WHAT A MEMORY CAN AND CANNOT CARRY: REPRESENTATION SHAPES WHAT AN EXPERIENCE LAYER CAN HOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model agents are increasingly given persistent memory, and the usual question is how much it helps. We report an experiment in which one archive of stored observations helps, does nothing, or hurts according to what is derived from it and how that is written down. A deployed observing agent at a solar telescope accumulates nightly experience over 62 days; we ask it a quantitative question its input cannot answer — how many minutes until an obscured target is observable again — and compare memories built from that one archive: the prose its consolidation cycle generates, a stratified statistical snapshot rebuilt nightly, and the snapshot plus a note on the configuration's own past error. Against truth from future frames the prose form is indistinguishable from having no memory, while the statistical form cuts mean absolute error from 22–25 to about 16 minutes and bias from $+16$ to $+5$; the note on past error reaches 15–16 minutes and $+2$ to $+3$. Both reproduce across two runs against a pre-fixed criterion. What that rung supplies is a level, not a conditional prediction: it does not raise the slope of truth on prediction, and an arithmetic version of the correction cannot be told apart from it. Replaying the 680 recorded prompts against five models from four laboratories sharpens this: the last rung beats no memory resolvably on all five, whereas the prose form raises the bias *above* the memory-free level on four of five, failing a pre-specified criterion. The one model the prose does not harm is the one that wrote it, which invites an author effect; a writer-by-reader matrix does not support it — across this three-writer panel the ablest reader writes the prose least useful to *every* reader. A second agent supplies a zero point predicted in advance: where the memory-free configuration is already near-unbiased, every rung was run and the ladder is flat, so the effect tracks the miscalibration it corrects. No configuration reaches a one-line statistical rule, whose bias-corrected form is the strongest predictor here. Four measurement settings that each invalidated an earlier run are reported with the released harness.

1 INTRODUCTION

A language-model agent with persistent memory is now a standard design. The question usually asked of such a memory is how much it improves the agent, and the usual answer is a single number from a single run. We report a case where that question is the wrong one, because one archive of stored observations improves the agent, does nothing, or makes it measurably worse according to what is derived from it and how that is written down.

The setting is a working one rather than a simulation. An observing agent deployed on a solar telescope decides, from an all-sky camera, whether to keep observing or to pause. Each night a consolidation cycle compresses the day's events into insights that later prompts retrieve. Over sixty-two days of real operation the store accumulates several hundred records. We can therefore ask a question about a memory that was filled by operation rather than by construction.

The decision the agent actually makes is a poor probe: at the instant a cloud state is confirmed the current frame is nearly decisive, and configurations with and without memory agree at 151 of 158 decision points. We therefore pose a second question on the same replay, chosen because a

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.

single frame cannot answer it: *given that the target is obscured now, how many minutes until it is observable again?* Answering it requires generalising from previous passages, which is what an experience store is for, and the answer is continuous, so each anchor contributes an error in minutes rather than one bit.

Two framings follow from that choice, and we state them before the results because they decide how the numbers should be read. The task is a **probe rather than a benchmark**: a one-line statistical rule — repeat the last observed duration — is more accurate than every configuration we test, and we report that rather than working around it, because what the probe is for is a situation in which the current observation is insufficient and accumulated experience could in principle help. And what memory turns out to supply is **a systematic level correction rather than a conditional prediction**: without it the agent is systematically wrong by $+13$ to $+28$ minutes against a truth median of 8, and the best representation removes most of that bias without ever becoming the most accurate predictor. The last rung is a self-calibration note by design and we keep that name; the conditional view that distinguishes a level from a calibration is in Appendix E. Stated generally — that an agent accumulating experience needs knowledge of its own systematic errors and not only of the world — that is a claim about long-lived agents which one deployment cannot establish. What we do establish is the quantity it turns on, how wrong the agent is before any memory reaches it, and that this quantity orders the two ladders we completed and the third we stopped.

Our contributions are two claims, followed by three results that constrain how either claim can be measured:

1. **Representation shapes what accumulated experience can carry — and across the two completed ladders, the effect appears where the agent is miscalibrated without it.** Holding one set of nightly observations as generated prose is indistinguishable from holding no memory; supplying a stratified statistical snapshot derived from the same archive is resolvably better, and adding a note on the configuration's own past error better still. Both results reproduce across two repeated runs against a criterion fixed in advance (Section 4). The condition is not an afterthought: if a memory acts on the agent's systematic error, the ladder should vanish where the agent begins near-unbiased. A second agent supplies such a target, answered by a probability: every rung was run, none is resolvably better and one is resolvably worse. A third, answered by a count, was screened out when its memory-free rung met a criterion fixed in advance — evidence that the gate works, not a measured null (Section 7).

2. **The quality of a written memory is not reducible to task capability, and the obvious explanation of why prose fails is the wrong one.** Replayed against five models from four laboratories, the prose form raises the systematic bias above the memory-free level on four of five, and the one model it does not harm is the one that composed it — which invites the reading that prose carries its author's idiosyncrasies. A writer-by-reader matrix does not support that account: within this three-writer panel the ablest model writes the prose least useful to *every* reader, itself included, which no evaluation of one model at a time could surface (Sections 5 and 6.1).

Three further controls delimit those claims rather than add to them: the pre-specified monotonicity criterion fails specifically at the narrative rung; a reasoning-off replay separates answerability from accuracy; and repeated runs expose serving variability that can exceed some configuration effects, one interval having reversed its own conclusion on replication. The controls and the failure analyses are in Sections 5 and N.

The harness, the recorded prompts and per-call transcripts for all 6,120 calls are released, so the cross-model comparison can be re-scored without re-running the agent.

The five are MiniMax-M3, Qwen3.8-Max, GLM-5.2 and DeepSeek V4 in its Pro and Flash variants: four laboratories within one model ecosystem. That reflects provider availability in the deployment jurisdiction rather than a selection preference, and Appendix C documents the constraint, the catalogue we checked and the alternatives considered.

2 RELATED WORK

Memory for language-model agents is usually organised as a store plus a retrieval policy (Park et al., 2023; Sumers et al., 2024), and recent systems augment retrieval with consolidation, structured storage and update policies (Chhikara et al., 2025; Packer et al., 2023; Hu et al., 2026). Those works ask how memories are selected, stored or maintained. Our question is prior to selection: given the same underlying observations and the same retrieval budget, what does the *form* of the stored unit permit the model to do with it. We do implement such a policy, and it improves every measurable property of the store while the end-to-end improvement sits at the edge of detectability — a gap between store-level and task-level effects that is part of what motivates this paper.

Zhang et al. (2026) find that memories an LLM consolidates from genuinely useful experience can degrade below the no-memory baseline, and locate the regression in the consolidation step rather than in the experience. Our narrative configuration reproduces that on operational data and on a quantitative question, and the cross-model replay adds what a single model cannot show: the damage transfers across readers, and Section 6.1 attributes the variation to writer quality rather than to a match between writer and reader. In the other direction, Petrov et al. (2026) and Tian et al. (2026) argue for schema-grounded and for analytic memory that supports filtering and aggregation; both are design arguments rather than measurements.

Which axis of representation. The controlled version of this experiment does exist, and being precise about it is what locates our contribution. Zeng et al. (2024) cross four stored forms — chunks, knowledge triples, atomic facts, summaries — and a mixed one with three retrieval methods. An (2026) holds a retrieval–rerank–reasoning pipeline fixed and swaps only the store, places verbatim chunks well ahead of LLM-extracted typed artifacts on two long-conversation benchmarks, and names the mechanism: lossy distillation rather than structure as such, because accuracy tracks how much source text survives. Yuan et al. (2026) cross three write strategies with three retrieval methods and find retrieval dominant by a wide margin, with raw chunks matching or beating the costlier lossy alternatives — so on their axis, how a memory is written is worth little, which is worth stating against ourselves. Our claim is therefore not that such a measurement was missing. It is that all three vary *fidelity* — source text against a lossy derivative of it — on question answering whose answers are stated somewhere in the source, as in the long-term memory benchmarks they target (Maharana et al., 2024; Wu et al., 2025). Both of our arms are lossy: the raw per-frame record enters no prompt on any line, both forms are built from the same summarised events, and the quantity being asked for is stated in no stored record and has to be obtained by aggregating across past events, so fidelity to the source is not the axis that decides whether it can be answered. The axes are complementary rather than competing: where fidelity can still be lost it dominates; once it has been lost, the form decides which kinds of knowledge the summary can carry. The effect sizes are in Appendix P. That distinction also settles what looks like a disagreement with An (2026) over augmenting a stored form rather than replacing it, which Section 4.1 takes up where the result raises it.

The benchmarking literature on variance (Bouthillier et al., 2021) states a conclusion we hit directly: a comparison is worth only as much as its accounting for uncontrolled variation. We differ from the closest methodological relatives in holding the source archive fixed while varying what is derived from it and how that is written, on a memory produced by operation rather than written by us.

3 SETUP

Testbed. The agent is the deployed cloud-analysis component of a multi-agent controller on an operational solar telescope, described by Tong et al. (2026); nothing about it was built for this experiment. Abstractly it receives, per timestep, a vector of directional darkness features around a target, a short-term trend and a scene label, and returns a structured decision; a nightly consolidation cycle compresses the day's events into insights that later prompts retrieve. Instrument detail beyond this does not bear on the result. What matters is that the agent, its prompts and the archive all pre-date the question asked here, so this is a measurement on a system in service.

Data and truth. We replay 64,641 archived frames spanning 62 observing days. Truth is self-supervised from future frames: the remaining obscuration time at an anchor is the interval until

Table 1: Duration prediction, two repeated runs at temperature 0. Errors in minutes over 170 scorable anchors on 22 observing days, with 95% percentile bootstrap intervals over observing days. "Replicated" requires the paired statistic to fall on the same side of zero and exclude it in both runs, a rule fixed before the second run.

Predictor	MAE r1	MAE r2	Bias r1	Bias r2	Paired vs L0, r1 [95% CI]	r2 [95% CI]	Repl.
persistence	8.31	8.31	+0.0	+0.0	−15.9 [−24.5, −8.7]	−13.5 [−18.8, −8.3]	yes
climatology	11.38	11.38	−4.0	−4.0	−12.6 [−20.2, −6.9]	−10.4 [−15.0, −6.5]	yes
conditional climatology	14.24	14.24	+2.9	+2.9	−10.2 [−17.0, −4.7]	−7.3 [−11.8, −2.9]	yes
L0 no memory	24.62	22.07	+16.4	+13.3	—	—	—
L1 narrative	21.21	18.79	+11.0	+8.5	−3.5 [−9.1, +2.8]	−2.5 [−6.0, +1.1]	**no**
L2 statistics	16.29	16.66	+5.7	+5.4	−8.0 [−14.6, −2.4]	−5.5 [−9.1, −1.9]	**yes**
L3 + own past error	14.78	15.73	+2.1	+3.1	−9.8 [−16.1, −4.5]	−6.3 [−9.9, −2.7]	**yes**

the first run of three consecutive unobscured frames. Anchors still obscured when the day ends are censored, not given the last frame as an answer. Two feature channels are kept disjoint at the type level so that the quantity truth reads can never enter a prompt. Scoring uses 170 anchors over 22 observing days; intervals are percentile bootstrap resampling *observing days*, not anchors, because a single decision fixes the state of several hundred consecutive frames. A paired effect is **resolvable** when its 95% interval excludes zero, and a repeated run re-calls the same anchors rather than collecting new ones.

The 170 anchors are every scorable occlusion in the archive rather than a draw from it; Appendix A states what this design cannot resolve, and two robustness checks on it: no verdict here changes when the most influential observing day is dropped, or under any unanswered-anchor charge from 0 to 90 minutes.

The representation ladder. Four configurations differ only in memory, with the situation report held identical across all of them including the memory-free line: `L0` features only; `L1` plus the narrative nightly insights the deployed consolidation cycle produces; `L2` plus a stratified statistical snapshot rebuilt nightly over season, onset scene, onset darkness band and residual duration conditioned on elapsed time; `L3` plus a note stating the median bias of this configuration's own previous answers against their outcomes. The snapshot is computed from those same nightly observations and no raw frame enters any prompt, so the rungs are cumulative rather than a change of form at fixed content: each adds quantities derived from the archive — medians the narrative does not state, and at `L3` the agent's own past error, which the archive does not contain. Section 4.1 isolates the one increment we can.

Model-free baselines. Three causal predictors set the bar, using only anchors preceding the one being predicted: persistence (the most recent observed duration), climatology (the running median) and conditional climatology (the median residual duration of past passages that had already survived as long as the current one). The third is the statistic later handed to `L2`, included by prior commitment so that crediting the agent for it would be visibly crediting the query.

Appendix A adds two baselines that bound what this archive permits — supervised predictors fitted on the agent's own features, and a look-ahead read from the image along the target's future track — and two protocol decisions: unanswered anchors, and the integrity checks on the replayed prompts. None changes a result below; all four bound how it may be read.

4 Representation decides what the memory can carry

Table 1 reports both runs. Every entry is over the same 170 scorable anchors; the paired column compares each configuration with `no_memory` on the anchors both answered, clustered by observing day.

Three findings, in descending order of effect size.

Information sufficiency dominates. The memory-free configuration answered a median of exactly 90 minutes — the largest occlusion in the archive — on all three runs of the deployed prompt, with a mean absolute error near 85. Under the explained situation report, with no memory of any kind,

it reaches 24.6 and 22.1. A model with nothing to reason from names a number large enough to be safe.

Representation decides whether a memory can carry a quantity. With the same underlying observations and the same situation report, the narrative nightly insights give an improvement that is not resolvable in either run, the stratified snapshot one that is resolvable in both, and the self-calibration note a larger one again (Table 1). The bias chain is monotone in both runs, against $+87$ on the deployed prompt. The penalised variant preserves the ordering, so the improvement is not obtained by declining the harder cases.

The agent does not beat the statistics, and what it corrects is a level. The best configuration reaches 14.8 and 15.7 against 14.24 for the conditional median it is handed, 11.38 for a running median, 8.31 for simply repeating the last observed duration, and 7.22 for that repetition with its own rolling bias subtracted (Figure 3 in Appendix S). Memory closed most of that gap; it did not cross it. And it closed a level, not a conditional response: the slope of truth on prediction reaches 0.22 and 0.12 against that rule's 0.72, and the calibration rung does not raise it over the rung below, so describing the bias cannot be told apart from subtracting it (Appendix E).

4.1 Does the snapshot need the narrative it was added to?

The rung above adds the statistics *on top of* the narrative, so what it compares is narrative against narrative-plus-statistics, and a reader is entitled to ask whether the gain comes from the form the information takes or from adding any further derived information at all. We removed the narrative instead: the body under the memory heading is deleted from the recorded structured prompts, everything outside that one block is byte-identical, and a third arm resends the unedited prompt so the endpoint's own jitter is measured rather than assumed. The criteria, including the width at which the result would be uninformative, were fixed before the first call (Appendix H).

Removing the narrative made the predictions *better* in both of two repeated runs: the paired shift is -1.4 minutes $[-3.3, -0.1]$ and -1.0 $[-3.1, +0.5]$, and mean absolute error is 14.8 and 15.7 against 16.0 and 15.8 for this session's unedited resend arm, not Table 1's run, on a prompt 42% of the length. **The direction reproduced and the resolvability did not** — only the first run's intervals exclude zero — while the resent arm includes zero in both, so the endpoint's jitter is not what produced either result. So what survives is the weaker reading and the one the paper needs: alongside the statistics the narrative block is *not required*, and no run gives evidence that it helps. That it is a net *cost* is resolvable in one run of two, and we do not claim it. None of this contradicts An (2026), whose recommendation concerns augmenting *verbatim source text*; a nightly narrative is already a lossy summary, which is precisely the case in which their mechanism predicts it carries no fidelity worth keeping.

5 Whose memory is it? Replay across five models

Everything above rests on one language model, so the finding could belong to that model rather than to the representation. Re-running the pipeline per model would cost ten times as much and would change the prompts, so the 680 advice prompts recorded in the first run — 170 anchors $\times$ 4 configurations — were replayed verbatim against four further models from three further laboratories, at temperature 0, and re-scored with the same code. Across the 6,120 calls nothing was truncated and nothing was lost to transport, so no request limit can have acted selectively on the longer-prompt configurations; the 24 replies that failed to parse are charged to their own configuration rather than removed from the denominator. Replay integrity and the failure accounting are in Appendix B.

The replay reproduces the original result. Replaying MiniMax-M3, the model that wrote the stored memory, is not cross-model evidence but a test of the method: if it does not reproduce the first run, nothing else in this subsection means anything. It reproduces every qualitative conclusion of Section 4 for a third time — the narrative rung again unresolvable, both structured rungs again resolvable — with the ladder at 22.6, 21.3, 16.0 and 15.0 minutes.

The pre-specified criterion fails on four of five models, and the reason is the finding. The rule fixed before any cross-model call required the bias chain to fall monotonically from L0 to L3 and both structured rungs to be resolvable; by that rule the ladder holds only for the control. What the

Table 2: Cross-model replay of the same 680 recorded prompts at temperature 0, as a snapshot of one week. Bias in minutes across the ladder, and the paired difference in mean absolute error against the memory-free configuration on the anchors both answered, with 95% percentile bootstrap intervals over observing days. A pre-specified rule required the bias chain to be monotone; the row-by-row departure from it is the result, not a nuisance. The DeepSeek V4 Pro row is the weights served that week; its provider updated the alias afterwards, and Appendix R.1 shows that every conclusion drawn here also holds on the newer weights.

		Bias (min)				Paired vs `no_memory`	
Model	Lab	L0	L1	L2	L3	L2	L3
MiniMax-M3*	MiniMax	$+13.7$	$+11.0$	$+5.1$	$+2.2$	-6.8 [-11.2, -3.0]	-7.4 [-12.3, -3.3]
Qwen3.8-Max†	Alibaba	$+13.4$	$+22.9$	$+7.4$	$+1.7$	-5.7 [-12.1, $+0.5$]	-9.2 [-16.9, -2.9]
GLM-5.2†	Zhipu	$+17.4$	$+19.7$	$+5.3$	$+2.0$	-8.3 [-14.8, -3.3]	-10.5 [-16.6, -5.3]
DeepSeek V4 Pro	DeepSeek	$+16.1$	$+28.7$	$+5.9$	$+7.9$	-9.8 [-14.2, -5.5]	-7.6 [-12.3, -2.8]
DeepSeek V4 Flash	DeepSeek	$+24.4$	$+52.1$	$+13.0$	$+10.3$	-8.7 [-18.2, $+1.0$]	-9.7 [-18.0, -1.1]

L0 features only, L1 narrative nightly insights, L2 stratified statistics, L3 L2 plus self-calibration. * the model that generated the stored memory, replayed as a control on the method. † called through an aggregating platform rather than the laboratory's own endpoint.

data show (Table 2, Figure 5 in Appendix S) is not a weaker version of the predicted pattern but a differently shaped one: the narrative rung *raises* the bias above the memory-free level on four of the five models, so the chain is V-shaped rather than monotone. We report the criterion as failed and the shape as observed, rather than redefining the criterion to fit. Figure 6 in Appendix Sgives all fifteen paired intervals; this table gives the narrative rung's only as a point.

What does replicate is the claim the experiment was built to test. Holding the nightly observations as stratified statistics lowers the bias on 5 of 5 models and the paired error on 5 of 5, resolvably on 3 of 5. Adding the self-calibration note is resolvably better than no memory on *all five* models, by 7.4 to 10.5 minutes. The penalised variant preserves every ordering in the table, so none of this is obtained by declining the harder cases.

Narrative memory is not merely useless; on other models it is harmful. Two of the four cross-laboratory models are resolvably *worse* with narrative memory than with none ($+14.2$ [$+0.1$, $+34.2$] and $+30.6$ [$+2.3$, $+72.2$]). A single-model experiment could only have concluded that narrative memory was indistinguishable from none; the panel gives the stronger and less comfortable result. That the one model it does not harm is the one that wrote the prose invites an authorship reading, which Section 6 tests directly and does not sustain.

The structured effect also survives on the weakest reader, DeepSeek V4 Flash, resolvably (-9.7), so it does not require a strong model — while the narrative harm is largest there, its bias rising from $+24.4$ to $+52.1$. One pre-specified reporting guard fired, on Qwen3.8-Max, where the answer rate varies by more than the permitted 0.10 across configurations; the imbalance runs against the claim rather than for it, since the configuration answering least often is the memory-free one (Appendix B).

5.1 Does the benefit require the model to reason?

A reasoning-off replay on three of the five models separates two effects. Structured memory raises the numeric answer rate to **0.99 on all three** even without reasoning, against 0.51–0.59 at the narrative rung, whereas its accuracy benefit survives reasoning-off only on the weakest model: representation decides whether the model can answer directly, while stronger readers need reasoning to turn the same statistics into accuracy. Full scores, both scoring variants, the answer-rate guard and the serving-weight provenance are in Appendices L and M.

Three limits bound this subsection. The replayed prompts contain memory that MiniMax-M3 wrote, so what this panel measures is whether another model can *use* these representations rather than whether it would *build* comparable ones; the production side is taken up separately in Section 6, on two models, and it agrees in direction. The two aggregated models were served by a platform whose quantisation and sampling implementation are not ours to control. And the reasoning contrast covers three of the five models rather than all of them, its error comparisons carrying a fired reporting flag; what it establishes firmly is the answerability result, measured on the flagged quantity itself.

Four settings — a request timeout, a completion budget, a sampling temperature and the identity of the weights behind a hosted alias — each invalidated an earlier version of this experiment while every loss was recorded as "the agent issued no advice". Appendix N reports them, together with the replication rule that follows from the third.

6 WHO SHOULD WRITE THE MEMORY, AND WHO SHOULD READ IT

Read as authorship, that pattern is an attractive story, and it is the one we set out to confirm: a pairwise self-versus-other stage came first and appeared to establish it. Three explanations fit equally well — *authorship* (prose carries its writer's habits), *readership* (one model uses retrieved prose better whoever wrote it) and *writer quality* (one model's insights are more useful for everyone). That stage cannot separate the first from the third, because in it "the other model" is always the same model: with one alternative writer, "better on my own prose" and "better than that particular writer" are the same observation — and here that one writer is the model Section 6.1 shows to be the worst of the three. Separating them needs a matrix, and Appendix D reports the precursor at its own values.

We therefore cross three writers with three readers, holding each row fixed except for the four sentences of the experience block. One measurement made before any call belongs here rather than in an appendix, because it decides whether the manipulation has a sample size at all: under the agent's own retrieval ordering a store with no weight evolution yields *two* distinct experience blocks across the 170 anchors, so the reported runs retrieve the night preceding each anchor's day instead. Appendix F gives the construction.

6.1 THE WRITER-BY-READER MATRIX

The matrix crosses three writers with three readers on the same 170 anchors, with each cell differing only in the four sentences of the experience block and each block carrying four records — a constant fixed from the products before any call, rather than equalised to the shortest writer, which would be a function of the writer set. No insight was re-consolidated, so the matrix cost only the reading. Appendix G gives the cell-level construction.

Figure 1 is the result, and it does not support authorship.

Three readings follow, given cell by cell in Appendix G. **The diagonal wins one row of three**: only DeepSeek V4 Pro is most accurate on its own prose, MiniMax-M3 is *least* accurate on its own, and of the six paired comparisons against a reader's own prose the one that excludes zero is a reader doing *worse* on M3's prose than on its own. **One reader has the lowest error against every writer, and that is the half of the matrix that is not controlled**: comparing readers compares models, and the three readers rank in exactly their memory-free order (22.6, 25.5 and 33.4 minutes, their `L0` column in Table 2), so we report it as reader skill or plain capability without separating the two. **One writer is worst for every reader, and that half is controlled**: within a row the reader, the anchors and every byte outside four sentences are fixed, and M3's insights are last in all three rows. The ordering is not identical — Flash and Pro change places in M3's row by 0.65 minutes — so what the panel fixes is which writer is worst for everyone rather than a complete ranking, which is enough, because a writer every reader does worst on is a writer effect rather than a match between writer and reader.

Taken together these say something more useful than the story they replace, and about a design decision a deployment actually has to make. **Consolidation quality is not predicted by task capability in this panel: the strongest reader is the worst writer.** Whether *reading* skill is likewise separable from capability we cannot say from this design — that is the uncontrolled half above, whose ordering matches the memory-free ordering exactly — so of the two axes one is attributable and the other is not shown either way. The consequences for a deployment that must choose one model for both roles are taken up in Section 8.

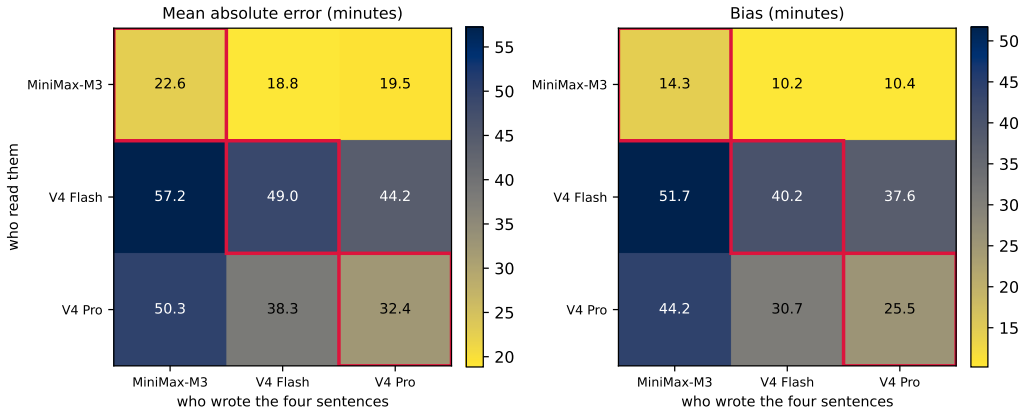

Figure 1: Every writer's nightly insights read by every reader, on the same 170 anchors over 22 observing days. Left, mean absolute error in minutes; right, signed bias. *The comparison that carries a causal reading is horizontal, within a row*: same reader, same anchors, same prompt except for who wrote the four sentences. A **column** compares different reader models and is therefore a ranking rather than an effect. The diagonal, outlined, is the reader's own prose. The diagonal is the best cell of 1 of the 3 rows. One reader has the lowest error in every column; since a column also changes the model, that is reader skill or plain capability, which this design does not separate. Of the 6 paired comparisons against a reader's own prose, 1 excludes zero over observing days.

The correction has a methodological form worth stating plainly, because we made the error before we caught it. **A self-versus-other contrast with one alternative author cannot establish an author effect**: whatever that one author's prose is worth is folded in, and here it was the worst writer of the three — the circumstance that makes the wrong conclusion look strongest. The precursor is not wrong as a measurement; it is wrong as evidence for what we read it as.

The first attempt at the stronger model was discarded because the provider re-pointed the alias partway through one arm; Appendix Q reports what that run said, why restricting it to the anchors that agreed was still not enough, and the free comparison of the two weight versions that keeping it made possible.

7 A ZERO POINT: THE SAME LADDER WHERE THERE IS NO BIAS TO REMOVE

Everything above is one question on one agent, so the ladder could be a property of this pipeline rather than of memory. The sharpest available check is not another model but another question: if the mechanism is that a memory acts on the agent's calibration, then on a target where the agent is already calibrated the same ladder should be absent — in either direction. We wrote that prediction down before running it, and it is what we observe.

A second agent on the same instrument supplies two such targets. On a probability target — three models, 477 anchors, 53 observing days — no rung improves on the memory-free configuration, whose bias is already -0.003 to -0.065 against a base rate of 0.281. On a count target, a memory-free run answers all 149 anchors — 61 of them scorable, the rest censored by nightfall — at $+0.26$ groups of bias against a threshold of 0.5 fixed before the calls, so the remaining rungs were never purchased. The two are of different kinds: the probability target is a completed null ladder, whereas the count target demonstrates the gating and estimates no memory effect, since the rungs above the first were never evaluated. Set beside the duration question's memory-free bias of $+13.4$ to $+28.2$ minutes and its resolvable gain on every model, what orders all three is the quantity measured first on each: where substantial miscalibration is there to correct, representation decides the outcome. Table 6 in Appendix K states them together, with the label artefact removed first and the two qualifications that belong with the reading.

The point is not a count of nulls. It is that the one completed ladder outside the duration question is flat, on a target chosen in advance for the reason that its agent begins calibrated, and that the

discriminating quantity — how wrong the agent is without any memory — can be measured on a few hundred calls before a ladder is designed on top of it. The count target is evidence for the second half only: the gate fired, and its own ladder went unmeasured.

8 DISCUSSION

Why prose fails, and why it is not about whose prose it is. A nightly narrative is a true and specific account of one day: three passages lasting 6, 26 and 34 minutes. It is neither a distribution nor filterable by condition at retrieval time, and those are the two properties the statistical form supplies. That is the account we can support, not a mechanism we isolated: separating the two would need a form that is a distribution but unfilterable, or filterable but not a distribution, and we ran neither. The tempting addition — that the wording carries its author's habits, so a reader fares best on its own — does not survive the matrix of Section 6.1. What that design does isolate is how good a summary a model writes, and the point of it is that this does not follow from how good the model is at the task.

Two consequences follow. Structured statistics outperform narrative-only memory across the tested readers, so an experience layer meant to outlive the model that filled it should hold statistics *as well as* whatever prose it keeps; that the prose is then dispensable rests on the one-model removal test of Section 4.1, resolved in one run of two, and is the weaker of the two. And the model that writes the nightly memory need not be the one that reads it: using one model for both, as our deployment and the obvious implementation do, fixes the writer as a side effect of choosing the decision model. What that is worth is smaller than the headline suggests — the best cell is a mixed pairing at 18.8 minutes against 22.6 for the best single-model configuration, and 3.8 minutes does not exclude zero, so it is a possibility to test rather than a gain to claim. What *is* resolvable concerns readers that are not the strongest: for a decision model that is not the ablest reader, which memory it is given matters resolvably — 18.0 minutes $[+3.5, +33.5]$ — and nightly consolidation is the cheaper of the two loops to reassign, tens of calls a night against hundreds in the decision path.

What bounds the agent lies outside its memory. A one-line rule — repeat the last observed duration — attains 8.3 minutes against 14.8 for the best configuration, and we do not read that as a ceiling on experience: persistence conditions on one variable that happens to match this site's day-scale autocorrelation, and what limits conditioning on more is the number of recorded passages per stratum rather than the representation. A larger effect than any of these came before all of them: replacing the deployed prompt's unlabelled JSON with an explained situation report moved the memory-free error from near 85 minutes to 22–25. Representation therefore matters only once information sufficiency is established, and of everything reported here this is the effect most likely to be present in another deployment.

9 LIMITATIONS

Four limitations bound what is above; Appendix J gives the reasoning behind each. First, the five readers are four laboratories but one language ecosystem, for the jurisdictional reason of Section 1, so the *magnitude* of the narrative harm need not carry to a differently trained model — the existence result, which needs one damaged reader and has two, is the more robust half. Second, the matrix is three models from two laboratories, and the replayed prompts hold memory that one model wrote, so we establish whether other models can *use* these representations and not whether they would construct comparable ones. Third, the statistics-only control is one model on constructed prompts whose endpoint exposes no serving fingerprint; it replicated in direction only, which is why we claim the narrative is not required rather than that it hurts. Fourth, the positive task is one quantitative question at one site with truth self-supervised from future frames, and the recorded passages per conditioning stratum are in single digits, which is why the finer conditioning that might beat a one-variable heuristic could not be tested.

A separate manuscript aimed at an astronomy venue covers the same deployment: the memory architecture and the operating cost of the suspend decision. It has not been submitted anywhere as of this submission, and Appendix O itemises the overlap so it can be judged rather than taken on trust.

ETHICS STATEMENT

The study replays an archive of sky images recorded by an instrument operated by the authors' institution. The images contain no people and no personally identifying information: the camera points at the sky, and the frames in which it does not — the intervals when the instrument is stowed — are excluded by a documented rule and are not otherwise examined. No human subjects were involved and no annotation labour was commissioned; the ground truth is derived from later frames of the same archive by a deterministic rule. The models were accessed through paid commercial endpoints under their published terms. Inference cost is reported rather than omitted: 6,120 recorded calls, whose token totals and wall-clock time are released with the transcripts, so the energy footprint of reproducing this work is knowable in advance. We report a negative result and a failed pre-specified criterion alongside the positive findings, which we regard as a requirement of publishing measurements of this kind rather than an optional courtesy. The authors have read the ICLR Code of Ethics and adhere to it.

REPRODUCIBILITY STATEMENT

The offline pipeline, the cross-model replay harness and the manuscript tooling are released, together with the recorded prompts and per-call transcripts for all 6,120 calls, so every number in Sections 4 and 5 can be re-scored without re-running the agent or spending any inference budget. A deterministic mock client makes the whole pipeline runnable, and byte-for-byte reproducible, with no credentials. Credentials are read from environment variables only and no released artefact contains one. The replication criterion, the primary comparisons and the bootstrap seed were committed to version control before any cross-model call was made. Appendix A states the protocol decisions that bound how the results may be read, and Appendix B the integrity checks on the replayed prompts. The code is attached as supplementary material. The repository and archive identifiers are withheld here to preserve anonymity, and will be supplied on acceptance; the stage products and per-call transcripts are data rather than code and are deposited there rather than attached.

USE OF AI TOOLS

AI coding assistants were used throughout the construction of the offline pipeline, the replay harness and the analysis code, and in drafting and revising this manuscript. Their use was extensive rather than incidental, so rather than characterise it in general terms we answer each category the call asks about.

- *Research question and hypotheses:* authors. The question of whether the form of a memory changes what an agent can do with it, and the choice to test it on a deployed instrument, are the authors'. An assistant contributed analysis when candidate targets were being screened, including the analysis that led us to reject a season-of-year memory feature as a known-harmful signal.
- *Methodology and experimental design:* yes, extensively, with every decision adjudicated and recorded by the authors before the corresponding calls. Assistants proposed and implemented the resampling unit, the unanswered-anchor variants, the precondition threshold and the replication rule; the authors fixed each in version control before the runs it governs, and the paper reports them whether or not they were met.
- *Implementation:* yes, extensively. The offline pipeline, the replay harness, the scoring code and the consistency suite were written with assistants.
- *Analysis and interpretation of results:* yes. Assistants drafted the interpretations, which the authors checked against the products. Several drafted readings were withdrawn when the products contradicted them, and the withdrawals are reported in the paper rather than removed — the pairwise author contrast in Section 6.1 and the mechanism once offered for the effect of reasoning length are both cases where an assistant's explanation did not survive the check.
- *Literature search and analysis:* yes. Retrieval and summarisation were done with assistants and every citation was verified against the record by the authors. Two otherwise relevant

works were left uncited because their author lists could not be established from an authoritative source; we did not reconstruct them.

- *Writing:* yes, extensively, for both drafting and revision.

Every quantitative claim in this paper is computed by released code from released products, and every number in the text is checked against those products by an automated consistency suite that fails the build on a mismatch; no figure in the paper was written down by a language model from its own recollection. The experimental design decisions that determine what the results mean — the pre-specified replication criterion, the choice of resampling unit, the treatment of unanswered anchors, the restriction of the second agent's label to one Stokes parameter, and the precondition threshold that stopped one experiment before it was scored — were made and recorded by the authors before the corresponding runs, and are reported in the paper whether or not they were met. Several defects that would have produced publishable-looking artefacts were found by the checks described in Appendix N rather than by inspection, and are reported.

The authors take full responsibility for the content, including for the parts of the code and prose that an assistant drafted. No tool was used to generate or alter any measurement, and no text in this manuscript describes a result that was not produced by the released pipeline.

REFERENCES

Tao An. Fidelity before structure: Verbatim chunks beat lossy artifact extraction in long-conversation LLM memory. *arXiv e-prints*, 2026. arXiv:2601.00821v4; earlier versions circulated under the title "CogCanvas: Verbatim-Grounded Artifact Extraction".

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, 2021.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*, 2025. arXiv:2504.19413.

Zhanghao Hu, Qinglin Zhu, Di Liang, Hanqi Yan, Yulan He, and Lin Gui. Beyond RAG for agent memory: Retrieval by decoupling and aggregation. *arXiv e-prints*, 2026. arXiv:2602.02007.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv e-prints*, 2023. arXiv:2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, et al. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. Article 2, 2023.

Alex Petrov, Alexander Gusak, Denis Mukha, and Dima Korolev. From unstructured recall to schema-grounded memory: Reliable AI memory via iterative, schema-aware extraction. *arXiv e-prints*, 2026. arXiv:2604.27906.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024.

Zhoujin Tian, Yao Tian, Hao Zhang, Cheng Chen, Yakun Li, Lei Zhang, and Xiaofang Zhou. Beyond retrieval: Analytic memory for multimodal agents. *arXiv e-prints*, 2026. arXiv:2607.29440.

Li-Yue Tong, Jia-Ben Lin, Yuan-Yong Deng, Ying-Zi Sun, Ming-Fu Shao, Hui Wang, and Chen Yang. JW-ASTClaw: A generalizable multi-agent framework for autonomous solar telescope and its implementation within chinese meridian project. *Research in Astronomy and Astrophysics*, 26 (8):085014, 2026.

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. arXiv:2410.10813.

Boqin Yuan, Yue Su, and Kun Yao. Diagnosing retrieval vs. utilization bottlenecks in LLM agent memory. *arXiv e-prints*, 2026. arXiv:2603.02473. MemAgents Workshop, ICLR 2026.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in LLM inference. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2506.09501.

Ruihong Zeng, Jinyuan Fang, Siwei Liu, and Zaiqiao Meng. On the structural memory of LLM agents. *arXiv e-prints*, 2024. arXiv:2412.15266.

Dylan Zhang, Yanshan Lin, Zhengkun Wu, Yihang Sun, Bingxuan Li, Dianqi Li, and Hao Peng. Useful memories become faulty when continuously updated by LLMs. *arXiv e-prints*, 2026. arXiv:2605.12978.

A Further baselines and protocol

A supervised upper bound, because the obvious objection deserves a measurement. We also fit four supervised predictors on exactly the features the agent is given — ridge regression, gradient-boosted trees, a GRU over the last twenty frames and a single-head attention pooling of the same sequence — under expanding-window cross-validation blocked by observing day, with hyper-parameters fixed in advance because 22 independent days afford no honest search. This is not a like-for-like system comparison and we do not present it as one: these models are trained on labels derived from the ground-truth channel that the agent never sees, each answers one question only, and none produces anything at all for a situation with no precedent. It is a bound on what supervised learning extracts from this archive. Cross-validation can score only the folds after the first training block, so every predictor in Table 3 — baselines and agent included — is re-scored on that common subset of 114 anchors.

Three things follow, and the third is a correction to how we would otherwise have read our own result. Every one of the four supervised models is resolvably *worse* than repeating the last observed duration, the best of them by 2.4 minutes; on 22 days of history that is what fitting buys, and the gradient-boosted trees — the most flexible of the four — are the worst by a wide margin, which is what over-fitting looks like when it is measured rather than assumed. Second, the two sequence models beat the two tabular ones, so what signal there is lies in the recent trajectory rather than in the instantaneous features; the attention model is also almost unbiased (-0.3), meaning it learnt the level of the distribution without learning the case-by-case variation. Third, the supervised models sit *between* persistence and the agent. That bounds our claim in both directions: the archive does contain a little more extractable signal than the agent captures, so "the agent is near the ceiling of this data" would be too generous; and no amount of fitting on this archive reaches a one-line rule, so "a supervised model would have solved this" is also wrong.

A second upper bound: is the answer already in the pixels? A reader may reasonably suspect that this is not a memory problem at all. The camera images the whole sky at once, and the sun's own motion converts time into angle — where the sun will be in T minutes is, at this instant, a different part of the same frame, displaced along its track by $0.2426\,T$ degrees at the solar rate measured on this archive. Under a still sky the remaining duration would then be readable from one image, with no memory and no model. Converting the sensing geometry into those units is what makes the limit legible: the features the agent receives come from an annulus 2.7 to 8.1 degrees around the sun, which is the window 11 to 33 minutes ahead, while the truth median is 8 minutes, that is 1.96 degrees — inside the disc that the ground truth reads and the agent may not. We measured the look-ahead profile at all 170 anchors regardless, in three readouts: one restricted to offsets whose one-degree patch fits inside the annulus (isolation-preserving, and reachable only between 15.3 and 29.3 minutes ahead), one reading the excluded inner disc as a bound on the image content, and one repeating the second while sampling upstream by the measured cloud velocity, which is necessary because that velocity (0.279 deg min^{-1}) is comparable to the sun's. All three are resolvably or

Table 3: Supervised upper bound. Errors in minutes over the 114 anchors that expanding-window cross-validation can score, on 22 observing days; the paired column is against persistence on the anchors both answered, with 95% percentile bootstrap intervals over observing days and * marking exclusion of zero. Hyper-parameters were fixed before fitting. The agent rows are the same answers scored in Table 1, restricted to this subset, which is why they differ from the figures quoted there.

	Predictor	MAE	Bias	Paired vs persistence
model-free	persistence	**7.02**	-0.0	—
	climatology	7.98	-1.6	$+0.96$ [-0.37, $+2.87$]
	conditional climatology	12.48	$+5.3$	$+5.46$ [$+0.83$, $+10.87$]*
supervised	attention over 20 frames	9.37	-0.3	$+2.36$ [$+0.42$, $+4.79$]*
	GRU over 20 frames	10.82	$+2.1$	$+3.81$ [$+2.21$, $+5.32$]*
	ridge regression	13.58	$+6.8$	$+6.56$ [$+2.88$, $+10.82$]*
	gradient-boosted trees	19.41	$+12.2$	$+12.39$ [$+7.12$, $+18.17$]*
agent	`L3` structured + calibration	12.42	$+4.9$	$+5.17$ [$+1.08$, $+10.17$]*
	`L2` structured	14.74	$+8.8$	$+7.45$ [$+3.58$, $+12.58$]*
	`L1` narrative	19.38	$+15.3$	$+11.79$ [$+5.24$, $+19.37$]*
	`L0` features only	19.95	$+15.3$	$+12.77$ [$+5.77$, $+23.43$]*

nominally worse than persistence (12.8, 19.7 and 14.9 minutes against 8.3), and the reason is a sensor property rather than a coding error: an absolute darkness read saturates by 3 degrees from the sun, and an exposure-invariant ring-relative read finds a median deficit of 0–4 grey levels with no separation between short and long occlusions — yet the same statistic evaluated at the sun's *own* position separates occluded from clear frames (-11 grey levels against 0, $n = 60$ each, the clear group clipped). This camera registers the extinction of the sun's own neighbourhood, not the presence of cloud a few degrees away. The spatial shortcut is therefore closed on this instrument, which is why the question remains one about what a memory can carry.

Unanswered anchors, and why 30 minutes. A configuration can decline to give a number, so every error statistic is reported twice: once dropping unanswered anchors and reporting the answer rate beside it, and once charging an unanswered anchor a fixed penalty. Reporting only the first would let a configuration improve its score by answering selectively; reporting only the second would confound accuracy with willingness. The penalty is 30 minutes, which was fixed in the single-model code before any cross-model call and is the order of a long passage rather than a tuned value: the true durations have a median of 8 minutes, a 90th percentile of 34 and a maximum of 90, so a refusal is charged roughly what a long occlusion costs. Two facts make the choice non-decisive. Every ordering we report holds under both variants, and the one place where the variants disagree is stated as such and is a finding rather than a nuisance (Section 5): with reasoning off, the dropped variant favours the narrative configuration and the penalised variant reverses it, because that configuration answered 57.6% of anchors.

Rather than defend the one value further, we sweep it. Charging 0, 5, 10, 15, 20, 30, 45, 60 and 90 minutes — from no charge at all to the longest occlusion in the archive — leaves the ordering of the four rungs unchanged at every point in both runs, so the choice is not decisive for anything the paper concludes from the ladder. The bottom of that range is included because it is absurd rather than because it is defensible: charging nothing rewards a configuration for answering only the anchors it finds easy, and the sweep shows how much that would have been worth here. The curve is in the released products; the invariance is what we rely on.

Does one observing day carry the result? Twenty-two independent days is few enough that a single day of unusual weather could produce a resolvable interval on its own, and a day-clustered bootstrap cannot reveal that, because resampling days with replacement keeps an influential day in most draws. We therefore recompute each paired difference with each day held out. In both runs every refit of every comparison keeps its sign, and for the two structured rungs — the two the paper reports as resolvable — the interval still excludes zero when the most influential day is removed altogether: -5.6 [-10.1, -1.5] and -7.6 [-12.2, -3.6] in the first run, -4.4 [-7.7, -1.2] and -5.3 [-8.6, -2.1] in the second, against full-sample values of -8.0, -9.8, -5.5 and -6.3. The narrative rung behaves the same way in the other direction: its sign is stable across refits and its interval contains zero both with and without that day, which is the same verdict we report from the

full sample. The most influential day is 20260203 or 20260205 in every case, and those are two of the four days on which the archive is missing frames — a day with fewer anchors moves a per-day mean further, which is a property of the resampling unit rather than of the weather.

Replay integrity. Each recorded prompt is digested with SHA-256 when it is read from the single-model products and again immediately before it is sent; a mismatch aborts the run rather than sending something that is not the recorded prompt. Requests use a 360-second timeout with three retries at 5, 15 and 30 seconds, applied to timeouts, connection errors and 429 or 5xx replies but not to authentication or other 4xx; roughly fifty requests hit the timeout and were recovered, and none became a permanent loss. The reporting thresholds were committed to version control as a frozen object before the first cross-model call: a configuration answering below 80% of anchors is flagged so that its answering choice is read alongside its error, and a within-model spread above 10 percentage points disqualifies that model from carrying a main claim. Both numbers are conventions rather than derivations; what matters is that they were fixed in advance, and one of them fires in the results and is reported.

Why 170 is the sample size, and what it is a sample of. This is not a benchmark comparison, and reading it as one invites the wrong objection. The 170 anchors are not drawn from a distribution that could be sampled more heavily: they are the *entire* set of scorable occlusion events in a 62-day operational archive, and the only way to enlarge the set is to wait for more weather. Every comparison is paired within those anchors — each configuration answers the same question at the same instant from the same observations — so the quantity being estimated is a within-anchor difference, not a between-sample mean, and the nuisance variance that a larger benchmark would average away is instead eliminated by construction. What the sample size does bound is the width of the interval, and the binding constraint there is not 170 but 22, the number of independent observing days, because that is the resampling unit the correlation structure permits. We therefore read effect sizes in the units the task is stated in: the paired differences that survive are about 6 to 10 minutes, against memory-free errors of 22.6 to 33.4 minutes and a true median duration of 8 minutes. That is the scale a reader should judge, rather than a standardised effect size, which over 22 clusters would be both less stable and harder to interpret than minutes. Conversely, this design cannot resolve small effects: a 2-minute difference is invisible here whatever it means, and we do not report one.

B REPLAY INTEGRITY

Prompt length is worth stating because it bounds two worries at once. Prompts run from 1,218 to 2,869 tokens, with the structured configurations about 1.6 times the memory-free one (median 2,604 and 2,644 against 1,580); completions run to a few thousand. Nothing here approaches a context limit or the 65,536-token completion ceiling, and no response was truncated, so neither the timeout nor the budget can have acted selectively on the longer-prompt configurations in this experiment — which is precisely the failure that invalidated an earlier run and is the reason the numbers are given.

Three of the five models were called at their own laboratory's endpoint and two through an aggregator, so the serving stack is controlled for the majority but not for all; the two aggregated models are named as such. Around fifty requests hit the 360-second timeout and were recovered by the retry schedule. Without it each would have become a permanently missing observation biased towards the slowest, hardest anchors — the exact mechanism of the earlier discarded run.

Every anchor lands in exactly one of *answered*, *parse failed*, *truncated* or *transport failed*, and the four counts are required to sum to the anchor count, because a silently dropped anchor is how a timeout once turned into a configuration effect. Over all 6,120 calls — the five-model panel, the three reasoning-off runs and the post-update replay — the counts are 6,096 answered and 24 parse failed, with no truncation, every response terminating on `stop` rather than at the completion ceiling, and no transport failure surviving the retries. The 24 are not evenly spread: 9 on DeepSeek V4 Flash, 5 each on MiniMax-M3 and pre-update DeepSeek V4 Pro, 3 on Pro with reasoning off, 2 on GLM-5.2 and none on either Qwen run, on Flash with reasoning off, or on the post-update Pro. They are counted as unanswered, which charges them to the configuration they occurred in rather than removing them from the denominator. The distribution is given because the concern it answers is specific: an earlier run of this experiment failed in a way that made the failure rate rise monotonically with the amount

Table 4: Within-model authorship. Errors in minutes on the narrative configuration, over the 170 anchors and 22 observing days of the main experiment; the paired column uses the anchors both arms answered, 147 for Flash and 150 for Pro, with 95% percentile bootstrap intervals over observing days and * marking exclusion of zero. The manipulation is realised 22 times, once per observing day, which is also the resampling unit. Every anchor of both models was answered by one set of weights, verified from the per-call serving fingerprints.

Model	Experience block written by	MAE	Bias	Paired, self − other
DeepSeek V4 Flash	the other model	68.06	$+63.30$	—
	itself	**46.17**	$+38.75$	-22.97 $[-66.35, +3.75]$
DeepSeek V4 Pro	the other model	49.56	$+43.53$	—
	itself	**32.69**	$+25.39$	-17.53 $[-36.68, -2.92]^*$

of injected context, so a breakdown that concentrated on the longest-prompt configurations would undermine the comparison. It does not.

C THE MODEL PANEL, AND THE TWO ALTERNATIVES THAT WERE REJECTED

Which models, and why these. The five are MiniMax-M3, Qwen3.8-Max, GLM-5.2 and DeepSeek V4 in its Pro and Flash variants: four laboratories, all of them Chinese. The reason is jurisdictional rather than one of preference or budget. The instrument, the archive and the authors are in mainland China, and the major non-Chinese model providers do not offer service in that jurisdiction; their terms exclude it. There is consequently no lawful route by which this study could have included them, and we would not report results obtained by an unlawful one.

Two alternatives were considered and rejected for stated reasons rather than left unexamined. We queried the catalogue of the aggregating platform that is reachable to us: of 236 served models, the only two whose base weights originate outside this ecosystem are Llama checkpoints distilled and post-trained by a Chinese laboratory. Instruction-following is precisely the behaviour a reader of a stored memory exercises, and for those two it comes from that post-training, so including them would have looked like coverage while adding none. Hosting an open-weight foreign model locally was rejected because it would confound the variable of interest: the hardware available to us admits only a small quantised model, so "trained outside this ecosystem" and "far weaker than every panel member" would vary together, and the panel already contains its own capability range in DeepSeek V4 Flash against V4 Pro. A confound is not a control.

The consequence for the claim has to be stated in both directions. It does not weaken the *existence* result: that a memory written by one model can be worse than no memory for a different model needs only one reader damaged by another writer's prose, and we observe it on two of the four cross-laboratory models with intervals excluding zero. It does bound the generality of the *magnitude*. Models trained within one language ecosystem may share a substantial part of their pretraining distribution, so some of what we read as an author effect could be a shared-corpus effect, and whether the size of the harm carries over to a model trained on a different distribution is a question this paper cannot answer. We therefore leave the discriminating prediction on record rather than only the caveat. Section 6.1 finds that what varies across writers is quality rather than a match with the reader, so the prediction is about that ordering: a reader from a different pretraining distribution should place these three writers in the same order, and should not favour any of them for being closer to its own corpus. A reader that ranked them differently would reopen the shared-corpus explanation, which our panel cannot rule out. Section 9 records it as the limitation we would most want removed. The constructive part is that removing it is cheap for someone who can: the recorded prompts and the scoring harness are released, and replaying a further model costs roughly half an hour of API time and no re-run of the agent. We would regard a group in another jurisdiction doing exactly that as the natural extension of this work, and the released products are organised to make it a single command.

Table 5: What a rolling bias correction adds, on the two repeated runs. *Left* is how much the correction still improves each predictor: a predictor that has already absorbed its own bias has little left to take. *Slope* regresses truth on prediction and is 1.0 for a predictor whose answers track the outcome, 0.0 for one that has learned only a level. Errors in minutes over 170 scorable anchors on 22 observing days.

	run 1			run 2		
Predictor	MAE	Left	Slope	MAE	Left	Slope
persistence	8.31	-1.09	0.72	8.31	-1.09	0.72
climatology	11.38	$+0.02$	0.61	11.38	$+0.02$	0.61
conditional climatology	14.24	-0.46	0.33	14.24	-0.46	0.33
`L0` no memory	24.62	-3.21	0.08	22.07	-4.02	0.17
`L1` narrative	21.21	-1.69	0.07	18.79	-1.24	0.06
`L2` statistics	16.29	-1.69	0.23	16.66	-1.07	0.13
`L3` + own past error	14.78	$\mathbf{-0.49}$	0.22	15.73	$\mathbf{-0.21}$	0.12

D THE PAIRWISE STAGE THE MATRIX CORRECTS

Table 4 reports the pairwise stage, and read alone it appears to confirm authorship: both models are more accurate on prose they wrote themselves, by 23.0 and 17.5 minutes, resolvably on the stronger one. These are not the same numbers as the matrix reports for the same two models in Section 6.1, and they should not be: this stage equalises the block to the shorter of two writers, five records for the stronger model and four for the weaker, whereas the matrix freezes it at four for all three, so the two stages score different prompts on overlapping but different anchor sets. Both are reported at their own values rather than reconciled. Two properties of that construction were checked before drawing any conclusion from it. The stores are closely matched on what is known to matter — 286 and 278 insights against the other author's 310, 261 and 245 distinct keys against 254, and 9.8% and 7.6% of records stating an elapsed duration against 11.0% — and the rebuilding itself accounts for very little, which is what the third arm exists to show: rebuilt against recorded is $+3.46$ [-19.37, $+27.61$] on the stronger model and $+6.99$ [-37.10, $+49.72$] on the weaker, both crossing zero. One asymmetry is not matched and is printed in the prompt: each model proposes weights for its own insights, and the medians differ, 0.75 against 0.85 for Flash.

E SUBTRACTING THE BIAS INSTEAD OF DESCRIBING IT

The last rung tells the model, in prose, how far its own earlier answers ran from the outcome. Subtracting that number is one line of arithmetic, so a reader is entitled to ask whether the rung is doing anything a subtraction would not. We answer it by making the subtraction a predictor: at each anchor, take the predictor's raw answer and remove the median residual of its own previously scored answers, floored at zero. The history closes at the anchor, following the ordering convention of the climatology and persistence baselines, so the corrected series is no more optimistic than they are, and it is defined on the same anchors as the raw one, which keeps the comparison paired.

Three things follow, and only the first is the one the objection anticipated.

The correction is a better predictor than the agent, and we report it as such. Applied to persistence it gives 7.22, below the 8.31 the paper already reports as the bar the agent does not reach. The bar moves down; the conclusion that the agent does not reach it is unchanged, and the strongest model-free predictor in this paper is now two lines of arithmetic rather than one.

Describing the bias does what subtracting it does, and that is the more useful finding. The direct contest is not the correction against the whole ladder but against the rung immediately below the calibration note: `L2` with the residual subtracted, against `L3`, which is `L2` told the same quantity in prose. The paired difference is -0.20 [-1.64, $+1.31$] in run 1 and -0.05 [-1.52, $+1.39$] in run 2, on 153 and 151 commonly answered anchors over 22 observing days. Neither run separates them. Read as a null this is unremarkable; read as a mechanism check it is informative. Had the prose rung beaten the arithmetic, it would be carrying something besides the bias, and we could not say what. Had it lost, the rung would be a worse implementation of a subtraction. It does neither, so the rung's incremental effect is *consistent with* arithmetic bias correction and we detect no benefit beyond it.

We stop at that wording rather than claiming equivalence: an equivalence claim needs a margin fixed before the measurement and a test against it, and an interval that happens to contain zero is not one. What the result does rule out, at this width, is a large additional effect from describing the quantity in prose rather than subtracting it.

An external check that the rung absorbs what it claims to absorb. The *Left* column was not defined to suit this argument — it is how much a bias correction can still improve a predictor, computable for any of them. On the rung that was told its own bias it finds -0.49 and -0.21 minutes; with no memory at all it finds -3.21 and -4.02, and the intermediate rungs sit between. The ordering holds in both runs. A rung that only appeared to calibrate would leave the correction with as much to take as `L0` does.

Why we say systematic bias and not calibration. A mean signed error near zero can be two opposite errors cancelling, so the mean is not on its own evidence of calibration. It is not here either. Within each tercile of its own prediction, `L3` runs -2.3, -5.1 and $+13.6$ minutes in run 1 and -1.6, -8.7 and $+19.4$ in run 2: still high on the long occlusions, and now low on the middle ones. The mean of $+2.1$ and $+3.1$ is what those cancel to. The slope tells the same story from the other side — it rises from 0.08 to 0.22 along the ladder, but that rise happens at `L2`, and the calibration note does not add to it. Against 0.72 for simply repeating the last duration, no rung is tracking the individual case.

This is a limit on what the mechanism is, not a retraction of the effect. The bias falls from $+16.4$ to $+2.1$ and from $+13.3$ to $+3.1$, the paired error improvements are resolvable in both runs, and an unconditional level is a real and useful thing to correct in a deployed estimator that was answering 90 minutes to every question. But a memory that supplies a level is a weaker claim than a memory that supplies a calibrated predictor, and the conditional view is what distinguishes them, so we report both.

F THE CONSTRUCTION BEHIND THE PAIRWISE STAGE

The construction gives one model two experience blocks built by identical code from identical observations, differing only in who wrote the sentences. Each model first replays the 62 nightly-compression prompts verbatim, so it writes its own insights from byte-identical input; only that nightly stage is used, because the other two consolidation stages read the store the author has been filling and would feed one author the other's material. Then the same model answers the same anchors twice, with the experience block rebuilt from each author's insights, every byte outside that block held identical and the number of records in it equalised per anchor. A third arm, the recorded prompts themselves, costs no calls and measures what rebuilding alone changes; it reproduces the published narrative bias of both models to within 0.05 minutes, which is what licenses reading the pair.

One design choice was forced by a measurement made before any call was spent. Under the agent's own retrieval ordering, a store with no weight evolution freezes: across the 170 anchors it yields *two* distinct experience blocks against 22 in the recorded prompts, so the manipulation would have had one realisation dressed as 170 samples. The reported runs instead retrieve the insights of the night preceding each anchor's day, which varies daily and makes the comparison tighter rather than looser, since both arms then summarise the same night from the same events. The number of distinct blocks is reported as a first-class quantity for that reason.

G THE WRITER-BY-READER MATRIX, CELL BY CELL

The diagonal wins one row of three. Only DeepSeek V4 Pro is most accurate on its own prose. MiniMax-M3 is *least* accurate on its own (22.6 minutes against 18.8 on Flash's and 19.5 on Pro's), and Flash does best on Pro's prose rather than its own (44.2 against 49.0). Of the six paired comparisons against a reader's own prose, one excludes zero, and it is a reader doing *worse* on M3's prose than on its own (Pro, $+18.0$ [$+3.5$, $+33.5$], over the 156 anchors both cells answered at the frozen block length of four — a different measurement from the -17.5 of Table 4, which the paragraph above that table explains, and not a rounding of it).

One reader has the lowest error against every writer — and this is the half of the matrix that is *not* controlled. M3 averages 20.3 minutes across the row against 40.3 for Pro and 50.2 for Flash, which is a large spread. But comparing readers means comparing different models, so their capability enters alongside any skill at using a memory, and this design cannot separate the two. The evidence says it does not: the three readers rank in exactly the order of the same three models with no memory at all — 22.6, 25.5 and 33.4 minutes at `L0` — so what the row means is "reader skill or plain capability, not distinguished here". We state it that way rather than as readership, because a row of a matrix is a controlled comparison and a column is not.

One writer is worst for every reader, and this is the controlled half. Every row holds the reader, the anchors and every byte outside four sentences fixed, so a difference within a row is attributable to the writer. Averaged down the columns, M3's insights are the least useful for everyone (43.4 minutes), then Flash's (35.4), then Pro's (32.0), and M3's prose is last in all three rows. The ordering is not identical across rows — Flash and Pro change places in M3's row, by 0.65 minutes — so what the panel fixes is which writer is worst for everyone rather than a complete ranking. That is enough, because a writer whom every reader does worst on is a writer effect rather than a match between writer and reader. Put beside the memory-free errors above, it says something sharper and it is the claim we rest on: **the ablest model of the three writes the least useful prose.** Evaluating models one at a time would never surface that, because it is not a property any single model exhibits on its own.

H THE STATISTICS-ONLY CRITERIA, FIXED BEFORE THE FIRST CALL

The criteria for the removal test of Section 4.1 were fixed before the first call and are released with the harness as `STATISTICS_ONLY_GATE.md`. Four branches were specified: the narrative block matters, it is a net cost, it is not required, or the measurement is uninformative. The last of these is the one that had to be written down in advance, because in this experiment the null result is the *convenient* one: an indistinguishable difference would itself support the claim that the statistics rather than the addition are doing the work. A bound wide enough to contain the 4.9-minute effect the test adjudicates — the size of the structured rung's advantage over the narrative one in the source run — therefore had to disqualify the measurement rather than be read as equivalence. It did not fire: the realised half-widths are 1.6 and 1.8 minutes. Had the criterion been chosen afterwards, the same numbers would have been available to support a conclusion the design could not licence, which is the reason for the order.

The construction is checked rather than asserted. Every "no memory" prompt in the source run keeps the memory heading above an empty body, in one identical byte pattern across all 170; the edited prompts are compared against that pattern, and everything outside the block — the statistical snapshot included — is required to be byte-identical to the recording. A first version of the edit left one newline too few and was rejected by that check.

I DISCUSSION, AT LENGTH

Two consequences follow, neither visible from a single-model experiment and the second not from a pairwise one either. The first is stated at two strengths, because the evidence comes at two strengths. Structured statistics outperform narrative-only memory across the tested readers, and relative to no memory the narrative raises the bias on four of five, so an experience layer meant to outlive the model that filled it should hold statistics *as well as* whatever prose it keeps. That the prose is then dispensable is supported by the removal test of Section 4.1 on one model, where two runs agreed on the direction and one of the two resolved it, and is the weaker of the two. And the model that writes the nightly memory need not be the one that reads it: using one model for both, as our deployment and the obvious implementation do, fixes the writer as a side effect of choosing the decision model. That side effect is the part we can size. For a decision model that is not the ablest reader, which memory it is given matters resolvably — DeepSeek V4 Pro reading the other model's prose is 18.0 minutes worse $[+3.5, +33.5]$ than reading its own — and nightly consolidation is the cheaper of the two loops to reassign, tens of calls a night against hundreds in the decision path.

The weakest model needs the calibration note, not just the statistics. On DeepSeek V4 Flash the statistics alone do not quite separate from having no memory (-8.7 [-18.2, $+1.0$]) while the statistics plus the self-calibration note do (-9.7 [-18.0, -1.1]). The point estimates barely differ, so this is about which side of an interval boundary they fall on rather than a difference in mechanism — but the direction is one a designer would want: telling a weak model how wrong it has been is cheaper than telling it more about the world.

What the agent still does not do. A one-line rule — repeat the last observed duration — attains 8.3 minutes against 14.8 for the best configuration. We do not read this as a ceiling on experience. Persistence conditions on exactly one variable and happens to match the day-scale autocorrelation of this site; our own data shows that conditioning on more is informative, and the limitation on doing so is the number of recorded passages per stratum, not the representation. What the memory does that no forward estimator can is calibrate itself from recorded outcomes: the bias runs $+87$ minutes on the deployed prompt, $+16$ with information but no memory, $+11$ with narrative memory, $+6$ with statistics and $+2$ once the agent is told how biased its own previous answers were. None of the baselines evaluated here maintains an estimate of its own historical error; one could be written, and the point is that nobody had, whereas the memory arrived at the correction from stored outcomes.

Information sufficiency dominates representation. The largest single effect came from stating what the input meant. The deployed prompt was 380 characters of JSON that never said which sector index faced the direction of travel, never gave units, and — for a question about remaining time — never said how long the current spell had already run. Replacing it with an explained situation report moved the memory-free configuration from a mean absolute error near 85 minutes, answering exactly 90 on every run, to 22–25. Most of what looked like a memory problem was a prompt that did not state what its own numbers meant, and of everything here this is the effect most likely to be present in another deployment.

J THE REASONING BEHIND EACH LIMITATION

The five models come from four laboratories but from one language ecosystem, for the jurisdictional reason given in Section 1, and this is the limitation we would most want removed. If those models share a substantial part of their pretraining distribution, then the *magnitude* of the harm we measure need not carry over to a model trained on a different distribution, and the consistent ordering of the three writers in Section 6.1 could in principle be an ordering by proximity to a shared corpus rather than by quality of summary. The existence result is more robust than the magnitude: it needs one reader that is damaged by another writer's prose, and we observe two with intervals excluding zero. The matrix is three models from two laboratories, so the writer ordering rests on three points and would be worth extending before it is relied on. Extending the panel costs about half an hour of replay per model against the released prompts, requires no access to the instrument, and is open to any group whose jurisdiction reaches a provider ours does not.

The replayed prompts contain memory that one model wrote, so we establish whether other models can *use* these representations and not whether they would *construct* comparable ones; the production side would require the full nightly cycle per model, at roughly ten times the cost. Two of the five models were reached through an aggregating platform rather than their own laboratory's endpoint, so for those two the serving configuration is not a variable we control. The reasoning contrast covers three models, but the weights behind the switch are established to different standards: for one, both sides carry the same serving fingerprint within the run; for a second the match rests on an out-of-band probe made afterwards, which is weaker evidence; for the third the endpoint returns no fingerprint at all, so that pair cannot be shown from the calls themselves to have held weights fixed.

The statistics-only measurement of Section 4.1 is one model and constructed prompts: it establishes what a model does with a statistics-only *prompt* rather than what a statistics-only *line* would have accumulated over 62 days. The nightly snapshot coincides, because it is rebuilt from the occlusion ground truth rather than from any model output, but the feedback layer does not — a real statistics-only line would write none. It was replicated, and only the direction survived, so the claim we take from it is that the narrative is not required rather than that it hurts. MiniMax's endpoint returns no serving fingerprint, so we also cannot show from the calls themselves that one set of weights answered all three arms. The task is a single quantitative question at a single site, and truth is self-

Table 6: The effect of a memory representation, against how miscalibrated the agent is without one. Each row is a different question put to a deployed agent on the same instrument; the bias column is the memory-free configuration's mean signed error, in the units of that row's answer. Intervals resample observing days throughout. The third row was stopped by a criterion fixed before the run: with the memory-free rung already near-unbiased, there is nothing for a quantitative memory to repair, so the remaining configurations were not purchased. *Anchors* counts anchors that can be *scored*, which for the third row is 61 of the 149 put to the model: a recovery still running at dusk has no later scan to confirm against, so 88 are censored, asked but never scored and never imputed. The other two rows have no censoring, so there the two counts coincide. *Resolvable on 6/6* in the first row counts six replays, not six independent models: five reader models plus one re-run of DeepSeek V4 Pro on the weights its provider served after updating the alias. The endpoints of that row's ranges, $+28.2$ and -19.0, are from that post-update replay rather than from the five-model snapshot.

Question	Answer is	Anchors	Days	Memory-free bias	Best rung against no memory
Remaining occlusion	minutes	170	22	$+13.4$ to $+28.2$	-7.4 to -19.0; resolvable on 6/6
Next scan affected	probability	477	53	-0.003 to -0.065	none resolvable; one resolvably worse
Scans until recovery	count	61	22	$+0.26$	not measured; precondition failed

supervised from future frames rather than labelled independently; the channel isolation described in Section 3 mitigates the circularity but does not remove it. Finally, the number of recorded passages per conditioning stratum is in single digits, which is the reason the finer conditioning by which a memory might exceed a one-variable heuristic could not be tested.

K THE SECOND AGENT'S TWO TARGETS

The second agent is the data-quality component of the same controller, on the same instrument. It flags wind-induced jitter from the dispersion of a limb annulus in the polarised difference image, and it makes two decisions a single frame does not settle: whether the next scan group, about thirty minutes later, will be flagged, and — once flagging has begun — how many more groups until it stops. The first answer is a probability, the second a count. Both were built on the same four-rung ladder, with the feature layer given to every rung, so that a difference between the two experiments is a difference in the target rather than in the design.

Two properties of the target had to be established before either could be read, and both turned out to matter. The flag rate varies by a factor of three across the eighteen acquisition slots of a scan, and the scan order is fixed and published in advance, so an agent shown recent frames can earn skill by learning the instrument's schedule rather than the weather; the target is therefore restricted to the circularly polarised frames, which removes the dependence and costs no anchors. And the archive's own answer is short: 57% of recoveries take exactly one further group, so the best model-free predictor is the constant 1.

Neither ladder appears, and the reason is measurable. On the probability question, three models over 477 anchors and 53 observing days place every rung within the noise of the memory-free one — twelve paired intervals, eleven crossing zero and the twelfth resolvable in the wrong direction. The memory-free rung is already at the best model-free rule (Brier 0.134–0.142 against the rule's 0.139), and its bias is -0.003 to -0.065 against a base rate of 0.281. On the recovery question a single memory-free run answers 149 of 149 anchors with a mean of 2.18 groups against a truth mean of 1.918: a bias of $+0.26$, against a threshold of 0.5 fixed in the analysis code before the calls were made. By that rule the experiment stops there, and the remaining three rungs were not run.

Why the second target is the informative one. A null on a probability can be dismissed as an instrument problem: a Brier difference is a blunter measure than an error in minutes, so perhaps the ladder was simply unmeasurable. The recovery question removes that reading. Its answer is a count with a real scale, it is scored by absolute error, and it has the same 22 observing days as the duration question that did show the ladder, so the shape of the target cannot account for the outcome. What accounts for it is that the memory-free rung arrives calibrated, at which point the pre-specified precondition fails and the rungs above it are never scored. Table 6 places the three side by side, and

the two negative outcomes in it are of different kinds, which we do not merge: on the probability target every rung was run and none is resolvable in either direction, whereas on the recovery target the ladder was not measured at all. What orders all three is the quantity measured first on each of them.

One consequence for how such an experiment should be planned. The condition is cheap to test and expensive to assume. Measuring the memory-free rung alone cost 149 calls and nineteen minutes, against the roughly 500 the full ladder would have needed on that target and the 2,000 the duration ladder did need. We would now put that measurement first on any new target: if the agent is not systematically wrong to begin with, a memory ablation on that target is measuring the absence of a precondition rather than a property of the memory.

Two qualifications belong with this. The recovery target has a narrow answer scale — the whole distribution runs from one group to seven, with a median of one — so a bias of a quarter of a group is small in absolute terms as well as against our threshold, and a target with more room might have failed the precondition less starkly. And the second agent's own harness was written after the duration experiment had established how much the explanation of the inputs matters, so its task description states the cadence, the units and the meaning of the threshold from the outset. Part of why its memory-free rung arrives near-calibrated is therefore a lesson learned on the first target and applied to the second, which is a confound in one direction and a design improvement in the other.

L THE REASONING SWITCH, MODEL BY MODEL

L.1 DOES THE BENEFIT REQUIRE THE MODEL TO REASON?

Every configuration above was answered by a model producing several thousand tokens of chain of thought, so the improvement might belong to the reasoning rather than to the representation. The control is within a single model: the same 680 prompts replayed against the same weights with the reasoning switch turned off, one boolean changed and nothing else. We ran it on three of the five — Qwen3.8-Max, and DeepSeek V4 in both its Pro and Flash variants — which costs almost nothing to add, because reasoning is the default on the DeepSeek models and the reasoning-on side therefore already exists above, and because a reasoning-off completion is a few hundred tokens rather than a few thousand. Two of the three are first-party endpoints, and Pro and Flash are a strong/weak pair from one laboratory, so the contrast is not confined to one capability level.

The switch has to be shown to have taken effect rather than merely to have been sent, since an endpoint that silently ignores an unrecognised field returns a normal response and would produce two arms that are identical and a conclusion that reasoning does not matter. Each vendor spells the switch differently, and each was verified on a short rehearsal before the full run: median completion length falls from 4,059 tokens to 182 on Qwen3.8-Max, from 1,566 to 118 on Pro and from 3,652 to 135 on Flash, with median latency falling from 36–41 seconds to about 4.4.

One distinction decides how much the arms being recorded hours apart matters. Two of the three findings below are measured *within a single run*, comparing configurations answered by whatever weights were serving at that moment, so nothing the provider does between runs can disturb them; the third compares one run against another and does require the two to be the same artefact. That is not hypothetical here: the provider re-pointed one alias between the arms of the Pro pair, which is why its reasoning-on arm was re-recorded under a separate key and why both Pro rows of Table 7 are the post-update weights `fp_v4pro_20260812`. Appendix M gives the fingerprint evidence, which two of the three pairs it establishes and which it cannot, the two harness changes that followed, and the answer-rate guard that fires on all three reasoning-off runs and is the reason the penalised column is the one to read.

The three pairs are given as Table 7 in Appendix L; every number the argument below rests on is stated here in the text.

Three things follow, and the second is the reason the penalised variant was specified in advance.

Most of the over-estimation is tied to reasoning, but not to the amount of it. (Cross-run.) Turning the chain of thought off collapses the bias on all three models: from $+13.4$ to $+2.0$ on Qwen3.8-Max, $+28.2$ to $+3.1$ on Pro and $+24.4$ to $+13.9$ on Flash at the memory-free rung, and

similarly at every other rung. On the weakest model the bias falls by more than half but stays clearly positive — Flash over-estimates whether it reasons or not, and reasoning makes it much worse. This is the finding that compares one run against another, so it is the one the provenance check above was needed for: on the two DeepSeek pairs both arms are the same weights, and on the Qwen pair that cannot be established. The obvious mechanism — that a model given room to argue talks itself towards a large number — is testable on the recorded transcripts and is not supported. Within the reasoning-on condition the length of the completion does not predict the magnitude of the answer: across the five reasoning runs the correlation between completion tokens and predicted minutes is -0.005 to -0.31, and splitting each run at its median completion length puts the *longer* completions at the *smaller* median prediction (12–15 minutes against 15–25). Whatever the reasoning mode changes, it is a property of the mode and not a dose effect of tokens spent, and we report the association without naming a mechanism.

"Reasoning off is more accurate" is an artefact of what it declined to answer, on all three models. (Within-run.) Every number in this paragraph is measured inside the reasoning-off run, comparing configurations against each other, so it does not depend on which weights the reasoning-on run met. Dropping unanswered anchors makes the reasoning-off narrative rung look like the best configuration in the study: 13.1, 11.1 and 16.7 minutes. It answered 57.6%, 50.6% and 58.8% of anchors. Charging an unanswered anchor 30 minutes moves those three to 20.3, 20.5 and 22.2 while barely moving the structured rungs, which answered 85–98%, and the ladder reappears with the same V shape on each model. Reading only the dropped column would have supported the opposite conclusion three times over, which is the clearest justification we can give for specifying both variants in advance.

The answerability benefit does not need reasoning; the accuracy benefit needs it on the stronger models but not on the weakest. (Within-run, and therefore the firmest of the three.) Both halves compare configurations inside one reasoning-off run — the structured rung against the memory-free rung, answered by the same weights within minutes of each other — so the provenance caveat does not reach them. Without reasoning, the structured representation makes the question answerable where the narrative one does not: the numeric answer rate rises from 0.51–0.59 at the narrative rung to *0.99 on all three models* at the self-calibrating rung. That is the part of the effect that survives switching reasoning off everywhere. Accuracy is the part that does not, and here three models say more than one did. On the two stronger models no paired difference against the memory-free configuration excludes zero with reasoning off (Qwen3.8-Max $+0.2$, $+1.8$, $+1.0$; Pro -0.0, -0.2, -0.2), which is the conclusion a single-model contrast supported — and on Pro the same two comparisons *with* reasoning are -19.0 [-33.5, -7.4] and -17.7 [-31.2, -6.7], so the contingency is stark rather than marginal. On Flash it does: the structured rungs are -6.0 [-9.1, -2.5] and -5.0 [-8.0, -1.6], resolvably better than no memory with the chain of thought switched off. Because all six of those intervals are within-run comparisons, this holds independently of the alias question. We therefore correct the single-model reading rather than repeat it: handing a model statistics instead of prose helps a weak model even when it answers directly, and helps a strong one only when it reasons over them. Stated as a design consequence, a quantitative memory earns its place twice over — it lets a model respond at all, and, for a model with little capacity to spare, it improves the answer without any reasoning budget.

Keeping the pre-update run rather than overwriting it also yields a controlled before-and-after measurement of one provider update on this frozen task, in which the newer weights reason 2.75 times longer, over-estimate almost twice as badly with no memory, and improve once handed the same observations as statistics, with every qualitative conclusion of the panel holding on both versions. Appendix R gives it.

M SERVING-WEIGHT PROVENANCE OF THE REASONING PAIRS

A second provenance check, which one pair failed and which is why the pair was re-run. A hosted model id is an alias, not a versioned artefact, so "the same model with one boolean changed" is an assumption about the provider and not a fact about the request. The arms of these pairs were first recorded 4 to 24 hours apart, and on one of them the provider re-pointed the alias in between: the serving fingerprint returned by the endpoint read `fp_9954b31ca7` for the reasoning-on arm of DeepSeek V4 Pro and `fp_v4pro_20260812` for its reasoning-off arm. That pair differed by the

Table 7: Three models with the reasoning switch on and off, over the same 680 recorded prompts each. Both DeepSeek pairs are version-controlled: each pair's two arms report one matching serving fingerprint. The Pro rows are therefore the post-update weights `fp_v4pro_20260812` on both sides, and differ from the Pro row of Table 2, which is the pre-update snapshot; Appendix R.1 compares the two. MAE charges an unanswered anchor 30 minutes; the dropped variant is discussed in the text and reverses the reading on the narrative rung. "Answered" is the fraction of occluded anchors given a numeric duration — every response parsed on all six runs, so these are explicit nulls and not losses. Bias in minutes.

		MAE, unanswered = 30 min				Bias				Answered			
Model	Reas.	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3
Qwen3.8-Max	on	26.0	34.1	18.9	16.2	+13.4	+22.9	+7.4	+1.7	0.82	0.88	0.95	0.94
	off	17.4	**20.3**	14.7	13.4	+2.0	−1.8	+3.0	+0.4	0.77	**0.58**	0.98	**0.99**
DeepSeek V4 Pro	on	34.8	44.5	16.9	18.0	+28.2	+35.9	+4.9	+5.2	0.92	0.91	0.95	0.96
	off	16.5	**20.5**	16.7	14.4	+3.1	+0.0	+2.9	+3.7	0.72	**0.51**	0.85	**0.99**
DeepSeek V4 Flash	on	33.0	58.1	24.9	22.7	+24.4	+52.1	+13.0	+10.3	0.89	0.91	0.93	0.91
	off	21.0	**22.2**	16.1	16.4	+13.9	+10.2	+5.1	+7.4	0.82	**0.59**	0.95	**0.99**

switch *and* by a model update. We therefore replayed the reasoning-on arm again against the post-update weights, under a separate key so that the pre-update measurements were preserved rather than overwritten, and Table 7 reports the pair in which both arms are served by `fp_v4pro_20260812`. The pre-update run is not discarded; Appendix R.1 uses it.

Two of the three pairs are now version-controlled by direct evidence: both DeepSeek arms report a single matching fingerprint, Flash's after a short probe of the same endpoint established that its alias had not moved and, in passing, that the fingerprint identifies the weights rather than the reasoning mode. The Qwen pair cannot be checked at all, because that platform returns no fingerprint; we report it as a same-model contrast on the weaker ground that no update to that model was announced in the window, and say so. Two changes followed. The harness records the fingerprint on every call — one short string, and the reason the problem was visible at all was that the affected run happened to contain three unparseable replies, which were then the only case for which the provider's fields were kept. And the report refuses to label a pair a within-model contrast unless both arms agree on the fingerprint, so the verdict is computed rather than assumed.

One distinction is worth stating because it applies to any replay against a hosted endpoint, and because it decides how much the above mattered. Two of the three findings below are measured *within a single run*, comparing configurations answered by whatever weights were serving at that moment; nothing the provider does between runs can disturb them. The third compares one run against another and does require the two to be the same artefact. Where we can, we put the load on the former.

One consequence has to be stated before the numbers, because our own pre-committed reporting rule fires on all three. With reasoning off, the numeric answer rate varies across configurations by 0.40 to 0.48, far above the 0.10 spread at which the framework refuses to let a model carry a main claim. That rule exists to catch a failure rate that tracks the configuration under test, and here it is doing its job: the reasoning-off *error* comparisons are confounded by which anchors each configuration chose to answer, which is why the penalised variant is the one to read. It does not disqualify the observation about answerability, because there the answer rate is the quantity of interest rather than a nuisance acting on something else.

N Measurement limits that decided earlier conclusions

Three settings were fixed rather than tuned, each after an earlier attempt had been discarded because the setting had become the independent variable. They are reported because none of them fails loudly.

A request timeout of 30 seconds removed 27%, 36% and 49% of answers from the three memory-bearing configurations against 2% from the memory-free one. We report that pattern without a mechanism: with the timeout raised, the memory-free configuration is the *slowest* of the four on average, so the loss was not simply longer prompts taking longer. A completion budget of 2048 tokens truncated roughly 30% of responses after the reasoning block had closed but before the an-

swer was emitted — deleting precisely the anchors the model had worked hardest on, in a different proportion per configuration. A hosted model identifier is an alias rather than a version, and the provider re-pointed one of ours between the two arms of a within-model comparison; we detected it only because that endpoint returns a serving fingerprint and because that run happened to contain a few unparseable replies, which were then the only case for which the harness kept the provider's own fields. A clean run left no evidence about its own weights at all. Three changes followed. The fingerprint is recorded on every call. The report refuses to call a pair a same-model contrast unless both arms agree on it, so the verdict is computed rather than assumed. And the affected arm was replayed against the current weights under a separate key, which preserved the earlier measurements instead of overwriting them — turning the accident into the before-and-after comparison of Section 5, in which the newer weights reason for 2.75 times as many tokens, over-estimate almost twice as badly with no memory, and are *better* once given the same observations as statistics. The analysis is also organised around the distinction this forces: a comparison made *within* one run cannot be disturbed by what the provider does between runs, while a comparison made *across* runs is only as sound as the assumption that the weights held still. A sampling temperature of 0.7 left only 6% of duration answers identical between two otherwise identical runs, with a mean per-anchor difference in absolute error larger than any effect under study; at temperature 0 the between-run spread of an unchanged configuration is 2.6 minutes.

That last figure is an aggregate, and a later experiment measured the per-anchor version of it, which is much larger. Sending byte-identical prompts twice in one session at temperature 0, against one endpoint reporting a single serving fingerprint for all 510 calls, changed the answer at 107 of 145 anchors with a median shift of 7 minutes and a maximum of 360. Greedy decoding is deterministic in principle; on a hosted endpoint it is not, and neither the temperature nor the fingerprint tells you by how much. The mechanism is established elsewhere — floating-point non-associativity under limited precision, driven by batch size and by which GPUs serve the request, and amplified in reasoning models because an early rounding difference cascades through the chain of thought (Yuan et al., 2025); what an API client can observe is only the size of the effect, which is why we measure it rather than assume it away. The consequence for design is concrete: an effect measured per anchor in a single run on such an endpoint is bounded below by this quantity, so a manipulation worth testing has to clear it. Ours did not — stripping every number out of the retrieved memory, which altered all 170 experience blocks, shifted the answers by $+7.8$ minutes $[-10.1, +32.4]$ against a same-prompt control of -1.2 $[-9.9, +5.3]$. We report the reproducibility measurement rather than the manipulation, because the measurement is what transfers.

In all three cases the framework recorded each loss as "the agent issued no advice" and continued, so nothing in the products looked like a failure. We think the transferable form of this is a reporting requirement: an ablation of a language-model agent should state its timeout, its completion budget and its sampling temperature alongside its statistics.

The stronger claim concerns replication. On an earlier version of this experiment a single run put one configuration 26.5 minutes $[-38.6, -10.9]$ below the memory-free baseline, apparently resolvable; an identical replay put the same comparison at 9.0 minutes $[-41.3, +33.9]$, not resolvable. A day-clustered bootstrap cannot reveal this, because it resamples within one realisation of the model's own sampling. **A single run of a language-model ablation, however carefully its intervals are constructed, cannot separate a configuration effect from the model's nondeterminism.** Every result in this paper is therefore reported from repeated runs against a rule fixed in advance: a paired statistic counts only if it falls on the same side of zero in both runs and excludes zero in both.

O THE ITEMISED OVERLAP WITH THE COMPANION MANUSCRIPT

Five elements are shared with the companion manuscript, generated from one source so they cannot disagree: the description of the archive and its observing days; the definition of the duration ground truth and its censoring rule; the two-run duration table (Section 4); the account of the measurement settings (Appendix N); and the dose-response table (Section 7).

Two analyses appear in both and are a contribution of neither: the supervised upper bound and the along-track look-ahead (Appendix A), which here bound what any method could extract from this archive — so that the agent's shortfall against a one-line rule is not read as a ceiling — and there bound what a fitted model could.

The second agent is divided rather than shared: that manuscript reports both targets as a result about the architecture, this one uses only the resulting dose-response row. The cross-model replay, the reasoning contrast and the matrix are absent there; the architecture, the suspend-decision ablation, the cost model and the portability analysis are absent here.

P EFFECT SIZES OF THE NEAREST NEIGHBOURS ON THE REPRESENTATION AXIS

An (2026) hold a retrieval–rerank–reasoning pipeline fixed and swap only the stored form, under six confound controls. Verbatim chunks lead LLM-extracted typed artifacts by 15.9 and 22.0 points on two long-conversation benchmarks. Their attribution is to lossy distillation rather than to structure: accuracy tracks how much of the source text survives into the store.

Yuan et al. (2026) cross three write strategies with three retrieval methods. Retrieval accounts for about 20 points of accuracy across methods against 3–8 points across write strategies, and raw chunks match or beat the costlier lossy alternatives. On that axis, how a memory is written is worth comparatively little — which is the comparison in this literature least favourable to the present paper, and is stated for that reason.

Zeng et al. (2024) cross four stored forms — chunks, knowledge triples, atomic facts and summaries — plus a mixed form, with three retrieval methods.

The reason these results do not settle our question is given in the main text: all three ask questions whose answers are stated somewhere in the source text, so the variable that decides the outcome is how much of that text survives. In our setting the raw per-frame record enters no prompt on any configuration, both arms are built from the same summarised nightly events, and the target quantity — the remaining duration of an obscuration — appears in no stored record at all. It has to be aggregated across past events, so there is no fidelity gradient to place the two arms on.

Q THE DISCARDED RUN, AND WHAT KEEPING IT MEASURED

An operational note that changed the result. The first attempt at the stronger model ran one arm to completion before the other, and the provider re-pointed the hosted alias part-way through the second: the per-call fingerprints show 87 of 170 anchors answered by one set of weights and 83 by another. Restricting to the 87 gave -11.95 $[-31.43, +4.06]$ and the unrestricted figure was -4.93. Neither is what the clean run reports. Two things caused the gap. Contamination dilutes, and the surviving 87 anchors were not a random half but the earlier half of the archive, on which the predicted quantity is seasonally different — so that estimate confounded the model with the season. The run was repeated on the settled weights, with the two arms issued pair by pair so that any further change lands on both sides, after an eight-call probe established that the alias was stable rather than alternating between serving stacks; had it been alternating, re-running arm by arm would have bought a second contaminated run.

The discarded products are retained rather than deleted, and that turns out to buy something. They are first the evidence that running both arms in one session — the mitigation we adopted after an earlier alias move *between* arms — is not sufficient on its own. They also hold answers to prompts that the clean run answered too, from the weights the update replaced: 79 of the 170 prompts in one arm are byte-identical across the two runs, the remainder differing because the block is truncated to what both arms can supply and the two runs wrote different numbers of insights per night. On those 79, over 11 observing days, the newer weights are resolvably *more* accurate on this task: 36.65 against 43.10 minutes, a paired difference of -7.37 $[-14.51, -2.34]$, with essentially unchanged completion length (median 4251 against 4227 tokens). We report that narrowly. It is one task and one metric on 79 anchors, it says nothing about whatever benchmark a report of a regression might refer to, and it exists only because a run we could not use was kept. The wider point is the one already made: the identifier is an alias, so a published result of this kind is reproducible only to the extent that the fingerprint it names is still served, and we name it.

Table 8: One hosted alias, `deepseek-v4-pro`, before and after a provider update, on identical prompts and ground truth. Errors and bias in minutes; "answered" is the fraction of occluded anchors given a numeric duration. The comparison exists only because the pre-update products were preserved under a separate key when the replacement run was made.

	Completion / latency		Bias by representation				MAE, L2 / L3	
Serving fingerprint	tokens	seconds	L0	L1	L2	L3	L2	L3
`fp_9954b31ca7` (before)	1,566	36.1	$+16.1$	$+28.7$	$+5.9$	$+7.9$	17.8	19.6
`fp_v4pro_20260812` (after)	4,309	78.3	$+28.2$	$+35.9$	$\mathbf{+4.9}$	$\mathbf{+5.2}$	**16.2**	**17.4**

R One provider update on a frozen task

R.1 What one provider update did to the same task

Keeping the pre-update run rather than overwriting it turned an accident into a measurement that is normally hard to obtain. The same 680 recorded prompts, the same self-supervised ground truth and the same scoring code were answered by one hosted model before and after its provider updated it, at the same temperature, the same completion ceiling and the same concurrency — so the elapsed times are comparable and nothing between the two runs was re-tuned, relabelled or re-picked. What differs is the weights, and the endpoint says so.

Three observations, and the second is the one we did not expect.

The newer weights think considerably longer, not shorter. Median completion rises from 1,566 tokens to 4,309 and median latency from 36.1 to 78.3 seconds, a factor of 2.75 and 2.17 on identical prompts. Reports circulating at the time of the update suggested the opposite — a denser chain of thought at lower token cost — and on this task that is not what the endpoint delivers. We note it as a single measurement on one task rather than a general claim, which is precisely what a frozen prompt set is good for: the observation is checkable because nothing else changed.

The update made the model worse without memory and better with a quantitative one. Left to reason from the situation report alone, its systematic over-estimation nearly doubles, from $+16.1$ to $+28.2$ minutes, and its narrative-memory bias worsens from $+28.7$ to $+35.9$. Given the same observations as stratified statistics, it improves: bias falls from $+5.9$ to $+4.9$ and from $+7.9$ to $+5.2$, and mean absolute error from 17.8 to 16.2 and 19.6 to 17.4 minutes. The paired advantage of the structured representation over no memory correspondingly grows from -9.8 [-14.2, -5.5] to -19.0 [-33.5, -7.4]. A model that argues at greater length talks itself further from the answer when it has nothing to anchor on, and gains more from being handed an anchor.

Every qualitative conclusion of the cross-model panel holds on both versions. On the post-update weights, as on the pre-update ones, prose memory leaves the bias above the memory-free level, both quantitative rungs lower it, and the self-calibrating rung is resolvably better than no memory. Format compliance also improved, from five unparseable replies to none. This is the closest thing we have to a test of whether the panel's findings are an artefact of one week's snapshot of one endpoint, and on the one model we can test it, they are not.

The obvious limitation is that this is one update to one model on one task, measured once, and our own between-run spread at temperature 0 is about 2.6 minutes — so the L2 and L3 bias changes of about one minute are within noise, while the memory-free change of twelve minutes and the token and latency ratios are not. We report the comparison because the protocol makes it possible at all, and we would encourage anyone maintaining a frozen prompt set to record the serving fingerprint alongside it: without that string, this section would have been an unexplained inconsistency between two of our own tables.

S Figures

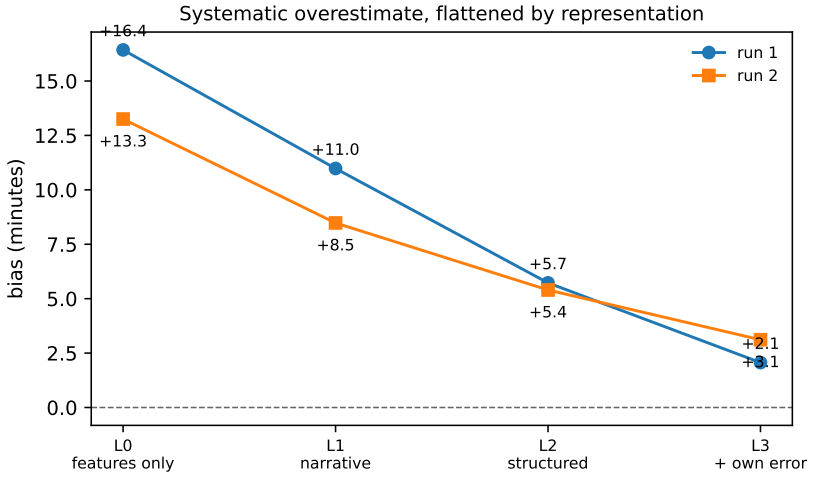

Figure 2: Systematic bias of the duration estimate across the representation ladder, for two repeated runs on the same anchors. One archive of nightly observations is held as generated prose (L1), then as a stratified statistical snapshot alongside it (L2), then with a note on the configuration's own past error (L3). The archive is fixed; the rungs differ in what is derived from it and how that is written. 170 scorable anchors on 22 observing days, sampling temperature 0. Point estimates; the paired differences and their day-clustered intervals are in Table 1.

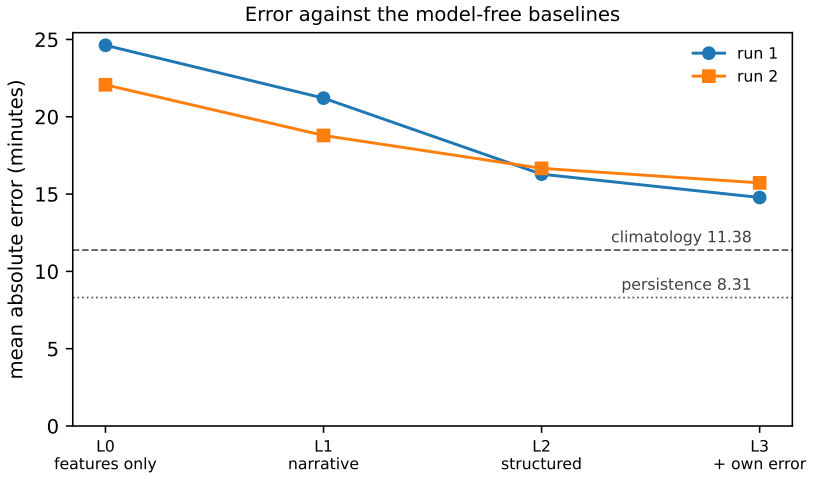

Figure 3: Mean absolute error across the representation ladder for two repeated runs, against the model-free baselines climatology (11.38), persistence (8.31). The ladder improves monotonically in bias and closes most of the gap, but does not cross it. The paired improvement of full_memory over the memory-free line is not resolvable in either run, so the first step is not an established effect. 170 scorable anchors on 22 observing days, sampling temperature 0. Point estimates; the paired differences and their day-clustered intervals are in Table 1.

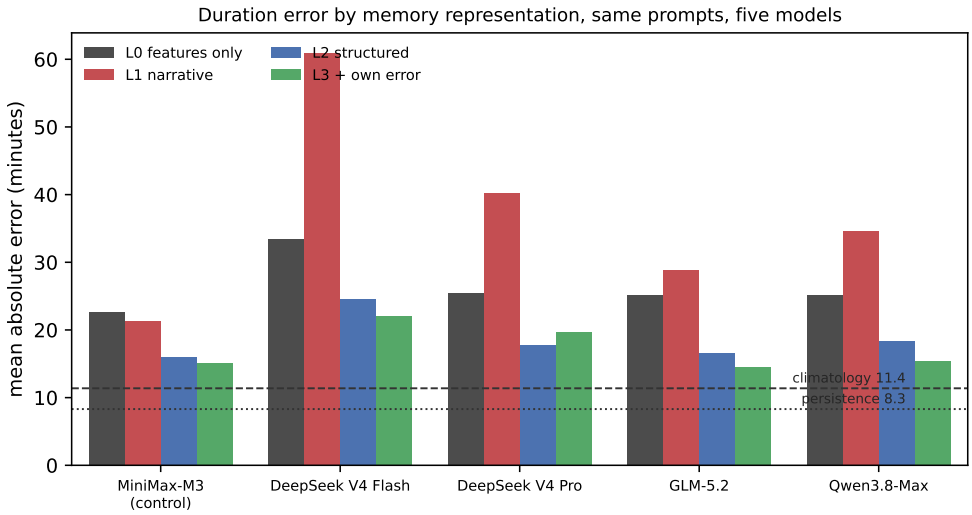

Figure 4: Mean absolute error of the remaining-occlusion estimate, by memory representation, for the same 170 recorded prompts replayed against 5 models at temperature 0. Holding the nightly observations as generated prose is worse than holding no memory at all on 4 of 5 models, and is the worst bar in its own group on 4 of 5. The 1 exception is the control, the model that wrote the prose. Supplying stratified statistics derived from the same archive beats the generated prose on 5 of 5, which is the comparison the figure is for. Horizontal rules are the model-free baselines (climatology 11.38, persistence 8.31), which no configuration reaches. Absolute levels differ between models and are not the comparison being made; the ordering within each group is.

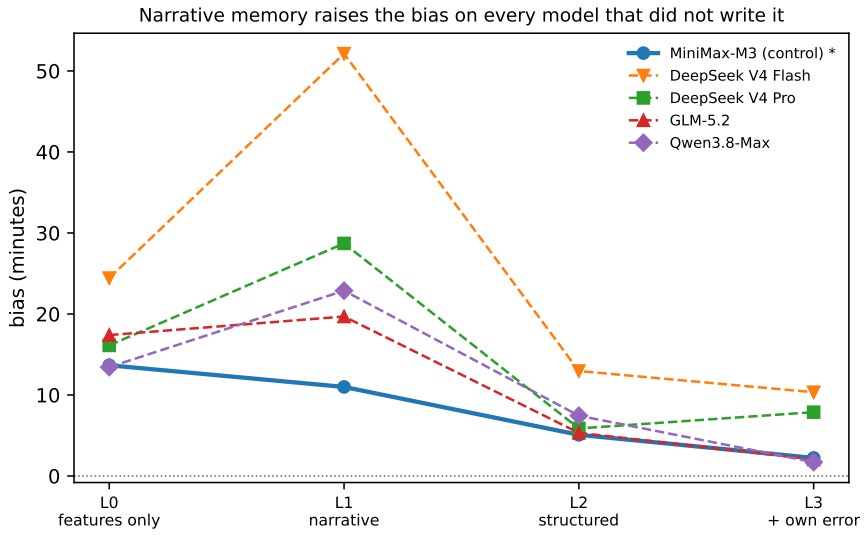

Figure 5: Systematic bias against memory representation for the same recorded prompts replayed against 5 models from 4 laboratories at temperature 0. The starred line is the model that generated the stored memory; it is the only one the narrative rung does not harm. Holding the same nightly observations as generated prose raises the bias above the memory-free level on 4 of 5 models, while holding them as stratified statistics lowers it on 5 of 5. A monotone fall was pre-specified and is not what the data show; the departure is reported rather than reinterpreted.

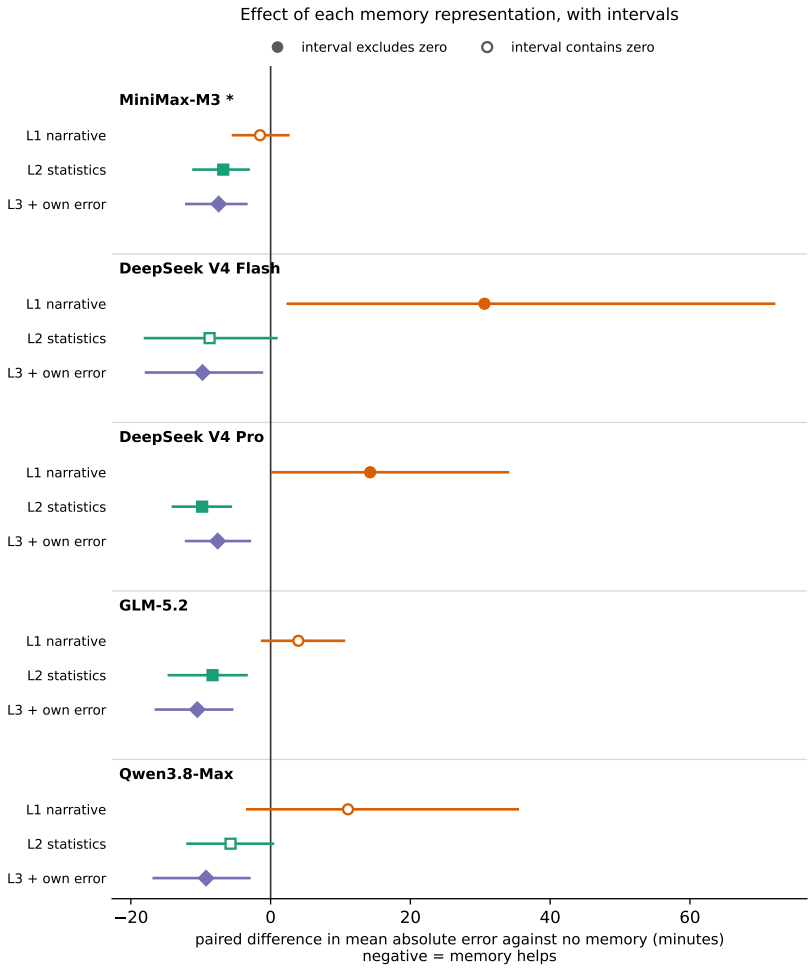

Figure 6: Paired difference in mean absolute error against the memory-free configuration, for each memory representation on each of 5 models replaying the same recorded prompts at temperature 0. Negative is better. Bars are 95% percentile bootstrap intervals resampling the observing days each comparison covers, 20 to 22 of the 22 in the archive, since a day on which either arm answered nothing contributes no paired anchor; filled markers exclude zero. The starred model wrote the stored memory and is a control on the replay method rather than cross-model evidence. Reading down a rung: the narrative form is worse than no memory on 4 of 5 models, resolvably on 2; the statistical form helps on 5 of 5, resolvably on 3; adding the note on the configuration's own past error is resolvable on 5 of 5. The two structured rungs were the pre-specified primary comparisons and the narrative rung is reported as the departure from the criterion; the three rungs of one model share its memory-free arm, so they are not independent tests and no multiplicity correction is applied or implied.