SORTED BY DATE: KNOWLEDGE-UPDATE SCORES CANNOT TELL DATES FROM POSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge-update benchmarks are used to compare assistants and measure the benefit of memory updates. We show that a routine choice in evaluation code can undermine both conclusions: presenting evidence chronologically places the current statement after the superseded one. Time reading and position following then earn the same score, even when averaged over time-ordered layouts. We identify this confound in three update settings across six audited memory benchmarks. Paired exchanges of time labels or positions preserve input length and format, yet expose opposite cue preferences among eleven reader settings with similar native scores. Four LongMemEval readers differ by just 1.4 percentage points natively but by 16.9 points when averaged over five random orders; the lowest native scorer ranks first. Placing the most relevant evidence last puts the superseded session last on 27–44% of questions; GPT-4o-mini loses 23 points overall with lexical retrieval. Even an oracle update note shows no significant native gain across five readers. Its gain is 11.7 points larger when the superseded session is last (95% CI [6.3, 17.1]; exploratory paired analysis). Four memory systems extend the diagnosis to arrival order and stored dates. We recommend retaining native accuracy, adding random-order accuracy split by the competing pair's orientation, and using paired exchanges as diagnostics.

1 INTRODUCTION

A user tells an assistant that their charity 5K best is 27:12. A week later they hope to beat their personal best of 25:50. Asked a month later for the current best, the assistant must resolve the update from these two dated conversations. LongMemEval (Wu et al., 2025) scores this as *knowledge update*.

Studies of order effects show that moving evidence can change model answers (Liu et al., 2024). We trace a shortcut to the evaluation pipeline itself. All six benchmarks we audited present full histories chronologically. Three also supply a time cue and score updates: LongMemEval and STALE (Chao et al., 2026) print session dates, and MemoryAgentBench (Hu et al., 2026) numbers its facts in the order it lists them (§3).

For current-state questions, these layouts place the current statement after the superseded one. Position alone thus resolves the choice between them. **The evaluation design makes time reading and position following indistinguishable in score** (Observation 1). High native accuracy cannot establish whether a reader used temporal evidence to resolve the update. The benchmark's ordering rule therefore limits what its scores can justify about reader choice and update benefits.

We expose this ambiguity with paired layouts in which the cues disagree. Exchanging either the time labels or the positions of competing statements preserves length, format, unrelated content and the final slot (Figure 1). Readers with similar native scores then follow opposite cues (Figure 2). Position-only results, an independent answer classifier and a blind human check support this interpretation (§4). Four memory systems extend the diagnosis to arrival order, stored dates and the answering model (§5).

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

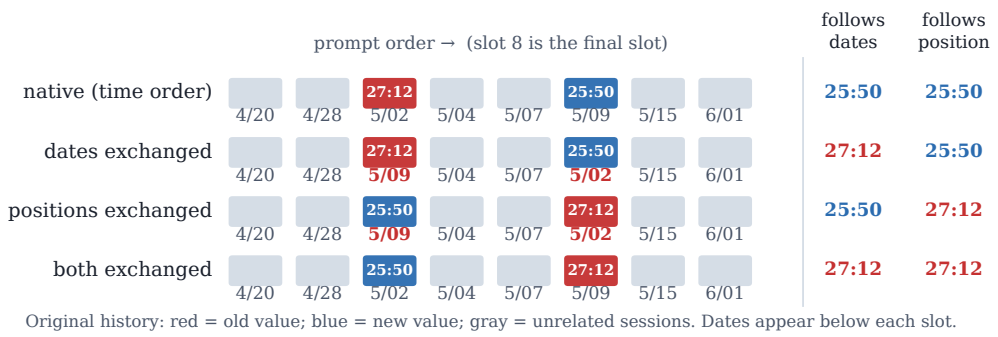


Figure 1: The four cells of Proposition 2 on the example of §1, eight sessions in prompt order. Each cell changes only the two slots of the competing statements; lengths, every other byte and the final slot stay fixed, and the two single-exchange cells carry the same sequence of dates. A reader that follows the dates and one that follows position answer alike natively and oppositely in those two cells. Values on the right are the answers selected by each rule.

These hidden differences have three measurable consequences (§6). *For model comparison*, four LongMemEval readers within 1.4 percentage points under the native layout span 62.0–78.9% mean accuracy over five random orders, a 16.9-point gap. The lowest-scoring reader under the native layout ranks first among these four. *For evidence assembly*, placing the most relevant session next to the question puts the superseded session last on 27–44% of questions; GPT-4o-mini loses 23 points overall under the lexical ranker. *For measuring update benefits*, an oracle note naming the superseded session gains 11.7 points more when that session is placed last than under the native layout (paired within question, five readers; exploratory).

These consequences motivate a reporting protocol: retain native accuracy, add random-order accuracy split by pair orientation, and use the paired exchanges as diagnostics (§7, Table 3).

Our contributions follow this chain:

C1 **Identify the evaluation confound.** We audit six evaluation pipelines and identify three update settings whose ordering rules make time reading and position following indistinguishable in score (§3).

C2 **Expose the hidden behavior.** Paired, format-preserving exchanges separate the two cues across three benchmarks, eleven reader settings and four memory systems (§4–§5).

C3 **Quantify the consequences.** We measure the effects on reader comparisons, exposure to assembly conventions and gains from an oracle update annotation (§6).

C4 **Make the diagnosis usable.** A protocol using existing pair annotations combines native and random-order accuracy with the two diagnostic exchanges (§7).

2 Related Work

Order effects. Readers use context unevenly by position (Liu et al., 2024); in knowledge conflicts, which of two passages a model adopts depends on their order (Xie et al., 2024). Common remedies marginalise order: permute options or candidates and aggregate (Zheng et al., 2024; Tang et al., 2024), or build order-invariant inference (Yoon et al., 2025; Egressy & Stühmer, 2025). Our use of random orders builds on these methods. Order is also chosen deliberately: FreshPrompt sorts search evidence from oldest to newest and reports higher accuracy than with a random order (Vu et al., 2024). A related analysis in video shows that temporal benchmark scores conflate reading event order from the frames with reading it from positional encodings, and models with similar scores fail on opposite inputs (Rahman, 2026). In an update benchmark the canonical order also coincides with the label, making the cue a reader follows unidentifiable and hiding measurable costs. Our two format-preserving exchanges identify that cue.

Confounds in evaluation. Prompt format (Sclar et al., 2024) and harness choices (Yao et al., 2026; Zhang et al., 2026) can move scores more than models do, and in memory evaluation the embedding

Table 1: Evidence order in the released evaluation code of six memory benchmarks. *aligned*: the later statement in time is also later in the prompt on every question; *sorted*: aligned by an explicit date sort; *score*: ordered by retriever score, which does not align.

Benchmark	Update construct	Cue shown to reader	Full context	Retrieval
LongMemEval (Wu et al., 2025)	knowledge update	session dates	sorted	sorted
STALE (Chao et al., 2026)	invalidated memory	session timestamps	aligned	not shipped
MemoryAgentBench (Hu et al., 2026)	conflict resolution	serial numbers	aligned	score
BEAM (Tavakoli et al., 2026)	knowledge update	dialogue order	aligned	score
AMemGym (Cheng et al., 2026)	evolving state	none	sorted	sorted
LoCoMo (Maharana et al., 2024)	none (temporal)	session dates	aligned	score

model, reader and retrieval pipeline confound reported gains (Wang, 2026a). A model that scores well from part of the input exposes a task solvable without the intended capability (Gururangan et al., 2018; Poliak et al., 2018); the broader failure is leakage (Kaufman et al., 2012; Kapoor & Narayanan, 2023), which for LLM benchmarks appears as question-only classifiers (Pacchiardi et al., 2024) and leaderboard-topping constant outputs (Zheng et al., 2025). A score that survives an input ablation need not be independent of the removed information (Feng et al., 2019), so we cross deletion with a contrast on the layout (Gardner et al., 2020). Where a model reaches its score through a cue the data leaves standing (McCoy et al., 2019; Geirhos et al., 2020), the way to find out is to set two cues against each other, as texture against shape (Geirhos et al., 2019); our two exchanges do that to a layout rather than to an item. We trace this shortcut to the ordering imposed by evaluation code.

Updates, time and memory. Long-term memory benchmarks evaluate recall across conversations (Wu et al., 2025; Maharana et al., 2024). Explicit update and conflict tasks appear in LongMemEval (Wu et al., 2025), supersession benchmarks (Chao et al., 2026), agent-memory suites (Hu et al., 2026; Tavakoli et al., 2026; Cheng et al., 2026), memory-system diagnostics (Tao et al., 2026; Patel, 2026), dynamic state tracking (Wu et al., 2026), and temporal reading studies (Özer & Yıldız, 2025; Wallat et al., 2026). EvolveBench (Zhu et al., 2025) pairs a query date with a single misaligned passage and evolveQA (Nakshatri et al., 2025) probes parametric knowledge; neither shows competing dated statements in one layout. Memory systems extract and update stored facts (Chhikara et al., 2025), organise linked notes (Xu et al., 2025), keep a temporal knowledge graph that invalidates facts by time (Rasmussen et al., 2025; Wang, 2026b), index events by time (Sen et al., 2026), or resolve conflicts by serial number after retrieval (Reddy & Challaram, 2026). Pooled accuracy hides their distinct failures (Long et al., 2026; Shi et al., 2026), and probing the state a memory holds brings those failures out (Ma et al., 2026). On evaluation itself, Bean et al. (2025) review construct validity and Akhtar et al. (2026) saturation. Our focus is the link between an evaluation pipeline's ordering rule, the ambiguity of its scores and the conclusions drawn about readers and update methods.

3 THE SORT, AND WHAT IT DOES TO THE SCORE

We read the released evaluation code of six memory benchmarks at pinned commits and traced the order in which evidence reaches the reader (Table 1; the file and line for every entry are in Table 6). Every full-context layout is chronological, the natural way to present a conversation, and LongMemEval's generation script goes further and sorts by date after all six of its context modes.[1] The argument below applies where a benchmark both prints a time cue and scores an update: LongMemEval and STALE print dates, and MemoryAgentBench's conflict-resolution task numbers its facts and states that the newer fact has the larger number, then lists them in number order. In BEAM and AMemGym no cue is printed, so order is the intended cue; LoCoMo scores no update.

LongMemEval assigns 78 of its 500 questions (15.6%) to knowledge update. The category includes both current-state and historical questions. Our formal argument below concerns current-state questions; the experiments retain the benchmark category and check the narrower case separately. A

[1] `src/generation/run_generation.py`, lines 224–225, sort `retrieved_chunks` on their first field, the session date. The README states that the S and M haystacks are sorted by timestamp; the cleaned S file is not (211 of its 500 questions contain inversions, 3,167 in all), but the script sorts after retrieval.

chronological layout is a reasonable presentation choice, but its score cannot establish which cue resolved the update. Relevance-based assembly can change that layout without changing the evidence (§6).

Identifiability. For a current-state question, fix competing statements o (superseded) and n (current), and everything else in the prompt, z. Two cues bear on which is current: the *time cue* $c \in \{n, o\}$, the statement whose printed time label is later, and the *position cue* $p \in \{n, o\}$, the statement placed later. A reader chooses n with probability $q(z, c, p)$. A layout is *aligned* if $p = c$; a chronological native layout has $c = p = n$. The native reference answer is n.

Observation 1. *Under any family of native aligned layouts ($c = p = n$), and any average over that family, accuracy depends on the reader only through $q(\cdot, n, n)$. The time rule $q = \mathbb{I}[c = n]$ and the position rule $q = \mathbb{I}[p = n]$ are therefore indistinguishable, however many aligned layouts are sampled.*

An aligned native layout never evaluates q off the diagonal $c = p$. Deleting the time labels under an aligned layout evaluates $q(z, \varnothing, n)$, so a reader that falls back on position loses nothing there even if it follows the labels whenever position misleads (§6). Under a uniformly random order, p equals n with probability $\frac{1}{2}$ independently of the item, so the position rule scores $\frac{1}{2}$ on the choice and the time rule 1.

The *native layout* is the benchmark's own order. The *dates-exchanged* version swaps the pair's time labels (dates or serial numbers); *positions-exchanged* swaps their slots, each with its own label; *both-exchanged* swaps their bodies, so both cues point to the superseded statement (Figure 1). Each version preserves prompt length and all bytes outside the two slots; session headers remain numbered by position.

Every version is scored against the *original* reference answer. Exchanging dates can change which answer the chronology supports; fixing the reference diagnoses answer dependence rather than correctness under the new dates. The *position preference* is the dates-exchanged score minus the positions-exchanged score: positive values favour position, negative values the time cue. This behavioral contrast does not identify an internal mechanism.

Proposition 2. *Write the item-averaged choice rate in the four cells $(c, p) \in \{n, o\}^2$ as $\beta_0 + \beta_d \mathbb{I}[c = n] + \beta_p \mathbb{I}[p = n] + \beta_{dp} \mathbb{I}[c = n]\mathbb{I}[p = n]$, a saturated model. Exchanging both cues ($c = p = o$), only the time labels ($c = o, p = n$), only the positions ($c = n, p = o$), and neither gives β_0, $\beta_0 + \beta_p$, $\beta_0 + \beta_d$ and $\beta_0 + \beta_d + \beta_p + \beta_{dp}$, so all four weights are identified. Without the both-exchanged cell, the difference of the time-exchanged and position-exchanged cells still identifies $\beta_p - \beta_d$, the position preference.*

The interaction β_{dp} depends on the scale (it vanishes for a reader that mixes a time policy, a position policy and content, and is negative near the ceiling for one whose log-odds add the cues), so we read it only where the probability and logit scales agree (Appendix K).

The *inversion control* swaps two unrelated sessions' labels, leaving the pair untouched. *Superseded-last* moves the superseded session to the final slot, a third cue held fixed in the exchanges; its *matched control* moves an unrelated session there. A *random order* is a uniform permutation that consults no annotation.

4 Time labels against position, with nothing deleted

We apply the exchanges to three benchmarks to test whether similar native scores conceal different cue preferences.

Items. On LongMemEval-S (cleaned release), 70 answerable knowledge-update questions have both pair members among the eight sessions retrieved under the official BM25 contract. The content-only rankers and their date-independence check are in Appendix B. Keeping the final slot fixed leaves 49 questions for the paired exchanges. Excluding six with an earlier-session gold value and two without a single competing-value choice leaves 41; all seven Holm-resolved directions remain (Appendix L; exploratory). We also test full histories (112k–134k tokens). On Stale, we use the first 190 of 291 eligible items in a seeded random order, with retrieval fixed. On

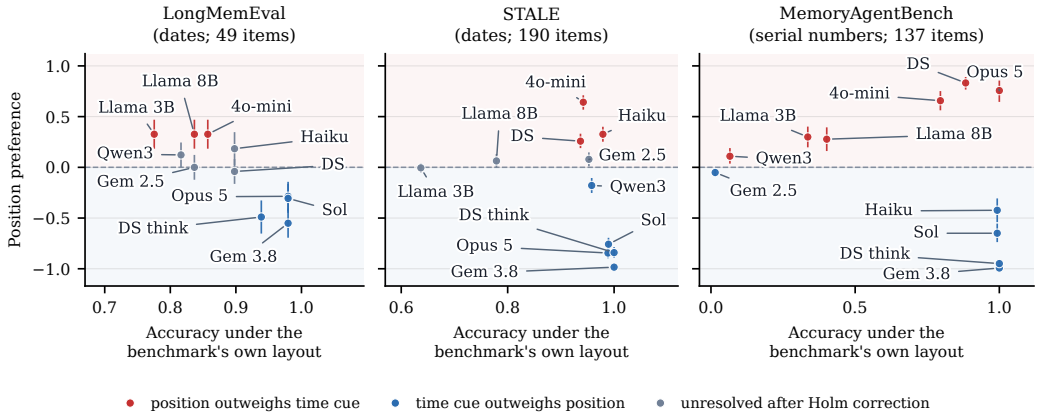


Figure 2: Readers with similar native scores follow opposite cues. Native accuracy (x) against *position preference* (y): the original-reference score with time labels exchanged minus that with positions exchanged (95% intervals). Each label connects to its own point. Red: position outweighs the time cue; blue: the time cue outweighs position; grey: not resolved after Holm correction. Table 28 gives the values; Haiku on MemoryAgentBench uses its 200-token re-run.

MemoryAgentBench, 137 single-hop conflict questions at 6k and 32k tokens contain one earlier and one edited fact (Appendix K).

Readers, judging and statistics. Eleven settings (Table 27): Claude Opus 5, GPT-5.6 Sol, Gemini 3.8 Flash, Claude Haiku 4.5, Gemini 2.5 Flash, Qwen3-235B, GPT-4o-mini, DeepSeek-Flash with and without thinking, Llama-3.1-8B and Llama-3.2-3B, each through one pinned provider at temperature 0. LongMemEval answers are scored by its official judge (`gpt-4o-2024-08-06`, the benchmark's per-category prompts), which sees the question, the gold answer and the response but never the history; STALE is scored from the emitted option letter and MemoryAgentBench by its own substring match. Intervals are percentiles of an item-cluster bootstrap (30,000 replicates, one resampling shared by the cells a contrast compares). A two-sided p is twice the smaller tail of the resampled differences, capped at 1. Holm's step-down correction at $\alpha = 0.05$ runs within a family: the settings of one benchmark, or the reader pairs of one layout. An interval excludes zero only if both bounds lie strictly on one side, and differences are computed before rounding. The predictions of every experiment were written into protocol documents before the runs they govern, with deviations logged (Table 4); analyses added later are labelled exploratory.

Similar scores, opposite cues. On STALE, GPT-4o-mini and GPT-5.6 Sol both sit near the ceiling in the benchmark's layout (0.942 and 0.989), yet their position preferences are $+0.642$ and -0.758. Exchanging the pair's positions sends most of GPT-4o-mini's answers to the superseded statement; exchanging the pair's dates does the same to Sol's. Native accuracy does not reveal the direction (Figure 2, Table 28).

After Holm correction within each benchmark, 4 settings follow the dates and 3 follow position on LongMemEval, 5 and 3 on STALE, and 5 and 6 on MemoryAgentBench. Excluding 2 MemoryAgentBench settings that answer mostly from world knowledge gives 4 and 5. These counts use Haiku 4.5's exploratory re-run; its pre-registered row gives 4 and 7. Applying the re-run's final-answer rule to every setting, instead of the harness's match, preserves all 11 signs (Table 36).

Checks of the cue interpretation. An exchanged date can contradict a relative expression ("last month"). The positions-exchanged version moves each session with its own date, avoiding this inconsistency, and separates the readers on its own (exploratory). On LongMemEval and STALE, the settings that follow the time cue lose at most 13.2 points there, and those that follow position at least 34.7 (Table 34). On MemoryAgentBench the cue is a serial number, avoiding this date–text inconsistency, and a setting resolved on it and on LongMemEval follows the same cue in 6 of 8 cases. Relative expressions occur in 35 of the 49 LongMemEval pairs and 112 of the 190 STALE pairs; on the rest, no resolved direction reverses (Table 37).

Table 2: Position preference under two within-model settings (95% intervals). Positive values favour position; negative values favour the time cue. DeepSeek: thinking off versus on, with the same weights and host. Haiku: value-only versus a 200-token answer budget that allows a written comparison, an exploratory prompt contrast.

Reader	Benchmark	Off / value only	On / 200-token budget
DeepSeek-Flash	STALE	$+0.258$ $[+0.189, +0.332]$	-0.841 $[-0.892, -0.786]$
DeepSeek-Flash	MemoryAgentBench	$+0.832$ $[+0.766, +0.891]$	-0.949 $[-0.985, -0.912]$
Claude Haiku 4.5	MemoryAgentBench	$+0.453$ $[+0.343, +0.555]$	-0.423 $[-0.540, -0.307]$

An exploratory judge check uses a classifier from another model family on LongMemEval: it preserves every preference sign the official judge resolves (8 of 8, Table 35). Among the 390 answers whose verdicts differ between the two versions, which determine the preference, the authors' blind labels of 60 agree with the judge on 59.

Cue preference can change within a model. The three settings that follow the time cue on all three benchmarks are the only ones that emit reasoning tokens (Table 38), and two within-model contrasts show that reasoning can move the cue (Table 2). Turning DeepSeek-Flash's thinking on takes it from position to the time cue with the same weights and host. Claude Haiku 4.5 emits no reasoning tokens, but with 200 answer tokens on MemoryAgentBench it writes the comparison of serial numbers into the answer itself, citing a fact number in 371 of 411 answers, and follows the serial numbers. Asked in the same cells for the value alone, it answers in 5 tokens and follows position. That is a within-item change of -0.876 $[-1.022, -0.730]$, in the direction we wrote down before the run (exploratory). The Haiku comparison changes the prompt as well as the visible reasoning; the DeepSeek comparison holds the prompt fixed. With explicit reasoning disabled, Claude Opus 5 follows the dates on LongMemEval (-0.286) and STALE (-0.845), but position on MemoryAgentBench ($+0.757$), although that benchmark's instructions state that the newer fact has the larger number. Settings resolved on LongMemEval and another benchmark follow the same cue in 16 of 18 cases. The 2 exceptions move opposite ways on MemoryAgentBench: Opus 5 to position, Haiku 4.5 to the serial numbers. With the full history (112k–134k tokens), every setting resolved there keeps its direction, while GPT-5.6 Sol's preference (-0.143 $[-0.286, +0.000]$) no longer excludes zero.

The exchanged versions have the same date sequence, controlling the visible timeline inversion. Swapping dates of unrelated sessions provides a separate inversion control; it changes accuracy by at most 2.0 points on LongMemEval and 5.8 on STALE. When positions are exchanged, 79%, 80% and 96% of the wrong answers from the three strongest position followers name the competing value. The both-exchanged cell and all pre-specified predictions are reported in Tables 33 and 42: 5 predictions held, 7 failed and 1 was untested. Failures include cross-benchmark exceptions, a full-history interval reaching zero and the Mem0 1.0 results below.

5 MEMORY SYSTEMS THAT INGEST THE HISTORY

The ambiguity can also arise during ingestion. We separate arrival order from stored dates in Mem0 1.0's extract-and-update pipeline (Chhikara et al., 2025), Mem0 2.0's additive pipeline, A-Mem's linked notes (Xu et al., 2025) and Graphiti, Zep's temporal knowledge graph (Rasmussen et al., 2025). Mem0 and A-Mem ingested the full histories of the 70 LongMemEval questions (3,336 sessions); Graphiti ran only the 49 conflict questions. Each writes with GPT-4o-mini and receives dates through its own channel. Retrieval uses a ten-item budget (Graphiti: up to ten facts and ten entity summaries); a fixed answer prompt lists retrieved items with their stored dates (Appendix K). Besides date order (C0) and a random order (C1), the exchanged versions become date order with the pair's arrival slots exchanged (C2: the superseded session arrives later, dates true) or with its dates exchanged (C3). Graphiti omits C1.

In date order the systems score 0.571 (Mem0 1.0), 0.857 (Mem0 2.0), 0.490 (A-Mem) and 0.122 (Graphiti) on the conflict questions (Figure 3). *A-Mem* ranks notes by relevance, which need not reflect recency. GPT-4o-mini, a position follower, answers 0.490 from those notes, while GPT-5.6 Sol and Gemini 3.8 Flash answer 0.837 and 0.898 and follow the exchanged dates (-0.265 and -0.490). *Mem0 2.0*'s store contains the gold value on 41/47 questions in date order and 41/47

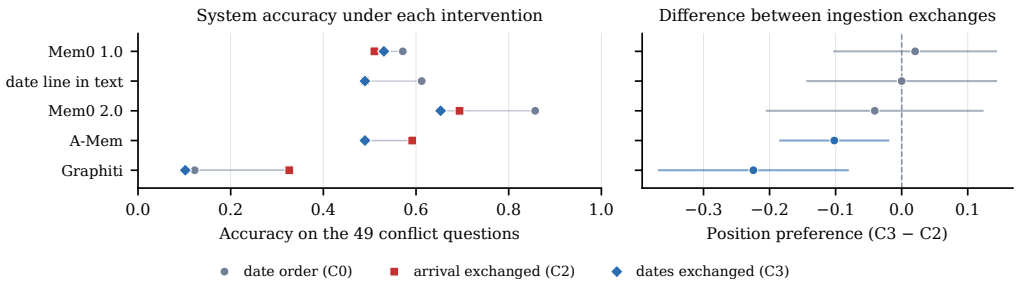


Figure 3: Memory systems on LongMemEval, answered by GPT-4o-mini and scored by the official judge. *Left:* accuracy on the 49 conflict questions ingested in date order (C0), with the pair's arrival slots exchanged (C2) and with its dates exchanged (C3). *Right:* the position preference, C3 − C2, with 95% intervals (grey: includes zero). Storage, retrieval and answering can all affect this contrast; read it alongside the cell scores. Table 39 gives random-order costs and other answerers.

with the arrival exchanged. Its score drops by 16.3 points when the older session arrives later and 20.4 when the dates are exchanged; all three answerers lose accuracy in both conditions. Its near-zero preference (-0.041) therefore hides two losses of similar size. Among questions where the gold value is retrieved in both conditions, GPT-4o-mini falls from $37/40$ to $30/40$ under the arrival exchange and from $38/41$ to $30/41$ under the date exchange. Answer quality falls even with the gold value available; the retrieved context and its ordering can still contribute (Appendix L). *Mem0 1.0*'s update step rarely acts: 1.4% of its operations were updates or deletions. Its store holds the gold value on 31/47 questions in date order and 26/46 with the arrival exchanged, and its search returns it on 31/47 and 25/47. Neither exchange moves its accuracy much ($+0.020$ $[-0.102, +0.143]$), and nor does printing each date into its input ($+0.000$). *Graphiti*, in its default configuration, stayed near the floor. Its search returned a median of 2 of the 70 facts stored, and the gold value on 8/47 questions; accuracy rises only when the superseded session arrives later (0.327), when its store keeps the gold value more often.

Three pre-specified predictions about Mem0 1.0 failed because its update step rarely acted (Table 42). A random order of ingestion costs Mem0 2.0 11.0 points over all 70 questions, and no more where the pair arrives reversed (5.5) than where it keeps its order (12.0). This pattern does not concentrate the loss on reversed pairs. That is the split §7 recommends, applied to ingestion.

6 What the layout costs

The following experiments quantify the blind spot's consequences for reader comparisons, evidence assembly and measured update benefits. Date deletion checks why a standard ablation can also miss the dependence.

Reader comparisons. The native layout resolves few differences between readers. Among 7 readers tested under three layouts, Holm-corrected comparisons separate 1 of 21 pairs natively, 9 under one random order per question and 12 under the superseded-last layout (Figure 9). In the original five-reader comparison, the five-order mean separates 8 of 10 pairs against 0 natively. Each draw's ranking agrees with the mean of the other four (Kendall's τ_b, accounting for ties, 0.8–1.0), while the native layout does not ($\tau_b = 0.120$). Across 9 readers with five officially judged orders, accuracy has s.d. at most 0.066, against a native-to-random gap of up to 23.7 points (Figure 4, Table 22).

A random order assigns each question to one of two groups, the pair's order preserved or reversed. In the five-reader comparison, accuracy falls by 27.5 points against the native layout where the pair was reversed and 5.0 where it was preserved (difference 22.4 points [15.4, 29.3]). The reversed questions alone separate 7 of 10 pairs against 3 for the preserved ones. The additional separation is concentrated on questions whose pair order was reversed. On MemoryAgentBench, pooled over the eleven settings, a random order lowers accuracy on the conflict items 10.1 points more than on questions without a conflicting fact (-0.101 $[-0.133, -0.069]$).

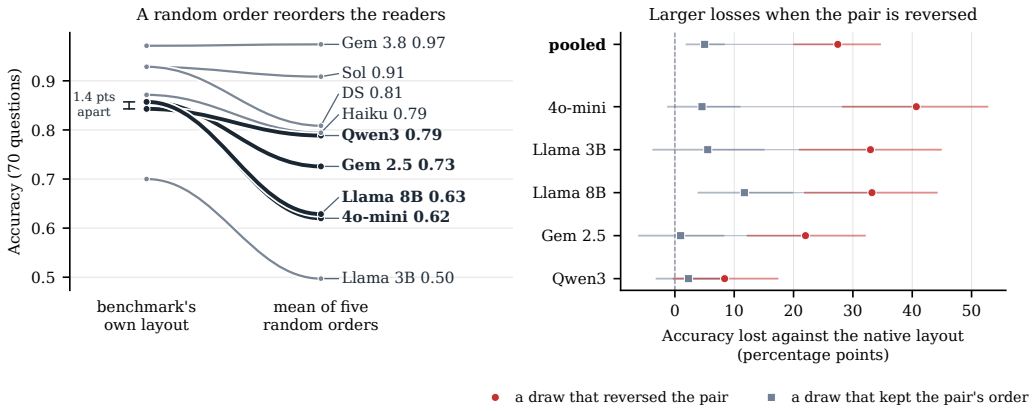


Figure 4: LongMemEval, 70 questions, official judge. *Left:* native accuracy against the five-order random mean; dark lines mark 4 readers within 1.4 points natively, whose order changes. *Right:* native-to-random accuracy loss for the five-reader comparison, averaged within question over draws reversing or preserving the pair (68 and 69 questions; 67 in both groups), with 95% intervals.

Assembly conventions. Retrieval conventions break the alignment. Pipelines often place the strongest match next to the question (Liu et al., 2024). Placing the most relevant session last puts the superseded session in the final slot on 27–44% of questions, depending on the ranker. With the lexical ranker, GPT-4o-mini loses 23 points overall; ordering most relevant first never puts the superseded session last and yields no resolved change (Table 7). Moving the superseded session last lowers accuracy for all 8 LongMemEval readers of the earlier runs, by 7.1 to 48.6 points, while moving an unrelated session there changes it by at most 5.7, every interval spanning zero. On five STALE panels the effect is 9.3 to 51.5 points, against a control 7 to 26 times smaller (Figure 7). The final slot is a cue of its own, if a small one for readers that follow the dates: pooled over the four such settings, the pair exchanged in place beats the superseded session moved last by $+0.020$ $[+0.005, +0.041]$.

Measured update benefits. An oracle note identifies the current and superseded sessions, providing an ideal update signal without extraction errors. Under the native layout its gain ranges from -2.9 to $+2.9$ points across five readers, with all intervals including zero. With the superseded session last, gains range from 4.3 to 24.3 points. The direct paired comparison is the relevant test: the note gains 11.7 points more under the latter layout $[6.3, 17.1]$, pooled over the five readers (Table 12; exploratory). Thus, evidence order changes how much benefit even an ideal update signal can demonstrate. Learned update methods may also incur extraction errors. Since knowledge update is 15.6% of LongMemEval, Llama-3.1-8B's gap under one random order would change the overall score by about 4.2 points if other categories stayed fixed.

Date deletion. A standard ablation can miss this dependence. A reader can use dates when position misleads yet lose nothing when dates are removed from an aligned layout (Observation 1). On LongMemEval (70 questions) the dates are worth 8.6 to 25.7 points to 5 of 7 readers once the superseded session is placed last, where the two cues disagree. Deleting them from the native layout costs the same readers at most 5.7 points (Table 13). Claude Haiku 4.5 loses nothing to deletion natively (0.871 either way) but gains $+0.257$ $[+0.143, +0.371]$ once position misleads; Llama-3.2-3B does the reverse, so a deletion test alone would credit it with date use (Feng et al., 2019). The readers that follow the dates show the pattern most clearly: the dates are worth -1.4 points to GPT-5.6 Sol and -1.4 to Gemini 3.8 Flash natively, but 38.6 and 20.0 with the superseded session last (Table 41).

7 RECOMMENDATIONS FOR UPDATE BENCHMARKS

Keep the native score: it measures performance on sorted input. Add accuracy under a uniform random order, which removes the association between relative position and recency. Report it separately for items whose pair order is preserved and reversed. This random split distinguishes losses

Table 3: A minimal report for an update benchmark, using the existing pair annotation. The random order adds one pass; each diagnostic version adds a pass over the annotated items.

Report	How	What it shows
Native accuracy	the benchmark's own layout	performance on sorted input
Random-order accuracy, split in two	one uniform permutation per item, reported separately where the pair arrives reversed and where it keeps its order	the loss with the pair order preserved and additional loss when it is reversed
Two exchanged versions	swap the pair's time labels; swap its slots, each with its label	the position preference: which cue the answers track
Optional	both exchanged; inversion control; superseded session last	the four weights; the cost of an out-of-order timeline; exposure to assembly conventions
For memory systems	ingest in a random order with true time stamps; exchange the pair's arrival slots or its stamps	the same questions, asked of ingestion rather than layout

shared by both groups from losses concentrated on reversed pairs (Table 3). Random order is an evaluation condition; deployment can retain its preferred assembly rule.

The two exchanged versions use the pair annotation the benchmark already carries, change neither length nor format, and identify the position preference; the both-exchanged version identifies each weight (Proposition 2), and the inversion control checks sensitivity to an unrelated timeline inversion. Where a harness keeps only its top-k items, exchange inside that set, as with the eight sessions here. For memory systems the analogues concern how a static log is replayed into the system; they do not ask an interactive agent to experience events out of order.

The same check applies when an evaluation pipeline orders, groups or truncates evidence in a way that reveals the target label. The check is a partial-input baseline (Gururangan et al., 2018) applied to evaluation code: use the benchmark's own annotation to find the competing statements, apply the rule the layout licenses (here, answer from the later-placed one), and score it. A final-session-only baseline misses this shortcut: Qwen3-235B scores 0.257 against 0.843 with the full context (Appendix D), because that slot often holds an unrelated session.

8 Limitations and conclusion

Scope. The argument applies where a benchmark prints a time cue and presents evidence in time order; where order is the intended cue (BEAM, AMemGym) there is nothing to separate. Three benchmarks qualify, all English; other domains, languages and modalities are untested, and the conflict items are few (49 on LongMemEval), so per-setting intervals are wide. *Exchanged labels.* Only 14 LongMemEval items are free of relative time expressions and the date-following preferences shrink on them, so the separation there rests on the positions-exchanged version and on MemoryAgentBench's serial numbers, which avoid date–text inconsistencies. *Readers.* Each setting is one model version through one provider, run in September 2026; the settings that follow the time cue everywhere all reason, reasoning was manipulated within a model only twice, and we did not test MemoryAgentBench without its instruction that the larger number is newer. *Systems.* Four systems ran with their defaults, one writer (GPT-4o-mini), one answer prompt, fixed retrieval budgets and a local embedding model; Graphiti ran on an embedded database, on the conflict questions only, and its low accuracy there does not measure Zep's hosted service. *Judging.* The official judge shares a family with GPT-4o-mini but never sees the history; a classifier from another family agrees with it on 99.0% of answers and the authors' blind labels on 59 of 60. *Predictions* were written before the runs and deviations logged, though with no third-party timestamp. The preferred deployment policy depends on the task and reliability of its time labels.

Time order makes date reading and position following indistinguishable to the native score. Paired exchanges expose the hidden behavioral differences; the experiments show their consequences for reader comparisons, evidence assembly and the measured benefit of update annotations. Reporting native and random-order accuracy together, with the exchanges as a diagnostic, makes these dependencies visible while retaining the benchmark's original evaluation.

Reproducibility statement

The supplement contains the manuscript sources, recorded responses and verdicts, analysis scripts, protocols and deviation log. Table 26 links quantitative results to their artifacts. Builders and manifests specify the omitted large prompt files and their expected hashes. The paper compiles from the archive without model calls; the added checks can also be recomputed from its recorded responses. Model identifiers, providers, decoding settings and available token and cost metadata accompany the runs. File hashes and numerical checks support the manual review of the paper and its figures.

Use of large language models

LLMs serve as readers, memory-system backbones, the official judge and the answer classifier; their settings are listed in Appendices B and K. An LLM coding assistant also helped draft protocols, implement and run experiments, analyse results, check literature, generate figures and revise prose under the authors' direction. The authors set the research question and scope, approved protocols and experiment choices, reviewed results and revised the manuscript. They checked the 70 value pairs, supplied the blind labels for the 60-answer judge check, and take responsibility for the paper and any errors.

References

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, et al. When AI benchmarks plateau: A systematic study of benchmark saturation. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. In *International Conference on Learning Representations*, 2026.

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

Beni Egressy and Jan Stühmer. Set-LLM: A permutation-invariant LLM. *arXiv preprint arXiv:2505.15433*, 2025.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 2012.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Xianxuan Long, Zhikai Chen, Shenglai Zeng, Shouren Wang, Kai Guo, and Jiliang Tang. MemTrace: Probing what final accuracy misses in long-term memory. *arXiv preprint arXiv:2606.17328*, 2026.

Enze Ma, Yufan Zhou, Wei-Chieh Huang, Jie Yang, Huanhuan Ma, Zixuan Wang, Chengze Li, Chunyu Miao, Philip S. Yu, and Zhen Wang. MEMPROBE: Probing long-term agent memory via hidden user-state recovery. *arXiv preprint arXiv:2606.24595*, 2026.

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

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Nishanth Sridhar Nakshatri, Shamik Roy, Manoj Ghuhan Arivazhagan, Hanhan Zhou, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiah. When facts change: Probing LLMs on evolving knowledge with evolveQA. *arXiv preprint arXiv:2510.19172*, 2025.

Atahan Özer and Çağatay Yıldız. Question answering under temporal conflict: Evaluating and organizing evolving knowledge with LLMs. *arXiv preprint arXiv:2506.07270*, 2025.

Lorenzo Pacchiardi, Marko Tesic, Lucy G. Cheke, and José Hernández-Orallo. Leaving the barn door open for clever hans: Simple features predict LLM benchmark answers. *arXiv preprint arXiv:2410.11672*, 2024.

Vedant Patel. Supersede: Diagnosing and training the memory-update gap in LLM agents. *arXiv preprint arXiv:2606.27472*, 2026.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018.

Farrukh Rahman. What does a temporal benchmark score measure? decomposing channel use in video VLM evaluation. *arXiv preprint arXiv:2607.12304*, 2026.

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

Vikas Reddy and Sumanth Reddy Challaram. Reliable post-retrieval assembly for agent memory: Separating evidence extraction from policy execution. *arXiv preprint arXiv:2606.01435*, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024.

Sahil Sen, Elias Lumer, Anmol Gulati, and Vamse Kumar Subbiah. Chronos: Temporal-aware conversational agents with structured event retrieval for long-term memory. *arXiv preprint arXiv:2603.16862*, 2026.

Zitong Shi, Yixuan Tang, and Anthony Kum Hoe Tung. A-TMA: Decoupling state-aware memory failures in long-term agent memory. *arXiv preprint arXiv:2607.01935*, 2026.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. MemConflict: Evaluating long-term memory systems under memory conflicts. *arXiv preprint arXiv:2605.20926*, 2026.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in LLMs. In *International Conference on Learning Representations*, 2026.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024.

Jonas Wallat, Wolfgang Nejdl, and Sandipan Sikdar. When facts change: Temporal knowledge conflict resolution in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2154–2184, 2026.

Kuan Wang. MemDelta: Controlled baselines and hidden confounds in agent memory evaluation. *arXiv preprint arXiv:2606.29914*, 2026a.

Ziming Wang. TOKI: A bitemporal operator algebra for contradiction resolution in LLM-agent persistent memory. *arXiv preprint arXiv:2606.06240*, 2026b.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813. arXiv:2410.10813.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues. *arXiv preprint arXiv:2605.12493*, 2026.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, 2025.

Yilun Yao, Xinyu Tan, Chao-Hsuan Liu, Yaoming Li, et al. Harness-bench: Measuring harness effects across models in realistic agent workflows. *arXiv preprint arXiv:2605.27922*, 2026.

Soyoung Yoon, Dongha Ahn, Youngwon Lee, Minkyu Jung, HyungJoo Jang, and Seung-won Hwang. RoToR: Towards more reliable responses for order-invariant inputs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Yunbei Zhang, Janet Wang, Yingqiang Ge, Weijie Xu, Jihun Hamm, and Chandan K. Reddy. Stop comparing LLM agents without disclosing the harness. *arXiv preprint arXiv:2605.23950*, 2026.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic LLM benchmarks: Null models achieve high win rates. In *International Conference on Learning Representations*, 2025.

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yuhao Wang, Yunfeng Guan, Yanfeng Wang, and Yu Wang. EvolveBench: A comprehensive benchmark for assessing temporal awareness in LLMs on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16173–16188, 2025.

A Deviations from protocol

Every deviation was logged when found, with the fix applied in the same step and numbers re-derived from artifacts. This table keeps the events that changed a reported number, a claim or the design. The complete log (81 entries, PE-1 to PE-96), including pipeline defects caught before any number was reported and editorial corrections, is in the supplementary material; identifiers are those of the full log.

Table 4: Deviations from protocol that changed a reported number, a claim or the design.

ID	Deviation and consequence
PE-21	Withdrawn: retrieval saturation was attributed to the benchmark's 2025 cleaning. The pre-cleaning release reproduces all 32 outcome cells, so saturation predates it.
PE-24	The first SLOTSHIFT generator was invalid: a shared random stream made length conditions differ in content, and statements took their dates from their slots. It was rebuilt with its guarantees checked on its artifacts; earlier SLOTSHIFT numbers are withdrawn.
PE-27, PE-43	Two claims were withdrawn as unsupported: that readers represent position as time, and that date sorting and order-invariant inference are mutually exclusive.
PE-33	The date-comparison instruction for the reasoning-capable reader was first reported on 39 questions; completed to 70 its interval reaches zero, and the claim changed from no effect to strong attenuation.
PE-46, PE-48	A provider outage cut the GPT-4o-mini STALE panel at 57 items and forced a substitute judge on arms added meanwhile. The panel was completed to 400 items and every substitute-judged LongMemEval row was re-scored by the official judge; no reported LongMemEval number rests on the substitute judge (agreement 0.986 over 427 rows).
PE-49	A reasoning reader on STALE returned unparseable answers far more often under the superseded-last layout; dropping them would have conditioned on easy items, so the pass was discarded and re-run with a larger output budget.
PE-51, 54, 58	Numbers that had drifted from their artifacts were corrected, among them two strand rates that no artifact produced; every number in the prose is now generated from artifacts and audited against the compiled paper.
PE-53	LoCoMo category labels were first taken from the order its paper introduces question types rather than from the code's `category` field, which inverted a conclusion; LoCoMo is now a manipulation check (Appendix F).
PE-60	Two cells of the crossing's protocol were first reduced for budget, by amendments recorded before those cells were analysed (random draws 2–5 for two readers, full history beyond 55 questions). Both were completed later (PE-73).
PE-61	The permutation marginal first paired substitute-judged permutation rows with an official-judged comparator; all rows are now official-judged.
PE-68	The STALE DeepSeek panels ran with thinking on: the runner sent a reasoning switch that the provider's API ignores. The specification had said thinking was off for every DeepSeek reader; it is off only on LongMemEval. No number changed.

continued on the next page

ID	Deviation and consequence
PE-69	The random-order arm file labelled each draw's sessions by permuting the retrieval-ranked id list, while the prompts permute the sessions in their native order. Prompts, outputs and verdicts are unaffected; the slot counts once quoted in Appendix B (36 to 51) were wrong, and the labels were rebuilt by exact content match.
PE-70	The provider serves V4.1-Flash under the requested identifier `deepseek-v4-flash` since 10 September 2026, before any DeepSeek call reported here; both names are given.
PE-71	The STALE exchanged versions were reduced to 190 items per thinking state by an amendment written after a discarded three-item dry run and before the main pass. An interim look at 70 items, not provided for in the protocol, changed nothing.
PE-73	A second pre-specified experiment completed the reduced cells, ran the no-date arms for Claude Haiku 4.5 and the exchanged versions on LongMemEval and on more STALE readers. Two amendments preceded the affected calls: the STALE cost cap was raised after the dry runs measured it, and both DeepSeek settings were reduced to the first 40 questions of the seeded order.
PE-77	The unified runs (a protocol with five dated amendments) pinned DeepSeek-Flash to a provider that does not serve this account; it was pinned to Fireworks instead.
PE-78, 87	Passes that reached their spend caps were resumed with added caps. GPT-5.6 Sol's full-history conflict pass, first stopped by its cap, was completed after the other passes; the preference on the questions covered before the resumption is reported beside the final one.
PE-80	One cell the final-slot prediction needs (DeepSeek-Flash with thinking, superseded session last) had never been run and was added before any analysis.
PE-81	Graphiti's search returns facts and entity summaries, as in Zep's own LongMemEval evaluation; for budget it ran on the conflict questions without a random order.
PE-82	The judge snapshot is rate-limited per account; part of the full-history answers were scored by a second host of the same snapshot, and the two hosts' agreement is reported.
PE-83	Memory systems on HaluMem (one prediction) were dropped for budget and time; the prediction is reported as not tested.
PE-84	The first Claude Opus 5 passes left the provider's default reasoning on, which consumed the answer budget; they were set aside unanalysed and Opus 5 was re-run with reasoning off, as in every earlier Opus run.
PE-85	The classifier's reference values were first extracted as current and superseded, which mislabels questions about the earlier state; they were re-extracted relative to the gold before any answer was classified.
PE-86	The protocol's date-line variant of Mem0 1.0 ran on the conflict questions. A memory leak in the embedding threads was fixed without changing any computation.
PE-88	For budget, the answer classifier labelled the eight-session conflict cells and the memory systems' primary answers, not the full-history cells or the secondary answerers; the store probe ran on the two ingestion orders its prediction compares.
PE-91, 94	Claude Haiku 4.5's MemoryAgentBench cells needed for the position preference were re-run with 200 answer tokens, since 32 cut most answers short. The re-run replaces its row where the paper summarizes cues; the predictions keep the pre-registered row. With 200 tokens Haiku quotes both facts before choosing, so the harness's substring match credits quoted values; the re-run is scored on the final answer, by a rule written after reading the answers and checked against the classifier.
PE-95, 96	Analyses and one pass added after the main runs, all marked exploratory in the text: a second within-model test of reasoning (Haiku's three MemoryAgentBench cells with the answer alone requested, predicted in writing before the run); the orientation split recomputed with five draws, with what each orientation group resolves on its own; and the final-answer rule applied to every MemoryAgentBench setting.

B EXPERIMENTAL SPECIFICATION

Readers. The unified runs of §4 and §5 use the settings of Table 27; this paragraph describes the earlier runs behind §6. `anthropic/claude-opus-5`, `anthropic/claude-haiku-4.5`, `google/gemini-2.5-flash`, `qwen/qwen3-235b-a22b-2507`, `openai/gpt-4o-mini-2024-07-18`, `meta-llama/llama-3.1-8b-instruct` and `meta-llama/llama-3.2-3b-instruct` through OpenRouter, and `deepseek-v4-flash` through its provider's API, which serves V4.1-Flash under that name (PE-70), with thinking disabled by the provider's own switch on LongMemEval and left on, the provider's default, for the STALE panels (PE-68); the conflict cells read

both states, with an 8,000-token budget when thinking is on. For the conflict cells only, `openai/gpt-5.6-sol` and `google/gemini-3.8-flash` through OpenRouter with their providers' default reasoning and an 8,000-token budget. For the date-comparison instruction only, `google/gemini-3.1-pro-preview`. Temperature 0. Answer budget: 200 tokens on LongMemEval; on STALE, the first-token distribution for GPT-4o-mini, 8 tokens for Qwen3-235B, and 6,000 for DeepSeek, whose thinking tokens count against the budget (16,000 on one retry in the conflict cells); the other STALE conflict-cell readers have 16 tokens, with one retry at 64. The initial no-date arms omitted the two Claude readers for budget; later runs added them (Table 5). At the time, their calls cost $0.025 and $0.18, against at most $0.007 for the other OpenRouter readers. **Judge:** `openai/gpt-4o-2024-08-06` with LongMemEval's per-category prompts, unmodified, for every LongMemEval row.

Reader prompt. The benchmark's own:

```
I will give you several history chats between you and a
user.  Please answer the question based on the relevant chat
history.
History Chats:
### Session i:
Session Date:  {date}
Session Content:
{session as a JSON list of role/content turns}
...
Current Date:  {question date}
Question:  {question}
Answer:
```

Sessions appear in the order the arm specifies. No-date arms delete every `Session Date` line and nothing else; the builder asserts that each no-date prompt equals its source with those lines removed, and that every arm shows the same session set.

Arms. In the main text the arms are called the native layout, the superseded-last layout, the matched control and the random order; "dates exchanged", "positions exchanged" and "both exchanged" are the conflict cells. The superseded-last layout moves the superseded member of the gold pair to the final slot. The matched control takes the retrieved sessions that are not members of the pair, listed in date order, and moves the middle one to the final slot. (An earlier description said "median retrieval rank"; the builder has always selected by date position.) The random order is the first of five seeded uniform permutations per question, consulting no annotation; over all draws the superseded session lands in each of the eight slots 29 to 58 times, and after the replacement in 32 to 34 of the 70 questions per draw (PE-69). Full-history arms apply the same constructions to every haystack session, sorted by date as the released script does.

Table 5: Which reader setting ran which analysis. Order arms: native, superseded-last and matched-control layouts; no-date: their twins without dates (the crossing of §6); draws: official-judged random orders; conflict: the five cells of §4; oracle note and instruction: Appendices H and G. A number in parentheses is the number of questions whose full history fits the setting's context. The earlier panels differed for cost and availability; the unified runs completed the conflict cells for every setting.

	LongMemEval, eight sessions				LongMemEval, full history		STALE		MemoryAgentBench		
Setting	order arms	no-date	draws	conflict	order arms	conflict	order panels	conflict	conflict	oracle note	instruction
Claude Opus 5	✓	✓	1	✓	–	–	–	✓[a]	✓[b]	✓	–
GPT-5.6 Sol	✓	✓	5	✓	–	✓	–	✓	✓	–	–
Gemini 3.8 Flash	✓	✓	5	✓	✓	✓	–	✓	✓	–	–
DeepSeek-Flash, thinking on	✓[c]	–	–	✓	–	–	2	✓	✓	–	–
Claude Haiku 4.5	✓	✓	5	✓	–	✓[a]	–	✓	✓	✓	✓
Gemini 2.5 Flash	✓	✓	5	✓	✓	✓	–	✓	✓	✓	✓
Qwen3-235B	✓	✓	5	✓	✓	✓	2	✓	✓	✓	✓
DeepSeek-Flash, thinking off	✓	✓	5	✓	✓	✓	–	✓	✓	–	–
GPT-4o-mini	✓	✓	5	✓	(20)	(14)	1	✓	✓	✓	✓
Llama-3.1-8B	✓	✓	5	✓	(53)	(32)	–	✓	✓	–	–
Llama-3.2-3B	✓	✓	5	✓	–	–	–	✓	✓	–	–
Gemini 3.1 Pro	–	–	–	–	–	–	–	–	–	–	✓

[a] native and the two exchanged cells only; [b] the 6k context only; [c] the superseded-last layout only (PE-80).

Retrieval. Session-level BM25 under the benchmark's official contract over user-turn text, a `BAAI/bge-base-en-v1.5` bi-encoder, and a `cross-encoder/ms-marco-MiniLM-L-6-v2` reranker over the fused union. Eight sessions throughout. Permuting every session date leaves all three rankings byte-identical on $78/78$ questions; timestamps do not enter these rankers.

C AUDIT EVIDENCE

Table 6 gives, for each benchmark of Table 1, the file and the lines at the audited commit that fix the order in which evidence reaches the reader, and the time cue the reader sees. We read the released evaluation code rather than the papers' descriptions: three of the six sort or keep a chronological order in code that no benchmark option turns off, and LongMemEval sorts by date after every one of its six context modes.

Table 6: Where each layout in Table 1 is set, at the commit audited.

Benchmark	Commit	Evidence (file, lines)
LongMemEval	`9e0b455f`	`src/generation/run_generation.py` 224–225: unconditional date sort after all six context modes.
STALE	`ea7d3911`	`STALE/Evaluation/run_target_model.py` 47–57 walks sessions in stored order; stored timestamps are non-decreasing in $400/400$ items and the superseded session comes first in $400/400$.
MemoryAgentBench	`fe1735de`	`agent.py` 287–296 appends each chunk to the long-context agent's context; `utils/templates.py` 80 states that the newer fact has the larger serial number; retrieval agents at 703–760.
BEAM	`b2da22ea`	`long_term_memory_methods.py` 245–265 keeps the latest messages in order and passes only role and content, so stored time anchors are not shown; 467–510 concatenates retrieved chunks in retriever order; `light.py` 112 and 464–545.
AMemGym	`ffcd1885`	`assistants/awi.py` 30–32 and 51–53; `evolvable.py` 155–158; `mem0.py` 18–27, 64 and 99 re-sort Mem0 results by time and print no dates.
LoCoMo	`3eb6f2c5`	`task_eval/gpt_utils.py` 172 builds the full conversation in order; 145–169 orders retrieved items by similarity, descending.

D THE RULE THAT EXPLOITS THE LAYOUT

A partial-input rule that reads only the canonical form is the check §7 prescribes. Keeping only the final session of the native context gives Qwen3-235B an accuracy of 0.257, a paired change of -0.586 $[-0.714, -0.457]$ against the full native context (0.843) on 70 questions, official judge. The final slot of a retrieved set is often an unrelated session, so what the layout supplies is not that the answer is last but that, of the competing statements, the later one is later in the prompt, which is what the two-by-two of §6 probes.

E ASSEMBLY CONVENTIONS

Assembly conventions determine the order in which retrieved evidence reaches the reader. Ordering the context least relevant first, so the strongest match sits next to the question as positional-utility work suggests (Liu et al., 2024), strands the superseded session in the final slot on 31 of 70 questions (44.3%) with the lexical ranker; on the same questions the rate is 27.1% ($19/70$) with a dense bi-encoder and 31.4% ($22/70$) with a cross-encoder. Within the same run, stranded questions lose 12.9 to 48.4 points against their native scores while questions that were not stranded do not lose accuracy, and the two groups score within 6.5 points of each other under the native layout, which argues against a large baseline difficulty imbalance. Ordering most relevant first strands the superseded session on 0 of 70 questions: it moves the current statement, usually the more relevant one, towards the start and leaves the superseded statement in the middle rather than beside the question. No reader's change against the native layout excludes zero under that convention (Table 7), which is what the position account predicts and what a general "reordering is harmful" account does not.

Table 7: Accuracy by assembly convention on the 70 questions, official judge, with within-question changes against each reader's native score. Most relevant last: the interval excludes zero for GPT-4o-mini and reaches zero at its upper end for Gemini 2.5 Flash, Claude Haiku 4.5 and Claude Opus 5. Claude Opus 5's most-relevant-first arm was not run.

	Date	Most relevant first		Most relevant last	
Reader	sorted	acc.	change	acc.	change
Qwen3-235B	0.843	0.786	-0.057 $[-0.157, +0.043]$	0.800	-0.043 $[-0.129, +0.043]$
Gemini 2.5 Flash	0.857	0.871	$+0.014$ $[-0.086, +0.114]$	0.743	-0.114 $[-0.229, +0.000]$
GPT-4o-mini	0.857	0.800	-0.057 $[-0.157, +0.043]$	0.629	-0.229 $[-0.343, -0.114]$
Claude Haiku 4.5	0.871	0.786	-0.086 $[-0.186, +0.014]$	0.771	-0.100 $[-0.200, +0.000]$
Claude Opus 5	0.986	—	—	0.929	-0.057 $[-0.129, +0.000]$

F LoCoMo deletion: a manipulation check

We ran the deletion operation on LoCoMo to check that removing `DATE` lines removes information readers use. Over the multi-hop, temporal and open-domain categories (699 questions), each answered twice over the same full conversation with and without every `DATE` line and scored by LoCoMo's own rule, deletion costs the temporal category 29.7 points (-0.297 $[-0.337, -0.256]$, n=321) and moves multi-hop (-0.015 $[-0.035, +0.005]$, n=282) and open-domain (-0.006 $[-0.032, +0.024]$, n=96) by amounts whose intervals span zero (Table 8). Many temporal answers are dates, so this shows that the deletion removes information and that readers use it when the answer is the date; it does not show that readers use dates to decide which of two statements is current, which is what the two-by-two of §6 measures. We did not run the single-hop (841) or adversarial (446) categories under deletion. Category numbers follow LoCoMo's released code rather than the order its paper introduces the types in (PE-53).

Table 8: Deleting every `DATE` line from LoCoMo (Qwen3-235B, within-question paired, LoCoMo's own scorer). Category numbers follow the released code (PE-53).

Category	n	Dated	Dates deleted	Paired change
Multi-hop	282	0.527	0.512	-0.015 $[-0.035, +0.005]$
Temporal	321	0.488	0.191	-0.297 $[-0.337, -0.256]$
Open-domain	96	0.405	0.399	-0.006 $[-0.032, +0.024]$

G The shape of the dependence (synthetic)

To characterise the dependence we built SlotShift, a generator in which each session is a unit carrying its own date, so a condition permutes order and nothing else; the released artifacts verify that items are identical across length conditions, printed dates identical across position conditions, and dates never move with position, on 100 of 100 items. Scoring is exact string match, no judge.

With the current statement at position 6 of 8, the error rate for Qwen3-235B, Gemini 2.5 Flash and GPT-4o-mini stays at or below 0.03 through slot 6 and rises at slots 7 and 8 (Figure 5), consistent with either "later than the current statement" or "near the end". Claude Haiku 4.5, on its 46 completed items, errs at 0.06 to 0.15 at every slot before 7 and peaks at slot 7 (0.48) rather than 8 (0.20). For GPT-4o-mini with the current statement at position 3, the slot immediately after stays clean (0.010) while error climbs monotonically to 0.830 at the final slot, so proximity to the end predicts the error and distance past the current statement amplifies it. With the superseded statement last, error grows with context length for four of five readers, from 0.020 at 525 tokens to 0.390 at 23k for GPT-4o-mini; Claude Opus 5 does not (0.050 to 0.025, 40 items).

Free generation and forced choice. Replacing free generation with a forced choice that names both candidate values at the decode point does not remove the dependence. With the superseded session first, all three readers choose the current value on all 100 items; with it last they choose it on 0.840, 0.830 and 0.730 of items, changes of -0.160 $[-0.230, -0.090]$, -0.170 $[-0.250, -0.100]$ and -0.270 $[-0.360, -0.190]$ in the share choosing the current value. (For GPT-4o-mini, whose option probabilities we read, the mean probability placed on the current value falls by 0.175.) Nor

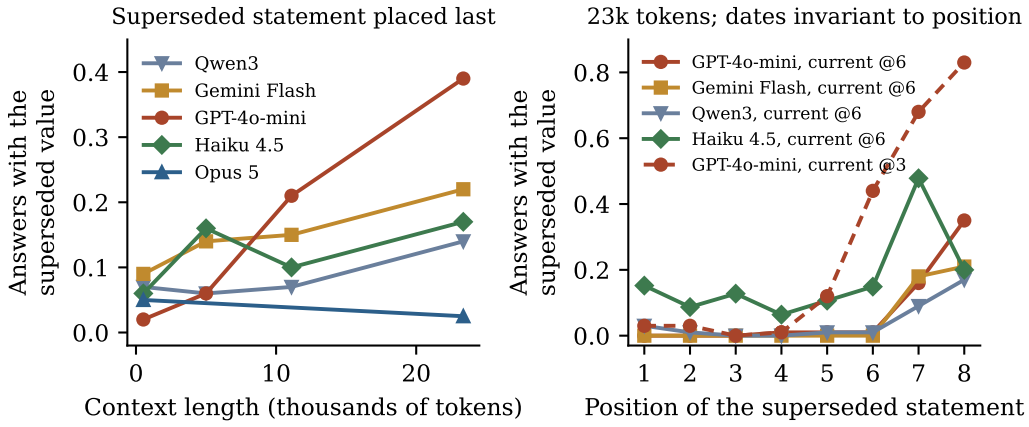


Figure 5: SLOTSHIFT, a *synthetic* generator, not pooled with the externally authored benchmarks. Left: superseded-value assertion rate with the superseded statement last, against context length (100 items per point; Claude Opus 5 40). Right: the position curve at 23k tokens with the *current* statement pinned at the slot named in the legend and the superseded one swept along the x-axis (100 items per point; Claude Haiku 4.5 46). The current-at-3 curve is GPT-4o-mini only.

does the forced choice repair the answers free generation got wrong: it recovers the current value on 0.692 $[0.543, 0.833]$ of them for GPT-4o-mini (n=39) but 0.545 $[0.333, 0.750]$ (n=22) and 0.500 $[0.231, 0.769]$ (n=14) for the others, intervals that include chance in a two-way choice (Table 10).

Date-comparison instruction. Re-running both orderings under an instruction to compare the printed dates and state which is most recent does not reduce the point estimate of the order effect for any of the four models run both ways (Figure 6, Table 9); traces name the superseded session's date as the most recent on 45.7% to 91.3% of wrong answers. Gemini 3.1 Pro, a reasoning model run only with the instruction, shows the smallest effect (-0.057 $[-0.129, +0.000]$), but has no direct-answer comparator. The results are consistent with the dissociation between verbalised temporal reasoning and prediction reported by Wallat et al. (2026).

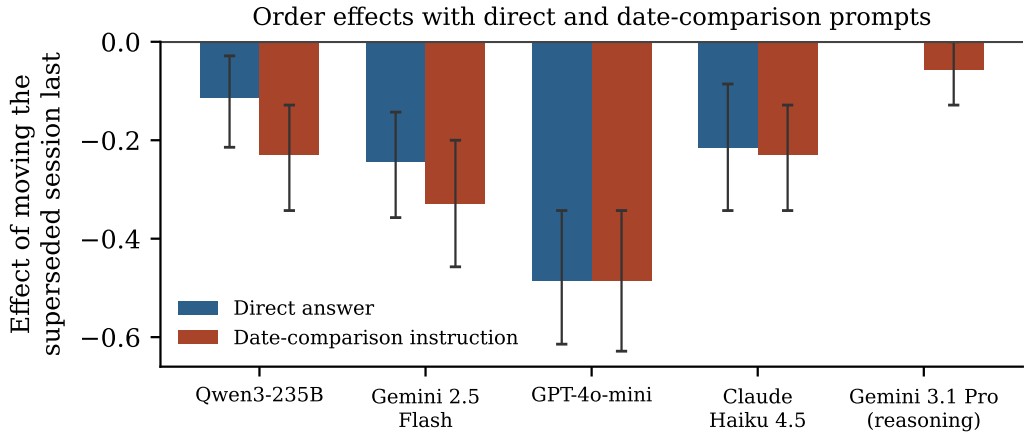


Figure 6: Effect of moving the superseded session last under direct answering and under an instruction to compare the printed `Session Date` lines before answering (70 LongMemEval questions, official judge). Gemini 3.1 Pro is reasoning-capable and is run only with the instruction.

Forced-choice probe. Appended to the unchanged SLOTSHIFT context at ≈ 23k tokens:

```
Question: {question}
Which of these is the user's current answer? (A) {a} (B)
```

Table 9: Effect of moving the superseded session last when answering directly and when instructed to compare the dates first. The last column is not the difference of the two effects: it is accuracy under the superseded-last layout with the instruction minus accuracy under the same layout without it, paired within question.

Model	Direct	Instructed	Instructed − direct
Qwen3-235B	-0.114 $[-0.214, -0.029]$	-0.229 $[-0.343, -0.129]$	-0.071 $[-0.171, +0.029]$
Gemini 2.5 Flash	-0.243 $[-0.357, -0.143]$	-0.329 $[-0.457, -0.200]$	-0.071 $[-0.186, +0.029]$
GPT-4o-mini	-0.486 $[-0.614, -0.343]$	-0.486 $[-0.629, -0.343]$	-0.043 $[-0.143, +0.057]$
Claude Haiku 4.5	-0.214 $[-0.343, -0.086]$	-0.229 $[-0.343, -0.129]$	$+0.071$ $[-0.057, +0.200]$
Gemini 3.1 Pro (reasoning)	—	-0.057 $[-0.129, +0.000]$	—

```
{b}
Reply with a single letter, A or B, and nothing else.
Answer:
```

The current value is option B on even-numbered items and A on odd ones; decoding is greedy with a four-token cap.

Table 10: Forced choice against free generation on the same 100 SLOTSHIFT items. Columns give the share naming the current value. "Recovered" is the share of items that free generation answered with the superseded value and the probe answers with the current one (n of 39, 22 and 14); chance is 0.5.

	Superseded first		Superseded last		
Reader	Free	Probe	Free	Probe	Recovered
GPT-4o-mini	0.890	1.000	0.580	0.840 [0.77, 0.91]	0.692 [0.54, 0.83]
Gemini 2.5 Flash	0.940	1.000	0.730	0.830 [0.75, 0.90]	0.545 [0.33, 0.75]
Qwen3-235B	0.950	1.000	0.820	0.730 [0.64, 0.81]	0.500 [0.23, 0.77]

H WHAT THE LAYOUT DOES TO A MECHANISM'S MEASURED VALUE

If the layout already points to the answer, an additional update signal may have less room to help. We test this with an *oracle* annotation built from the gold pair labels, inserted before the `Current Date` line under both orderings: *"Currency notes: Session i is SUPERSEDED by Session j; Session j is the CURRENT statement on that topic. Answer from the CURRENT statement unless the question asks about the past."* Session numbers refer to the positions as rendered in each arm.

Under the native layout the annotation is not distinguishable from nothing for any of the five original readers: $+0.014$, $+0.029$, -0.014, $+0.029$ and -0.029, all intervals spanning zero, though they still permit benefits of up to 12.9 points. With the superseded session last, on the same questions with the same sentence, it is worth $+0.129$, $+0.243$, $+0.071$, $+0.129$ and $+0.043$, with intervals excluding zero for three of five readers (Table 11). The annotation supplies an ideal signal; extraction errors of a deployed method are outside this test. To compare its gains directly, we compute a within-question difference in differences. The gain with the superseded session last exceeds that under the native layout by 11.7 points $[6.3, 17.1]$, pooled over five readers. Intervals exclude zero for 2 of the 5 readers individually (Table 12; exploratory).

The same instruction on LoCoMo's adversarial category. Where correct behaviour is abstention, a turn-level version of the note, whose superseded pair is chosen by the pre-registered lexical rule rather than by gold labels, reduces abstention from 0.794 $[0.756, 0.832]$ to 0.733 $[0.693, 0.774]$, a paired change of -0.061 $[-0.087, -0.036]$ (GPT-4o-mini, n=446). The prompt appends, after the retrieved excerpts, *"Currency notes: – The statement in [a] is SUPERSEDED by [b], which is the CURRENT statement on that topic."* or, when the rule finds no superseded pair, *"– No excerpt in this set was superseded by a later excerpt on the same topic."*, followed by *"Answer from the CURRENT statement unless the question asks about the past."* (`experiments/run_locomo_e2_adversarial.py`, lines 106–128).

Table 11: The oracle annotation under both orderings (70 questions, official judge): accuracy without and with it, and the paired difference. Under the native layout no interval excludes zero; with the superseded session last three of five do.

	Native layout			Superseded last		
Model	without	with	change	without	with	change
Qwen3-235B	0.843	0.857	$+0.014$ $[-0.071, +0.100]$	0.729	0.857	$+0.129$ $[+0.029, +0.229]$
Gemini 2.5 Flash	0.857	0.886	$+0.029$ $[-0.071, +0.129]$	0.614	0.857	$+0.243$ $[+0.129, +0.357]$
GPT-4o-mini	0.857	0.843	-0.014 $[-0.043, +0.000]$	0.371	0.443	$+0.071$ $[+0.014, +0.143]$
Claude Haiku 4.5	0.871	0.900	$+0.029$ $[-0.043, +0.114]$	0.657	0.786	$+0.129$ $[+0.000, +0.257]$
Claude Opus 5	0.986	0.957	-0.029 $[-0.071, +0.000]$	0.914	0.957	$+0.043$ $[-0.043, +0.129]$

Table 12: The oracle annotation's gain with the superseded session last minus its gain under the native context (Table 11), paired within question: a difference in differences, official judge. Exploratory.

Reader	n	Difference of the gains
Qwen3-235B	70	$+0.114$ $[+0.000, +0.229]$
Gemini 2.5 Flash	70	$+0.214$ $[+0.086, +0.343]$
GPT-4o-mini	70	$+0.086$ $[+0.029, +0.157]$
Claude Haiku 4.5	70	$+0.100$ $[-0.043, +0.243]$
Claude Opus 5	70	$+0.071$ $[-0.014, +0.171]$
Pooled over readers	70	$+0.117$ $[+0.063, +0.171]$

I Additional results

This section collects the earlier runs behind §6 and the earlier conflict runs: the layout-by-deletion crossing (Table 13, per reader in Figure 8, the rest in Table 14); the order contrasts on LongMemEval and StALE (Figure 7, Tables 15 and 16, item-level transitions in Figure 10); the final slot as a third cue (Table 23); the LongMemEval conflict cells without the questions that mention explicit calendar dates (Table 24); and the earlier conflict runs with their predictions (Tables 18 to 21).

Table 13: Layout crossed with date deletion on LongMemEval (70 questions, official judge, earlier runs): accuracy with and without every `Session Date` line, under the native and the superseded-last layout, and the within-question value of the dates under each. The random order and $M = D_{\text{adv}} - D_{\text{nat}}$ are in Table 14; the readers added in the unified runs are in Table 41.

	Native		Superseded last			
Reader	dates	none	dates	none	D_{nat}	D_{adv}
DeepSeek-Flash	0.929	0.900	0.686	0.557	$+0.029$ $[-0.029, +0.086]$	$+0.129$ $[+0.029, +0.229]$
Claude Haiku 4.5	0.871	0.871	0.657	0.400	$+0.000$ $[-0.071, +0.071]$	$+0.257$ $[+0.143, +0.371]$
Gemini 2.5 Flash	0.857	0.800	0.614	0.500	$+0.057$ $[+0.014, +0.114]$	$+0.114$ $[+0.014, +0.214]$
GPT-4o-mini	0.857	0.843	0.371	0.286	$+0.014$ $[-0.043, +0.071]$	$+0.086$ $[+0.029, +0.157]$
Llama-3.1-8B	0.857	0.800	0.371	0.386	$+0.057$ $[-0.029, +0.143]$	-0.014 $[-0.114, +0.071]$
Qwen3-235B	0.843	0.786	0.729	0.600	$+0.057$ $[-0.014, +0.143]$	$+0.129$ $[+0.029, +0.229]$
Llama-3.2-3B	0.700	0.586	0.329	0.371	$+0.114$ $[+0.014, +0.214]$	-0.043 $[-0.129, +0.043]$

Where the random-order loss falls (exploratory). The first draw reversed the update pair on 32 of the 70 questions and preserved it on the rest; presentation order is taken from the content-verified labels of PE-69. Table 17 gives each reader's loss against the native layout in the two groups, first under that draw and then under all five draws for the 5 readers whose five draws are official-judged. Across five draws, 68 questions enter the reversed group and 69 the preserved group; 67 enter both. We average available draws within question and resample question IDs jointly across groups. The pooled difference is 22.4 points [15.4, 29.3] (loss 27.5 where reversed, 5.0 where preserved). Read on their own, the reversed questions separate 7 of the 10 reader pairs and the preserved questions 3, against 8 for the five-draw mean and 0 for the native layout on the same questions. The additional

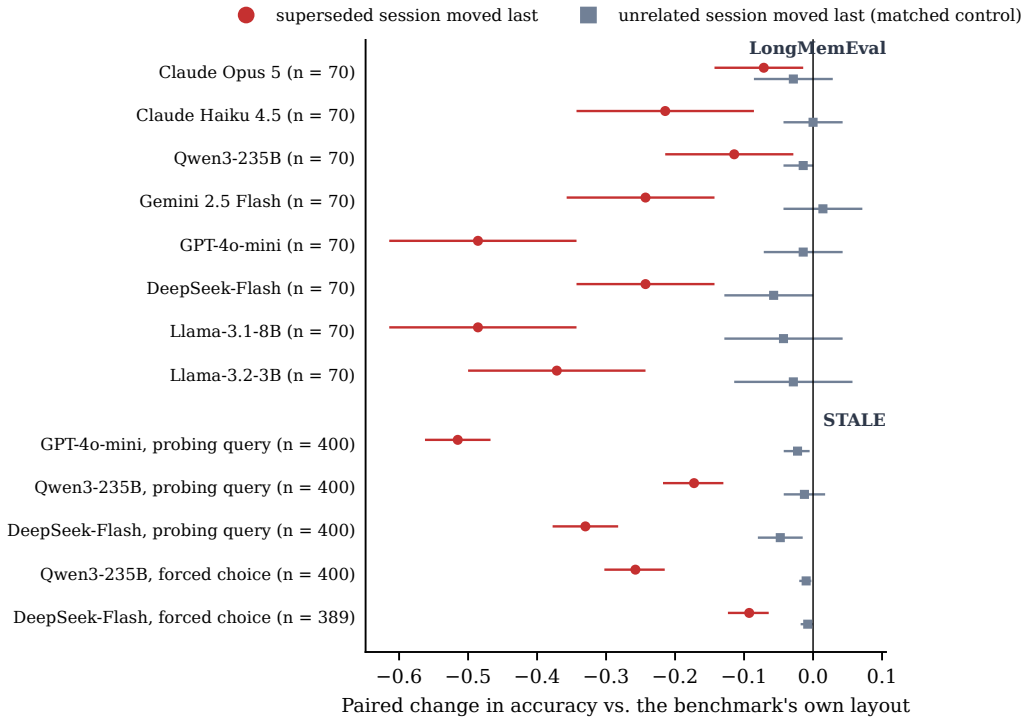


Figure 7: Every order contrast of the earlier runs, with its matched control. Moving the superseded session to the final slot lowers accuracy for every LongMemEval reader and every STALE panel; moving an unrelated session to the same slot costs at most 5.7 points. Within-item changes against each benchmark's own layout.

Table 14: The rest of the two-by-two (Table 13): accuracy under one uniformly random order per question with and without the dates, the value of the dates there (D_{neu}), and $M = D_{\text{adv}} - D_{\text{nat}}$. With random order and no dates, neither of these two cues consistently identifies the newer statement.

	Random order			
Reader	dates	none	D_{neu}	M
DeepSeek-Flash	0.829	0.729	$+0.100$ $[+0.029, +0.186]$	$+0.100$ $[-0.014, +0.214]$
Claude Haiku 4.5	0.829	0.657	$+0.171$ $[+0.071, +0.271]$	$+0.257$ $[+0.114, +0.400]$
Gemini 2.5 Flash	0.686	0.657	$+0.029$ $[-0.057, +0.114]$	$+0.057$ $[-0.057, +0.171]$
GPT-4o-mini	0.629	0.614	$+0.014$ $[-0.043, +0.071]$	$+0.071$ $[-0.014, +0.171]$
Llama-3.1-8B	0.586	0.629	-0.043 $[-0.157, +0.071]$	-0.071 $[-0.200, +0.057]$
Qwen3-235B	0.814	0.714	$+0.100$ $[+0.014, +0.186]$	$+0.071$ $[-0.043, +0.186]$
Llama-3.2-3B	0.514	0.543	-0.029 $[-0.114, +0.057]$	-0.157 $[-0.271, -0.043]$

resolution is concentrated in the reversed group. Pooled over the 7 readers of the crossing without selection, M is $+0.047$ $[-0.006, +0.100]$.

Earlier runs of the conflict cells. Before the unified runs, the same cells were run on STALE for DeepSeek-Flash (Tables 18 and 25) and on LongMemEval and STALE for further readers, without pinned providers (Tables 19 and 20); Table 21 gives their predictions. Of the 17 settings and benchmarks run both times, 12 have position preferences whose intervals exclude zero in both runs, and 12 of these agree in sign. The two runs' intervals fail to overlap only for Gemini 2.5 Flash on STALE ($+0.237$ in the earlier run, $+0.079$ in the unified run, where the provider was pinned). Native answers on LongMemEval agree between the runs on 0.88 to 1.00 of the questions, per setting.

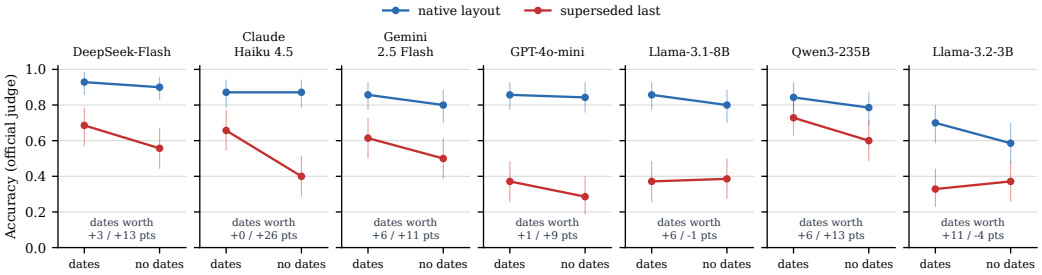


Figure 8: The layout-by-deletion crossing of Table 13, per reader: accuracy with and without the dates under the native layout and with the superseded session last. Bars are 95% intervals; the annotation gives the value of the dates under each.

Table 15: LongMemEval order contrasts (70 questions, official judge): accuracy under the benchmark's layout, paired changes when the superseded or an unrelated session is moved last, and accuracy under one uniformly random order per question.

Reader	Native	Superseded last	Matched control	Random order
Claude Opus 5	0.986	−0.071 [−0.143, −0.014]	−0.029 [−0.086, +0.029]	–
DeepSeek-Flash	0.929	−0.243 [−0.343, −0.143]	−0.057 [−0.129, +0.000]	0.829
Claude Haiku 4.5	0.871	−0.214 [−0.343, −0.086]	+0.000 [−0.043, +0.043]	0.829
Gemini 2.5 Flash	0.857	−0.243 [−0.357, −0.143]	+0.014 [−0.043, +0.071]	0.686
GPT-4o-mini	0.857	−0.486 [−0.614, −0.343]	−0.014 [−0.071, +0.043]	0.629
Llama-3.1-8B	0.857	−0.486 [−0.614, −0.343]	−0.043 [−0.129, +0.043]	0.586
Qwen3-235B	0.843	−0.114 [−0.214, −0.029]	−0.014 [−0.043, +0.000]	0.814
Llama-3.2-3B	0.700	−0.371 [−0.500, −0.243]	−0.029 [−0.114, +0.057]	0.514

Cue conflict: sensitivity. Table 25 repeats the estimates of Table 18 with the one non-answer dropped and on the 174 items whose two competing sessions mention no explicit year or month and day, which removes explicit calendar references but can leave relative time expressions. Position preference is the share choosing the replacement with dates exchanged minus with positions exchanged; it is positive when position outweighs the dates.

Table 16: STALE order contrasts. Native is accuracy under the benchmark's chronological layout; the last two columns are within-item paired changes when the superseded memory (superseded last) or an unrelated session (matched control) is moved last. DeepSeek-Flash ran with thinking on for these panels.

Readout	Reader	n	Native	Superseded last	Matched control
Probing query	GPT-4o-mini	400	0.573	-0.515 $[-0.562, -0.468]$	-0.022 $[-0.043, -0.005]$
Probing query	Qwen3-235B	400	0.345	-0.172 $[-0.218, -0.130]$	-0.013 $[-0.043, +0.018]$
Probing query	DeepSeek-Flash	400	0.672	-0.330 $[-0.377, -0.282]$	-0.048 $[-0.080, -0.015]$
Forced choice	Qwen3-235B	400	0.968	-0.258 $[-0.303, -0.215]$	-0.010 $[-0.020, -0.002]$
Forced choice	DeepSeek-Flash	389	1.000	-0.093 $[-0.123, -0.064]$	-0.008 $[-0.018, +0.000]$

Table 17: Loss against the native layout under one random order per question, split by whether the draw placed the superseded session after its replacement (reversed) or before it (preserved). The split is random, so baseline difficulty is balanced in expectation. The last column uses five draws, with 68 questions in the reversed group, 69 in the preserved group and 67 shared; resampling is clustered by question. Exploratory.

	Reversed		Preserved		Difference	
Reader	n	loss	n	loss	one draw	five draws
GPT-4o-mini	32	$+0.438$	38	$+0.053$	$+0.385$ $[+0.166, +0.600]$	$+0.361$ $[+0.245, +0.477]$
Llama-3.1-8B	32	$+0.438$	38	$+0.132$	$+0.306$ $[+0.070, +0.533]$	$+0.215$ $[+0.103, +0.324]$
Llama-3.2-3B	32	$+0.344$	38	$+0.053$	$+0.291$ $[+0.056, +0.518]$	$+0.274$ $[+0.158, +0.386]$
Claude Haiku 4.5	32	$+0.188$	38	-0.079	$+0.266$ $[+0.074, +0.470]$	–
Gemini 2.5 Flash	32	$+0.281$	38	$+0.079$	$+0.202$ $[+0.012, +0.399]$	$+0.210$ $[+0.117, +0.305]$
DeepSeek-Flash	32	$+0.125$	38	$+0.079$	$+0.046$ $[-0.143, +0.238]$	–
Qwen3-235B	32	$+0.031$	38	$+0.026$	$+0.005$ $[-0.137, +0.145]$	$+0.061$ $[-0.006, +0.132]$
Mean of readers	32	$+0.263$	38	$+0.049$	$+0.215$ $[+0.110, +0.317]$	$+0.224$ $[+0.154, +0.293]$

Table 18: Cue conflict on STALE for DeepSeek-Flash (190 items, within-item contrasts). Left: share of answers choosing the replacement memory in each version. Right: E_p^+, the cost of exchanging positions with dates intact; E_d^+, the cost of exchanging dates with positions intact; I, how much more exchanging the dates costs when positions are also exchanged than when they are not.

	Chooses the replacement				Within-item effects		
Reader	native	dates	positions	both	E_p^+	E_d^+	I
Thinking on	0.995	0.097	0.968	0.026	$+0.026$ $[+0.000, +0.053]$	$+0.897$ $[+0.853, +0.937]$	$+0.045$ $[+0.000, +0.092]$
Thinking off	0.921	0.784	0.500	0.179	$+0.421$ $[+0.347, +0.495]$	$+0.137$ $[+0.089, +0.189]$	$+0.184$ $[+0.095, +0.274]$

Table 19: Conflict cells on LongMemEval (official judge), ordered by position preference. Accuracy under the benchmark's layout (native), with the competing sessions' dates exchanged, and with their positions exchanged; E_p^+ and E_d^+ as in Table 18; position preference is accuracy with dates exchanged minus accuracy with positions exchanged.

Reader	n	native	dates	positions	E_p^+	E_d^+	Pos. pref.
Gemini 3.8 Flash	49	0.980	0.469	0.980	$+0.000$ $[+0.000, +0.000]$	$+0.510$ $[+0.367, +0.653]$	-0.510 $[-0.653, -0.367]$
DeepSeek-Flash (thinking on)	40	0.925	0.525	0.900	$+0.025$ $[-0.054, +0.114]$	$+0.400$ $[+0.250, +0.553]$	-0.375 $[-0.526, -0.225]$
GPT-5.6 Sol	49	0.918	0.592	0.918	$+0.000$ $[-0.082, +0.082]$	$+0.327$ $[+0.184, +0.469]$	-0.327 $[-0.469, -0.184]$
DeepSeek-Flash (thinking off)	40	0.900	0.750	0.750	$+0.150$ $[+0.049, +0.268]$	$+0.150$ $[+0.049, +0.268]$	$+0.000$ $[-0.122, +0.119]$
Qwen3-235B	49	0.837	0.714	0.653	$+0.184$ $[+0.061, +0.327]$	$+0.122$ $[+0.020, +0.224]$	$+0.061$ $[-0.082, +0.204]$
Gemini 2.5 Flash	49	0.857	0.755	0.653	$+0.204$ $[+0.082, +0.327]$	$+0.102$ $[+0.020, +0.184]$	$+0.102$ $[-0.020, +0.224]$
Claude Haiku 4.5	49	0.878	0.816	0.653	$+0.224$ $[+0.102, +0.347]$	$+0.061$ $[-0.041, +0.163]$	$+0.163$ $[+0.020, +0.306]$
GPT-4o-mini	49	0.878	0.837	0.551	$+0.327$ $[+0.184, +0.469]$	$+0.041$ $[+0.000, +0.102]$	$+0.286$ $[+0.143, +0.429]$
Llama-3.2-3B	49	0.776	0.776	0.429	$+0.347$ $[+0.224, +0.490]$	$+0.000$ $[-0.102, +0.102]$	$+0.347$ $[+0.204, +0.490]$
Llama-3.1-8B	49	0.878	0.837	0.408	$+0.469$ $[+0.306, +0.633]$	$+0.041$ $[-0.061, +0.163]$	$+0.429$ $[+0.265, +0.592]$

Table 20: Conflict cells on STALE for every reader setting (190 items), ordered by position preference. DeepSeek-Flash rows are those of Table 18.

Reader	native	dates	positions	both	E_p^+	E_d^+	I	Pos. pref.
DeepSeek, thinking on	0.995	0.097	0.968	0.026	$+0.026$ $[+0.000, +0.053]$	$+0.897$ $[+0.853, +0.937]$	$+0.045$ $[+0.000, +0.092]$	-0.871 $[-0.916, -0.821]$
Qwen3-235B	0.963	0.626	0.863	0.463	$+0.100$ $[+0.053, +0.153]$	$+0.337$ $[+0.268, +0.405]$	$+0.063$ $[-0.021, +0.147]$	-0.237 $[-0.311, -0.163]$
Llama-3.2-3B	0.763	0.721	0.663	0.611	$+0.100$ $[+0.037, +0.163]$	$+0.042$ $[+0.005, +0.079]$	$+0.011$ $[-0.047, +0.068]$	$+0.058$ $[+0.000, +0.116]$
Llama-3.1-8B	0.732	0.679	0.574	0.516	$+0.158$ $[+0.089, +0.226]$	$+0.053$ $[-0.005, +0.111]$	$+0.005$ $[-0.074, +0.089]$	$+0.105$ $[+0.026, +0.179]$
Gemini 2.5 Flash	0.926	0.832	0.595	0.353	$+0.332$ $[+0.263, +0.400]$	$+0.095$ $[+0.042, +0.147]$	$+0.147$ $[+0.063, +0.232]$	$+0.237$ $[+0.168, +0.311]$
DeepSeek, thinking off	0.921	0.784	0.500	0.179	$+0.421$ $[+0.347, +0.495]$	$+0.137$ $[+0.089, +0.189]$	$+0.184$ $[+0.095, +0.274]$	$+0.284$ $[+0.211, +0.358]$
GPT-4o-mini	0.934	0.895	0.263	0.211	$+0.671$ $[+0.605, +0.737]$	$+0.039$ $[+0.011, +0.071]$	$+0.013$ $[-0.032, +0.058]$	$+0.632$ $[+0.563, +0.700]$

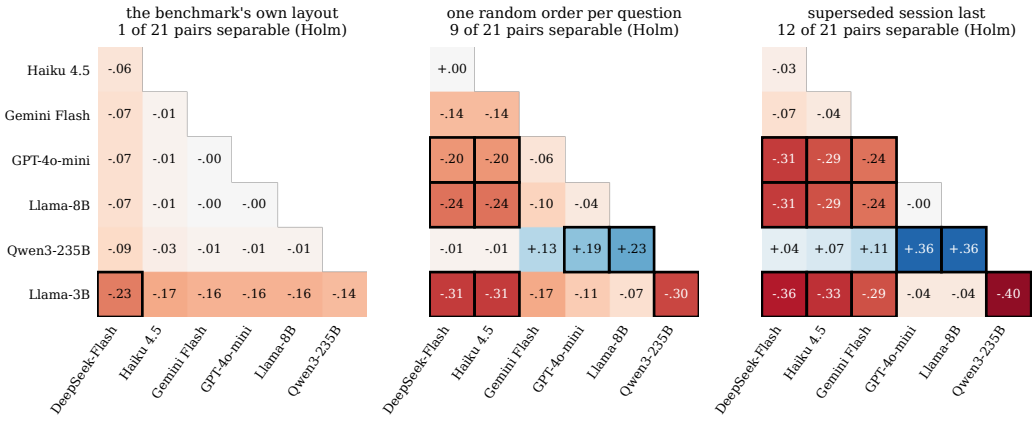


Figure 9: Pairwise accuracy differences (row minus column) among the readers run under all three layouts, ordered by accuracy under the native layout. Outlined cells are pairs separable after Holm correction within the layout (§6).

Table 21: Every pre-specified prediction of the experiments reported here, and its outcome. The protocol documents are in the supplementary material; they carry no third-party timestamp.

Experiment	Prediction	Outcome
Crossing	For most readers the dates are worth more when position misleads ($D_{\text{adv}} > D_{\text{nat}}$)	Held for 5 of 7 readers
Crossing	A random order separates more reader pairs than the native layout, else the claim is withdrawn	Held: 9 against 1 of 21
Conflict, STALE	Thinking on: exchanging positions changes answers ($E_p^+ > 0$)	Not held: $+0.026$ $[+0.000, +0.053]$
Conflict, STALE	Thinking on: masking interaction $I > 0$	Not held: $+0.045$ $[+0.000, +0.092]$
Conflict, STALE	Thinking off (secondary): $E_p^+ > 0$ and $I > 0$	Held: $+0.421$ $[+0.347, +0.495]$; $+0.184$ $[+0.095, +0.274]$
Five draws	The five-draw mean separates more pairs than the native layout	Held: 8 against 0 of 10
Conflict, LongMemEval	Among 10 settings, position preferences of both signs survive Holm	Held: 3 positive, 3 negative
Conflict, LongMemEval	Thinking lowers DeepSeek-Flash's position preference	Held: -0.375 $[-0.550, -0.190]$
Conflict, LongMemEval	Position preference < 0 for GPT-5.6 Sol and Gemini 3.8 Flash	Held: -0.327 $[-0.469, -0.184]$; -0.510 $[-0.653, -0.367]$
Conflict, STALE	Interaction $I > 0$ averaged over 5 further readers	Held: $+0.048$ $[+0.017, +0.080]$
Conflict, STALE	$I > 0$ for each such reader whose $E_p^+ > 0$	Not held: only Gemini 2.5 Flash of 5

Table 22: Accuracy under the native layout and under each of five uniformly random orders per question (LongMemEval, the 70 questions, official judge), with the mean and standard deviation over the five orders, for every reader with five judged orders. The first five are the readers of the pairwise comparison in §6. Exploratory.

Reader	native	order 1	order 2	order 3	order 4	order 5	mean	s.d.
Gemini 2.5 Flash	0.857	0.686	0.714	0.729	0.786	0.714	0.726	0.037
Qwen3-235B	0.843	0.814	0.771	0.771	0.771	0.814	0.789	0.023
GPT-4o-mini	0.857	0.629	0.614	0.643	0.614	0.600	0.620	0.016
Llama-3.1-8B	0.857	0.586	0.614	0.686	0.629	0.629	0.629	0.036
Llama-3.2-3B	0.700	0.514	0.429	0.571	0.429	0.543	0.497	0.066
GPT-5.6 Sol	0.929	0.900	0.871	0.943	0.929	0.900	0.909	0.028
Gemini 3.8 Flash	0.971	0.971	0.986	0.971	0.971	0.971	0.974	0.006
Claude Haiku 4.5	0.871	0.829	0.800	0.786	0.786	0.771	0.794	0.022
DeepSeek-Flash (thinking off)	0.929	0.829	0.814	0.843	0.771	0.786	0.809	0.030

Table 23: The final slot as a third cue: accuracy with the superseded session moved to the final slot, accuracy with the pair exchanged in place (relative order reversed, final slot unchanged), and their difference on the same items. The two cells come from separate runs of the same reader with the same settings. Exploratory.

Benchmark	Reader	n	moved last	in place	in place − moved last
LongMemEval	Claude Haiku 4.5	49	0.633	0.653	$+0.020$ $[-0.122, +0.163]$
LongMemEval	Gemini 2.5 Flash	49	0.592	0.653	$+0.061$ $[-0.061, +0.184]$
LongMemEval	Qwen3-235B	49	0.735	0.653	-0.082 $[-0.204, +0.041]$
LongMemEval	GPT-4o-mini	49	0.367	0.551	$+0.184$ $[+0.082, +0.306]$
LongMemEval	Llama-3.1-8B	49	0.347	0.408	$+0.061$ $[-0.061, +0.184]$
LongMemEval	Llama-3.2-3B	49	0.367	0.429	$+0.061$ $[-0.102, +0.204]$
LongMemEval	DeepSeek-Flash (thinking off)	40	0.725	0.750	$+0.025$ $[-0.054, +0.114]$
STALE	DeepSeek-Flash (thinking on)	187	0.845	0.968	$+0.123$ $[+0.075, +0.175]$
STALE	Qwen3-235B	190	0.674	0.863	$+0.189$ $[+0.126, +0.253]$

Table 24: Position preference on LongMemEval without the 8 questions whose competing sessions mention an explicit year or month and day. Exploratory.

Reader	n	Position preference
Gemini 3.8 Flash	41	-0.488 $[-0.641, -0.333]$
GPT-5.6 Sol	41	-0.390 $[-0.553, -0.225]$
DeepSeek-Flash (thinking on)	34	-0.382 $[-0.548, -0.222]$
DeepSeek-Flash (thinking off)	34	$+0.000$ $[-0.118, +0.114]$
Qwen3-235B	41	$+0.073$ $[-0.089, +0.233]$
Gemini 2.5 Flash	41	$+0.098$ $[-0.050, +0.250]$
Claude Haiku 4.5	41	$+0.122$ $[-0.026, +0.279]$
GPT-4o-mini	41	$+0.293$ $[+0.140, +0.450]$
Llama-3.2-3B	41	$+0.366$ $[+0.222, +0.513]$
Llama-3.1-8B	41	$+0.415$ $[+0.244, +0.581]$

Table 25: Cue-conflict estimates under the pre-specified sensitivity analyses.

Reader	Analysis	n	E_p^+	E_d^+	I	Pos. pref.
On	primary	190	$+0.026$ $[+0.000, +0.053]$	$+0.897$ $[+0.853, +0.937]$	$+0.045$ $[+0.000, +0.092]$	-0.871 $[-0.916, -0.821]$
	non-answer dropped	189	$+0.026$ $[+0.000, +0.053]$	$+0.899$ $[+0.856, +0.941]$	$+0.042$ $[+0.000, +0.089]$	-0.873 $[-0.920, -0.824]$
	no calendar dates	174	$+0.023$ $[+0.000, +0.052]$	$+0.899$ $[+0.853, +0.942]$	$+0.043$ $[-0.003, +0.090]$	-0.876 $[-0.923, -0.826]$
Off	primary	190	$+0.421$ $[+0.347, +0.495]$	$+0.137$ $[+0.089, +0.189]$	$+0.184$ $[+0.095, +0.274]$	$+0.284$ $[+0.211, +0.358]$
	non-answer dropped	190	$+0.421$ $[+0.347, +0.495]$	$+0.137$ $[+0.089, +0.189]$	$+0.184$ $[+0.095, +0.274]$	$+0.284$ $[+0.211, +0.358]$
	no calendar dates	174	$+0.420$ $[+0.343, +0.497]$	$+0.138$ $[+0.086, +0.193]$	$+0.178$ $[+0.083, +0.272]$	$+0.282$ $[+0.203, +0.360]$

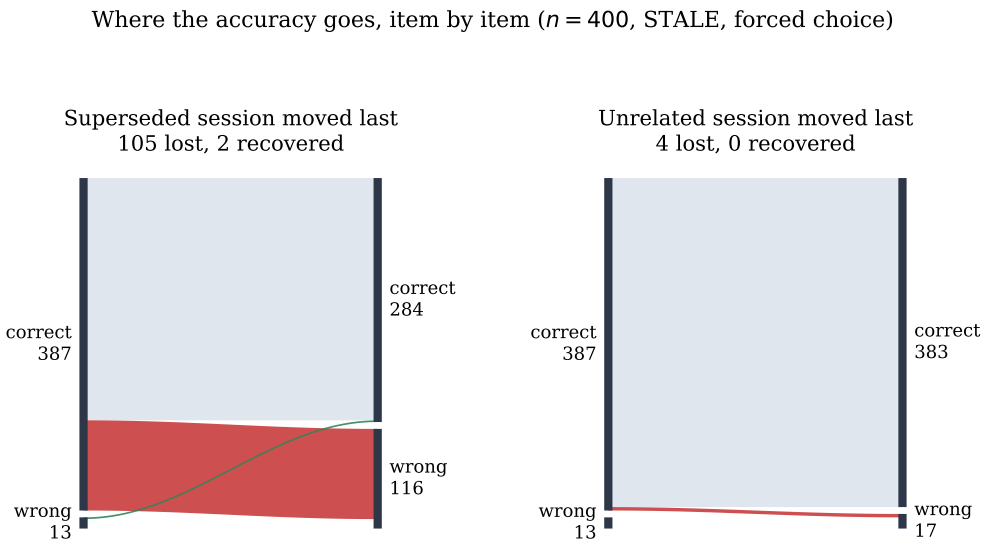


Figure 10: Item-level transitions on the STALE forced choice for Qwen3-235B (n=400). Moving the superseded session last turns 105 correct answers wrong and recovers 2; moving an unrelated session to the same slot changes four items in total.

J CLAIM-TO-EVIDENCE MAP

Hashes are the first eight hex digits of the SHA-256 of each file, computed when the paper is built.

Table 26: Quantitative results and their source artifacts. Hashes identify the recorded files; large prompt files are specified by builders and manifests rather than included in the archive.

Claim	Where	Artifact	SHA-256
Evidence order in six benchmarks (Table 1)	§3	e15_harness_layout_audit_v1.json	3ff5c660
Pair recovery at four and eight sessions	§4	e1_pair_recovery_v1.json	16fdd15b
Rankers ignore permuted timestamps, 78/78	§4	e2_c1a_date_leak_gate.json	7857848f
E14 arms and their guarantees	§4	e14_arms_cap8.jsonl	3f472e23
Pre-registration of E14 and its amendments	§4	E14_PROTOCOL_2026-09-16.md	ab391d68
Two-by-two, order contrasts, resolving power, judge agreement	§6	e14_results_v1.json	cc8f7d12
Official-judge verdicts, E14 arms	§6	judged_cap8_primary.jsonl	0005e1dc
Official-judge verdicts, random order and LastOnly	§6	judged_rejudge_primary.jsonl	4d1c5fa1
Official-judge verdicts, original dated arms	§6	e3_judged_primary.jsonl	1ffbe092
Official-judge verdicts, original dated arms (continued)	§6	e3_judged_extra.jsonl	91f4c2d2
Detectable-effect floor and equivalence bounds	§6	e5_null_power_v1.json	4963cd0e
Full-history arms and their guarantees	§6	e14_arms_fullhistory.jsonl	3cf2d385
Official-judge verdicts, full history	§6	judged_fullhistory_primary.jsonl	550de769
Random-order family (seeded permutations)	§6	e13_marginal_arms.jsonl	79c9ca61
Random-order draws, verified presentation order	§6	e13_marginal_orders_v2.json	905ba91e
Orientation split and pooled M (Table 17)	§6	e17_orientation_v1.json	58503c3a
Pre-registration of the cue-conflict experiment	§4	E16_PROTOCOL_2026-09-17.md	4d41258d
Cue-conflict arms and their guarantees	§4	cueconflict_arms_v1.jsonl	4234d43c
Cue-conflict reads, thinking on	§4	e16_read_deepseek_on.jsonl	0cf34172
Cue-conflict reads, thinking off	§4	e16_read_deepseek_off.jsonl	489253ea
Cue-conflict analysis (Table 18)	§4	e16_cueconflict_v1.json	ad5ffd1c
Pre-registration of the second experiment and its amendments	§4	E18_PROTOCOL_2026-09-17.md	8a57eb76
Conflict cells on LongMemEval, arms and guarantees	§4	e18_cueconflict_arms.jsonl	e13b0988
Conflict cells on LongMemEval, official-judge verdicts	§4	judged_B_primary.jsonl	197e516a
Completed cells: draws, full history, no-date arms (verdicts)	§6	judged_e18a_primary.jsonl	8e19ccb8
Conflict cells, five draws, pre-specified outcomes (Table 21)	§4	e18_results_v1.json	c1522b41
STALE conflict reads, GPT-4o-mini	§4	stale_read_or-gpt-4o-mini.jsonl	503a9a37
STALE conflict reads, Qwen3-235B	§4	stale_read_or-qwen3.jsonl	eb5cfcda
STALE conflict reads, Gemini 2.5 Flash	§4	stale_read_or-gemini-flash.jsonl	c95fdb21
STALE conflict reads, Llama-3.1-8B	§4	stale_read_or-llama31-8b.jsonl	f6e59a3c
STALE conflict reads, Llama-3.2-3B	§4	stale_read_or-llama32-3b.jsonl	7d1aaca2
STALE panel (Table 16)	§6	e8_stale_panel_v2.json	4fa24f59
STALE order arms and their invariants	§6	order_arms_v1.jsonl	9898ecc3
STALE forced-choice arms, balanced gold	§6	forcedchoice_arms_v1.jsonl	a4b145cb
STALE reader outputs, GPT-4o-mini (400 items)	§6	order_read_gpt4omini.jsonl	40fe7f0a
STALE transitions (Figure 10)	§6	e12_transitions_grid_v1.json	ebf71f1f
Strand rates by ranker	App. E	e12_ranker_strand_rates_v1.json	b706efdd
Assembly conventions (Table 7)	App. E	e6_assembly_conventions_v1.json	fd03f76d
Stranded versus not stranded, paired	App. E	e3_pipeline_paired_v1.json	423515ad
LoCoMo deletion (Table 8)	App. F	e11_deletion_v1.json	885c99e7
SLOTSHIFT length sweep and position curves	App. G	slotshift_results_v2.json	1a2ec353
SLOTSHIFT current-early curve	App. G	slotshift_freshearly_v2.json	d4a83c4a
Forced choice, three readers (Table 10)	App. G	forced_choice_v1.json	d40055da
Date-comparison instruction grid (Table 9)	App. G	e5_cot_grid_v1.json	2242db60
Traces naming the superseded date as most recent	App. G	e5_cot_traces_v1.json	269a4807
Oracle annotation grid (Table 11)	App. H	e4_annotation_grid_v1.json	8aa44336
LoCoMo abstention under the currency note	App. H	e2_adversarial_v1.jsonl	5c53f3e7
Post-review analyses: calendar-free subset, final slot (Tables 23–24)	App. I	e19_post_review_v1.json	e1450243
Protocol of the unified runs, five amendments	§4	E20_PROTOCOL_2026-09-19.md	91e9d623
Conflict cells, LongMemEval eight sessions: arms and guarantees	§4	lme_cc8_arms.jsonl	d0af9fbb
Conflict and order cells, LongMemEval full history: arms	§4	lme_fh_cc_arms.jsonl	19f58226
Conflict cells, STALE: arms	§4	stale_cc_arms.jsonl	45e13e2b
Conflict cells, MemoryAgentBench: arms and control questions	§4	mab_cc_arms.jsonl	51565150
Every file of the unified runs (reads, verdicts, systems, Parts G and H), hashed	§4	manifest_v1.json	b3ffeabe
Position preferences, weights, predictions (Tables 28, 42)	§4	e20_results_v1.json	0335426e
Memory systems: items and ingestion orders	§5	systems_items.json	a3ab842b

Claim	Where	Artifact	SHA-256
Memory systems (Figure 3)	§5	`e20_results_v1.json`	`0335426e`
Competing values for answer classification, checked by the authors	App. K	`competing_values.json`	`48d3af42`
Relative time expressions	App. K	`relative_time.json`	`c6b67e89`
Oracle gain difference; five random orders (Tables 12, 22)	§6	`review_extras_v1.json`	`2c2627e1`
README ordering invariant versus the released files	§3	`e2_hygiene_diff_v1.json`	`587fcd65`
Complete deviation log	App. A	`PROTOCOL_EVENTS_FULL.md`	`899a4ff1`
Current-state subset, ties and orientation counts	App. L	`analysis_v7.json`	`efb4fd14`
Missing outputs and paired retrieval checks	App. L	`sensitivity_audit.json`	`b30349ee`

K DESIGN DETAILS OF THE UNIFIED RUNS

The runs of §4–§5 follow one protocol document written before any of them (with five dated amendments, each stating what had been seen when it was written); deviations are PE-77 to PE-96 in the log (Appendix A).

Readers and providers. Every pass pins one provider with fallbacks disabled and records the serving provider of every call (Table 27); all cells that one analysis compares come from one pass. Temperature is 0. A reasoning reader whose answer is empty because reasoning consumed the budget is retried once at the larger budget in parentheses. *Claude Opus 5.* Through the provider's default, Opus 5 reasons before answering, and a first pass lost most of its answer budget to it (empty answers on 97 of 245 LongMemEval prompts); those answers are kept in the supplementary material and not analysed. As in every earlier Opus run in this project, reasoning is switched off and 400 answer tokens are allowed; this pass was sent through the provider's batch interface, which serves the same model at half price. *DeepSeek-Flash* is V4.1-Flash on every row (the model the earlier official-API runs were redirected to); the pinned host is Fireworks.

Table 27: Reader settings of the unified runs. Reasoning: as requested from the provider (–: not requested); the reasoning tokens providers reported are in Table 38. Budgets: answer tokens on LongMemEval / STALE / MemoryAgentBench, with the retry budget in parentheses; Claude Haiku 4.5's MemoryAgentBench cells were re-run with 200 (PE-91).

Setting	Model	Reasoning	Provider	Budgets
Claude Opus 5	anthropic/claude-opus-5	off	Anthropic	400 / 16 / 32
GPT-5.6 Sol	openai/gpt-5.6-sol	default	OpenAI	8000 (16000)
Gemini 3.8 Flash	google/gemini-3.8-flash	default	Google AI Studio (flex)	8000 (16000)
Claude Haiku 4.5	anthropic/claude-haiku-4.5	–	Anthropic	200 / 16 (64) / 32
Gemini 2.5 Flash	google/gemini-2.5-flash	–	Google AI Studio	200 / 16 (64) / 32
Qwen3-235B	qwen/qwen3-235b-a22b-2507	–	DeepInfra (fp8)	200 / 16 (64) / 32
GPT-4o-mini	openai/gpt-4o-mini-2024-07-18	–	OpenAI	200 / 16 (64) / 32
DeepSeek-Flash, thinking off	deepseek/deepseek-v4.1-flash	off	Fireworks	200 / 16 (64) / 32
DeepSeek-Flash, thinking on	deepseek/deepseek-v4.1-flash	on	Fireworks	8000 / 6000 (16000) / 8000
Llama-3.1-8B	meta-llama/llama-3.1-8b-instruct	–	DeepInfra (fp8)	200 / 16 (64) / 32
Llama-3.2-3B	meta-llama/llama-3.2-3b-instruct	–	Parasail (bf16)	200 / 16 (64) / 32

Which items the exchanges can use. An exchange needs the later-dated statement outside the final slot. Where it already sits there, moving the pair would also move the final-slot cue, which the design holds fixed, so those items are excluded: 21 of the 70 LongMemEval questions and 109 of STALE's 400. This restriction separates relative pair order from the final-slot cue; it defines the eligible subset rather than all knowledge-update questions.

The five cells. Per item: the native layout; the pair's date labels (serial numbers) exchanged; the pair's sessions (numbered lines) exchanged with their labels; the pair's bodies (fact texts) exchanged with the labels left in place, so both cues point to the superseded statement; and an *inversion control* that exchanges the labels of two unrelated sessions (the one nearest each pair member, never the final slot) and leaves the pair untouched. Every cell has the native prompt's length and every untouched byte, which the builders assert. Session headers are numbered by position in every layout (`### Session 1` is always the first slot), so the header carries no time. The two exchanged cells carry the same date sequence (one inversion between the pair's slots), controlling that aspect of the layout.

The inversion control checks the effect of exchanging unrelated date labels. Scores against the unchanged original reference in every cell are in Tables 29 (LongMemEval, eight sessions), 30 (full history), 31 (STALE) and 32 (MemoryAgentBench).

Tokens and what the judge sees. The exchanged cells are equal in characters by construction. With every provider pinned, the provider-reported prompt tokens are identical across the cells of every answered item for all 11 settings on the three benchmarks (4,057 of 4,057 setting-items; the only differences, in calls stopped by a provider's content filter, are in unanswered cells that the analysis drops). In the earlier, unpinned runs, differences for the two Llama models and Qwen3 varied between calls to the same prompt, which is provider routing. The official judge sees the question, the gold answer and the response, never the history (`evaluate_longmemeval_official.py`), so a layout cannot reach it except through the answer. Two hosts serve the same judge snapshot. 849 answers, 843 of them from the full history, were scored by the Azure host before the account's rate limit was known (PE-82), and 224 verdicts of the first pass do not record their host; every other verdict comes from the OpenAI host. On a seeded sample of 300 Azure-scored answers re-scored by the OpenAI host, the two agree on 98.7%.

STALE prompt. The benchmark's six sessions in the LongMemEval template (Appendix B), dates as `YYYY-MM-DD HH:MM`, then:

```
Question: Which of these describes the user's current
situation?
(A) {one memory}
(B) {the other memory}
Reply with a single letter, A or B, and nothing else.
Answer:
```

The two options are the superseded and the current memory in counterbalanced order; the emitted letter is scored.

MemoryAgentBench. We reproduce the released long-context agent (commit `fe1735de`): the fact list is split into chunks of at most 4,096 tokens, each wrapped in the benchmark's factconsolidation memorize template with one timestamp common to all chunks, followed by its long-context query template (which states that the newer fact has the larger serial number) and its system message. The harness joins sentences with spaces; we keep the dataset's line breaks, so only whitespace differs. A single-hop question is a conflict item if its gold answer ends exactly one fact that shares subject and relation with exactly one other fact; 74 items at 6k tokens and 63 at 32k qualify. Two further items have gold on the earlier, smaller-numbered fact, contrary to the benchmark's own rule, and are excluded as label noise; 61 questions without a conflicting fact are asked under the native and random layouts as a control. Answers are scored with the harness's substring exact match against the gold and, separately, against the superseded value. The benchmark's new facts are counterfactual edits of true facts; a reader that answers the true value is scored superseded in every cell, which pulls its position preference toward zero without changing its sign (Gemini 2.5 Flash and Qwen3 answer from world knowledge on most items). Claude Haiku 4.5 opens its answers with a reasoning preamble and was cut at 32 tokens (the harness allows 10), so most of its pre-registered answers are truncated (native accuracy 0.204); the three cells that the position preference needs were re-run with 200 tokens (PE-91). Given that room, Haiku quotes both facts before choosing (115 of 137 native answers contain both values), and the harness's substring match then credits the gold value even when the answer chooses the other. The re-run is therefore scored on its final answer, taken by a fixed rule (the text after the last "answer:", else the last bold span, else the whole response, cut to its first line without parentheses; PE-94); reading each full response, the classifier below names the same value in 410 of 411 answers (the exception is an item whose gold value is contained in the superseded one, which the substring match cannot separate). Scored this way, Haiku's native accuracy is 0.993 and its position preference -0.423 $[-0.540, -0.307]$; the substring match on the full responses gives $+0.066$. The same rule applied to every MemoryAgentBench setting (Table 36) leaves 9 of 11 settings untouched cell for cell and every sign as it was; answers stopped at the answer budget occur in Claude Haiku 4.5, Llama-3.1-8B and nowhere else, and no setting names both values in more than 6 of its cells. A third pass (PE-95) repeats Haiku's three cells with one line added to the user message, asking for the value alone: it answers in 5 tokens, cites a fact number in

0 of 411 answers, and its position preference is $+0.453$ $[+0.343, +0.555]$ against -0.423 with the comparison written out. Opus 5 runs the 6k context only.

Memory systems. Mem0 1.0.11 (extract and update: every extracted fact is compared with similar stored memories and added, updated, deleted or left), Mem0 2.0.19 (additive extraction, its default install without the optional keyword index), A-Mem (released source; one note per session, with its memory evolution) and Graphiti 0.30.2 (Zep's open-source engine; one episode per session on embedded Kuzu, with facts invalidated by their reference time). All write with `gpt-4o-mini-2024-07-18`. Each session's true date reaches the system through its own channel: Mem0's metadata (printed with each retrieved memory at answer time, as Mem0's own evaluation does; the open-source `add` rejects a timestamp), A-Mem's note time, Graphiti's `reference_time`. Retrieval uses each system's own search with $k = 10$ (for Graphiti, 10 facts with their validity and 10 entity summaries, the facts-and-entities context of Zep's own LongMemEval evaluation). The answer prompt lists the retrieved items in the order returned, each with its date, then the question date and the question. Two Graphiti driver bugs were patched with Graphiti's own code from two upstream fixes, recorded in the released runner. Ingestion orders: C0 date order; C1 a seeded uniform permutation (three seeds for Mem0, one otherwise); C2 date order with the pair's two arrival slots exchanged, dates true; C3 date order with the pair's dates exchanged. Graphiti ran C0, C2 and C3 on the 49 conflict questions (PE-81). A variant of Mem0 1.0 also opens each session's first user turn with its date line (`Session Date: ...`), so that its extraction and update prompts see the date, which metadata never reaches; it ran C0, C2 and C3 on the conflict questions. The answer prompt is:

```
I will give you several memories retrieved from past
conversations between you and a user, each with the date it
refers to.  Please answer the question based on the relevant
memories.
Memories:
[1] ({date}) {memory}
...
Current Date:  {question date}
Question:  {question}
Answer:
```

For Graphiti the date is the fact's validity interval ("valid from … until …"). Graphiti's hybrid search keeps facts above its default similarity threshold; on these histories it returned a median of 2 facts per question, and 11 of its 147 searches returned nothing. The store probe gives the classifier below the question, its gold answer and every item a system stored, and asks whether the gold value is among them; A-Mem keeps each session verbatim, so the probe does not apply.

Answer classification and relative time. A model outside the judge's family (`gemini-3.1-pro-preview`) read the two sessions of every pair and extracted the gold value and the competing value; the authors checked all 70 pairs and excluded four (two questions ask for both states, one states no competing value, one has the same answer in both sessions). The same model labelled answers as the gold value, the competing value, both, or neither: every answer in the eight-session cells with dates or positions exchanged, and the memory systems' GPT-4o-mini answers under C0, C2 and C3 (PE-88), and it ran the store probe, the retrieval probe (PE-92) and the check of Claude Haiku 4.5's final answers on MemoryAgentBench (PE-94). The same model and a fixed regular expression flagged relative time expressions ("last month", "two weeks ago") in each pair session; an item is flagged if either flags either session: 35 of 49 LongMemEval and 112 of 190 STALE items. On the unflagged items (Table 37), no setting resolved on all items reverses its position preference: 6 of 7 keep its sign on LongMemEval (Claude Opus 5's falls to zero on the 14 items) and 8 of 8 on STALE, and 5 and 8 of them still have intervals excluding zero on the smaller samples.

Human check of the judge. On LongMemEval, the answers that decide a position preference are those whose official verdicts differ between the dates-exchanged and the positions-exchanged version of the same setting and question: 390 answers over the eleven settings. A sample of 60, drawn with a fixed seed, was shown to an author with the question and the gold answer but without the verdict, the version or the setting, and the author marked whether each response gives the

gold answer (a response that states both values without choosing counts as not giving it). The author and the judge agree on 59 of the 60: the judge accepted 1 response that the author did not, a hedged one, and rejected 0 that the author accepted (PE-93; sample, verdicts and labels in `results/e20/human_audit/`).

Table 28: Position preference: original-reference score with time labels exchanged minus that with positions exchanged. Positive: position outweighs the time cue. Bold: nonzero after Holm correction within the benchmark. LongMemEval (49 questions, eight sessions), its full history (49; settings whose context cannot hold it are missing), STALE (190), MemoryAgentBench (137; Opus 5 on the 6k context only). ‡: re-run with 200 answer tokens and scored on the final answer; the pre-registered 32-token run cut most answers short (+0.080). †: native accuracy below 0.25; these settings answer mostly from world knowledge. Intervals and every cell in Appendix K.

Setting	LongMemEval	LongMemEval, full history	STALE	MemoryAgentBench
Claude Opus 5	**−0.286**	–	**−0.845**	**+0.757**
GPT-5.6 Sol	**−0.306**	−0.143	**−0.758**	**−0.650**
Gemini 3.8 Flash	**−0.551**	**−0.408**	**−0.984**	**−0.993**
DeepSeek-Flash (thinking on)	**−0.490**	–	**−0.841**	**−0.949**
Claude Haiku 4.5	+0.184	**+0.571**	**+0.326**	**−0.423**‡
Gemini 2.5 Flash	+0.000	+0.143	+0.079	**−0.051**†
Qwen3-235B	+0.122	−0.061	**−0.179**	**+0.109**†
DeepSeek-Flash (thinking off)	−0.041	+0.000	**+0.258**	**+0.832**
GPT-4o-mini	**+0.327**	+0.429	**+0.642**	**+0.657**
Llama-3.1-8B	**+0.327**	**+0.594**	+0.063	**+0.277**
Llama-3.2-3B	**+0.327**	–	−0.005	**+0.299**

Table 29: Original-reference score per cell, LongMemEval with eight retrieved sessions (official judge).

Setting	n	native	dates exch.	positions exch.	both exch.	inversion	position preference
Claude Opus 5	49	0.980	0.673	0.959	0.531	0.959	−0.286 [−0.429, −0.143]
GPT-5.6 Sol	49	0.980	0.592	0.898	0.327	0.980	−0.306 [−0.449, −0.163]
Gemini 3.8 Flash	49	0.980	0.429	0.980	0.388	0.980	−0.551 [−0.694, −0.408]
DeepSeek-Flash (thinking on)	49	0.939	0.429	0.918	0.306	0.918	−0.490 [−0.653, −0.327]
Claude Haiku 4.5	49	0.898	0.796	0.612	0.531	0.918	+0.184 [+0.020, +0.347]
Gemini 2.5 Flash	49	0.837	0.714	0.714	0.367	0.816	+0.000 [−0.122, +0.122]
Qwen3-235B	49	0.816	0.714	0.592	0.531	0.837	+0.122 [+0.000, +0.245]
DeepSeek-Flash (thinking off)	49	0.898	0.735	0.776	0.551	0.878	−0.041 [−0.163, +0.082]
GPT-4o-mini	49	0.857	0.816	0.490	0.347	0.837	+0.327 [+0.184, +0.469]
Llama-3.1-8B	49	0.837	0.796	0.469	0.408	0.816	+0.327 [+0.184, +0.469]
Llama-3.2-3B	49	0.776	0.755	0.429	0.429	0.776	+0.327 [+0.184, +0.469]

Table 30: Original-reference score per cell, LongMemEval with the full history (official judge). Missing settings could not hold the history or were not run (Table 5); GPT-4o-mini's 128k-token context holds 14 of the 49 histories and Llama-3.1-8B's 32. GPT-5.6 Sol's pass was completed after the others (PE-87); on the 44 questions covered before, its position preference was −0.159 [−0.311, +0.000].

Setting	n	native	dates exch.	positions exch.	both exch.	inversion	position preference
GPT-5.6 Sol	49	0.959	0.694	0.837	0.408	0.939	−0.143 [−0.286, +0.000]
Gemini 3.8 Flash	49	0.980	0.571	0.980	0.469	0.980	−0.408 [−0.551, −0.265]
Claude Haiku 4.5	49	0.796	0.816	0.245	–	–	+0.571 [+0.388, +0.735]
Gemini 2.5 Flash	49	0.755	0.714	0.571	0.449	0.776	+0.143 [+0.000, +0.286]
Qwen3-235B	49	0.735	0.612	0.673	0.673	0.735	−0.061 [−0.224, +0.102]
DeepSeek-Flash (thinking off)	49	0.837	0.735	0.735	0.612	0.857	+0.000 [−0.122, +0.122]
GPT-4o-mini	14	0.857	0.786	0.357	0.429	0.786	+0.429 [+0.083, +0.750]
Llama-3.1-8B	32	0.625	0.688	0.094	0.281	0.613	+0.594 [+0.419, +0.759]

Table 31: Original-reference score per cell, STALE forced choice (190 items).

Setting	n	native	dates exch.	positions exch.	both exch.	inversion	position preference
Claude Opus 5	174	0.989	0.126	0.971	–	–	-0.845 $[-0.897, -0.789]$
GPT-5.6 Sol	190	0.989	0.211	0.968	0.042	0.989	-0.758 $[-0.816, -0.695]$
Gemini 3.8 Flash	189	1.000	0.016	1.000	0.016	1.000	-0.984 $[-1.000, -0.963]$
DeepSeek-Flash (thinking on)	182	1.000	0.121	0.962	0.027	0.989	-0.841 $[-0.892, -0.786]$
Claude Haiku 4.5	190	0.979	0.853	0.526	0.221	0.974	$+0.326$ $[+0.253, +0.400]$
Gemini 2.5 Flash	190	0.953	0.832	0.753	0.384	0.926	$+0.079$ $[+0.011, +0.147]$
Qwen3-235B	190	0.958	0.647	0.826	0.463	0.900	-0.179 $[-0.253, -0.105]$
DeepSeek-Flash (thinking off)	190	0.937	0.795	0.537	0.179	0.921	$+0.258$ $[+0.189, +0.332]$
GPT-4o-mini	190	0.942	0.895	0.253	0.195	0.937	$+0.642$ $[+0.568, +0.711]$
Llama-3.1-8B	190	0.779	0.679	0.616	0.558	0.747	$+0.063$ $[-0.005, +0.132]$
Llama-3.2-3B	190	0.637	0.621	0.626	0.605	0.637	-0.005 $[-0.053, +0.042]$

Table 32: Original-reference score per cell, MemoryAgentBench FactConsolidation (single hop, 6k and 32k).

Setting	n	native	numbers exch.	positions exch.	both exch.	inversion	random	position preference
Claude Opus 5	74	1.000	0.892	0.135	0.000	0.986	0.743	$+0.757$ $[+0.645, +0.855]$
GPT-5.6 Sol	137	0.993	0.073	0.723	0.007	0.993	1.000	-0.650 $[-0.737, -0.555]$
Gemini 3.8 Flash	137	1.000	0.007	1.000	0.007	1.000	1.000	-0.993 $[-1.000, -0.978]$
DeepSeek-Flash (thinking on)	137	1.000	0.036	0.985	0.007	0.993	1.000	-0.949 $[-0.985, -0.912]$
Claude Haiku 4.5, 32 tokens	137	0.204	0.139	0.058	0.015	0.270	0.117	$+0.080$ $[+0.015, +0.146]$
Claude Haiku 4.5, 200 tokens	137	0.993	0.270	0.693	–	–	–	-0.423 $[-0.540, -0.307]$
Gemini 2.5 Flash	137	0.015	0.015	0.066	0.029	0.015	0.044	-0.051 $[-0.095, -0.015]$
Qwen3-235B	137	0.066	0.182	0.073	0.066	0.226	0.124	$+0.109$ $[+0.036, +0.190]$
DeepSeek-Flash (thinking off)	137	0.883	0.891	0.058	0.153	0.964	0.526	$+0.832$ $[+0.766, +0.891]$
GPT-4o-mini	137	0.796	0.818	0.161	0.044	0.847	0.394	$+0.657$ $[+0.562, +0.752]$
Llama-3.1-8B	137	0.401	0.496	0.219	0.299	0.423	0.314	$+0.277$ $[+0.161, +0.394]$
Llama-3.2-3B	137	0.336	0.431	0.131	0.161	0.372	0.270	$+0.299$ $[+0.197, +0.401]$

Table 33: Proposition 2's weights and the interaction on three scales. *: interval excludes zero. Last column: the interaction has the same sign, with intervals excluding zero, on the probability and logit scales. Claude Haiku 4.5's MemoryAgentBench row is the pre-registered 32-token run, since the re-run of PE-91 covers only the three cells the position preference needs.

Benchmark	Setting	β_0	β_d	β_p	β_{dp} (prob.)	logit	probit	robust
LongMemEval	Claude Opus 5	0.531	$+0.429$	$+0.143$	-0.122	-0.056	-0.129	
	GPT-5.6 Sol	0.327	$+0.571$	$+0.265$	-0.184^*	$+0.313$	-0.013	
	Gemini 3.8 Flash	0.388	$+0.592$	$+0.041$	-0.041	-0.165	-0.103	
	DeepSeek-Flash (thinking on)	0.306	$+0.612$	$+0.122$	-0.102	-0.245	-0.184	
	Claude Haiku 4.5	0.531	$+0.082$	$+0.265$	$+0.020$	$+0.439$	$+0.216$	
	Gemini 2.5 Flash	0.367	$+0.347$	$+0.347$	-0.224^*	-0.737	-0.484	
	Qwen3-235B	0.531	$+0.061$	$+0.184$	$+0.041$	$+0.311$	$+0.172$	
	DeepSeek-Flash (thinking off)	0.551	$+0.224$	$+0.184$	-0.061	$+0.088$	$+0.001$	
	GPT-4o-mini	0.347	$+0.143$	$+0.469$	-0.102	-0.294	-0.202	
	Llama-3.1-8B	0.408	$+0.061$	$+0.388$	-0.020	$+0.017$	-0.005	
	Llama-3.2-3B	0.429	$+0.000$	$+0.327$	$+0.020$	$+0.110$	$+0.064$	
LongMemEval (full)	GPT-5.6 Sol	0.408	$+0.429$	$+0.286$	-0.163^*	$+0.195$	-0.033	
	Gemini 3.8 Flash	0.469	$+0.510$	$+0.102$	-0.102	-0.402	-0.252	
	Gemini 2.5 Flash	0.449	$+0.122$	$+0.265$	-0.082	-0.279	-0.181	
	Qwen3-235B	0.673	$+0.000$	-0.061	$+0.122$	$+0.547$	$+0.333$	
	DeepSeek-Flash (thinking off)	0.612	$+0.122$	$+0.122$	-0.020	$+0.044$	$+0.008$	
	GPT-4o-mini	0.429	-0.071	$+0.357$	$+0.143$	$+0.698$	$+0.412$	
	Llama-3.1-8B	0.281	-0.188	$+0.406$	$+0.125$	$+0.958$	$+0.524$	
STALE	GPT-5.6 Sol	0.042	$+0.926$	$+0.168$	-0.147^*	-0.777	-0.502	
	Gemini 3.8 Flash	0.016	$+0.984$	-0.005	$+0.005$	$+0.342$	$+0.134$	
	DeepSeek-Flash (thinking on)	0.027	$+0.934$	$+0.093$	-0.055^*	$+1.238$	$+0.319$	
	Claude Haiku 4.5	0.221	$+0.305$	$+0.632$	-0.179^*	$+0.628$	$+0.116$	
	Gemini 2.5 Flash	0.384	$+0.368$	$+0.447$	-0.247^*	-0.209	-0.278	
	Qwen3-235B	0.463	$+0.363$	$+0.184$	-0.053	$+0.768^*$	$+0.299$	
	DeepSeek-Flash (thinking off)	0.179	$+0.358$	$+0.616$	-0.216^*	-0.344	-0.312	
	GPT-4o-mini	0.195	$+0.058$	$+0.700$	-0.011	$+0.298$	$+0.120$	
	Llama-3.1-8B	0.558	$+0.058$	$+0.121$	$+0.042$	$+0.269$	$+0.154$	
	Llama-3.2-3B	0.605	$+0.021$	$+0.016$	-0.005	-0.021	-0.013	
MemoryAgentBench	Claude Opus 5	0.000	$+0.135$	$+0.892$	-0.027	-0.242	-0.129	
	GPT-5.6 Sol	0.007	$+0.715$	$+0.066$	$+0.204^*$	$+1.547$	$+0.846^*$	
	Gemini 3.8 Flash	0.007	$+0.993$	$+0.000$	$+0.000$	$+0.000$	$+0.000$	
	DeepSeek-Flash (thinking on)	0.007	$+0.978$	$+0.029$	-0.015	$+0.295$	$+0.048$	
	Claude Haiku 4.5	0.015	$+0.044$	$+0.124$	$+0.022$	-0.811	-0.297	
	Gemini 2.5 Flash	0.029	$+0.036$	-0.015	-0.036	-0.785	-0.359	
	Qwen3-235B	0.066	$+0.007$	$+0.117$	-0.124^*	-1.228^*	-0.640^*	yes
	DeepSeek-Flash (thinking off)	0.153	-0.095	$+0.737$	$+0.088^*$	$+0.963^*$	$+0.493^*$	yes
	GPT-4o-mini	0.044	$+0.117$	$+0.774$	-0.139^*	-1.510^*	-0.770^*	yes
	Llama-3.1-8B	0.299	-0.080	$+0.197$	-0.015	$+0.034$	$+0.008$	
	Llama-3.2-3B	0.161	-0.029	$+0.270$	-0.066	-0.170	-0.122	

Table 34: The cost of exchanging positions alone (native minus positions exchanged). This version moves each competing statement with its own time label, so no label can contradict the text. Exploratory.

Setting	LongMemEval	STALE	MemoryAgentBench
Claude Opus 5	$+0.020$ $[-0.041, +0.102]$	$+0.017$ $[-0.006, +0.045]$	$+0.865$ $[+0.781, +0.938]$
GPT-5.6 Sol	$+0.082$ $[+0.000, +0.184]$	$+0.021$ $[+0.005, +0.042]$	$+0.270$ $[+0.197, +0.343]$
Gemini 3.8 Flash	$+0.000$ $[+0.000, +0.000]$	$+0.000$ $[+0.000, +0.000]$	$+0.000$ $[+0.000, +0.000]$
DeepSeek-Flash (thinking on)	$+0.020$ $[-0.041, +0.082]$	$+0.038$ $[+0.011, +0.069]$	$+0.015$ $[+0.000, +0.036]$
Claude Haiku 4.5	$+0.286$ $[+0.143, +0.429]$	$+0.453$ $[+0.384, +0.521]$	$+0.299$ $[+0.226, +0.372]$
Gemini 2.5 Flash	$+0.122$ $[+0.000, +0.265]$	$+0.200$ $[+0.142, +0.263]$	-0.051 $[-0.095, -0.015]$
Qwen3-235B	$+0.224$ $[+0.102, +0.347]$	$+0.132$ $[+0.079, +0.184]$	-0.007 $[-0.066, +0.051]$
DeepSeek-Flash (thinking off)	$+0.122$ $[+0.020, +0.224]$	$+0.400$ $[+0.332, +0.468]$	$+0.825$ $[+0.759, +0.883]$
GPT-4o-mini	$+0.367$ $[+0.204, +0.531]$	$+0.689$ $[+0.621, +0.753]$	$+0.635$ $[+0.540, +0.723]$
Llama-3.1-8B	$+0.367$ $[+0.204, +0.531]$	$+0.163$ $[+0.100, +0.226]$	$+0.182$ $[+0.066, +0.292]$
Llama-3.2-3B	$+0.347$ $[+0.204, +0.490]$	$+0.011$ $[-0.037, +0.058]$	$+0.204$ $[+0.109, +0.299]$

Table 35: LongMemEval position preference scored by the official judge and by the answer classifier from another model family (`gemini-3.1-pro-preview`), counting an answer as correct only when the classifier labels it the gold value; 47 questions with checked values. Exploratory.

Setting	Official judge	Classifier
Claude Opus 5	-0.286 [-0.429, -0.143]	-0.383 [-0.533, -0.229]
GPT-5.6 Sol	-0.306 [-0.449, -0.163]	-0.362 [-0.511, -0.213]
Gemini 3.8 Flash	-0.551 [-0.694, -0.408]	-0.745 [-0.867, -0.617]
DeepSeek-Flash (thinking on)	-0.490 [-0.653, -0.327]	-0.553 [-0.702, -0.395]
Claude Haiku 4.5	$+0.184$ [$+0.020$, $+0.347$]	$+0.213$ [$+0.043$, $+0.378$]
Gemini 2.5 Flash	$+0.000$ [-0.122, $+0.122$]	$+0.064$ [-0.085, $+0.213$]
Qwen3-235B	$+0.122$ [$+0.000$, $+0.245$]	$+0.085$ [-0.043, $+0.213$]
DeepSeek-Flash (thinking off)	-0.041 [-0.163, $+0.082$]	-0.043 [-0.174, $+0.085$]
GPT-4o-mini	$+0.327$ [$+0.184$, $+0.469$]	$+0.340$ [$+0.184$, $+0.500$]
Llama-3.1-8B	$+0.327$ [$+0.184$, $+0.469$]	$+0.319$ [$+0.174$, $+0.468$]
Llama-3.2-3B	$+0.327$ [$+0.184$, $+0.469$]	$+0.298$ [$+0.156$, $+0.444$]

Table 36: MemoryAgentBench scored by the harness's substring match and by the final-answer rule of PE-94, with the answers stopped at the answer budget, the answers that name both values, and the cells whose score the rule changes. Only the two settings whose answers are cut have cells that change, and no sign moves. Opus 5's budget is not recorded in its batch summary. Exploratory.

Setting	budget	at budget	both values	cells changed	harness	final answer
Claude Opus 5	–	0/222	3	0	$+0.757$	$+0.757$ [$+0.645$, $+0.855$]
GPT-5.6 Sol	8000	0/411	2	0	-0.650	-0.650 [-0.737, -0.555]
Gemini 3.8 Flash	8000	0/411	2	0	-0.993	-0.993 [-1.000, -0.978]
DeepSeek-Flash (thinking on)	8000	0/411	3	0	-0.949	-0.949 [-0.985, -0.912]
Claude Haiku 4.5	32	369/411	0	4	$+0.080$	$+0.066$ [$+0.000$, $+0.131$]
Gemini 2.5 Flash	32	0/411	6	0	-0.051	-0.051 [-0.095, -0.015]
Qwen3-235B	32	0/411	6	0	$+0.109$	$+0.109$ [$+0.036$, $+0.190$]
DeepSeek-Flash (thinking off)	32	0/411	2	0	$+0.832$	$+0.832$ [$+0.766$, $+0.891$]
GPT-4o-mini	32	0/411	2	0	$+0.657$	$+0.657$ [$+0.562$, $+0.752$]
Llama-3.1-8B	32	25/411	2	17	$+0.277$	$+0.226$ [$+0.109$, $+0.343$]
Llama-3.2-3B	32	0/411	2	0	$+0.299$	$+0.299$ [$+0.197$, $+0.401$]

Table 37: Position preference on all items and on the items without relative time expressions (14 of the 49 LongMemEval items, 78 of the 190 STALE items), with 95% intervals for the latter. Exploratory.

	LongMemEval		STALE	
Setting	all items	without relative time	all items	without relative time
Claude Opus 5	-0.286	$+0.000$ [-0.286, $+0.286$]	-0.845	-0.877 [-0.947, -0.795]
GPT-5.6 Sol	-0.306	-0.286 [-0.538, -0.067]	-0.758	-0.821 [-0.901, -0.732]
Gemini 3.8 Flash	-0.551	-0.357 [-0.625, -0.100]	-0.984	-0.987 [-1.000, -0.957]
DeepSeek-Flash (thinking on)	-0.490	-0.143 [-0.429, $+0.133$]	-0.841	-0.867 [-0.939, -0.785]
Claude Haiku 4.5	$+0.184$	$+0.071$ [-0.167, $+0.333$]	$+0.326$	$+0.333$ [$+0.225$, $+0.444$]
Gemini 2.5 Flash	$+0.000$	$+0.143$ [$+0.000$, $+0.357$]	$+0.079$	$+0.090$ [-0.013, $+0.193$]
Qwen3-235B	$+0.122$	$+0.143$ [$+0.000$, $+0.357$]	-0.179	-0.128 [-0.239, -0.015]
DeepSeek-Flash (thinking off)	-0.041	$+0.000$ [-0.200, $+0.200$]	$+0.258$	$+0.269$ [$+0.174$, $+0.370$]
GPT-4o-mini	$+0.327$	$+0.286$ [$+0.067$, $+0.545$]	$+0.642$	$+0.692$ [$+0.582$, $+0.797$]
Llama-3.1-8B	$+0.327$	$+0.429$ [$+0.167$, $+0.700$]	$+0.063$	$+0.128$ [$+0.014$, $+0.244$]
Llama-3.2-3B	$+0.327$	$+0.429$ [$+0.167$, $+0.700$]	-0.005	-0.077 [-0.149, -0.013]

Table 38: Reasoning and the cue a setting follows. Left: median provider-reported reasoning tokens per answered call (GPT-5.6 Sol reasons on 50% to 59% of calls). Right: the cue followed after Holm correction within the benchmark (–: not resolved). Exploratory.

	Reasoning tokens			Cue followed		
Setting	LongMemEval	STALE	MemoryAgentBench	LongMemEval	STALE	MemoryAgentBench
Claude Opus 5	0	0	0	time cue	time cue	position
GPT-5.6 Sol	41	18	27	time cue	time cue	time cue
Gemini 3.8 Flash	355	439	231	time cue	time cue	time cue
DeepSeek-Flash (thinking on)	281	610	94	time cue	time cue	time cue
Claude Haiku 4.5	0	0	0	–	position	time cue
Gemini 2.5 Flash	0	0	0	–	–	time cue
Qwen3-235B	0	0	0	–	time cue	position
DeepSeek-Flash (thinking off)	0	0	0	–	position	position
GPT-4o-mini	0	0	0	position	position	position
Llama-3.1-8B	0	0	0	position	–	position
Llama-3.2-3B	0	0	0	position	–	position

Table 39: Memory systems, every answerer. Left: the 49 conflict questions under date order (C0), arrival exchanged (C2) and dates exchanged (C3), and the position preference C3 − C2. Right: all 70 questions under date order and random order (C1, mean over seeds), and the cost of the random order (C0 − C1) where the pair arrived reversed and where its order was preserved. The date-line variant ran C0, C2 and C3 only; Graphiti ran no random order (PE-81).

		Conflict questions				All questions			
System	Answerer	C0	C2	C3	Pos. pref.	C0	C1	reversed	preserved
Mem0 1.0 (extract, update)	GPT-4o-mini	0.571	0.510	0.531	$+0.020$ $[-0.102, +0.143]$	0.586	0.567	$+0.046$	-0.025
	GPT-5.6 Sol	0.633	0.531	0.551	$+0.020$ $[-0.102, +0.143]$	0.657	0.619	$+0.060$	$+0.014$
	Gemini 3.8 Flash	0.571	0.551	0.551	$+0.000$ $[-0.122, +0.122]$	0.614	0.619	$+0.060$	-0.063
date line in text	GPT-4o-mini	0.612	0.490	0.490	$+0.000$ $[-0.143, +0.143]$	0.612	–	–	–
Mem0 2.0 (additive)	GPT-4o-mini	0.857	0.694	0.653	-0.041 $[-0.204, +0.122]$	0.857	0.748	$+0.055$	$+0.120$
	GPT-5.6 Sol	0.878	0.735	0.714	-0.020 $[-0.163, +0.122]$	0.871	0.805	$+0.044$	$+0.060$
	Gemini 3.8 Flash	0.918	0.755	0.735	-0.020 $[-0.163, +0.143]$	0.900	0.824	$+0.074$	$+0.074$
A-Mem	GPT-4o-mini	0.490	0.592	0.490	-0.102 $[-0.184, -0.020]$	0.529	0.571	-0.026	-0.062
	GPT-5.6 Sol	0.837	0.816	0.551	-0.265 $[-0.388, -0.143]$	0.829	0.814	$+0.000$	$+0.031$
	Gemini 3.8 Flash	0.898	0.918	0.429	-0.490 $[-0.633, -0.347]$	0.929	0.943	-0.026	$+0.000$
Graphiti (Zep)	GPT-4o-mini	0.122	0.327	0.102	-0.224 $[-0.367, -0.082]$	0.122	–	–	–
	GPT-5.6 Sol	0.143	0.306	0.122	-0.184 $[-0.327, -0.061]$	0.143	–	–	–
	Gemini 3.8 Flash	0.163	0.347	0.102	-0.245 $[-0.367, -0.122]$	0.163	–	–	–

Table 40: What each system's search put in front of the answering model (GPT-4o-mini) on the conflict questions with checked values: the questions whose ten retrieved items state the gold value, those whose items state only the competing value, accuracy, and accuracy on the questions where the gold value was retrieved. The Part G model read the retrieved items (PE-92); A-Mem keeps whole sessions and is traced by session dates in §5. Exploratory.

System	Order	n	gold retrieved	only competing	accuracy	given gold retrieved
Mem0 1.0 (extract, update)	C0	47	31	12	0.574	0.871
	C2	47	25	17	0.511	0.960
	C3	47	30	14	0.532	0.833
date line in text	C0	47	30	12	0.617	0.967
	C2	47	26	15	0.489	0.885
	C3	46	26	13	0.500	0.885
Mem0 2.0 (additive)	C0	47	41	4	0.851	0.927
	C2	46	42	4	0.696	0.762
	C3	47	41	6	0.660	0.732
Graphiti (Zep)	C0	47	8	8	0.106	0.625
	C2	47	16	5	0.319	0.875
	C3	47	7	5	0.106	0.714

Table 41: The layout-by-deletion crossing of §6 for the readers added in these runs (70 eligible questions, official judge): the value of the dates under the native layout (D_{nat}) and with the superseded session last (D_{adv}), and $M = D_{\text{adv}} - D_{\text{nat}}$ on the probability and logit scales (0.5 added to each count). Opus 5 has 67 complete questions after provider filtering; its dated cells come from earlier runs (PE-16).

Reader	n	D_{nat}	D_{adv}	M	M (logit)
Claude Opus 5	67	$+0.045$ $[-0.015, +0.114]$	$+0.134$ $[+0.029, +0.246]$	$+0.090$ $[-0.029, +0.206]$	-0.118 $[-1.967, +1.716]$
GPT-5.6 Sol	70	-0.014 $[-0.071, +0.029]$	$+0.386$ $[+0.271, +0.500]$	$+0.400$ $[+0.286, +0.514]$	$+2.029$ $[+1.075, +3.352]$
Gemini 3.8 Flash	70	-0.014 $[-0.043, +0.000]$	$+0.200$ $[+0.100, +0.300]$	$+0.214$ $[+0.129, +0.314]$	$+2.641$ $[+1.650, +4.047]$

Table 42: Every prediction of the unified runs, as written in the protocol before them, and its outcome with 95% intervals. These are the confirmatory tests. The protocol also fixed the measures and corrections of the other tables of this appendix; analyses added after the runs are marked exploratory. "Resolved": interval excluding zero.

	Prediction	Outcome
P1	Claude Opus 5 follows the dates on LongMemEval: position preference < 0	held: -0.286 $[-0.429, -0.143]$
P2	Signs agree between LongMemEval and each other benchmark for every setting resolved on both	failed: 17 of 18 pairs agree; not Claude Opus 5 (LongMemEval -0.286, other $+0.757$)
P3	MemoryAgentBench: GPT-5.6 Sol and Gemini 3.8 Flash follow the numbers; GPT-4o-mini and Llama-3.1-8B follow position (Holm)	held: GPT-5.6 Sol: cue; Gemini 3.8 Flash: cue; GPT-4o-mini: position; Llama-3.1-8B: position
P4	Full history: GPT-4o-mini follows position; Gemini 3.8 Flash and GPT-5.6 Sol the dates	failed: GPT-4o-mini $+0.429$ $[+0.083, +0.750]$ ($n = 14$); Gemini 3.8 Flash -0.408 $[-0.551, -0.265]$ ($n = 49$); GPT-5.6 Sol -0.143 $[-0.286, +0.000]$ ($n = 49$)
P5	The inversion control moves accuracy by less than half of each resolved position preference	failed: 28 of 30; not Claude Haiku 4.5 on MemoryAgentBench, Qwen3-235B on MemoryAgentBench
P6	MemoryAgentBench: a random order costs more on conflict items than on control questions (pooled)	held: -0.101 $[-0.133, -0.069]$
P7	Date readers: positions exchanged beats the superseded session moved last (pooled)	held: $+0.020$ $[+0.005, +0.041]$
P8	Mem0 1.0 follows arrival: position preference (C3 $-$ C2) > 0	failed: $+0.020$ $[-0.102, +0.143]$
P9	Graphiti's position preference is below Mem0 1.0's (paired)	held: -0.245 $[-0.429, -0.061]$
P10	Mem0 1.0 stores the current value on fewer questions in C2 than in C0 (paired)	failed: -0.109 $[-0.234, +0.021]$
P11	Mem0 1.0: random ingestion lowers accuracy where the pair arrives reversed	failed: $+0.046$ $[-0.081, +0.175]$
P12	HaluMem: random ingestion hurts update questions more (Part F)	not tested: dropped (PE-83)
P13	At least 80% of the wrong answers of the three strongest position followers, positions exchanged, name the competing value	failed: GPT-4o-mini 19/24; Llama-3.1-8B 20/25; Llama-3.2-3B 25/26

L ADDITIONAL CHECKS OF THE REPORTED CONTRASTS

These analyses use recorded answers and labels. They were added after the original runs and are exploratory; they do not replace the pre-specified analyses.

Current-state scope and historical questions. The formal argument in §3 uses a question whose target is the current state. The benchmark's knowledge-update category also contains historical questions. Of the 49 eligible LongMemEval questions, six have a checked gold value in the earlier session (`0ddfec37`, `07741c44`, `50635ada`, `618f13b2`, `9bbe84a2`, `dfde3500`). Several explicitly ask for an initial or previous state. We retain the full benchmark subset and separately exclude these six using the existing value annotations, without consulting model responses to choose the subset. The remaining 43 questions retain all seven previously Holm-resolved directions. Two further questions were already flagged in the author checks: `031748ae` asks for both states, and `22d2cb42` has no competing value. Excluding these yields 41 checked current-state conflicts, which also retain all seven directions (Table 43). Results for all three subsets, including any newly resolved directions, are released as exploratory analyses. Original gold answers remain fixed in every cell. The contrast measures which intervention changes those answers, rather than scoring correctness against a newly assigned answer after a date swap.

Table 43: LongMemEval position preference on all 49 eligible questions and on the 41 checked current-state conflicts. Intervals are pointwise 95% item-bootstrap intervals; the last column uses Holm correction across the eleven settings on the 41-item subset. Exploratory.

Setting	All 49	Subset of 41	Subset interval	Resolved cue
Claude Opus 5	-0.286	-0.268	$[-0.439, -0.098]$	time
GPT-5.6 Sol	-0.306	-0.341	$[-0.512, -0.171]$	time
Gemini 3.8 Flash	-0.551	-0.610	$[-0.756, -0.463]$	time
Claude Haiku 4.5	$+0.184$	$+0.220$	$[+0.073, +0.390]$	position
Gemini 2.5 Flash	$+0.000$	$+0.024$	$[-0.122, +0.171]$	unresolved
Qwen3-235B	$+0.122$	$+0.171$	$[+0.049, +0.317]$	position
GPT-4o-mini	$+0.327$	$+0.390$	$[+0.244, +0.537]$	position
DeepSeek-Flash (off)	-0.041	-0.049	$[-0.195, +0.098]$	unresolved
DeepSeek-Flash (on)	-0.490	-0.561	$[-0.732, -0.390]$	time
Llama-3.1-8B	$+0.327$	$+0.366$	$[+0.220, +0.512]$	position
Llama-3.2-3B	$+0.327$	$+0.317$	$[+0.171, +0.488]$	position

The check uses 30,000 paired question-bootstrap replicates with seed 20260915, sharing one resampling matrix across settings within each subset. Two-sided bootstrap probabilities are twice the smaller zero-inclusive tail, bounded by $1/30{,}000$ and 1; Holm correction is applied separately within each eleven-setting family. Exact item lists, probabilities, input hashes and results are in `results/submission_checks/analysis_v7.json`.

Retrieval transitions on the same questions. Table 44 compares Mem0 2.0 on questions probed under both conditions. The probe checks for the gold value, which may be a historical value; it does not require the rest of the retrieved context to be identical. When the gold value remains available, GPT-4o-mini still loses accuracy under both exchanges. This rules out absence of the gold value as the sole explanation for these losses, while allowing effects of other retrieved content, its ordering and its date metadata.

Table 44: Mem0 2.0 paired retrieval audit. "Both" means the gold value is retrieved under C0 and the exchanged condition. Lost/gained refer to gold-value availability. Correct answers are GPT-4o-mini's counts on the "both" subset, scored by the official judge. Exploratory.

Contrast	Paired probes	Both	Lost	Gained	Correct: C0 $\rightarrow$ exchanged
Arrival (C2 vs C0)	46	40	1	2	$37/40 \rightarrow 30/40$
Dates (C3 vs C0)	47	41	0	0	$38/41 \rightarrow 30/41$

Missing outputs. The primary STALE estimates use complete native/date-exchanged/position-exchanged responses within each setting. The 190 selected items yield 174 complete items for

Opus, 189 for Gemini 3.8, 182 for DeepSeek with thinking, and 190 for the other settings. Opus loses responses to provider filtering; Gemini has one unparsed choice; DeepSeek reaches its output limit without a parsed choice. We also score invalid outputs as wrong on all 190 items and restrict every setting to the 166 items complete across all eleven settings. Both checks retain all eight originally Holm-resolved directions. The common-item check additionally resolves Gemini 2.5 toward position. Neither check makes the original missingness ignorable: invalid-as-wrong evaluates the output contract, while common-item estimates condition on shared complete outputs. The released `results/rewrite_sensitivity/` directory gives all item IDs, estimates, intervals and corrected probabilities. Opus's LongMemEval date-deletion crossing similarly uses 67 complete questions from 70 eligible ones after provider filtering.

Ties and orientation counts. Native scores of three of the five ranking readers are tied at $60/70$. Kendall's τ_b retains these ties: the native ranking agrees with the five-draw mean at 0.120, while each single draw agrees with the other four at 0.800–1.000. Across the five draws, 68 questions appear in the reversed group and 69 in the preserved group, with 67 shared. The orientation analysis averages draws within question and resamples question IDs jointly, retaining this overlap.