papers, with overlap in papers across answer types. Construction uses the source paragraphs and question; gold answers, alternative IDs, and support sizes are reserved for evaluation.

Support structure. The task invalidates paragraphs within one paper. In 45/105 questions, the designated alternative is a subset of the active evidence. In 46/105, active evidence alone still contains some annotated support after withdrawal. Twenty questions have two multi-paragraph annotation sets, of which 13 retain support inside the active state after withdrawal. Released annotations define support-set sufficiency in this task. The panel combines questions that require additional evidence with questions whose current evidence remains sufficient after withdrawal. All original questions remain in the primary denominator.

Selection and grounded answers. The eight-test primary family concerns retention and grounded-answer superiority over relevance and review-aware selection at both budgets. Estimated survival–relevance grounding differences are $+2.381\,[-3.968, 8.871]$ and $+2.381\,[-0.877, 7.031]$ points; neither passes the declared correction. Estimated survival–review-aware differences are $-7.937\,[-16.129, -0.781]$ and $-8.730\,[-17.188, -0.833]$. The intervals give two-sided effect es-

Table F.12: QASPER yes/no evaluation, 63 questions. Retention and annotation-grounded answer rates are percentages. Whole-group and review-aware comparators use the same source inputs.

K	Selector	Retention	Grounded answer
512	Active-preserving relevance	44.44	32.54
	Review-aware	58.73	42.86
	Whole-support	53.97	39.68
	Estimated survival + Pair-rescue	47.62	34.92
1024	Active-preserving relevance	53.97	37.30
	Review-aware	76.19	48.41
	Whole-support	68.25	44.44
	Estimated survival + Pair-rescue	58.73	39.68

timates, and the declared tests assess superiority. Inference clusters by paper, gives equal question weight, and averages two reviewers within question.

On the 42 extractive questions, annotation-grounded answer F1 (multiplied by 100) is 33.25, 48.10, 41.26, and 37.74 at $K = 512$ for relevance, review-aware, whole-support, and estimated survival; at $K = 1024$ it is 46.65, 51.43, 52.33, and 48.74. Corresponding support retention is 64.29, 80.95, 73.81, and 69.05% at the smaller budget and 88.10, 92.86, 92.86, and 90.48% at the larger budget. These F1 comparisons are secondary and are kept separate from yes/no accuracy.

Available evidence and identified alternatives. Candidate pools contain a complete post-withdrawal annotated support in 103/105 questions. A new, withdrawal-free annotated support appears in proposed groups on 26/105 questions and survives qualification on 16/105. Among 208 accepted estimated groups, 41 equal the active set and 52 are its strict supersets. The 103/105 count concerns any surviving support in the source pool; the new-alternative counts additionally check overlap with the active evidence and withdrawal. These diagnostics locate the gap between available evidence and identified alternatives.

Realizing support through selection. On the same estimated families, whole-support retention exceeds estimated survival by 6.35 points at 512 tokens and 9.52 at 1,024 tokens on yes/no questions. Holding these families fixed exposes a further difference in how selection realizes the proposed supports. The results distinguish source availability, proposal and qualification, and complete admission as conditions for translating argument survival into grounded answers.

F.8 Basis repair at memory interfaces

F.8.1 All-case evaluation

Table F.13: All-case source-match estimates and actual state changes. Scores are percentages; effects and paired 95% intervals are percentage points.

Interface	Changed /600	Base	Repaired	Δ	95% CI
A-MEM	53	37.944	45.389	$+7.444$	$[5.556, 9.444]$
LongLLMLingua	52	33.667	40.500	$+6.833$	$[4.889, 8.889]$
Structured event	5	48.667	49.111	$+0.444$	$[0.056, 0.944]$
LightMem	4	54.778	55.056	$+0.278$	$[-0.111, 0.833]$
Raw records (reference)	184	16.722	43.333	$+26.611$	$[23.611, 29.611]$

All 600 cases remain in each mean. Raw records have no memory interface; their current accuracy increases from 30.50% to 52.61% under repair. Differences use unrounded values.

We apply candidate qualification, budgeted record replacement, and self-verification to four memory interfaces: A-MEM, LongLLMLingua, structured event memory, and LightMem. The original state is retained when verification fails. Budget-matched raw records provide a reference condition. All use the same 600 public cases, with 200 each from FEVEROUS, Climate-FEVER, and FactNet.

Qwen, Gemma, and GPT-OSS review each case; their outcomes are averaged within case before paired inference over the 600 original cases.

Table 2 reports the four memory interfaces and the budget-matched raw-record reference. Table F.13 gives higher-precision pooled source-match estimates. Table F.14 gives all three outcomes, and Table F.15 gives the source-specific source-match changes.

Table F.14: Complete outcomes for basis repair at four memory interfaces and the raw-record reference. Base and repaired values are percentages; paired changes and 95% intervals are pp. Each row uses the same 600 independent cases with three reviewers averaged within case.

Interface	Outcome	Base	Repaired	Change [95% CI]
Raw, budget-matched	Current accuracy	30.50	52.61	$+22.11$ $[+18.94, +25.39]$
	Review accuracy	19.06	45.11	$+26.06$ $[+22.72, +29.50]$
	Source-match score	16.72	43.33	$+26.61$ $[+23.61, +29.61]$
A-MEM	Current accuracy	70.06	70.56	$+0.50$ $[+0.17, +0.89]$
	Review accuracy	39.89	47.44	$+7.56$ $[+5.50, +9.72]$
	Source-match score	37.94	45.39	$+7.44$ $[+5.56, +9.44]$
LongLLMLingua	Current accuracy	64.61	67.11	$+2.50$ $[+1.39, +3.72]$
	Review accuracy	36.39	42.72	$+6.33$ $[+4.50, +8.28]$
	Source-match score	33.67	40.50	$+6.83$ $[+4.89, +8.89]$
Structured event memory	Current accuracy	68.61	68.61	0.00 $[-0.17, +0.17]$
	Review accuracy	49.94	50.50	$+0.56$ $[+0.11, +1.11]$
	Source-match score	48.67	49.11	$+0.44$ $[+0.06, +0.94]$
LightMem	Current accuracy	72.17	72.22	$+0.06$ $[0.00, +0.17]$
	Review accuracy	55.67	55.94	$+0.28$ $[-0.11, +0.83]$
	Source-match score	54.78	55.06	$+0.28$ $[-0.11, +0.83]$

Changes use unrounded values. All five current-accuracy contrasts meet the -2 pp non-inferiority criterion. Failed parses remain in the denominator and score as incorrect.

Table F.15: Source-specific source-match changes from basis repair. Each cell uses 200 cases and three reviewers per case. Percentage points with paired 95% intervals.

Memory interface	FEVEROUS	Climate-FEVER	FactNet
Raw, budget-matched	6.33 $[3.17, 9.83]$	16.67 $[12.00, 21.83]$	56.83 $[50.00, 63.50]$
A-MEM	0.00 $[0.00, 0.00]$	0.00 $[-1.50, 1.50]$	22.33 $[16.83, 28.17]$
LongLLMLingua	1.00 $[0.00, 2.33]$	6.50 $[3.17, 10.17]$	13.00 $[8.50, 18.00]$
Structured event memory	0.00 $[0.00, 0.00]$	0.50 $[0.00, 1.33]$	0.83 $[0.00, 2.17]$
LightMem	0.00 $[0.00, 0.00]$	0.33 $[-0.50, 1.50]$	0.50 $[0.00, 1.50]$

The aggregate benefits differ in their source composition. A-MEM's 7.4-point increase is concentrated on FactNet, with a 22.3-point gain and zero point differences on the other two sources. LongLLMLingua gains 6.8 points overall, with larger effects on Climate-FEVER and FactNet. Structured event memory and LightMem change by 0.4 and 0.3 points, respectively; the LightMem interval includes zero. All five conditions satisfy the two-point current-accuracy non-inferiority criterion.

All-case estimates include unchanged states as well as applied repairs. Trigger-conditioned analyses use smaller subsets, including five final structured-event changes and four LightMem changes. Thirteen event-memory proposals were accepted before the final token reconstruction; only five changed the final state. The procedure operates through budgeted replacement and retention of the original state at each interface; the independently matched full-prompt-cost experiment is reported in Appendix F.6.

The corresponding current-accuracy changes are $22.11\,[18.94, 25.39]$ points for raw evidence, $0.50\,[0.17, 0.89]$ for A-MEM, $2.50\,[1.39, 3.72]$ for LongLLMLingua, $0.00\,[-0.17, 0.17]$ for structured event memory, and $0.06\,[0.00, 0.17]$ for LightMem. Parsing failures remain in the denominator and count as incorrect.

F.8.2 ACTUAL STATE CHANGES AND CONTENT-LEVEL COVERAGE

Our structured event-memory baseline writes each source record separately with event extraction, associates the extraction with a source alias, retrieves the stored memory payload and speaker field by embedding score, deduplicates alias/content pairs, and admits complete records under 256 tokens. It performs no cross-event consolidation in this experiment. Repair checks whether the remaining content supports reconsideration, qualifies unretained non-active candidates, replaces non-active tail records, and falls back after failed self-verification. Final strict-token reconstruction protects the active records and attempts accepted additions as complete units.

Final changed states are 184/600 for raw records, 53/600 for A-MEM, 52/600 for LongLLMLingua, 5/600 for event memory, and 4/600 for LightMem. The accepted-proposal counts are 216, 130, 104, 13, and 4. The counts summarize proposals, acceptances, and final changed states at their respective processing stages.

For content-level structural and grounding checks, raw and A-MEM states are jointly decidable on all 600 cases. LongLLMLingua has 271 jointly decidable cases (74 FEVEROUS, 64 Climate-FEVER, 133 FactNet). Event memory and LightMem have no jointly decidable content subset under this check; aliases alone leave all 600 cases unresolved. Unresolved content remains a separate category from absent support. The all-case source-match outcomes remain defined on all 600 cases.

A-MEM's source-specific score differences are 0.00 pp on FEVEROUS, 0.00 pp on Climate-FEVER, and +22.33 pp on FactNet. The Climate-FEVER interval is $[-1.50, 1.50]$ pp; paired gains and losses can cancel while producing a nonzero interval. Final state changes are 0, 6, and 47 across those sources. LongLLMLingua changes 4, 22, and 26 states and gives source-specific score gains of 1.00, 6.50, and 13.00 points. These source-specific changes show where the aggregate repair gains arise; the content checks above assess retained support directly.

G REPRESENTATION AND TEMPORAL SCOPE

These evaluations describe how the argument-survival question relates to summaries, repeated updates, and native temporal tasks. Appendix G.1 examines representations and constructed update sequences; Appendices G.2 and G.3 evaluate released temporal questions and chronologically maintained Mem0 states under their own task metrics.

G.1 ADDITIONAL REPRESENTATIONS AND WITHDRAWAL SETTINGS

G.1.1 CITATION-BOUND SUMMARIES

This experiment uses synthetic citation-bound summaries, a 256-token writer budget, and executor evaluation of the cited records. It is a distinct task from the 180-token ProofWriter record-selection comparison. The matched natural-evidence comparison against review-aware selection appears in Table F.11; its effect is $15.67\ [10.67, 20.67]$ percentage points of grounded review.

The controlled summary experiment contains 150 cases, three writers, and two record orders, yielding 900 outputs per objective. At a 256-token writer budget, current-focused, neutral-generic, and review-aware summaries retain the requisite cited records on 570/900 (63.3%), 818/900 (90.9%), and 883/900 (98.1%) outputs. The corresponding citation-bound executor review counts are also 570, 818, and 883. Mean writer lengths are 244.62, 248.43, and 246.19 tokens. The current-focused summary objective permits alternative evidence. The active-derivation prompt belongs to the separate evidence-role experiment.

Neutral-generic summaries outperform current-focused summaries under the executor in all 18 public source–constructor cells, with a pooled difference of $8.07\ [7.33, 8.84]$ points over 1,306 source clusters (Table G.1). An eQASC paired-answer diagnostic gives a different ordering, favoring current-focused selection by $1.87\ [0.13, 3.78]$ points. The two results concern different representations and review endpoints.

Table G.1: Neutral-generic minus current-focused selection in citation-bound summaries. The endpoint is annotation-bound executable review. Percentage-point effects use original source cases as clusters.

Source	Case clusters	Review difference	95% CI
FEVEROUS	323	$+5.78$	$[4.41, 7.15]$
Climate-FEVER	483	$+4.02$	$[2.93, 5.12]$
FactNet	500	$+13.47$	$[12.23, 14.68]$
Combined	1,306	$+8.07$	$[7.33, 8.84]$

G.1.2 Multiple withdrawals from one state

Across two valid withdrawals from the same controlled state, mean grounded review is 12.0% for native memory, 21.3% for record-wise selection, 39.8% for basis-group selection, and 82.8% for full history. Additional insufficient-support and irrelevant-withdrawal conditions test appropriate abstention and decision stability.

G.1.3 Repeated updates and reapplication

A constructed lifecycle diagnostic uses 75 public cases and three constructors, giving 225 trajectories. A single initial selection covers 90.7%, 45.3%, 22.7%, and 0.0% of live evidence groups at time steps 1, 4, 16, and 32. Reapplying the initial qualification judgments through supplied group lineage yields 96.0% final coverage at an eight-record budget; the full oracle reaches 98.7%. Their paired gap is $2.67\,[0.44, 5.78]$ points over the 75 source cases.

At four records, final coverage is 48.0% for reapplication and 49.3% for the oracle. A separate native-state panel records five complete-basis losses in 180 case–system trajectories; withdrawal-conditioned requery changes grounded review by $1.94\,[-0.83, 5.00]$ points over 90 paired cases.

The lifecycle schedules are constructed from static evidence. Evaluation on naturally occurring updates additionally requires linked evidence snapshots, update payloads, and revised-answer annotations.

Multi-record coverage. Single-record alternatives account for much of the broad public extension. A screen of 654 current-state units for a missing multi-record alternative yields 25 eligible FEVEROUS units and none from Climate-FEVER or FactNet. Joint-completeness results therefore come from the dedicated FEVEROUS, eQASC, RuleTaker, and controlled panels.

G.2 Native questions about changing preferences and facts

PersonaMem-v1 supplies a released conversational prefix for each question (Jiang et al., 2025). We evaluate 100 questions from 16 held-out personas at a 2,048-token read budget chosen on development data, with mean candidate size 6,497 tokens. Relevance, review-aware, whole-support, and estimated survival with Pair-rescue score 80.2, 77.1, 78.0, and 79.0%, respectively, using persona-macro estimates. All outputs parse. Paired persona-cluster intervals for estimated survival minus review-aware and whole-support are $+1.93\,[0.00, 4.91]$ and $+1.04\,[0.00, 3.12]$ pp (Holm $p = 0.4332$ and 0.6203). Nineteen questions have no qualified family; the support-based methods select identical states on 71/100 questions. These native questions test temporal answering at their released prefixes without a subsequent withdrawal intervention.

MemoryAgentBench FactConsolidation supplies 30 single-hop and 30 multi-hop questions over the same released 32K fact stream (Hu et al., 2026). Each of the four policies scores 29/30 and 2/30 in the respective categories under both substring and strict exact scoring. With mean candidate size 7,864 tokens, the 2,048-token budget omits approximately 24 records. Qualified families are empty on 26/60 questions, and whole-support and Pair-rescue states coincide on 96.7%. All 60 questions share one fixed knowledge stream, which defines the common history for these estimates. Across these native tasks and the evidence-use comparison, 1,655 physical generations consume 9,320,073 input and 120,079 output tokens.

G.3 CONVERSATIONAL APPLICABILITY IN NATIVE MEM0

We evaluate chronological memory on 200 LoCoMo questions, 20 from each of the ten released conversations (Maharana et al., 2024). Native Mem0 OSS maintains the history through its normal write operations (Chhikara et al., 2025). The inputs comprise 272 sessions and 5,882 turns, or 386,347 Qwen tokens after rendering speaker, date, and record information. One fixed Qwen reader answers from a 4,096-token evidence allocation using four read policies: relevance, review-aware selection, whole-support admission, and estimated survival. This panel evaluates conversational applicability without an evidence-withdrawal step.

Native task and common inputs. The 156 answerable questions use answer F1; the remaining 44 adversarial or unanswerable questions use the native detection score. Their composite is averaged over all 200 questions. Since each conversation contributes 20 questions, question-weighted and conversation-macro means coincide. Paired bootstrap draws resample conversations with all their questions and policy outcomes.

All 200 complete candidate pools fit within $K = 4{,}096$, so every policy retains every candidate record. Candidate costs range from 1,338 to 1,693 tokens, with mean 1,466.3; no question requires a record to be omitted. The four policies therefore supply identical evidence on every question and have identical native outcomes (Table G.2). This configuration evaluates reading from the maintained conversational memory with every candidate record available.

Table G.2: Conversational applicability on the same 200 LoCoMo questions and ten conversations. All scores and intervals are percentages; answerable F1 and adversarial detection use their respective question subsets.

Read policy	Native score	95% CI	Answerable F1	Adversarial
Relevance	35.98	$[32.48, 39.65]$	18.57	97.73
Review-aware	35.98	$[32.48, 39.65]$	18.57	97.73
Whole-support	35.98	$[32.48, 39.65]$	18.57	97.73
Estimated survival	35.98	$[32.48, 39.65]$	18.57	97.73

Each row contains 156 answerable and 44 adversarial/unanswerable questions, with 146/200 abstentions, 0/200 parse failures, and 0/200 invalid-ID outputs. The composite averages the native F1 and adversarial-detection scores.

The two primary estimated-survival contrasts, against review-aware and whole-support selection, are both $0.00\,[0.00, 0.00]$ pp with Holm $p = 1.000$. These degenerate intervals follow from identical inputs and paired scores on every question. Table G.3 preserves the native task categories.

Table G.3: Native LoCoMo categories in the 200-question panel. The four policies have the same score in each category.

Category	Questions	Conversations	Common score (%)
Multi-hop	29	10	18.59
Temporal	33	10	13.79
Open-domain	9	9	27.78
Single-hop	85	10	19.45
Adversarial/unanswerable	44	10	97.73
All categories	200	10	35.98

The first four rows report answer F1; the fifth reports detection. The last row is the question-weighted composite across the two scoring rules.

Estimated supports. Qualified support families are empty on 169/200 questions. Estimated active-set sizes are zero, one, two, three, and six records on 156, 32, seven, four, and one questions. The size-six estimate has 15 active pairs but no qualified support family, so it follows relevance fallback. The native estimated-survival policy includes all active pairs; this choice changes no final state in the panel. Model-estimated groups and source identifiers are kept separate from executable support certificates.

Resource accounting. Table G.4 reports the requests newly executed for this evaluation; previously computed histories and query objects are reused. Including those reused checkpoints, attributable native writing comprises 598 requests, 1,319,027 input tokens, and 298,535 output tokens. That writing attribution includes the measured write row; the newly executed total remains 416 requests. The evaluation contains all 800 question–policy rows, all 200 state-identity comparisons, and the 240 cached CPU timing trials described in Appendix D.10.

Table G.4: Newly executed model requests for the 200-question conversational evaluation. Columns count physical requests and tokens; they exclude reused requests. Complete-history writing attribution is stated separately in the text.

Stage	Requests	Input tokens	Output tokens
Native Mem0 writing	59	132,790	30,544
Provisional answer	79	125,455	1,447
Support proposal	79	127,562	34,089
Independent support check	41	9,077	1,848
Review-aware selection	79	126,693	880
Native answer	79	122,532	1,543
Newly executed total	416	644,109	70,351

G.4 Broader memory context

Logical foundations. Assumption-based truth maintenance tracks alternative justifications and their recorded dependencies (de Kleer, 1986); OWL explanation algorithms enumerate subset-minimal justifications (Kalyanpur et al., 2007). Resource-bounded belief revision relates efficient belief change to rationality postulates (Alechina et al., 2006). We select records for a language-model reader before withdrawal under a token budget, measuring alternative loss despite current-answer success and its effect on grounded responses. Argument survival scores a prespecified withdrawal family and captures complementary records' joint value.

Temporal evaluation. Supersede examines current values under bounded memory (Patel, 2026). Memora evaluates recall and forgetting (Uddin et al., 2026); StateMemBench distinguishes current from superseded state (Fan et al., 2026); Zep (Rasmussen et al., 2025) models temporal relations, and MemoryArena (He et al., 2026) links memory to action. These settings complement our principal selection task, which keeps the full-history answer fixed while withdrawing its current evidence.

Construction, consolidation, and revision. A-MEM links notes (Xu et al., 2025), LightMem stages consolidation (Fang et al., 2026), and MemGPT manages context tiers (Packer et al., 2023). Mem0 updates records (Chhikara et al., 2025); BeliefMem maintains probabilistic conclusions (Liao et al., 2026). Kumiho (Park, 2026), NeuSymMS (Sultan et al., 2026), and TMMA (Phadke et al., 2025) study revision and write verification. KV-cache selection reduces inference memory use (Zhang et al., 2023), while OAS chooses between retention and consolidation under budget pressure (Kang et al., 2026). Argument survival concerns the evidence that these operations leave available to the reader.

Broader evaluation and use. AgingBench examines reliability over prolonged deployment (Zhu et al., 2026); MemGym evaluates long-horizon agent memory (Xu et al., 2026), MemDelta isolates model and retrieval confounds (Wang, 2026), and AMemGym tests memory in interactive conversations (Cheng et al., 2026). Multi-turn studies examine failures over continued interaction (Laban et al., 2026). The controlled withdrawal task separates evidence availability from answer realization within a fixed state; the native evaluations above describe where its selection opportunities arise in broader workflows.

Prompt compression (Jiang et al., 2024; Pan et al., 2024) reduces reader input. Evidence-insufficiency studies examine missing support (Atanasova et al., 2022), while position studies examine access to available evidence (Hsieh et al., 2024); +VeriRel supplies verification feedback that can inform discovery (Deng et al., 2025). Our interventions ask whether complete arguments survive selection and whether the reader produces a grounded answer from them after withdrawal.