Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Geometric Indexing Does Not Improve Versioned Memory Retrieval: Three Pre-Registered Negative Results and a Silent Total-Failure Mode

Maximiliano Rodrigo Speranza

maximiliano.speranza@gmail.com

Independent Researcher https://orcid.org/0009-0005-0413-8554

Research Article

Keywords: conversational memory, retrieval-augmented generation, knowledge updates, embedding geometry, pre-registration, negative results, coreference resolution

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10669947/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Geometric Indexing Does Not Improve Versioned Memory Retrieval: Three Pre-Registered Negative Results and a Silent Total-Failure Mode

Maximiliano Speranza
Independent Researcher, Buenos Aires, Argentina
ORCID: 0009-0005-0413-8554
maximiliano.speranza@gmail.com

August 2026

Abstract

Conversational agents are increasingly expected to remember what a user told them across sessions, including the case where the user *corrects* something said earlier. A natural proposal is to exploit the geometry of the embedding space: rather than overwriting a memory, deposit the revised version *near* its predecessor, so that proximity encodes the relation between versions. We test that proposal directly, in a non-parametric index over a frozen encoder, with every prediction registered and hash-frozen before the corresponding data existed.

The result is a consistent negative with an identified mechanism. With a fixed displacement step, the revision cluster performs an unbounded random walk away from its anchor: cosine to the query falls from 0.811 to 0.324 after four revisions and to -0.139 after six, and coverage collapses to 0.000 from $K \geq 4$. Bounding the displacement repairs the drift exactly as intended — the cosine becomes flat at 0.8177, slope $+0.0000$ per revision — and the mechanism *still* loses, by -0.0864 (95% CI $[-0.1014, -0.0714]$) at $K = 8$. The reason is a constant toll of ≈ 0.036 cosine: the embedding of the revision's own text is already optimally placed, because it contains the entity the query mentions, so any artificial displacement can only move it away.

That last clause is a property of the *generator*, not of corrections. Real conversational corrections are elliptical: "no, it's Beto" does not name the entity. We therefore ran an adversarial third experiment in the regime that should favour geometry, against the honest baseline (coreference hydration, as deployed by production memory systems), sweeping its failure rate τ. Geometry only wins above $\tau^* \approx 0.45$ — a coreference resolver worse than a coin flip — so the negative extends rather than reverses.

The useful positive finding is independent of the mechanism under test: **raw elliptical corrections are unrecoverable, scoring** 0.0000 **recall at** $k = 5$ **across all ten seeds** — not low, never. In a separate check, their top-1 rank is likewise indistinguishable from chance. A memory system that archives conversational turns without resolving coreference loses **100% of corrections**, and loses them silently — the index always returns something, just never the right thing.

Keywords: conversational memory, retrieval-augmented generation, knowledge updates, embedding geometry, pre-registration, negative results, coreference resolution

1 Introduction

The goal that motivates this work can be stated in one sentence: *that a language model never forgets what I told it.* The emphasis is on *told.* Retaining what a model *read* — long documents,

large contexts, a knowledge base — is a crowded field with enormous resources behind it. Retaining what a specific person *said*, across sessions, including the case where that person later *corrects* it, is a structurally different problem: it persists beyond the forward pass, it belongs to one person rather than a corpus, and it has versions. A corpus does not contradict itself; a conversation does, constantly.

The canonical task of that goal is a versioned fact: *"the director of X is Ana"*, later *"no, it's Beto"*, then *"who directs X?"*. This is precisely what the **knowledge updates** category of LongMemEval [1] measures at the system level. Our contribution is not the task but the *instrument*: we isolate the **indexing mechanism** with everything else held fixed, which a full-system benchmark cannot do.

The mechanism under test is what we call **budding** (*gemación*): when a memory is revised, do not overwrite it — deposit the new version in a *nearby* direction, so that proximity itself encodes the correlation between versions. Its lineage is honourable: Kanerva's sparse distributed memory writes to neighbouring addresses and reads the neighbourhood, and the Differentiable Neural Computer solves temporal linking with an $N \times N$ matrix, which is exactly why it did not scale. The contribution of budding, if it worked, would be that **geometry replaces the link matrix** — $O(1)$ per write instead of $O(N^2)$.

It does not work. This paper reports why, in three pre-registered experiments, and reports the one finding that survived and is useful to practitioners.

1.1 Contributions

1. A pre-registered negative for geometric indexing of versioned memory, replicated across **three regimes** — fixed-step displacement, bounded displacement, and elliptical corrections — with a distinct mechanism identified in each (Sections 4, 5, 6).

2. An **adversarial test of our own conclusion**: the third regime was designed specifically to favour the mechanism, by removing the property that explained its failure in the second. It still lost.

3. A quantitative crossover point: geometric anchoring only outperforms coreference hydration when the hydrator fails on $\tau^* \approx 0.45$ of corrections or more.

4. An independent, actionable finding: **raw elliptical corrections are never retrieved** (0.0000 recall at $k = 5$, 10/10 seeds; top-1 at chance in a separate check), so archiving conversational turns without coreference resolution loses every correction, silently.

5. Two methodological results: a four-part **encoder admission gate** that would have saved two lost nights, and a case study in **a mean concealing its distribution** — the third such case in this research programme.

2 Related work

Benchmarks. LongMemEval [1] defines five abilities for long-term conversational memory, of which *knowledge updates* — a fact stated in one session is changed in a later one — is precisely our task. It measures complete systems (LLM + memory + prompt), where the source of an effect cannot be isolated. We deliberately work one level below. We therefore do not claim the task as novel: it is a controlled, synthetic instance of a category that already has a benchmark, and our contribution is the isolation of one component within it.

Conflict resolution in deployed systems. Graphiti/Zep invalidates a superseded fact while retaining the historical record with a validity window. REMem [3] builds a hybrid memory graph with gists and timestamped facts. NeuSymMS [4] curates a neuro-symbolic memory for

persistent agents, and Supersede [5] diagnoses and trains for the specific *memory-update gap*. Mem0 and Letta offer memory as a service. All of them operate in the *application layer*: graphs, metadata, explicit pointers, symbolic invalidation. None places the index *inside* the network.

Deterministic freshness. A recent line of work [2] argues that resolving memory conflicts should not be delegated to the model's judgement but to a deterministic rule. Our own geometric measurements converge on the same design verdict from a different direction — *geometry to cluster, metadata to order* — and our negative results are therefore an independent, quantified confirmation of that position rather than a competitor to it.

Lineage of the mechanism under test. Budding descends from Kanerva's sparse distributed memory [6], which writes to neighbouring addresses and reads the neighbourhood, and whose read operation is formally close to softmax attention [8]. The Differentiable Neural Computer [7] solves temporal linking with an explicit $N \times N$ link matrix, which is precisely the cost that budding aims to avoid by letting geometry carry the relation instead.

Retrieval indices inside the network. Memorizing Transformers [10] attach a kNN memory that is not differentiable; Memory Layers [11] train the index but with parametric values rather than retrieved content; HOLA [12] and Hybrid Associative Memories [13] combine a recurrent state with an exact bounded cache. All of these live *within* a sequence: their memory is born and dies in the forward pass. Persistence across sessions — our setting — is left to the application layer, and the two obstacles to closing that gap are that gradients do not flow through top-k selection and that a frozen index goes stale as the encoder moves [9, 14].

Reported numbers are not comparable. On LongMemEval with GPT-4o, one report gives Zep 63.8% against Mem0 49.0%, while Mem0 reports 94.4% under a different configuration. The discrepancy is methodological, not a difference in capability. This is exactly the situation that motivates measuring a single mechanism on a controlled bench.

3 Method

3.1 Task

$N = 3000$ synthetic entities, each with one attribute drawn from five templates (*director*, *headquarters*, *founder*, *chief engineer*, *main supplier*), a value drawn from that attribute's pool, and a natural-language query (*"Who directs Helios Laboratory?"*). A revision assigns a new value to the same entity. The index stores directions; a query retrieves the top-k ($k = 5$) by cosine similarity.

Three metrics, all computed per seed and aggregated with Student's t 95% confidence intervals on 9 degrees of freedom (10 seeds $\times$ 1000 sub-sampled entities):

- **CURRENT** — the entry holding the latest revision is in the top-k.
- **PREVIOUS** — the entry holding the preceding revision is in the top-k.
- **COVERAGE** — both are, simultaneously.

3.2 Conditions

- `none` — no archive. Floor control.
- `overwrite` — the revision replaces the original. The old version is lost by construction; its PREVIOUS score must be 0, which is our task-validity control.

- `duplicates` — both versions stored at the true embedding of their own text. This is the baseline budding must beat.
- `budding` — the revision's *content* stored at a direction *anchored* to the original: $\hat{v} = \text{normalize}(E_0 + \varepsilon\, t)$, with t a random unit tangent to E_0 and $\varepsilon = 0.30$ inherited from earlier geometric measurements and never re-tuned.

3.3 Pre-registration protocol

Every prediction, threshold, and falsification clause in this paper was written and **frozen with a SHA-256 hash before the corresponding data existed**. Hashes are recorded in `PREREG_HASH.txt` with timestamps in the repository's commit history. Three properties of the protocol are worth stating explicitly, because they are what make the negatives interpretable:

1. **Effective margin fixed in advance.** A difference counts only if it exceeds 0.02 absolute *and* its CI does not cross zero.
2. **Blocking controls run first.** Each experiment declares a control whose failure voids the rest. They are evaluated before the primary prediction, and the analysis script aborts if one fails.
3. **Falsification committed in advance.** Each pre-registration states what conclusion follows if the primary prediction fails, *including* the commitment to close the line rather than try another variant.

3.4 Encoder admission gate

Before any data is generated, the encoder must pass four checks. This gate exists because its absence cost two nights of work on a misdiagnosed failure, and one line of it would have caught the problem immediately:

1. **C1 — unique-vector census.** How many distinct vectors does the corpus produce?
2. **C2 — value discrimination.** Do two versions of the same fact differ?
3. **C3 — entity discrimination.** Do two different entities differ?
4. **C4 — identification by top-1 and rank, never by AUC.**

C4 is not pedantry. In one failed attempt an AUC of 0.97 coexisted with a top-1 of 0.13: the score distribution separated well on average while the correct item was almost never ranked first. AUC is the wrong instrument for a retrieval gate.

A concrete encoder failure. `nomic-embed-text` served through Ollama collapses *every capitalised token to a single vector*. Measured cosine between unrelated proper nouns:

pair	capitalised	lower-cased
Helios / Vantor	**1.000000** (bit-identical)	0.407
Ana / Beto	**1.000000**	0.349
cat / dog (control)	0.603	0.603

This is not an endpoint artefact (reproduced on `/api/embeddings` and `/api/embed`, singly and batched) nor machine-specific. It explains an earlier corpus that yielded only 75 unique vectors out of 3000 — exactly 5 attributes × 15 numeric suffixes, the only non-capitalised things that varied. **The fix is to lower-case the text before encoding**, after which the gate opens: 400/400 unique vectors, top-1 0.975, median rank 0. All experiments below use lower-cased text.

4 Experiment 1: fixed-step budding

4.1 Single revision

With one revision per entity, all archive conditions sit at the ceiling:

condition	CURRENT	PREVIOUS	COVERAGE
`none`	0.0000	0.0000	0.0000
`overwrite`	1.0000	0.0000	0.0000
`duplicates`	0.9992	0.9988	0.9988
`budding`	0.9928	0.9928	0.9928

P1 (primary) does not confirm: budding − duplicates in COVERAGE = −0.0060, 95% CI [−0.0077, −0.0043], against a required +0.02.

P2 (control) passes: `overwrite` scores exactly 0.0000 on PREVIOUS, so the task measures what it claims. **P3 passes**: anchoring costs nothing on retrieving the current value (−0.0072, floor −0.02).

We stress the correct reading: with both conditions above 0.99, this is a **negative without power**, not a demonstration of equivalence. The ceiling is what motivated Experiment 1b.

4.2 Scaling the number of revisions

Populating the cluster forces the geometry to work. COVERAGE as a function of K revisions:

K	`duplicates`	`budding`	difference
1	0.9991	0.9949	−0.0042 [−0.0050, −0.0034]
2	0.9914	0.0567	**−0.9347** [−0.9390, −0.9304]
4	0.9482	0.0000	**−0.9482** [−0.9537, −0.9427]
8	0.2822	0.0000	**−0.2822** [−0.2912, −0.2732]

The mechanism, measured. Cosine of the current entry against the query, by revision index:

r	`budding` (fixed step)	`duplicates` (true text)
0	0.8106	—
1	0.7777	0.7887
2	0.6792	0.7683
4	**0.3235**	0.8064
6	**−0.1391**	0.8187
8	**−0.1391**	0.8015

With a fixed step ε per revision the cluster performs an **unbounded random walk**: the arc grows without limit (≈ 0.30 rad per step; 2.36 rad at $r = 8$), so from $r \approx 4$ the current entry is farther from the query than an arbitrary distractor and stops entering the top-k. Under `duplicates` each revision is the real embedding of its own text and stays at ≈ 0.80 for all r: there is no drift because there is no induced displacement.

This is not an artefact — it is a property of budding *as specified*, with ε fixed and explicitly non-re-tunable by the pre-registration.

5 Experiment 2: bounded budding

If unbounded drift is the problem, bounding it should suffice. We pre-registered a variant, `orbit`, in which every revision is placed at distance ε from the *original anchor* rather than from the previous revision, so the displacement cannot accumulate. A decaying-step variant and the original fixed-step condition were included as controls.

The repair works exactly as designed. P-A4 (mechanistic): cosine slope per revision is $+0.0000$ for `orbit` against -0.1410 for the fixed step; `orbit` sits flat at 0.8177 for every K. The blocking control P-A3 also passes: the fixed step reproduces its collapse (0.0000 at $K = 4$), so the harness is comparable across campaigns.

And the mechanism still loses. P-A1 (primary): COVERAGE at $K = 8$, `orbit` − `duplicates` = $\mathbf{-0.0864}$, 95% CI $[-0.1014, -0.0714]$. It loses at all four values of K.

K	`orbit`	`decay`	`fixed`	`duplicates`
1	0.9946	0.9946	0.9946	0.9991
2	0.9747	0.6805	0.0546	0.9914
4	0.8628	0.0282	0.0000	0.9482
8	0.1958	0.0000	0.0000	0.2822

Why. The cosine table gives it in one line: `duplicates` sits at 0.8540, `orbit` at 0.8177. Both are flat. The difference is a **constant toll of** ≈ 0.036 paid for displacing the entry ε away from the anchor, and that toll is enough to lose the top-k competition against entries belonging to other entities.

The underlying reason is that $\text{emb}(v_r)$ — the embedding of that version's real text — **is already optimally placed**, because it contains the entity the query mentions. Any artificial displacement, however bounded, can only move it away. *The geometry had nothing to add: the encoder had already put each version where it belonged.*

Per the falsification clause committed in advance, budding was closed as an indexing mechanism without trying a third geometry.

6 Experiment 3: elliptical corrections

6.1 The objection to our own conclusion

The closing argument of Section 5 rests on a property of the *generator*, not of corrections. Our generator builds every revision as a fully-specified template, entity included. Real conversational corrections are **elliptical**: *"no, it's Beto"*, *"sorry, I meant Beto"*. They do not name the entity, because the interlocutor already knows it.

In that regime the argument inverts: if the text does not carry the retrieval key, the encoder *cannot* place it near the query, and geometry would be the only available source of that key. So we tested our own conclusion where it should fail.

Viability check. Cosine against the query, $N = 3000$: self-contained revision 0.8541, elliptical revision 0.4247. An elliptical correction belonging to a *different* entity scores 0.4064 — statistically indistinguishable from the right one. In a smaller pilot ($N = 60$) the top-1 rank of the correct elliptical correction was 0.0167, i.e. exactly the chance rate 1/60.

6.2 The honest baseline

In this regime the rival is **not** bare `duplicates`, which we already know fails. It is `duplicates` + **coreference hydration**: rewriting *"no, it's Beto"* into *"The director of Helios is Beto" be-*

fore archiving, which is what deployed systems do and which they can do cheaply, since the conversation is at hand at write time.

We therefore swept the hydrator's failure rate τ: each entry is stored hydrated with probability $1 - \tau$ and raw with probability τ. Note that $\tau = 0$ reproduces the regime of Section 5 exactly, which gives us a blocking comparability control.

6.3 Results

CURRENT at $K = 8$:

τ	0.00	0.05	0.10	0.40	1.00
`hydrated`$_\tau$	0.5414	0.5368	0.5312	0.4884	**0.0000**
`orbit` (flat)			0.4723		

P-E0 (blocking) passes: at $\tau = 0$, hydration beats orbit by $+0.0691$ $[+0.0587, +0.0795]$, reproducing the toll measured in Section 5.

P-E2 (regime control) passes, and is the paper's most useful number: raw elliptical corrections score **0.0000**, with a CI of exactly $[0.0000, 0.0000]$ across all ten seeds. Not *low* — *never*.

P-E3 (mechanistic) passes: the cosine of the stored entry falls linearly in τ with slope -0.4291 (predicted -0.426), while `orbit` is flat at 0.8177.

P-E1 (primary) does not confirm. We required a crossover at $\tau^* \leq 0.25$. A finer grid, run after the verdict and reported as exploratory, locates it between $\tau = 0.40$ (-0.0161) and $\tau = 0.50$ ($+0.0349$): $\tau^* \approx$ **0.45**.

Practical statement. Geometric anchoring is worth its toll only if your coreference resolver fails on **45% or more** of corrections — worse than a coin flip. No deployed system is in that regime. The negative therefore *extends* to the regime designed to favour the mechanism, rather than being confined to self-contained text.

6.4 The finding that survives

Independently of budding: **a memory system that archives conversational turns without resolving coreference loses 100% of corrections.** The failure is total (0.0000, 10/10 seeds) and **silent**: the index always returns five candidates, ranked confidently, none of them right. There is no error signal, no empty result, no low-confidence flag — the correct correction sits at 0.4237 and a stranger's at 0.4064.

This is measurable, cheap to check, and applies to every system in the application-layer table of Section 2.

Structural note. An elliptical correction carries no entity information by construction, so the space of possible texts is tiny: four phrasings $\times$ 60 distinct values = **240 possible strings** for 24,000 stored entries. Two different entities corrected to the same value produce *bit-identical* vectors. That degeneracy is not an artefact of our design — it is the arithmetic form of the phenomenon under study, and it is why exact ties had to be broken explicitly in the ranking code rather than assumed away.

7 A methodological result: means conceal distributions

The pre-registration recorded a *point* prediction, $\tau^* \approx 0.075$, derived from a linear model of the cosine curves. The cosine mechanism behaved exactly as predicted — P-E3 passes, slope -0.4291 against -0.426 predicted — and the crossover *still* landed at 0.45, six times later.

The reason is worth stating because it generalises. $\texttt{hydrated}_\tau$ does not degrade uniformly: each entry is either **intact** (0.854) or **ruined** (0.42), never in between. The *mean* crosses `orbit` at $\tau \approx 0.08$, but top-k is decided entry by entry, and at $\tau = 0.10$ ninety per cent of queries still have a perfectly placed entry. Orbit, by contrast, pays its toll on *every* entry.

A second-order effect pushes the same way: as competing entries degrade, they stop competing for the top-k, which *helps* the healthy entry. This is why $\texttt{hydrated}_{0.40}$ scores 0.4884 rather than the $0.6 \times 0.5414 = 0.325$ that a multiplicative model predicts.

Rule of thumb: when a prediction is derived from an average, check the shape of the distribution before trusting the number. This is the third instance of a mean concealing its distribution in this research programme.

8 Limitations

1. **Non-parametric index over a frozen encoder.** Nothing here speaks to an index **co-trained inside the network**, where directions would be learned representations rather than fixed text embeddings, and where "the encoder already placed it correctly" stops applying by construction. That remains the open problem, with its two known obstacles: gradients do not flow through top-k selection, and the index goes stale as the encoder moves.
2. **Synthetic corpus.** Entities, attributes, and the four elliptical phrasings are synthetic and fixed; they are not a sample of human corrections.
3. **Hydration is modelled as all-or-nothing.** A real resolver can fail in worse ways — hydrating with the *wrong* entity — which we do not model and which would be more damaging than not hydrating at all.
4. **τ is imposed, not measured.** The experiment yields the crossover point; it does not tell you where a given production system sits on that axis.
5. **Single encoder family.** All results use `nomic-embed-text`; the capitalisation pathology of Section 3.4 is a reminder that encoder-specific behaviour is real.

9 Conclusion

We set out to test whether the geometry of an embedding index can carry the relation between versions of a remembered fact. Across three pre-registered experiments the answer is no, each time for an identifiable reason: with a fixed step the cluster walks away; bounded, it pays a constant toll; and in the elliptical regime designed to favour it, it only wins when the alternative is already broken beyond usefulness.

The negative is strengthened, not weakened, by the third experiment: we built the condition under which our own explanation should fail, and it did not. What remains is a design verdict that agrees with the deterministic-freshness position from an independent direction — *geometry to cluster, metadata to order* — and one practical warning with a number attached: if your memory system stores conversational turns as written, every correction that does not name its subject is already lost, and nothing in the system will tell you.

Data and code availability

All pre-registrations, hashes, analysis scripts, raw per-seed results, and the reports generated directly by the analysis code are archived in the project repository. Every verdict in this paper is emitted by a script that was frozen before the corresponding data existed; none was computed by hand.

Declarations

Funding. None.

Competing interests. The author declares no competing interests.

Author contributions. The author is solely responsible for the design, implementation, analysis, and writing.

References

[1] Wu, D. et al. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. *International Conference on Learning Representations (ICLR)*, 2025.

[2] Don't Ask the LLM to Track Freshness: A Deterministic Recipe for Memory Conflict Resolution. *arXiv preprint* arXiv:2606.01435, 2026.

[3] REMem: hybrid memory graph with gists and timestamped facts. *International Conference on Learning Representations (ICLR)*, 2026.

[4] NeuSymMS: self-curated neuro-symbolic memory for persistent agents. *arXiv preprint* arXiv:2605.17596, 2026.

[5] Supersede: diagnosing and training the memory-update gap. *arXiv preprint* arXiv:2606.27472, 2026.

[6] Kanerva, P. *Sparse Distributed Memory*. MIT Press, 1988.

[7] Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* 538, 471–476, 2016.

[8] Attention and sparse distributed memory: formal correspondence. *arXiv preprint* arXiv:2303.11934, 2023.

[9] Basu et al. On the instability of coordinate systems learned jointly with the model. *arXiv preprint* arXiv:2603.22858, 2026.

[10] Wu, Y. et al. Memorizing Transformers. *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2203.08913.

[11] Memory Layers at Scale. *arXiv preprint* arXiv:2412.09764, 2024.

[12] Cui, W. HOLA: A Hippocampus for Linear Attention. *arXiv preprint* arXiv:2607.02303, 2026.

[13] Lufkin, L. et al. Hybrid Associative Memories. *arXiv preprint* arXiv:2603.22325, 2026.

[14] Query Drift Compensation. *arXiv preprint* arXiv:2506.00037, 2025.

Appendix A: reference provenance

Entries above are cited with the identifiers recorded in this project's literature review at the time of writing. Where an entry lists no author, the corresponding record in that review preserved the identifier and short title but not the full author list; the arXiv identifier is given so the reader can resolve it directly. No bibliographic field in this list was reconstructed from memory.