

Construction before search: a pre-registered factorial test of two repairs to LLM-extracted agent memory

Luiz Antonio Busnello

Version 1.0 · doi:10.5281/zenodo.23716613

Status. Version 1.0. §11 is the active design; §4 is kept as the blocked first attempt. Not peer reviewed. §11 supersedes §4 as the design; the Codex pass (`audits/2026-10-27-codex-challenge-s11.md`) then found two literal errors in §11 — the dose in the wrong unit and the truncation direction inverted — plus a power calculation that does not cover its own estimand, decision labels contradicting their intervals, two controls a constant NAO SEI would pass, and four claims asserted rather than derived. All are addressed, and the two items §11.7 had left open were closed the same day — both making the study harder: **n=193 rather than 52** for production’s own effect size, and the dose demoted from a point to a three-level factor because its median rests on **six** sessions spanning [0.020 .. 0.894]. A later run then bracketed the criterion effect across judge families at **[+8.1 · +21.2] pp** (§11.5 item 7), which is pre-specified here as the reporting form. §4 stays in the document as the record of a design its own sampling code refuted — removing it would hide why §11 looks the way it does.

[defect] **On §4 as written — blocked 2026-10-27.** Writing the sampling code revealed that the corpus cannot support §4. The window factor has variance in **12 of 194** eligible sessions and **8 of those 12 were already used** in the exploratory run; the >1M stratum holds **zero** sessions and 50–200k holds **three**. §4 asks for 60 sessions, 15 per stratum. A confirmatory run here would re-use the sessions that generated the hypothesis.

The correction is in `audits/2026-10-27-neither-cause-explains-it-alone.md`, and the choice of redesign is Toto’s — see **§10**. Everything below §4 stands as written and is reusable; §4 is the part that broke.

Status. Written 2026-10-27 after the collision check (`README.md`), and **revised the same day after reading 2601.00821 in full** — which took the faithfulness half off our side of the ledger and replaced it with a named prediction to test. H2 and H3 are the revision. **Nothing has been run under this plan.** It is submitted for Toto’s approval; on approval it gets a Zenodo DOI *before* any arm executes, and the DOI goes in the paper. The exploratory run that motivated it (`audits/2026-10-27-*`) is cited as exploratory and never as confirmation.

1. The question, and why it is not settled by the literature

Four studies find that LLM-extracted memory loses to retrieval over raw turns, and conclude that the answer is better search rather than better construction (2601.00821, 2609.34227, 2610.29823, 2609.34160). All four hold the extraction step at its defaults — 2610.29823 says so: “*Mem0 and Zep run with their default extraction and retrieval, and other settings may do better*” — and 2601.00821, which is the closest and the one that isolates the mechanism, states the open question as a prediction:

*“Better extraction prompts — explicit date/time targeting, temporal attribute modeling — could shrink this class, **but cannot reach the unconditional retention of a verbatim store.**”*

This plan tests that prediction, and adds a factor no benchmark study can have: how much of the conversation the builder was shown. **factorial** appears **zero** times in all four.

H1 (primary). The window the builder is shown and the criterion it extracts under **interact**: the joint effect on answer accuracy exceeds the sum of the two main effects.

H2 (the prediction). With both changed, extracted memory is **non-inferior** to naive RAG over the raw transcript. Pre-specified in **two budget regimes**, because the exploratory run split on exactly this and reporting one alone would be choosing the answer:

- **H2a, equal context budget** — the memory arm truncated to the characters the deployed memory serves. The exploratory estimate was 48.5% against 63.6%, i.e. **the prediction holds**.
- **H2b, native budget** — the memory arm serving all the units a repaired builder produces, which is 2.4× the deployed volume and is what the fixed system would actually serve. The exploratory estimate was 63.6% against 63.6%, i.e. **the prediction fails**.

Both are reported whatever they say. A result that holds in one regime and fails in the other is the outcome, not a problem with it.

H3 (mechanism — identifier survival, not text survival). 2601.00821 establishes that accuracy tracks how much **source text** survives the store. We predict the governing quantity is narrower: the survival of the **identifier the question asks for**. Two consequences are tested:

- **H3a.** The interaction is concentrated in facts whose answer is an identifier — file name, path, command, config key, numeric value — and is small or absent for lessons and decisions.
- **H3b.** Accuracy draws level **well below** full retention. Mechanical identifier-survival rate is measured by string match, with no judge, before any accuracy is read; the exploratory run saw 6.6% → 12.9% against a verbatim store’s 100%.

[limit] **H3a is confirmatory of An, not novel.** That extraction loses the literal detail is already established there (“*use type hints everywhere*” → “*the user prefers type hints*”, 14.0% against 91.0%), with a 69%/31% write-time/query-time split. H3a is included because the factorial needs it, and it is labelled as replication. **H3b is the new claim.**

2. Design — 2×2 factorial, within-question

Every arm answers the **same** questions, so every contrast is paired.

factor	level 0	level 1
window	the prose the builder actually saw, cut at its processed_at	the session’s complete prose
criterion	DISTILL_SYSTEM as deployed	the same, plus one paragraph requiring identifiers verbatim

Four memory arms, plus three reference arms that are **controls, not results**:

arm	role
W0C0, W0C1, W1C0, W1C1	the factorial
oracle — the raw window around the evidence	positive control. A question it cannot answer is a bad question and the item is dropped before any memory arm is read
naive RAG — BM25 and dense over raw 2000-char chunks, truncated to the memory arm’s characters	the comparator for H2
replica check — W0C0 against the memory the live system actually holds	fidelity. W0C0 must reproduce the deployed system

Fixed across all arms, and named here so no later change is silent: extractor `gemini-2.5-flash-lite` at temperature 0.1 with `responseMimeType: application/json`; the **10-item ceiling**; importance ≥ 0.5 ; blocking at `BLOCK_SIZE = 2000` by the deployed algorithm (which never splits a line); answerer `gemini-2.5-flash-lite` at temperature 0 with the deployed `ANSWER_SYSTEM`, which permits NAO SEI; judge `gpt-6-luna` via the Decisions API with the predicates and 4-level scale already in use.

3. Primary endpoint and test

[scope] **Boundary** — §§1–10 are the first attempt; §11 is what will run. Everything from here to §10 describes the production-corpus design that §4 records as blocked, and it is kept because removing it would hide why §11 has the shape it has. **It is history, not instruction.** Where §11 specifies something this section also specifies — the estimand, the test, the α , the sample, the stopping rule, the cost — **§11 governs and this section does not.** Nothing below needs to be reconciled with §11 before the run.

Endpoint. Per (question, arm): the judge’s **contem** probability thresholded at 0.5 — the candidate conveys the reference fact. Pre-specified; the three-cut sensitivity (0.3/0.5/0.7) is secondary.

Primary test — **H1.** The interaction contrast on paired data:

$$I = (W1C1 - W1C0) - (W0C1 - W0C0)$$

estimated by a logistic GLM with question-level random intercepts, `window * criterion` as fixed effects, and the interaction coefficient’s Wald test at **two-sided** $\alpha = 0.05$. H1 is supported only if the interaction term is positive and significant. **A positive joint effect with a null interaction term refutes H1**, and that outcome is written up as such.

H2. One-sided non-inferiority of W1C1 against the better naive-RAG arm at equal budget, margin **−5 percentage points**, the margin 2609.34227 used, chosen so the two studies are comparable and not because it flatters us. H2 fails if the lower bound falls below −5 pp.

H3. Interaction estimated within each of `identifier / lesson / decision`, pre-declared as **exploratory-confirmatory**: directional, with Holm correction over the three, and reported as suggestive if only the direction holds.

4. Sample, and the power it buys

Unit. A (question, session) pair. Questions are generated from the **raw** prose by a model that never sees any memory arm, under the deployed DISTILL_SYSTEM’s own criterion, and must carry a verbatim citation that is verified by string match against the source before the item enters.

Target n = 400 questions over 60 sessions, stratified by final transcript prose (<50k / 50–200k / 200k–1M / >1M), 15 sessions per stratum, drawn with a seed recorded in this plan (**20261028**) from the 194 sessions that hold both a raw transcript and distilled chunks.

Power. The exploratory run gave main effects of +4.0 pp each and an interaction of +13.2 pp at n=99. A paired interaction contrast carries roughly twice the variance of a main effect of the same size, so n=400 gives >0.90 power for a 13 pp interaction and ~0.80 for 8 pp at $\alpha=0.05$. **n=400 is not powered to resolve an interaction below ~6 pp**, and a null at that scale will be reported as inconclusive rather than as absence.

Stopping rule. Fixed n. No interim look, no extension on a near-miss. If the pipeline fails on a session it is replaced by the next draw in the same stratum, and replacements are reported.

5. Exclusions, declared in advance

1. Items the **oracle** cannot answer (bad questions). The exploratory run lost 5 of 104 this way.
2. Items whose verbatim citation is not found in the source prose.
3. Sessions whose turns are under 90% timestamped — the window cut would be unmeasurable.
4. **Cron sessions** ([cron:...], HEARTBEAT_OK), which the deployed builder filters by design.

Nothing else is excluded, and in particular **no item is dropped for being hard**.

6. What would falsify each claim

claim	falsified by
H1	<i>(first attempt’s wording; §11 governs the leg that runs)</i> interaction term not significant, or negative
H2a	lower bound above –5 pp at equal budget — i.e. An’s prediction would fail where we expect it to hold
H2b	lower bound below –5 pp at native budget — the prediction holds everywhere and we have no accuracy result
H3a	interaction as large on lesson/decision as on identifier
H3b	accuracy drawing level only once identifier survival approaches a verbatim store’s, which would make the mechanism text survival after all and leave An’s law unrefined
the whole instrument	W0C0 failing to reproduce the deployed system’s accuracy within 5 pp

Note that H2a and H2b are falsified by **opposite** results, deliberately: the claim is that the budget regime decides, so a plan that could only be confirmed by one direction would not be testing it.

7. Known weaknesses of this plan, stated before the result

- **The judge is one model family.** Deterministic on this corpus (540 pairs, 0 label changes) but single. A blind human grading of a 60-item stratified subsample is **required** for the strong form of H2; without it H2 is reported as judge-only. Human grading is Toto’s to schedule — and the grader must not be whoever built the arms, which is the limitation 2609.34227 had to declare.
- **The replica is not the deployed builder.** It omits the embedding dedup (isDuplicate), which needs the live database, and the learned noise prototypes, because isNoise writes to noise_prototypes. Both omissions give the memory arms **more** units, so they bias against our own hypothesis. The exploratory run’s replica matched the deployed accuracy exactly (42.4%) in aggregate but missed two cells; per-cell claims therefore carry that caveat.

- **One corpus, one language, six agents.** External validity is not claimed. The window factor is specific to systems that build memory incrementally from live sessions.
- **The production system must not be repaired before this runs.** Fixing `session-distill.ts:141-144` or the prompt would destroy the W0C0 cell, which *is* the deployed system. Repair is sequenced after the run.

8. Release

Plan DOI before execution. On publication: the code, the generated questions in redacted form, the judge keys and all per-item labels. **The corpus prose is not releasable** — it holds live credentials this repository has already catalogued, and 2.66M characters of the user’s own conversation. The paper states plainly which numbers a third party cannot reproduce from the release and why.

9. Cost

step	estimate
distillation, 4 cells over ~60 sessions	~US\$ 1.30
answering, 7 arms $\times$ 400	~US\$ 0.80
judging, ~2,800 items	~US\$ 0.25
embeddings for the dense RAG arm	~US\$ 0.25
total	~US\$ 2.60

Anchored on measured costs from 2026-10-27, not on list prices: 563 blocks cost US\$ 0.2436, 99 pairs over 4 arms cost US\$ 0.1082, 396 judgements cost US\$ 0.0318.

10. The redesign the corpus forces — three options, one recommendation

Added 2026-10-27 after §4 was found unexecutable.

The binding constraint is sessions, not questions. The unit of inference is the session: questions from one session share its prose, its builder window and its memory. $n=400$ questions over 12 sessions is not $n=400$. And the window factor — the one production has and benchmarks do not — varies in only those 12.

option	what it buys	what it costs
(a) publish the exploratory run as exploratory	honest, cheap, available today; 2610.29823 publishes on one conversation and says so	no confirmation of anything, and the interaction rests on 8 sessions
(b) manipulate the window experimentally on this corpus — truncate the prose artificially instead of using the window the builder happened to have	gives the factor variance in all 194 sessions, and makes it a manipulated variable rather than one confounded with session size	only 28 sessions hold ≥ 2 blocks of prose and 16 are unused; the truncated window is not the window production produces
(c) two legs: production for existence, a public corpus for the estimate	the interaction is estimated on LoCoMo/LongMemEval with the window truncated under our control — large n , reproducible by third parties , which also answers the release problem — while production supplies what no benchmark can: that the window occurs naturally, and its magnitude (median 24.0% seen in the 200k–1M band)	two instruments to build and validate; the public-corpus leg inherits the criticism we make of benchmarks

[key] **Recommendation: (c), with (a) as its production leg.** It separates the two claims by the evidence each needs. *That the window exists and bites* is a fact about production and can only be shown there —

observationally, on 12 sessions, as an existence result with a magnitude, which is exactly what 12 sessions can carry. *That window and criterion interact* is a quantitative claim that needs n and a manipulated factor, and a public corpus gives both plus third-party reproducibility.

It also fixes the release problem from `README.md` §Open decisions: the confirmatory numbers would rest on data anyone can download, and the production leg would be cited for existence, where the unreleasable prose matters least.

[limit] And it keeps us honest about something (b) would hide: a truncated window is **our** truncation. Production’s window is set by a 7-day prune crossed with a 30-day lookback and a 50-session-per-run ceiling, and reproducing that shape on a benchmark is a modelling choice to declare, not a detail.

11. Option (c), designed — the public-corpus leg

Written 2026-10-27, after §10’s recommendation was approved. §4 was written and *then* found unexecutable, so this section prices and powers the design **first** and lets the arithmetic choose n . Every number below comes from a census of the downloaded data (`paper3-construction/lme-census.py`), not from a README — this repository’s ROADMAP said SHIPPED for features the product did not have.

11.1 The corpus, and why not the other one

LongMemEval s_cleaned, revision 98d7416c24c778c2fee6e6f3006e7a073259d48f, 277.4 MB, sha256 prefix d6f21ea9d60a0d56. MIT.

	LoCoMo	LongMemEval s_cleaned
independent units	1,986 questions over <code>list[10]</code> conversations	500 instances, 0 sharing a haystack
sessions per unit	19 per conversation	38–62, median 48
chars per unit	—	454,980–513,953, median 489,763
units $\geq$ 200k chars	—	500 of 500
licence	CC BY-NC 4.0 (non-commercial)	MIT

[defect] **LoCoMo would repeat exactly the defect that killed §4.** Ten conversations is fewer units than production’s twelve, and the unit of inference is the conversation, not the question.

And two properties of `s_cleaned` that matter more than its size:

1. **Every instance sits in the 200k–1M band**, which is where production showed the damage: -35.0 pp against naive RAG, median seen-fraction 24.0%. The benchmark is not easier than the case of interest; it is entirely inside it.
2. **The char distribution is narrow** — 455k to 514k, a 13% spread. In production the window is *confounded* with session size (a longer session means a smaller seen-fraction). Here size is near-constant and the window is manipulated, so **the factor is unconfounded**. That is the one thing a benchmark can give us that production cannot.

[limit] **Against the README: there are no _abs abstention variants.** It lists them for seven categories; the census finds **0 of 500** in both `oracle` and `s_cleaned`, across 6 categories. So this leg cannot carry an abstention claim, and nothing here is planned on one.

11.2 Unit, design, factors

[defect] **The instrument, named here because §11 did not name it.** Every parameter below lived only in §2, and §3’s boundary note declares §§1–10 *history, not instruction* — so the section that governs specified no model, no temperature, no thresholds, and the controls in §11.3 were validating a machine the plan never described. A third adversarial family named it (GLM-5.3). A pre-registration whose instrument is unnamed makes any implementation conforming, which is the freedom pre-registration exists to remove. The values are not new; they are the deployed ones, moved into the section that binds:

component	fixed value
extractor	gemini-2.5-flash-lite, temperature 0.1 , responseType: application/json
item ceiling	10 per block
salience cut	importance >= 0.5
blocking	BLOCK_SIZE = 2000, the deployed algorithm, which never splits a line
answerer	gemini-2.5-flash-lite, temperature 0 , deployed ANSWER_SYSTEM, which permits NAO SEI
judge	gpt-6-luna via the Decisions API, the predicates and 4-level scale already in use
criterion level 0	DISTILL_SYSTEM v1, verbatim as production ran it
criterion level 1	DISTILL_SYSTEM_V2 — v1 plus the verbatim-identifier paragraph

[limit] **Serving is the one piece still not pinned here.** §11.5 records the served context length per cell, which presupposes a retrieval-and-serving procedure over the distilled units; that procedure is described only in §§1–10. Until it is written into this section, the per-cell length record is a measurement of something the plan does not define, and that limit is declared rather than hidden.

Unit of inference: the instance — with the coupling measured, not asserted.

[defect] The first draft said “500 of them, no shared haystacks, so n is n ”. That compared the sorted tuple of `haystack_session_ids`, which only detects haystacks identical **as a whole**; the adversarial pass (finding 6) pointed out that instances can share **component** sessions. Measured (`lme-independence.py`, `lme-answer-overlap.py`):

coupling	measured
session ids appearing in more than one instance	20.5% of 19,195 distinct
instances with zero shared sessions	0 of 500 — median 36.0% of a haystack is non-exclusive
pairs sharing ≥ 1 session	4.30% · max 3 of ~48, median 1 · 0 pairs share ≥ 10
pairs sharing an ANSWER session	5 of 124,750 = 0.004%
instances in an evidence-coupled pair	10 of 500 = 2.0%
answer sessions that are a distractor elsewhere	0 of 940

⇒ “ **n is n** ” is **withdrawn**. But the coupling that can propagate into the estimate runs through *evidence*, and it touches **2.0% of instances in 5 pairs** — not the structure that killed §4, where 12 units carried 400 observations. The consequence is procedural, not fatal: **the standard error is cluster-robust over answer-session groups**, pre-specified here, and the 5 coupled pairs are listed in the final text so a reader can drop them. Distractor-level sharing is left in deliberately: it changes the haystack both instances must search, which is a property of the corpus, not a dependence between outcomes.

2×2 within-instance — every instance passes through all four cells, the contrast is paired, and between-instance variance drops out of the interaction estimate.

factor	level 0	level 1
W — window	truncated : the distiller sees the most recent sessions whose cumulative prose reaches 0.240 of the haystack’s characters	full : all 38–62 sessions
C — criterion	DISTILL_SYSTEM v1, verbatim as production ran it	DISTILL_SYSTEM_V2, v1 plus the verbatim-identifier paragraph

[defect] **Both halves of that cell were wrong in the first draft, and the adversarial pass caught them** (audits/2026-10-27-codex-challenge-s11.md, finding 1).

- **The unit was wrong.** Production’s 0.240 is a fraction of **prose characters**; the draft kept 0.240 of the **number of sessions**, with a ceiling. Sessions are not equal in length, so that does not preserve even the nominal dose. It is now a cumulative-character cut.
- **The direction was inverted.** The draft said “*production’s prune removes the oldest sessions, so a leading prefix is the right shape*” — but if the prune removes the oldest, what the distiller sees is the **most recent**, and a leading prefix keeps precisely what the prune discarded. The truncation now takes the recent tail.

The 0.240 is production’s measured median seen-fraction in the 200k–1M band, and it remains an *approximation of the mechanism*, not the mechanism: production’s window is a 7-day prune crossed with a 30-day lookback and a 50-session-per-run ceiling, and it cuts mid-session at **processed_at** rather than at a session boundary. [limit] **What this approximation cannot separate** is unavailability from recency and from context composition: a recency-tail cut removes old evidence, contradictory updates and redundancy together. A positive interaction under it is causal **for this truncation**; attributing it to production’s window needs the sensitivity arms of §11.5 to agree, and if they do not, the finding is about our cut.

[defect] **The denominator is 6, and it dissolves the point anchor** (measured 2026-10-27, *see-once-denominator.py*). Every seen-fraction in the 200k–1M band:

session	total chars	visible	fraction
ea80140b	248,614	4,867	0.020
9c75f331	631,998	18,999	0.030
d37a5964	344,003	62,954	0.183
5a944bec	268,055	79,625	0.297
1a5d8840	625,030	263,634	0.422
a35348df	349,560	312,441	0.894

n = 6. min 0.020, max 0.894, spread 0.874 — nearly the whole of [0,1]. And **0.240 corresponds to no observation at all**: it is the interpolation between 0.183 and 0.297. The neighbouring 50–200k band is n=3 at [0.392, 0.721, 1.000].

⇒ **the dose cannot be a point.** The ends are where half the band sits, and the sensitivity arms first pre-specified (0.12 and 0.48) sat *inside* the observed range and would have missed both. This is the §4 failure caught one step earlier: a statistic resting on fewer units than it appeared to, used to parameterise something much larger.

[defect] **But promoting W to a three-level factor dissolved the design, and the second adversarial pass caught it** (finding 2): with $W \in \{0.020, 0.240, 0.894\}$ the **full window disappears from the matrix** — while the estimand, the fidelity control and the whole notion of repair all depend on it. Worse, no primary contrast was named for a three-level factor, and a multi-degree-of-freedom joint test is not “superadditivity”. The matrix is therefore fixed, once, here:

	C0 = DISTILL_SYSTEM v1	C1 = v2 (+identifier paragraph)	role
W = 0.240 (median seen-fraction)	cell 00	cell 01	primary 2×2
W = full (all 38–62 sessions)	cell 10	cell 11	primary 2×2
W = 0.020 (band minimum)	sens	sens	dose sensitivity
W = 0.894 (band maximum)	sens	sens	dose sensitivity

Four primary cells on every instance; four sensitivity cells on a pre-specified n=60 subsample of stage 1. The primary contrast is the one §11.4 fixes — full+v2 against truncated+v1 — and the dose arms answer a different question (does the sign of the effect survive across the observed band?) at the n that question needs, not at the primary’s n.

The cut rule, executable rather than described. “Sessions whose cumulative characters reach the fraction” does not determine a dose on its own, and at W=0.020 a single session can overshoot it entirely. So, in order: sessions are taken **most-recent-first**; the running total is of `len(prose)` over the session’s turns; the session that **crosses** the threshold is **included whole**; the denominator is the haystack’s total prose characters, excluding the question and the answer. The realised fraction is **recorded per instance** and reported as a distribution, because with 38–62 unequal sessions the realised dose will not equal the nominal one and the paper should say by how much rather than claim the nominal figure.

Response: was the generated answer correct, judged by the bounded judge (Decisions API, no free-text field, so the prose-contradicts-the-field failure cannot occur). [limit] Not GPT-4o, which is the LongMemEval paper standard — so **our numbers are not comparable to the published leaderboard**, and the paper must say so rather than invite the comparison. The judge is held identical across all four cells, which is what the interaction needs.

11.3 Controls

control	construction	required result
positive	truncate so that every answer-bearing session is dropped (evidence is session-level; median 2 answer sessions, 0 instances with none , so this is always constructible)	accuracy at floor. If a cell with no evidence still answers, the pipeline is reading something it should not
negative	truncation that drops only non-answer sessions, keeping all evidence	accuracy $\approx$ the full-window cell. A drop here means truncation hurts through something other than losing evidence, and the interaction estimate is contaminated
fidelity	the W1C1 cell re-run against our own replay harness	must reproduce the arm-C/D numbers on our prose, as C_visible reproduced B’s 42.4%/7.1% exactly

[defect] **Both controls can pass with a broken pipeline, and the adversarial pass gave the counterexample (finding 5): a system that always answers NAO SEI.** It sits at the floor in the positive control — pass — and keeps the full-window accuracy in the negative one — pass. So the required results are necessary and **not** sufficient, and two more legs are added:

control	construction	required
answerable floor	the full-window cell on the same instances	accuracy at or above 50.0% . A pipeline that cannot answer <i>with</i> the evidence makes the other controls vacuous
paired diagnosis	per-item, not aggregate	the negative control’s equality must hold item-wise within a margin , because aggregate equality hides compensating losses and gains

And two asymmetries to state rather than assume:

- in the positive control, low accuracy can also come from a failing reader or judge, so it is diagnostic only once the answerable floor passes;

- answering **without** the annotated evidence does not prove improper access — prior knowledge, cues in the question, unannotated redundancy and judge false positives all produce it. “*No marked session*” is not demonstrably “*no usable information*”, and the completeness of the evidence annotation is itself unverified.

[defect] **The floor is 50.0%, fixed here, and here is where it comes from.** For two revisions this read “*a pre-set minimum*” and promised that “the final text” would state a number — which is deciding the pass **after** seeing the data, the exact freedom a pre-registration exists to remove. The number is derived, not picked: the deployed system scores **42.4%** on this corpus and the repaired one **63.6%**, so a full-window cell below 50.0% sits closer to the unrepaired system than to the repaired one while holding all the evidence, and nothing it then says about repair is interpretable. At $n=200$ the standard error of a proportion near 0.5 is 3.5 pp, so 50.0% is about 2.2 standard errors above 42.4% — far enough from the unrepaired accuracy to be a real test rather than a rounding of it. **If the floor fails, the run is reported as an instrument failure and the primary is not interpreted**, rather than reported as a null.

[defect] **The added legs are still not sufficient, and the second adversarial pass exhibited the pipeline that beats all of them** (audits/2026-10-27-codex-and-deepseek-on-the-prereg.md, finding 7):

```
in the benchmark loader:  keep answer_session_ids and the reference answer on the item
when answering:          if any session marked as evidence is still present
                          -> return the reference answer
                          else
                          -> return NAO SEI
```

It sits at the floor in the positive control, matches the full window in the negative one, clears the answerable floor, and passes the paired diagnosis **item by item** — while never reading the constructed context at all. Every control above tests the *outcome* of the causal path; none tests that the path was walked. One more leg does:

control	construction	required
counterfactual evidence	on a seeded 40-item subsample, rewrite the answer-bearing session so the fact itself differs (a date shifted, an identifier changed), leaving the question, the haystack size and the annotation untouched	the answer must follow the altered fact. A pipeline that returns the original reference answer has not read the context, and no aggregate control can see that

This is the only control here that distinguishes *answering correctly* from *answering from the constructed context*, and it is cheap: the benchmark ships the evidence spans, so the rewrite is mechanical and the expected answer is known by construction. It costs US\$ 0.03 per item at the measured rates, so the leg adds about **US\$ 1.20** to either stage.

The positive control is cheap to construct, which is why this corpus was chosen over manufacturing one — but cheap to construct is not the same as sufficient to validate, and three rounds of this section have now each been defeated by a pipeline simpler than the one before.

11.4 Power, cost, and a declared stopping rule

Unit costs are **measured, not assumed**: US\$ 6.22e-4 per distiller block (arm C: 563 blocks for US\$ 0.35) and US\$ 6.4e-5 per judge call (152 calls for US\$ 0.0098). Blocks per instance: 245 full, 59 truncated. **US\$ 0.380 per instance for all four cells.**

[never] **Superseded twice — kept as the record of what was wrong, not as a plan.** Its δ column is mislabelled: 21.4 pp is production’s **total** effect, not its interaction, and its $\text{Var}(D)$ was the raw discordance. The governing numbers are in *Resized twice* below.

effect to detect (<i>column mislabelled “interaction” in the original</i>)	paired n	cost	distiller calls	wall-clock at 8 concurrent
21.4 pp — production’s total effect, mislabelled here as interaction	52	US\$ 20	31.6k	~2.2 h
15 pp	105	US\$ 40	63.8k	~4.4 h
10 pp	236	US\$ 90	143k	~10 h
8 pp	368	US\$ 140	224k	~16 h
5 pp	942	—	—	exceeds the corpus

[defect] **This table does not yet power the primary contrast, and the adversarial pass showed why (finding 3).** The estimand is the additive interaction

$D_i = Y_{11,i} - Y_{10,i} - Y_{01,i} + Y_{00,i}$, with $D_i \in \{-2, -1, 0, 1, 2\}$ and target $E[D_i]$ in percentage points.

Its standard error depends on $\text{Var}(D_i) = \text{Var}(A) + \text{Var}(B) - 2 \cdot \text{Cov}(A, B)$ with $A = Y_{11} - Y_{10}$ and $B = Y_{01} - Y_{00}$. **A single “discordance of 0.30” does not specify that distribution:** it fixes each pair’s discordance and says nothing about the covariance **between** the two contrasts, and the arithmetic above silently treated 0.30 as $\text{Var}(D)$ itself. Pairing removes some between-instance variance, not all of it.

Three further defects in that table, all real:

- **α mismatch.** It uses $\alpha = 0.05$; the stopping rule below spends 0.03 / 0.02. The power must be computed for the procedure actually run, including the two-stage structure.
- **Scale mismatch.** §3 prescribes a logistic interaction coefficient, which is interaction in **log-odds**, not superadditivity in percentage points. The two can disagree in sign of significance. The estimand for the leg that will run is fixed as the **additive** one in §11.4, and §3 is **not** amended: it is the first attempt’s endpoint and stays as the record of it. See the boundary note at the head of §3 — a plan that says “§3 is amended to match” without carrying the replacement text leaves the reader to guess which sentences still bind.
- **An inverted sentence.** The draft said the plan is “*revised up if the discordance comes in lower*”. Holding the effect fixed, **lower variance needs fewer observations, not more**. If the intent was to adapt on the observed *effect*, that is a different rule and has to be written as one.

Resized twice, 2026-10-27 — and the second pass moved the primary

The first resize fixed the variance and left the effect wrong. It recomputed $\text{Var}(D)$ properly and took n from 52 to 193, and it kept sizing the test of the **interaction** with the **total** effect as δ . Two adversarial voices from different families found this independently on the same day (`audits/2026-10-27-codex-and-deepseek-on-the-prereg.md`), and the arithmetic is the document’s own:

quantity	value
main effect, window alone	+4.1 pp
main effect, criterion alone	+4.1 pp
interaction	+13.0 pp
total effect of repairing both	+21.2 pp

$4.1 + 4.1 + 13.0 = 21.2$. The table’s row labelled “*21.4 pp — production’s measured interaction*” was production’s **total** effect wearing the interaction’s name, and sizing an interaction test on it overstates power by the ratio of the squares. Ten lines below that row the same section already said “*the honest ceiling of what 500 instances can resolve is ≈ 13 pp*” — so the document stated the ceiling and then sized past it, in one page.

Every number below is computed by `estimand-and-power.py`, which **exits non-zero** if the decomposition stops closing, if the serial null stops being beaten, or if the interaction starts fitting the corpus — the three facts that this design now rests on.

The interaction does not fit, so it is not the primary

estimand	δ	n at $\alpha=0.03$, Var=0.973	n at Var=3.347
total effect — repairing both against repairing neither	21.2 pp	196	675
interaction against 0	13.0 pp	522 — exceeds the corpus	1,796
interaction against the serial null (below)	12.6 pp	556 — exceeds the corpus	1,911

So the primary endpoint is the total effect, which is also the claim the README actually makes: repairing window *and* prompt together recovers the deficit. The interaction becomes a **pre-registered secondary** whose power is stated out loud rather than assumed:

n	power for the total effect	for the interaction
200 (stage 1)	80.8%	38.0%
240	87.7%	44.9%
500 (whole corpus)	99.6%	78.1%

[limit] **The interaction is underpowered at every n this corpus allows**, 78.1% at the full 500 against the 80% the plan targets. That is pre-registered as a limitation, not discovered afterwards: a secondary that fails to reach significance will be reported as *underpowered*, with the interval, and never as evidence of additivity.

“Conservative” was false, and is withdrawn

The first resize varied an exchangeable within-instance correlation $\rho \in [0, 0.6]$ and called the maximum the worst case:

ρ	0.0	0.2	0.4	0.6
Var(D)	0.973	0.779	0.584	0.390

[defect] **That is a maximum inside a dependence family chosen by the author, not an upper bound for paired data with those marginals.** The adversarial pass supplied the counterexample, and `estimand-and-power.py` reproduces it exactly: with $U \sim \text{Uniform}(0,1)$, $Y_{11} = 1[U < 0.636]$, $Y_{00} = 1[U < 0.424]$, $Y_{10} = Y_{01} = 1[U > 0.535]$, all four marginals are the measured ones, $E[D] = 0.130$ is the measured interaction, and

$$\text{Var}(D) = 3.3471 - 3.4 \times \text{the number the draft called conservative.}$$

The word is therefore removed. The plan sizes on $\rho=0$ **because stage 1 measures the joint pattern**, not because $\rho=0$ bounds anything, and the second column of the table above reports what the counterexample would cost so that the exposure is visible before the run rather than after it.

The serial null — a positive interaction is not evidence of joint repair

A two-stage system needs no joint mechanism to produce superadditivity. Let

$$P(\text{correct} \mid w, c) = e_w \cdot q_c$$

where the window governs whether the evidence arrives (e_w) and the prompt governs whether it is used once present (q_c). The additive interaction is then $(e_1 - e_0)(q_1 - q_0) > 0$ whenever both factors help, with nothing jointly repaired: the prompt can only act on information that reached the system. So “**interaction > 0**” **does not identify the mechanism**, and this design stops claiming it does.

The model is not merely a caveat, because it is **fitted and testable**. Scale-free, three cells determine the fourth: $Y_{11} = Y_{01} \cdot Y_{10} / Y_{00}$. On our own measured cells that predicts

	Y11	implied interaction
serial chaining predicts	0.5100	+0.4 pp
production measured	0.6360	+13.0 pp

33× the serial prediction, and the direct check fails in the right direction: $Y_{00} \cdot Y_{11} = 0.2697$ against $Y_{01} \cdot Y_{10} = 0.2162$, which would be equal if chaining were the whole story. The secondary hypothesis is therefore stated against the serial null rather than against zero, which is what finally gives this leg a **refutable prediction**: if the benchmark returns an interaction indistinguishable from I_{serial} computed on its own three cells, the joint-repair thesis is **refuted**, and that is a result this design can produce.

Cost

[defect] **The missing full-window cell was also a missing cost.** The draft quoted 565 blocks per instance for “six cells” — a matrix in which $W=\text{full}$ did not appear (§11.2). With the primary 2×2 restored, blocks per instance are **608** for the four primary cells alone, and the honest figures are these, from measured unit costs (US\$ 6.22e-4 per block, US\$ 6.4e-5 per judge call) and blocks proportional to prose seen (5 / 59 / 219 / 245 at $W = 0.020 / 0.240 / 0.894 / \text{full}$):

component	n	blocks	cost
primary 2×2	200	121,600	US\$ 75.7
dose sensitivity, 4 cells	60	26,880	US\$ 16.7
counterfactual-evidence control	40	24,320	US\$ 15.1
stage 1 total		172,800	US\$ 108
stage 2, primary only	500	304,000	US\$ 189
both stages, worst case			US\$ 297

[limit] **This is US\$ 108 for stage 1, not the US\$ 71 the draft promised**, and up to US\$ 297 if stage 2 runs. The increase is not inflation: it is the price of the cell the design needs and the controls that make it interpretable. Wall-clock for stage 1 is ~6 h at 8 concurrent.

Two stages, every state with an action

- **Stage 1: $n = 200$** , sampled by seed, stratified by question category. **US\$ 108.** $\alpha = 0.03$.
- **Stage 2: all 500** (+US\$ 189), $\alpha = 0.02$. No third stage, whatever stage 2 returns.
- **Stop at stage 1** if the primary CI excludes 0 **and** lies entirely above **+5 pp**. Otherwise go to stage 2. (The earlier wording, “*lies entirely above the equivalence margin*”, read as either -5 or $+5$ pp; under the first reading a CI of (1, 4) would have declared a confirmed repair and an absent effect at the same time.) At $n=200$ and $\text{Var}(D)=0.973$ the half-width is 15.1 pp, so this stop requires an observed effect above ~ 20 pp — about what production showed. There is no futility stop on the primary: at $n=200$ it has 80.8% power, so a non-significant stage 1 is uninformative rather than negative.
- [defect] **This design cannot produce evidence of absence, and says so instead of pretending otherwise.** An earlier version of this section pre-set an equivalence margin of 5 pp and treated a CI inside $(-5, +5)$ pp as evidence of absence. That state is unreachable here, and the arithmetic is the plan’s own: at $\text{Var}(D)=0.973$ the half-width is **10.3 pp at $n=500$** ($\alpha=0.02$) and 15.1 pp at $n=200$, so an interval narrow enough to sit inside ± 5 pp needs $n \approx 2,106$ against a corpus of 500. Widening the margin until it is reachable does not rescue the test: a margin above 10.3 pp would swallow the whole 13.0 pp interaction, so “equivalence” there would assert nothing. Therefore **a primary CI containing 0 is reported as inconclusive, never as absence** — the form §4 already used for the same reason. The cost is stated plainly: **no outcome of this study refutes the repair thesis by showing the repair does nothing.** What the study can refute is the *mechanism* claim, through the serial null, and that leg is intact.
- **If no valid CI is obtainable** (separation, non-convergence): report the point estimate with a percentile bootstrap over **instances** (10,000 resamples, the instance as the resampling unit), and say which path produced the interval.
- **Intervals are adjusted to the spent alpha.** The two-sided interval reported for each decision is the $1-\alpha$ interval at that stage’s α (0.03, then 0.02) — not a 95% interval relabelled. This is fixed here because leaving it to “the final text” is leaving an inferential choice until after the results exist, which is the freedom a pre-registration exists to remove.

What each primary interval licenses — the labels are what the interval supports, not what the author hoped:

primary CI (total effect)	conclusion
excludes 0, contains 21.2 pp	repair confirmed; the interval covers production’s point estimate
excludes 0, entirely below 21.2 pp	repair confirmed, and smaller on this corpus than production’s
excludes 0, entirely above 21.2 pp	repair confirmed, and larger on this corpus than production’s
excludes 0, entirely negative	repairing both hurts on this corpus, contrary to the thesis
contains 0	inconclusive ; report the interval, never a null. See the note above on why absence is not demonstrable at this n

And for the secondary, against the serial null:

secondary CI (interaction — I_{serial})	conclusion
excludes 0 from above	joint repair beyond serial chaining; the mechanism claim survives
excludes 0 from below	the interaction is smaller than serial chaining predicts; the mechanism claim is refuted
contains 0	inconclusive. Equivalence to the serial null is not declarable either, for the reason above: it would need $n \approx 2,106$

11.5 Pre-specified sensitivity and threats

1. **The 0.240.** Repeat at **0.020 and 0.894** on stage 1’s sample — the observed min and max of the six sessions the median rests on, which is the matrix §11.2 fixes. If the interaction’s sign flips inside that range, the result is about our truncation and not about windows, and must be reported that way. (This item ordered *0.12 / 0.48* for two revisions after §11.7 recorded those as the corrected error: they were chosen before the order statistics were computed and sit **inside** the observed range, so they would have missed both ends.)
2. [defect] **The cut’s shape is fixed, and the shape check is NOT bought.** This item used to read “*production’s prune removes the oldest sessions, so a leading prefix is the right shape*” — the exact sentence §11.2 quotes as **wrong**. It is wrong for the reason given there: if the prune removes the oldest, what the distiller sees is the most **recent**, so a leading prefix keeps precisely what the prune discarded. The cut is the **recency tail**, in §11.2 and here. `haystack_dates[]` would allow a scattered truncation as a shape check, and this item used to promise one arm of stage 1 would run it — **no such arm is in the cost table**, at either stage. Rather than carry an unbudgeted promise into a frozen document: the shape check is **not run**, and the result’s sensitivity to cut shape is **unmeasured**, declared here as a limit.
3. **The benchmark inherits the criticism we make of benchmarks.** LongMemEval sessions are synthesised dialogue, not an agent working; identifier density is lower than in our prose, and the criterion factor acts precisely on identifiers. $\Rightarrow$ the criterion effect measured here is **an estimate on this corpus**, with no demonstrated ordering against production — see item 6, which withdraws the lower-bound reading. (This sentence asserted the floor for two revisions after item 6 retracted it; the second adversarial pass found the pair still contradicting each other.)
4. **No _abs variants** $\Rightarrow$ no abstention claim from this leg (§11.1).
5. **Not the paper’s judge** $\Rightarrow$ no leaderboard comparison (§11.2).
6. [defect] **The “lower bound” claim about the criterion effect is withdrawn.** The draft argued that because this corpus has lower identifier density than our prose, the criterion’s main effect here is a **floor** on what it would be in production. That ordering was asserted, not derived: transportability also depends on difficulty, redundancy, question mix and competition for capacity, and higher identifier density could as easily **saturate** the store and reduce the gain. Without a stated additional assumption there is no demonstrated ordering between the two corpora, so the effect here is **an estimate on this corpus** and nothing more.
7. [defect] **A constant judge is not a constant error — and this is now MEASURED, not feared.** Holding the judge identical across cells removes one source of variance; it does not validate the labels. A second family was run over the same 99 replay items (Google `gemini-2.5-flash-lite` against OpenAI `gpt-6-luna`, 594 judgements; audits/2026-10-27-a-second-family-measured-the-problem-it- was-meant-to-remove.md):

contrast	family 1	family 2	difference
d_full - c_all — full window both sides, only the prompt differing	+8.1 pp	+21.2 pp	13.1 pp

The difference between families exceeds the effect one of them reports. The disagreement is almost one-directional (family 2 adds 123 acceptances, removes 1 — a calibration difference, not noise) and **asymmetric toward D**: it adds 17.2% on `c_all` against **30.3%** on `d_full`.

Two readings fit and they have opposite consequences — **literalness bias** (D attracts credit it has not earned, so +21.2 is inflated) or **severity** (family 1 rejects correct answers, so +21.2 is truer). The asymmetry is consistent with both, and two LLM judges cannot separate them.

[defect] **And the pair does not cover the primary.** The two families were compared on `d_full - c_all` — the criterion effect with the window held full — so the bracket bounds **that** contrast, not

the total effect this plan now tests, and not the interaction. Worse, the disagreement is asymmetric **toward the cell the thesis needs**: family 2 adds 30.3% on `d_full` against 17.2% on `c_all`, and `d_full` is the W1C1 corner. A judge that rewards literalness therefore inflates the corner that carries both the total effect and the interaction. $\Rightarrow$ **the blind human grading of §7 must reach the primary contrast**, on a stratified subsample of all four primary cells, or the confirmatory claim is judge-conditional and the paper says so in those words.

8. [defect] **Context length is not controlled, and it is an alternative explanation for the whole effect** (second adversarial pass, DeepSeek finding 4). W1C1 serves more prose than W0C0 by construction, so a positive total effect is consistent with “more context helps” and needs no construction story at all. Three things are therefore pre-specified: the **served context length is recorded per item and reported per cell**; the primary is also estimated with served length as a covariate, reported alongside the unadjusted estimate; and a **length-matched reference arm** runs on the stage-1 sample, serving W0C0’s selection padded to W1C1’s length with the same corpus’s non-evidence prose. If the length-matched arm reproduces W1C1, the effect is capacity, not construction, and the paper reports that instead.
9. **A positive interaction does not identify the mechanism.** §11.4’s serial null makes this testable rather than conceded: two factors acting in series on availability and on use produce superadditivity with nothing jointly repaired. The identifier-survival mechanism is **removed from the confirmatory claims** of this leg — it stays a hypothesis that H3 probes descriptively. This is a change to the thesis, not a caveat on it, and the README carries the same change.
 $\Rightarrow$ **Pre-specified consequence**: the criterion effect is reported as the **bracket [+8.1 · +21.2] pp across judge families**, never as one family’s figure, unless a human pass picks between them. The human validation of §7 must therefore reach the **primary contrast**, and its job is now stated precisely: collapse this bracket.
10. [defect] **The factorial does not identify the mechanism.** The prompt changes identifier survival *and* unit count, length, selection and retrievability together. A positive interaction does not establish that identifier survival governs the effect, so the mechanism claim is not among the confirmatory conclusions.
11. [defect] **Pointwise equality of accuracy does not refute a retention claim.** The README notes this and then presents the result as a refutation of 2601.00821’s prediction anyway. The defensible claim is narrower: our intervention closes the *accuracy* gap under native budget on our corpus, while verbatim retention remains unreached ($12.9\% \neq 100\%$).

11.6 What the production leg carries, and only it

Unchanged from §10’s recommendation, and deliberately modest: **that the window occurs naturally, and its magnitude**. Observational, 12 sessions with window variance of which 8 are used, reported as an existence result with a magnitude (**+21.2 pp total effect** on 56 pairs, of which **+13.0 pp** is interaction) and with its limits stated — exactly what 12 sessions can carry and no more. The confirmatory claim lives in §11.4, on data anyone can download.

11.7 What remains before this can be deposited

The adversarial pass (Codex, `gpt-6-astra`, `exit`: 0, 155 s, full parecer in `audits/2026-10-27-codex-challenge-s11.md`) raised six findings. Four were measured or corrected in place; two need work that is not textual.

#	finding	state
1	manipulated factor changed meaning — dose in sessions not characters, truncation inverted	[closed] corrected (§11.2); the mechanism-vs-truncation gap now declared, not hidden
2	a production statistic parameterises the public leg $\Rightarrow$ legs not independent in the broad sense	[closed] declared ; the 0.240's denominator and spread are now in §11.2 (n=6, [0.020 .. 0.894])
3	power not demonstrated for the additive contrast; α and scale mismatched; one inverted sentence	[closed] closed — resized twice (§11.4); the primary is the total effect, the interaction a secondary that no n this corpus allows powers
4	decision labels contradicted their own intervals	[closed] rewritten (§11.4); 8 pp demoted to a power target
5	both controls pass for a constant NAO SEI	[limit] PARTIAL — two legs added (§11.3), but the answerable floor is still a word, not a number . See the row below
6	independence asserted; four claims not derived	[closed] measured (§11.2: 2.0% evidence-coupled $\Rightarrow$ cluster-robust SE) and the four claims withdrawn or narrowed (§11.5)

[limit] **Read this table against the prose below it.** Until 2026-10-27 the table said rows 3 and 5 were open while the paragraph under it announced that both blocking items were closed — a state marker and a claim about the same state, contradicting each other in the same screen. Row 3 is genuinely closed. Row 5 is **not**, and saying so is the correction: the floor's number is owed, and the remaining open item is listed with the others at the end of this section.

The two items that blocked the first pass are closed, 2026-10-27, and each changed the design:

1. [limit] **Resized, then resized again — the first pass fixed one leg of two** (power-additive.py, then estimand-and-power.py). It corrected Var(D) to 0.973 and took n from 52 to 193, and it kept **21.4 pp — the total effect — as the δ of an interaction test**. The second adversarial pass caught that (audits/2026-10-27-codex-and-deepseek-on-the-prereg.md): the interaction is **13.0 pp**, needs **n=522** against a corpus of 500, and is therefore no longer the primary. n=193 would have delivered **36.7%** power for it. The n=60 stage 1 was gravely underpowered either way. **10 pp is not detectable in this corpus** (883 > 500). The stopping rule was rebuilt around n=200 $\rightarrow$ 500 and the 8 pp threshold removed, because it was a power target dressed as a confirmation boundary.
2. [closed] **Denominator stated** (see-once-denominator.py) — and it **dissolved the point anchor: n = 6**, fractions [0.020, 0.030, 0.183, 0.297, 0.422, 0.894], and **0.240 corresponds to no observation**, being the interpolation of the two middle ones. W is now a **three-level factor** at the observed order statistics. The sensitivity arms I had pre-specified (0.12 / 0.48) sat *inside* the observed range and would have missed both ends, where half the band lives.

[key] **Both closures made the study harder, not easier** — $3.7\times$ the n and six cells instead of four. That is the expected direction when a design is checked rather than defended, and it is the second time today this document has moved that way: §4 was blocked by its own sampling code, §11 by a voice and then by the two measurements the voice demanded.

[key] **And the finding that is about me rather than the plan:** §11 was written *after* §4 was found unexecutable, with that failure explicitly in view, and it still shipped a dose in the wrong unit and a truncation pointing the wrong way. Pricing and powering a design before writing it — which §11 did, and which was the lesson from §4 — does not catch an intervention that means something other than what it is named. **Adversarial review and mechanical checks catch disjoint defects:** every mechanical gate I ran on this document passed.

[limit] A second process note, from the same hour. The commit that first marked this document superseded used `if old in s: s = s.replace(...)` instead of an assert — a silent conditional that applied nothing and reported nothing, so the header kept saying something untrue for two commits. Every substitution in this file now asserts its anchor.