Beyond Retrieval

Source, Supersession, and Binding in Long-Horizon Agent Memory

The LBrain Architecture and Early Evidence

Tad MacPherson

Founder, Metavolve Labs, Inc.

Technical White Paper · August 22, 2026

Status: Technical white paper. Not peer-reviewed and not under review.

Software: LBrain 0.1.6, open source under BSD-3-Clause.

Evidence horizon: Public literature and product documentation reviewed through 22 August 2026.

Contents

0.1 Abstract . . . 3

1 The Record Problem **3**

2 Three Claims, Kept Separate **4**

3 The Result That Motivated LBrain **5**

4 LBrain's Architectural Answer **6**
4.1 Hybrid retrieval . . . 6
4.2 Binding-aware serving . . . 6
4.3 Supersession without erasure . . . 7
4.4 Honest dating . . . 7

5 Retrieval Quality Is Not Enough: Binding and Abstention **7**

6 Temporal Validity: Relevant Does Not Mean Current **8**

7 From Memory to Accumulated Judgment **9**

8 Memory Can Preserve Error Too **9**

9 Toward Belief Provenance **10**

10 Context Economics: A Secondary, Falsifiable Thesis **11**
10.1 Why context growth matters . . . 11
10.2 Archive analysis . . . 11
10.3 Controlled boundary replicate . . . 12
10.4 The failures matter more than the ranking . . . 12
10.5 The crossover hypothesis . . . 12
10.6 What the authors do in practice . . . 13

11 The Next Decisive Experiment **14**

12 Correction Persistence Should Be Tested Separately **15**

13 Related Work **15**

14 Current Limits **16**
14.0.1 The cross-architecture result is not a universal model comparison . . . 16
14.0.2 Earlier confabulation interpretations were retracted . . . 16
14.0.3 Binding is conservative and imperfect . . . 16
14.0.4 Supersession is only as complete as the record . . . 16
14.0.5 Context economics remain unresolved . . . 17
14.0.6 Correction persistence has not been established here . . . 17
14.0.7 Belief provenance is incomplete . . . 17
14.0.8 Security maturity remains beta . . . 17

15 Open Replication **17**

16 Conclusion **18**

17 References **18**

0.1 Abstract

Long-horizon AI is usually described as having a memory problem. That description is incomplete.

Retrieval can succeed technically and still give a model the wrong record from which to answer. A retrieved passage may be authentic, topically relevant, semantically similar, and even contain a true historical value, while still being stale, attributable to the wrong entity, superseded by a later decision, or incapable of answering the question actually asked.

Finding the nearest text and finding an answer are not the same operation.

LBrain is an open-source, local-first memory layer built around that distinction. It combines hybrid lexical and vector retrieval with explicit source and date metadata, non-destructive supersession, deterministic binding checks, evidence annotations, bounded source-grounded rendering, and conservative abstention signals. Its purpose is not to make a language model remember everything. Its purpose is to improve the record from which a model decides what deserves to govern an answer.

This paper separates LBrain's thesis into three claims. First, the **epistemic claim**: retrieval relevance is insufficient because relevant records need not be answer-bearing or governing records. Second, the **temporal and provenance claim**: persistent agent memory requires explicit structure for source, time, supersession, and evidence status. Third, the **economic claim**: an external memory layer may eventually repay its fixed retrieval overhead by slowing active-context growth, but only beyond a workload-specific crossover depth.

The evidence for these claims is intentionally uneven. Metavolve Labs' current public eight-architecture record-structure matrix reports aggregate failure rates spanning **1.3%, 35.7%, and 16.7% across three served-record conditions**, a 34.4 percentage-point range, with the same ordering in all eight tested architectures and approximately zero variance attributed to architecture. [MEASURED] Separately, an archive analysis found approximately **19% lower context growth per turn** among comparable LBrain sessions, while a controlled calibration experiment found no economic crossover within nine turns and exposed five retrieval-miss cascades. [MEASURED]

The latter failures matter. They show why memory infrastructure should be evaluated not as a feature but as a system with its own measurable failure modes.

The central proposition of this paper is therefore deliberately narrow:

> **Long-horizon AI needs more than access to prior text. It needs memory infrastructure that helps distinguish relevant evidence from governing evidence, preserves the ancestry of changing records, and can say when the available record does not answer the question.**

1 The Record Problem

Consider two perfectly valid notes:

“‘text March: Deploys use –safe-mode. “‘

“‘text August: –safe-mode is retired. Deploys now use –verify-first. Supersedes [[deploy-flag]]. “‘

A semantic retriever has done nothing wrong if it returns both.

A language model has done nothing obviously irrational if it treats both as evidence.

The missing information is not semantic relevance. It is **record status**.

Which instruction governs now? Which is historical? Which record replaced which? Does the passage actually bind the requested entity and attribute? What evidence established the current rule? Is the record direct observation, sourced reporting, synthesis, or something whose credibility has never been graded?

These questions produce a useful hierarchy:

“‘text Relevant ≠ current Relevant ≠ attributable Relevant ≠ answer-bearing Retrieved ≠ admissible Faithful ≠ true “‘

The last distinction is especially important. A model may faithfully summarize the text it was given and still produce a wrong answer because the text belonged to the wrong entity or represented a superseded state.

Independent work now demonstrates versions of this problem. LongMemEval treats knowledge updates and abstention as core long-term-memory capabilities and reports substantial degradation under sustained interaction. [6] Deceptive Grounding shows that RAG responses can remain well-cited and document-faithful while attributing evidence to the wrong entity, with failures ranging from 8% to 87% under peak adversarial conditions across thirteen models. [10] Research on temporal validity likewise finds that similarity alone performs poorly when stale and current values remain semantically close. [11,12]

The problem is therefore not simply:

> Can the agent retrieve something relevant?

It is:

> **Can the system determine what the retrieved record means for the present question?**

That is the problem LBrain is designed around.

2 Three Claims, Kept Separate

LBrain’s thesis becomes clearer when three different claims are not allowed to borrow evidence from one another.

Claim	Proposition	Current status
1. Epistemic	Retrieval relevance is insufficient. Record structure and binding materially affect whether a model can use retrieved evidence correctly.	[**MEASURED**]
2. Temporal / provenance	Long-lived memory benefits from explicit source, date, supersession, credibility, and answer-binding structure.	[**IMPLEMENTED + MEASURED limitations**]
3. Economic	External memory may repay fixed retrieval overhead by reducing active-context growth in sufficiently long workloads.	[**HYPOTHESIS**]

This separation is consequential.

The third claim is currently the least established. LBrain has **not** demonstrated a cost crossover at the tested depth.

That does not invalidate the first two claims.

If an external memory system never saved a single API dollar but materially reduced wrong-record answers, preserved corrections across sessions, exposed provenance, and improved appropriate abstention, it could still be valuable infrastructure.

Context efficiency is worth measuring. It should not be asked to carry the entire argument.

3 The Result That Motivated LBrain

A simple observation drove the project:

> **We put the answer in the retrieved record. The models could still fail to use it correctly.**

The important variable was not merely whether information was present. It was how that information was structured and served.

Metavolve Labs' current frozen public summary reports an experiment spanning **eight model architectures from seven organizations** on the same near-domain retrieval task. Across three record-serving conditions, aggregate failure rates were **1.3%, 35.7%, and 16.7%**, a **34.4 percentage-point absolute spread**. The ordering of those conditions was identical across **8 of 8 architectures**, while architecture itself explained approximately **0% of the observed variance** in the reported analysis. [MEASURED]

This finding does not mean models are interchangeable. It does not establish that every retrieval failure is caused by record structure, nor that better models provide no advantage elsewhere.

It establishes something narrower and more actionable:

> **For the tested failure class, changing the model did not remove the effect. Changing the served record changed the effect.**

That points engineering attention toward the input side of the generation boundary.

It also requires an important correction to an earlier interpretation. During development, Metavolve Labs initially attributed part of an observed failure to spontaneous retrieval confabulation. Later calibration showed that a prompt nudge had contaminated that interpretation. The current binding-aware-serving design explicitly retracts that claim and does **not** claim that near-match records inherently cause spontaneous confabulation. Its supported motivation is serve quality, extraction utility, ambiguity disclosure, and the ability to distinguish an answer-bearing record from a neighboring one. [MEASURED / RETRACTED CLAIM]

That distinction belongs in this paper because the mechanism is more interesting after the retraction, not less.

The problem does not require a model to hallucinate.

The model can be wrong while reading entirely real records.

4 LBrain's Architectural Answer

LBrain is a local-first memory engine in which the human-readable source remains authoritative and the searchable index is derivative.

The public 0.1.6 package is distributed under BSD-3-Clause and uses SQLite, FTS5, sqlite-vec, and hybrid lexical/vector retrieval. Its default embedding path runs locally, requiring no hosted API for ordinary indexing and retrieval. [IMPLEMENTED]

Its conceptual read path is:

"'text Authoritative source records ↓ Hybrid lexical + vector retrieval ↓ Temporal / supersession state ↓ Deterministic binding check ↓ binds / near-miss / superseded ↓ Source + date + evidence annotations ↓ Bounded source-grounded rendering ↓ Agent context "'

The terminology matters.

LBrain does **not** presently treat the binding check as an infallible blocker. Its implemented serving design annotates records non-destructively. A result may be marked `binds` or `near-miss`; records are not silently withheld or reordered simply because the check dislikes them. [IMPLEMENTED]

This makes the mechanism inspectable.

The model sees the evidence. The operator can see the same evidence. The deterministic layer states what it believes about that evidence without pretending to become another oracle.

4.1 Hybrid retrieval

Lexical and semantic retrieval fail differently.

Literal search is strong for identifiers, error messages, filenames, version strings, and exact values. Vector retrieval helps when a query and a source describe the same concept using different vocabulary. LBrain combines both rather than treating semantic similarity as sufficient.

The output of retrieval, however, remains a set of **candidates**.

Candidate generation and evidence qualification are deliberately separate operations.

4.2 Binding-aware serving

For question-shaped queries, LBrain can run a deterministic check against the exact excerpt being served and annotate it as answer-bearing or near-match. [IMPLEMENTED]

This leads to a fundamental distinction:

> **Retrieval asks whether a record is about the question. Binding asks whether the record contains evidence that answers it.**

A record can score highly under the first test and fail the second.

That failure is not exotic. A note about the same project, customer, endpoint, dosage, benchmark, version, or deployment procedure can be semantically excellent and operationally wrong.

4.3 Supersession without erasure

A replaced record can remain historically valid while no longer being operationally current.

LBrain therefore separates **persistence** from **activation**. Superseded information is not deleted merely because a newer record governs current behavior. [IMPLEMENTED]

That provides a primitive for auditability:

“‘text what governs now ↓ what governed before ↓ what changed it “‘

A memory system that erases the first version loses history.

A memory system that serves both versions without status loses authority.

The useful system must preserve one while distinguishing the other.

4.4 Honest dating

Time metadata has different evidentiary strength.

A claim date embedded in content is not equivalent to filesystem modification time. Copying a directory, cloning a repository, or synchronizing an archive can make old records appear newly modified.

LBrain therefore distinguishes stronger date evidence from weaker file dating rather than silently treating every timestamp as claim time. [IMPLEMENTED]

That small distinction becomes important the moment a query contains words such as *current*, *latest*, *still*, *now*, or *when*.

5 Retrieval Quality Is Not Enough: Binding and Abstention

Traditional retrieval evaluation asks questions such as Recall@k or Mean Reciprocal Rank.

Those metrics remain useful, but they do not measure the entire system LBrain is trying to build.

A retriever may return three topically excellent records when none contains an answer. Calling that a successful memory operation creates pressure on the generator to manufacture certainty from proximity.

LBrain therefore motivates a second measurement layer.

Metric	Question
Retrieval recall	Did the system retrieve the record containing the answer?
Binding precision	When the system labels a record answer-bearing, does it actually answer the question?
Binding recall	When an answer-bearing record is present, does the system recognize it?
False-binding rate	How often is neighboring evidence incorrectly treated as governing evidence?
Correct abstention	When the available record contains no answer, does the system say so?
False abstention	How often does it reject evidence that really does answer the question?

[PROPOSED EVALUATION FRAMEWORK]

LongMemEval independently supports the importance of this distinction by treating abstention as one of five core long-term-memory abilities, alongside extraction, multi-session reasoning, temporal reasoning, and knowledge updates.

The engineering objective is not maximum refusal.

It is **calibrated refusal**.

An agent that never abstains is dangerous.

An agent that abstains whenever records disagree is useless.

A useful memory layer should make that trade measurable.

6 Temporal Validity: Relevant Does Not Mean Current

The March/August deployment example exposes one of the central weaknesses of ordinary similarity retrieval.

Both records are relevant because both are genuinely about the deploy flag.

The stale record does not become semantically dissimilar when it becomes obsolete.

External results reinforce this point. MemStrata reports that cosine similarity was nearly uninformative for distinguishing contradictions from duplicates in its calibrated setting, with AUROC 0.59, and reports ordinary RAG serving superseded values 15% to 40% of the time in its evolving-knowledge tests. Separate 2026 work found substantial gains on explicit fact-conflict tasks when current-value resolution was moved from model judgment into deterministic post-retrieval assembly.

LBrain has also failed here.

In one internal probe, its own index served a retired figure **2:1 over the current figure** because the necessary supersession marker existed for a neighboring claim but not for the one being queried. [MEASURED]

On another probe, where the retirement marker was present, LBrain returned the current value and flagged the stale record. [MEASURED]

That is not evidence that supersession "solves" stale memory.

It establishes its boundary:

> **The mechanism works exactly as far as the record structure reaches.**

This converts a fuzzy model-behavior problem into an engineering problem.

Supersession coverage can be measured.

Missing markers can be detected.

Conflicting active values can become a regression test.

The failure can enter CI instead of remaining a story about the model being confused.

7 From Memory to Accumulated Judgment

Persistent memory becomes more consequential when it stores not only facts, but **corrections to prior behavior**.

Suppose an agent acts on an outdated deployment rule. A human corrects it. The current conversation acknowledges the correction.

The interesting question is not whether the agent said "understood."

The interesting question is whether a later session, a later context window, or a different model instance can recover and apply that correction before repeating the mistake.

This suggests a distinction between conversational memory and **persistent judgment**.

LBrain's current interface already includes a pre-action mechanism intended to check consequential actions against stored corrections and preferences. [IMPLEMENTED]

The broader behavior should now be measured explicitly.

Two proposed metrics are:

Error carry-forward rate

\{}[E_{cf} = \{}frac{\{}text{previously corrected errors repeated later}} {\{}text{opportunities to repeat them}} \{}]

Recovery-after-correction rate

\{}[R_{ac} = \{}frac{\{}text{later opportunities in which the correction is correctly applied}} {\{}text{later opportunities requiring the correction}} \{}]

[**PROPOSED**]

These metrics intentionally avoid anthropomorphic claims about persistent identity.

The scientific question is simpler:

> **Can judgment survive the session that produced it?**

If yes, agent memory begins to accumulate operational precedent rather than merely accumulating transcript.

8 Memory Can Preserve Error Too

Persistence is not intrinsically beneficial.

A memory system can preserve a correction.

It can also preserve a confident mistake.

Once persistent records influence future generations, repeated internal reuse creates a new failure mode: an unsupported assertion may acquire apparent authority because it has been copied, summarized, retrieved, or cited many times.

The system can become very efficient at remembering the wrong thing.

For that reason, persistence must be paired with evidence structure.

LBrain's newer two-axis evidence-grading design begins separating two questions:

1. **How well is this claim known?**
2. **How reliable is the source that authored it?**

The credibility axis is implemented. Records can distinguish direct observation, sourced information, synthesis, and ungraded material. The source-reliability axis is explicitly not treated as solved because it depends on verified authorship and identity binding that are not yet complete. [IMPLEMENTED / PARTIAL]

This is an important discipline.

A citation should not automatically upgrade the credibility of an assertion.

A repeated assertion should not become independent corroboration.

And multiple models operating over one common record should not be mistaken for multiple independent sources.

A useful research principle follows:

> **Three model votes over one shared record are not necessarily stronger evidence than one model contradicted by an external test.**

[HYPOTHESIS / DESIGN PRINCIPLE]

The improving system must not control the entire definition of improvement.

Memory can help an agent accumulate better judgment only if the system can distinguish internal repetition from independent evidence.

9 Toward Belief Provenance

Document provenance answers:

> Which artifact supplied this statement?

A deeper form of provenance asks:

> **Why does the system currently believe this?**

Those questions are related, but they are not identical.

For consequential operational beliefs, an eventual lineage may resemble:

“‘text Current decision ↓ Superseded decision ↓ Correction or incident ↓ Evidence that caused the correction ↓ Original source “‘

Or:

“‘text Model recommendation + External test + Human adjudication ↓ Revised doctrine “‘

This is the beginning of **belief provenance**, rather than document provenance.

LBrain does not claim that complete architecture today.

The current system implements several primitives required for it: source-bearing records, claim dates, non-destructive supersession, binding annotations, and an implemented credibility grade. Verified source reliability remains dependent on stronger identity binding. A separate authority-verification design in the repository is explicitly marked design-only and documents why a hash alone proves integrity but not authorship. [IMPLEMENTED + PROPOSED]

This distinction should remain sharp.

The long-term direction is not merely a database of what an agent has seen.

It is a recoverable ancestry of how an operational belief came to govern.

10 Context Economics: A Secondary, Falsifiable Thesis

External memory has another possible benefit: active-context efficiency.

This is the most speculative of LBrain's three principal claims.

It is also straightforward to falsify.

10.1 Why context growth matters

Prompt caching substantially changes the economics of long agent sessions. A 2026 study spanning more than 500 agent sessions across OpenAI, Anthropic, and Google reported API cost reductions of **45% to 80%** and time-to-first-token improvements of **13% to 31%**, while also finding that naive full-context caching can increase latency.

Anthropic currently prices a five-minute prompt-cache write at **1.25**× the normal input rate and a cache hit at **0.1**×, making a write token 12.5 times as expensive as a cache-hit token under that schedule. A separate cache-aware compression study reports a two-tier cache effect on Sonnet 4.6 and demonstrates that compression and caching can interact nonlinearly enough to produce negative ROI in some strategies and substantial savings in others.

The relevant question for LBrain is therefore not whether fewer tokens are aesthetically preferable.

It is whether the fixed cost of retrieving external memory can be repaid by slower growth of expensive active context.

10.2 Archive analysis

Across an archive of **344 sessions**, restricting analysis to working sessions of at least 20 turns produced **31 comparable sessions**, 21 with LBrain and 10 without.

Sessions using LBrain showed approximately **19% lower context growth per turn**, with a median ratio of **0.79**× and aggregate ratio of **0.81**×. [MEASURED]

Compaction incidence was **14% with LBrain versus 40% without**, but only ten compaction events occurred, making that result too weak for a strong inference. [MEASURED, LOW N]

Cache-read share was statistically indistinguishable between the archival groups, providing no evidence in that sample that retrieval injection was materially destroying prompt-cache prefix reuse. [MEASURED]

This was observational data, not a randomized experiment.

10.3 Controlled boundary replicate

A separate calibration experiment branched three arms from the **same session at the same context boundary**:

Arm	Continuation
C	Continue from native compaction summary
L	Begin fresh with LBrain over the agent's own notes
B	Begin fresh with files on disk but no LBrain

Eight probes were run per arm.

Result	C	L	B
Probe passes	7/8	6/8	5/8

On the pre-registered primary L-versus-C comparison, there were four matched pairs and **zero discordant outcomes**. No quality separation can be inferred.

Graders were blinded. Their arm guesses were **32.6% correct**, compared with **33.3% chance**.

The replicate cost **$157.48** in metered API spend.

[MEASURED: one workload, one seed, eight probes per arm. Calibration only.]

10.4 The failures matter more than the ranking

Two pre-registered defeat detectors fired.

Short-session overhead: FIRED. No cost crossover occurred within nine turns. At the tested depth, LBrain cost more than the bare-floor continuation.

Retrieval-miss cascade: FIRED. Five events occurred, all in the LBrain arm, in which an indexed retrieval failed and the agent then read source files directly. The resulting context carried both the unsuccessful lookup and the fallback.

The always-on core block added approximately **173 tokens per turn**, a median **0.9%** of per-turn output.

One harness criterion also failed because a run exceeded its registered budget hold.

These are useful results.

They tell us external memory is not a free optimization.

10.5 The crossover hypothesis

The internal archive result and controlled replicate initially appear contradictory:

- lower context growth per turn;
- higher cost at the observed short depth.

They are compatible under a simple mechanism.

Let \{}(F\{}) be fixed per-turn memory overhead and \{}(g_C - g_L\{}) be the reduction in context-growth slope produced by external memory.

A conceptual cumulative difference is:

\{}[\{}Delta C(T) \{}approx F T - P_{\{}mathrm{effective}} \{}frac{(g_C-g_L)T(T+1)}{2} + C_{\{}mathrm{miss}} + C_{\{}mathrm{boundary}} \{}]

This is **not** a fitted model.

It is a mechanistic hypothesis.

The fixed term grows approximately linearly. Avoided accumulated context can grow faster with session depth. Retrieval misses push in the wrong direction. Cache pricing and compaction behavior modify the economic value of each avoided token.

The resulting prediction is simple:

> **An external memory layer should lose on sufficiently short sessions. If it has an economic advantage, that advantage should appear only beyond a workload-specific crossover depth.**

[HYPOTHESIS]

The controlled experiment observed the predicted early loss.

It has not observed the crossover.

Therefore:

> **No cost saving is claimed in this paper.**

10.6 What the authors do in practice

Before adopting LBrain, the authors' own long-horizon sessions typically passed through **three to five compaction events** each. Since adopting it, comparable work routinely runs for days without reaching a compaction boundary. [OBSERVED]

Two qualifications belong with that statement. It is the authors' operational experience rather than a controlled measurement, and it comes from a single team working on one codebase.

The second qualification is more useful, because the practice changed alongside the tooling. The authors found that **compaction itself produced drift**: after a boundary, the successor would re-open settled questions and lose constraints established earlier in the work. They now treat compaction as something to be avoided rather than relied upon, and retire sessions deliberately instead. **The mechanism by which continuity is carried across that retirement is outside the scope of this paper.** [OBSERVED]

That practice resolves an otherwise awkward archive result. In the 344-session archive, LBrain sessions were **shorter** in wall-clock than non-LBrain sessions — a 30.9-hour median against 40.3 hours — which reads as disconfirming until the practice is known. Sessions are shorter because they are deliberately retired rather than compacted. The archive's compaction incidence, 14% with LBrain against 40% without, is consistent with the same practice, though it rests on only ten events. [MEASURED, LOW N]

The claim this supports is narrower than "longer sessions," and more interesting. It is that **continuity stops depending on session length**. When the record lives outside the transcript, a session becomes a disposable unit of work rather than the vessel that must hold everything — and the question shifts from how long a context window can be stretched to how faithfully a successor can be reconstituted. That reframing, not the duration, is what the authors report having changed in day-to-day practice. [OBSERVED]

11 The Next Decisive Experiment

The record-structure evidence now suggests a more important experiment than another memory-versus-no-memory comparison.

The next study should isolate **which structural intervention recovers which portion of the failure**.

Use the same underlying facts, model, and questions across four arms:

Arm	Served record
A	Clean current record
B	Raw historical record containing stale and overlapping evidence
C	LBrain-curated record
D	LBrain-curated record with provenance and supersession signals removed

[**PROPOSED**]

Then rotate the underlying values so that no arm benefits from lexical artifacts or a memorized answer.

Repeat across model families.

Measure:

- answer accuracy;
- answer-bearing retrieval recall;
- binding precision and recall;
- stale-value error;
- wrong-entity attribution;
- correct and false abstention;
- served tokens;
- context growth;
- latency;
- cumulative task cost.

This design allows the experiment to separate several competing explanations.

If C beats B, curated memory helps.

If D matches C, provenance and supersession metadata may be unnecessary in that workload.

If C beats D, structure itself contributes beyond token reduction.

If A remains much stronger than C, the memory layer is still failing to reconstruct the clean-record ideal.

The result becomes more useful than:

> Messy records hurt models.

It becomes:

> **These specific structural interventions recover this fraction of the loss.**

That is a system result.

12 Correction Persistence Should Be Tested Separately

A second benchmark should study **durable correction**.

A model receives a record, acts incorrectly, and receives an explicit correction.

The originating session ends.

A new session, model instance, or model architecture later encounters a condition requiring the same judgment.

Experimental conditions should vary:

“‘text Correction absent from persistent memory

Correction stored as ordinary text

Correction stored with source and date

Correction stored with supersession relationship

Correction stored with evidence grade

Correction contradicted by later external evidence “‘

The key outcomes are not conversational acknowledgment.

They are:

- whether the error recurs;
- whether the correction is recovered;
- whether newer contradictory evidence replaces it correctly;
- whether the system can explain why the present rule governs;
- whether the result transfers across models.

[PROPOSED]

This would test persistent judgment directly.

13 Related Work

The broader literature is converging on a similar premise: long-term memory is becoming a systems layer with its own representation, update, retrieval, temporal, and assembly problems.

LoCoMo demonstrated persistent difficulty with long-range temporal and causal dialogue even when using long-context or RAG strategies. LongMemEval formalized extraction, multi-session reasoning, temporal reasoning, updates, and abstention as separable capabilities.

Recent structured-memory systems reinforce the value of selection and temporal organization. APEX-MEM combines append-only history with temporally grounded structured retrieval and reports strong LoCoMo and LongMemEval performance. TiMem similarly treats temporal continuity as a first-class organizing principle and reports improved benchmark accuracy while reducing recalled memory length in its LoCoMo evaluation.

LBrain is not presented as a replacement for these architectures.

Its narrower contribution is the **serve boundary**:

> after retrieval has found candidate memory, what structure should accompany a record before that record is allowed to influence generation?

LBrain emphasizes original-source recovery, explicit dating, supersession, binding status, evidence annotation, conservative abstention, and local inspectability.

The research question is not whether structured memory matters.

Increasingly, the question is **which structure matters, under which failure condition, and how much of the error it recovers.**

14 Current Limits

The following limitations define the boundary of the claims in this paper.

14.0.1 The cross-architecture result is not a universal model comparison

The eight-architecture matrix addresses a particular record-structure failure class. It should not be generalized to overall model capability. The governing artifact is the internal `B1-REFIT-RESULTS` record, which states that the cross-architecture convergence is **not evidence of an architecture-general constant**: the intraclass correlation attributable to model is approximately zero (ICC ≈ 0.003), stimulus accounts for the large majority of variance, and every jackknife confidence interval includes zero. The per-condition rates are reported as measured; the inference that the effect is architecture-general was examined and retired.

14.0.2 Earlier confabulation interpretations were retracted

An earlier prompt-nudged result was found to overstate spontaneous confabulation. The current system does not claim that binding-aware serving prevents general hallucination or that near-match retrieval inherently induces confabulation.

14.0.3 Binding is conservative and imperfect

A deterministic binding check can produce false abstention or false binding. Those rates need independent measurement.

14.0.4 Supersession is only as complete as the record

The internal stale-value failure demonstrates that missing supersession structure can allow obsolete values to survive.

14.0.5 Context economics remain unresolved

The archive association is not causal. The controlled replicate is small. No crossover was observed within nine turns. No dollar savings are claimed.

14.0.6 Correction persistence has not been established here

LBrain can store and query corrections and exposes a pre-action check, but this paper does not report a controlled cross-session recovery-after-correction benchmark.

14.0.7 Belief provenance is incomplete

Credibility grading is implemented. Verified source reliability remains dependent on identity binding. Full ancestry from current doctrine through corrections, evidence, and authoritative identity is a research direction, not a completed product claim.

14.0.8 Security maturity remains beta

LBrain 0.1.6 is public beta software. Its documentation states that it has not yet undergone an independent security audit. Retrieved-content fencing and sanitization reduce specific injection surfaces but should not be mistaken for a complete security boundary.

15 Open Replication

The useful form of this work is not a private benchmark win.

It is a set of failure classes other operators can try to break.

For long-running agent workloads, the highest-value external measurements would include:

1. record-structure accuracy under clean, stale, overlapping, and wrong-entity conditions;
2. binding precision, binding recall, and correct abstention;
3. stale-fact and wrong-entity error rates;
4. correction recovery across fresh sessions or different model instances;
5. context growth per turn;
6. turns between compaction boundaries;
7. retrieval-miss cascade rate;
8. cumulative cost per completed task;
9. crossover depth, if one exists.

Runs in which LBrain loses should be published alongside runs in which it wins.

A failure that survives replication is not an inconvenience to this research program.

It is a specification.

16 Conclusion

Long-horizon AI does not merely need more memory.

It needs **memory that can be questioned**.

A useful record should expose where it came from, when it applied, what it refers to, whether it has been superseded, how well its claim is supported, and whether it actually contains evidence for the question being asked.

When the evidence is absent, the system should be capable of saying so.

LBrain is an attempt to move those properties out of conversational convention and into infrastructure.

The present evidence supports some parts of that ambition more strongly than others.

Record structure has produced a large and architecture-consistent effect in Metavolve Labs' cross-model experiments.

Binding, dating, supersession, source-grounded rendering, and evidence grading are now concrete engineering mechanisms rather than abstract aspirations.

Those mechanisms still fail.

A missing supersession marker has served stale information.

A retrieval miss has added context rather than removing it.

A conservative binding check can reject usable evidence.

The economic crossover remains unobserved.

Those failures are precisely why the architecture is worth measuring.

The remaining questions are empirical:

How much error comes from retrieval?

How much comes from the structure of the record after retrieval?

How much can deterministic serving recover?

How reliably can corrections survive across sessions and model architectures?

How much authority should come from source, evidence, time, or independent adjudication?

At what depth, if any, does external memory repay its operational cost?

Those questions are measurable.

That is the standard this work should be held to.

The goal is not to make an AI remember everything.

It is to give it a better record from which to decide what should be remembered, what should be trusted, and when it should say: "I don't know."

17 References

[1] **Metavolve Labs, Inc.** *LBrain.* Public source repository and README, 2026. Current documentation of hybrid retrieval, provenance, binding annotations, supersession, local

embeddings, MCP interfaces, limitations, and the frozen cross-architecture record-structure summary.

[2] Metavolve Labs, Inc. *Design: Binding-Aware Serving*, v2 post-red-team, implemented July 24, 2026. Documents query-centered excerpts, deterministic admissibility/binding analysis, non-destructive annotations, rendering budgets, sanitization, security posture, and experimental corrections.

[3] Metavolve Labs, Inc. *Design: Two-Axis Evidence Grading*, 2026. Documents implemented claim-credibility grading and the planned identity-dependent source-reliability axis.

[4] Metavolve Labs, Inc. *Design: gcx:// Authority Verification*, 2026. Design-only analysis of integrity versus authorship and cryptographic authority binding.

[5] Metavolve Labs, Inc. *LBrain Context-Economics Calibration Archive and C/L/B Boundary Replicate.* Internal measurements, August 2026. Includes the 344-session archive analysis, controlled boundary experiment, defeat detectors, and crossover hypothesis.

[6] Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., & Yu, D. "LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory." *International Conference on Learning Representations*, 2025.

[7] Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. "Evaluating Very Long-Term Conversational Memory of LLM Agents." *Proceedings of ACL 2024*, pp. 13851–13870. DOI: 10.18653/v1/2024.acl-long.747.

[8] Banerjee, P., Moshtaghi, M., Subramanian, S., Misra, A., & Chadha, A. "APEX-MEM: Agentic Semi-Structured Memory with Temporal Reasoning for Long-Term Conversational AI." *Proceedings of ACL 2026.* [identifier not independently confirmed at time of writing]

[9] Li, K., Yu, X., Ni, Z., et al. "TiMem: Temporal-Hierarchical Memory Consolidation for Long-Horizon Conversational Agents." *Findings of ACL 2026*, pp. 21700–21720. DOI: 10.18653/v1/2026.findings-acl.1091.

[10] Caruzzo, C., Yoo, D., & Kim, T. S. "Deceptive Grounding: Entity Attribution Failure in Clinical Retrieval-Augmented Generation." arXiv:2607.09349, 2026.

[11] Reddy, V., & Challaram, S. "Reliable Post-Retrieval Assembly for Agent Memory: Separating Evidence Extraction from Policy Execution." arXiv:2606.01435, 2026.

[12] Yadav, N. "Temporal Validity in Retrieval Memory: Eliminating Stale-Fact Errors for AI Agents over Evolving Knowledge." arXiv:2606.26511, 2026.

[13] Lumer, E., Nizar, F., Jangiti, A., et al. "Don't Break the Cache: An Evaluation of Prompt Caching for Long-Horizon Agentic Tasks." arXiv:2601.06007, 2026.

[14] Song, Y. "Cache-Aware Prompt Compression: A Two-Tier Cost Model for LLM API Caching." arXiv:2607.15516, 2026.

[15] Anthropic. *Claude Platform Pricing: Prompt Caching.* Current vendor documentation accessed August 2026. Five-minute cache writes are priced at 1.25× base input and cache hits at 0.1×.

LBrain 0.1.6 is open source under the BSD-3-Clause license and available through PyPI. It is beta software and has not yet undergone an independent security audit.