Construction before search: a pre-registered test of what an LLM memory builder keeps

Luiz Antonio Busnello

Version 1.1 · doi:10.5281/zenodo.23966758

Version 1.1 (2026-11-04), doi:10.5281/zenodo.23966758. Not peer reviewed. Every number in this text comes from a file in the repository directory `paper3-construction/` or in the data archives of the records in §9; §10 lists the sources.

Abstract

Several recent studies compare memory extracted by a language model with retrieval or search over the raw conversation, and in most of their settings the raw record wins. Their conclusions carry conditions: one limits its recommendation to the extraction designs it tested, one finds extraction more accurate at generous context budgets, and one finds designed memory helping only a weak reader. One of them, *Fidelity before Structure* (arXiv 2601.00821), also varies the extraction itself, down to a near-verbatim, full-coverage design that still trails verbatim chunks, and predicts that better extraction prompts can narrow the gap but cannot reach what a verbatim store retains. None of them varies how much of the conversation the builder is shown.

We pre-registered a test of two repairs to the construction step: how much of the conversation the builder is shown (the window), and the instructions it extracts under (the criterion). The first instrument failed its own registered floor. With the full window, the pipeline answered 21% of LongMemEval questions against a required 50%, so its primary contrast was not interpreted. Two more amendments followed, each fixed before its calibration calls. The first of them tried four readers, and none passed. An exploratory run then found the larger loss between the raw evidence and the units distilled from it. Units distilled from the evidence sessions scored 0.475, and the raw turns of the same sessions scored 0.830. A revised criterion, C2, asks the builder to keep facts said in passing, the event's own time and what the assistant said. It passed calibration on the development instances.

The confirmatory run used 242 instances with no prior exposure, defined by provenance and not by question id alone. All four registered controls passed. Repairing both factors against repairing neither raised accuracy by +66.5 pp [+60.6, +72.5] under judge 1 and +66.9 pp under a second judge family. At the truncated window, 120 of the 242 instances kept no annotated evidence session, and the truncated cells answered none of them. That pattern is consistent with evidence availability as a large part of the window contrast, though the design does not estimate a mediated share. The criterion contrast at the full window is +24.8 to +27.3 pp across judges. A length-matched arm rules out its own form of padding as the explanation. The result confirms a revised intervention, not the original hypotheses. A raw-evidence arm, registered as a further amendment before it ran, used the same 242. C2 units of the evidence sessions scored below the raw turns of those sessions. Its two judges fell on opposite sides of the decision rule (−5.4 pp [−10.3, −0.8] and −5.0 pp [−9.9, 0.0]), so a last amendment, written after that split and fixed before any call of the new judges, added two judges and a committee label. Under it the units lose to raw evidence by −5.8 pp [−10.4, −1.2], and non-inferiority at 10 pp is not shown. Both arms are restricted to the annotated evidence and differ in length and order as well as content, so the arm compares two served materials and does not test retention.

1. Introduction

An agent that remembers usually does it in two steps. A builder reads past conversations and writes memory units, often facts or lessons with an importance score. A reader later answers from the units that are served to it. When such systems are compared against retrieval over the raw transcript, the raw transcript often wins. Three recent studies report this in different settings, each with a condition attached. 2601.00821 recommends that structure augment verbatim text rather than replace it, for the extraction designs it tested. 2609.34227 finds raw turns non-inferior to an extraction memory at a tight budget, and extraction systems more accurate at generous ones. 2610.29823 finds that designed memory beats full context only with the weakest reader it tried. A fourth study, 2609.34160, works on the construction step itself: processing the same records in a different order changes what memory keeps.

2609.34227 and 2610.29823 keep the extractor at its default; 2610.29823 says so directly: Mem0 and Zep "run with their default extraction and retrieval, and other settings may do better". 2601.00821 goes further. Besides its main ablation, it sweeps the compression rate of a real extractor and tests a near-verbatim extraction with full coverage and no salience filter, which still saturates about 11 points below verbatim chunks on LoCoMo; it locates the mechanism in lossy distillation. None of the three varies how much of the conversation the builder is shown. 2601.00821 also states, as a prediction rather than a result, that "better extraction prompts — explicit date/time targeting, temporal attribute modeling — could shrink this class, but cannot reach the unconditional retention of a verbatim store."

That prediction can be tested, and our own system gave us a reason to test it. In a production multi-agent memory system (§3), the builder skips sessions it has already logged, so in the largest sessions it saw a median of 24% of the prose. Its criterion also turned specific facts into general lessons. In an exploratory, unregistered measurement, repairing either factor alone moved little, and repairing both moved a lot. Our contribution is narrower than a claim about extraction in general: a registered test, in one system, of the construction window crossed with one concrete revision of the extraction instructions.

This paper reports a pre-registered test of that claim on a public benchmark, LongMemEval `s_cleaned`. It reports the path to the result as well: the first instrument failed its own floor, the first amendment selected no reader, and the criterion that passed was written after looking at failures on development data. Every registration record except v1.4 and v1.5 was deposited before the runs it governs; those two were fixed by pushed tags and deposited afterwards (§5), so a reader can check what was fixed when.

What this paper claims. On 242 LongMemEval instances the test had never touched, a full construction window plus the revised criterion C2 raised answer accuracy from 0.153 to 0.818, against the deployed criterion at a truncated window, with a fixed reader. All four registered controls passed. The effect is not explained by the length-matched padding arm.

What it does not claim.

- It does not show that C2 beats the first repaired criterion, C1, on new data, because C1 did not run in the confirmatory stage.
- It does not confirm the interaction hypothesis H1 or the identifier-survival hypothesis H3 as first registered.
- It does not settle 2601.00821's prediction. On the 242 a four-judge committee put C2 units 5.8 pp below the raw evidence of the same sessions (§6.7), but that comparison holds retrieval perfect and does not measure retention.
- It does not show that the gain holds at an equal context budget or under selective retrieval, because every unit of a cell is served (§8).
- It makes no comparison with the LongMemEval leaderboard, because our judges are not the benchmark's judge.

2. Related work

study	what it varies	finding relevant here	what it holds fixed
2601.00821 *Fidelity before Structure*	memory representation, with six confound controls; a real extractor's compression rate (App. F); a near-verbatim, full-coverage extraction (App. H)	verbatim chunks beat lossy artifact extraction; of diagnosable failures, 69.0% are facts never written down and 31.0% retrieval misses; the near-verbatim extraction stays about 11 points below chunks	the retrieval and answer stack; how much of the conversation the extractor sees
2609.34227 *When Does Selection Replace Extraction?*	context budget; pre-registered, blind human grading	raw turns non-inferior at a tight budget; reranking adds 17.4 points there and 1.5 at a generous budget, where extraction systems are more accurate	the extractor
2610.29823 *MemoryTax*	reader tier, cost	designed memory beats full context only with the weakest reader; a stronger reader does not recover what a weaker writer left out; build cost can exceed the rollout it saves	default Mem0 and Zep extraction
2609.34160 *RoutePrism*	construction order; memory policy	restoring one displaced record recovers more than 60 pp; memory policies leave distinct failure signatures	record content, timestamps, policy and answer model, within each order comparison

The studies above already show, in their settings, that extracted memory can lose to raw retrieval and that the writer limits the reader. What we add is narrower: a registered test of a change to the builder's configuration, with controls that can fail. 2609.34227 is the methodological model, with a registered plan, a primary test and blind human grading. We have the first two and not yet the third (§8).

`gbrain-evals` (not peer reviewed) uses the same `longmemeval_s_cleaned.json` file, identified by sha256. It repaired a writer's triage and quoting and measured page recall, not downstream answer accuracy. Two of its observations shaped our controls: LongMemEval's official judge accepts a "not available" answer to an answerable question 19 times in 100, and preference questions defeat readers told to abstain (§8).

3. Motivation: an exploratory production measurement

This section is not evidence for the confirmatory claim. It explains why the factors are the window and the criterion.

The system is a hybrid memory (BM25, dense embeddings, rank fusion) for a group of agents. A builder, `session-distill`, cuts each session transcript into blocks of 2,000 characters and asks `gemini-2.5-flash-lite` for at most ten units per block, each with a category and an importance. Units below 0.5 are dropped. The builder skips any session it has already logged. Long sessions keep growing after their first distillation, so in the 200k to 1M character band the builder saw a median of 24.0% of the session's prose. That median rests on six sessions ranging from 0.020 to 0.894.

On 99 questions over that corpus (2026-10-27, one judge family, not pre-registered):

arm	accuracy
oracle (raw window around the evidence)	88.9%
deployed builder	42.4%
window repaired alone	46.5%
criterion repaired alone (identifiers kept verbatim)	46.5%
both repaired	63.6%

Each repair alone added 4.1 pp. Both together added 21.2 pp, of which 13.0 pp is the additive interaction. At an equal context budget the deployed memory was 21.2 pp below BM25 over the raw text. The repaired memory, at its own budget of 2.41× the deployed context, was level with it.

The limits are serious. Only 12 sessions had any variance in the window, 8 of them were used to form the hypothesis, the corpus is private, and one judge family graded it. A second judge family on the same items moved the criterion effect at the full window from +8.1 pp to +21.2 pp. The disagreement between judges was larger than the effect one of them reported. That is why the registered analysis reports criterion effects as a bracket across judge families (§4.7).

4. Design

4.1 Corpus

LongMemEval `s_cleaned` (revision `98d7416c`, MIT licence) has 500 instances. Each instance is a haystack of 38 to 62 chat sessions (455k to 514k characters) and one question, with the evidence sessions annotated. The 30 abstention instances are excluded, and 8 more were spent on a declared smoke run, which leaves 462. LoCoMo was rejected because its 1,986 questions sit on 10 conversations, and the conversation is the unit of inference.

Instances share distractor sessions: 20.5% of session ids appear in more than one haystack. Only 5 pairs share an evidence session. Standard errors are therefore cluster-robust over answer-session groups.

4.2 Factors and cells

Each instance passes through every cell, so all contrasts are paired.

	criterion C0 (deployed)	criterion C2 (revised)
window 0.240	cell 00	cell 01
full window	cell 10	cell 11

- **Window:** at level 0 the builder sees the most recent sessions whose cumulative prose reaches 0.240 of the haystack's characters. The session that crosses the threshold is included whole. The realised dose in cell 00 had median 0.251 [0.240, 0.274]. At level 1 it sees all sessions. The cut approximates production's window by its median seen-fraction. It does not reproduce production's mechanism, which is a prune crossed with a lookback and a per-run ceiling.
- **Criterion:** C0 is the deployed `DISTILL_SYSTEM`, verbatim. C1, the repair registered in v1.0, adds one paragraph requiring identifiers verbatim. C2 adds a second paragraph to C1 (§5.4). The prompts are in Portuguese, the deployed system's language.

Two dose cells (W = 0.020 and W = 0.894, the band's observed extremes) run on a seeded subsample of 60. A length-matched cell serves cell 00's material, padded to cell 11's length with raw text from sessions that hold no answer.

Each block is distilled once per criterion, and every window that contains it reuses the same units. The truncated memory is therefore an exact subset of the full one, and a difference between windows comes from which sessions the builder saw. Sampling noise in the extractor cancels out of that contrast.

4.3 Serving and reading

All units of a cell are served, sorted by importance, each as `[session date] [category] text`. The question carries LongMemEval's own question date. The reader in the confirmatory run is R3: `gemini-3.8-flash` with prompt B and thinking set to low. Prompt B tells the reader the material is dated memories, asks it to count, compare and do date arithmetic where the question needs it, and reserves the abstention answer `NAO SEI` for material in which nothing is relevant.

4.4 Judges

Judge 1 is `gpt-6-luna` through a structured decision API with no free-text field. An answer is correct when the judge's probability for "contains the reference fact" exceeds 0.5. Judge 2 is `gemini-2.5-flash-lite` with a fixed prompt, and it grades every row. Before any run, a preflight checks that judge 1 accepts none of 16 abstentions offered against real references. The reader and judge 2 are both Gemini models, so the judge bracket is read on contrasts only (§8).

4.5 Primary hypothesis and analysis

H1′: E[Y(full window, C2, R3) − Y(0.240 window, C0, R3)] > 0 on the 242.

This is the total effect of repairing both factors against repairing neither. The estimate is a paired difference in accuracy with a cluster-robust standard error. The test is two-sided at $\alpha = 0.05$, in one stage, with no interim look. The interaction is a registered secondary, tested against a serial null, not against zero. A model where the window decides whether the evidence arrives and the criterion decides whether it is used produces a positive additive interaction with no joint mechanism. The serial null predicts Y11 = Y01 · Y10 / Y00.

4.6 Controls

control	construction	passes when
answerable floor	cell 11 on every instance	acc ≥ 0.50
positive	full window minus every annotated answer session, 40 instances	acc ≤ 0.10
negative	the 0.240 cut applied to non-answer sessions only, all answer sessions kept, 40 instances	$\lvert\text{acc} - \text{acc}(11)\rvert \leq 0.10$ and item-wise discordance ≤ 0.20
counterfactual	31 instances whose numeric answer was rewritten inside the evidence sessions (+3 for years, +7 otherwise)	in cell 11, answers with the original number $\leq 10\%$ and fewer than answers with the altered number

A pipeline that always abstains passes the positive and negative controls, and the floor catches it. A pipeline that reads the benchmark's annotations instead of the constructed memory can pass the first three. The counterfactual is the control that checks the answer came from the memory. If the floor fails, the run is reported as an instrument failure and the primary contrast is not interpreted.

4.7 Pre-specified reporting rules

- Criterion contrasts are reported as the bracket across the two judge families.
- An interval that contains zero is inconclusive. It is never read as a null or as equivalence.
- Served length is recorded per row. The primary is also estimated with length as a covariate.
- A missing annotated session does not prove that no usable information remained.

5. How the registration changed, and why

Each amendment was fixed by a git tag of this repository before the calls it governs. Where a DOI is given, a Zenodo deposit followed. The v1.3 calibration was fixed by tag on 2026-11-03, and the v1.3 deposit, which includes the calibration result, was published before the confirmatory run. v1.4 was fixed by its tag before its first call and deposited after the run, so for v1.4 the evidence of precedence is the push to the public repository. GitHub's activity log records the push of commit `bca912c`, which holds §15 and the arm's runner, at 21:55:05 UTC on 2026-11-04; the arm's first call started at 21:55:12 UTC. v1.5 was fixed the same way: the push of commit `977ec5f`, which holds §16 and its runner, is logged at 22:48:26 UTC, and the new judges' first call on the data started at 22:48:32 UTC. The tag's own date is set by the author and proves nothing alone.

version	fixed	DOI	what it changed	outcome
v1.0	2026-10-27	10.5281/zenodo.23716613	the 2×2 factorial on LongMemEval with C1	not run as written
v1.1	2026-11-03	10.5281/zenodo.23914868	excluded the 30 abstention instances; pinned the seed, serving, controls and judges; declared a smoke run	stage 1: instrument failure
v1.2	2026-11-03	tag `paper3-prereg-v1.2-calibration`	the reader, calibrated over four candidates	no reader selected
v1.3	2026-11-04	10.5281/zenodo.23934489	criterion C2, reader R3, confirmatory sample of 242	controls pass, H1′ confirmed
result	2026-11-04	10.5281/zenodo.23949145	the confirmatory result as its own record	
v1.4	2026-11-04	tag `paper3-prereg-v1.4`; deposited after the run as 10.5281/zenodo.23964906	the raw-evidence arm on the 242, margin 10 pp (§6.6)	judges split on the decision rule
v1.5	2026-11-04	tag `paper3-prereg-v1.5`; deposited after the run as 10.5281/zenodo.23966699	two more judges on H4 and a committee label, at least 3 of 4 (§6.7)	committee: units lose to raw evidence

5.1 v1.0 to v1.1: a census error

v1.0 stated that the corpus held no abstention instances. The census script had looked for the `_abs` suffix in the question type, while the marker is in the question id. There are 30. The v1.1 amendment excluded them before any arm ran. It also pinned what v1.0 had left to the runner. A declared smoke run on 8 excluded instances checked the plumbing and was not evidence.

5.2 Stage 1: the floor failed

Stage 1 ran on 200 instances with C1 and the deployed reader (R0, `gemini-2.5-flash-lite`) and cost US$ 18.80. The answerable floor was $\mathrm{acc}(11) = 0.210$ against the required 0.50. The other three controls passed. Under the registered rule the run is an instrument failure. The printed primary of +14.5 pp is not a finding, and stage 2 was not run.

Two defects of the registration were exposed. The 0.50 floor had been justified by numbers from the production corpus of §3, not from LongMemEval. The smoke had passed the floor at exactly 4/8, a sample that cannot test a threshold.

5.3 v1.2: no reader passed

v1.2 calibrated four readers on the spent stage-1 instances, as a 2×2 of model and prompt. The pass bar was 0.55, set above the floor to absorb selection over four candidates.

reader	model	prompt	acc(11), judge 1
R0	gemini-2.5-flash-lite	A	0.210
R1	gemini-2.5-flash-lite	B	0.245
R2	gemini-3.8-flash	A	0.360
R3	gemini-3.8-flash	B	0.505

None passed, and v1.2 stopped as registered (US$ 19.03). The prompt helped the larger model (+14.5 pp) far more than the smaller one (+3.5 pp).

5.4 Exploratory: the loss sits between raw evidence and distilled units

An unregistered run on the same spent 200 asked how much of the gap was the reader and how much was the material (US$ 1.71). R3 read three materials:

material	mean characters	acc, judge 1
cell 11: all sessions as C1 units	~161,700	0.505
evidence sessions only, as C1 units	9,178	0.475
evidence sessions only, as raw turns	26,458	0.830

Removing roughly fifty distractor sessions made no detectable difference (12 against 6 discordant pairs, exact McNemar $p = 0.24$, which is not an equivalence result). Replacing the units with the raw turns they came from did (6 against 77, $p < 10^{-16}$). The two materials differ in length, in representation (scored units against chronological turns) and in order as well as in content, so the contrast measures the loss of the whole distill-and-serve step under this protocol, of which omitted facts are one part. The largest losses were in temporal reasoning (0.37 against 0.94) and single-session-assistant questions (0.46 against 1.00). With the weaker reader R0 the same comparison gave 0.205 and 0.340. A weak reader hides the construction loss, and that is why stage 1 seemed to point at the reader.

A cache-only diagnosis compared the 77 instances R3 answered from raw turns and missed from units with the 89 it answered from both. The answer-bearing block was rarely distilled to nothing (5 of 77). The losses were in what the units kept. A fact said in passing became a lesson: “the commute takes 45 minutes each way” became “the user retains more from audiobooks”. The date of the conversation replaced the time of the event. Items the assistant had listed were not kept.

C2 adds one paragraph to C1 that asks for exactly those three things. It wants concrete facts about the user’s life kept even when said in passing, as facts and not lessons, at importance of at least 0.5. It wants the event’s own time expression copied with the event. And it wants specific information the assistant gave and the user may ask back, such as lists, names and numbers.

5.5 v1.3: calibration, then a sample defined by provenance

v1.3 fixed the reader at R3 and replaced C1 with C2. Calibration on the spent 200 (pass bar 0.55) gave acc(11) = 0.770 and acc(pos) = 0.000. C2 was right on 56 instances that C1 under R3 missed and wrong on 3 that C1 got. C2 was written from these instances' failures, so 0.770 was declared likely optimistic.

Before deposit, two adversarial reviews from other model families read the amendment and both advised against depositing it as written. One found that "untouched" had been defined by question id only. Seventeen of the 262 instances left after stage 1 had run as counterfactual variants, and 3 more shared an answer session with a spent instance. The confirmatory sample became the remaining 242, the counterfactual control was reclassified as a regression control on known examples, and the hypothesis was renamed H1′ to say that it tests a revised intervention. A third review found two defects in the spending cap, which were fixed in the runner. v1.3 was deposited before the confirmatory run began.

6. Results

The confirmatory run produced 1,654 rows, all with a normal finish. Of 155,506 distinct blocks, 87 failed to distil and were served empty (0.056%). It ran as three processes on one output directory, resumed from its own cache after a rate limit and after the provider's prepaid credit ran out. It cost US$ 70.73 against a cap of US$ 110. No row was read for its outcome until every row existed. `lme-analyze.py runs/confirm-2026-11-04 --check` reproduces the analysis byte for byte.

6.1 Controls

control	result	required
answerable floor	acc(11) = 0.818 (n = 242)	≥ 0.50
positive	0.000 (n = 40)	≤ 0.10
negative	0.775 against 0.775, discordance 0.100	≤ 0.10, ≤ 0.20
counterfactual	original number 1, altered number 28, of 31	original ≤ 3.1, altered $>$ original

All four pass, so the primary is interpreted. The counterfactual instances had been exposed in earlier stages, so that control is a regression check on known examples. Three of the four controls test unseen data.

6.2 Primary

	estimate	95% interval
judge 1	**+66.5 pp**	[+60.6, +72.5]
judge 2	+66.9 pp	[+61.0, +72.9]

acc(00) = 0.153 and acc(11) = 0.818, with G = 241 answer-session clusters. The interval excludes zero, so H1′ is confirmed. The two judge families agree on 922 of 968 primary-cell judgments.

6.3 Contrasts behind the total

These are reported by the registered analysis, and none is a confirmatory test.

contrast	judge 1	judge 2
window, 11 − 01 (C2 held)	+59.5 [+53.1, +65.9]	+61.6 [+55.2, +67.9]
criterion, 11 − 10 (full window held)	+24.8 [+18.8, +30.8]	+27.3 [+21.1, +33.4]
criterion, 01 − 00 (0.240 window held)	+7.0 [+3.6, +10.4]	+5.4 [+2.1, +8.6]

Cell accuracies under judge 1 are 0.153 (00), 0.223 (01), 0.570 (10) and 0.818 (11). The contrasts are conditional and interact, so they are not shares of the total. In one order the total splits as +59.5 (window) plus +7.0 (criterion), and in the other as about +41.7 plus +24.8. The criterion contrast at the full window is reported as the bracket [+24.8 · +27.3] pp. It compares C2 with the deployed C0, not with C1.

Interaction. The additive interaction is +17.8 pp. Against the serial null it is −1.4 pp [−20.2, +11.8]. The interval contains zero, so the secondary is inconclusive. It licenses neither a null nor an equivalence claim.

Served length. Cell 00 served a mean of 36,199 characters and cell 11 served 231,416. The length-matched arm serves cell 00's material padded to cell 11's length. It scored +0.8 pp [−0.8, +2.4] above cell 00 and −65.7 pp below cell 11. The registered length adjustment regresses the paired accuracy difference on the mean-centred length difference and reports the intercept. On the full sample that intercept equals the unadjusted mean by construction, so its value (+66.5 pp) carries no evidence about length. The registered capacity reading asks whether the padded arm gains at least half of the total effect over cell 00, with an interval above zero, and it does not. That rules out this form of padding as the explanation. It does not rule out every explanation by length.

Dose. On the subsample of 60, Y(full, C2) − Y(dose, C0) was +83.3 pp at W = 0.020 and +35.0 pp at W = 0.894. The sign is the same across the observed band. The analysis output labels these levels `W=0.20` and `W=8.94`. That is a formatting defect in the frozen analysis script, and the cells it reads are `d020` and `d894`.

6.4 Where the window contrast comes from (descriptive)

The design predicted that the window contrast would be dominated by evidence availability (v1.1 §12.8). The table below splits the 242 by how many annotated evidence sessions the 0.240 cut kept in cell 00. Each instance gets one group, computed from its sessions and annotations under the reference cut before any model call, and the same group is used in all four cells, so the split depends on neither the extraction nor the answers. It describes how the contrasts vary across instances. It does not estimate how much of the window contrast is mediated by evidence availability, and it is not a registered analysis.

evidence kept by the cut	n	00	01	10	11
none	120	0.000	0.000	0.550	0.808
some	79	0.177	0.278	0.595	0.861
all	43	0.535	0.744	0.581	0.767

When the cut kept no evidence session, the truncated cells answered nothing under either criterion. This is consistent with the positive control, which removed every answer session from the full window, although the two interventions serve different material. When the cut kept all the evidence, the window added +4.7 pp under C0 and +2.3 pp under C2, and the criterion added +20.9 pp at the truncated window and +18.6 pp at the full one. The pattern is consistent with evidence availability being a large part of the window contrast on this corpus, with the criterion contrast present at both windows once the evidence reached the builder. The groups may also differ in recency, question type and difficulty, and the window changes distractors and length along with evidence. The "all kept" group has 43 instances, these figures carry no interval, and the small window difference in that group is not evidence that the two windows are equivalent.

6.5 By question type, and against calibration

question type	n	00	01	10	11
knowledge-update	30	0.333	0.467	0.733	0.900
multi-session	68	0.059	0.147	0.603	0.838
single-session-assistant	28	0.250	0.321	0.536	0.786
single-session-preference	16	0.062	0.062	0.312	0.438
single-session-user	32	0.281	0.312	0.781	0.906
temporal-reasoning	68	0.088	0.147	0.441	0.824

Abstention falls from 0.760 in cell 00 to 0.074 in cell 11. On the 242 unseen instances cell 11 scored 0.818, against 0.770 in calibration on the 200 spent ones. The drop that §5.5 expected did not appear. The two numbers come from different samples, so the comparison is descriptive.

6.6 The raw-evidence arm (v1.4)

Both arms of v1.4 serve only the instance's annotated evidence sessions, so both are given perfect retrieval. E-raw serves the raw turns, chronologically. E-dist(C2) serves the C2 units of the same sessions from the confirmatory cache, with nothing re-distilled, ordered by importance and then recency. The two materials therefore differ in length (13,115 against 28,153 characters on average), order and format as well as in what they keep, and H4 compares them as served packages. The new responses were not seen before §15 was fixed, but the 242 instances and their primary-cell results were. The instrument check passed (acc(E-raw) = 0.872 against the required 0.60). The completed run recorded US$ 2.06; a first attempt, stopped at its first judge call by a stale API key, made answer calls that were reused and not ledgered.

arm	judge 1	judge 2	mean characters
cell 11, all sessions, C2 units	0.818	0.826	231,416
E-dist(C2), evidence sessions, C2 units	0.818	0.835	13,115
E-raw, evidence sessions, raw turns	0.872	0.884	28,153

H4: E-dist(C2) − E-raw	estimate	95% interval	discordant pairs
judge 1	−5.4 pp	[−10.3, −0.8]	11 / 24
judge 2	−5.0 pp	[−9.9, 0.0]	12 / 24

Under judge 1 the upper bound is below zero, so C2 units lose to raw evidence, and the lower bound misses the −10 pp margin by 0.3 pp. Under judge 2 the lower bound clears the margin, so the units are non-inferior at 10 pp, and the upper bound is exactly zero (0.0 in the unrounded output, `raw-arm.json`), so a loss is not shown. §15.4 says the weaker claim stands when the judges disagree, but these two outcomes are not nested and the rule does not order them. We therefore make neither claim across judges from these two, and report both estimates and intervals. The bracket [−5.4 · −5.0] pp is variation between judges, not an uncertainty interval. §6.7 reads H4 under the rule v1.5 registered for this split.

Descriptively, cell 11 and E-dist(C2) score the same under judge 1 (0.0 pp [−3.7, +3.7], 10 against 10 discordant pairs; judge 2 −0.8 pp [−4.1, +2.5]), so the roughly fifty distractor sessions per instance cost nothing detectable under C2. §15.6 also asked for intervals on the by-type accuracies; the runner printed them without intervals, and with 16 to 68 instances per type the differences below are descriptions only. The gap sits in multi-session questions (0.853 against 0.956) and single-session-assistant questions (0.821 against 0.964). Temporal reasoning, where C1 units lost most in development (0.37 against 0.94), is within 3 pp (0.824 against 0.853).

6.7 H4 with four judges (v1.5)

The split in §6.6 was known when v1.5 was written. To keep the rule from being chosen after seeing the new judges, v1.5 fixed them, their gate and the aggregation before any of their calls on the arm's rows. Judges 3 (`glm-5.3`) and 4 (`gemini-4-argon`) graded the same 484 rows with judge 2's prompt; no answer was regenerated. Both passed the gate: each rejected all 16 smoke abstentions and accepted the reference answer on 8 of the 8 smoke instances. The primary of v1.5 is a committee label, correct when at least 3 of the 4 judges accept, with a 2-2 tie counted as not correct, read with the §15.4 table.

H4: E-dist(C2) − E-raw	E-dist(C2)	E-raw	estimate	95% interval	discordant pairs	§15.4 row
judge 1	0.818	0.872	−5.4 pp	[−10.3, −0.8]	11 / 24	loss
judge 2	0.835	0.884	−5.0 pp	[−9.9, 0.0]	12 / 24	non-inferior
judge 3	0.835	0.893	−5.8 pp	[−10.4, −1.2]	9 / 23	loss
judge 4	0.839	0.901	−6.2 pp	[−10.8, −1.7]	9 / 24	loss
committee, at least 3 of 4	0.839	0.897	−5.8 pp	[−10.4, −1.2]	10 / 24	loss

The committee's interval lies below zero, so under v1.5 the C2 units lose to the raw evidence of the same sessions, by about 6 pp. Its lower bound misses the −10 pp margin, so non-inferiority is not shown. Three of the four judges land in that row on their own; judge 2 is the only one that does not. The 3 tied rows (1 in E-dist(C2), 2 in E-raw) would move the estimate by −0.4 pp if counted as correct, so the reading does not depend on the tie rule. Pairwise agreement between judges ranges from 0.957 (judges 1 and 2) to 0.981 (judges 3 and 4). The run made 976 calls. This settles the reading of the arm under the rule registered for it; it does not change what H4 compares (§6.6).

7. Discussion

Extraction against raw retrieval. On this corpus, with this reader, changing how the builder is configured moved accuracy from 15% to 82%. A study that holds the extraction step at its defaults measures one configuration of the builder, and the gap it reports between extraction and raw retrieval holds for that configuration. We ran no retrieval arm in the confirmatory stage, so nothing here says extraction beats retrieval. What the run shows is that the builder's configuration is a large effect modifier in this system. Whether it changes the conclusions of the studies in §2 was not tested here.

The prediction of 2601.00821. It concerns retention, which this study does not measure. The v1.4 arm compares accuracy on two materials under perfect retrieval, with new responses on the same 242 instances, so its result is at most compatible with the prediction and cannot adjudicate it. Descriptively, the gap between units and raw evidence of the same sessions was 35.5 pp for C1 in development, 8.5 pp for C2 in development and 5.4 pp for C2 on the 242 under judge 1 (5.8 pp under the committee). These come from different samples and criteria, and no run estimates C2 against C1 on new data. For context, 2601.00821's near-verbatim extraction trails verbatim chunks by about 11 points on LoCoMo, under retrieval and a different pipeline. On the 242, the committee of v1.5 puts the loss at −5.8 pp [−10.4, −1.2], and three of the four judges show a loss on their own. The committee was added after the first two judges disagreed, with its rule fixed before the new judges ran, and it is four language models, two of them in the reader's family. It adds measurements, not a ground truth; the blind human grading of the discordant pairs (§8) is still the step that would validate it.

The window. The production finding that motivated the window factor is a property of a deployed system. The benchmark imposes it by construction. The descriptive split in §6.4 is consistent with the window acting largely by deciding whether the evidence sessions reached the builder, which is the serial mechanism the analysis plan named. It does not identify that mechanism. The interaction against the serial null was inconclusive, so the run neither selects the serial model nor supports a joint one.

The instrument. Three things went wrong before the result. The floor was justified with numbers from another corpus. A smoke of 8 instances was treated as if it could check a threshold. And a clean sample was

defined by question id when exposure also travels through rewritten variants and shared evidence sessions. The first two were found by the run itself and the third by external review before deposit. A pre-registration let these be stated as failures rather than absorbed into the analysis. It did not prevent them.

8. Limitations

- **Judges:** the confirmatory contrasts have two judges and H4 has four, all language models. R3, judge 2 and judge 4 are all Gemini models. The bracket is read on contrasts. The blind human pass the plan provides for has not been done, so the conclusions are conditional on the judges. It matters most for H4: its first two judges parted on 0.3 pp and on 35 to 36 discordant pairs, and the committee that settled the split was registered after the split was seen.
- **Margin of H4:** the 10 pp margin was fixed after the development gap of C2 (8.5 pp) was known. It is a registered threshold, not a margin of practical irrelevance, and a non-inferiority result at 10 pp would still allow a loss larger than the one observed in development.
- **Adaptive development:** R3 and C2 were chosen on development data after earlier versions failed. The confirmatory sample was untouched, but the intervention it tests was shaped by what failed before.
- **Preference questions:** their references are rubrics rather than answers. In the exploratory run on the spent 200, the 13 preference questions scored 0.077 under R3 even on raw evidence. The 242 hold 16: cell 11 scored 0.438 on them and raw evidence 0.250. We read the low scores as a sign that reading or grading this category is hard; they do not establish a ceiling.
- **Window shape:** the cut takes whole sessions, most recent first. Sensitivity to the shape of the cut was not measured.
- **One corpus, one reader, one extraction model:** LongMemEval sessions are synthetic dialogue. The extractor is `gemini-2.5-flash-lite` throughout.
- **Language:** the builder, reader and abstention prompts are in Portuguese, the deployed system's language, while LongMemEval is in English. Units mix both languages. The distillation-loss diagnosis found no difference in the share of Portuguese units between lost and kept instances (0.10 against 0.14).
- **Context budget:** every unit of a cell is served. Cell 11 serves about 6.4 times the characters of cell 00 (231,416 against 36,199), and at the full window C2 serves more than C0 (231,416 against 142,733). The result says nothing about accuracy at an equal budget, and the padded arm does not test a competing use of that budget.
- **No selective retrieval:** no retriever chooses among the units. The gain may not survive a context limit or a retriever that serves only part of the memory.
- **No unanswerable questions:** the 30 abstention instances were excluded. The fall in abstention on answerable questions does not show a better balance between correct answers and wrong ones where no answer exists.
- **No leaderboard comparison:** neither judge is LongMemEval's official judge.
- **No production confirmation:** C2 has not run on the production corpus that motivated it.

9. Release

- Pre-registration, cumulative through v1.3: 10.5281/zenodo.23934489 (concept 10.5281/zenodo.23716612).
- Pre-registration v1.4, the raw-evidence arm: 10.5281/zenodo.23964906 (deposited after the run; the tag `paper3-prereg-v1.4` precedes it).
- Confirmatory result with data: 10.5281/zenodo.23949145. It holds the rows, receipt, analysis, manifests and runner. `lme-analyze.py --check` reproduces the analysis without any API key.
- Raw-evidence arm result with data, as a new version of that record: 10.5281/zenodo.23964931. `raw-arm.py --analyze` reproduces its analysis without any API key.
- Pre-registration v1.5, two more judges on H4: 10.5281/zenodo.23966699 (deposited after the run; the tag `paper3-prereg-v1.5` precedes it).
- Four-judge H4 result with data, as version 1.2 of the result record: 10.5281/zenodo.23966704. `h4-judges.py --analyze` reproduces its analysis without any API key.

- Code: the public repository `totobusnello/memoria-nox`, directory `paper3-construction/`, at tags `paper3-prereg-v1.0` through `paper3-prereg-v1.5`.

10. Sources of the numbers in this text

section	file
§1, §2, §7	the arXiv abstract pages of the four studies, read on 2026-11-04; 2601.00821v4 Appendices F and H (Tables 9 and 11)
§3	`README.md`, "Numbers it rests on"; `PREREGISTRATION-DRAFT.md` §11.2, §11.5 item 7, §11.6
§4	`PREREGISTRATION-DRAFT.md` §11.1–§11.5, §12.3–§12.7, §13.3, §14.3, §14.9
§5.2	`STAGE1-RESULT.md`, `runs/stage1-2026-11-03/`
§5.3	`PREREGISTRATION-DRAFT.md` §13.10, `runs/calibration-2026-11-03/`
§5.4	`READER-CEILING.md`, `runs/ceiling-2026-11-03/`; `PREREGISTRATION-DRAFT.md` §14.1
§5.5	`PREREGISTRATION-DRAFT.md` §14.8–§14.10, `runs/c2-calibration-2026-11-03/`
§6.1–§6.3, §6.5	`CONFIRM-RESULT.md`, `runs/confirm-2026-11-04/analysis.txt`
§6.3 cell accuracies, §6.4	`manuscript-tables.py` (descriptive, keyless)
§5 (v1.4), §6.6, §7	`PREREGISTRATION-DRAFT.md` §15, `RAW-ARM-RESULT.md`, `runs/raw-arm-2026-11-04/raw-arm.txt`
§5 (v1.5), §6.7, §7, §8	`PREREGISTRATION-DRAFT.md` §16, `H4-JUDGES-RESULT.md`, `runs/h4-judges-2026-11-04/h4-judges.txt`

References

- *Fidelity before Structure: Verbatim Chunks Beat Lossy Artifact Extraction.* arXiv 2601.00821.
- *When Does Selection Replace Extraction? A Pre-Registered Test.* arXiv 2609.34227.
- *MemoryTax: The Bitter Lesson of Agent Memory.* arXiv 2610.29823.
- *RoutePrism: Tracing Construction Order Effects in Agent Memory.* arXiv 2609.34160.
- *LongMemEval*, `s_cleaned` release, revision `98d7416c24c778c2fee6e6f3006e7a073259d48f`.
- `gbrain-evals`, workstreams W8, W10b and Cat 35 (not peer reviewed).