WHEN DOES AGENT MEMORY MATTER? SOURCE ACCESS CONDITIONS ITS MARGINAL CONTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent memory is typically evaluated by comparing performance with and without memory, but the meaning of that comparison depends on what information remains available through other sources. We study source accessibility (whether information stored in memory remains obtainable from its original sources) as one condition shaping memory's marginal contribution. In a purposive full-text review of 19 agent-memory evaluations, none reports memory's contribution across multiple source-access regimes. We introduce *Memory Opportunity Audit (MOA)*, which freezes acquired memory and execution state and crosses downstream memory use with source access. In our most controlled document study, memory is constructed before any evaluation question is seen and source loss is assigned independently of memory content. NOTES improves strict-EM accuracy by $+3.33$pp with intact access and by $+6.33$pp under random denial, yielding a $+3.00$pp interaction $[+0.33, +5.74]$. A disjoint second cohort shows positive access-conditioned interactions for both tested memories, while adjudicated human scoring and semantic rescoring preserve the direction of the primary interaction. Scope studies across conversational memory, tool use, and video show substantial heterogeneity across settings. These results imply that a memory-on/off ablation measures contribution within the source-access regime in which it is conducted, rather than an access-invariant property of the memory system. Claims that span access regimes should therefore evaluate the same acquired memory across those regimes and report regime-specific contributions and their interaction.

1 INTRODUCTION

Large language models have become foundational components of AI systems that reason, generate content, and interact with external tools and information (Brown et al., 2020; Bommasani et al., 2021). As these systems move beyond single-turn prediction, many applications require information gathering, tool use, and decision-making over extended interactions. Coupled with tools and iterative feedback, language models can operate as agents across diverse settings, including software engineering (Yang et al., 2024), web research (OpenAI, 2025), interactive decision-making (Yao et al., 2023), and multimodal tasks (Yang et al., 2025; Yeo et al., 2026). Such agents are increasingly studied and deployed in practice (McCain et al., 2026).

Sustaining such interactions creates a corresponding state-management problem. Agents accumulate observations, tool outputs, intermediate decisions, and information from earlier interactions, while the language models that power them operate with bounded context. Explicit memory mechanisms address this mismatch by preserving information beyond the immediate context and making it available for later reuse (Packer et al., 2023; Chhikara et al., 2025). A-MEM organizes linked memory notes (Xu et al., 2025), Mem0 extracts and consolidates conversational information (Chhikara et al., 2025), and benchmarks such as LongMemEval (Wu et al., 2025) and MemoryArena (He et al., 2026) evaluate long-term recall and interdependent multi-session action.

Yet building and benchmarking memory systems leave a more fundamental issue unresolved: results from one setting do not tell us under what conditions the same memory will matter. This motivates our central question: *when does agent memory actually contribute to downstream performance?*

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.

We focus on one condition that can shape the answer: *source accessibility*: whether the information stored in memory can still be obtained from its original sources. If those sources remain accessible, memory is one of several routes to the same information; when source access is restricted, the stored information may become substantially more consequential. Source access can also change in practice through retention and deletion policies or changes in access authorization (Boeckl & Lefkovitz, 2020); our experiments isolate this axis through controlled restrictions rather than modeling natural outages. We show that source accessibility is a moderator of agent-memory contribution: the marginal contribution of the same frozen memory can change across source-access conditions in our controlled document studies; conversational, tool, and video checks show mixed extensions and boundaries.

Existing memory-on/off evaluations in our coded sample do not report memory's contribution across multiple source-access regimes: 15 of 19 keep the original information available in a comparison condition, and none reports memory's contribution under more than one such regime (Section 2; Appendix J.2). With the original information recoverable through retrieval or tools, an ablation estimates memory's contribution alongside that alternative path, not whether it holds when the path is restricted. We therefore ask three research questions. **RQ1:** Does the marginal contribution of the same frozen memory change with source access? **RQ2:** Does this access dependence persist when memory is constructed before the downstream question and source loss is assigned independently of memory content? **RQ3:** What does a conventional memory-on/off ablation identify when evaluated under only one source-access regime?

Our contribution is to make this dependence experimentally identifiable: we hold acquired memory fixed, vary source access independently where possible, and measure the resulting change in marginal contribution. A single memory-on/off ablation identifies a conditional contribution, not one that transfers across access regimes.

To isolate this dependence, we introduce *Memory Opportunity Audit (MOA)* (Figure 1). MOA freezes acquired memory and execution state, then crosses downstream memory use with source access. This design asks a simple question: does the same acquired memory contribute differently when the original information becomes harder to obtain? Audits verify that blocked sources are absent from delivered inputs, that the frozen memory is delivered as intended, and that paired branches use matched access rules, budgets, and delivery contracts.

Our most controlled test addresses RQ2 directly. On 300 new M3DocVQA text/table questions (Cho et al., 2024), factual notes are built from the entire corpus before any question is seen, and document loss is assigned independently of memory content. NOTES improves strict-EM accuracy by $+3.33$pp with intact access and by $+6.33$pp under denial, a $+3.00$pp interaction $[+0.33, +5.74]$. A disjoint second cohort shows the same evaluation shift: neither tested memory has a resolved positive contribution with sources intact, whereas both do after document denial (Figure 2). Human and semantic rescoring show that the primary direction is not specific to strict lexical matching. Exploratory support-hit analyses further suggest that the aggregate interaction is concentrated when denial reaches annotated supporting documents. Supplier-source, conversational, tool, and video studies then probe where this access dependence extends and where it does not (Sections 4–5).

Contributions.

- **When agent memory matters.** We identify source accessibility as a moderator of memory's marginal contribution: the same frozen memory can contribute differently depending on whether the information it stores remains obtainable from its original sources. This pattern persists under question-independent memory construction and content-independent source loss, and remains positive under adjudicated human scoring.
- **Conditional estimand.** We formalize the regime-specific contribution identified by a memory-on/off ablation and distinguish it from claims that transfer across access regimes.
- **MOA protocol.** We introduce a crossed experimental protocol that freezes acquired memory and execution state, varies downstream source access and memory use, and audits source delivery, branch isolation, and matched execution.

2 Related Work

Memory evaluation and retained information paths. Many memory evaluations leave the underlying information available through an alternative path in their comparison conditions. Examples include the preceding conversation in Xu et al. (2025), the full history in Chhikara et al. (2025), and interaction history for the long-context agents in He et al. (2026). Our full-text coding of a purposive sample of 19 agent-memory evaluations finds that 15 retain the original information in context or make it retrievable in a comparison condition, and none reports memory's contribution separately under two or more source-access conditions (Appendix J.2). Sixteen of the 19 were coded in two independent AI-assisted passes, with disagreements adjudicated and all final codes verified by an author. ExpeL (Zhao et al., 2023) is a partial exception: it crosses extracted insights with retrieval of the trajectories they summarize, but both are components of its experience memory and the insight set is not stated to be fixed. ReadAgent (Lee et al., 2024a) fixes its gist memory while varying how many raw pages may be re-read; this changes the lookup budget rather than source availability, since every page remains accessible. These are meaningful baselines for the access conditions they study.

Source conditions and attribution. Related work also studies how evidence is distributed and whether it remains sufficient, as well as how contributions are attributed when evidence or intermediate components change (Table 1; Appendix J.2). SMMBench (Chai et al., 2026) varies how fixed gold evidence is distributed across sources, while Selective QA (Yang et al., 2026) distinguishes source reachability from extraction and resolution and crosses resolver choice with input quality. Shah (2026) and Zhang (2026) attribute contributions to agent steps, and Li et al. (2026) use re-execution during training. These designs address evidence organization, sufficiency, or component attribution. MOA instead estimates how the contribution of the same frozen memory changes across regimes.

Table 1: Evaluation contrasts in related work; rows summarize the relevant reported comparison.

Study	Varied	Measured
MemoryArena (He et al., 2026)	Memory system	Later-task completion
SMMBench (Chai et al., 2026)	Placement of fixed gold evidence	Effect of evidence distribution
Selective QA (Yang et al., 2026)	Resolver $\times$ input quality	Extraction vs. resolution
MemAudit (Tan et al., 2026)	Removal of selected memory entries	Contribution to harmful outputs
Shen (2026)	Eviction; gold evidence restored	Recoverable vs. irreversible forgetting
19 evaluations (App. J.2)	Memory on/off; sources kept in 15	No cross-regime contribution estimate
MOA (ours)	Access $\times$ memory, memory frozen	Contribution change across regimes

Memory behavior and policies. A separate line of work examines how agents use and manage stored information, including experience-following and error propagation (Xiong et al., 2025), obsolete-memory use (Uddin et al., 2026), and self-consistency gating (Zhang & Li, 2026). Other work studies policies for when to surface or retain memory, including proactive reminders (Wu et al., 2026) and constrained retention (Kang et al., 2026). Adaptive retrieval consults non-parametric memory when parametric knowledge is weak (Mallen et al., 2023); MOA instead varies access to the sources of fixed agent memory, complementing these approaches.

3 Measuring Source-Conditioned Memory Contribution

MOA evaluates memory contribution as conditional on source access. Conceptually, this is an effect-modification question (VanderWeele, 2009), related to transporting effects across environments (Pearl & Bareinboim, 2011). The memory interaction value (MIV) is an ordinary interaction contrast; MOA supplies the experimental contract, freezing acquired memory and execution state, under which it measures a change in the contribution of the same memory.

Notation. For evaluation unit i, let h_i denote the acquisition history, s_i the frozen downstream state, and M_i the saved memory. The core audit crosses source access $a \in \{A, U\}$ with downstream memory use $m \in \{0, 1\}$ (off/on): A leaves sources available and U restricts targeted sources. Study-specific conditions add matched access arms: P removes matched non-target sources as a placebo,

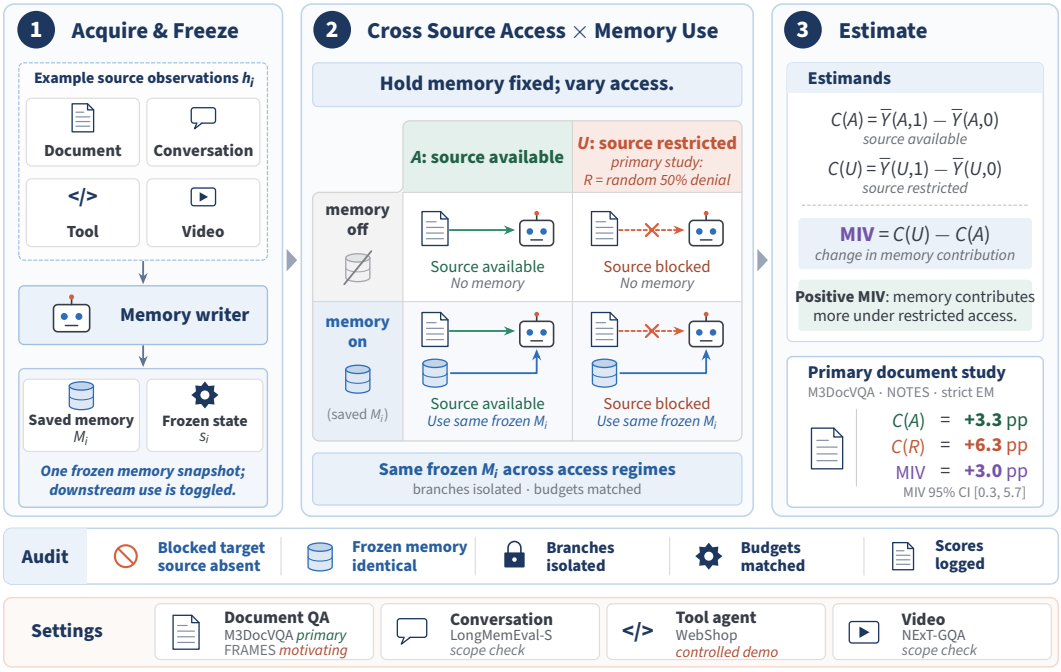


Figure 1: Memory Opportunity Audit (MOA). Memory and execution state are frozen after acquisition, then downstream source access and memory use are crossed under matched execution. Memory contribution is measured within each access regime, and MIV is the difference between those contributions. The primary document study instantiates restricted access as random 50% denial.

and R denotes random loss or retention. A continuation under these conditions is $\tau_i(a, m)$, scored by a predeclared utility function u as $Y_i^u(a, m) = u(\tau_i(a, m))$. With fixed weights w_i, the mean score is $\bar{Y}_u(a, m) = \sum_i w_i Y_i^u(a, m)$ with $\sum_i w_i = 1$.

Acquisition and evaluation. The agent first acquires memory from source observations, and we save the memory and execution state. Four continuations then start from that state, crossing available or restricted sources with memory on or off; memory-off disables memory only after acquisition, so the contrast measures the use of already acquired memory (Figure 1).

Estimands. All cells in a study use the same frozen population and weights. Memory contribution in access condition a is $C_u(a) = \bar{Y}_u(a, 1) - \bar{Y}_u(a, 0)$. The difference between restricted- and intact-access contributions is the source-access interaction, or MIV. The placebo-referenced version compares restricted access with non-target removal:

$$\mathrm{MIV}_u = C_u(U) - C_u(A) = L_0 - L_1, \qquad \mathrm{MIV}_{u,src} = C_u(U) - C_u(P),$$

where $L_m = \bar{Y}_u(A, m) - \bar{Y}_u(U, m)$ is the source-access loss at memory condition m; for random loss or retention, the interaction is $C(R) - C(A)$. A positive MIV means that memory makes a larger marginal contribution under restricted than under intact access; a negative MIV means the reverse. An MIV of zero means that memory's measured contribution is unchanged across the two access regimes, not that memory itself has zero contribution. Like any interaction contrast, MIV is scale-dependent through u; all claims refer to the stated outcome scale. We report both conditional contributions beside MIV so that the change can be traced to the underlying cells. Because MIV shares outcome cells with L_0, their correlation is not evidence of a recovery mechanism.

Inference. Intervals are paired 95% bootstrap percentile intervals over the stated resampling unit, and accuracy differences are reported in percentage points (pp). An interval containing zero leaves the sign unresolved and is not evidence of no effect; nor do differing significance outcomes by themselves establish a difference. Rescoring the same outputs is a sensitivity analysis, not a replication.

Scoring. Each study retains the primary scoring family specified by its protocol before outcome inspection: normalized strict EM for the two new document cohorts, substring EM for the supplier-source study (with an implementation defect later corrected uniformly; Appendix D.2), and a semantic judge fixed before the corrective LongMemEval rerun. Other scorers are secondary, and effect magnitudes are not compared across studies.

Operational validity. Audits verify that blocked sources are absent from reader inputs and that execution branches remain isolated. Each memory-on/off comparison matches the environment instance, context window, action budget and memory-delivery procedure; budget or delivery changes define separate contrasts. A valid restriction may leave inputs unchanged if that source would not have been delivered. Repeatability and intervention-validity checks are in Appendices I.1 and I.2.

4 When Source Access Changes Memory Contribution

We test whether the marginal contribution of fixed memory changes with source access, focusing on designs that separate access from memory construction. Table 6 (Appendix A) orders the evidence by the confounds each design removes; Figure 2 summarizes the main evaluation consequences.

4.1 Pre-question memory with independent loss

Design. Our most controlled study removes both question-dependent memory construction and memory-dependent source loss. We evaluate 300 new M3DocVQA text/table questions (Cho et al., 2024; 281 groups). NOTES stores question-independent factual extractions and FLAT retains raw text, both constructed from all 3,311 corpus documents before any evaluation question is seen. The writer receives no question, answer, evidence label, or loss mask. Qwen3-VL-32B is selected using only intact-access, no-memory strict EM on development questions.

At evaluation time, the reader receives the same ten-action source budget in every condition. Source access is either intact or restricted by one of three fixed random masks that deny half of the documents; each mask depends only on seed, group, and document ID. The memory-on payload is identical across access conditions. The predeclared primary estimand is the NOTES interaction under strict EM. Intermediate summaries were inspected during execution, but all 3,600 cells were completed under the fixed analysis plan without outcome-based exclusions; execution amendments and replica changes are documented in Appendix B.1.

Results. NOTES improves strict-EM accuracy by $+3.33$pp with intact access and by $+6.33$pp under random denial. The same frozen memory therefore contributes $+3.00$pp more when source access is restricted $[+0.33, +5.74]$ (Figure 2A; Table 2). Denial lowers no-memory accuracy by 5.22pp but memory-on accuracy by only 2.22pp; descriptively, NOTES offsets 57% $[9\%, 92\%]$ of the observed loss.

The interaction reflects a change in the information environment rather than a change in memory content. When the original source path remains available, some stored information is redundant with information the reader can still obtain directly. When that path is restricted, the same frozen memory can become more consequential. An exploratory support-hit analysis is consistent with this interpretation: the interaction is $+4.69$pp when a random mask denies at least one annotated supporting document and $+0.00$pp when it denies none, with a $+4.69$pp hit-minus-miss difference $[+0.34, +9.11]$. Because this split was defined after the primary result and uses gold support labels, we treat it as explanatory rather than predictive.

The magnitude is scorer-dependent. Token-F1 yields a positive interaction for both memories, whereas FLAT under strict EM and both memories under substring EM have intervals that include zero (Appendix B.1). Across scorers, the NOTES interaction remains positive in point estimate, while its precision varies.

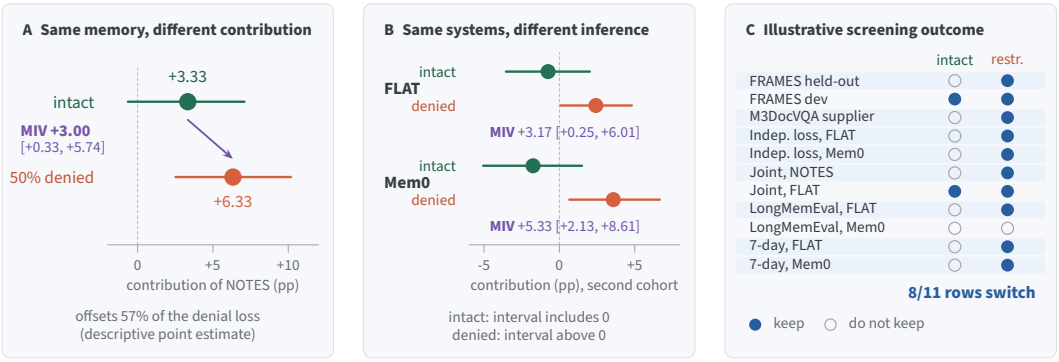


Figure 2: Source access can change both the measured contribution of memory and the conclusion an evaluation supports. **A**: primary joint study (NOTES, strict EM; Table 2). **B**: second cohort (Table 14). **C**: keep-if-the-interval-excludes-zero rule, a decision rule rather than a test of $C(A) \neq C(R)$ (which MIV tests); shaded rows switch (Appendix H.2). *Do not keep* is not evidence of no effect; scorers differ.

Table 2: Primary joint-study estimates of memory contribution under intact and random-denial (R) access. Contributions (pp) with paired 95% group-bootstrap intervals; human labels are adjudicated on a pre-outcome sample; LLM-judge rescoring is in Table 13.

				Contribution of memory		Interaction
Scorer		n	Memory	$C(A)$: intact	$C(R)$: denied	MIV $= C(R) - C(A)$
Strict EM	primary	300	**NOTES**	$+3.33$ $[-0.64, +7.09]$	$+6.33$ $[+2.51, +10.18]$	$\mathbf{+3.00}$ $[+0.33, +5.74]$
			FLAT	$+6.33$ $[+2.67, +10.17]$	$+8.89$ $[+5.22, +12.71]$	$+2.56$ $[-0.22, +5.30]$
Human	adjudicated	150	NOTES	$+10.00$ $[+3.33, +16.77]$	$+16.44$ $[+10.44, +22.60]$	$\mathbf{+6.44}$ $[+1.35, +11.63]$
			FLAT	$+12.00$ $[+5.37, +18.83]$	$+18.44$ $[+12.36, +24.72]$	$+6.44$ $[+1.55, +11.48]$

Source-coupled deletion. A follow-up tests whether notes derived from documents that a mask denies are especially valuable under restriction. Removing those NOTES units reduces strict-EM accuracy by 4.22pp $[+1.46, +7.06]$ when their source documents are denied, compared with 2.56pp $[+0.44, +4.78]$ when the same deletion is forced while those documents remain readable. The resulting retention interaction is $+1.67$pp $[-0.76, +4.09]$. Deleting source-coupled notes is therefore more costly under restricted access at the point estimate, but the interaction interval includes zero and the KEEP and DROP cells ran in different epochs (Appendix B.2).

4.2 Independent loss on a second cohort

A disjoint second cohort shows why the access regime matters for evaluation, not just for effect size. We evaluate 400 previously unused M3DocVQA questions, grouped so that no supporting document is shared across groups. Three hash masks independently deny documents at nominal rate $\rho = 0.5$. A single two-cycle acquisition run over the unmasked corpus is used to construct both FLAT and Mem0. Unlike the joint study, acquisition sees the question.

The ten-action budget was fixed on a separate 50-question development cohort. An earlier four-action run produced negative interactions, so we treat the budget choice as development tuning; no final-cohort outcome informed it. The final and development cohorts share neither questions nor supporting documents.

Results. The evaluation conclusion changes with the access regime. Under intact access, neither FLAT nor Mem0 has a resolved positive contribution (-0.75pp and -1.75pp). Under denial, both contribute positively ($+2.42$pp and $+3.58$pp), yielding interactions of $+3.17$pp $[+0.25, +6.01]$ for FLAT and $+5.33$pp $[+2.13, +8.61]$ for Mem0 (Figure 2B; Table 14). An intact-only evaluation would therefore characterize both memories differently from an evaluation conducted after source restriction. Under equal-group weighting, Mem0's interval remains above zero, whereas FLAT's includes zero.

A post-hoc modality split shows that a positive interaction need not reflect recovery of lost information. On text/table questions, FLAT has a $+4.17$pp interaction $[+0.42, +7.84]$, whereas Mem0 is $+1.48$pp $[-2.48, +5.44]$. Mem0's larger aggregate interaction is concentrated in image questions, where its contribution changes from -9.21pp under intact access to $+2.41$pp under denial. Here, the positive interaction is driven by the disappearance of a negative intact-access contribution rather than by a large restricted-access gain. Source access can therefore change how memory affects performance, not merely how much lost information it recovers.

The exploratory support-hit split repeats the primary-study pattern. Interactions are larger when denial removes an annotated supporting document (hit-minus-miss: $+5.10$pp $[+0.17, +10.04]$ for FLAT and $+7.96$pp $[+2.83, +12.92]$ for Mem0), including on text/table questions. Because the split is post hoc and uses gold support labels, we treat it as explanatory rather than predictive.

4.3 Human validation

The primary interaction is not specific to strict lexical matching. Before inspecting outcomes, we fixed a hash-selected sample of 150 final questions, corresponding to 647 distinct question–response pairs and $1{,}800$ experimental cells. Two external, non-author annotators independently scored the same blinded item list without seeing experimental conditions, automatic scores, or each other's labels. They agreed on 94.6% of responses (Cohen's $\kappa = 0.885$); a third external annotator adjudicated the 35 disagreements.

On this same 150-question sample, the NOTES interaction is $+4.89$pp under strict EM and $+6.44$pp $[+1.35, +11.63]$ under adjudicated human scoring. For FLAT, the corresponding interactions are $+3.78$pp and $+6.44$pp $[+1.55, +11.48]$. Both human-scored interactions remain positive under either reviewer's labels alone. Two semantic LLM judges also yield positive NOTES interactions on the full 300-question output set (Table 13). These rescoring results address dependence on the scoring rule; they are sensitivity analyses on the same outputs, not independent replications.

5 Motivating Evidence and Scope Checks

Section 4 provides the most controlled evidence for access-conditioned contribution. We next probe its scope and boundary conditions across different memory constructions, access interventions, and tasks.

5.1 Supplier-source blocking on FRAMES

Design. A PlanRAG-style agent (Lee et al., 2024b) answers FRAMES (Krishna et al., 2025) and M3DocVQA questions from retrieved excerpts ("cards") and records intermediate answers as memory. We freeze the acquired memory and trajectory, then re-answer each question with Qwen3-VL-8B. The restricted arm blocks supplier documents, defined as documents whose retrieved cards match a fact stored in memory. A placebo arm removes the same number of high-ranked non-supplier documents, with retrieval budgets matched across arms. We evaluate disjoint FRAMES development ($n = 218$) and held-out ($n = 211$) cohorts. M3DocVQA ($n = 195$) is reported as supporting evidence because it falls five questions short of the predeclared minimum.

Results. Blocking the sources that supplied stored facts substantially increases memory contribution. On held-out FRAMES, memory contributes $+3.32$pp while supplier documents remain accessible and $+13.74$pp when they are blocked, an interaction of $+10.43$pp $[+5.21, +16.11]$ (Table 21). The matched placebo leaves memory contribution unchanged. Memory contribution also rises monotonically with the fraction of supplier documents blocked, from $+3.32$pp with no supplier blocking to $+13.74$pp with all supplier documents blocked; the prespecified slope is $+10.62$pp per full dose $[+5.50, +16.11]$. The interaction remains positive under strict EM, and all six tested readers have positive point estimates.

Two design features limit this interpretation. Supplier identity is derived from memory content, coupling the access intervention to what the memory stores. In addition, the delivered memory contains the reference answer for 12 of the 211 held-out questions, which account for 80% of the

strict-EM interaction. We therefore use this study to motivate the controlled designs in Section 4 rather than as primary evidence for the main claim.

5.2 Conversational, tool, and video scope checks

Conversational memory. LongMemEval (Wu et al., 2025) tests whether access-conditioned contribution extends from documents to conversational history. Memory is constructed before the question, and source access is varied either through random session loss or through retention windows. On the predeclared 186-question stratum, random half-session loss yields interactions of $+8.33$pp $[+0.90, +15.50]$ for FLAT and $+8.96$pp $[+1.78, +15.95]$ for Mem0. Seven-day retention produces the same qualitative pattern, whereas 30- and 90-day retention remove much less history and yield small or unresolved interactions.

These results indicate that the extent of source restriction matters. Seven-day retention removes about 44% of sessions in the final cohort, compared with 13% and 4% under the 30- and 90-day policies, respectively. Interactions are larger under seven-day retention and weak under the milder policies. These questions were previously inspected. A 50-question human audit is directionally consistent for random loss and seven-day retention but imprecise; we use it as a scoring sensitivity rather than an independent replication (Appendix E.2).

Tool agent. In a controlled WebShop task, the previous purchase determines the correct current action. Removing access to that receipt while preserving its memory yields a $+32.5$pp interaction $[+15, +50]$. The task is deliberately constructed around this information dependency, so the result is an information-path demonstration rather than a leaderboard estimate. A separate canonical-prefix held-out WebShop study yields 0pp $[-10, +10]$ (Table 50). Thus, the interaction observed in the controlled task does not automatically transfer to a different WebShop task and execution contract.

Video. Post-observation video memory provides a boundary case. On NExT-GQA, blocking the target segment does not produce a resolved increase in memory contribution. EgoLifeQA follow-ups are likewise mixed across question blocks and are confined to a single life record. These results provide no consistent evidence that restricting access to past video makes post-observation video memory more useful under the tested contracts (Appendix G; Figure 3).

Scope. Across these settings, restricting source access is not sufficient by itself to increase memory contribution. The interaction depends on what information memory retains, how much relevant source information the intervention removes, and how the downstream reader or agent uses the resulting memory. Section 4 establishes access-conditioned contribution under controlled document loss; the studies here delineate where this dependence extends and where it does not.

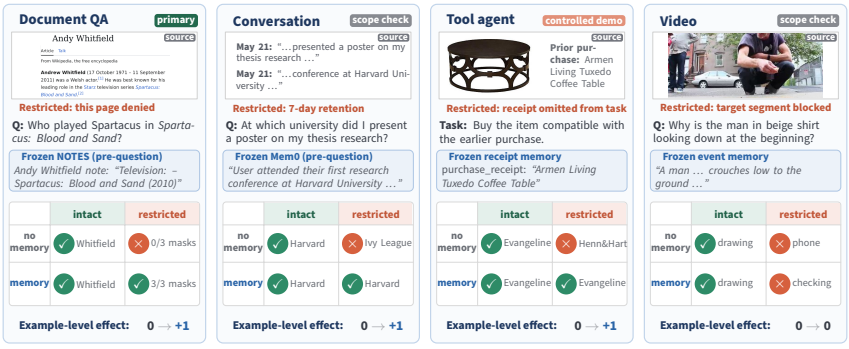


Figure 3: Illustrative MOA instances in document QA, conversation, tool use, and video. Examples were selected post hoc to show access-conditioned contribution and its boundaries, not to estimate prevalence.

6 Implications for Memory Evaluation

A memory-on/off ablation measures contribution relative to the information available to the agent at evaluation time. When source access changes, the measured contribution of the same frozen

memory can change with it. An intact-access ablation therefore estimates $C(A)$, not an access-invariant property of the memory system.

Claims that span source-access regimes should evaluate the same acquired memory under each regime and report both regime-specific contributions and their interaction. At minimum, this requires specifying the memory-acquisition contract, the source-access intervention, the four outcome cells, and the downstream execution and scoring contract (Table 5 in Appendix A).

This conditionality can change the conclusion supported by an evaluation. In the second document cohort, neither FLAT nor Mem0 has a resolved positive contribution with sources intact, whereas both do after document denial. Across the broader set of study–memory pairs, an illustrative interval-based screening rule changes its decision in 8 of 11 rows (Figure 2C; Appendix H.2). The rule illustrates evaluation sensitivity to access regime; it is not a deployment criterion. Intact-only evaluation can also miss costs of removing retained memory after its sources become unavailable.

7 Discussion and Limitations

Source access changes the opportunity for memory to matter. Across the controlled document studies, the marginal contribution of fixed memory varies with source availability. The interaction is not positive by construction. NOTES stores question-independent extracted facts rather than source copies; FLAT has an unresolved strict-EM interaction in the primary study; and the four-action development configuration in the second cohort produces negative interactions. The mixed retention, WebShop, and video results likewise show that source restriction can make memory more consequential without guaranteeing a positive interaction.

The observed interactions need not arise from a single recovery mechanism. In the primary document study, the interaction is concentrated on masks that remove annotated supporting documents. In the second cohort, by contrast, part of the Mem0 interaction arises because a negative intact-access contribution disappears under denial rather than from a large gain under restricted access. A positive MIV therefore indicates that marginal contribution changes across access regimes; it does not, by itself, identify intrinsic memory quality or information recovery as the sole mechanism.

Evaluation and deployment are separate questions. Whether a memory system is useful in practice depends on how often restricted access occurs, its contribution in each regime, and its computational or operational cost. If restriction occurs with probability p, the expected contribution is $(1-p)C(A) + pC(R)$. MOA estimates the regime-specific contributions and their interaction; it does not estimate p or deployment cost.

Scope. The strongest controlled result uses Qwen3-VL-32B, whose intact source-only strict-EM accuracy is 30.0% on the full primary cohort. The result therefore establishes access-conditioned contribution for this reader and execution contract, not invariance across reader capability. Our most controlled identification uses corpus-wide, pre-question document memory on text/table questions. Finally, the source-access changes studied here are controlled interventions rather than observations of naturally occurring access failures or policy changes. Effects are conditional on the tested readers, tasks, memories, and interventions.

8 Conclusion

The marginal contribution of agent memory can depend on whether the information it stores remains obtainable from its original sources. In our controlled document studies, the contribution of the same frozen memory changes with source access, while scope studies show that this dependence varies across tasks, memory constructions, and access interventions.

A memory-on/off ablation therefore measures contribution within the source-access regime in which it is conducted, rather than an access-invariant property of the memory system. MOA operationalizes this distinction by holding acquired memory fixed while crossing memory use with source access. Claims that span access regimes should report memory contribution within each regime and the interaction between them. More broadly, when memory matters depends not only on what the agent remembers, but also on what information remains accessible without that memory.

AI USE STATEMENT

We used generative AI tools for the following tasks with required disclosure: designing and providing feedback on research methodology and experiments (including follow-up evaluation design), implementing methods (analysis code and experimental orchestration), translation, and interpreting results. We did not use generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide ingredients for or write proofs, propose or refine hypotheses, clean or reformat datasets, or support qualitative and thematic data analysis. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing parts of the paper, creating and modifying figures, creating and editing code, and assisting the coding of existing literature. Separately, automated semantic judges scored reader responses as part of the study. Both human audits use two independent human reviews and third-reviewer adjudication; analysis code normalizes and imports the submitted label files without substituting model judgments. An author verified every final literature code against the full text. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The artifact records each study separately. Supplier-source document experiments include frozen populations, reader outputs, scoring rules and delivery checks (Appendices D, D.2, I.2, and I.3). The content-independent study records loss masks, question groups and budget selection (Appendix C). The pre-question LongMemEval study records memory construction, session masks, continuation outputs, both judge labels and paired analysis (Appendix E.1). The joint document follow-up records the corpus memory, disjoint cohort, reader selection, masks and all 3,600 outputs (Appendix B.1). Its 150-question human audit ships both reviewers' labels, the adjudication and the response-to-cell mapping. The source-coupled deletion follow-up records its protocol and analysis code, frozen before new outputs, the deletion payloads and all 1,800 outputs (Appendix B.2). Scripts in the supplement recompute the screening-rule table, offset shares and pooled interaction from the archived analysis files; the literature coding ships its protocol, both passes and adjudications, and the model-judge rescoring ships its inputs, both judges' labels and analysis.

The human audit links the fixed list of 50 questions to 463 distinct responses and 1,350 execution cells. The response-to-cell mapping and final adjudicated labels support recomputation of the human scores and matched automatic comparisons (Appendix E.2). The analysis averages repeated runs within question and resamples history groups together. Original submitted labels and documented file conversions are retained separately from the normalized analysis inputs.

The study registry records source hashes, scoring rules and analysis configurations. Appendix I.4 maps execution and determinism records; Table 3 distinguishes current evidence from earlier analyses.

REFERENCES

Katie Boeckl and Naomi Lefkovitz. NIST Privacy Framework: A tool for improving privacy through enterprise risk management, version 1.0. NIST Cybersecurity White Paper CSWP 01162020, National Institute of Standards and Technology, Gaithersburg, MD, 2020.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Huacan Chai, Yukai Wang, Yingxuan Yang, Dan Peng, Yuanyi Song, Zhihui Fu, Weiwen Liu, Jianghao Lin, Jun Wang, and Weinan Zhang. SMMBench: A benchmark for source-distributed multimodal agent memory. *arXiv preprint arXiv:2605.15710*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3DocRAG: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, et al. MemoryArena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*, 2026.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025.

Eric Hanchen Jiang, Zhi Zhang, Yuchen Wu, Levina Li, Dong Liu, Xiao Liang, Rui Sun, Yubei Li, Edward Sun, Haozheng Luo, Zhaolu Kang, Aylin Caliskan, Kai-Wei Chang, and Ying Nian Wu. Memory as a controlled process: Learned adaptive memory management for LLM agents. *arXiv preprint arXiv:2607.13591*, 2026.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. *arXiv preprint arXiv:2506.06326*, 2025.

Qingcan Kang, Mingyang Liu, Shixiong Kai, Kaichao Liang, Tao Zhong, and Mingxuan Yuan. Learning what to remember: Observability-safe memory retention via constrained optimization for long-horizon language agents. *arXiv preprint arXiv:2606.10616*, 2026.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of NAACL-HLT*, 2025.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*, 2024a.

Myeonghwa Lee, Seonho An, and Min-Soo Kim. PlanRAG: A plan-then-retrieval augmented generation for generative large language models as decision makers. In *Proceedings of NAACL-HLT*, pp. 6537–6555, Mexico City, Mexico, 2024b.

Mukai Li, Qingcheng Zeng, Tianqing Fang, Zhenwen Liang, Linfeng Song, Qi Liu, Haitao Mi, and Dong Yu. Verified critical step optimization for LLM agents. In *Findings of the Association for Computational Linguistics: ACL*, San Diego, California, 2026.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable LLMs with long-term memory. *arXiv preprint arXiv:2311.08719*, 2023.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.

Miles McCain, Thomas Millar, Saffron Huang, Jake Eaton, Kunal Handa, Michael Stern, Alex Tamkin, Matt Kearney, Esin Durmus, Judy Shen, Jerry Hong, Brian Calvert, Jun Shern Chan, Francesco Mosconi, David Saunders, Tyler Neylon, Gabriel Nicholas, Sarah Pollack, Jack Clark, and Deep Ganguli. Measuring AI agent autonomy in practice. https://www.anthropic.com/research/measuring-agent-autonomy, February 2026.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, February 2025.

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.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 247–254, 2011.

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.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Chen Shen. What eviction destroys: A restore-counterfactual audit of forgetting in agent memory. *arXiv preprint arXiv:2609.08279*, 2026.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Zhewen Tan, Yilun Yao, Huiyan Jin, Wenhan Yu, Guoan Wang, Mengyuan Fan, Liang Lu, Feng Liu, Xiangzheng Zhang, Duohe Ma, Tong Yang, and Lin Sun. MemAudit: Post-hoc auditing of poisoned agent memory via causal attribution and structural anomaly detection. *arXiv preprint arXiv:2605.23723*, 2026.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. *arXiv preprint arXiv:2604.20006*, 2026.

Tyler J. VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009.

Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. SCM: Enhancing large language model with self-controlled memory framework. *arXiv preprint arXiv:2304.13343*, 2023.

Yu Wang and Xi Chen. MIRIX: Multi-agent memory system for LLM-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.

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 (ICLR)*, 2025.

Yifan Wu, Lizhu Zhang, Yuhang Zhou, Mingyi Wang, Bo Peng, Serena Li, Xiangjun Fan, and Zhuokai Zhao. Remember when it matters: Proactive memory agent for long-horizon agents. *arXiv preprint arXiv:2607.08716*, 2026.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tiancheng Yang, Matthias Schonlau, and Ilia Sucholutsky. Selective QA over conflicting multi-source personal memory: A diagnostic testbed and method comparison. *arXiv preprint arXiv:2605.30087*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Haiyue Zhang. Credit without ground truth: Auditing step-level credit assignment in LLM agents against executed replay. *arXiv preprint arXiv:2608.19760*, 2026.

Yan Zhang and Shibo Li. ConsistencyGate: Preventing memory contamination in LLM agents via self-consistency admission control. *arXiv preprint arXiv:2607.22962*, 2026.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. *arXiv preprint arXiv:2308.10144*, 2023.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.

A Appendix Roadmap and Evidence Status

Guide to the appendices. The appendices follow the evidence hierarchy of the main text. Appendix B holds the primary joint pre-question document study, its human audit, the source-coupled deletion follow-up and model-judge rescoring. Appendix C holds the disjoint-cohort confirmation under independent loss. Appendix D holds the motivating supplier-source study with its answer-exposure, scoring, reader and implementation-transfer checks. Appendices E, F, and G give the conversational, tool-agent, and video scope checks. Appendix H collects cross-study summaries, Appendix I reproducibility and intervention audits, Appendix J the literature coding, and Appendix K the full limitations. Write-ups of the archived census, predictor, packer and failed-probe studies are omitted from this PDF; their code, result records and phase reports are in the supplementary artifact. They document earlier analyses and are not additional independent confirmations of the main results. Table 3 distinguishes current evidence from superseded or withdrawn estimates and links each to its detailed record.

Global conventions. Unless stated otherwise, intervals are paired 95% bootstrap percentile intervals over the stated resampling unit. Each study keeps its own predeclared primary scorer; other scorers are secondary, rescoring the same outputs is a sensitivity analysis rather than an independent replication, and analyses marked exploratory or post hoc do not support confirmatory claims.

Key result tables are the primary joint result (Table 2), supplier-source estimates (Table 21), six-reader comparison (Table 28), independent-loss study (Table 14), joint-test results (Table 8), source-coupled deletion (Table 11), screening-rule verdicts (Table 60), model-judge rescoring (Table 13), offset shares (Table 59), coding of 19 memory evaluations (Table 67), pre-question absolute cells (Table 41), human audit (Table 43), and exploratory answer-exposure split (Table 23).

Table 3: Status and location of current analyses. Historical, superseded and withdrawn analyses are listed separately in Table 4.

Status	Study or analysis	Role and detailed record
Primary	Pre-question document memory with independent loss	New text/table cohort; NOTES strict EM is the predeclared primary; adjudicated 150-question human audit (Section 4; Appendix B.1).
Disjoint-cohort confirmation	Independent document loss, second cohort	Loss independent of memory; acquisition sees the question (Appendix C).
Motivating case	Supplier-source Document QA	Held-out FRAMES was the confirmatory cohort of its own protocol; here it motivates the controlled tests because supplier identity depends on memory content. M3DocVQA is supporting (Section 5.1; Appendix D).
Scope	Corrective pre-question LongMemEval with human audit; WebShop; Video	Conversational memory, retention and boundaries (Appendices E.1, E.2, F and G).
Current deletion follow-up	Source-coupled deletion of NOTES (2×2)	Same cohort, store and reader; predeclared strict-EM retention interaction; planned KEEP-cell re-execution was not run (Appendix B.2).
Model-judge rescoring	Joint and deletion outputs, Llama and Qwen judges	Secondary to the predeclared strict EM; protocol and analysis fixed before judge output (Appendix B.3).
Post-hoc summaries	Offset shares; pooled independent-loss and joint primaries	Specified after the results; each component is a predeclared primary (Appendix H.1).
Literature coding	19 agent-memory evaluations, 16 double-coded	Protocol fixed before coding; AI-assisted passes with verbatim quotes; adjudications recorded (Appendix J.2).
Current reader sensitivities	Qwen32B; Gemma27B; native-input Llama8B/70B and Mistral7B	All reuse held-out FRAMES. Native/legacy input controls and delivery audits are in Appendices D.3 and D.4.
Corrected scoring	Supplier-source document endpoints	Primary estimates use corrected substring scoring; legacy scores are retained for comparison in Appendix D.2.

Table 5: Reporting checklist for memory claims that span source-access conditions.

	Item	Report	Guards against
❑	**Acquisition**	What the writer saw; when memory was frozen	Question/answer leakage into memory
❑	**Source access**	Restriction and assignment rule, affected population, realized extent	Memory-dependent loss
❑	**Outcomes**	Four absolute scores; regime-specific contributions, MIV, paired intervals	Reading one contribution as invariant
❑	**Execution**	Source/memory delivery, input format, branch isolation, scoring, budgets	Leakage or unmatched execution

Table 6: Evidence for source-conditioned memory contribution, ordered by the confounds each study removes. Scorers differ by study (Section 3); FACT is the supplier study's fact memory. †Directional only: all four human MIV intervals include zero (Appendix E.2).

	Confound removed					
Study (§)	Pre-question memory	Loss indep. of memory	Human audit n/N	Memory	Scorer	MIV, pp [95% CI]
Joint (4.1)	✓	✓	150/300	NOTES	strict EM	**+3.00** [+0.33, +5.74]
				NOTES	human	+6.44 [+1.35, +11.63]
Independent loss (4.2)	✗	✓	–	FLAT	strict EM	+3.17 [+0.25, +6.01]
				Mem0	strict EM	+5.33 [+2.13, +8.61]
LongMemEval (5.2)	✓	✓	50/186†	FLAT	Llama judge	+8.33 [+0.90, +15.50]
				Mem0	Llama judge	+8.96 [+1.78, +15.95]
FRAMES supplier (5.1)	✗	✗	–	FACT	substring EM	+10.43 [+5.21, +16.11]

Table 4: Historical, superseded, withdrawn and archived analyses, retained as an audit record.

Status	Study or analysis	Role and detailed record
Historical reader configurations	Earlier Llama/Mistral runs	Original estimates are retained alongside matched current wrapper comparisons (Appendix D.4).
Superseded	Original LongMemEval construction and substring estimates	The corrective rerun supplies current evidence. Original reader-transfer results remain historical sensitivities, not validation of the corrected stores or semantic endpoints (Appendix E.1).
Replaced or withdrawn	WebShop implementation-transfer contributions	The direct-utility bridge replaces the mismatched direct comparison. Source-access projection contributions are withdrawn; their within-study contrast remains. The original controlled MOA result is unaffected. See Appendix D.6 and Table 52.
Archived diagnostics	Single-write census, packer repairs, predictors	These document earlier analyses and their limits. Their write-ups are omitted from this PDF; code, result records and phase reports are in the supplementary artifact.

No superseded or withdrawn estimate supports any main-text claim.

B Primary Document Evidence

Status. Primary evidence for Section 4.1 and Section 4.3. The joint study, its human audit, the predeclared deletion follow-up and the secondary model-judge rescoring are reported together.

B.1 Design, Results and Human Audit

Question and population. This follow-up tests access dependence when question-independent memory construction and content-independent loss hold together. We exclude 750 M3DocVQA questions found in the prior MOA output archive and questions sharing their supporting documents. Of 959 eligible text/table questions, a fixed group split assigns 60 to development and 300 to final evaluation. The final cohort has 281 supporting-document/duplicate groups. Supporting documents

are disjoint between the two splits and from the scanned prior questions. This is a new evaluation cohort, not a claim about pretraining exposure. Sample size and primary analysis are fixed before final outcomes.

Memory and access. Qwen3-VL-8B processes all $3{,}311$ indexed corpus documents as $4{,}674$ units of at most $12{,}000$ characters. The writer receives titles and text, but no question, reference answer, evidence label or loss mask. NOTES stores generic factual notes, capped at 384 output tokens per unit; FLAT retains the same units verbatim. NOTES is an extraction baseline, not Mem0 or a new memory algorithm. Memory is frozen before final queries. BM25 ($k_1 = 1.5$, $b = 0.75$) selects up to eight units once per question, with a $2{,}048$-token cap; each arm's payload is identical across access conditions. Three fixed hash masks use only group ID, document ID and seed to deny documents at nominal rate 0.5. Denial applies before search ranking and to direct reads.

Reader and execution. Development selection uses intact-access NONE strict EM, without memory outcomes. Qwen3-VL-32B scores 40.00% with sources and 30.00% with all sources closed; Gemma-4-31B scores 25.00% and 21.67%, respectively. Qwen3-VL-32B is selected. Final partial denial differs from development closure; final intact source-only strict EM is 30.00%. The reader has ten source actions and a final answer call. SEARCH returns five 320-character snippets; READ returns the selected indexed section card, up to $2{,}400$ characters. Source and context caps are $12{,}000$ and $24{,}576$ tokens, with at most 200 output tokens per call, temperature zero and prefix caching disabled. Source trajectories may differ across memory arms; access rules and budgets are matched. This changed acquisition and reader contract is not a matched model-scale comparison with the earlier study.

Paired analysis. All $3{,}600$ cells are present: 300 questions by NONE/FLAT/NOTES by intact access/three masks. The predeclared primary contrast is NOTES MIV under normalized strict EM. Masks are averaged within question; $20{,}000$ paired bootstrap draws resample the 281 groups with equal question weights. FLAT, alternative scorers and equal-group weighting are secondary. There are no outcome-based exclusions. Tables 7 and 8 report absolute scores and all scorer contrasts.

Table 7: Joint follow-up: strict-EM accuracy (%) and source-loss drop (pp). Restricted scores average three masks within each of 300 questions. Memory reduces the observed accuracy drop without eliminating it.

Memory	Intact accuracy	Restricted accuracy	Loss L_m
NONE	30.00	24.78	5.22
FLAT	36.33	33.67	2.67
NOTES	33.33	31.11	2.22

Table 8: Joint follow-up: conditional contributions and paired interactions in pp for EM and 0–100 points for token F1. NOTES strict EM is the predeclared primary contrast; other rows are secondary. All 300 questions are retained.

Memory	Scorer	$C(A)$	$C(R)$	MIV [95% CI]
FLAT	Strict EM	$+6.33$	$+8.89$	$+2.56$ $[-0.22, +5.30]$
NOTES	Strict EM	$+3.33$	$+6.33$	$+3.00$ $[+0.33, +5.74]$
FLAT	Token F1	$+7.66$	$+10.81$	$+3.15$ $[+0.49, +5.72]$
NOTES	Token F1	$+4.54$	$+7.69$	$+3.15$ $[+0.48, +5.87]$
FLAT	Substring EM	$+10.67$	$+12.78$	$+2.11$ $[-0.91, +5.20]$
NOTES	Substring EM	$+7.67$	$+9.11$	$+1.44$ $[-1.90, +4.78]$

Interpretation and sensitivity. NOTES contributes $+3.33$pp $[-0.64, +7.09]$ under intact access and $+6.33$pp $[+2.51, +10.18]$ under denial. Its interaction is $+3.00$pp $[+0.33, +5.74]$; equal-group weighting gives $+3.04$pp $[+0.20, +5.93]$. FLAT's strict-EM interaction is $+2.56$pp $[-0.22, +5.30]$ (equal-group: $+2.61$pp $[-0.36, +5.46]$). Token-F1 intervals exclude zero for both memories under either weighting; both substring intervals include zero. The primary result extends the evidence to the joint conditions, with a modest effect and scorer-dependent precision. It does not establish a recovery mechanism, ranking reversal or frontier-model generality. Different significance outcomes do not establish a difference between memory systems.

Audit and provenance. Checks find zero blocked-source deliveries, zero changed memory payloads, complete paired cells and unchanged acquired memory. Resources were reduced from two identical TP2 reader replicas to one after 2,625 cells; some groups therefore span physical replicas. Original outputs and contracts are preserved, and later rows record the change. Intermediate summaries were inspected, but all 300 questions and 3,600 cells were completed under the fixed analysis plan. Device and batch changes may affect numerical replay. The artifact directory `joint_document/` contains the protocol, reader selection, execution amendments, frozen inputs, outputs, `analysis.json` and `RESULTS.csv`.

Human audit. A pre-outcome hash sample selects 150 final questions for independent paired human scoring. Its 1,800 cells map to 647 distinct question–response pairs, shown without conditions or automatic scores. Two external annotators, neither an author, labelled every pair independently, without seeing each other's labels; reviewer B worked from reviewer A's row order rather than a separately shuffled order. They agreed on 612 of 647 pairs (94.59%; Cohen's $\kappa = 0.885$), and a third reviewer, shown neither label, adjudicated the 35 disagreements. Identical question–response pairs share labels across cells, and no missing label was replaced by a model score. The analysis averages masks within question, assigns equal question weights and uses 20,000 paired group-bootstrap draws, as in the automated analysis. Table 9 reports human MIV beside every automated scorer on the same questions. Human scoring credits many answers that strict matching rejects: intact no-memory accuracy is 49.3% under human labels and 29.3% under strict EM. The human interaction is positive for both memories under either reviewer's labels alone, so it does not depend on the adjudication.

Table 9: Human and automated scoring of the same 150-question sample (pp; points for token F1; paired 95% group-bootstrap intervals). The sample was fixed before any outcome; its automated effects exceed those of the full cohort by sampling variation, so human estimates are compared within the sample. Model-judge values for the sample are in Table 13.

Scorer (150 questions)	$C(A)$ NOTES	$C(R)$ NOTES	MIV NOTES	MIV FLAT
Adjudicated human	+10.00	+16.44	+6.44 [+1.35, +11.63]	+6.44 [+1.55, +11.48]
Reviewer A only	+12.00	+18.00	+6.00 [+1.10, +11.04]	+8.00 [+3.53, +12.67]
Reviewer B only	+9.33	+15.56	+6.22 [+1.36, +11.11]	+6.22 [+1.55, +10.96]
Strict EM	+3.33	+8.22	+4.89 [+1.12, +8.84]	+3.78 [−0.44, +7.95]
Token F1	+4.72	+10.70	+5.97 [+2.38, +9.74]	+5.17 [+1.40, +8.98]
Substring EM	+8.67	+13.11	+4.44 [+0.44, +8.61]	+5.56 [+1.79, +9.46]

An observed example. Figure 4 shows one question from the human-audited sample. With intact access, all three configurations answer correctly. Under each of the three masks the supporting document is denied; without memory the reader then names the lead of a different production or the actor who later took over the role, while both memories still deliver the relevant facts.

Question: who is the actor who played spartacus in spartacus blood and sand
Reference answer: Andy Whitfield

Delivered NOTES entries (verbatim excerpts; built before any question)

1. *Andy Whitfield* note: "Television: / – Spartacus: Blood and Sand (2010) – filmed in New Zealand"
2. *Liam McIntyre* note: "Succeeded Andy Whitfield as Spartacus in the series after Whitfield's death from Non-Hodgkin lymphoma."

Access change: The single supporting document (*Andy Whitfield*) is denied under all three random masks; the frozen memory is identical across conditions.

Source access	Without memory	With FLAT	With NOTES
Intact access	Andy Whitfield Correct	Andy Whitfield Correct	Andy Whitfield Correct
Mask 1	Kirk Douglas Incorrect	Andy Whitfield Correct	Andy Whitfield Correct
Mask 2	Liam McIntyre Incorrect	Andy Whitfield Correct	Andy Whitfield Correct
Mask 3	Kirk Douglas Incorrect	Andy Whitfield Correct	Andy Whitfield Correct

Figure 4: A qualitative example from the primary joint study (M3DocVQA). Strict EM and both independent human reviewers agree on every displayed judgment. The case was selected after inspecting outcomes, from the pre-outcome human-audited sample, to illustrate access dependence; it does not estimate prevalence, and here both memory conditions continue to answer correctly under denial.

Support-hit stratification (exploratory). Random masks deny half of the documents, so a question–mask pair need not lose any gold supporting document. We classify each of the 900 pairs using only the frozen masks and the dataset's gold support labels, never outcomes: 576 pairs deny at least one supporting document ("hit"), 324 deny none ("miss"), and 351 deny every supporting document. MIV is computed per pair and averaged within stratum, with 20,000 group-bootstrap draws (Table 10); the hit and miss strata recombine exactly to the reported full-cohort estimates. The split was defined after the primary result and is not a confirmatory test; another document can still supply the same value, so a hit does not guarantee loss of the information.

Table 10: Joint-study MIV (pp, 95% group-bootstrap intervals) by whether the random mask denies a gold supporting document. Strata use masks and gold labels only; exploratory.

Scorer	Memory	Hit (576)	Miss (324)	All denied (351)	Hit − miss
Strict EM	NOTES	+4.69 [+0.71, +8.66]	+0.00 [−2.29, +2.28]	+7.12 [+2.70, +11.90]	+4.69 [+0.34, +9.11]
	FLAT	+3.99 [−0.17, +8.06]	+0.00 [−1.52, +1.52]	+6.27 [+1.95, +10.81]	+3.99 [−0.17, +8.07]
Token F1	NOTES	+4.99 [+1.02, +9.04]	−0.11 [−2.34, +2.07]	+8.17 [+3.69, +12.93]	+5.11 [+0.67, +9.56]
	FLAT	+4.78 [+0.84, +8.63]	+0.27 [−1.20, +1.73]	+7.66 [+3.50, +12.11]	+4.51 [+0.55, +8.37]
Substring EM	NOTES	+3.30 [−1.54, +8.23]	−1.85 [−4.69, +0.83]	+5.98 [+0.00, +12.12]	+5.15 [−0.17, +10.59]
	FLAT	+3.99 [−0.36, +8.33]	−1.23 [−3.59, +0.93]	+6.84 [+1.49, +12.15]	+5.23 [+0.86, +9.56]

B.2 Source-Coupled Deletion of Pre-Question Notes

Question and design. This follow-up asks whether the joint-study NOTES interaction is carried by notes derived from the documents that become unavailable, and what a source-coupled deletion policy costs. It reuses the frozen 300-question cohort, NOTES store, Qwen3-VL-32B reader, tools, budgets, prompts, masks and scorer of Appendix B.1. Units are document-local, so provenance is exact. For question i and mask k, DROP recall deletes every unit whose source document mask k denies for that question's group and ranks the remaining units with the same BM25 recall; the next-ranked units fill the eight-unit, 2,048-token payload. The design crosses this memory factor with access to the same documents: KEEP or DROP$_k$ memory, with all documents accessible (o) or mask k applied (s). KEEP cells are the original joint-study NOTES cells, whose payload bytes the new recall reproduces exactly; the 1,800 DROP cells are new. Deletion changes the delivered payload for 96.4% of question–mask pairs, removing on average 2.7 of 5.4 delivered units.

Estimands and inference. With masks averaged within question, $K(o)$, $K(s)$, $D(o)$ and $D(s)$ denote mean accuracy. The predeclared primary estimand is the retention interaction $I = [K(s) -$

$D(s)] - [K(o) - D(o)]$ under normalized strict EM: the extra value of notes from denied documents when those documents are denied. $K(s) - D(s)$ is the loss-regime cost of source-coupled deletion. Under the conditional policy itself, intact access triggers no deletion, so its intact-access cost is zero by construction. The experimental $D(o)$ arm instead forces deletion of the mask-indexed notes while their documents remain readable; $K(o) - D(o)$ is therefore a placebo deletion contrast, not the policy's intact-access action. With NONE cells N, the joint interaction decomposes as $\text{MIV} = I + \{[D(s) - N(s)] - [D(o) - N(o)]\}$, where the second term is the interaction of the remaining notes. Intervals use 20,000 paired bootstrap draws over the 281 groups with equal question weights (seed 20260923). The protocol and analysis code were frozen before any new output.

Table 11: Source-coupled deletion on the joint cohort (300 questions, three masks). Top: accuracies (%; token F1 on a 0–100 scale). Bottom: contrasts in pp (points for token F1) with paired 95% intervals. Strict EM is primary.

Quantity	Strict EM	Token F1	Substring EM
$K(o)$ / $D(o)$	33.33 / 30.78	40.99 / 38.35	49.67 / 45.89
$K(s)$ / $D(s)$	31.11 / 26.89	38.96 / 33.99	48.22 / 41.56
Placebo $K(o) - D(o)$	+2.56 [+0.44, +4.78]	+2.64 [+0.61, +4.79]	+3.78 [+1.43, +6.29]
Deletion cost $K(s) - D(s)$	+4.22 [+1.46, +7.06]	+4.97 [+2.31, +7.75]	+6.67 [+3.59, +9.85]
Retention interaction I	+1.67 [−0.76, +4.09]	+2.32 [+0.05, +4.61]	+2.89 [+0.22, +5.65]
Joint MIV (KEEP)	+3.00 [+0.33, +5.74]	+3.15 [+0.48, +5.87]	+1.44 [−1.90, +4.78]
Remaining-notes interaction	+1.33 [−0.68, +3.41]	+0.83 [−1.24, +2.86]	−1.44 [−4.00, +1.08]

Results. Source-coupled deletion costs 4.22pp $[+1.46, +7.06]$ strict EM under loss, and its interval excludes zero under all three scorers. An intact-access evaluation cannot observe this cost. Part of it reflects the value of these notes regardless of access: deleting them while their documents remain readable costs 2.56pp $[+0.44, +4.78]$. The primary retention interaction is positive, $+1.67$pp $[-0.76, +4.09]$, but its strict-EM interval includes zero; the token-F1 and substring intervals exclude zero. In the strict decomposition, 1.67 of the 3.00pp joint interaction is attributed to notes from denied documents and 1.33pp to the remaining notes; neither component is individually resolved. These results are consistent with information retained past source loss but do not establish it as the sole mechanism.

Exploratory analyses. The following analyses were specified after the primary result and are not predeclared. Stratum membership uses only the frozen KEEP payload, the mask and the gold supporting-document labels, which enter no memory, recall, mask or reader input; it does not depend on outcomes. For each question–mask pair, we ask whether deletion removes a delivered KEEP unit derived from a supporting document. It does in 457 of 900 pairs (242 questions). Table 12 reports the contrasts within each stratum, averaging a question's qualifying masks, and their difference with groups resampled jointly. Where a supporting-document note is deleted, the placebo cost is 5.51pp and the cost under loss is 10.81pp, so these notes are worth about twice as much once their documents are denied; elsewhere all three contrasts are near zero. The pattern holds under all three scorers. Per mask, the primary interaction is $+2.33$, -1.00 and $+3.67$pp, so the aggregate is not uniform across masks; equal-group weighting gives $+1.72$pp $[-0.83, +4.27]$.

Table 12: Exploratory split of the source-coupled deletion contrasts by whether deletion removes a delivered note derived from a gold supporting document. Cost is $K(s) - D(s)$ and placebo $K(o) - D(o)$ (pp; points for token F1; paired 95% group-bootstrap intervals). Not predeclared.

Stratum	Contrast	Strict EM	Token F1	Substring EM
Supporting note deleted (457 pairs, 242 questions)	I	+5.30 [+0.80, +9.79]	+7.12 [+2.85, +11.38]	+8.75 [+3.97, +13.62]
	Cost	+10.81 [+6.08, +15.67]		
	Placebo	+5.51 [+1.96, +9.21]		
No supporting note deleted (443 pairs, 248 questions)	I	−1.61 [−3.80, +0.45]	−1.49 [−3.67, +0.57]	−0.87 [−3.27, +1.51]
	Cost	−1.75 [−4.20, +0.60]		
	Placebo	−0.13 [−2.37, +2.03]		
Difference	I	+6.92 [+1.82, +11.94]		

Execution and amendments. All 1,800 DROP cells completed with zero blocked-source deliveries and zero empty answers. The vLLM server used the frozen model snapshot, dtype, context length, seed, temperature and disabled prefix caching; for throughput, two replicas used 16 concurrent sequences with CUDA graphs instead of the original 4 and eager execution, and one replica reverted to the original settings after an outcome-blind throughput comparison. Replicas were added and repaired by GPU availability; job-to-replica assignment followed queue order. Four transport timeouts were re-run from the start as permitted transport retries. A predeclared 1,200-cell re-execution of the KEEP cells under the new servers was not run, by an author decision recorded before any analysis; the 100 KEEP cells produced during shutdown are excluded. KEEP and DROP cells therefore come from different execution epochs; the interaction is less sensitive to epoch effects common to the two access conditions, whereas the deletion and placebo level contrasts may also reflect epoch differences.

B.3 Model-Judge Rescoring of the Joint and Deletion Studies

Design. The joint study and the deletion 2×2 predeclare lexical scorers. As a secondary semantic check, we rescored their outputs with the two judges of the LongMemEval study: Llama-3.1-8B-Instruct (primary) and Qwen3-8B (sensitivity). Both use the official LongMemEval answer-check prompt in its generic factual form, with temperature zero, eight output tokens and thinking disabled. The 3,600 joint cells and 1,800 deletion cells yield 1,643 distinct question–response pairs; each judge labels every pair once, with no unparseable output. The protocol and analysis code were fixed before any judge output; one coding error in the analysis script, which scored unparseable labels as zero rather than missing, was corrected before any output existed and did not affect results. Estimands, weights and resampling match the lexical analyses.

Table 13: Lexical and model-judge scores of the same outputs (pp, paired 95% group-bootstrap intervals). Strict EM is the predeclared primary; the judges are secondary. The 150-question rows are the pre-outcome sample prepared for human scoring.

Contrast	Cohort	Strict EM	Llama judge	Qwen judge
NOTES $C(A)$	300	+3.33 [−0.64, +7.09]	+5.67 [+0.98, +10.48]	+7.00 [+2.07, +11.80]
NOTES $C(R)$	300	+6.33 [+2.51, +10.18]	+9.67 [+4.91, +14.46]	+11.78 [+7.38, +16.16]
NOTES MIV	300	+3.00 [+0.33, +5.74]	+4.00 [+0.10, +8.00]	+4.78 [+1.00, +8.61]
FLAT MIV	300	+2.56 [−0.22, +5.30]	+5.44 [+1.53, +9.40]	+6.33 [+2.58, +10.04]
NOTES MIV	human-audit 150	+4.89 [+1.12, +8.84]	+4.22 [−1.34, +9.75]	+8.67 [+3.33, +14.22]
FLAT MIV	human-audit 150	+3.78 [−0.44, +7.95]	+8.22 [+2.49, +14.04]	+9.78 [+4.61, +15.21]
Deletion I	300	+1.67 [−0.76, +4.09]	+3.22 [+0.34, +6.13]	+1.78 [−1.06, +4.65]
Deletion cost	300	+4.22 [+1.46, +7.06]	+7.11 [+3.64, +10.61]	+6.11 [+2.94, +9.43]

Reading the judges. The judges credit semantically equivalent answers that strict matching rejects: cell accuracy is 58.1% (Llama) and 52.8% (Qwen) against 30.7% under strict EM, with cell-level agreement of 71.4% and 77.9%. On the 647 human-labelled responses of this cohort (Appendix B.1), they agree with the adjudicated human labels on 81.0% (Llama) and 91.8% (Qwen) of responses, against 73.9% for strict EM; on LongMemEval the corresponding figures were 77.97% and 94.82%. All five automated scorers give positive interactions for both memories and for the deletion contrast. The judges do not replace the predeclared strict-EM estimands. The inputs, both judges' labels and the analysis are in the supplement.

C Independent-Loss Confirmation on a Second Cohort

Status. Disjoint-cohort confirmation for Section 4.2. Acquisition sees the question, so this study removes memory-dependent loss but not question-dependent construction.

Table 14: Independent source loss on a second cohort: both memories contribute more once documents are denied (strict EM; paired 95% intervals over 330 question groups; Appendix C).

	Accuracy (%)		Contribution of memory (pp)		Interaction (pp)
Memory	A: intact	R: denied	$C(A)$: intact	$C(R)$: denied	MIV $= C(R) - C(A)$
None	29.25	25.17	–	–	–
FLAT	28.50	27.58	-0.75 [$-3.55, +2.02$]	$+2.42$ [$+0.08, +4.81$]	$+3.17$ [$+0.25, +6.01$]
Mem0	27.50	28.75	-1.75 [$-5.06, +1.50$]	$+3.58$ [$+0.66, +6.67$]	$\mathbf{+5.33}$ [$+2.13, +8.61$]

Weighting and paired inference. The estimand gives each question equal weight. Intervals use a cluster bootstrap over the 330 question groups, preserving this weighting and moving all arms, conditions and mask seeds of each sampled group together. Equal-group weighting changes both estimates and intervals. Mem0's interaction remains positive under all three scorers; FLAT's strict-exact-match interval becomes $[-0.31, +6.00]$ and includes zero (Table 16). Each question's three mask seeds are averaged within it, not treated as independent samples.

Population. The pool is the 2,441-question M3DocVQA sample set. A frozen exposure manifest over every artifact of every earlier study in this paper marks 300 questions as exposed; the remaining 2,141 are the draw pool. Questions are grouped by duplicate question–answer key and by shared supporting document, 49 are dropped for supporting documents absent from the card index, leaving 2,092 eligible questions in 1,662 groups. From these 450 were drawn by seeded hash of the group key and split into 50 development and 400 final questions, with the split constrained so that no supporting document is shared across it. The intersection of the final cohort with the exposed set is empty, as is its intersection with the development cohort and with the screen. All 400 final questions yield complete six-cell outcomes; none is excluded after the fact.

Table 15: Order in which the budget was chosen and the final cohort was run. The two cohorts are disjoint by question (0 shared) and by supporting document (0 shared), and the final cohort shares no question with any earlier study in this paper. The freeze precedes the final run's first episode. Because the selection used only development questions, the final estimate is not an adaptively selected one; had a final-cohort quantity entered the choice, the result would instead have to be reported as a follow-up analysis.

step	cohort	what was observed	artifact
four-call run	development, 50	all arms, development only	`budget4_superseded/`
access screen	development, 50	memory-off arms only	`corpus_screen_summary`
budget frozen at 10	–	–	`E1_LOCK.json`
development run	development, 50	all arms, development only	`episodes_development`
final run	final, 400	all arms	`episodes_final`

Execution. One vLLM server hosts Qwen3-VL-8B-Instruct at a fixed snapshot for both the memory builder and the consumer, greedy, with prefix caching disabled. Retrieval is BM25 over 12,839 section cards spanning 3,311 documents, one hit per document, five results per search. Acquisition runs two cycles at a ten-action budget over the unmasked corpus, keeps the full history regardless of outcome, and excludes the gold answer, the evaluation label and the loss mask from its inputs. What is handed to both memory systems is the raw observed history (the query issued, the text returned and the source identifier) rather than the short fact projection the earlier studies used. After the branch the snapshot is read-only, and `Snapshot.verify()` re-derives its digest after every episode.

Budget. The ten-action cap counts `search` and `read` actions. An invalid reply costs a turn. The automatic memory lookup runs at every decision, is identical in shape across arms, returns nothing in the memory-off arm, and does *not* count against the budget; model calls are therefore not capped at ten and are reported separately. The collector and the consumer are given the same budget, so that the memory arms do not enter the branch with less experience than the memory-off arm can gather for itself afterwards. Table 15 records the order in which the budget was fixed and the final cohort was run.

Analysis. For each question the three mask seeds are averaged into a single R cell before any contrast is formed, so the design has 400 units and not 1,200. The estimand is the mean over questions

of $C(R) - C(A)$. Intervals resample the 330 question groups with replacement, 20,000 draws; every arm, access condition and mask seed of a drawn group moves together. The primary interval is a cluster ratio bootstrap, which preserves the per-question weighting of the estimand; the group-mean variant, which weights each group equally regardless of size, is given below. Table 16 reports both weightings. The FLAT strict-EM interval excludes zero only with equal question weights.

Table 16: The interaction under the two weightings of the group bootstrap (pp for EM; 0–100 points for token F1). The point estimates differ because 115 of the 400 questions sit in the 45 groups that hold more than one, so the two weightings do not coincide: the group sizes are 285 groups of one, 30 of two, 8 of three, 4 of four and 3 of five.

		question-weighted		group-mean	
path	scorer	MIV	95% CI	MIV	95% CI
FLAT	strict EM	$+3.17$	$[+0.25, +6.01]$	$+2.90$	$[-0.31, +6.00]$
	corrected substr	$+4.00$	$[+0.72, +7.27]$	$+4.17$	$[+0.65, +7.64]$
	token F1	$+4.51$	$[+1.60, +7.37]$	$+4.36$	$[+1.15, +7.52]$
Mem0	strict EM	$+5.33$	$[+2.13, +8.61]$	$+4.99$	$[+1.70, +8.29]$
	corrected substr	$+6.75$	$[+3.21, +10.34]$	$+6.97$	$[+3.22, +10.72]$
	token F1	$+6.98$	$[+3.76, +10.26]$	$+6.77$	$[+3.35, +10.24]$

Where the interaction comes from: evidence modality. The retrieval tools index text section cards only, so for a question whose evidence is an image they cannot supply it. `metadata.modalities` is a dataset field fixed before any outcome existed, so splitting on it uses no model output. We report the split because it qualifies the E1 estimate rather than supporting it.

Table 17: The content-independent study by evidence modality, strict exact match. Absolute cells are percentages, contributions and interactions percentage points. The two strata resample together, so the between-stratum interval accounts for both being estimated on one draw. Question-weighted, they pool exactly to the reported $+3.17$ and $+5.33$. L_0 is the source-access loss without memory. It is *not* an upper bound on MIV: since $\text{MIV} = L_0 - L_1$, a negative L_1 makes MIV exceed it, which is what happens for Mem0 on image questions.

stratum	path	$C(A)$	$C(R)$	MIV
text or table, n=248, $L_0 = +4.57\ [+1.05, +8.19]$				
absolute A/R: NONE 30.24/25.67, FLAT 29.03/28.63, Mem0 33.06/29.97				
	FLAT	-1.21	$+2.96$	$+4.17$
		$[-4.86, +2.42]$	$[+0.14, +5.76]$	$[+0.42, +7.84]$
	Mem0	$+2.82$	$+4.30$	$+1.48$
		$[-1.20, +6.97]$	$[+0.69, +7.94]$	$[-2.48, +5.44]$
image, n=152, $L_0 = +3.29\ [-1.60, +8.15]$				
absolute A/R: NONE 27.63/24.34, FLAT 27.63/25.88, Mem0 18.42/26.75				
	FLAT	$+0.00$	$+1.54$	$+1.54$
		$[-4.49, +4.51]$	$[-2.56, +6.00]$	$[-3.47, +6.22]$
	Mem0	-9.21	$+2.41$	$+11.62$
		$[-15.48, -3.54]$	$[-2.43, +7.75]$	$[+6.10, +17.57]$

The two memory paths respond differently across the strata (Table 17). Both have positive loss-condition contributions on text or table ($+2.96\ [+0.14, +5.76]$ for FLAT and $+4.30\ [+0.69, +7.94]$ for Mem0), so neither is useless where the tools work. What differs is the increase over intact access: it is resolved from zero for FLAT, $+4.17\ [+0.42, +7.84]$, and not for Mem0, $+1.48\ [-2.48, +5.44]$. We do not report a within-stratum contrast between the two systems, so this does not establish that they differ.

Mem0's aggregate is concentrated in the other stratum: $+11.62\ [+6.10, +17.57]$ on image questions against $+1.48$ on text or table, a difference of $-10.14\ [-17.31, -3.18]$. The two secondary scorers agree in the direction of that difference, -7.15 and -7.95, though one of their intervals contains zero ($[-14.83, +0.31]$ and $[-15.34, -0.92]$). Arithmetically, of the $+5.33$ E1 estimate about $+4.42$ points come from the image stratum and about $+0.92$ from text or table.

What produces the image figure is not a large loss-condition contribution ($+2.41$ there, interval containing zero) but a strongly negative intact-access one, $-9.21\ [-15.48, -3.54]$. The corresponding

L_1 is -8.33 points: the memory-on arm is *more* accurate with documents denied ($18.42 \rightarrow 26.75$) than with them reachable. So this is not memory compensating for information the consumer lost. It is a different response pattern, in which restricting sources coincides with the removal of a cost that Mem0's payload imposes while the corpus is reachable. We have not isolated why. Distraction, a change in what retrieval returns, and a change in tool use are all consistent with it, and we do not claim a mechanism.

A caution on the stratum labels. That the text-only tools cannot supply image evidence is established. It does not follow that blocking textual documents changes nothing for those questions: the image stratum's L_0 is $+3.29$ $[-1.60, +8.15]$, uncertain rather than demonstrably zero. We therefore read the text-or-table stratum as the one where the intervention acts on evidence the tools can actually serve, not as the only one where it acts at all. The decomposition was not predeclared, we do not treat it as a test, and the population of record remains all 400 questions.

The budget contrast. Both budgets ran the same 50 development questions under the same three mask seeds, the same ρ, the same prompts and the same tools, so $\Theta = \mathrm{MIV}_{10} - \mathrm{MIV}_4$ is paired at the question group over 37 groups and needs no new inference; it re-analyses the two stored episode files. What differs besides the consumer budget is the collector budget, which moved with it; the acquired histories are byte-identical on 0 of 50 questions between the two runs, so the memory contents differ as well and the contrast does not isolate the consumer. Table 18 gives it for both paths and all three scorers.

Table 18: Paired budget contrast on the development cohort, 50 questions in 37 groups (pp for EM; 0–100 points for token F1). These intervals do not account for the fact that the budget was selected on this same development set, so they are descriptive rather than confirmatory, and are reported to bound the E1 estimate rather than to support a claim about budgets in general. The disjoint 400-question final is unaffected: no final outcome entered the selection.

path	scorer	MIV_4	MIV_{10}	Θ	95% CI
FLAT	strict EM	-5.33	$+4.67$	$+10.00$	$[+1.23, +19.87]$
	corrected substr	-5.33	$+6.00$	$+11.33$	$[+2.42, +21.38]$
	token F1	-6.11	$+5.24$	$+11.36$	$[+2.62, +20.53]$
Mem0	strict EM	-7.33	$+3.33$	$+10.67$	$[+1.55, +21.43]$
	corrected substr	-7.33	$+6.67$	$+14.00$	$[+4.55, +25.19]$
	token F1	-6.22	$+5.03$	$+11.26$	$[+1.86, +22.17]$

Audit. All 4,800 continuations and 1,200 closed-book continuations are present with no failure, no duplicate and no re-run. Across the development run the mask suppressed 14,726 search hits and refused 61 reads; no denied document reached a consumer through a chunk, an alias, a cached search hit or a direct read, and no mask edited a memory snapshot. Evidence modality is a dataset field rather than an outcome, and the final split is 248 text or table and 152 image questions; the text card index cannot serve image evidence, which bounds what the tools can contribute on the latter. Plan, lock, screen, code hashes, episode files, per-cell outputs and the two re-analyses are under `e1_content_independent/` in the supplement, whose `REPRODUCE.md` gives the two commands that regenerate every number in this section without a model, and the question-id check that the development and final cohorts are disjoint.

Support-hit stratification (exploratory). As in Appendix B.1, each question–mask pair is classified using only the frozen masks and the dataset's gold support labels: a "hit" denies at least one supporting document. The hit and miss strata recombine exactly to the reported estimates. Table 19 gives strict EM; token F1 and corrected substring EM give larger hit-minus-miss differences with intervals above zero for both memories, overall and on text/table questions (supplementary artifact). On text/table questions Mem0's hit-stratum interval includes zero, although its hit-minus-miss difference does not. The split is exploratory.

Table 19: Second-cohort MIV (strict EM, pp, 95% group-bootstrap intervals) by whether the random mask denies a gold supporting document. Exploratory.

Questions	Memory	Hit MIV (pairs)	Miss MIV (pairs)	Hit − miss
All (400)	FLAT	$+4.85$ [$+1.17$, $+8.62$] (804)	-0.25 [-4.05, $+3.45$] (396)	$+5.10$ [$+0.17$, $+10.04$]
	Mem0	$+7.96$ [$+3.76$, $+12.26$] (804)	$+0.00$ [-3.44, $+3.97$] (396)	$+7.96$ [$+2.83$, $+12.92$]
Text/table	FLAT	$+7.28$ [$+1.92$, $+12.78$] (467)	-1.08 [-4.63, $+2.46$] (277)	$+8.36$ [$+2.30$, $+14.56$]
	Mem0	$+4.07$ [-1.68, $+9.94$] (467)	-2.89 [-6.13, $+0.33$] (277)	$+6.96$ [$+1.17$, $+12.89$]

Ordering of the two memories (post hoc). Table 20 contrasts Mem0 with FLAT within each access regime, using the same paired group bootstrap (20,000 draws, seed 20260926). On the full cohort the point estimates reverse the ordering under every scorer: FLAT's contribution is larger under intact access and Mem0's under denial. No interval excludes zero, so the reversal is directional only. It is confined to image questions, where FLAT's resolved intact-access lead (-9.21pp for Mem0 minus FLAT) disappears under denial; on text/table questions Mem0's contribution is larger in both regimes, and the MIV difference has the opposite sign. This analysis was specified after the results and does not establish a ranking reversal. Code and output are `memory_ordering.py` and `memory_ordering.json` under `e1_content_independent/`.

Table 20: Second-cohort contrasts between the two memories (pp, 95% paired group-bootstrap intervals). $C(R)$ averages the three masks. Post hoc.

Questions	Scorer	$C(A)$: Mem0 − FLAT	$C(R)$: Mem0 − FLAT	MIV: Mem0 − FLAT
All (400)	strict EM	-1.00 [-4.09, $+2.06$]	$+1.17$ [-1.30, $+3.75$]	$+2.17$ [-0.58, $+5.00$]
	token F1	-1.04 [-4.09, $+2.03$]	$+1.44$ [-0.96, $+3.96$]	$+2.48$ [-0.17, $+5.20$]
	substring EM	-1.25 [-4.71, $+2.28$]	$+1.50$ [-1.08, $+4.20$]	$+2.75$ [-0.25, $+5.80$]
Text/table (248)	strict EM	$+4.03$ [$+0.00$, $+8.09$]	$+1.34$ [-1.82, $+4.69$]	-2.69 [-5.71, $+0.27$]
Image (152)	strict EM	-9.21 [-14.49, -4.52]	$+0.88$ [-3.14, $+4.94$]	$+10.09$ [$+4.97$, $+15.75$]

D Motivating Supplier-Source Study and Robustness

Status. Motivating case for Section 5.1. Supplier identity is derived from memory content, so these results motivate rather than establish the main claim; the subsections report answer exposure, scoring, reader and implementation-transfer checks.

Table 21: Supplier-source study (motivating case). Blocking the documents that supplied stored facts increases memory contribution. Held-out FRAMES is the confirmatory cohort within this supplier-source protocol; M3DocVQA is supporting. Corrected substring EM with paired 95% bootstrap intervals.

	Contribution of memory (pp)		Interaction (pp)
Cohort (n)	$C(A)$: available	$C(U)$: blocked	MIV $= C(U) - C(A)$
FRAMES held-out (211)	$+3.32$ [-1.42, $+8.06$]	$+13.74$ [$+8.53$, $+19.43$]	$\mathbf{+10.43}$ [$+5.21$, $+16.11$]
FRAMES dev (218)	$+5.05$ [$+0.46$, $+10.09$]	$+20.64$ [$+14.68$, $+26.61$]	$+15.60$ [$+10.09$, $+21.56$]
M3DocVQA (195)	-0.51 [-5.13, $+4.10$]	$+20.00$ [$+13.33$, $+26.67$]	$+20.51$ [$+13.33$, $+27.69$]

Replay is the measurement instrument here, so its settings are part of the method rather than configuration. All runs use vLLM 0.16.0. Except where a reader is named otherwise (the scale and family comparisons below), they use a pinned snapshot of `Qwen3-VL-8B-Instruct`, revision `0c351dd0. . .`, with

`enable_prefix_caching`	`False`
`enforce_eager`	`True`
`tensor_parallel_size`	1
`temperature` / `top_p`	0.0 / 1.0
max new tokens	64 reader, 96 planner
`token_budget`	14000
`max_model_len`	16384
`top_k` cards	10

The first line is not a tuning choice. vLLM enables prefix caching by default, and in the tested engine configuration replay agreement depended on that setting. Because the study replays thousands of trajectories, cache state differs systematically between the base and suppressed conditions, and averaging over runs does not remove a bias that is correlated with the condition. We did not trace the discrepancy to logits and do not identify its numerical mechanism.

Shard integrity. Each source-access run is sharded four ways and merged only after every shard reports completion. Expected counts are recorded in the manifest ($44{+}47{+}52{+}52 = 195$ for M3DocVQA), and the scoring script asserts them, together with zero blocked-source cards delivered, zero memory/no-memory packet mismatches, and the exclusion of smoke outputs from the merge. Two defects during this work motivated the checks: a patch that truncated a script's entry point while leaving its exit status zero, and a merge glob that silently included a smoke file. Neither reached a reported number, and the assertions now fail loudly where both were silent.

The scale-comparison reader. The within-family scale runs use `Qwen/Qwen3-VL-32B-Instruct` at snapshot `0cfaf48183f594c314753d30a4c4974bc75f3ccb`, from the same local cache and the same vLLM version as the primary reader, at tensor-parallel size 2 across two GPUs, maximum model length 16,384, greedy decoding ($T{=}0$, 64 new tokens), the same prompt text, the same `limit_mm_per_prompt` and the same refusal-parsing rule. Population and metric are unchanged and nothing was retuned. All three arms were run: $C(\text{full}) = +4.27$pp, $C(\text{blocked}) = +13.74$pp, $C(\text{placebo}) = +2.84$pp, giving $\text{MIV} = +9.48$pp $[4.27, 14.69]$ and $\text{MIV}_{\text{src}} = +10.90$pp $[5.21, 16.59]$. Outputs are `p28_32b_{0,1}.jsonl`.

Interval and test conventions. Bracketed intervals are 95% percentile bootstrap intervals resampling questions as clusters, 20,000 resamples. Where a p-value is given it is a two-sided paired sign-flip permutation test on the same clustered differences, not a normal approximation.

Historical Llama reader. The original cross-reader runs use `meta-llama/Llama-3.1-8B-Instruct` at snapshot `0e9e39f249a16976918f6564b8830bc894c89659`, loaded from the same local cache with the same vLLM version, the same maximum model length (16,384), the same greedy decoding contract ($T{=}0$, 64 new tokens), the same prompt text and the same refusal-parsing rule as the primary reader. The text-only model omits `limit_mm_per_prompt`. Task instructions, budget, population and metric are unchanged, and no threshold was retuned. The input-serialization check below qualifies the comparison of the resulting model inputs (Appendix D.5).

Historical Mistral reader and external-proxy test. The original runs use `mistralai/Mistral-7B-Instruct-v0.3` at snapshot `c170c708c41d. . .` on the same 211 questions, prompt, 14k evidence budget, greedy decoder, and six cells. The historical input exceeds 16,384 tokens at this fixed evidence budget, requiring a 32,768-token model window. Before Mistral inference, we freeze $d_i = Y_i^{\text{Qwen}}(A, 0) + Y_i^{\text{Llama}}(A, 0) \in \{0, 1, 2\}$ and regress Mistral's question-level L_0 on this external ordinal score; MIV slope is secondary. One complete immediate-EOS output in the placebo/no-memory cell exposed the inherited scorer property that an empty string matched every gold substring. We retain the raw cell without rerun and apply the universal rule "empty prediction $\mapsto$ EM 0" before effect analysis. The cell cannot affect available-versus-unavailable MIV or either proxy slope. A subsequent audit of 44,331 scored cells across all top-level non-smoke document artifacts finds no other empty false-positive EM; six other empty predictions were already scored zero. That audit covers empty outputs. The separate refusal correction of Appendix D.2 was applied to the 3,744 primary six-cell outcomes; the two scopes are not the same.

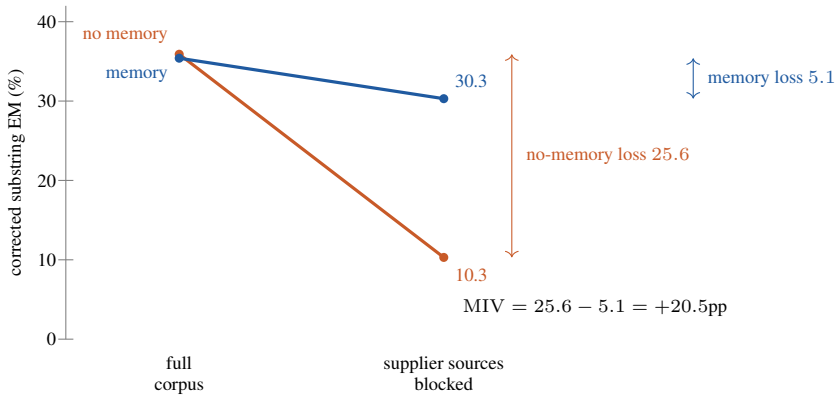


Figure 5: M3DocVQA supporting study (n=195). Source blocking exposes the value of the same memory: its contribution changes from -0.51pp with the corpus available to $+20.00$pp when supplier documents are blocked. Scores use the corrected primary scorer; the legacy-to-corrected score changes are retained in the audit artifact.

Table 22: Pre-specified out-of-reader proxy test (percentage points). The score d is fixed from Qwen and Llama available/no-memory correctness before Mistral inference. Mistral's L_0 slope is $+17.58$pp per score point $[+6.66, +28.68]$; its MIV slope is $+12.14$pp $[+1.60, +23.14]$. The top stratum contains only ten questions; the association does not causally decompose the between-reader difference.

prior-reader score d	n	Mistral A_0	Mistral U_0	L_0	MIV
0	168	3.57	4.76	-1.19	-1.19
1	33	21.21	9.09	$+12.12$	$+6.06$
2	10	50.00	10.00	$+40.00$	$+30.00$

D.1 Exploratory Reference-Answer Exposure in Delivered Memory

We split the 211 held-out FRAMES questions by whether the delivered memory contains the normalized reference-answer token sequence within a single Q/A entry. We reconstruct the primary reader's FACT entries using the original acquisition records and prompt builder, and verify their complete preambles against all 1,266 retained follow-up requests. Matching excludes the current question and retrieved source text. It uses the frozen scorer's lowercase and whitespace normalization, retaining punctuation. Matching uses consecutive whitespace-delimited tokens, with no aliases, stemming or partial-answer expansion. References are used only for this offline diagnostic.

This analysis is post-hoc. We fixed its matching and resampling rules before computing subgroup outcomes, after inspecting the aggregate results. No questions or outcomes are excluded. Restricting matching to the entries' answer fields gives the identical 12/199 split, so sub-question text alone does not account for membership. The three scorers evaluate the same frozen primary-reader outputs. Each bootstrap draw resamples complete six-cell question vectors within each fixed stratum; independent stratum draws yield present-minus-absent differences. We use 20,000 draws and seed 20260923.

Table 23: Post-hoc lexical-exposure strata for primary Qwen3-VL-8B (pp for EM; 0–100 points for token F1). Difference means present minus absent, with a directly bootstrapped interval; it is not inferred from separate significance tests. All conditional and absolute-score intervals are retained in the analysis artifact.

Scorer	Answer string	n	$C(A)$	$C(U)$	MIV [95% CI]
Substring EM	Present	12	$+33.33$	$+100.00$	$+66.67$ $[+41.67, +91.67]$
Substring EM	Absent	199	$+1.51$	$+8.54$	$+7.04$ $[+2.01, +12.06]$
Substring EM	Difference	–	$+31.83$	$+91.46$	$+59.63$ $[+31.62, +85.64]$
Strict EM	Present	12	$+33.33$	$+100.00$	$+66.67$ $[+41.67, +91.67]$
Strict EM	Absent	199	-2.01	-1.01	$+1.01$ $[-1.51, +3.52]$
Strict EM	Difference	–	$+35.34$	$+101.01$	$+65.66$ $[+39.15, +91.16]$
Token F1	Present	12	$+33.33$	$+100.00$	$+66.67$ $[+41.67, +91.67]$
Token F1	Absent	199	$+0.95$	$+3.91$	$+2.96$ $[-0.32, +6.33]$
Token F1	Difference	–	$+32.38$	$+96.09$	$+63.70$ $[+36.82, +89.20]$

In the 12 answer-present questions, source-only accuracy falls from 8/12 to 0/12 under blocking, while memory-on accuracy is 12/12 in both conditions. All three scorers agree on these cells. The resulting $C(U) = 100$pp has a degenerate empirical-bootstrap interval because the observed differences are unanimous; it does not imply certainty outside this small sample. The present-minus-absent MIV intervals are positive under all three scorers. Weighting this stratum by 12/211 yields 3.79 points of the full-cohort interaction under each scorer: 36.4% of substring EM, 80.0% of strict EM and 57.6% of token F1. These are additive contributions to observed means, not causal mediation shares. For the other 199 questions, only the substring MIV interval excludes zero.

Under corrected substring scoring, intact/no-memory accuracy is 66.67% in the answer-present stratum and 14.07% in the other stratum. Membership is observational, so the larger interaction cannot be attributed to answer exposure alone. Groups may also differ in memory quality and which sources are removed. An answer string can be copied without demonstrating a general recovery process; its absence does not exclude useful intermediate facts, punctuation variants or paraphrases. These results do not identify recovery versus reduced interference. The artifact records every question's membership, delivered entries, cell scores, both matching rules, all intervals and source hashes.

D.2 Scoring Robustness of the Source-Access Interaction

Every cell of the three primary document cohorts stores the reader's raw answer alongside its recorded score, so the rescoring below required no new inference. Gold answers come from the same recorded trajectories the original scoring used. The archived figures reproduce exactly under the original scorer and its own seed, including $\mathrm{MIV}_{src} = +18.97$pp on M3DocVQA. The primary-score entries for the three document cohorts use the corrected substring scorer. Strict exact match, token F1, and numeric matching retain their separately specified rules. The legacy substring row is preserved for audit. Scoring versions for the other analysis families are recorded separately in the study-level registry, which also identifies every changed cell. Table 24 summarises the scoring rules and the refusal-correction audit for the listed analysis families.

Table 24: Scoring rule and re-scoring outcome per analysis family. Changed cells count logical cells whose score differs from the archived one under the corrected rule. All changed cells in this refusal-correction audit belong to the same M3DocVQA question, whose gold is a short negative that substring containment credits inside a refusal; the earlier Mistral empty-output correction is a separate audit with its own scope, reported in Appendix D. The common-record row counts the M3DocVQA memory-on cells of the three paths; those questions' memory-off cells are the primary study's, reused through the score sidecar, where the same question accounts for three further changed cells applied when the family was regenerated. The FRAMES common-record row covers the same cell groups, read from the four retained shard files; the batched attempts the contract records as discarded are excluded. A zero means the family was re-scored and nothing moved, not that it was left unexamined. Counts are family-specific and may include reused outcomes; they are not additive across rows. The archived predictor analyses keep legacy labels by design; their write-ups are in the supplementary artifact, not this PDF.

analysis family	scoring rule	changed cells	reported in
Primary document six-cell	corrected substring	$3/3{,}744$	61
Common-record transfer, M3DocVQA	corrected substring	$3/1{,}755$	33
Common-record transfer, FRAMES	corrected substring	$0/2{,}532$	33
Donor-pool transfer, both pools	corrected substring	$24/14{,}616$	36
Write census	corrected substring	$5/3{,}628$	archived (artifact)
Final-step activation	corrected substring	$2/2{,}496$	archived (artifact)
Reader proxies (32B, Llama)	corrected substring	$0/2{,}532$	reader comparisons
Graded dose response	corrected substring	$0/3{,}798$	§5.1
Archived predictor	legacy labels	–	archived (artifact)

Table 25: All three document cohorts under four scoring rules. Percentage points, except token F1 in 0–100 points. Numeric match is reported as a diagnostic: it scores only the leading number, so it is at floor wherever answers are not numeric, and its small estimates are not comparable in magnitude to the others. The corrected substring rows use the corrected primary scoring rule; other rows use the named alternative rule. The M3DocVQA legacy row is retained only for comparison with the archived scoring, and both interactions are identical under the two rules.

corpus	scorer	$C(A)$	$C(U)$	MIV [95% CI]	MIV_{src} [95% CI]
FRAMES dev	corrected substring EM	$+5.05$	$+20.64$	$+15.60$ [$+10.09$, $+21.56$]	$+14.68$ [$+8.72$, $+21.10$]
	strict EM	$+4.13$	$+14.22$	$+10.09$ [$+5.50$, $+14.68$]	$+11.93$ [$+6.88$, $+16.97$]
	token F1	$+4.00$	$+17.13$	$+13.13$ [$+8.34$, $+18.12$]	$+13.41$ [$+7.96$, $+18.92$]
	numeric match	$+1.83$	$+5.50$	$+3.67$ [$+0.92$, $+6.88$]	$+4.13$ [$+0.92$, $+7.80$]
FRAMES held-out	corrected substring EM	$+3.32$	$+13.74$	$+10.43$ [$+5.21$, $+16.11$]	$+10.43$ [$+4.74$, $+16.11$]
	strict EM	0.00	$+4.74$	$+4.74$ [$+1.42$, $+8.53$]	$+4.74$ [$+1.42$, $+8.53$]
	token F1	$+2.79$	$+9.38$	$+6.59$ [$+2.70$, $+10.68$]	$+6.00$ [$+1.66$, $+10.52$]
	numeric match	$+0.95$	$+2.84$	$+1.90$ [$+0.00$, $+4.27$]	$+1.90$ [-0.47, $+4.74$]
M3DocVQA	corrected substring EM	-0.51	$+20.00$	$+20.51$ [$+13.33$, $+27.69$]	$+18.97$ [$+11.79$, $+26.67$]
	strict EM	$+0.51$	$+17.95$	$+17.44$ [$+10.77$, $+24.62$]	$+15.90$ [$+8.72$, $+23.08$]
	token F1	$+0.45$	$+20.66$	$+20.21$ [$+13.41$, $+27.32$]	$+19.29$ [$+12.21$, $+26.62$]
	numeric match	0.00	$+4.62$	$+4.62$ [$+1.03$, $+8.21$]	$+4.10$ [$+1.03$, $+7.69$]
	legacy substring EM	-1.03	$+19.49$	$+20.51$ [$+13.33$, $+27.69$]	$+18.97$ [$+11.79$, $+26.67$]

Disagreements between the two exact-match rules are one-directional in all three cohorts: under the corrected scorer 263 cells are scored correct by substring matching and incorrect by strict matching, 108 of 1,308 development cells, 109 of 1,266 held-out cells and 46 of 1,170 M3DocVQA cells, and no cell disagrees the other way. Under the legacy scorer the M3DocVQA count is 49 and the total is 266; the three additional cells are the refusal defect described below. Classified by string-level structure, with categories tested in a fixed order so that each cell lands in exactly one, 93.9% are containment cases and 6.1% differ only in an article or punctuation. The partition is complete: no cell is left unclassified. This classification is not a semantic correctness adjudication. Raw model outputs were unchanged, and the scorer correction was applied uniformly rather than to selected questions or treatment arms.

The refusal defect is a scorer fault rather than a reader success: the gold no occurs inside the refusal phrase. Applying the corrected rule to all 3,744 cells rather than only the cells an earlier audit surfaced changes three cells, the no-memory scores of the same M3DocVQA question in all three access arms. It therefore changes the access-specific contributions, to -0.51pp and $+20.00$pp, but cancels in both interaction contrasts, which stay at $+20.51$pp and $+18.97$pp. This invariance is

checked on the question-level contrasts under one shared resampling, not inferred from rounded means.

D.3 Additional Gemma Reader

We evaluate `google/gemma-3-27b-it` at snapshot `005ad3404e59. . .` on the same 211 held-out FRAMES questions. The reader was selected before its outcomes were available, using a locally cached model from a different family that fits the available hardware in BF16. The acquisition records, supplier and placebo rules, retrieval card counts, 14k-token evidence budget, strict answer prompt and one-shot repair policy match the existing reader experiment. We use greedy decoding with at most 64 output tokens, two-way tensor parallelism, a 32,768-token model window and disabled prefix caching. Evidence is packed with Gemma's tokenizer, so token-matched budgets do not imply identical text across reader families. The chat template is encoded once with a single beginning-of-sequence token; the supplied token IDs are checked against the engine's input IDs. This is a reader sensitivity on a shared cohort.

Table 26: Gemma-3-27B-IT on held-out FRAMES: corrected substring accuracies (%) and paired 95% intervals for memory contribution.

Source access	Memory off	Memory on	Contribution (pp)
Available	14.22	19.43	$+5.21$ $[+0.47, +9.95]$
Supplier blocked	6.16	20.38	$+14.22$ $[+9.00, +19.91]$
Placebo blocked	18.96	17.06	-1.90 $[-6.16, +2.37]$

Table 27: Gemma scoring sensitivity on the same generated responses (pp for EM; 0–100 points for token F1). Intervals use 20,000 question-paired bootstrap draws.

Scorer	MIV	Placebo-referenced MIV
Corrected substring	$+9.00$ $[+3.79, +14.69]$	$+16.11$ $[+9.95, +22.27]$
Strict EM	$+4.27$ $[+0.95, +7.58]$	$+7.11$ $[+3.32, +11.37]$
Token F1	$+8.32$ $[+4.34, +12.51]$	$+12.65$ $[+8.03, +17.44]$

All 1,266 logical cells are complete. No memory-on/off source-packet mismatch or card-budget mismatch occurs. An offline replay reconstructs every logged model prompt and reproduces every final cell from the recorded model answers. The delivered-document audit includes headers and sibling expansion; it finds zero blocked supplier or placebo documents in their restricted arms. The artifact contains raw prompts and outputs, the frozen protocol, six-cell scores under all three scorers, bootstrap summaries and the replay audit.

Interpretation. Under the primary scorer, the placebo-minus-intact contribution, $C(P)-C(A)$, is -7.11pp $[-11.85, -2.84]$. Thus, placebo removal also changes memory contribution for this reader. The three positive interaction intervals extend the evidence beyond Qwen, but the low intact-access accuracy still limits conclusions about higher-performing readers.

D.4 Native Reader Inputs

Table 28: Readers on the same held-out FRAMES cohort under corrected substring EM. A_0 is intact/no-memory accuracy; L_0 and L_1 are access losses without and with memory. Llama and Mistral use native chat-template inputs. Historical runs remain in Table 29.

Reader	A_0 (%)	L_0 (pp)	L_1 (pp)	MIV [95% CI] (pp)
Qwen3-VL-8B	17.06	$+9.00$	-1.42	$+10.43$ $[+5.21, +16.11]$
Qwen3-VL-32B	18.01	$+10.43$	$+0.95$	$+9.48$ $[+4.27, +14.69]$
Gemma-3-27B-IT	14.22	$+8.06$	-0.95	$+9.00$ $[+3.79, +14.69]$
Llama-3.1-8B	9.95	$+5.21$	$+2.37$	$+2.84$ $[-1.90, +7.58]$
Mistral-7B	10.90	$+5.21$	-1.90	$+7.11$ $[+1.90, +12.80]$
Llama-3.1-70B	22.75	$+10.90$	$+5.21$	$+5.69$ $[+0.00, +11.37]$

We rerun Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3 on all 211 held-out questions after correcting chat serialization. The user message contains plain text, and the engine receives native chat-template token IDs directly. Both readers retain the original source packets in all 1,266 cells. Acquisition, retrieval, evidence budgets, instructions, decoding and the bounded repair rule are fixed.

To separate input-format changes from the current execution setup, we also reconstruct the legacy inputs and run them under the same settings. Native and legacy requests are interleaved in fixed order, with initial generations and required repairs batched separately. Historical runs were serial; their estimates remain separate. The native-minus-current-legacy contrast measures the combined serialization correction, not the isolated effect of a BOS token.

Table 29: Reader input follow-up on the same held-out FRAMES cohort. Contributions and interactions are in percentage points under corrected substring EM. Historical rows rescore archived outputs with the same scorer and bootstrap; current legacy and native inputs share the new execution settings.

Reader	Input/run	A_0 (%)	$C(A)$	$C(U)$	MIV [95% CI]
Llama-3.1-8B	Historical	8.06	$+9.00$	$+11.37$	$+2.37$ $[-1.90, +6.64]$
Llama-3.1-8B	Legacy	8.06	$+9.00$	$+11.37$	$+2.37$ $[-1.90, +6.64]$
Llama-3.1-8B	Native	9.95	$+4.74$	$+7.58$	$+2.84$ $[-1.90, +7.58]$
Mistral-7B	Historical	8.06	$+0.47$	$+1.90$	$+1.42$ $[-2.37, +5.69]$
Mistral-7B	Legacy	8.53	$+0.00$	$+2.37$	$+2.37$ $[-1.90, +6.64]$
Mistral-7B	Native	10.90	$+5.21$	$+12.32$	$+7.11$ $[+1.90, +12.80]$
Llama-3.1-70B	Native	22.75	$+0.47$	$+6.16$	$+5.69$ $[+0.00, +11.37]$

Table 30: Native minus reconstructed legacy input under matched current execution, in percentage points with question-paired 95% intervals. Both input modes and all access/memory cells remain paired in each bootstrap draw.

Reader	Change in $C(A)$	Change in $C(U)$	Change in MIV
Llama-3.1-8B	-4.27 $[-8.53, -0.47]$	-3.79 $[-8.53, +0.47]$	$+0.47$ $[-4.74, +5.69]$
Mistral-7B	$+5.21$ $[+0.00, +10.43]$	$+9.95$ $[+5.21, +15.17]$	$+4.74$ $[-1.42, +10.90]$

Input-format effects. These comparisons use corrected substring EM. For Llama-3.1-8B, correcting serialization leaves the direction of the MIV change unresolved ($+0.47$pp $[-4.74, +5.69]$). For Mistral-7B, correcting serialization leaves the direction of the MIV change unresolved ($+4.74$pp $[-1.42, +10.90]$). These contrasts assess the proposed explanation for the earlier small interactions. An interval containing zero does not establish equivalence between the two input modes.

Larger reader. Llama-3.1-70B-Instruct uses the cached BF16 snapshot `1605565b47bb...`, four-way tensor parallelism and native inputs. Its initial token IDs and repair text match the native 8B run in all 1,266 held-out cells. The model and analysis were fixed before its outcomes. A hash selected 20 development questions without consulting responses; its source-only accuracy is 20.00%, compared with 25.00% for the recorded primary reader on the same questions. The full held-out study was required regardless of this diagnostic. Its intact/no-memory accuracy is 22.75%, with a paired difference of $+5.69$pp $[+0.00, +11.37]$ from the primary Qwen reader. Parameter count alone does not establish a high-accuracy evaluation regime.

The 70B interaction has a positive point estimate under all three scorers (Table 32). The primary interval includes both zero and positive interactions; it establishes neither a positive interaction nor its absence. Strict EM also includes zero, while the token-F1 interval is positive. Placebo-referenced interaction intervals include zero under all three scorers. The results extend the reader comparison with uncertainty about the interaction and its specificity to supplier-source loss.

Table 31: Native-input absolute accuracies (%) and conditional contributions (pp) under corrected substring EM.

Reader	Source access	Memory off	Memory on	Contribution [95% CI]
Llama-3.1-8B	Available	9.95	14.69	$+4.74$ $[+0.00, +9.48]$
Llama-3.1-8B	Supplier blocked	4.74	12.32	$+7.58$ $[+2.84, +12.32]$
Llama-3.1-8B	Placebo blocked	11.37	18.01	$+6.64$ $[+1.90, +11.37]$
Mistral-7B	Available	10.90	16.11	$+5.21$ $[+0.47, +9.95]$
Mistral-7B	Supplier blocked	5.69	18.01	$+12.32$ $[+7.11, +18.01]$
Mistral-7B	Placebo blocked	9.48	17.54	$+8.06$ $[+3.32, +12.80]$
Llama-3.1-70B	Available	22.75	23.22	$+0.47$ $[-5.21, +6.16]$
Llama-3.1-70B	Supplier blocked	11.85	18.01	$+6.16$ $[+0.47, +11.85]$
Llama-3.1-70B	Placebo blocked	22.27	23.22	$+0.95$ $[-4.27, +6.16]$

Table 32: Scoring sensitivity for the native-input follow-up (pp for EM; 0–100 points for token F1). Intervals use 20,000 question-paired bootstrap draws. These are multiple scores of the same outputs, not independent reader replications.

Reader	Scorer	MIV [95% CI]	Placebo-referenced MIV [95% CI]
Llama-3.1-8B	substring EM (corrected)	$+2.84$ $[-1.90, +7.58]$	$+0.95$ $[-4.27, +6.16]$
Llama-3.1-8B	strict EM	$+3.79$ $[+0.95, +6.64]$	$+2.37$ $[-0.95, +6.16]$
Llama-3.1-8B	token F1	$+3.22$ $[-0.02, +6.79]$	$+1.56$ $[-2.63, +5.84]$
Mistral-7B	substring EM (corrected)	$+7.11$ $[+1.90, +12.80]$	$+4.27$ $[-1.42, +9.95]$
Mistral-7B	strict EM	$+3.79$ $[+0.95, +6.64]$	$+3.32$ $[+0.00, +6.64]$
Mistral-7B	token F1	$+5.33$ $[+1.48, +9.38]$	$+6.67$ $[+2.31, +11.12]$
Llama-3.1-70B	substring EM (corrected)	$+5.69$ $[+0.00, +11.37]$	$+5.21$ $[-0.47, +10.90]$
Llama-3.1-70B	strict EM	$+3.79$ $[-0.47, +8.06]$	$+1.90$ $[-2.37, +6.16]$
Llama-3.1-70B	token F1	$+6.06$ $[+1.50, +10.88]$	$+4.02$ $[-0.91, +9.10]$

Execution and interpretation. All readers use greedy decoding, 64 output tokens, a 14,000-token evidence budget and disabled prefix caching. The maximum model windows are 16,384 for Llama8B and 32,768 for Mistral7B and Llama70B. For the 70B run, a synthetic communication check before held-out inference selects `NCCL_P2P_DISABLE=1`. Engine inputs match the supplied token IDs. Memory-on/off source packets and card counts match; no blocked supplier or placebo document is delivered, including headers and sibling expansion. Recorded-answer replay reproduces final repair decisions. This checks execution reconstruction rather than fresh-generation repeatability. These follow-ups reuse inspected questions and do not establish reader-invariant effects or internal recovery.

D.5 Input Serialization in Earlier Text-Reader Runs

A tokenizer-only check using the recorded model revisions and runtime finds that converting the chat template to a string and encoding it again preserves the Qwen token sequence, but changes the Llama and Mistral sequences. The Llama path adds a beginning-of-sequence token and renders the structured content as a string representation; the Mistral path re-encodes a display of the native tokens. These are limitations of the recorded wrappers. Their estimates and the proxy analysis in Table 22 remain descriptions of those configurations, rather than controlled comparisons of model families. The tokenizer check alone does not measure an accuracy effect. The matched input experiments in Appendix D.4 test that question directly.

D.6 Public Memory Implementation Transfer

Frozen scope. We evaluate Mem0 OSS at commit `02f7a9b2c4fe38dedb96631e48c85c74ad58b605` and A-MEM at commit `ceffb860f0712bbae97b184d440df62bc910ca8d`, plus a deterministic FLAT vector control. Mem0 runs native inference; A-MEM runs its released note construction, linking, and evolution. The systems receive identical ordered pre-branch records: document FACT question–answer entries, the frozen rendered purchase receipt for WebShop, or pre-query

10-minute timestamped visual–audio caption rows for EgoLifeQA. The main question, answer, correctness, access arm, and post-query observations are unavailable during construction. A snapshot is built once and held fixed across all access arms.

Table 33: Public-implementation document transfer in percentage points. All four MIV intervals exclude zero: A-MEM/FLAT and Mem0 are respectively $[6.16, 17.06]$ and $[0.95, 11.37]$ on FRAMES, and $[11.79, 25.64]$ and $[12.82, 26.67]$ on supporting M3DocVQA. †Payloads are byte-identical, so these are not independent architecture replications.

corpus	memory path	$C(A)$	$C(U)$	MIV	MIV_{src}
FRAMES	A-MEM/FLAT†	$+2.37$	$+13.74$	$+11.37$	$+10.90$
	Mem0	$+4.74$	$+10.90$	$+6.16$	$+6.64$
M3DocVQA	A-MEM/FLAT†	$+0.51$	$+18.97$	$+18.46$	$+16.41$
	Mem0	0.00	$+19.49$	$+19.49$	$+18.46$

Table 34: Projection-controlled public-memory contract. Native internal state is retained for audit, but only the listed text projection reaches the common Qwen3-VL-8B reader.

component	Mem0	A-MEM	FLAT control
native processing	inferred ADD/UPDATE/DELETE events and memory IDs	note context, keywords, tags, links, and evolution	none
retrieval	OSS semantic search, $k = 5$, isolated namespace	released `search_agentic`, $k = 5$	MiniLM cosine, $k = 5$
projected fields	retrieved memory text	retrieved original note content	original record text
shared delivery	question-only query; MiniLM embeddings where replacement is supported; stable order; exact-string deduplication; neutral heading; item-boundary truncation at 1,024 Qwen tokens; no scores, tags, system name, gold fields, or generated final answer		

The FLAT control embeds the original records with `all-MiniLM-L6-v2`. Mem0 uses the same 384-dimensional embedding space through a local isolated store and does not enable its optional lexical path. A-MEM uses the released OpenAI-compatible path. Every query uses an isolated document qid, WebShop history instance, or EgoLifeQA life-and-cutoff namespace. The common reader is the pinned Qwen3-VL-8B snapshot from Appendix D, with greedy decoding and prefix caching disabled. All intervals use 20,000 paired percentile-bootstrap draws, seed 20260910; one draw resamples the complete cross-system, cross-arm vector for a question or counterfactual pair. No cross-benchmark estimate is pooled.

Document absolute cells. On FRAMES ($n = 211$), A-MEM/FLAT accuracy is $17.06 \to 19.43$ in A, $8.06 \to 21.80$ in U, and $15.64 \to 18.48$ in P; Mem0 is $17.06 \to 21.80$, $8.06 \to 18.96$, and $15.64 \to 19.91$. Their respective $C(P)$ values are $+2.84$pp $[-1.42, +7.11]$ and $+4.27$pp $[+0, +8.53]$. On supporting M3DocVQA ($n = 195$), A-MEM/FLAT is $35.90 \to 36.41$, $10.26 \to 29.23$, and $33.85 \to 36.41$; Mem0 is $35.90 \to 35.90$, $10.26 \to 29.74$, and $33.85 \to 34.87$. Their $C(P)$ values are $+2.56$pp $[-1.54, +6.67]$ and $+1.03$pp $[-2.56, +4.62]$. The full contribution and interaction intervals are summarized in Table 33.

Table 35: Non-document transfer results. WebShop direct utility uses 50 pairs (100 histories). Matched contributions come from Table 52. Crossed WebShop rows subtracted baselines from different environments and context windows; their estimates are withdrawn and not reported here, and the original outputs remain in the supplementary artifact. The between-projection contrast retains validity because the shared baseline cancels. Table 53 records each cell's execution contract. The separate controlled study in Table 47 is unaffected. EgoLifeQA source access uses 50 questions from one life record. Direct and MIV rows are distinct studies and are never combined.

study	path	$A_0 \to A_1$	$U_0 \to U_1$	$P_0 \to P_1$	G or MIV	95% CI
WebShop direct, matched	A-MEM/FLAT	–	–	–	$G = +16$	$[+8, +24]$
	Mem0	–	–	–	$G = -4$	$[-9, +0]$
WebShop direct, crossed	both	*withdrawn*; replaced by the matched rows above				
WebShop source, crossed	both	*withdrawn*; see Appendix F.5				
EgoLife source	A-MEM/FLAT	$52 \to 52$	$40 \to 44$	$30 \to 38$	$+4$	$[-10, +18]$
	Mem0	$52 \to 56$	$40 \to 44$	$30 \to 36$	0	$[-14, +14]$

For WebShop, the paired A-MEM/FLAT-minus-Mem0 direct-gain contrast is $+20$pp $[+13, +27]$ under the matched contract, against $+19$pp $[+12, +26]$ as first computed; the two agree because both projections subtracted the same baseline, so the environment mismatch cancels in their difference even though it does not cancel in either gain. The paired MIV contrast is $+22.5$pp $[+2.5, +45]$. For the EgoLifeQA direct block ($n = 110$), A-MEM/FLAT changes $31.82\% \rightarrow 40.00\%$ ($+8.18$pp $[-1.82, +18.18]$), while Mem0 changes $31.82\% \rightarrow 33.64\%$ ($+1.82$pp $[-8.18, +11.82]$); their paired difference is $+6.36$pp $[-1.82, +14.55]$. These text-memory transfers neither reproduce WorldMM's visual memory nor establish person-level generalization.

Payload and execution audit. A-MEM and FLAT retrieval produce byte-identical payloads for all 211 FRAMES, 195 M3DocVQA, 100 WebShop-direct, 40 WebShop-source, 110 EgoLifeQA-direct, and 50 EgoLifeQA-source units. A-MEM's native metadata and links are therefore not separately identified by the downstream comparison. Mem0 has empty retrievals on 22/211 FRAMES and 20/195 M3DocVQA questions, and none in the other studies. Every delivered payload is within the 1,024-token cap. All final rows complete without inference error. Discarded batched, mismatched-configuration, stochastic-decoding, and invalid context-window attempts are retained with the outcome-blind amendment record; none is used in an estimate.

D.6.1 Independent Memory Construction under Competitive Donor Pools

Controlled inputs and system-specific construction. The common-record study above gives each system the same acquired records; it does not force Mem0 to share A-MEM's or FLAT retrieval's delivered payload. This follow-up introduces competing acquired material. For each question, FLAT retrieval, A-MEM, and Mem0 independently build and query a memory store from the same ordered donor pool. The 20-record pool is nested within the 100-record pool, with the target record held fixed. Acquisition, source-access interventions, the delivery contract, and the pinned reader remain fixed. The study therefore tests memory-module behavior under a controlled contract, not native end-to-end agent acquisition.

The evaluation retains 211 FRAMES and 195 M3DocVQA questions at both pool sizes. The manifest contains 48,720 input records per path and 14,616 logical reader cells across three paths, two pool sizes, and six access–memory conditions. Logical cell counts are not claims about the number of distinct model calls. Intervals use 20,000 paired question-bootstrap draws with seed 20260911. The earlier study uses seed 20260910; bootstrap seeds affect Monte Carlo interval estimates, not the definition of the estimand or its point estimate. Mem0 failures follow the declared fallback without removing questions from the analysis.

Table 36: Document memory-module transfer under fixed acquisition and delivery, in percentage points. Each path constructs its own memory from the same ordered 20- or 100-record pool. MIV and MIV_{src} are reported with individual paired 95% intervals. The cohorts contain 211 FRAMES and 195 supporting M3DocVQA questions. Delivered-payload identity is audited in Table 40.

corpus	pool	path	$C(A)$	$C(U)$	MIV	95% CI	MIV_{src}	95% CI
FRAMES	20	FLAT	$+3.32$	$+14.22$	$+10.90$	$[+6.16, +16.11]$	$+11.37$	$[+5.69, +17.54]$
		A-MEM	$+3.32$	$+14.22$	$+10.90$	$[+6.16, +16.11]$	$+11.37$	$[+5.69, +17.54]$
		Mem0	$+2.84$	$+12.80$	$+9.95$	$[+4.27, +15.64]$	$+10.90$	$[+4.74, +17.06]$
	100	FLAT	$+3.79$	$+13.74$	$+9.95$	$[+5.21, +14.69]$	$+10.90$	$[+5.69, +16.59]$
		A-MEM	$+3.79$	$+13.74$	$+9.95$	$[+5.21, +14.69]$	$+10.90$	$[+5.69, +16.59]$
		Mem0	$+2.84$	$+13.74$	$+10.90$	$[+6.16, +15.64]$	$+11.37$	$[+6.16, +17.06]$
M3DocVQA	20	FLAT	$+0.51$	$+18.97$	$+18.46$	$[+11.79, +25.13]$	$+17.95$	$[+10.77, +25.13]$
		A-MEM	$+0.51$	$+18.97$	$+18.46$	$[+11.79, +25.13]$	$+17.95$	$[+10.77, +25.13]$
		Mem0	0.00	$+17.44$	$+17.44$	$[+10.26, +24.62]$	$+16.92$	$[+9.74, +24.10]$
	100	FLAT	$+1.54$	$+22.05$	$+20.51$	$[+13.85, +27.69]$	$+20.00$	$[+12.82, +27.18]$
		A-MEM	$+1.54$	$+22.05$	$+20.51$	$[+13.85, +27.69]$	$+20.00$	$[+12.82, +27.18]$
		Mem0	$+1.54$	$+19.49$	$+17.95$	$[+11.28, +24.62]$	$+18.46$	$[+11.78, +25.64]$

Table 37: Absolute six-cell accuracy (%), from memory off ($\cdot_0$) to memory on ($\cdot_1$). The final column partitions unavailable-arm questions by the sign of the paired utility difference: helpful, harmful, or tied. A tied utility does not necessarily imply an identical answer string.

corpus	pool	path	$A_0 \to A_1$	$U_0 \to U_1$	$P_0 \to P_1$	$C(P)$	helpful/harmful/tied in U
FRAMES	20	FLAT	$17.06 \to 20.38$	$8.06 \to 22.27$	$15.64 \to 18.48$	$+2.84$	$31/1/179$
		A-MEM	$17.06 \to 20.38$	$8.06 \to 22.27$	$15.64 \to 18.48$	$+2.84$	$31/1/179$
		Mem0	$17.06 \to 19.91$	$8.06 \to 20.85$	$15.64 \to 17.54$	$+1.90$	$30/3/178$
	100	FLAT	$17.06 \to 20.85$	$8.06 \to 21.80$	$15.64 \to 18.48$	$+2.84$	$30/1/180$
		A-MEM	$17.06 \to 20.85$	$8.06 \to 21.80$	$15.64 \to 18.48$	$+2.84$	$30/1/180$
		Mem0	$17.06 \to 19.91$	$8.06 \to 21.80$	$15.64 \to 18.01$	$+2.37$	$30/1/180$
M3DocVQA	20	FLAT	$35.90 \to 36.41$	$10.26 \to 29.23$	$33.85 \to 34.87$	$+1.03$	$42/5/148$
		A-MEM	$35.90 \to 36.41$	$10.26 \to 29.23$	$33.85 \to 34.87$	$+1.03$	$42/5/148$
		Mem0	$35.90 \to 35.90$	$10.26 \to 27.69$	$33.85 \to 34.36$	$+0.51$	$40/6/149$
	100	FLAT	$35.90 \to 37.44$	$10.26 \to 32.31$	$33.85 \to 35.90$	$+2.05$	$46/3/146$
		A-MEM	$35.90 \to 37.44$	$10.26 \to 32.31$	$33.85 \to 35.90$	$+2.05$	$46/3/146$
		Mem0	$35.90 \to 37.44$	$10.26 \to 29.74$	$33.85 \to 34.87$	$+1.03$	$41/3/151$

Table 38: Paired between-system contrasts under native construction. Every contrast resamples one shared question-level draw, so the intervals are paired rather than two independent bootstraps. Acquisition, donor pool, delivery contract and reader are held fixed, so a difference is attributable to the memory system under that fixed contract and not to the amount or order of experience. It does not decompose which component of a system (construction, linking, retrieval) produced the difference.

corpus	pool	contrast	$\Delta C(U)$	95% CI	ΔMIV	95% CI
FRAMES	20	FLAT − A-MEM	0.00	$[0.00, 0.00]$	0.00	$[0.00, 0.00]$
		FLAT − Mem0	$+1.42$	$[-1.42, +4.74]$	$+0.95$	$[-2.37, +4.27]$
		A-MEM − Mem0	$+1.42$	$[-1.42, +4.74]$	$+0.95$	$[-2.37, +4.27]$
	100	FLAT − A-MEM	0.00	$[0.00, 0.00]$	0.00	$[0.00, 0.00]$
		FLAT − Mem0	0.00	$[-3.32, +3.32]$	-0.95	$[-4.74, +2.84]$
		A-MEM − Mem0	0.00	$[-3.32, +3.32]$	-0.95	$[-4.74, +2.84]$
M3DocVQA	20	FLAT − A-MEM	0.00	$[0.00, 0.00]$	0.00	$[0.00, 0.00]$
		FLAT − Mem0	$+1.54$	$[-2.05, +5.64]$	$+1.03$	$[-2.56, +4.62]$
		A-MEM − Mem0	$+1.54$	$[-2.05, +5.64]$	$+1.03$	$[-2.56, +4.62]$
	100	FLAT − A-MEM	0.00	$[0.00, 0.00]$	0.00	$[0.00, 0.00]$
		FLAT − Mem0	$+2.56$	$[-1.03, +6.67]$	$+2.56$	$[-1.54, +7.18]$
		A-MEM − Mem0	$+2.56$	$[-1.03, +6.67]$	$+2.56$	$[-1.54, +7.18]$

Table 39: Within-system changes from the 20-record to the nested 100-record pool. Entries report paired changes in contribution and MIV, in percentage points. All reported ΔMIV intervals include zero, so this experiment does not establish an advantage from increasing the pool size.

corpus	path	$\Delta C(A)$	$\Delta C(U)$	ΔMIV	95% CI
FRAMES	FLAT	$+0.47$	-0.47	-0.95	$[-2.84, +0.95]$
	A-MEM	$+0.47$	-0.47	-0.95	$[-2.84, +0.95]$
	Mem0	0.00	$+0.95$	$+0.95$	$[-3.32, +5.21]$
M3DocVQA	FLAT	$+1.03$	$+3.08$	$+2.05$	$[-1.54, +5.64]$
	A-MEM	$+1.03$	$+3.08$	$+2.05$	$[-1.54, +5.64]$
	Mem0	$+1.54$	$+2.05$	$+0.51$	$[-3.59, +4.62]$

How the audit columns are defined. Every column of Table 40 counts questions, over the full cohort denominator; no question is dropped for failing. *Native fail.* counts store-level construction failures: the status is recorded per question, not per input record, and a failed store invokes the declared empty-memory fallback, so every failed question also has an empty delivered payload. *Empty retr.* counts questions whose retrieval returned nothing *among successfully built stores*, which is why it can be zero while construction failures are not. Final empty delivered payloads, fallback units included: FRAMES pool 20 Mem0 1; FRAMES pool 100 Mem0 2; every other group has none. The last three columns count questions whose target record survives lexically in the full store, in what retrieval returns, and in what reaches the reader; they do not count nonempty stores or payloads and do not establish semantic sufficiency, since a payload can be nonempty while the target is absent.

Table 40: Native construction coverage and target survival, in questions, over the full cohort denominator. Column definitions are given above. Identical payloads across paths: FRAMES pool 20 FLAT/A-MEM $211/211$; FRAMES pool 20 FLAT/Mem0 $0/211$; FRAMES pool 20 A-MEM/Mem0 $0/211$; FRAMES pool 100 FLAT/A-MEM $211/211$; FRAMES pool 100 FLAT/Mem0 $0/211$; FRAMES pool 100 A-MEM/Mem0 $0/211$; M3DocVQA pool 20 FLAT/A-MEM $195/195$; M3DocVQA pool 20 FLAT/Mem0 $0/195$; M3DocVQA pool 20 A-MEM/Mem0 $0/195$; M3DocVQA pool 100 FLAT/A-MEM $195/195$; M3DocVQA pool 100 FLAT/Mem0 $0/195$; M3DocVQA pool 100 A-MEM/Mem0 $0/195$.

corpus	pool	path	native fail.	empty retr.	stored	retrieved	delivered
FRAMES	20	FLAT	0	0	211	211	211
		A-MEM	0	0	211	211	211
		Mem0	1	0	196	194	194
	100	FLAT	0	0	211	211	211
		A-MEM	0	0	211	211	211
		Mem0	2	0	198	194	194
M3DocVQA	20	FLAT	0	0	195	195	195
		A-MEM	0	0	195	195	195
		Mem0	0	0	180	178	178
	100	FLAT	0	0	195	195	195
		A-MEM	0	0	195	195	195
		Mem0	0	0	184	179	179

Observed downstream identity of A-MEM and FLAT retrieval. A-MEM and FLAT retrieval deliver byte-identical memory payloads in all four corpus–pool groups (Table 40). Their recorded question-level outcome contrasts are also zero, yielding a mechanically zero estimate and a degenerate resampling interval under this contract. The contrast is defined, but it does not identify the incremental value of A-MEM's structural components. Mem0 produces different delivered payloads. These observations do not imply equivalence between A-MEM and FLAT retrieval outside the tested contract. Payload identity, full-request identity, and outcome identity were confirmed separately across 4,872 evaluated cells.

Why the payloads coincide in this configuration. The identity is not a construction failure and not an evolution failure. Across 812 A-MEM units and 48,720 notes, evolution attached links on 27.6% and tags on 28.4% of the 47,908 attempts, changed neighbour context on 1.1%, and failed to parse the model response on 0.22% of them; a unit's first note is returned unevolved without a model call and is excluded from that denominator. In the audited revision, search first collects the base vector results and then appends linked candidates before applying the final size limit. When the base results already fill that limit, the appended candidates do not survive the final selection. That is what happened here: of 4,060 delivered retrieval entries, 0 are linked candidates, and every unit returned a full set of base results. The tested delivery contract then renders only the selected content fields. Together with the recorded indexing and evolution behavior, this explains the observed payload identity in this configuration. The conclusion is specific to the audited revision and local adapter, and does not establish that A-MEM's structure is inert under a contract that leaves room for linked candidates.

Artifacts. The plan, generation ledgers, per-unit payloads, frozen execution inputs, reader outputs, analysis, diagnostics, generation-aware verification and the equivalence audit are under `GeneralMOA/robustness_followup/`. The tables above are emitted by `p44_emit_transfer_tables.py` from the analysis JSON and the equivalence figures by `p44_a_mem_equivalence.py`; both record the SHA-256 of their inputs. Memory construction spanned three execution generations after two worker losses; each payload is validated against the ledger that built it, and the per-group generation map is recorded beside each frozen audit.

E Conversational Memory and Retention

Status. Scope check for Section 5.2. The questions were previously inspected; the corrective rerun supplies current evidence and the original construction is superseded.

E.1 Memory Formed Before the Question: Corrective Contract

Primary-stratum definition. The original cohort script `e3_build_cohort.py` classifies a question as arithmetic if its reference answer matches the numeric-format rule or its question contains "how many," "how much," "how long," or "how often" (case-insensitive, with word boundaries). The numeric rule matches the entire answer after trimming whitespace: a digit sequence with optional commas and a decimal part, an optional dollar, euro, or pound prefix, an optional minus sign, and an optional percent suffix. All remaining questions form the non-arithmetic stratum, including 186 in the final split. This predeclared label denotes a screening rule, not a manual assessment of whether calculation is required. The corrective rerun retains the original stratum membership.

Archived results and correction. The original $100/400$ split groups shared answer sessions; the broader distractor pool is shared. The archived Mem0 writer received answer-bearing raw identifiers despite opaque consumer IDs. Its snapshot export used an ignored `limit` argument, yielding a default 20 items while native recall searched the full store. These issues motivated a corrective rerun, not an equivalence claim about the old result. Full-store exports are checked after every continuation in the rerun. The old substring scorer could credit an incorrect response when only an extracted number matched; the rerun uses the semantic judges described below. The original reader-transfer results and the old substring-score estimates remain historical sensitivity results and do not independently validate the newly constructed stores or the semantic endpoints.

Isolation and immutability. A field allowlist retains only role and content from sessions. IDs become `S01`, `S02`, and so on before extraction. The writer seed is 20260910 and its prompt clock is fixed at 2026-09-18; this is a fixed extraction context, not a replay of a historical writer clock. Retention uses the corpus's question and session timestamps. The original 1,800-character observation window, ordered input, embedding model, native Mem0 extraction, and delivered-memory cap are retained. Full exports use `top_k=10000, show_expired=True`; all stores are smaller than that cap. The audit checks 436 stores and 21,974 actual extraction calls, with zero original-ID prompt exposures and unchanged full stores during consumption. Native extraction parsing behavior is retained, including failed extractions; no favorable rebuild is selected. Non-JSON raw-response counts are not equated with native extraction failures because wrappers may be parsed.

Repeatability diagnostic. Twelve fixed development questions receive three independent Mem0 builds and two consumers per frozen snapshot, alongside repeated NONE and FLAT conditions: 480 episodes. The no-numeric-expansion substring diagnostic estimates writer variance 0.00617 and zero within-store consumer variance in this sample. A predeclared rule compares projected writer standard error at 186 questions with 0.015, selecting one final build. This small variance estimate is not a guarantee of determinism or adequate precision. Final standard random-mask conditions have two consumer repeats; retention and matched-random conditions have one.

Retention exposure. On the 400 final questions, $7/30/90$ days remove $44.02/12.94/3.87\%$ of sessions and remove at least one annotated evidence session for $204/63/31$ questions. On the primary stratum the analogous counts are $78/24/9$ out of 186. These labels are used only for offline description, never to construct memory or choose a mask. Each retention mask has three random controls with the same question-specific removal count.

Scoring audit. The official task-specific prompts are fixed across conditions. Llama-3.1-8B is the primary automated judge; Qwen3-8B is a required sensitivity judge, both fixed before corrective inference. Their full-cohort results remain automated estimates; human validation covers only the final-cohort subset below. They produce 8,532 valid correctness judgments for 4,266 distinct question–response pairs. Human validation of these judges uses the completed final-cohort audit below.

Mechanism audit. All 800 question–memory traces under 30-day retention have a constant recalled payload matching the cached payload used for controls. Both judges identify seven supported rescue combinations across four distinct questions (two combinations belong to the primary stratum); this is a behavioral pattern, not a causal trace of internal reasoning. Of 1,600 support judgments, one fails the required yes/no format and remains missing; it does not change that count. The

fixed 24-question foreign-memory control has 48 episodes: 22 payloads exactly match the receiver's token count and 26 are shorter. All own-minus-foreign confidence intervals include zero. No strong information-recovery mechanism is claimed.

Selection and version boundaries. The original observed selectors both choose FLAT, with zero realized improvement. Bootstrap selection disagreements and gain ranges are post-hoc sensitivities, not confidence intervals establishing a selection benefit. The delivery-budget extension also has no observed crossing. The separate state-dependent recall study uses a development-frozen rule and disjoint conversation groups (Appendix E.3). The archived corpus is not silently replaced by the official cleaned release: references match, but 454 haystacks differ. Reusing those references with new histories would not create an independent unseen-question test.

Table 41: LongMemEval-S automated results on the 186-question primary stratum. Absolute accuracies are percentages; MIV is in percentage points. Subscripts $0/1$ denote memory off/on. These are automated Llama scores on 186 questions; Table 43 reports the separate 50-question human audit. This is a rerun of previously inspected questions.

loss regime	memory	A_0	A_1	R_0	R_1	MIV [95% CI]
random half	FLAT	48.66	46.51	38.17	44.35	$+8.33$ $[+0.90, +15.50]$
random half	Mem0	48.66	45.16	38.17	43.64	$+8.96$ $[+1.78, +15.95]$
7-day retention	FLAT	48.66	46.51	32.80	44.09	$+13.44$ $[+5.14, +21.54]$
7-day retention	Mem0	48.66	45.16	32.80	45.16	$+15.86$ $[+8.24, +23.66]$

Complete numerical results. Table 42 includes both judges, all retention policies, and the full-cohort sensitivity. Intervals are paired group-bootstrap intervals, not episode-level intervals. Multiple policy comparisons are exploratory and unadjusted.

Table 42: Corrective automated semantic results in percentage points. A separate 50-question human comparison appears in Table 43. Count-matched contrasts and all absolute cells are also preserved in the machine-readable analysis artifact.

n	judge	memory	loss	$C(A)$	$C(R)$	MIV [95% CI]
186	Llama	FLAT	random_half	-2.15	+6.18	$+8.33$ $[+0.90, +15.50]$
186	Llama	FLAT	retention_7	-2.15	+11.29	$+13.44$ $[+5.14, +21.54]$
186	Llama	FLAT	retention_30	-2.15	-0.54	$+1.61$ $[-2.43, +5.62]$
186	Llama	FLAT	retention_90	-2.15	-3.23	-1.08 $[-4.01, +1.62]$
186	Llama	MEM0	random_half	-3.49	+5.47	$+8.96$ $[+1.78, +15.95]$
186	Llama	MEM0	retention_7	-3.49	+12.37	$+15.86$ $[+8.24, +23.66]$
186	Llama	MEM0	retention_30	-3.49	-3.23	$+0.27$ $[-2.97, +3.51]$
186	Llama	MEM0	retention_90	-3.49	-3.23	$+0.27$ $[-2.43, +3.21]$
186	Qwen	FLAT	random_half	-0.54	+8.24	$+8.78$ $[+2.15, +15.14]$
186	Qwen	FLAT	retention_7	-0.54	+12.90	$+13.44$ $[+5.91, +21.35]$
186	Qwen	FLAT	retention_30	-0.54	-1.61	-1.08 $[-5.38, +2.70]$
186	Qwen	FLAT	retention_90	-0.54	-1.08	-0.54 $[-2.16, +1.08]$
186	Qwen	MEM0	random_half	-2.69	+9.59	$+12.28$ $[+6.04, +18.63]$
186	Qwen	MEM0	retention_7	-2.69	+9.68	$+12.37$ $[+5.91, +18.65]$
186	Qwen	MEM0	retention_30	-2.69	-3.76	-1.08 $[-3.21, +0.81]$
186	Qwen	MEM0	retention_90	-2.69	-1.08	$+1.61$ $[+0.00, +3.78]$
400	Llama	FLAT	random_half	+0.50	+6.33	$+5.83$ $[+1.21, +10.52]$
400	Llama	FLAT	retention_7	+0.50	+10.00	$+9.50$ $[+4.50, +14.66]$
400	Llama	FLAT	retention_30	+0.50	+1.25	$+0.75$ $[-1.88, +3.38]$
400	Llama	FLAT	retention_90	+0.50	+0.25	-0.25 $[-2.14, +1.63]$
400	Llama	MEM0	random_half	-1.62	+7.50	$+9.13$ $[+4.36, +13.83]$
400	Llama	MEM0	retention_7	-1.62	+10.50	$+12.12$ $[+6.77, +17.50]$
400	Llama	MEM0	retention_30	-1.62	-2.25	-0.62 $[-3.47, +2.01]$
400	Llama	MEM0	retention_90	-1.62	-1.50	$+0.12$ $[-1.88, +2.12]$
400	Qwen	FLAT	random_half	+0.62	+8.12	$+7.50$ $[+3.44, +11.68]$
400	Qwen	FLAT	retention_7	+0.62	+10.00	$+9.38$ $[+4.64, +14.14]$
400	Qwen	FLAT	retention_30	+0.62	+1.00	$+0.38$ $[-2.23, +2.86]$
400	Qwen	FLAT	retention_90	+0.62	+1.00	$+0.38$ $[-1.13, +2.00]$
400	Qwen	MEM0	random_half	-0.62	+10.00	$+10.62$ $[+6.49, +14.82]$
400	Qwen	MEM0	retention_7	-0.62	+12.25	$+12.88$ $[+8.62, +17.17]$
400	Qwen	MEM0	retention_30	-0.62	-0.25	$+0.38$ $[-1.63, +2.38]$
400	Qwen	MEM0	retention_90	-0.62	+0.25	$+0.88$ $[-0.75, +2.63]$

Table 43: MIV (pp) and paired 95% intervals on the identical 50-question LongMemEval-S human audit subset. Positive human point estimates coexist with limited precision; zero-containing intervals are not evidence of no effect.

loss	memory	human	Llama judge	Qwen judge
random half	FLAT	$+5.33$ $[-7.33, +17.00]$	$+6.67$ $[-11.33, +23.67]$	$+1.33$ $[-12.00, +13.67]$
7-day	FLAT	$+8.00$ $[-2.00, +18.00]$	$+1.00$ $[-14.00, +16.00]$	$+8.00$ $[-4.00, +20.00]$
random half	Mem0	$+9.00$ $[-1.67, +20.00]$	$+14.33$ $[-1.33, +30.00]$	$+13.33$ $[+0.67, +25.67]$
7-day	Mem0	$+11.00$ $[+0.00, +22.00]$	$+12.00$ $[-1.00, +25.00]$	$+11.00$ $[+1.00, +22.00]$

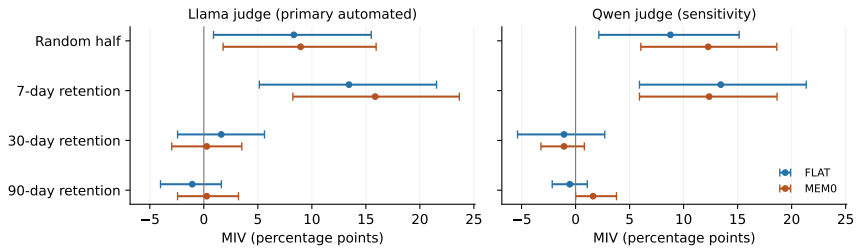


Figure 6: Access-conditioned interaction on the 186-question LongMemEval-S primary stratum under two semantic judges; paired 95% intervals. Positive effects at seven days coexist with weak effects at longer retention windows. Both judges score the same reader responses. The 50-question human audit is reported separately in Table 43.

E.2 Completed Follow-up Human Effect Audit

This final-cohort audit is distinct from the unfinished 180-response development judge audit. Automated aggregate results had already been inspected when this follow-up was designed; it is not a preregistration before the automated experiment.

A fixed hash ordering of the 186 primary final question IDs selects 50 questions from 50 history groups. Sampling uses no response, judge label, disagreement or observed effect. The question list was locked before the preparation script read the response file. For each question we retain NONE, FLAT and Mem0 under intact access, three random masks, and seven-day retention, including both original repetitions where available: 1,350 matched episode cells. Deduplication within question yields 463 distinct responses per reviewer. Review materials show question, reference and response while hiding conditions and automatic judgments. Two independent external annotators, neither an author, agreed on 439/463 responses (94.82%; Cohen's $\kappa = 0.856$). All 24 disagreements received third-reviewer adjudication; 104 final labels are correct and 359 incorrect. Identical question–response pairs share labels across cells. No missing labels were replaced with model scores.

Analysis averages repeats and random masks within question, assigns equal question weights, and uses 10,000 paired history-group bootstrap draws (seed 20260920). Table 44 reports the four human-scored outcome cells and conditional contributions; Table 43 compares human and automated MIV on exactly the same questions. This budgeted secondary sample checks scoring sensitivity and directional consistency. It is not a power-certified confirmation or a census of the 186 questions; intervals are exploratory and unadjusted across contrasts.

Interpretation. Human MIV is positive in all four contrasts, $+5.33$ to $+11.00$pp. All four intervals include zero, with the lower bound for Mem0 seven-day exactly zero. This uncertainty does not establish absence of an effect or failure of the scoring audit: separate significance for each contrast was not its success criterion. It supports a positive directional pattern while leaving the magnitude uncertain. On this same subset, all four Llama intervals and both FLAT Qwen intervals include zero; both Mem0 Qwen intervals exclude zero. In contrast, on the full 186-question primary stratum, all four intervals under each automated judge exclude zero. Comparing human results on 50 questions with automatic results on 186 would mix sampling variation with judge differences.

Table 44: Adjudicated human scores on the same 50-question subset. Accuracies are percentages and contributions are percentage points. Masks and repetitions are averaged within question.

loss	memory	A_0	A_1	R_0	R_1	$C(A)$	$C(R)$
random half	FLAT	34.00	38.00	22.67	32.00	4.00	9.33
7-day	FLAT	34.00	38.00	22.00	34.00	4.00	12.00
random half	Mem0	34.00	35.00	22.67	32.67	1.00	10.00
7-day	Mem0	34.00	35.00	22.00	34.00	1.00	12.00

Judge agreement and differential errors. Agreement with final human labels is 361/463 (77.97%) for Llama and 439/463 (94.82%) for Qwen. Llama has 75 false positives and 27 false negatives; Qwen has 15 and 9. These counts concern distinct responses, not question-weighted task accuracy. Paired human-minus-automatic MIV intervals include zero for every contrast, which does not establish judge equivalence. Memory-on minus memory-off error-rate intervals also include zero in each access regime, without ruling out differential error in a sample this size. Higher Qwen agreement on this audit does not justify replacing the primary judge after seeing these results. Condition-level errors, paired differences and source hashes are preserved with the machine-readable audit.

An observed example. Figure 7 shows how access changes the conclusion for one human-audited question. All three configurations answer correctly with intact sources. After seven-day retention blocks the relevant sessions, Mem0 retains the correct answer while NONE and FLAT do not. The example shows the delivered memory alongside the observed answers.

Question: At which university did I present a poster on my thesis research?
Reference answer: Harvard University

Delivered Mem0 entries (verbatim excerpts)

1. "User presented a poster on their thesis research about education technology at their first research conference over the summer"
2. "User attended their first research conference at Harvard University and observed interesting projects on AI in education"

Access change: Both evidence sessions (S09 and S13, May 21) are blocked by seven-day retention at the May 30 question. Memory was built before the question and remains available.

Source access	Without memory	With FLAT	With Mem0
Intact access	Harvard University Correct	Harvard University Correct	Harvard University Correct
7-day retention	Ivy League University Incorrect	Ivy League University Incorrect	Harvard University Correct

Figure 7: A qualitative MOA example from LongMemEval-S (`a96c20ee`). Both intact repetitions give the answers shown; retention has one run per configuration. Two independent human reviewers agree on every displayed judgment. Mem0 delivers the two relevant facts; FLAT's delivered memory omits the university. This case was selected after inspecting outcomes to illustrate access dependence and its boundary, not to estimate prevalence.

E.3 Exploratory Recall Selection on Disjoint Conversations

Scope and adaptation. We use the ten released conversations of LoCoMo (Maharana et al., 2024), pinned to repository commit `3eb6f2c`. This is a text-plus-provided-caption adaptation, not an official multimodal leaderboard reproduction. We include categories 1–4 and exclude category 5, whose adversarial answer contract differs. A fixed hash selects 50 eligible questions per conversation and allocates four conversations to development (200 questions), with the remaining six for evaluation (300 questions). There is no conversation overlap. These data are new to this MOA experiment; we make no claim that they are absent from model pretraining.

Frozen construction and access. Each conversation supplies one shared memory snapshot constructed before any of its questions are given to the writer. All dialogue text and provided captions are preserved in dated chunks of at most 1,700 characters; the existing 1,800-character observation window therefore contains the complete chunk. Gold answers, QA annotations, evidence labels,

dialogue identifiers, summaries and event annotations do not enter construction. Native Mem0 extraction failures are retained, without selective rebuilding. Access interventions deny entire original sessions, including all their chunks. The evaluation clock is fixed to one day after the final session; it is a constructed post-conversation clock rather than an observed deployment query timestamp. We retain the same reader, top-five retrieval, 1,024-token memory cap and ten-action budget as the corrected study.

Decision rule. The observable input d is the fraction of original sessions denied by the access policy. It is available before the reader runs and uses no gold evidence or response outcome. Candidate policies are the three constants NONE, FLAT and MEM0, and rules recalling FLAT or MEM0 only when $d > t$, for $t \in \{0, 0.1, 0.25, 0.5\}$. Otherwise they use NONE. We compare intact-only selection of a constant, mixture-based selection of a constant, and mixture-based selection from constants and gates. Accuracy is the objective; no cost or latency coefficient is fitted. Ties within 10^{-12} prefer constants in the order NONE, FLAT, MEM0, followed by FLAT gates, then MEM0 gates, with thresholds in descending order.

The primary loss-mixture probability is $p = 0.5$, with 0.25 and 0.75 reported as sensitivities. Restricted conditions give weight $1/12$ to each of three fixed half-session random masks and $1/4$ to each of 7-, 30-, and 90-day retention. Intact access has weight $1 - p$. Development generation and scoring finish before rules and input hashes are locked; only then does evaluation generation begin. The 4,200 development and 6,300 evaluation episodes cover every question–arm–access cell. Policy evaluation selects the corresponding matched arm outcome; it is not an additional independent online agent run. Memory construction cost remains when recall is skipped.

Scoring and inference. Frozen LongMemEval type-specific prompts provide a shared semantic rubric, not the official LoCoMo metric. Rules are selected using the provisional Llama judge on development data; the same frozen choices are assessed by both judges on evaluation data. This study has no human scoring; its estimates rest on the two automated judges. Paired intervals resample the six evaluation conversations, using 10,000 bootstrap draws; leave-one-conversation-out values are preserved in the analysis artifact. Six clusters give limited precision. The multiple mixtures and contrasts are exploratory and unadjusted.

Table 45: Choices fixed on the four development conversations before evaluation generation. The Qwen sensitivity evaluates these same choices.

p	Intact-only constant	Mixture constant	Mixture with gates
0.25	MEM0	MEM0	MEM0
0.5	MEM0	MEM0	MEM0
0.75	MEM0	MEM0	MEM0

Table 46: Evaluation accuracy (%) and paired gains (percentage points) for the policy chosen from the set including gates. That choice may itself be a constant. Intervals use conversations, not individual questions, as clusters.

p	judge	gate-set accuracy	vs. intact choice [95% CI]	vs. mixture constant [95% CI]
0.25	Llama	65.07	$+0.00$ $[+0.00, +0.00]$	$+0.00$ $[+0.00, +0.00]$
0.25	Qwen	61.16	$+0.00$ $[+0.00, +0.00]$	$+0.00$ $[+0.00, +0.00]$
0.5	Llama	63.81	$+0.00$ $[+0.00, +0.00]$	$+0.00$ $[+0.00, +0.00]$
0.5	Qwen	60.32	$+0.00$ $[+0.00, +0.00]$	$+0.00$ $[+0.00, +0.00]$
0.75	Llama	62.54	$+0.00$ $[+0.00, +0.00]$	$+0.00$ $[+0.00, +0.00]$
0.75	Qwen	59.48	$+0.00$ $[+0.00, +0.00]$	$+0.00$ $[+0.00, +0.00]$

This separate exploratory study does not replace the original LongMemEval selection result of zero observed gain. A degenerate zero contrast when two procedures select the same policy is an identity on these replay outcomes, not evidence that all possible selection policies have zero value. Results do not establish deployment cost savings or broad generalization beyond the six evaluation conversations and the stated access mixture. Shared-server timing does not identify deployment latency or cost-adjusted utility.

F TOOL-AGENT STUDIES: WEBSHOP

Status. Controlled demonstration and scope check for Section 5.2. The controlled MOA study is an information-path test, not a leaderboard result; replaced and withdrawn implementation-transfer contributions are marked in the last subsection.

F.1 CONTROLLED UTILITY AND EXECUTION AUDITS

A derived WebShop task from released MemoryArena bundles (He et al., 2026) pairs identical instructions and candidates with different prior purchases, so the history-compatible action changes. Receipt contribution is $+22.5$pp with replay available and $+55.0$pp without, giving MIV $+32.5$pp $[+15, +50]$ on 20 counterfactual pairs; a separate 50-pair study raises exact success from 15.0% to 84.0%. The construction deliberately makes the prior purchase determine the correct action, so this is a controlled information-path test, not a leaderboard result (Appendix F).

Public-implementation results use the correction recorded in Appendix F.5. The matched direct study gives $+16$pp $[+8, +24]$ for FLAT and -4pp $[-9, +0]$ for Mem0. Source-access projection contributions remain withdrawn; only their difference, $+22.5$pp $[+2.5, +45.0]$, is retained. Within the matched contract, changing receipt delivery costs 19.0pp; the expected record tag recovers 13.0pp at unchanged layout. The controlled result in Table 47 remains valid: each of its 40 units contains all four cells from one execution configuration.

F.2 CONTROLLED WEBSHOP CELLS

Table 47: Controlled WebShop exact history-compatible purchase success and percentage-point effects. Intervals use 20,000 counterfactual-pair bootstrap draws. For the direct study these do not account for dependence between pairs sharing a source bundle; the bundle-connectivity sensitivity is in Appendix F.3. The direct gain and source-access interaction come from separate, disjoint studies and are not interchangeable estimands.

study	unit	off	on	estimand	95% CI
fixed-use direct utility	50 pairs	15.0	84.0	$G = +69.0$	$[+60, +78]$
receipt available (A)	20 pairs	67.5	90.0	$C(A) = +22.5$	$[+12.5, +32.5]$
receipt unavailable (U)	20 pairs	32.5	87.5	$C(U) = +55.0$	$[+42.5, +67.5]$
source-access interaction	20 pairs	disjoint A/U factorial		MIV $= +32.5$	$[+15, +50]$

This appendix reports two studies that are not pooled: the initial pipeline-conditioned realised-prefix evaluation and a later canonical-prefix confirmation prompted by a semantic-feasibility audit. The original artifacts and registered analyses remain unchanged; the confirmation uses disjoint development and final bundles under a new task-coherent contract.

Frozen execution configuration. Both Shopping pipelines use the local OpenAI-compatible endpoint name `agent-llm`, backed by the pinned Qwen3-VL-8B-Instruct snapshot `0c351dd0...`, greedy decoding, 256 output tokens per call, a 65,536-token context window, a 1,024-token margin, and at most 20 actions per task. Hence the checked input cap is $65{,}536 - 1{,}024 - 256 = 64{,}256$. The server uses tensor parallelism 2 with prefix caching disabled and eager execution. Long-context supplies the recorded history; BM25 uses native `RAGMemorySystem` with 2,048-token chunks and top-3 lexical retrieval. These settings are distinct from the document contract above.

Recruitment and eligibility. The frozen candidate order includes item-2 through item-14 by category; item-0 and item-1 were reserved for pilots. The manifest SHA-256 is `24b7cbadd8be0a310cdcff2fe5777953958d9af04788b492d96c924d00554fc0`. Eligibility requires a realised ancestor purchase and an interpretable frozen compatibility rule before any factorial outcome. Long-context attempted 53 bundles and retained 49 (153 steps). Its original minimum of 50 was not met; the scorer first failed closed, and analysis proceeded only after an outcome-blind amendment that retained all pre-eligible bundles at the frozen cap. BM25 attempted 54 and retained 50 (143 steps), meeting its original minimum. The excluded long-context bundles are `beauty_item_3`, `beauty_item_5`, `beauty_item_6`, and `beauty_item_7`; BM25 excludes `baking_item_10`, `beauty_item_6`, `beauty_item_12`, and `beauty_item_13` under the same prefactorial rule.

Table 48: Initial realised-prefix Shopping absolute utilities and contributions under the same bundle-macro weights (% or percentage points). Brackets are paired 95% percentile intervals. The corresponding primary interactions and all original observations are retained in the frozen artifact.

pipeline	endpoint	arm	memory on	memory off	contribution
long	K	A	$40.65\,[31.80, 49.32]$	$38.10\,[29.42, 46.77]$	$+2.55\,[-7.65, +12.24]$
		U	$37.76\,[28.91, 46.77]$	$31.12\,[22.79, 39.97]$	$+6.63\,[-4.25, +16.84]$
		P	$37.59\,[29.08, 45.92]$	$34.01\,[26.02, 42.18]$	$+3.57\,[-6.80, +13.27]$
	V	A	$6.97\,[2.89, 11.90]$	$5.27\,[1.70, 9.35]$	$+1.70\,[-3.40, +6.80]$
		U	$6.46\,[2.72, 10.88]$	$8.50\,[3.40, 14.63]$	$-2.04\,[-8.50, +3.74]$
		P	$6.97\,[3.23, 11.39]$	$6.12\,[2.55, 10.71]$	$+0.85\,[-4.08, +5.78]$
	VK	A	$4.25\,[1.02, 8.67]$	$3.74\,[1.02, 7.31]$	$+0.51\,[-3.57, +4.59]$
		U	$3.74\,[1.02, 7.48]$	$5.27\,[1.02, 10.88]$	$-1.53\,[-7.14, +3.06]$
		P	$4.25\,[1.19, 7.99]$	$3.06\,[0.51, 6.63]$	$+1.19\,[-2.55, +5.10]$
BM25	K	A	$22.33\,[14.67, 30.67]$	$36.33\,[27.50, 45.17]$	$-14.00\,[-23.67, -4.67]$
		U	$21.50\,[14.33, 29.33]$	$34.33\,[25.33, 43.83]$	$-12.83\,[-22.83, -2.67]$
		P	$23.00\,[15.83, 30.67]$	$36.33\,[26.50, 46.00]$	$-13.33\,[-24.67, -2.16]$
	V	A	$3.50\,[1.17, 6.33]$	$5.50\,[1.33, 11.00]$	$-2.00\,[-8.00, +3.00]$
		U	$2.33\,[0.00, 5.33]$	$7.67\,[3.00, 13.33]$	$-5.33\,[-11.00, -0.83]$
		P	$2.67\,[0.50, 5.50]$	$6.83\,[3.17, 11.00]$	$-4.17\,[-7.83, -0.83]$
	VK	A	$1.50\,[0.00, 3.50]$	$3.33\,[0.67, 6.50]$	$-1.83\,[-5.33, +1.50]$
		U	$0.00\,[0.00, 0.00]$	$5.00\,[1.67, 8.83]$	$-5.00\,[-8.83, -1.67]$
		P	$1.00\,[0.00, 2.50]$	$4.17\,[1.50, 7.33]$	$-3.17\,[-6.33, -0.33]$

Endpoint decomposition. All endpoints retain the same eligible steps, and the scorer verifies $K = VK + (1 - V)K$ at cell, bundle, and contrast levels. For invalid-compatible outcomes, long-context has MIV $+6.12$pp $[-2.72, +15.31]$ and $\mathrm{MIV}_{src} = +5.78$pp $[-4.25, +15.82]$; BM25 has $+4.33$pp $[-6.50, +15.00]$ and $+2.33$pp $[-8.67, +13.50]$. These intervals include zero. An all-zero observed cell gives a mechanical percentile-bootstrap $[0, 0]$; this does not establish a zero population probability. All intervals are individual and unadjusted, and displayed decompositions may differ by 0.01pp because values are rounded to two decimal places.

Execution and source landing. The 918 long-context and 858 BM25 cells have zero execution, cell, or cleanup errors. All 306 long-context and 286 BM25 available-arm cells serve the ancestor; the corresponding unavailable cells deny it; placebo serves it while denying the frozen non-ancestor. Prefix snapshot/clone hashes pass for all 49/50 eligible bundles, every downstream step has six cells, exact token-count mismatches are zero, and no message is trimmed.

Semantic-feasibility scope correction. A post-hoc CPU census enumerates, for every initially eligible step, the frozen products satisfying both canonical current-task validity and realised-ancestor compatibility. The joint set is nonempty for 77/153 long-context steps and 61/143 BM25 steps; 76 and 82 steps, respectively, are structural zeros for VK. Every such step has zero contribution and interaction in all observed cells, so these steps dilute rather than generate an interaction. The registered K analyses remain arithmetically defined, but the consumer acted under a current target that could conflict with realised-history compatibility. We therefore treat the negative BM25 contribution as specific to this operational adapter. The retrospective feasible subset is diagnostic only and does not replace any registered estimate.

Pipeline comparison audit. The common-ID subset contains 46 bundles. BM25-minus-long is -6.88pp $[-20.47, +6.70]$ for MIV and -6.52pp $[-22.29, +9.06]$ for MIV_{src}. Identical build-purchase trajectories occur in 0/46 bundles, identical eligible-step sets in 16/46, and identical ancestors in 38/116 shared steps. This is therefore a descriptive comparison of pipeline-conditioned estimands. Step-micro values are retained only as descriptive statistics: long-context MIV/MIV_{src} are $+3.27/+1.96$pp and BM25 values are $0.00/-4.90$pp.

Table 50: Valid-compatible purchase in the canonical-prefix held-out final ($B = 50$). MIV is 0pp $[-10, +10]$, MIV_{src} is -4pp $[-16, +8]$, and $C(U)$ is -2pp $[-12, +8]$. Intervals are individual 20,000 replicate bundle-bootstrap intervals with seed 20260908.

access	memory on	memory off	contribution
available	18%	20%	−2pp
unavailable	14%	16%	−2pp
placebo	20%	18%	+2pp

Table 49: Exploratory Shopping category estimates (pp). Categories are strata within one task family, not independent tool domains.

category	long MIV	long MIV_{src}	BM25 MIV	BM25 MIV_{src}
baking	+13.33	+6.67	0.00	+5.00
beauty	+7.41	+1.85	+3.33	+11.67
electronics	−12.50	−10.00	+7.50	−5.83
grocery	−5.00	+7.50	+10.83	+7.50
home	+17.50	+9.17	−15.83	−15.83

All intervals in the initial realised-prefix Shopping study use 20,000 bundle bootstrap resamples with seed 0. The frozen scorer and pipeline-comparison analyses are `GeneralMOA/code/g_moa_shop_score.py` and `GeneralMOA/code/g_moa_shop_compare_pipelines.py`; their strict contracts require the prefix audit, valid-compatible endpoint, memory cleanup, and manifest hash. The associated test suite reports 25/25 passing tests.

Canonical-prefix confirmation contract. The follow-up scripts the benchmark's canonical purchases through the real environment, records only the observed receipt, and branches from the resulting state. The original downstream instruction, candidate set, validity target, and compatibility rule then refer to the same canonical chain. Memory-on receives exactly the routed ancestor receipt by controlled delivery; memory-off receives no persistent payload. A pre-factorial census finds 84 coherent steps across 25 development bundles and 165 across 50 held-out bundles. The frozen rule selects the earliest coherent step per bundle. Coherence is therefore established for the selected 50 final steps, not for every candidate step or for autonomous end-to-end acquisition.

One generic development revision instructs the consumer to bind the previous receipt to the compatibility rules before applying product preference. Its 25-bundle MIV point estimate changes from 0pp $[-12, +12]$ to $+12$pp $[-4, +32]$; the paired revision contrast is $+12$pp $[-8, +36]$. This is used to freeze the consumer, not claimed as a confirmed effect.

The unavailable-arm paired outcomes are success in both cells for four bundles, memory-only success for three, no-memory-only success for four, and failure in both cells for 39. All 300 scripted prefix purchases match their target and all 300 factorial cells complete with no intervention, payload, isolation, trimming, or scoring violation; the maximum input is 26,156 of 64,256 tokens. The scripted prefix count is evaluation-experience construction, not consumer task success. Two trusted-prefix search failures are repaired under versioned, outcome-blind acquisition amendments before factorial outcomes are inspected; ordinary factorial failures are not retried. The held-out final does not confirm the positive development estimate, but its interval also does not establish absence or equivalence. We freeze this canonical-prefix final and close additional prompt, memory-schema, category, and sample search for that configuration. The separately specified derived-task study follows Section F.1.

F.3 Direct Memory-Utility Follow-up

This study was opened after the original Shopping and NExT-GQA factorial results were frozen. Its estimand is direct paired utility gain, not MIV, and its sample and artifacts are not pooled with the factorial studies.

Controlled WebShop construction. All 150 released MemoryArena bundles had been exposed, so the new evaluation is explicitly a controlled derived task. A deterministic census identifies 509 eligible source rows and 1,158 counterfactual history pairs. In each pair, two prior purchases imply different compatible current purchases while the public instruction and candidate set remain identical. Memory-on receives the observed receipt; memory-off does not. The derived-task pool reserves 30 source bundles for development and 120 for final construction with zero overlap, although all 150 had appeared in earlier studies. "Held-out" therefore describes disjoint new tasks and uninspected outcomes within this follow-up, not unseen original benchmark material. Development selects 10 pairs from 6 source bundles; the frozen final selects 50 pairs and 100 history instances from 48 source bundles. Each history carries one recorded source-bundle id, and in every pair the two histories come from different bundles; between one and six distinct pairs touch a given bundle, with a median of two. Because bundles are shared across pairs, the pair-clustered intervals we report do not model that dependency: resampling pairs that share a bundle together leaves seven connected components, the largest containing 24 of the 50 pairs. Exact history-compatible purchase is the binary endpoint.

The held-out paired outcomes are 70 memory-only successes, one no-memory-only success, 14 successes in both conditions, and 15 failures in both. Thus memory-on succeeds in 84/100 instances and memory-off in 15/100. The $+69$pp interval $[+60, +78]$ resamples the 50 counterfactual pairs as clusters for 20,000 percentile-bootstrap draws with seed 20260908. That primary interval conditions on the selected source-bundle composition and does not model the dependence induced by shared bundles, which is not a marginal structure here: the largest component holds 24 of the 50 pairs. Leaving each component out in turn moves the gain between $+64.5$ and $+75.0$pp, and resampling whole components with replacement gives $[+62.5, +81.1]$pp. The first is an influence range and the second a sensitivity over only 7 blocks; neither is a confidence interval for a new population of bundles, and both are reported beside the primary interval rather than in place of it. The same diagnostics for the two public projections and their paired difference are in Table 51.

Table 51: Shared-bundle sensitivity of the 50-pair WebShop direct-utility study. All contrasts preserve equal pair weighting and use the same component-resampling draws. Leave-one-component-out ranges measure influence. Whole-component percentile ranges are exploratory sensitivity summaries over seven unequal blocks, not calibrated confidence intervals for a new population of source bundles. The point estimator is unchanged, but the resampling unit differs; with only seven unequal components these ranges are exploratory sensitivity summaries rather than stronger evidence of a nonzero effect. The nominal pair-bootstrap intervals do not account for dependence between pairs sharing bundles. The separate 20-pair source-access study is not included.

contrast	gain	nominal pair CI	leave-one-component-out	component resampling
Raw receipt − no memory	$+69$	$[+60, +78]$	$[+64.5, +75.0]$	$[+62.5, +81.1]$
A-MEM/FLAT − no memory	$+6$	$[-2, +14]$	$[+4.3, +9.6]$	$[+2.1, +17.3]$
Mem0 − no memory	-13	$[-19, -7]$	$[-15.8, -7.7]$	$[-18.0, -3.9]$
A-MEM/FLAT − Mem0	$+19$	$[+12, +26]$	$[+17.3, +21.1]$	$[+13.6, +25.0]$

All 100 instances complete without infrastructure failure or outcome-based exclusion. This result is conditional on the derived task generator and is not an official MemoryArena leaderboard result.

Execution scope. Table 51 retains the superseded comparisons documented in Appendix F.5. Beyond that environment/context mismatch, 4 of 100 memory-off episodes exceed 16,384 tokens and 1 of 100 FLAT episodes exceeds its served window. Fixed task manifests do not fix the searchable catalogue; the second environment is harder for every condition evaluated in both. The model snapshot, decoding settings, instruction file and disabled prefix caching are shared.

A payload audit over the same 100 histories separates what is deliverable from what is delivered. The A-MEM and FLAT payloads are byte-identical on 100/100 tasks, so they are one condition rather than two. The raw receipt and the FLAT projection each carry the compatibility-relevant ASIN on 100/100 tasks, and no payload reaches the 1,024-token delivery cap, so neither content loss nor truncation distinguishes them. Mem0 delivers the ASIN on 70/100; of the 30 losses, 0 are stored but not retrieved and 30 are never stored, which places the loss at extraction rather than selection. Identifier survival is not the whole of the gap: restricted to the 70 payloads that do keep the ASIN, Mem0 succeeds on 2, where FLAT succeeds on 11 of those same 70. The wider-looking 2/70 against 21/100 compares different subsets, since FLAT keeps the identifier on 100/100; on the

matched subset the gap is 9 tasks rather than the 19-point difference in the unmatched rates. Both arms ran in the same environment instance here, so this within-study comparison is unaffected by the mismatch described in Appendix F.5.

The dominant failure under the common contract is non-completion rather than a wrong purchase. The original delivery ends 11/100 episodes without any purchase; the two common-contract projections end 78/100 that way, after roughly twice as many turns (7.7 and 7.6 against 3.7). One candidate explanation is visible in the shared instruction, whose decision procedure names a served `source_purchase_record` or `purchase_receipt` and then instructs the agent not to assume the content of an unavailable record, while the common contract delivers the same receipt under a heading matching neither name. We report that as a hypothesis consistent with the pattern, not as an identified cause.

A bridge study under one execution contract. We re-ran every condition on the same 50 pairs in environment 13037486 with a 65,536-token context, varying only the prefix attached to the user prompt. Seven conditions: no payload in the original framing (O0) and in the common framing (C0); the unmodified receipt in the original `memory_context` wrapper (OR), under the common heading (CR), and under the common heading but tagged as a `purchase_receipt` (CRt); and the FLAT (CF) and Mem0 (CM) projections under the common heading. Within this contract the bridge reproduces the archived projection arms: 22 of 100 history instances reach the target under CF against 21 in the archived run, and 2 under CM against 2. It does not reproduce the archived raw-receipt level, which was measured in the other environment.

Table 52: WebShop bridge: 50 counterfactual pairs, one environment instance, one context window, equal pair weighting, 20,000 paired draws. Percentage points. The last three rows are the within-environment counterparts of the corresponding rows of Table 51, whose archived values are $+6$, -13 and $+19$. All three move, but by very different amounts: the two that crossed environments move by 10 and 9 points, while the system difference, in which the mismatched baseline cancels, moves by 1. The last two columns are the shared-bundle sensitivity of each row: the 50 pairs draw on 31 distinct source bundles, and the ranges resample those bundles with a ratio estimator that keeps equal pair weighting within a draw. They are exploratory sensitivity summaries over unequal blocks, not calibrated intervals for a new population of bundles. Under that resampling the Mem0-minus-no-memory range reaches 0.0, so its negative sign is not robust to bundle dependence. The delivery-contract row is formed per pair as the difference of the two contrasts above it and then passed through the same draws, rather than by differencing their endpoints, which would not be valid. Full per-row values are in `webshop_bridge_dependency.tsv` and `e2_dlayout_sensitivity.json` of the supplement.

contrast	cells	gain	nominal	leave-one-out	component
receipt, original contract	OR−O0	+28.0	[+19, +36]	[+26.5, +29.6]	[+20.4, +35.4]
receipt, common contract	CR−C0	+9.0	[+3, +16]	[+7.3, +10.4]	[+2.1, +16.0]
cost of delivery contract	CR−C0 − (OR−O0)	−19.0	[−27, −11]	[−20.8, −17.7]	[−26.5, −11.7]
recovered by record tag	CRt−CR	+13.0	[+6, +20]	[+11.5, +13.8]	[+6.8, +20.2]
FLAT over raw receipt	CF−CR	+7.0	[+2, +13]	[+6.1, +7.4]	[+2.9, +11.9]
Mem0 over raw receipt	CM−CR	−13.0	[−20, −6]	[−14.3, −11.5]	[−19.3, −6.7]
FLAT − no memory	CF−C0	+16.0	[+8, +24]	[+14.6, +17.7]	[+8.0, +23.9]
Mem0 − no memory	CM−C0	−4.0	[−9, +0]	[−5.1, −3.1]	[−8.7, 0.0]
FLAT − Mem0	CF−CM	+20.0	[+13, +27]	[+18.8, +20.8]	[+13.3, +26.7]

Three readings follow, all within this contract. First, with no payload attached the two framings score identically at 6.0%, so the framing by itself does not move the endpoint. That is a statement about the empty conditions only: equal accuracy without a payload does not imply equal behaviour, and it does not rule out a layout effect once a payload is present. Second, the same receipt is worth 19.0pp less under the common heading, and tagging it as the record type the instruction names recovers 13.0pp of that while leaving the heading and bullet layout unchanged. This supports a contribution from instruction–payload alignment; it does not isolate every layout effect, and we do not claim the residual 6.0pp is or is not one. Third, Mem0 remains 13.0pp behind the raw receipt under identical delivery, which is consistent with the extraction-stage identifier loss above and is not explained by wrapper or capacity.

Repeated cells provide a reproducibility bound. A server restart caused 21 cells to run twice under the same contract: success agreed on 20 of 21, the purchased item on 17, and the turn count on 13. Individual trajectories in this environment are therefore not reproducible even with greedy decoding

and prefix caching disabled, although the binary endpoint is nearly so. We keep the first run of each repeated cell.

F.4 Source-Access Follow-up

This study uses four cells, A0/A1/U0/U1, on units disjoint from its earlier two-cell direct-efficacy sample. It estimates MIV $= (U1 - U0) - (A1 - A0)$ and does not contain a placebo arm. All intervals are individual, unadjusted percentile intervals from 20,000 paired-cluster bootstrap draws with seed 20260909. The frozen artifacts and a checksum registry are under `GeneralMOA/source_access_followup/`.

Controlled WebShop receipt replay. The source-available arm materializes the realised acquisition receipt in the environment instruction; the unavailable arm omits it. Memory-on independently injects the persistent receipt. An initial development implementation logged the source only in driver bookkeeping and therefore produced a technical no-op; a second diagnostic retained an unrelated suffix from the older MemoryArena workflow. Both were corrected before any held-out cell was run. The frozen final comprises 20 new counterfactual pairs, 40 history instances from 40 source bundles, and 160 cells. All cells complete without error, and an artifact prompt audit finds the source tag only in A and the memory tag only in memory-on (0/160 violations). Memory changes 11 failures to successes and 2 successes to failures in A; in U, the corresponding counts are 23 and 1. MIV is positive for 12 pairs, zero for 6, and negative for 2. Table 47 reports the pair-cluster estimate.

F.5 Execution Provenance and Replaced Results

Execution provenance of the WebShop cells. A later audit of which run produced each cell found that neither WebShop study executed all of its cells under one contract. Table 53 records what we found. In the direct study the raw-receipt and memory-off arms ran in environment instance 25101970 at a 65,536-token context and the projections in 13037486 at 16,384; we re-ran every condition under one contract, and the bridge of Table 52 supersedes those gains. In the source-access study the projections' memory-on cells ran in a third instance, 45722587, at 16,384, while their A_0 and U_0 were joined by task identifier from the earlier run in 48844776 at 65,536; the reuse is visible in the scoring code and in the fact that all three systems report byte-identical $A_0 = 67.5$ and $U_0 = 32.5$. The transfer plan allowed shared off cells "after equivalence"; no equivalence artifact exists, and the verification script written to keep the manuscript aligned with the artifacts reads only the direct study's files. We report the source-access projection contributions as withdrawn rather than re-run them. The original controlled WebShop study is not affected: each of its 40 units carries all four cells from one run under one configuration hash.

Table 53: Which run produced each WebShop cell. A contrast is identified only where the cells it subtracts share a row's environment and context. Withdrawn rows are marked in Table 35 without estimates; the difference between the two projections is retained, since the imported baseline cancels in it.

study	cells	env instance	context	status
controlled WebShop	all four	48844776	65,536	reported
transfer, direct	receipt, off	25101970	65,536	superseded
	projections	13037486	16,384	superseded
	bridge, all seven	13037486	65,536	reported
transfer, source	projections on	45722587	16,384	withdrawn
	off, imported	48844776	65,536	withdrawn

The frozen plan, lock, amendment, report, code, manifests, outputs, audits, and SHA-256 registry are under `GeneralMOA/public_memory_system_followup/`.

G Video Memory Scope and Boundaries

Status. Scope and boundary check for Section 5.2. Post-observation video memory shows no consistent benefit; EgoLifeQA follow-ups use a single life record.

Table 54: Task-generated Video tool use and delivered payloads, reported as the number of question-cells with at least one indicated event (not total calls). Technical probes separately verify the intended source policy in every cell. Requests are agent actions; served counts reflect the requested segment, source-access policy, and frozen execution budgets.

cell	raw call	target req.	target served	control req.	control served	memory req.	memory payload	target-memory
A, on	110	54	49	49	43	16	16	9
A, off	110	53	48	49	43	16	0	0
U, on	110	54	0	65	59	24	24	15
U, off	110	53	0	63	57	24	0	0
P, on	110	59	53	49	0	22	22	14
P, off	110	59	53	49	0	22	0	0

The long-video boundary. On one EgoLifeQA record (Yang et al., 2025), a local pipeline (Yeo et al., 2026) gains +20pp $[+10, +30]$ on a frozen 110-question block but not on a later 50-question one ($-$2pp $[-16, +12]$); every estimate conditions on that single record (Appendices G, G.1 and G.2).

Frozen runtime and acquisition. Both the question-free event writer and downstream consumer use Qwen3-VL-8B-Instruct revision `0c351dd0. . .` , vLLM 0.16.0, greedy decoding, eager execution, and disabled prefix caching on one GPU. The server context is 65,536 tokens; after a 1,024-token safety margin and 256 output tokens, the checked input cap is 64,256. Acquisition divides decoded video into 8-second half-open PTS intervals and samples at most eight chronological frames per interval. The writer sees neither question, choices, answer, nor gold temporal span. It writes one short event per segment; the immutable snapshot stores the frame IDs, actual PTS values, contact sheet hash, writer prompt hash, and memory hash.

Recruitment and precision plan. Of 567 valid raw validation videos, 165 original videos exposed during prior development were excluded before the clean hash order. Twelve pilot videos and four additional acquisition-only pilot candidates were then excluded from final recruitment. The outcome-blind pilot progression rule required source and memory exposure across independent videos but never used accuracy, MIV sign, or significance; it passed. Separately, the larger of the two pilot per-video contrast dispersions was $s = 0.12975$. A 5pp normal-approximation planning target for the half-width, with 1.5-fold variance inflation, fixed $N = 39$ and an acquisition-attempt cap of 48 before final evaluation; this target was not a guarantee of realised precision. All 48 acquisitions completed. The first 39 candidates were structurally eligible; their 215 eligible questions were reduced to 110 only by the frozen maximum of three hash-ordered questions per video. No question was filtered by downstream tool use or outcome. The final manifest SHA-256 is `335bce5c0cba6f2c. . . .`

Intervention and isolation audit. The 660 expected cells occur exactly once and all complete without infrastructure or answer-parse failure. All clone IDs are unique; each question's six branches share one initial environment hash and one memory-snapshot hash. Synthetic source probes pass in every cell, U never returns a target frame, P never returns a control frame, and memory-off returns no persistent payload. Inference runs in a mount namespace where the public question manifest is readable but scorer gold, the original annotation directory, and a legacy gold-bearing bundle are not. Contact-sheet hashes, half-open frame boundaries, and frame-origin uniqueness pass for every selected video. The 22-check integration audit passes. Of the completed cells, 106 reach the six-call tool cap and 179 reach the four-read raw cap; these are retained as normal agent outcomes. In allowed arms, every requested-but-unserved target or control is explained by the frozen four-read raw cap: 5 target and 6 control question-cells in each A memory condition, 6 control question-cells in each U memory condition, and 6 target question-cells in each P memory condition. Gate denials are reported separately for U targets and P controls.

Table 56: EgoLifeQA paired direct-utility results. The non-overlapping 110-question extension is the primary new estimate. The cumulative row is secondary, and the embedded 100-question row is descriptive rather than a stopping point. Intervals are paired question-bootstrap percentile intervals from 20,000 draws with seed 20260908.

question set	N	memory on	memory off	paired gain (pp)
initial held-out, IDs 11–30	20	7	4	$+15.00$ $[-10.00, +40.00]$
new extension, IDs 31–140	110	57	35	$+20.00$ $[+10.00, +30.00]$
cumulative, IDs 11–140	130	64	39	$+19.23$ $[+9.23, +29.23]$
embedded milestone, IDs 11–110	100	54	30	$+24.00$ $[+13.00, +35.00]$

Table 55: NExT-GQA original-video-macro accuracy and memory contribution (percentage points). Intervals are individual 95% percentile intervals from 20,000 original-video cluster resamples.

access	accuracy, memory on	accuracy, memory off	$C(a)$
A	66.67	66.67	-0.00 $[-2.56, 2.56]$
U	64.96	66.67	-1.71 $[-5.13, 1.71]$
P	67.95	66.24	$+1.71$ $[0.00, 4.27]$

Results, decomposition, and budget. Table 55 reports all six video-macro utilities. The source changes are $\bar{Y}(U,1) - \bar{Y}(A,1) = -1.71$pp $[-5.98, +2.56]$ and $\bar{Y}(U,0) - \bar{Y}(A,0) = 0.00$pp $[-3.42, +3.42]$; their difference is MIV. The placebo contribution change $C(P) - C(A)$ is $+1.71$pp $[+0.00, +4.27]$, and together with $\mathrm{MIV}_{src} = -3.42$pp it recovers MIV exactly before rounding. Question-micro descriptive MIV and MIV_{src} are -1.82pp and -3.64pp, respectively, and are not mixed with the primary video-macro analysis. Across 2,598 consumer requests, mean server input is 2,091.4 tokens and the maximum is 7,325; the locally audited maximum visual component is 4,788 tokens. The maximum server-versus-local processor difference is 168 tokens. Accordingly, server-reported input-budget compliance passes with a 56,931-token margin, whereas exact server–local count agreement is not met and is not claimed. The originally planned exact-count-agreement criterion was not satisfied. We report this validation deviation separately from server-reported input-budget compliance; no threshold, sample, cell, or effect estimate was changed.

The frozen implementation, report, video table, and checksum registry are under `GeneralMOA/video/`. Intervals use 20,000 original-video cluster bootstrap resamples with seed 0 and are individual and unadjusted. Supplementary recomputation deterministically reproduces scoring and aggregation from the same frozen execution artifacts; it is not an independent model rerun.

G.1 EgoLifeQA Direct Memory-Utility Follow-up

This study was opened after the original Shopping and NExT-GQA factorial results were frozen. Its estimand is direct paired utility gain, not MIV, and its sample and artifacts are not pooled with the factorial studies.

EgoLifeQA configuration and ordering. The local run uses WorldMM commit `3a55b65235e4f9618626a91c28e7e45baa6d8bdf`, released A1_JAKE annotations, multiscale captions, semantic memory, visual embeddings, and the released 30-second OpenIE artifact. Memories later than each question's query time are excluded. Memory-on uses WorldMM's iterative memory policy and memory-off provides the same Qwen3-VL-8B responder no retrieved context. The local Qwen retriever and structured-output compatibility adapter differ from the published WorldMM configuration.

The earliest three questions form an engineering smoke, IDs 4–10 form a seven-question development extension, and IDs 11–30 form the first frozen held-out sample. The first held-out sample yields 7/20 versus 4/20, a $+15$pp gain $[-10, +40]$. Only after observing it, its paired variance is used to freeze the complete next 110-question list, IDs 31–140, for an approximately 10pp question-level half-width. No extension prediction is observed before that list and stopping rule are fixed.

Table 57: Descriptive later-block results by question type. Counts are correct answers. High transfer accuracy alone is not evidence of memory utility; the paired no-memory comparator is required. Small strata are not confirmatory and were not used to select a new primary population.

question type	N	no memory	WorldMM	EgoCITE transfer	transfer − no memory
TaskMaster	7	4	3	6	$+28.6$pp
HabitInsight	4	1	1	4	$+75.0$pp
RelationMap	14	10	10	9	-7.1pp
EventRecall	11	2	3	2	0.0pp
EntityLog	14	3	2	1	-14.3pp

The 110-question extension contains 30 off-fail/on-success, 8 off-success/on-fail, 27 both-success, and 45 both-fail outcomes. Three infrastructure attempts terminate without result artifacts while exposing schema-invalid null triples, JavaScript-style comments outside JSON strings, and redundant graph rebuilding; outcome-blind compatibility amendments retain the same questions, models, retrieval depths, prompts, and analysis. The first complete artifact then contains CUDA out-of-memory errors in only IDs 31/on and 32/on during initial shared-GPU contention. The independent scorer rejects rather than scores these cells. Before repair, a separate contract freezes all and only the two errored cells; they are rerun from the clean ID-30 cache under the unchanged scientific settings, and the merge preserves every original memory-off and non-error memory-on result. The final 110 pairs contain no inference errors.

All EgoLifeQA rows concern one A1_JAKE life record. Question-level paired resampling therefore quantifies uncertainty across these questions, not across people or independent lives. The intervals do not additionally model dependence among questions that share events or temporal context within that record. The positive result supports this enabled local memory pipeline against its no-retrieved-context baseline; it does not identify which WorldMM component causes the gain. That two-cell result alone does not establish a source-access interaction; the disjoint test of that question follows.

Block usage and ordering. EgoLifeQA questions are consumed in dataset order and each block is used once as a frozen estimate, so we state the order rather than leave it to be inferred. IDs 1–3 are an engineering smoke; 4–10 a development extension; 11–30 the first frozen held-out sample; 31–140 the extension list, frozen after 11–30's paired variance was observed and before any of its own outcomes; 141–145 development and 146–195 the frozen held-out for the forced still-frame study, both disjoint from 1–140. IDs 196–245 are used twice, and the order matters: the repaired audiovisual-source final ran first and was frozen before selector generation or prediction, and the EgoCITE evidence-transfer comparison ran afterwards on the same block, with that final's outcomes already observed. That is why the transfer comparison is reported as development throughout and is never treated as a confirmatory estimate or used to select a population; no held-out claim in this paper rests on it.

Later-block evidence-transfer development. On IDs 196–245, the same native WorldMM path yields 19/50 correct versus 20/50 without retrieved context, a -2pp gain $[-16, +12]$. A frozen development comparison transfers the raw dense captions selected by the public EgoCITE evidence provider to the same Qwen3-VL-8B response prompt. It obtains 22/50 correct: $+4$pp $[-10, +18]$ over no memory (8 helpful, 6 harmful flips) and $+6$pp $[-6, +18]$ over WorldMM (7 helpful, 4 harmful). Public predictions, answers, correctness fields, gold timestamps, and final reasoning are excluded; all transferred captions end no later than query time. The public artifact omits ID 198, which remains in the denominator under the predeclared fallback to its completed no-memory response. This is an evidence-provider transfer with an upstream 27B pipeline, unequal context and compute, and no independent final; it neither reproduces EgoCITE under the local model nor estimates MIV.

G.2 EgoLifeQA Source-Access Follow-up

This study uses four cells, A0/A1/U0/U1, on units disjoint from its earlier two-cell direct-efficacy sample. It estimates $\text{MIV} = (U1 - U0) - (A1 - A0)$ and does not contain a placebo arm. All intervals are individual, unadjusted percentile intervals from 20,000 paired-cluster bootstrap draws with seed 20260909. The frozen artifacts and a checksum registry are under `GeneralMOA/source_access_followup/`.

Table 58: Repaired EgoLifeQA audiovisual-source final on 50 A1_JAKE questions. MIV is 0pp $[-14, +14]$ and MIV_{src} is -10pp $[-26, +6]$. Intervals use 20,000 paired-question percentile-bootstrap draws.

source payload	memory off	memory on	$C(a)$	individual 95% CI
A: routed audiovisual caption+frames	52.0	50.0	-2.0	$[-12, +8]$
U: no source payload	40.0	38.0	-2.0	$[-16, +12]$
P: matched non-target caption+frames	30.0	38.0	+8.0	$[-6, +22]$

EgoLifeQA forced still-frame payload. Development uses IDs 141–145 and the held-out estimate uses IDs 146–195, all disjoint from direct-efficacy IDs 1–140. For each question, a frozen BM25 router selects the top three positive-scoring pre-query 30-second captions from the question and choices. The A arms receive eight sampled raw frames per interval (24 frames total); the U arms receive none. Routing captions are not consumer inputs. The memory-on arms use the same local WorldMM retrieval path as the direct-efficacy study, whereas memory-off has no retrieved items.

The first development launch was mistakenly placed on GPU4 while an unrelated process occupied about 20 GB. Its four OOM/dependent-error cells for ID 141 are preserved and excluded. Before the valid rerun, the episodic cache was restored from a clean clone and the runner was amended to verify a root-independent cache tree hash. Development and final preflights then pass without changing the question set or scientific settings. The held-out run contains exactly 50 questions and 200 cells, with zero inference errors, zero delivery violations, and at least one retrieved item in every memory-on cell. One logged invocation of the frozen compatibility rule drops non-string triple rows but causes no cell error or replacement.

In A, memory changes 14 failures to successes and 6 successes to failures; in U, the counts are 9 and 5. Per-question MIV is positive for 8 questions, zero for 30, and negative for 12. The resulting negative MIV point estimate is uncertain. More importantly for interpretation, U0 exceeds A0. Moreover, the question-and-choice router selects transcript-bearing visual–audio captions while the consumer receives only sampled frames. The experiment therefore estimates the interaction for forced still-frame provision, not memory's contribution under loss of all usable past audiovisual information. The result is conditional on one A1_JAKE record and does not test full loss of past-video replay. These are design problems with results consistent with them; this analysis does not isolate their individual causal contributions.

EgoLifeQA audiovisual-caption repair. The old 50 questions are used only for development. A question-only Qwen3-Embedding-4B top-60 retriever and a frozen Qwen3-VL-4B text reranker select three pre-query captions. The A payload contains those same released timestamped visual–audio captions and two sampled frames from each interval. P selects disjoint low-relevance, same-date caption rows matched on word count and uses the same six-frame budget; U contains no source payload. The need-audio annotation is retained only for predeclared descriptive stratification and never routes the payload. A development evaluator-only audit records target-time overlap in 15/50 target selections and 1/50 P selections, but gold time is not inspected for final selection, routing, eligibility, or progression. Development A0/U0/P0 is 60/52/48%, passing the frozen $\geq$ 4pp point-estimate progression rule for A0-U0 and A0-P0.

The disjoint final freezes dataset-order IDs 196–245 before selector generation or prediction. It contains 50 questions and 300 complete cells with no inference errors. Two audits verify exact IDs and six-cell completeness, A/P delivery and U blocking, three captions and six frames in every A/P cell, no source payload in U, target/P disjointness, pre-query timing, matching frame budgets, no retrieval in memory-off, and identical retrieval count and reasoning history in A1/U1/P1. The result artifact SHA-256 begins `6bc42df5`; deterministic recomputation produces a byte-identical summary with SHA-256 `331aeee2`. Memory changes three failures to successes and four successes to failures in A; the corresponding U counts are six and seven, and P counts are nine and five. Table 58 reports the locked estimands. Question-type and need-audio breakdowns are descriptive only. The intervals do not model dependence among questions sharing events in the same life record.

H Cross-Study Summaries and Evaluation Implications

Status. Descriptive and post-hoc summaries across studies. They combine or compare primaries estimated under different scorers and do not add confirmatory evidence.

H.1 Descriptive Offset Shares and a Post-hoc Pooled Interaction

Offset shares. Since $\mathrm{MIV} = L_0 - L_1$, the ratio $\mathrm{MIV}/L_0 = 1 - L_1/L_0$ is the fraction of the no-memory accuracy loss from restricted access that memory offsets. Values above 100% mean that memory-on accuracy rises under restriction ($L_1 < 0$). The ratio shares cells with MIV and is undefined when L_0 is near zero, so we report it only as a descriptive scale for the interactions in Table 59. For the joint cohort, a group bootstrap of the ratio of means (20,000 draws, seed 20260923) gives 57.4% $[9.1\%, 92.0\%]$ for NOTES and 48.9% $[-7.4\%, 82.2\%]$ for FLAT.

Table 59: No-memory access loss L_0, interaction and offset share MIV/L_0. Descriptive; intervals for the joint cohort are in the text.

Study (scorer)	Memory	L_0 (pp)	MIV (pp)	Offset share
FRAMES held-out, supplier (corrected substring)	FACT	9.00	+10.43	116%
Independent loss (strict EM)	FLAT	4.08	+3.17	78%
Independent loss (strict EM)	Mem0	4.08	+5.33	131%
Joint, random half (strict EM)	NOTES	5.22	+3.00	57%
Joint, random half (strict EM)	FLAT	5.22	+2.56	49%
LongMemEval, random half (Llama judge)	FLAT	10.49	+8.33	79%
LongMemEval, random half (Llama judge)	Mem0	10.49	+8.96	85%
LongMemEval, 7-day (Llama judge)	FLAT	15.86	+13.44	85%
LongMemEval, 7-day (Llama judge)	Mem0	15.86	+15.86	100%

Pooled interaction. The independent-loss study and the joint follow-up use disjoint question cohorts, both assign document denial independently of memory content at nominal rate 0.5, and both predeclare a strict-EM primary: MIV for Mem0 ($+5.33$pp $[+2.13, +8.61]$) and for NOTES ($+3.00$pp $[+0.33, +5.74]$). A fixed-effect inverse-variance combination, with standard errors approximated as interval width divided by 3.92, gives $+3.96$pp $[+1.88, +6.03]$; the heterogeneity statistic is $Q = 1.17$ on one degree of freedom. The combination was specified after both results were known and is not a predeclared estimand. The studies differ in memory construction, reader and memory system, so the pooled value summarizes the direction and precision of the two primaries rather than estimating one common effect. Adding either LongMemEval random-loss interaction, which uses a semantic judge, gives $+4.28$pp $[+2.29, +6.28]$ for FLAT or $+4.35$pp $[+2.36, +6.35]$ for Mem0. The script `tables/pooled_and_offset.py` in the supplement recomputes the pooled value and the joint-cohort offset shares from the archived analysis files.

H.2 Screening Decisions Under an Illustrative Interval Rule

A common screening rule keeps a memory component only when its paired 95% interval for contribution excludes zero. Table 60 applies this rule separately to the intact-access and restricted-access contributions reported in the main studies, without new inference. A *do not keep* verdict means that the interval includes zero; it is not evidence that memory has no contribution. The rule is illustrative: we do not recommend it as a deployment policy, and the verdicts inherit each study's scorer, population and access intervention.

The verdict changes in eight of eleven study–memory rows. In each change, intact access leaves the contribution unresolved while restriction yields an interval above zero. Intact-access evaluation would therefore not certify these memories, even though the same frozen memory contributes measurably once the corresponding sources are restricted. This is a statement about what an intact-access ablation can support, not about which memory to deploy, which additionally depends on the probability of restricted access (Section 7).

Table 60: The "keep if the interval excludes zero" rule applied to intact-access and restricted-access contributions (pp). Supplier-source rows use corrected substring EM, independent-loss and joint rows strict EM, and LongMemEval rows the primary Llama judge on the 186-question stratum. The verdict changes in 8 of 11 rows; in no row does it change from keep to do not keep.

Study (restriction)	Memory	$C(A)$ [95% CI]	Rule	$C(R)$ or $C(U)$ [95% CI]	Rule
FRAMES held-out (supplier)	FACT	$+3.32\,[-1.42, +8.06]$	do not keep	$+13.74\,[+8.53, +19.43]$	keep
FRAMES dev (supplier)	FACT	$+5.05\,[+0.46, +10.09]$	keep	$+20.64\,[+14.68, +26.61]$	keep
M3DocVQA (supplier)	FACT	$-0.51\,[-5.13, +4.10]$	do not keep	$+20.00\,[+13.33, +26.67]$	keep
Independent loss	FLAT	$-0.75\,[-3.55, +2.02]$	do not keep	$+2.42\,[+0.08, +4.81]$	keep
Independent loss	Mem0	$-1.75\,[-5.06, +1.50]$	do not keep	$+3.58\,[+0.66, +6.67]$	keep
Joint (random half)	NOTES	$+3.33\,[-0.64, +7.09]$	do not keep	$+6.33\,[+2.51, +10.18]$	keep
Joint (random half)	FLAT	$+6.33\,[+2.67, +10.17]$	keep	$+8.89\,[+5.22, +12.71]$	keep
LongMemEval (random half)	FLAT	$-2.15\,[-9.41, +5.14]$	do not keep	$+6.18\,[+0.36, +11.98]$	keep
LongMemEval (random half)	Mem0	$-3.49\,[-10.27, +3.23]$	do not keep	$+5.47\,[-0.00, +11.08]$	do not keep
LongMemEval (7-day)	FLAT	$-2.15\,[-9.41, +5.14]$	do not keep	$+11.29\,[+3.76, +19.15]$	keep
LongMemEval (7-day)	Mem0	$-3.49\,[-10.27, +3.23]$	do not keep	$+12.37\,[+4.30, +20.43]$	keep

I Reproducibility and Intervention Audits

Status. Audits of determinism, intervention validity, populations, predeclared rules and artifacts that apply across studies.

I.1 Determinism Verification

Both conditions are evaluated under the same pinned local engine, and the estimates are conditional on that engine. Agreement with the archived reference trajectories, recorded under a different engine configuration, is 71.1% overall; this is a cross-engine comparison rather than a treatment-effect invariance check, and it does not determine the paired within-engine contrast. In the tested engine configuration, replay agreement depended on the prefix-caching setting, and disabling prefix caching restored exact agreement in the repeat controls, so the paired comparisons use the cache-disabled configuration. We did not trace the discrepancy to logits and do not claim that it occurs in every engine configuration. The in-run repeat control agreed on 73 of 73 sampled trajectories, a baseline rerun in a later session reproduced 505 of 505, and conditional resynthesis, which holds the recorded prefix fixed and generates only the synthesis call, reproduces the archived outcome on 97.6% of trajectories, against 71.1% for full replay. Per-agent replay diagnostics are in the supplementary artifact.

I.2 Intervention Audit

Table 64 audits the supplier-source intervention of Section 5.1. No blocked-source card reaches the reader, memory-on and memory-off packets match, and card budgets are matched across arms in all three cohorts. In 24–41% of trajectories the retriever still returns the committed value from a document that was not blocked, which is why we call the manipulated quantity source access rather than exhaustive unavailability.

A second intervention family. Five interventions inside the rendered evidence pack give smaller contrasts than source access, with adjacent gaps that do not exclude zero, so we treat them as intervention-validity probes. Their audit matters more than their estimates: card removal often leaves the value available, so a candidate-list manipulation can fail its suppression at the consumer. The detailed pack-level diagnostics are archived in the supplementary artifact.

The binary contrast extends to a graded response. Removing nested hash-selected fractions $q \in \{0, .25, .5, .75, 1\}$ of each question's supplier documents, each dose matched by non-supplier removals, raises observed mean contribution from 3.32 to 13.74pp, a pre-specified slope of $+10.62$pp per full dose $[+5.50, +16.11]$; no-memory accuracy falls 9.00pp across the range while memory accuracy moves $+1.42$pp, so the curve is preservation under progressive loss rather than a growing normal-access penalty. It reuses the analysed cohort and provides within-cohort dose-response evidence. It does not isolate internal evidence use or constitute an independent replication.

Table 61: Retrieval is re-run at synthesis in all three arms; only the corpus differs, and the card budget is identical by construction. The placebo blocks as many of the highest-ranked documents that are *not* the agent's sources. Each condition carries its own paired interval, not only their difference. In both tabulated cohorts the full-corpus contribution interval includes zero, whereas the source-blocked interval does not. The FRAMES development cohort (n=218), omitted for space, is the exception: its full-corpus interval also excludes zero. Its contributions are $+5.05$ $[+0.46, +10.09]$, $+20.64$ $[+14.68, +26.61]$ and $+5.96$ $[+1.38, +11.01]$, from cells $28.9/23.9$, $28.9/8.3$ and $28.9/22.9$.

arm	EM_{mem}	EM_{no}	contribution
held-out, n=211, disjoint questions			
full corpus	20.4	17.1	$+3.32$ $[-1.42, +8.06]$
original sources blocked	21.8	8.1	$+13.74$ $[+8.53, +19.43]$
highest-ranked non-source placebo	19.0	15.6	$+3.32$ $[-0.95, +7.58]$
M3DocVQA, n=195, second corpus			
full corpus	35.4	35.9	-0.51 $[-5.13, +4.10]$
original sources blocked	30.3	10.3	$+20.00$ $[+13.33, +26.67]$
highest-ranked non-source placebo	34.9	33.8	$+1.03$ $[-3.08, +5.13]$

A pre-treatment version of the same question. Counting, before any intervention, how many documents in the corpus carry the committed value gives a covariate the treatment cannot have caused. The effect is larger where that count is small: below the median the source-access interaction is $+22.43$pp $[+14.02, +31.78]$ against $+10.26$pp $[+1.28, +19.23]$ above it, a difference of $+12.17$pp $[+3.76, +21.52]$; held out the ordering repeats at $+12.50$pp against $+8.14$pp but the difference does not clear zero ($+4.36$pp $[-3.59, +12.32]$). Blocking matters most where the fact had few independent sources to begin with. This is an exploratory pre-treatment analysis: the interaction clears zero on development and only repeats in direction held out, and source multiplicity may also proxy source importance, so we read it as supportive heterogeneity rather than a decisive mechanism test.

Nested source-loss dose response. For each held-out question we assign each supplier document a deterministic SHA-256-derived score in $[0, 1)$ from its question and document IDs, then block it at nominal dose q when that score is below q (with all suppliers blocked at $q = 1$). This produces nested integer sets; because most questions have few suppliers, adjacent nominal doses can select the same set. For the control, unique non-supplier documents are taken in their first available-retrieval rank order and paired to the lexically ordered suppliers; the partner of each selected supplier is then blocked. Thus P40 uses the same ranked non-supplier pool as the binary placebo but a source-hash-induced nested subset at intermediate doses.

All nine access arms (5 target and 4 placebo) and both memory cells were newly executed for each of the 211 existing questions: 3,798 scored outputs. No P20 prediction is reused. The newly executed $q = 0$ and $q = 1$ target answers match their P20 counterparts exactly, while the three intermediate target and placebo doses are new interventions. The paired bootstrap resamples each question's complete 18-cell vector. All records pass blocked-document delivery, packet identity, matched-count, nesting, and endpoint-reproduction checks.

Table 62: FRAMES held-out dose response in percentage points. The pre-specified linear slope of $C(q)$ is $+10.62$pp per full nominal dose $[+5.50, +16.11]$. This intervention-based follow-up reuses the binary-study cohort; it does not isolate internal evidence use.

nominal q	realised blocked	$C(q)$	MIV vs. q=0	MIV_{src}
0	0.00%	3.32	0.00	0.00
.25	26.47%	6.16	2.84	2.37
.50	48.41%	9.00	5.69	5.21
.75	71.98%	11.85	8.53	9.00
1	100.00%	13.74	10.43	10.43

What the source placebo does and does not match. The placebo blocks the highest-ranked non-source documents the full-corpus arm retrieved, taken in rank order; it is not drawn from each source's own rank bin. Measured, the two sets sit close at the median and diverge in the tail: source

documents reach a best rank of 2 (median) with a 90th percentile of 53 on development and 61 held out, against 4 and 12/11 for the placebo, and 74–75% of matched pairs fall within ten ranks. The divergence runs in the conservative direction: the placebo removes documents the retriever ranks at least as highly, on average, as the sources, so the contrast is not explained by having blocked less retrievable material. It does not rule out a difference in semantic relevance that retrieval rank does not capture.

Every condition in §I.2 and §5.1 is audited against the context string the reader received, not against the candidate list the packer was handed. Per row and per cell we record whether the committed value survives, whether the memory and no-memory cells saw byte-identical packets, the card and document counts, the retrieval-rank distribution, and the number of removals or redactions applied.

Table 63: Landing rate is the percentage of trajectories in which the committed value is absent from the rendered context (development / held-out). C0, C3 and C4 do not attempt to remove it, so their few percent are trajectories whose value was never in the pack. Packet match is the percentage of rows where both cells saw the same context; it is 100 everywhere.

condition	landed	packet match	cards	docs	edits
C0 intact	3.4 / 4.9	100	18.6 / 18.8	10.1 / 10.2	0
C1 supplier cards removed	20.3 / 26.9	100	18.6 / 18.8	11.3 / 11.4	3
C2 value redacted	**100.0 / 99.6**	100	18.6 / 18.8	10.1 / 10.2	10 / 12
C3 same-card sham	4.6 / 6.1	100	18.6 / 18.8	10.1 / 10.2	10 / 12
C4 matched-card placebo	3.8 / 6.9	100	18.6 / 18.8	11.0 / 11.2	3

Table 64: Audit of the source-access intervention. "Value returned from another source" is the share of trajectories in which the retriever found the committed value in a document other than the ones blocked; it is the reason we call the manipulated quantity source access rather than exhaustive unavailability. Both MIV estimands are reported per reader; percentile cluster-bootstrap intervals on the held-out sample are $[5.21, 16.11]$ and $[4.74, 16.11]$ for the primary reader and $[4.27, 14.69]$ and $[5.21, 16.59]$ at 32B. Other readers are compared in Table 28.

	FRAMES dev	FRAMES held-out	M3DocVQA
eligible cohort	218	211	195
blocked-source cards delivered	0	0	0
memory/no-memory packet mismatch	0	0	0
card budget matched across arms	yes	yes	yes
value returned from another source	40.8%	26.5%	24.1%
MIV / MIV_{src}, primary (8B)	+15.60 / +14.68	+10.43 / +10.43	+20.51 / +18.97
same family at 32B	–	+9.48 / +10.90	–

Two consequences. First, the card-removal manipulation in this harness (removing every card that supplies the value) leaves the value in the rendered pack four times in five, because the packer expands past the cards it is given to their siblings under the same parent. Any effect measured that way is the effect of removing value-supplying cards together with whatever context they carried, which is a real intervention but not the one the text claims. Second, redaction has to touch four surfaces to land: chunk bodies, the card inline-text fallback used when no chunk fits, the ancestry header built from the document title and section path, and the unit identifier, which is derived from Wikipedia titles and therefore often spells the value. Masking chunk text alone removes 17.5% of the value's occurrences.

One row of the held-out sample did not converge; it is enumerated and excluded by the frozen rule, and the estimates change in the second decimal.

I.3 Predeclared Decision Rules

The source-access cohort and eligibility rules below were fixed before the runs they judge.

Source-access cohort. Fixed before any outcome was generated: a sample is admissible if at least 200 trajectories yield all three arms at a matched card budget, 300 preferred. FRAMES gives 218

and 211 and is confirmatory. M3DocVQA gives 195; it misses the minimum by five and is reported as supporting evidence, not as a replication. We did not add trajectories to reach the threshold after seeing the estimate.

Corpus admissibility. A corpus is usable for the source-access design only if blocking the agent's sources is narrower than deleting the task. QASPER fails this: its questions concern a single paper, so the source-document count is one for every trajectory, and blocking it removes the only document that could answer. We report the preflight and exclude the corpus rather than take the large effect it would produce.

FACT typing on M3DocVQA. That corpus's trajectory records predate the `memory_record` field. Values are typed with `public_memory_record`, the rule frozen in the preprocessing script, and the choice was made before outcomes were generated. It decides the eligible cohort, so we record it here rather than as an implementation detail: 58 of 300 trajectories are excluded for carrying no FACT write under this rule.

The source-access eligibility funnel. FACT typing is the first of three sequential filters, not the only one, so Table 65 gives the whole descent from the 300 trajectories of each cohort to the eligible cohorts of 218, 211 and 195. The filters are applied in the order shown and each is evaluated only on what survives the previous one, so the counts partition the drops exactly and are not overlapping conditions. Two further tests (an empty union pack, and a union containing no document that carries the value) drop zero trajectories on all three cohorts and are omitted from the table. The largest single cost is the placebo: a trajectory is dropped when the available arm did not retrieve enough non-source documents to build a rank-matched control blocking the same number of documents at the same card budget. Dropping the placebo arm would leave the first two filters only, raising eligibility to 261 and 245 on the two FRAMES cohorts and 231 on M3DocVQA; we kept the control and reported the smaller cohorts, which is why M3DocVQA misses the predeclared minimum by five. The counts are read from `reports/P1_A_preflight*.txt` in the supplement, which also carries the QASPER preflight behind that corpus's exclusion.

Table 65: Source-access eligibility, from scanned trajectories to the cohorts reported throughout. The second filter removes trajectories with nothing to suppress; the third removes those where the available, source-blocked and rank-matched placebo arms cannot all be built at one card budget.

stage, applied in order	FRAMES dev	FRAMES held-out	M3DocVQA
trajectories scanned	300	300	300
− no FACT write	−25	−26	−58
− no committed value supported in the pack	−14	−29	−11
− an arm short of the matched card budget	−43	−34	−36
eligible cohort	218	211	195

I.4 Adapter and Artifact Reference

Adapter contrasts. The document studies use the six cells $A/U/P\times$memory with questions weighted equally, P blocking matched non-supplier documents; the controlled WebShop study uses $A/U\times$memory on 20 counterfactual pairs, estimating MIV but not $C(P)$. We resample document questions and WebShop pairs as clusters. Separate fixed-use follow-ups compare memory-enabled and memory-disabled execution under one regime, giving a direct utility gain G rather than an interaction. For implementation transfer, every system receives the same ordered records, is frozen before the branch, and shares source sets, question-only top-five retrieval, the 1,024-token cap, delivery envelope, reader and endpoint; only retrieved memory text reaches the consumer, so the estimand excludes the systems' native acquisition policies. Full adapter contracts are in Appendices F and G.

Artifact registry and auxiliary studies. Appendix D records the model revision, decoding and replay contract, and Appendix I.1 reports determinism checks. Appendix I.3 defines the source-access cohort and eligibility rules. Appendix I.2 reports, for every condition, whether the intervention

was present in the context the reader received, and both cells' packet identity. Appendix F separately records the Shopping population, configuration, six-cell endpoints, and branch-isolation audit. Appendix G records the Video acquisition and source-selection contract, gold isolation, first-stage delivery, six-cell results, and token audit. Appendices F.3 and G.1 record the separately frozen direct-utility Shopping and EgoLifeQA studies, including the Video precision-extension and error-only repair sequence. Appendices F.4 and G.2 record the disjoint source-access follow-ups, including the six-cell audiovisual repair, and their delivery audits. Appendix D.6 records the public implementations, common projection contract, absolute cells, paired system contrasts, and payload audit. A study-level registry records the raw-output and gold-manifest hashes, scoring implementation and configuration, changed-cell counts, and table generator for each analysis family. Primary document, common-record transfer, donor-pool transfer, write census, activation, and predictor analyses are listed separately. The joint document follow-up and its execution amendments are recorded in Appendix B.1. The 32B and Llama six-cell outcomes and the graded dose-response outcomes were re-scored with the corrected substring rule, with no score changes. Archived predictor analyses retain their legacy training, evaluation, and selection labels. Corrected-scoring results and archived legacy analyses are explicitly distinguished; predictor training, evaluation, and selection labels are recorded separately. WebShop and Video retain their study-specific endpoints. Appendix D.2 records the scoring specification, the legacy-to-corrected score changes, the alternative scorers, and the disagreement classification. The WebShop pair-to-history-to-bundle map and its dependency structure, and the coding of the related studies, are provided with the artifact. The supplementary artifact provides the scripts, frozen manifests, and per-intervention outputs needed to recompute the reported document, shopping, and video contrasts. Where a family was re-scored, the artifact also carries the reader answers and gold that the refusal-correction audit used; the FRAMES common-record cohort is covered through its retained shard files rather than a single consolidated result file.

J Literature Audit and Axis Placement

J.1 Coding of the Closest Designs

Each study below was read in full text rather than from its abstract; per-axis answers and the section or table each rests on are in the revision record.

Table 66: What the six closely related studies report, coded from their full texts. A baseline that removes a dedicated memory module is distinguished from one that removes the past information itself: in all three memory benchmarks the history remains fully available to the consumer. *not reported* means the full text we read does not describe it, not that the design is absent from the wider literature. Evidence locators are in the revision record's `related_work_comparison.tsv`.

study	baseline or ablation	mem. removed	history/source retained	source interv.	mem. × access
Tan et al. (2026)	ranked subset of stored memories	selected entries	not stated	not reported	not reported
Shah (2026)	five do-operations over steps	n/a	not stated	not reported	not reported
Zhang (2026)	sampled alternative actions	n/a	not stated	not reported	not reported
Chhikara et al. (2025)	full-context processing	yes	entire conversation in context	not reported	not reported
Xu et al. (2025)	LoCoMo baseline	yes	complete preceding conversation	not reported	not reported
He et al. (2026)	long-context agents	yes	verbatim interaction history	not reported	not reported
MOA (ours)	memory-use path off	yes	crossed: A / U / P	yes	yes

J.2 Coding of Agent-Memory Evaluations

Sample and fields. We coded a purposive sample of agent-memory systems and benchmarks, including those the paper builds on: 16 new papers plus the three memory evaluations coded above. For each paper, the coding records three fields from the full text: (i) whether a comparison condition without the evaluated memory keeps the original information the memory would be built from, in context or retrievable; (ii) whether source access is varied while the acquired memory is held fixed;

and (iii) whether memory's contribution is reported separately under two or more access conditions. A field that the text does not describe is coded *not stated*, not inferred. The protocol was fixed before coding.

Procedure and agreement. Each of the 16 new papers was coded twice, in independent AI-assisted passes that read the full arXiv text and recorded a locator and a verbatim quote of at most 25 words for every code; the quotes were string-checked against the text. Raw agreement was 15/16 for field (ii) and 14/16 for field (iii); after normalizing label spellings, agreement on field (i) was 15/16. Disagreements were adjudicated against the full text, and an author verified every final code against the full text. Two were substantive. ExpeL's Figure 5 crosses extracted insights with retrieval of successful trajectories, but both belong to its experience memory and the insight set is not stated to be fixed, so we code field (ii) as partial and field (iii) as no. Reflexion reports gains under chain-of-thought, ground-truth-context and ReAct settings, which change the agent rather than source access with memory fixed, so field (iii) is no. We also overrode one agreed code. Both passes coded ReadAgent's field (ii) as yes, but its Table 1 varies a maximum look-up budget of one to six raw pages while every original page remains stored and selectable. MOA treats budget changes as configuration contrasts rather than source restriction (Section 3), so we code it no. The protocol, both passes, quotes and adjudications are in the supplement.

Table 67: Full-text coding of 19 agent-memory evaluations. Field (i) is *yes* if at least one comparison condition without the evaluated memory keeps the original information in context or retrievable. *No condition* means no quantitative without-memory condition is reported. Counts: (i) 15 of 19; (ii) none, with one partial; (iii) 0 of 19.

System or benchmark	(i) original info kept	(ii) access varied, memory fixed	(iii) memory × access
ExpeL (Zhao et al., 2023)	no	partial	no
Generative Agents (Park et al., 2023)	yes	no	no
MemGPT (Packer et al., 2023)	yes	no	no
MemoryBank (Zhong et al., 2023)	no condition	no	no
Reflexion (Shinn et al., 2023)	yes	no	no
SCM (Wang et al., 2023)	yes	no	no
Think-in-Memory (Liu et al., 2023)	yes	no	no
Agent Workflow Memory (Wang et al., 2024)	yes	no	no
LoCoMo (Maharana et al., 2024)	yes	no	no
LongMemEval (Wu et al., 2025)	yes	no	no
ReadAgent (Lee et al., 2024a)	yes	no	no
A-MEM (Xu et al., 2025)	yes	no	no
Experience-following (Xiong et al., 2025)	no	no	no
MIRIX (Wang & Chen, 2025)	yes	no	no
Mem0 (Chhikara et al., 2025)	yes	no	no
MemoryAgentBench (Hu et al., 2025)	yes	no	no
MemoryOS (Kang et al., 2025)	not stated	no	no
Zep (Rasmussen et al., 2025)	yes	no	no
MemoryArena (He et al., 2026)	yes	no	no
MOA (ours)	yes	yes	yes

These codes describe what each paper reports, not whether its baselines are appropriate. Keeping the original information available is a meaningful baseline for the access condition each paper studies; the coding shows only that contribution under restricted access is not what these comparisons measure.

Additional source-oriented comparisons. The six-study coding above is a selected comparison, not an exhaustive account of work on source conditions. SMMBench (Chai et al., 2026), Sections 4.1, 4.3 and Appendix C.3, compares retrieval and memory baselines and relocates fixed gold evidence between concentrated and distributed sources. Its Appendix B.6 also masks individual evidence items to verify necessity during dataset construction. Yang et al. (2026), Table 1 and Sections 4–5, distinguish source-free, single-source and fusion baselines and diagnose source reachability separately from extraction and resolution. Their Section 4.2 crosses resolver (DSNBF/GPT) with input quality (direct-readout/extracted atoms) and reports a difference-of-differences attribution. These are substantive controls for evidence organization, sufficiency and component attribution. MOA instead estimates a memory-use contrast within each source-access regime and compares those contrasts after acquisition is frozen. We distinguish these reported designs without claiming that source manipulation or evidence ablation itself is new.

The same split appears in published numbers. Jiang et al. (2026) report gains over an empty-memory baseline on three interactive benchmarks and three QA and tool-use benchmarks, across three agent frameworks. Split by whether the stored content can be obtained again (a past trajectory or an acquired skill cannot; an answer in a searchable corpus can), the mean gain is $+15.1$pp against $+2.4$pp, a ratio of 6.2 that holds in each framework separately (5.6, 7.9, 5.5), with eight of the nine QA gains falling below the smallest interactive one. This ordering is consistent with the source-access account in numbers collected by a paper not testing for it. It is also only a correlation: task type stands in for re-obtainability, and we did not run our instrument on their tasks.

K Full Limitations and Archived Analyses

Measurement validity changes the conclusion. Supplier-card removal suppresses the value completely in only 20.3% and 26.9% of document cases (Table 63), so delivery, execution and determinism checks identify the effect rather than tidy it. Once the intervention is enforced, a wide or negative estimate is still a result: a positive MIV is an outcome of MOA, not a validity gate.

Implementation transfer is bounded. With acquired records and the consumer contract fixed, positive document interactions persist for the tested Mem0 path and for the shared A-MEM/FLAT projection; because A-MEM and FLAT deliver byte-identical payloads here, the latter identifies no contribution from A-MEM-specific structure. WebShop effects differ and EgoLifeQA estimates remain imprecise. Because A-MEM and FLAT deliver identical payloads and acquisition is held common, this neither isolates representation nor compares native agents. WebShop supports the matched direct-utility contributions and the between-projection source-access contrast. Individual source-access projection contributions remain withdrawn. Appendix F.5 records the execution mismatch and the missing equivalence check; Table 52 reports the replacement direct comparison.

Aggregate interaction is not a gate. MIV averages over a declared population and identifies neither which memory to store nor when to inject it.

LongMemEval scope. The archived corpus differs from the official cleaned release in 454 haystacks. Questions share distractor pools, sessions are limited to $1{,}800$ characters, and the final study uses one writer build per question. These constraints limit generalization. The development repeatability check does not establish universal determinism; the full contract is in Appendix E.1.

Scope and interventions. All MOA adapters impose controlled source restrictions rather than measure naturally occurring outages, access-control changes or deployment cost, and the content-independent study is a simulated denial process, not a claim about real outages. The WebShop construction deliberately makes prior purchase determine the correct action and its MIV interval remains broad; earlier Shopping estimates are preserved separately (Appendix F). Document sources are reconstructed lexically rather than by complete semantic provenance, and other documents reintroduce the value in 24.1–40.8% of samples. EgoLifeQA estimates condition on one life record, model no shared-event dependence and show no person-level generalization. Public-system transfer reuses existing cohorts under one projection and reader; it broadens implementation coverage, not population coverage or native end-to-end generality.

Archived analyses. Superseded, replaced and withdrawn estimates are listed in Table 4: the original LongMemEval construction (Appendix E.1), historical reader configurations (Appendix D.4) and WebShop implementation-transfer contributions (Appendix F.5). None supports a main-text claim.