WHEN DO BRAIN-INSPIRED COMPONENTS IMPROVE AGENT MEMORY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term agent memory increasingly uses brain-inspired storage, forgetting, reinforcement, and retrieval, yet system-level gains leave the advantage of each mechanism unclear. Removing a component tests system dependence; it does not establish superiority over a functioning alternative. We formulate component attribution as a comparison defined by the replacement, measured endpoint, and state or resource constraint. Our matched-substitution protocol applies this formulation to storage substrates, memory dynamics, and retrieval routing, checks mechanism activation, and separates shared-state readout contrasts from closed-loop comparisons that allow retrieval feedback. These controls change what the observations support. Hybrid sparse distributed memory lowers answer accuracy by about 14.8 percentage points relative to dense storage on frozen synthetic streams, while favorable readout cells disappear against compact equal-byte dense controls. Temporal decay helps in tested regimes with an advancing clock and distinguishable trace states, whereas joint reinforcement can revive stale rules. Chronological answer replay leaves the joint benefit of dense readout and restricted reinforcement unresolved. Prospective graph retrieval improves bridge-question F1 by 9.24 points under a read constraint; a separate lifecycle study measures substantial construction and maintenance costs. The resulting design guidance is conditional: compare compact storage alternatives, test when temporal updates affect ranking, and evaluate routing quality alongside cost. Reliable attribution requires functioning substitutions within a stated operating regime, beyond removal effects alone.

1 INTRODUCTION

Long-term memory lets language-model agents retain experience across interactions. Existing systems combine several design choices: Generative Agents store experiences and synthesize reflections (Park et al., 2023), MemGPT manages information across memory tiers (Packer et al., 2023), and A-Mem links new memories and updates existing representations (Xu et al., 2025). Brain-inspired designs add associative storage, forgetting and reinforcement, or graph-based retrieval (Nechesov & Ruponen, 2025; Zhong et al., 2024; Gutiérrez et al., 2024).

Benchmarks such as LongMemEval evaluate information extraction, temporal reasoning, and knowledge updates over interaction histories (Wu et al., 2025). These system designs and evaluation tasks motivate a complementary attribution question: a system-level gain does not identify each mechanism's advantage over a functioning alternative. Two questions guide attribution and deployment: **does the system depend on a component, and does that component outperform a functioning alternative serving the same role?**

A removal ablation addresses dependence. In our reference system, removing feedback costs 67–77 percentage points (pp), yet append-only storage reproduces the tested plasticity controller's immediate question-answering (QA) outputs. Retaining corrections matters; this fixture cannot credit elapsed-time retention. Likewise, under a byte constraint, sparse distributed memory (SDM)'s apparent advantage depends on the dense alternative's representation. A mechanism can also be present but inactive if its clock is fixed or the evaluated path bypasses it. Removal effects leave these comparisons unresolved.

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

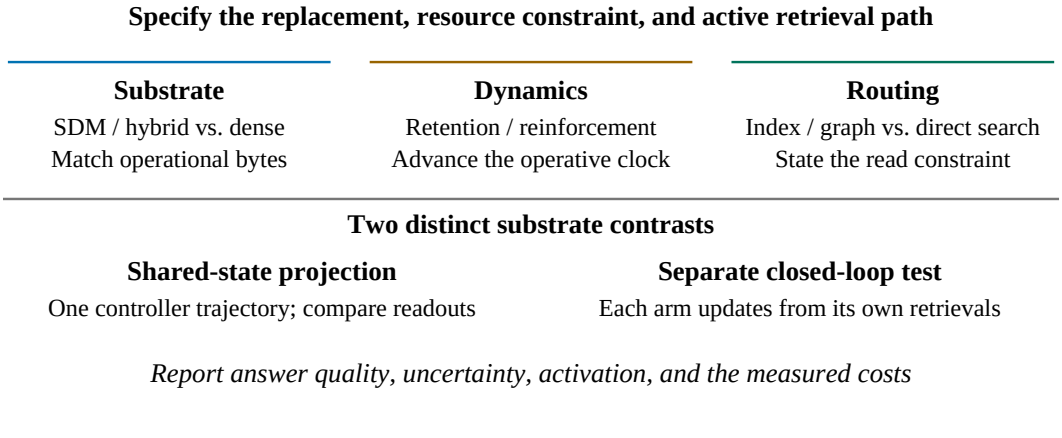


Figure 1: Matched-substitution workflow. Shared-state projection compares readouts along one controller trajectory; the separate closed-loop test restores retrieval-induced updates. Each comparison states its replacement, resource constraint, and activation check.

Our central insight is that **component advantage is a property of a specified comparison**. Its target mechanism, functioning replacement, endpoint, workload, and state or resource constraint define what is held comparable and when the mechanism can affect the outcome.

This formulation yields a matched-substitution protocol for **storage substrate, memory dynamics, and retrieval routing**, with activation checks and explicit controls. Constructive projection presents one controller trajectory to both substrates, isolating their direct readout contrast. A separate closed-loop comparison allows retrievals to change subsequent reconsolidation. Representation sweeps and advancing retention clocks test why contrasts arise. Figure 1 shows the workflow; Table 1 maps its implementations.

The controls change the interpretation at each layer. Hybrid SDM verification lowers accuracy by about 14.8 pp on 1,280 frozen synthetic streams, and compact equal-byte dense representations remove SDM's favorable cells in a separate readout sweep. Thus capacity assigned to the alternative can change an apparent substrate advantage. For dynamics, advancing the clock reveals useful decay, while joint reinforcement can erase the distinction between old and new traces. The combined dense-readout/top-1 update has unresolved answer-level benefit on RealMem: a mechanistic diagnosis need not yield a system-level gain. For routing, prospective HippoRAG testing finds a 9.24 pp bridge-question F1 gain under a read constraint, while separate lifecycle measurements reveal substantial construction and maintenance costs under the measured policies.

The resulting design choices are conditional: compare compact storage, check whether temporal updates change ranking, and assess routing gains against deployment costs. This supports attributing a gain to an implementation under stated conditions, beyond dependence established by removal.

Contributions. (1) We formulate component attribution through functioning replacements, activation checks, state/resource controls, and distinct direct and closed-loop contrasts. (2) We explain how representation, time, and update semantics change the attributed effect, supported by controlled readout and answer experiments. (3) We test transfer through external systems, chronological replay, prospective routing predictions, and replicated lifecycle measurements, identifying positive gains as well as unresolved benefits and policy-specific costs.

2 Problem Setup and Attribution Protocol

We specify a component-attribution comparison as $(M, M_0, Y, \mathcal{W}, \mathcal{C})$: target mechanism M, functioning alternative M_0, endpoint Y, workload $\mathcal{W}$, and state or resource controls $\mathcal{C}$. The reported effect is the endpoint contrast within that comparison. The protocol checks activation and distinguishes shared-state readout from retrieval-induced feedback.

Table 1: Systems and matched substitutions. Each comparison has its own resource constraint; endpoints, patches, and study status are detailed in Table N1.

System	Replaced mechanism	Matching criterion
MAM	SDM readout; controller; episodic index	Bytes and state; feedback; access path
MemoryBank	SDM readout; forgetting	Bytes or candidate pool
HippoRAG 2	Graph/dense; full/one-step propagation	Read exposure or shared ceiling
MG^2-RAG	Graph/dense; one-step; randomization	Read exposure
HippoMM	Segmentation; summaries; filters; gates	Segment, token, or frame budget

Our reference implementation, Modular Agent Memory (MAM), exposes trace storage, temporal updates, and retrieval routing for the substitutions studied here. We compare SDM and dense readouts of the trace store. The temporal controller applies salience-dependent decay,

$$w(t) = w_0(1 + \Delta t/\tau)^{-\alpha(s)}, \tag{1}$$

plus reconsolidation, replay, dormancy, and interference. Routing selects cross-agent trace handles or searches the cold archive for retrieval-augmented generation (RAG). Appendix A.6 states the implementation and release boundary.

Specify and activate. Check that the intervention reaches the evaluated path: an absent hot store, fixed clock, or unused gate can make a mechanism inactive. Operational bytes, read exposure, and segment/frame budgets define different comparisons, not one lifecycle budget (Table 1).

Couple state, then restore feedback. Constructive projection computes the controller trajectory from a common event schedule and maps it onto both substrates, fixing encoder, seeds, writes, and state. Its 240-pair audit tests a constructed identity, not an absence of statistical significance (Appendix I; Proposition 1). The separate closed-loop experiment instead gives each arm its own retrieval-to-reconsolidation trajectory. We report these effects separately and compare substrate contrasts across controllers to assess interaction. Projection requires a common state representable by both stores; jointly rewritten traces may require another intervention.

Figure 2 makes the separation concrete: identical active-trace trajectories coexist with different readout accuracy. Equality in the upper panel follows from the shared-state construction and is an implementation check, not independent evidence that feedback never matters. The lower panel is the empirical readout comparison. Releasing the coupling in the closed-loop study tests the path that projection deliberately removes. Figure B1 provides a layer-wise quantitative overview, with its distinct endpoints kept separate.

3 Layer 1: Storage Substrate

Does associative readout improve retrieval at the same storage budget? We compare readout at shared state, restore feedback in a closed-loop test, and vary representation to distinguish retrieval quality from capacity.

3.1 Readout controls and closed-loop confirmation

The pre-registered primary grid spans independent and identically distributed (i.i.d., labeled iid), clustered, and drifting populations, six loads, five cue conditions, and 20 seeds. Controls match operational bytes; reservoir sampling is primary. Equivalence requires the paired-seed 90% confidence interval (CI) to lie within ± 2 pp. Across 180 comparisons, dense wins 178, one is equivalent, and SDM wins one (Table S1). That exception is iid/N=1024/clean (+35.83 pp). Held-out tuning, cleanup/address controls, and an independent C++ implementation support the tested readout diagnosis (Appendix E).

An earlier three-reader answer matrix does not establish equivalence: two one-sided tests (TOST) fail at ± 2 and ± 3 pp, with minimum detectable effect (MDE) about 4 pp. Its post-hoc benefit bounds and pure-SDM extension remain exploratory (Table S3). We therefore froze a separate 1,280-stream generator, 11 arms, two answer families, exact-match scoring, and 12 tests before observing outcomes. Each stream supplies 20 writes and 12 queries at days 10, 40, 120, and 730,

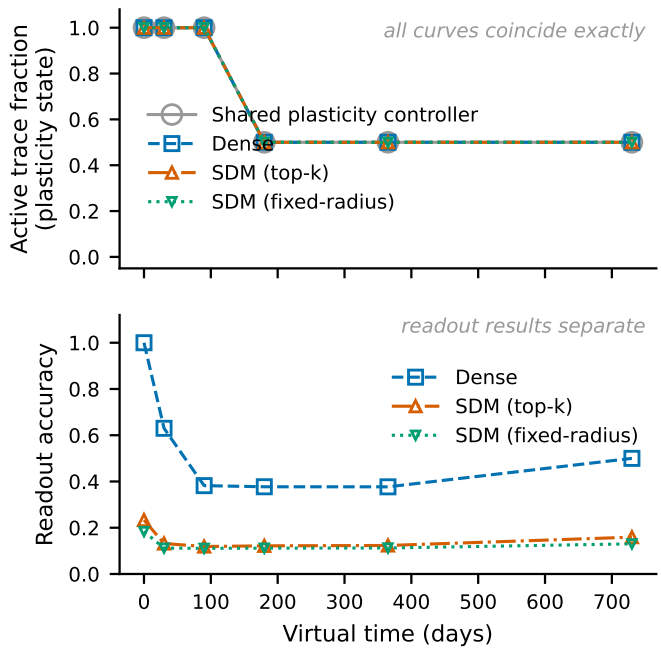


Figure 2: Constructive projection: plasticity state coincides across substrates (top, 240/240 paired units), while readout results separate (bottom, 0/240 identical units). The upper identity is constructed; the lower contrast identifies the controlled direct effect along this fixed trajectory. Appendix I gives the audit.

Table 2: Closed-loop answer effects with the full controller. Differences are arm minus dense in pp; the last column is the one-sided 95% stream-bootstrap upper bound. All 12 frozen tests, including other controllers, exclude a +2 pp benefit (Table S10).

Reader	Readout	Difference (pp)	Upper bound (pp)
Qwen	Hybrid SDM	−14.857	−14.310
DeepSeek	Hybrid SDM	−14.805	−14.251
Qwen	Pure SDM	−58.210	−57.897
DeepSeek	Pure SDM	−58.060	−57.734

including current, historical, stable, and unanswerable questions. Each arm follows its own update trajectory.

Across 337,920 terminal answers, hybrid verification lowers accuracy by 14.857 pp for Qwen (two-sided 95% CI $[-15.514, -14.199]$) and 14.805 pp for DeepSeek ($[-15.469, -14.147]$). Table 2 shows the full-controller contrasts. The ordering persists with no reconsolidation, no decay, append-only, and time/version controllers: every one-sided upper bound is below +2 pp. Retrieval feedback does not reverse the substrate result on this generator; it is not a population estimate for deployed agents.

3.2 Representation budget explains the observed exception

We hold population, queries, seeds, and bytes fixed and change only the dense representation (Figure 3(a)). At the exception cell, equal-operational-byte binary dense reaches 1.0000 and reverses the contrast to -7.01 pp (90% CI $[-7.79, -6.20]$); int8 and float16 do the same. Across the separate 48-cell sweep, SDM wins 6/3/1/0 cells against float32/float16/int8/binary. These counts use a CI-excluding-zero rule and a different grid from the 180-comparison primary test.

The explanation is capacity: SDM uses $256 \times 1024 \times 5 = 1{,}310{,}720$ substrate bytes; a bit-packed dense row uses 156 bytes. Even equal substrate bytes buy 8,402 rows. Under the operational budget,

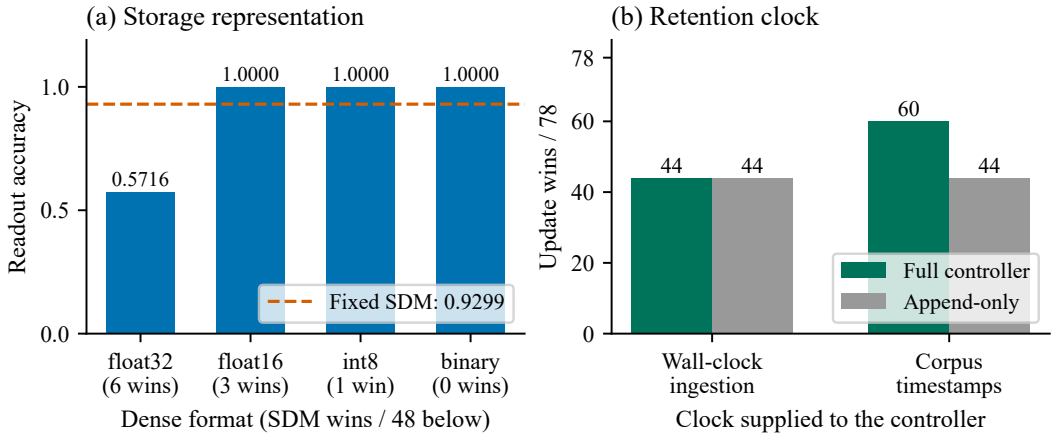


Figure 3: Two controls that change the attributed effect. (a) Dense representation at the iid/N=1024/clean exception, with equal operational bytes and fixed SDM readout; counts below the axis summarize SDM wins in the separate 48-cell sweep. (b) On 78 LongMemEval update questions, an operative corpus clock raises full-controller update wins from 44 to 60, versus 44 for append-only: 16 gains, no losses (exact McNemar $p = 3.1 \times 10^{-5}$). These are distinct retrieval endpoints, not generated-answer accuracy. Representation heatmaps and intervals remain in Figure D1 and Table S4.

$\lfloor (1{,}310{,}720 + 1052N)/156 \rfloor \geq N$, so binary dense is never capacity-constrained. This identifies a representation dependency in the tested comparisons, not the source of gains in every SDM system. Include compact dense controls and state which bytes count; keep source evidence separately from its retrieval representation.

3.3 Public workloads and a second implementation

Sign random projection locality-sensitive hashing (SRP-LSH) maps angle to expected Hamming distance as θ/π (Lemma 1). Under both tested encoders, all 510 public query points lie outside the grid's empirical SDM-favorable region: the corpora are correlated, underloaded (0.11–0.19×), and noisy (0.31–0.37; Table S8). On LoCoMo and LongMemEval, the same-state hot-store swap, with no large language model (LLM) in the loop, favors dense on every retrieval metric except LoCoMo any-gold recall (hybrid +0.10 pp); dense is faster and returns less context (Table S2).

Figure 4 separates the data geometry from load: changing store size moves the vertical coordinate, but not correlation or cue noise. The comparison remains specific to the tested encoders and segmentation.

MemoryBank's fixture provides a second implementation: matched-byte SDM lowers hit@2 (a gold item in the top two) from 0.7714 to 0.1714, reaching only 0.1857 with sixfold bytes (20 seeds; Appendix D). These are retrieval-level extensions, not further closed-loop answer confirmations. Geometric placement likewise does not fully determine answer quality.

4 Layer 2: Memory Dynamics

When does changing trace state affect retrieval? We keep substrate and feedback intact, replace the controller, then advance time and ablate updates. Chronological answer replay tests transfer beyond the controlled stale-rule fixture.

4.1 Activation depends on the operative clock

On 1,200 paired rows from the immediate-QA fixture, dense-hot with the full controller and append-only storage return identical results; only measured latency varies (Table S9). Proposition 2 com-

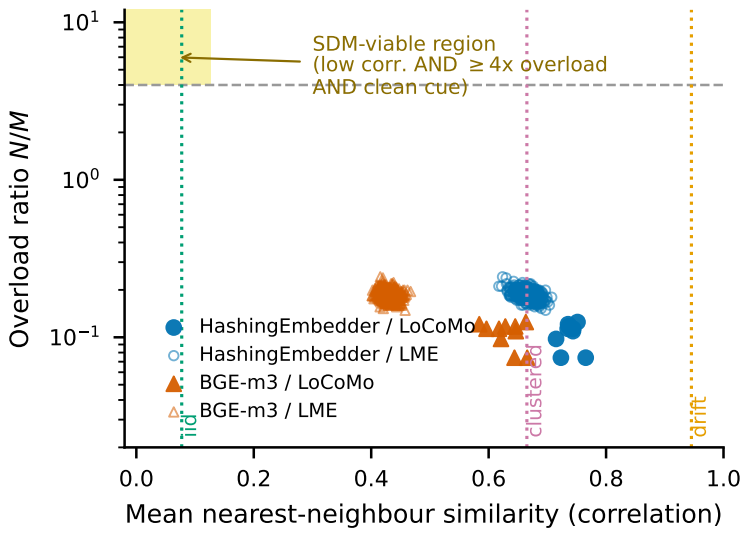


Figure 4: Public workloads in the empirical phase-diagram coordinates. None of 510 workload instances satisfies the joint low-correlation, high-load, and clean-cue conditions under either encoder. The two plotted coordinates do not show cue noise; Appendix G reports all three. This placement is not a universal answer-quality guarantee.

bines Eq. (1) with the score factor

$$\varphi(w) = \tfrac{1}{2} + \tfrac{1}{2}\min(w, 1). \tag{2}$$

At $\Delta t = 0$, decay equals one; with initial $w_0 = 1$ and reinforcement that only increases it, the score stays saturated. In one instrumented cell, reconsolidation fires 384 times and changes w_0 on 105 occasions without changing ranking. The 67–77 pp feedback-removal loss therefore cannot be assigned to elapsed-time retention in this fixture.

Figure 3(b) shows why corpus timestamps must reach the implemented clock. A static history can test retention once re-clocked. It can also test mechanisms that act without elapsed time: removing Generative Agents' in-corpus recency term reshuffles 28–32% of its top-30. Reflection, trace-writing replay, and index reorganization are not ruled out by the fixed-clock identity.

External activation check. MemoryBank's shipped configurations disable forgetting; when enabled, it deletes candidates probabilistically at index construction. At zero time deletion is inactive. On 566 memories, 15 users, and 35 date-anchored questions, its nonzero-time settings instead lower hit@2 by 14.9 pp at the fixture's last day and 77.1 pp at the wall-clock default (20 seeds; Appendix F). Deletion can improve or worsen retrieval in general (Proposition 3); these observed losses are not a theorem about forgetting.

4.2 Decay helps, but joint reinforcement can revive stale rules

A post-hoc 730-day fixture writes 12 old/new rule pairs, with supersession at day 90. On dense storage across 20 seeds, decay prevents stale top-1 returns when rules are unused, while append-only storage returns the old rule. Old/new saliences are 0.7/0.9; the exact separation is a controlled tie-regime demonstration, not deployed prevalence. Table 3 localizes the benefit to decay and the reuse failure to co-reconsolidation. Replay does not enter this fixture's path.

Co-recall raises both base weights and saliences:

$$w_0 \mapsto \min(w_{\max}, b\, w_0), \qquad s \mapsto \min(1, s + \delta), \qquad k^* = \left\lceil \frac{\ln w_{\max}}{\ln b} \right\rceil = 4. \tag{3}$$

Base weights saturate after four co-recalls; salience equality additionally needs six for the stale trace. With matching reset times, decay curves then coincide and content decides ranking while traces remain eligible (Theorem 1). Top-1 reinforcement avoids this paired update; Theorem 2 gives

Table 3: Which dynamics mechanism changes stale-rule retrieval? Dense substrate, 730 days, 20 seeds. Entries are zombie rates (%); each is identical across seeds. Full intervals and effects remain in Table S7.

Controller	Without reuse	With reuse
Full controller	0	100
Append-only (all disabled)	100	100
Disable decay	100	100
Disable salience grading	0	100
Disable reconsolidation	0	0
Disable dormancy	0	100

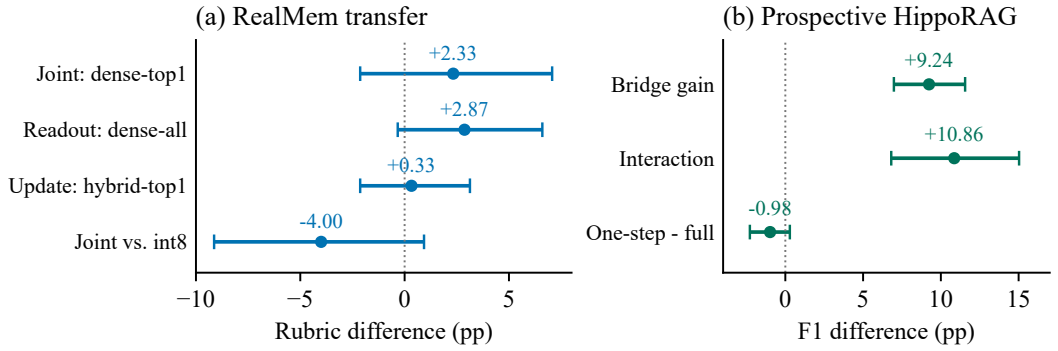


Figure 5: Answer-level transfer and prospective routing evidence, with descriptive 95% intervals. Left: paired-persona RealMem bootstrap (ten personas); the first three contrasts use hybrid-all as reference, the last uses timestamp-aware int8. Right: prospective HotpotQA contrasts. Its frozen decisions use multiplicity-adjusted one-sided bounds, reported in Section 5, not these descriptive intervals.

sufficient separation conditions, not an unconditional guarantee after a single win. The ablations neither compare decay families nor establish dormancy as generally ineffective; salience grading helped on the separate hybrid fixture (Appendix D).

Check the endpoint. In a separate substrate comparison, dense/hybrid/pure-SDM zombie rates are 100/30.83/3.75%, but pure SDM retrieves unrelated rules on 73.75% of probes and the correct rule on only 22.50% (Table D1). A low zombie rate can reflect retrieval failure. Explicit supersession relations are therefore useful validity signals; historical questions still require earlier evidence.

4.3 Chronological answer transfer remains uncertain

RealMem (Bian et al., 2026) contributes ten simulated personas, 500 selected queries, and 2,500 answers across five arms. We replay earlier sessions, advance the clock, withhold target references, and let each arm update from its own retrievals. Qwen answers; a separate, provider-reported DeepSeek family grades a 0–3 rubric. Figure 5 separates the joint dense-top1 intervention from readout-only and reinforcement-only changes. Every RealMem interval spans zero. The results establish neither superiority nor equivalence; the joint contrast cannot isolate storage, and normalized rubric scores are not answer accuracy (Appendix O).

5 Layer 3: Retrieval Routing

Which routing choices improve evidence access, and what do they cost? We test cross-agent substitution at answer parity, graph retrieval under a read constraint, and separate construction and maintenance policies.

Access versus implementation. Removing MAM's episodic handle index costs 27.08/70.83/75.00 pp across three readers. Replacing it with direct archive search instead reaches 1.000 cross-agent

Table 4: Lifecycle phase-wall time, mean of three repetitions. Read phases include worker setup. Full and one-step each incur the shared graph's build and rebuild costs as logical alternatives. These cost runs are separate from the fixed-reset quality test; the measured maintenance policies differ (Appendix Q).

Phase (unit)	Dense	Full graph	One-step graph
Initial construction (h)	0.45	3.31	3.31
1,000 initial reads (s)	293.37	1,157.59	1,099.19
Mixed maintenance	10.31 s	4.02 h	4.02 h

accuracy for all three, but requires 1,200 cold searches versus zero with the index (Table S3). The measured contribution is avoided cold search at answer parity, not a requirement to use SDM locally. The isolation audit retains its one cue-prefix exception (Appendix B).

External graph retrieval. HippoRAG 2 (Gutiérrez et al., 2025) uses personalized PageRank (PPR) over an open information extraction (OpenIE) graph. On MuSiQue-1000 and 2Wiki-1000 (Trivedi et al., 2022; Ho et al., 2020), full graph retrieval beats dense by +3.86/+9.81 pp token-F1 (one-sided lower bounds +1.63/+8.04; Appendix N). Corpus, encoder, reader and exact-match (EM)/F1 scoring are shared; read exposure meets the frozen signed-median criterion, not per-query identity. Graph construction additionally uses 14.7M/8.0M LLM tokens versus dense's zero. These are read-constrained gains, not equal-lifecycle-budget gains.

Applying PPR once while holding graph, linking, and reset fixed retains the full observed MuSiQue margin and about three quarters of 2Wiki's. This localizes much of that margin to the retained pipeline without separating graph and linking. The score-vector bound in Proposition 4 (Eq. (K4)) is not a bound on answer quality.

Prospective predictions. A frozen HotpotQA test uses 1,000 questions, 9,811 passages, and a common top-five/6,000-token ceiling. Full/dense/one-step F1 is 68.14/60.95/67.16%. Figure 5 shows the contrasts: full graph improves the 811 bridge questions by +9.24 pp, with adjusted lower bound +6.79; the bridge-versus-comparison interaction also passes (+10.86, lower bound +6.41). However, one-step minus full has adjusted lower bound -2.38 pp, missing the -2 pp noninferiority criterion. This neither establishes equivalence nor proves that further propagation helps. Only full-versus-one-step isolates extra propagation; full-versus-dense changes the whole retrieval bundle (Appendix P).

Lifecycle measurement. Three fresh repetitions add construction, reads and a fixed mixed update batch (Table 4). Dense's native updates take 10.31 s; the conservative graph-rebuild policy takes 4.02 h. One-step retains graph construction and most measured read cost. These are policy costs, not an intrinsic graph-maintenance lower bound. The original exact-vector maintenance check failed; its approved post-observation validation amendment remains a reported limitation, with the failure retained.

6 Related Work

Associative storage and memory architectures. SDM and related associative memories motivate alternatives to explicit dense retrieval (Kanerva, 1988; Bricken & Pehlevan, 2021; Nechesov & Ruponen, 2025). Universal Hopfield Networks decompose associative retrieval into similarity, separation, and projection (Millidge et al., 2022); agent-memory surveys organize designs by forms, functions, and dynamics (Hu et al., 2025b). Our decomposition serves an experimental purpose: it specifies replaceable components and separates the effect of readout at a shared controller state from the effect of retrieval-induced updates. It does not replace these mathematical frameworks or taxonomies.

Temporal memory and graph retrieval. MemoryBank uses forgetting and reinforcement (Zhong et al., 2024), while Generative Agents combine stored experiences with reflection and retrieval (Park et al., 2023). HippoRAG and HippoRAG 2 develop graph retrieval inspired by hippocampal indexing (Gutiérrez et al., 2024; 2025). We examine activation of the implemented mechanisms and compare functioning alternatives: an elapsed-time rule requires an operative clock, whereas a graph-versus-dense gain need not identify the contribution of further propagation.

Evaluation. LoCoMo, LongMemEval, and MemoryAgentBench test long conversations, memory abilities, and incremental interactions (Maharana et al., 2024; Wu et al., 2025; Hu et al., 2025a). Our contribution concerns how to interpret a component contrast on such workloads: specify the replacement, verify activation, and report the relevant state and resource controls.

Controlled evaluation is also central to MemDelta, which varies components within a common pipeline and examines model, embedding, and retrieval confounds (Wang, 2026). Analyses of agentic memory discuss metric validity, judge sensitivity, and maintenance overhead (Jiang et al., 2026); scale-conditioned evaluation holds task evidence fixed while irrelevant sessions accumulate (Shao et al., 2026). These approaches address complementary sources of variation. We focus on controller coupling and mechanism activation, then test selected predictions prospectively and measure implementation costs separately. This positions our work as a component-attribution study within a broader evaluation effort, rather than treating matched controls as a new idea in themselves. Further discussion appears in Appendix R; Table 1 inventories the system-specific comparisons.

7 Discussion

The findings turn component attribution into testable design choices. **Storage:** include compact dense alternatives before crediting associative retrieval; representation changes capacity under a byte constraint. **Time:** advance the operative retention clock and query multiple future times to establish when decay affects ranking. **Updates:** restrict joint reinforcement when it would refresh superseded traces, while testing answer-level transfer; RealMem leaves the combined benefit unresolved. **Routing:** report bridge-question gains alongside construction, reading, and maintenance costs, keeping the measured policies explicit.

These choices follow from the replacement and operating regime, not biological similarity alone. Output identity, an uncertain estimate, and equivalence within a prespecified margin remain distinct findings. Functioning substitutions show which implementation can be credited for an observed gain; activation and resource checks delimit that claim. The different endpoints should not be pooled into one cross-layer effect size.

8 Limitations

Our synthetic streams control writes, supersession, feedback, and query time but do not reproduce deployed-agent distributions. Stress tests fix encoders, controller settings, and content geometry. External implementations broaden the coverage without exhausting these choices. Shared-state projection also assumes that both stores can represent a meaningful common controller state.

RealMem replays observed dialogue from ten simulated personas and uses an automated judge; it is not an autonomous interaction study with real users. Not every retrieval diagnosis has an answer-level counterpart. HippoMM's summary, segmentation, and flat-context controls do not establish equivalence; its filter estimate becomes uncertain under tighter exposure (Appendix N).

The lifecycle study covers one synthetic update batch and one conservative rebuild policy, omitting shared-service GPU use, energy, and currency. Maintained vectors failed exact equality, and the amended comparison is descriptive cost, not equal-representation quality (Appendix Q). Alternative update policies and matched lifecycle budgets are needed for a deployment-efficiency claim.

9 Conclusion

Brain-inspired components warrant credit for a gain when they outperform a functioning alternative under explicit conditions. Our protocol tests activation, controls state or resources, and separates readout from retrieval feedback. The experiments show why representation, operative time, and reinforcement semantics matter, while routing requires both quality and lifecycle measurements. Chronological replay leaves the combined configuration's benefit unresolved. The contribution is conditional attribution: identify which implemented mechanism improves the measured endpoint, and make the boundary of that evidence explicit.

REPRODUCIBILITY STATEMENT

The supplement contains saved outcomes, protocols, the frozen MAM core, and analysis scripts. The frozen confirmation can be statistically recomputed from the audited terminal rows using the supplied analysis entry point. The new chronological and prospective studies include saved prompts, outputs and offline scoring; lifecycle audit coverage and reconstruction instructions are stated in Appendix Q and the supplement. Some experiments require external inputs or model services not bundled here; the package should not be interpreted as a complete end-to-end installation. Appendix A.6 specifies this boundary.

AI USE STATEMENT

LLMs assisted sentence polishing, paragraph structuring, statistical script validation, artifact auditing, and LaTeX typesetting. Model use for answer generation, synthetic fixtures, and rubric scoring is documented in the corresponding experimental protocols.

The authors reviewed and approved the protocols, authorized data and API use, audited the saved outputs, and take responsibility for all claims and research artifacts.

REFERENCES

Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang, and Ronghao Chen. RealMem: Benchmarking LLMs in real-world memory-driven interaction, 2026. URL https://arxiv.org/abs/2601.06966.

Trenton Bricken and Cengiz Pehlevan. Attention approximates sparse distributed memory. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. URL https://papers.nips.cc/paper/2021/hash/8171ac2c5544a5cb54ac0f38bf477af4-Abstract.html.

Trenton Bricken, Xander Davies, Deepak Singh, Dmitry Krotov, and Gabriel Kreiman. Sparse distributed memory is a continual learner. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2303.11934. arXiv:2303.11934.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions, 2025a. URL https://arxiv.org/abs/2507.05257. arXiv preprint; latest revision June 2026.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of AI agents, 2025b. URL https://arxiv.org/abs/2512.13564. arXiv preprint.

Dongming Jiang, Yi Li, Songtao Wei, Jinxin Yang, Ayushi Kishore, Alysa Zhao, Dingyi Kang, Xu Hu, Feng Chen, Qiannan Li, and Bingzhe Li. Anatomy of agentic memory: Taxonomy and empirical analysis of evaluation and system limitations, 2026. URL https://arxiv.org/abs/2602.19320. arXiv preprint.

Pentti Kanerva. *Sparse Distributed Memory*. MIT Press, Cambridge, MA, 1988. ISBN 9780262111324. URL https://mitpress.mit.edu/9780262111324/sparse-distributed-memory/.

Pentti Kanerva. Sparse distributed memory and related models. Technical Report RIACS TR 92.10 (NASA-CR-190553), Research Institute for Advanced Computer Science (RIACS), NASA Ames Research Center, Moffett Field, CA, April 1992. URL https://ntrs.nasa.gov/citations/19920021480.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal Hopfield networks: A general framework for single-shot associative memory models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15561–15583. PMLR, 2022. URL https://proceedings.mlr.press/v162/millidge22a.html.

Andrey Nechesov and Janne Ruponen. CALM: Continual associative learning model via sparse distributed memory. *Technologies*, 13(12):587, 2025. doi: 10.3390/technologies13120587. URL https://www.mdpi.com/2227-7080/13/12/587.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22. ACM, 2023. doi: 10.1145/3586183.3606763.

Jiaqi Shao, Yiyi Lu, Yunzhen Zhang, and Bing Luo. When stored evidence stops being usable: Scale-conditioned evaluation of agent memory, 2026. URL https://arxiv.org/abs/2605.07313. arXiv preprint.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Kuan Wang. MemDelta: Controlled baselines and hidden confounds in agent memory evaluation, 2026. URL https://arxiv.org/abs/2606.29914. arXiv preprint.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0593. URL https://papers.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/AAAI.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

SUPPLEMENTARY TABLES

Tables S1–S10 collect detailed results cited in the main text. Figures 1 and 3(a) present the protocol and representation comparison; appendix tables and figures use their appendix letter.

Table S1: Substrate verdict — phase diagram, frozen tuning split, external canonical implementation

SDM arm	Comparisons	SDM wins	Dense wins	Equivalent	Inconclusive
sdm_topk	90	1	88	1	0
sdm_radius	90	0	90	0	0

Independent tuning, then frozen (activation-width selection on training seeds)

k	2	4	8	12	16	24	32
Train top-1 mean	0.1661	0.1703	0.1776	0.1747	0.1319	0.0641	0.0280

External canonical KanervaSDM cross-check (independent C++ implementation; a cross-check of the observed direction)

Byte-budget convention	Cells	SDM wins	Dense wins	Inconclusive
dense_equal_substrate	72	5	61	6
dense_equal_operational	72	4	63	5

Table S2: Same-question hot-layer ablation on public benchmarks (gold-evidence recall, no LLM). MRR: mean reciprocal rank.

Dataset	Metric	Hybrid-SDM	Dense-only	Diff
LoCoMo	Any-gold recall	0.9319	0.9309	+0.10pp
LoCoMo	All-gold recall	0.8370	0.8446	-0.76pp
LoCoMo	MRR	0.7380	0.7422	-0.0042
LoCoMo	Mean retrieval (ms)	0.7431	0.5659	+0.1773
LoCoMo	Mean context chars	26,117	25,295	+821
LoCoMo	Any-gold paired	hybrid-only wins 27	dense-only wins 25	ties 1930
LongMemEval	Any-gold recall	0.9560	0.9740	-1.80pp
LongMemEval	All-gold recall	0.8480	0.8700	-2.20pp
LongMemEval	MRR	0.8674	0.8921	-0.0247
LongMemEval	Mean retrieval (ms)	4.4416	2.9478	+1.4938
LongMemEval	Mean context chars	110,758	106,747	+4,011
LongMemEval	Any-gold paired	hybrid-only wins 2	dense-only wins 11	ties 487

Table S3: Frozen held-out — one-sided non-inferiority bound, plus the pure-SDM arm

(a) SDM verification layer (Full MAM - dense_hot). Exploratory statistic: one-sided 95% upper bound. The two-sided TOST fails at both +/-2 and +/-3 pp, which we disclose rather than omit.

Answer model	Delta	90% CI	SDM benefit bound	MDE	Equiv 2pp	Equiv 3pp
Qwen	-3.91pp	[-6.25,-1.56]	**-1.56pp**	4.02pp	no	no
gpt-5.6-sol	-0.91pp	[-3.39,+1.56]	**+1.56pp**	4.19pp	no	no
Gemini	+1.43pp	[-0.91,+3.78]	**+3.78pp**	4.05pp	no	no

Removal effects measured at the same sample size (these do not establish power for a small substitution effect)

Answer model	vs rag-only	vs no-feedback	vs no-hippocampus
Qwen	+2.60pp (MDE 6.1)	+67.45pp (MDE 7.5)	+27.08pp (MDE 24.7)
gpt-5.6-sol	+6.38pp (MDE 5.2)	+76.17pp (MDE 6.1)	+70.83pp (MDE 19.3)
Gemini	+10.16pp (MDE 5.1)	+76.82pp (MDE 6.7)	+75.00pp (MDE 19.0)

(b) pure-SDM arm, added after the freeze (post-hoc, not pre-registered; the frozen protocol records why it had been excluded). Pure-SDM post-feedback accuracy 0.4556 (n=720)

Comparator	Comparator acc	Delta (pure - comp)	90% CI	McNemar p
Full MAM	0.8056	-35.42pp	[-40.36,-30.34]	4.71e-45
dense_hot	0.8472	-39.32pp	[-44.27,-34.24]	1.38e-53
rag_only	0.7778	-32.81pp	[-38.15,-27.34]	7.77e-34
no_feedback	0.1417	+32.03pp	[+27.47,+36.59]	9.59e-47

Table S4: Compact dense controls reverse the selected SDM exception under stated byte budgets

Exception cell: iid / N=1024 / clean. SDM is fixed. Op.: equal operational bytes; Sub.: equal substrate bytes. Δ: SDM minus dense.

Budget	Format	Capacity	Accuracy	Δ (pp)	90% CI	Winner
Op.	binary	1024	1.0000	-7.01	[-7.79,-6.20]	**Dense wins**
Op.	float16	1024	1.0000	-7.01	[-7.79,-6.22]	**Dense wins**
Op.	float32	579	0.5716	+35.83	[+34.84,+36.85]	SDM wins
Op.	int8	1024	1.0000	-7.01	[-7.81,-6.22]	**Dense wins**
Sub.	binary	1024	1.0000	-7.01	[-7.79,-6.22]	**Dense wins**
Sub.	float16	631	0.6206	+30.94	[+29.27,+32.55]	SDM wins
Sub.	float32	317	0.2971	+63.28	[+61.48,+64.97]	SDM wins
Sub.	int8	1024	1.0000	-7.01	[-7.81,-6.22]	**Dense wins**

Wins evaporate monotonically as the dense rows are compressed (main grid, sdm_topk, 48 cells per representation)

Dense repr	SDM wins	Dense wins
float32	6	42
float16	3	45
int8	1	47
binary	0	48

Table S5: Tested cleanup caps and address schemes on iid and clustered populations

(a) Iterative-cleanup cap (random addresses; raising the cap lowers accuracy monotonically in all four cells)

Family/condition	cap=3	cap=10	cap=50	Delta (50-3)	cap-3 truncation
iid/clean	0.9299	0.6445	0.6443	-28.57pp	87.37%
iid/flip15	0.1583	0.0828	0.0828	-7.55pp	92.68%
clustered/clean	0.0182	0.0120	0.0104	-0.78pp	95.36%
clustered/flip15	0.0128	0.0078	0.0062	-0.65pp	99.51%

(b) Hard-address decorrelation (clustered / N=1024, cap=3; equidistant addresses do not improve these cells)

Address scheme	clean	flip15	Pairwise Hamming std	Min distance
random	0.0182	0.0128	16.0103	435
decorrelated	0.0156	0.0104	0.0000	512
data_adaptive	0.0159	0.0159	48.2672	124

Table S6: Layer-wise contribution — three regimes, not one

The substrate QA entry comes from the closed-loop confirmation; the retention QA entry comes from the fixed-clock fixture. Long-horizon entries are zombie rates, not answer accuracy. The retention row uses the dense substrate and 20 seeds; the substrate row compares stores. Retention arrows denote full-controller to append-only replacement.

Layer	QA and operational endpoints	Long-horizon, unused (730d)	Long-horizon, reused (730d)
Substrate	Closed loop: hybrid −14.86/−14.80 pp and pure SDM −58.21/−58.06 pp (Qwen/DeepSeek; Table S10)	SDM 0.00% vs Dense 0.00% (Δ=+0.00pp)	**SDM 3.75% vs Dense 100.00% (Δ=-96.25pp)**
Elapsed-time retention	**Exact output identity in the tested dense fixture**	**0.00% → 100.00% (+100.00pp)**	100.00% → 100.00% (+0.00pp)
Routing	Answer parity with cold search; avoids 1,200 archive searches	—	—

Table S7: Second-order attribution within the dynamics (dense substrate, 730 days, 20 seeds)

Configuration matches dense_hot in Section 4.1. **An earlier version of this table was run on the hybrid substrate** (the released configuration's default), which did not match the dense configuration the text described; v2 unifies them and raises the seed count from 5 to 20.

Ablation	Unused rate (90% CI)	Reused rate (90% CI)	Δ unused (pp)	Δ reused (pp)
Full controller	0.0000 [0.0000, 0.0000]	1.0000 [1.0000, 1.0000]	baseline	baseline
Naive (all disabled)	1.0000 [1.0000, 1.0000]	1.0000 [1.0000, 1.0000]	+100.00	+0.00
Disable decay	1.0000 [1.0000, 1.0000]	1.0000 [1.0000, 1.0000]	+100.00	+0.00
Disable salience grading	0.0000 [0.0000, 0.0000]	1.0000 [1.0000, 1.0000]	+0.00	+0.00
Disable reconsolidation	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	+0.00	-100.00
Disable dormancy	0.0000 [0.0000, 0.0000]	1.0000 [1.0000, 1.0000]	+0.00	+0.00

Table S8: Real workloads placed in the phase-diagram coordinate system

HashingEmbedder — synthetic reference scale: iid 0.0773 / clustered 0.6651 / drift 0.9454

Corpus	Cases	NN similarity	Nearest family	Overload	Cue noise	In SDM-viable region
locomo	10	0.7386	**clustered**	0.106×	0.3722	**0/10**
longmemeval	500	0.6660	**clustered**	0.186×	0.3650	**0/500**

BGE-m3 — synthetic reference scale: iid 0.0773 / clustered 0.6651 / drift 0.9454

Corpus	Cases	NN similarity	Nearest family	Overload	Cue noise	In SDM-viable region
locomo	10	0.6312	**clustered**	0.106×	0.3096	**0/10**
longmemeval	500	0.4291	**clustered**	0.186×	0.3176	**0/500**

Table S9: Recomputation of numbers quoted outside the main tables

(a) Retrieval-side figures for the pure-SDM arm (Section 3.1)

Arm	Post-feedback accuracy	Evidence recall	Retrieval latency (ms)	Rows
pure-SDM	0.4556	0.3708	249.65	1200
dense_hot	0.8472	0.8306	5.79	1200
Full MAM	0.8056	0.7944	62.62	1200

(b) Hard-address decorrelation (Section 3.1; clustered / N=1024, cap=3)

Address scheme	clean	flip15	Pairwise Hamming std	Min distance
random	0.0182	0.0128	16.0103	435
decorrelated	0.0156	0.0104	0.0000	512
data_adaptive	0.0159	0.0159	48.2672	124

Iterative-cleanup cap (Section 3.1; random addresses)

Family/condition	cap=3	cap=10	cap=50	Delta (50-3)	cap-3 truncation
iid/clean	0.9299	0.6445	0.6443	-28.57pp	87.37%
iid/flip15	0.1583	0.0828	0.0828	-7.55pp	92.68%
clustered/clean	0.0182	0.0120	0.0104	-0.78pp	95.36%
clustered/flip15	0.0128	0.0078	0.0062	-0.65pp	99.51%

(c) Field-by-field comparison on the held-out matrix (Section 4.1; deterministic policy, no LLM)

Scenario	Rows	Post-feedback	Field diffs	Post-feedback correct (dense_hot / naive)
capacity-growth	432	288	0	288 / 288
cross-agent-relay	144	96	0	60 / 60
delayed-feedback	192	48	0	48 / 48
partial-cue-budget-curve	432	288	0	240 / 240

Table S10: Closed-loop substrate confirmation on 1,280 frozen synthetic streams
Each arm follows its own retrieval-to-reconsolidation trajectory. Differences are arm minus matched dense control in percentage points. The upper bound is the one-sided 95% stream-bootstrap bound; the fixed criterion is upper bound $< +2$ pp.

Reader	Contrast	Difference	One-sided 95% upper	Excludes +2 pp
Qwen	hybrid_full − dense_full	−14.857	−14.310	yes
Qwen	sdm_full − dense_full	−58.210	−57.897	yes
Qwen	sdm_no_recon − dense_no_recon	−51.823	−51.335	yes
Qwen	sdm_no_decay − dense_no_decay	−63.945	−63.659	yes
Qwen	sdm_append_only − dense_append_only	−55.514	−55.039	yes
Qwen	sdm_time_version − dense_time_version	−51.139	−50.566	yes
DeepSeek	hybrid_full − dense_full	−14.805	−14.251	yes
DeepSeek	sdm_full − dense_full	−58.060	−57.734	yes
DeepSeek	sdm_no_recon − dense_no_recon	−51.667	−51.172	yes
DeepSeek	sdm_no_decay − dense_no_decay	−63.861	−63.568	yes
DeepSeek	sdm_append_only − dense_append_only	−55.286	−54.811	yes
DeepSeek	sdm_time_version − dense_time_version	−51.126	−50.553	yes

Dense/hybrid/pure-SDM full-controller accuracy is 92.949/78.092/34.740% for Qwen and 92.793/77.988/34.733% for DeepSeek. All 168,960 Qwen answers and 168,960 DeepSeek answers are present; 158 DeepSeek length terminations are retained and scored zero.

APPENDIX A — PROTOCOLS AND FREEZE RECORDS

A.1 The frozen held-out protocol. `sdm-feedback-heldout-v2` was frozen before execution (status field `frozen-before-heldout-run`), specifying held-out seeds {2711, 3917, 5521}, an excluded development seed (1701), a model contract (temperature 0, thinking disabled, 64 answer tokens), and a non-Cartesian experiment matrix of four experiments. Every result file records the protocol SHA-256, the runner script SHA-256, and the generator version. The protocol also records, in the `conditions` field, that a pure-SDM arm was *excluded* from the expensive held-out run "after adverse development result", a decision we later revisited (Appendix D).

A.2 Three-model replication and its limits. The same frozen matrix was executed on three answer-model families (111 cells and 4,944 calls each; 14,832 calls total, 0 failures). Row-level experiment keys are identical across all three, as are correct labels, feedback timing, source and target agents, `retrieval_relevant`, and `cold_searches`. Two caveats are recorded rather than smoothed over: (i) the three runs do **not** share a runner script SHA-256: the Qwen run predates the addition of the API-backed policies, which were appended for the cross-model replication (the diff touches only policy classes and concurrency, not question generation, which we verified by direct comparison); (ii) `evidence_relevant` agrees at 98.42% (Qwen/Codex), 98.30% (Qwen/Gemini), and 99.88% (Codex/Gemini) rather than exactly. We therefore treat **within-model paired mechanism contrasts** as the primary evidence and never interpret cross-model absolute differences as model-quality claims.

A.3 The synthetic phase-diagram protocol. `substrate-crossover-v3`, paper profile: dim 1024, 256 hard addresses, top-k 12, loads $\{0.125, 0.25, 0.5, 1, 2, 4\}\times$, 20 seeds, three population families, five cue conditions, matched cell/row/column lesions, and a pre-registered equivalence rule (paired seed-level 90% bootstrap CI wholly inside ± 2 pp) with `dense_reservoir` designated in advance as the primary control.

A.4 Default-off discipline. Every flag introduced for this paper (`naive_dynamics`, `dynamics_ablation`, the figure/table builders) defaults to the pre-existing behaviour, and we verify bit-exactness of the default path before use. For `dynamics_ablation` the gate is a full re-run of the 730-day fixture with the flag unset, compared field-by-field against the pre-change artifact.

A.5 The HippoRAG routing replication. Design, arms, and four predictions frozen at commit `64ea52b`; amendment 1 (reader token cap, deterministic answer parsing, three repository patches, diffs shipped) at `e71bb03`; the statistics script at `3782332`; MuSiQue was settled before any 2Wiki execution, and the recalibrated 2Wiki gate was registered at `9a779d8` before it ran. One quality gate missed its band: the full arm's MuSiQue recall at five retrieved passages (R@5) of 59.19 fell 0.81 below the pre-registered [60, 85], a range set against published absolute levels under a stronger embedder. The contrast the study is about, full minus dense, reproduced the published gap, and all arms share the embedder; we proceeded on that basis and record the miss here. The recalibrated gate (contrast lower bound above zero; sanity floor R@5 above 45) passed at 84.10. Per-type contrasts (Section 5): 2Wiki compositional/inference +18.26 pp (CI [+14.71, +21.86]) vs. comparison +0.62 (CI spans zero); MuSiQue 2-hop +8.71 vs. 3–4-hop -1.36, difference -10.07 (CI $[-14.60, -5.21]$). Out-of-memory remediation between datasets (embedding batch size, device selection) touched no statistical procedure.

A.6 Reference implementation and release scope. MAM provides the interfaces used in the controlled interventions. The supplement includes its 16-module public-API import closure, evaluation helpers, and an optional two-field compatibility patch. Public code and result identifiers are normalized to MAM; `MANUSCRIPT-NAMING.md` and the normalization ledger document this editorial change. Current release checksums differ from historical anchors where identifiers changed. CPU checks cover the patched interfaces; numerical results and privately retained originals are unchanged. Trajectories and checkpoints cover all 4,800 sweep cells, and cached binary inputs support multimodal placement replay. Large float embeddings and external assets are excluded under the archive limit. The runbook distinguishes saved-result replay from fresh model execution. A *rule* is a natural-language lesson written to the hot layer with a salience and a context string (encoding specificity: the address key hashes situation plus lesson). One translated zombie-fixture pair is: old, salience 0.70, day 0, "Lesson: tenant Gipimoz's full-backup window must be set to 04:00"; new, salience 0.90, day 90, same context, "Correction: the window is now 11:00; the old value is void."

Tenant names are synthetic; the twelve pairs span twelve operational topics. The released defaults are `recall_mode="hybrid"` and `reconsolidate_top_k=0`; individual experiments state their overrides. Controller constants as shipped: τ = 30 days; the salience-to-half-life interpolation is log-linear from 7 days (salience 0) to 15 years (salience 1); decay mode, reconsolidation breadth, dormancy threshold, and replay window are exactly the config fields the Section 4 ablations exercise. The encoder behind the Section 3.3 placement is the system's default `HDEncoder`, an SRP-LSH binarizer (sign of random Gaussian projections); "default" means the encoder the code ships with, not a claim of commercial deployment.

A.7 The second system's representation-budget sweep. Frozen at commit `2c984e0`, nineteen minutes before the settlement commit `caaf38d` and with no result files in its diff: arms S0 (float32) through S3 (byte-matched binary) plus two exploratory arms, the projection drawn once from `default_rng(4242)` at shape 32768×1024, the statistics driver imported unchanged from the frozen HippoRAG script, quality gate G1 (the float32 control must reproduce the archived dense arm exactly), four numbered predictions with stated confidences, and a deviation policy. One deviation was logged and is reported in Appendix L.

A.8 Closed-loop synthetic confirmation. Before outcome inspection we froze generator `synthetic-evolution-v1`, stream IDs 0–1279, 11 arms, 12 queries per stream, exact-match scoring, two reader families, 10,000 stream bootstraps (seed 20260915), and the family rule that all 12 one-sided 95% upper bounds must lie below +2 points. Each stream contains four entities, 20 writes, four checkpoints through day 730, superseding updates, and current, historical, stable, and abstention queries. The audit covers all 337,920 terminal answers and verifies source, request, first-terminal receipt, model identity, failure retention, and matrix completeness. DeepSeek execution required five disclosed infrastructure recoveries involving TLS failures, quota-period HTTP 403 responses, and a final 8–16–32 concurrency ramp. The original six-attempt execution gate therefore failed; existing terminal answers were never resampled, and the model, payload, sample, scoring, hypotheses, and sample size remained fixed. Artifacts are in `supplementary/artifacts/confirmatory-1280/`.

APPENDIX B — ABLATION INVENTORY, LEAKAGE AUDIT, AND LAYER-WISE OVERVIEW

Cross-agent leakage audit (Section 5). Across 12 cross-agent episodes and 37 participating turns, zero secret tokens, protocol codes, or lesson bodies reach hippocampal persistent state, while the same scanner recovers all 37 from the global archive and 12 of 12 episodes reconstruct their participants from handles alone. One boundary: the index records the first 20 characters of a query cue, so a caller placing a secret in a cue prefix would leak it; hashing that prefix is required. Artifact: `results-hippocampus-audit-v1.json`.

Figure B1 summarizes the layer-wise attribution; Table B1 lists the arms behind each of its panels.

Table B1: Ablation inventory and default settings.

Layer	Ablation	Where	Default
Substrate	SDM top-k / SDM fixed-radius / dense × 4 eviction policies	Section 3.1	—
Substrate	dense representation: float32 / float16 / int8 / binary	Section 3.2	float32 (frozen)
Substrate	iteration cap 3 / 10 / 50	Section 3.1	3
Substrate	address scheme: random / decorrelated / data-adaptive	Section 3.1	random
Substrate	byte budget: equal-substrate / equal-operational	Section 3.1, Section 3.2	equal-operational
Dynamics	full controller vs. `naive_dynamics`	Section 4.1–4.2	off
Dynamics	disable decay / salience grading / reconsolidation / dormancy	Section 4.2	off
Dynamics	reconsolidation breadth: all hits / top-1 only, ± write interference	Section 4.2	all hits
Routing	with / without hippocampal index	Section 5	with
Encoder	HashingEmbedder / BGE-m3	Section 3.3	hashing

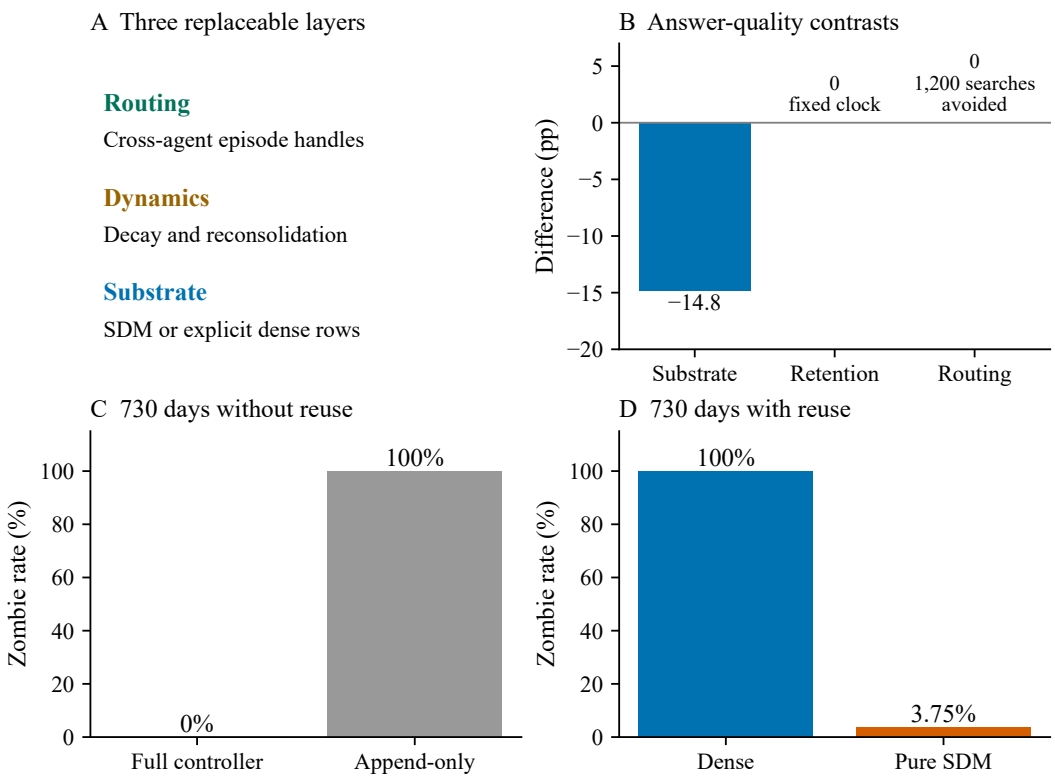


Figure B1: Three replaceable layers and their measured effects. Panel B reports the closed-loop hybrid–dense substrate contrast (mean over two readers), the static-fixture retention contrast, and routing's answer-quality contrast against cold search. Panels C and D are mechanistic stress tests on a separate fixture and use zombie rate rather than answer accuracy. pp: percentage points.

APPENDIX C — REPRODUCTION MANIFEST

The released package contains protocols, saved results, analysis and figure scripts, and the frozen MAM core. Its root `MANIFEST-SHA256.txt` records file digests; the earlier `APPENDIX-D-manifest.md` documents the historical analysis bundle. These manifests establish byte integrity relative to their recorded digests, not the timing of experiments or completeness of the implementation. Remaining dependencies and core compatibility conditions are disclosed in Appendix A.6. The HippoRAG replication ships alongside: six arm-level result files, the frozen statistics script, the run harness with repository patch diffs, and the pre-registration with both amendments.

APPENDIX D — POST-HOC ANALYSES

Running the grid on all three substrates fills a cell left empty so far — the substrate's effect on the *long* horizon. The no-reuse arms nearly coincide; the **reuse** arms separate (Table D1):

Table D1: Zombie rates by substrate under the 730-day fixture (20 seeds).

Substrate (full controller, 20 seeds)	No-reuse	Reuse	Pairs at ceiling
Dense	0.00%	**100.00%**	12.00 / 12
Hybrid (dense routing + SDM verification)	1.25%	30.83%	4.45 / 12
SDM only	0.00%	**3.75%**	2.25 / 12

This changes the interpretation of the zombie endpoint, rather than reversing the answer-quality result in Section 3. Once base weights and saliences jointly saturate and reset times agree, the pair has no differential plasticity signal; ranking then depends on whatever else the substrate can discriminate; dense has nothing left, since two same-context rules are near-indistinguishable in semantic space. SDM lowers the zombie rate for a different reason: on reuse probes its top-1 is an unrelated rule 73.75% of the time and the correct rule only 22.50% (dense 0.00%, hybrid 37.92%), and it reaches the ceiling on 2.25 of 12 pairs. It usually avoids the saturated comparison by failing to retrieve the pair. Restricting reconsolidation removes the pathology entirely (dense 0.00%; Section 4.2), and this paired-rule stress test is distinct from the readout and closed-loop experiments in Section 3. The result is therefore narrow evidence that the substrate changes long-horizon behavior on a correlated, language-like distribution.

Appendix A documents five protocol freezes: the held-out matrix (A.1), synthetic phase diagram (A.3), HippoRAG routing replication (A.5), second-system representation-budget sweep (A.7), and closed-loop 1,280-stream confirmation (A.8). Additional registered predictions, campaigns, and amendments are identified with their results below and in Appendix N. The following earlier analyses were designed post hoc:

1. **The earlier held-out pure-SDM arm (Section 3.1).** The frozen protocol explicitly excluded it; we added it after review feedback identified the gap between "SDM verification adds nothing'' and "SDM as a substrate is unnecessary". Question matrix, row keys, and per-item ground truth are identical to the pre-registered arms, which we verified rather than assumed.
2. **The earlier held-out one-sided bound (Section 3.1).** The pre-registered statistic was a two-sided TOST at ± 2 pp, which fails. We report that failure and adopt a one-sided bound on the argument that the proposition under test is one-sided. For that earlier matrix alone, a reader who rejects this reframing should read the result as "no evidence of benefit, with an MDE of $\approx$4 pp" rather than as an equivalence claim.
3. **The representation-budget sweep (Section 3.2) and the canonical-readout study (Section 3.1).** Both were designed after the phase-diagram result, to stress-test the exception cell and to give SDM its most favourable readout. Both changed our conclusions: the first reversed the selected exception with compact dense controls, and the second showed the frozen iteration cap was best among the tested caps in those cells.
4. **The entire 730-day zombie fixture (Section 4.2).** It has no freeze record. It was built to probe a failure mode we had already observed, its salience values (0.9 new / 0.7 old)

were chosen by us, and both the four-factor ablation (Section 4.2) and the substrate × long-horizon grid (Section 4.2) were added after the main results, in response to review feedback on an earlier version of this analysis; the Section 4.2 grid exists because that feedback identified the substrate claim as untested on the long horizon, and its results show why a lower zombie rate need not indicate better retrieval (Section 4.2).

5. **The natural-language workload placement (Section 3.3) and the third-party replication (Section 4.1).** Both post-date the substrate result and were designed to test whether it generalizes.
6. **The formal statements (Appendix K).** The controller analysis was developed after the measurements it organizes. Archived numerical checks test finite state identities; the revised general statements require the assumptions and proofs in Appendix K. Neither those checks nor the proofs replace measurements of deployment effects.

Registration status is campaign-specific. Appendix A documents the earlier held-out protocol, phase diagram, routing replication, representation sweep, and the subsequent 1,280-stream confirmation. The long-horizon stress fixture and other analyses listed here are exploratory. Appendix N identifies additional protocol-frozen external-system campaigns; its MemoryBank results remain post hoc.

Figure 3(a) summarizes the representation sweep of Section 3.2; Figure D1 shows every cell under both budget conventions.

Sensitivity sweep for the zombie fixture (Section 4.2). Because the fixture's parameters were chosen by us (item 4 above), we swept them: salience gap $\in \{-0.2, 0, +0.1, +0.2, +0.4\}$ (including an inverted arm), contradiction day $\in \{45, 90, 180, 365\}$, rule pairs $\in \{6, 12, 24\}$: 60 cells × 20 seeds × three configurations on the dense substrate (3,600 runs, `zdense/`). The two Table S7 claims are not knife-edge: *disabling decay drives the no-reuse arm to 1.0000 in 60 of 60 cells*, and *the full controller's reuse-arm pathology holds at ≥ 0.95 in 59 of 60 cells* (exactly 1.0000 in 47; lowest 0.91 at gap +0.4, day 365, 24 pairs). The attribution of the pathology to reconsolidation holds throughout the gap ≥ +0.1 region — but **not below it**: at gap < 0, stale rules resurface at 100% *with reconsolidation disabled and on the no-reuse arm alike*, because salience-graded decay itself preserves a rule more salient than its correction; at gap = 0 the measured rate is 47.2%, structured by contradiction day. Equal exponents do not make finite-time weights equal: the newer trace remains heavier, while the weight ratio tends to one (Corollary 1). Content scores, score saturation, and eligibility determine whether that advantage changes the retrieved rule. The sweep therefore motivates an explicit supersession mechanism; it does not prove that every lower-salience update must fail under decay. Section 4.2's claims are scoped accordingly. An auxiliary sweep of the same grid on the *hybrid* substrate (1,200 runs, `zsweep/`) shows the reuse-arm rate there falls from 0.71 (6 pairs) through 0.26 (12) to 0.04 (24), consistent with Section 4.2's substrate-dependence.

Re-clocked knowledge-update replay (Section 4.1; Figure 3(b)). The carve-out was verified constructively on real data (`lme_ku_reclock.py`; `results-lme-ku-reclock.json`; predictions pre-registered before execution). 75 of LongMemEval's 78 knowledge-update questions carry evidence on two distinct in-corpus dates (median evidence span 40.5 days; the surrounding haystack spans a median 70.5): the superseded value on the earlier date, the current one on the later. Each session enters as one hot trace at uniform salience, so weight differences can come only from timestamps. Under the wall clock ($\Delta t \approx 0$), the full controller and append-only storage produce identical update-win verdicts on every question (44/78 each, zero discordant pairs): the Section 4.1 indistinguishability, on a public benchmark. Re-clocked to the corpus's own timestamps, the full controller wins sixteen additional questions and loses none (60/78; b/c = 16/0; exact McNemar $p = 3.1 \times 10^{-5}$): decay demotes the staler superseded session (old-evidence@5 falls 66 → 50) without touching the current one (new-evidence@5: 74 → 75). Append-only is clock-invariant, as it must be. The prescription follows: an immediate-QA benchmark becomes able to credit retention dynamics the moment its retention clock is bound to in-corpus time.

Generative Agents recency probe (Section 4.1). A third-system characterization, on its released memory streams (`ga_recency_probe.py`; `results-ga-recency-probe.json`; scoring replicated line-for-line from its `retrieve.py`, constants from each persona's `scratch.json`). Generative Agents scores memories by a weighted sum of recency, relevance and importance with global weights [0.5, 3, 2], recency the smallest, its recency term ordinal in in-corpus game time (d^{rank} by last access, $d \in \{0.99, 0.995\}$ per persona). Removing that term changes

28–32% of top-30 retrievals (Jaccard 0.69–0.72), 17–19% of top-10, and flips the top-1 memory on 6–24% of focal points across the three shipped personas (50 seeded focal points each; zero LLM calls). This is a characterization, not a criticism: the system's temporal mechanism is live on its own data, which is precisely why an evaluation that freezes the clock measures a different system.

Second-system substrate swap (Section 3.3). On MemoryBank's fixture, with its corpus, embedder (all-MiniLM-L6-v2), per-user candidate pools, and gold-date hit@k metric unchanged (`mb_substrate.py`; `results-mb-substrate-v1.json`): the native float32 readout reproduces the shipped hit@2 = 0.7714 bit-exactly, validating the harness; a binarized control (1,024-bit SRP-LSH rows, Hamming ranking) reaches 0.7414 $\pm$ 0.031 at one-twelfth the bytes; SDM readout over the same hypervectors reaches 0.1714 $\pm$ 0.057 at matched operational bytes (median 42 locations, int8 counters) and 0.1857 $\pm$ 0.060 at a fixed 256 locations, a $\approx$6$\times$ byte advantage, disclosed as such. The corpus lies outside the empirically defined favorable region on the correlation and cue-noise axes; the favoured arm is also *under*-loaded (0.15$\times$). This is consistent with the workload locator, but the swap does not independently vary those axes or identify their separate causal contributions.

Overload extension beyond the pre-registered grid (Section 3.2). The pre-registered phase diagram stops at 4$\times$ overload, so where SDM eventually wins was untested. We ran an iid extension at 16$\times$/32$\times$/64$\times$/128$\times$ overload, 10 seeds, both byte budgets, all four dense representations (`results-substrate-exception-showdown-overload.json`). Under **equal-operational** budgets SDM never wins: it trails binary dense by 59.01, 62.50, 65.36 and 66.98 pp at the four loads. Under **equal-substrate** budgets it beats float32 dense throughout (+34.27 / +34.01 / +32.55 / +32.19 pp) but still trails binary at 64$\times$ ($-$15.05 pp) and first wins only at **128**$\times$, by +7.08 pp. This qualifies the favorable region by both representation and budget definition: within this extension, SDM first beats binary dense at 128$\times$ under the equal-substrate convention, but never under the equal-operational convention.

APPENDIX E — CANONICAL READOUT FOR SDM (DETAIL FOR SECTION 3.1)

Tuning. SDM's activation width was tuned on five training seeds (k=8); on ten held-out validation seeds the frozen arm reaches **0.1803** against **0.7811** for the equal-byte dense control. The phase-diagram grid froze earlier at k=12 (Appendix A.3), a mismatch running against SDM by 0.29 pp, two orders below the 60 pp the tuned arm concedes.

Kanerva's own readout. Neither of the tested cleanup and address interventions rescues it (Appendix E, Table S5): raising the cleanup cap makes accuracy fall *monotonically* in every cell, so the frozen cap of 3 is best among the tested caps in those cells, and decorrelation-optimal Sylvester–Hadamard addresses move clustered accuracy by $\leq$0.3 pp, intervals spanning zero. These controls implicate load and capacity under the tested settings; they do not exclude every contribution of correlation.

Implementation. KanervaSDM 1.0.1, an independent C++ fixed-radius implementation, reproduces **both** sides across 72 cells: SDM wins 4, loses 63, the wins confined to low-correlation, overloaded, clean-cue cells.

A reader who knows Kanerva (1988; 1992) will ask whether SDM was run with its canonical readout. Two aspects deserve scrutiny: iterative cleanup, and hard-address decorrelation (Table S5).

Iterative cleanup. Our readout already iterates with a convergence test, but the cap is 3, and at that cap **93.73%** of queries are truncated before converging, which sounds like a handicap. Raising the cap reverses the intuition: accuracy falls **monotonically** in every cell. At iid/clean, 0.9299 (cap 3) $\rightarrow$ 0.6445 (cap 10) $\rightarrow$ 0.6443 (cap 50), a drop of 28.56 pp (Table S5 reports $-$28.57 pp from the full-precision values); the three other cells fall by 0.65–7.55 pp. Under this load, cleanup does not pull the state toward the stored trace but toward crosstalk-dominated spurious attractors, and early stopping protects SDM. **The frozen cap of 3 is thus SDM's per-cell optimum among $\{3, 10, 50\}$, not a limitation imposed on it.** A load sweep from 0.125$\times$ to 4$\times$ confirms no cell improves with a higher cap.

The convergence diagnostics also supply a mechanism for the clustered collapse: at cap 50, iid converges on 100% of queries in 5.5 iterations on average, while **clustered converges on only 62.2–**

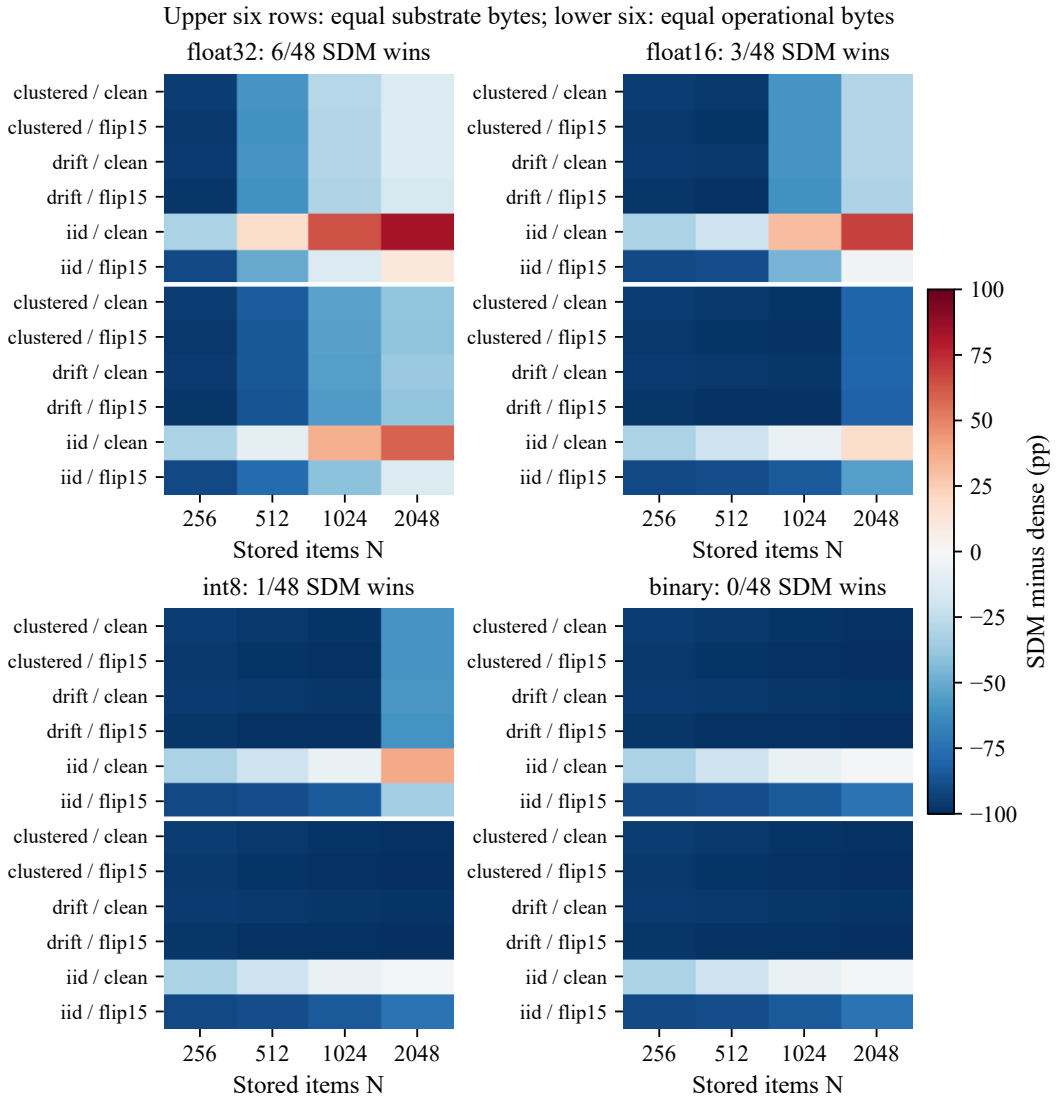


Figure D1: Full representation sweep underlying Figure 3(a). Each representation panel contains 48 cells: the upper six rows use equal substrate bytes and the lower six use equal operational bytes. Color denotes the SDM-minus-dense readout difference in percentage points.

64.1% and 26.5–30.5% of queries enter limit cycles — roughly a third of clustered queries enter a cycle rather than converge in the tested readout.

Address decorrelation. Kanerva (1992, Sec. 8) describes how clustered data can waste uniformly distributed hard locations. As a geometric control, we construct decorrelated addresses from Sylvester–Hadamard rows, achieving the theoretical optimum: pairwise Hamming distance standard deviation falls from **16.01 to exactly 0.000** and the minimum pairwise distance rises from **435 to 512**, i.e. all 32,640 pairs are exactly equidistant at $\dim/2$ (the Plotkin bound for 256 codewords of length 1024 is 514, so this is within 2 bits of provably optimal). It does not help: on clustered/1024 accuracy moves from 0.0182 to 0.0156 (clean) and 0.0128 to 0.0104 (flip15), both changes within noise and both in the wrong direction. A data-adaptive scheme that deliberately places addresses on the cluster structure does improve SDM at every load tested, but never enough to matter — dense is at 1.0000 in all twelve cells — and most at $0.125\times$ load (+19.06 pp), where dense is already at 1.0000, and the effect vanishes by $0.5\times$ load.

The tested address and cleanup alternatives do not rescue SDM in these cells. This makes the result more robust to those implementation choices and is consistent with a load/capacity limitation. It does not exhaust every way in which correlation, addressing, or readout could affect other SDM designs.

Scope note. The external KanervaSDM check (Section 3.1, "Implementation") uses radius activation with single-shot readout, whereas this appendix studies our top-k family with iteration. The two agree in direction on clustered populations but their numbers are not interchangeable.

APPENDIX F — MEMORYBANK FORENSICS (DETAIL FOR SECTION 4.1)

On MemoryBank's own fixture (566 memories, 15 users, 35 date-anchored questions; the system of Section 4.1), with embedder, candidate pools, and metric unchanged and only the readout replaced: dense retrieval reproduces the shipped hit@2 of 0.7714 exactly (hit@k: at least one gold item among the top k results); a binarized-dense control reaches 0.7414 ± 0.031 at one-twelfth the bytes; SDM readout collapses to 0.1714 ± 0.057 at matched bytes and to $\mathbf{0.1857 \pm 0.060}$ **with a** $\mathbf{6\times}$ **byte advantage** (256 locations; 20 seeds; Appendix D). The collapse lands where Section 3.3's diagram places this corpus — correlated text under noisy cues, outside the viable region on two axes — on a system we did not build. Separate dense-store experiments in HippoRAG 2 and a multimodal knowledge graph (KG) test representation and budget sensitivity (Appendices L and N). The workload locator also places 0 of 1,510 image–text points inside the empirically defined SDM-favorable region (Appendix M); this geometric placement is distinct from answer-level validation.

A reasonable objection is that Section 4.1 and Section 4.1 describe a quirk of our own implementation. We therefore examined **MemoryBank** (Zhong et al., 2024), an independent system that presents an Ebbinghaus-motivated memory updater as one of its three central pillars, alongside a memory store and a retriever. It reaches the same endpoint by an entirely different route, which makes the replication more informative than a matching score would have been.

The forgetting curve is disabled in every shipped configuration. The flag defaults to off (`enable_forget_mechanism: bool = field(default=False)`); of the four launch scripts in the repository, three never pass it and one passes `–enable_forget_mechanism False` explicitly. With the flag off, retrieval is pure vector similarity with no date, weight, or decay term — i.e. exactly our `naive_dynamics` arm.

When enabled, retention changes candidate availability. The curve is evaluated during index construction and consumed as a one-shot Bernoulli delete (`if random.random() > retention_probability`). Surviving memories retain their similarity-based order, but removing candidates can change the top-k set. This differs from continuous score weighting and does not imply monotonic degradation (Proposition 3, Appendix K).

At $\Delta t = 0$ the mechanism is a no-op. Retention is $\exp(0) = 1$, so the deletion test is false for every item. On the examined English fixture (566 memories, 15 users, 35 date-anchored questions, 20 seeds, top-2 retrieval), forgetting-on at zero time matches forgetting-off: hit@2 is 0.7714 and the retrieval signatures agree across all seeds.

The observed nonzero-time settings reduce retrieval quality. At the fixture's last day, deletion removes 55.9% of memories and reduces hit@2 by 14.9 pp. At the shipped wall-clock default, it removes 100% and costs all 77.1 pp. These measurements apply to the tested dates and release; they do not establish that every deletion policy or query date is harmful. Whether retention discriminates useful from obsolete evidence depends on ages, strengths, relevance, and the update semantics.

Scope. This is a retrieval-level experiment on MemoryBank's fixture, not an answer-quality experiment on LoCoMo or LongMemEval. No direction of downstream answer change is inferred from candidate deletion alone. The zero-time identity transfers only when the corresponding clock and update conditions hold. The release's disabled forgetting flag is an implementation fact, not evidence of the designers' motives.

Two incidental defects, reported for reproducibility. The curve is written `math.exp(-t / 5*S)`, which Python evaluates as $\exp(-tS/5)$ rather than the documented $\exp(-t/5S)$; the docstring states that higher memory strength forgets more slowly, while the code makes it forget exponentially faster — and since recall increments strength, reinforcement accelerates deletion of the memories just used. Separately, four of fifteen user keys in the English fixture carry stray whitespace absent from the probing-question file, so a naive join silently drops 11 of 35 anchored questions; we repaired the join before computing the numbers above.

APPENDIX G — ENCODER ROBUSTNESS FOR THE WORKLOAD PLACEMENT (DETAIL FOR SECTION 3.3)

We next ask whether the tested public corpora resemble the synthetic conditions with favorable SDM cells. MAM's sign random projection locality-sensitive hashing (SRP-LSH) encoder maps angle to expected Hamming distance as θ/π (Lemma 1, Eq. (K5)), permitting measurement of correlation, overload N/M, and effective cue noise in the same coordinates. This geometric bridge does not itself establish that these three statistics fully determine retrieval quality. Table S8 places both on the synthetic scale: real corpora sit far from the empirical SDM-favorable region on **all three axes**: clustered rather than iid, *under*-loaded at 0.11–0.19$\times$ vs. the locator's $\geq$4$\times$ band, cue noise 0.31–0.37. **Zero of 510 fall inside** under either encoder (Appendix G).

Two axes are corpus properties; the third, overload, is ours, inviting the small-M edge-deployment objection — right about N/M and immaterial, since either remaining axis alone excludes them (correlation 0.1273 vs 0.4016; clean-cue 0.05 vs 0.1992) and shrinking M moves points up, not left or down. The placement motivates a no-benefit prediction for these tested corpora, consistent with the separate retrieval measurements in Section 3.1.

The bridge is available because MAM's reference encoder is itself an SRP-LSH: `HDEncoder` projects a dense embedding into a d-dimensional **binary hypervector**, structurally identical to the synthetic patterns of Section 3.1. The phase diagram's coordinates can therefore be applied verbatim to real corpora:

Figure 4 plots the placement: real corpora against the shaded intersection of the three SDM-viable conditions, under both encoders.

- **correlation** — mean nearest-neighbour similarity, the same function used by the substrate router;
- **overload** — N/M, where N is the number of sessions written as hot traces;
- **effective cue noise** — normalized Hamming distance between a query hypervector and its gold session hypervector, directly readable as an equivalent bit-flip rate.

Table S8 places LoCoMo (10 conversations) and LongMemEval (500 cases) against the synthetic reference scale (iid 0.0773, clustered 0.6651, drift 0.9454). Real corpora sit far from the SDM-viable region on **all three axes simultaneously**: they are clustered rather than iid (0.43–0.74 across encoders, in every case nearer clustered than iid; Table S8); they are *under*-loaded at 0.11–0.19$\times$ rather than the $\geq$4$\times$ band of this empirical locator; and their effective cue noise of 0.31–0.37 is comparable to or worse than the harshest synthetic condition we tested (flip35), exceeding it under the hashing encoder. **Zero of 510 corpora fall inside the SDM-viable region**, under either encoder (Figure 4).

Two of the three axes are properties of the corpus; one is not, and that distinction matters. Correlation and effective cue noise are measured from the text itself: they do not change if we resize the store. The overload ratio N/M is different — M is ours to choose, so a reader can object that we manufactured the null by picking a large M, and that a fixed-capacity edge deployment with small M would push the same corpora into the SDM-viable region. The objection is correct about N/M and immaterial for the conclusion, because the region is an **intersection** and the other two axes exclude these corpora on their own. Of 510 corpora, the number satisfying the correlation condition is **0** and the number satisfying the clean-cue condition is **0**, under either encoder. The correlation threshold is 0.1273 (the iid reference of 0.0773 plus the 0.05 margin) and the lowest corpus sits at 0.4016 — 3.2× outside; the clean-cue threshold is 0.05 and the lowest corpus sits at 0.1992 — 4.0× outside. Shrinking M moves points upward on the vertical axis but cannot move them left or down. For an edge deployment, load must still be measured; smaller capacity alone does not place a workload inside this empirically defined region. Being outside it is a geometric classification, not a universal performance guarantee for other SDM implementations.

We replicate under two encoders precisely because a single-encoder result would be an artifact risk. The qualitative conclusion is identical; the quantitative difference is disclosed: BGE-m3 yields systematically lower nearest-neighbour similarity than character n-gram hashing (LME 0.4291 vs. 0.6660), as a semantic encoder discriminates same-topic documents that lexical overlap conflates. The BGE arm truncates sessions at 2048 tokens. The reported coordinates therefore also depend on this encoding configuration; they need not remain identical under longer contexts or different session segmentation.

The locator recommends dense storage for all 510 tested corpora under both encoders. This agrees in direction with the public-benchmark hot-store comparisons, which include the small LoCoMo any-gold exception reported in Appendix H. Combined with the equal-operational-byte sweep, these findings give no measured reason to activate the adaptive SDM router on these workloads. They do not establish that such routing is unnecessary in every deployment.

APPENDIX H — PUBLIC-BENCHMARK ABLATION (DETAIL FOR SECTION 3.1)

We use LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2025). We write each benchmark session as one identical hot trace and query the same in-memory state twice — once with hybrid SDM verification, once dense-only — reporting gold-evidence-session retrieval with no LLM in the loop (Table S2). On LoCoMo (1,982 questions with gold evidence) any-gold recall is 0.9319 (hybrid) vs. 0.9309 (dense), a +0.10 pp difference; the paired comparison is 27 hybrid-only wins against 25 dense-only wins with 1,930 ties. On LongMemEval (500 questions) dense is **higher**: 0.9740 vs. 0.9560, with 11 dense-only wins against 2. The remaining metrics point the same way and we report them rather than only the headline one: on LoCoMo dense is also better on all-gold recall (0.8446 vs. 0.8370) and MRR (0.7422 vs. 0.7380) despite the marginal any-gold deficit, and on LongMemEval it leads on every metric. Dense is faster on both (0.57 vs. 0.74 ms; 2.95 vs. 4.44 ms) and returns less context.

APPENDIX I — AUDITING THE CONSTRUCTIVE PROJECTION (DETAIL FOR SECTION 2)

The trajectory agreement in Figure 2's upper panel is a design property. It removes controller-state imbalance and identifies the substrate's controlled direct effect on readout for the fixed event schedule. It does not include effects mediated by a substrate changing later reconsolidation. Section 3.1 estimates that closed-loop quantity separately.

Three things about the construction are not free, and we state them so the design can be audited. **(1) Which feedback paths are severed.** In deployment the readout influences the controller — recalled traces are reconsolidated — so a substrate retrieving different traces would induce a different weight trajectory. The projection breaks exactly this path: the controller advances on a fixed, substrate-independent event schedule. This is what makes the comparison clean and is also a departure from deployment; Section 4 shows it is inconsequential for the immediate-QA fixture, where reconsolidation has no effect on ranking at all, and treats the long-horizon fixture separately. **(2) Byte matching is a modeling choice with consequences.** SDM's fixed substrate cost buys a specific number of

dense rows, and changing that exchange rate reverses the selected exception and removes SDM wins in the tested representation sweep (Section 3.2). We report both budget conventions because the comparison depends on them. **(3) What the active-trace check does and does not show.** The active-trace fraction agrees exactly across all 240 paired units. This is an implementation guard confirming that the single projected weight vector was in fact shared, not an independent falsifier: the quantity is computed once from that shared vector and copied into each arm's record, so no substrate-dependent effect could move it. The construction's falsifiable content is the lower panel — coincident controller state is compatible with a large, stable, directional gap in readout quality, and a substrate that closed that gap would refute the section's claim.

APPENDIX J — CLAIMS AND EVIDENCE

Key claims, with their evidence, location, and provenance: registration status (pre-registered or post hoc, per Appendix D), replication coverage, and measurement mode.

Table J1: Key claims, evidence, and provenance.

Claim	Evidence	Where	Provenance
Closed-loop SDM benefit is below +2 pp across readers and controller variants	1,280 streams; 337,920 answers; all 12 fixed upper bounds below +2 pp	Section 3.1, Table S10	frozen before outcomes; synthetic generator; independently audited
Earlier held-out matrix gives an exploratory upper bound of +3.78 pp	three answer models; $\leq$ +0.45 pp after excluding the unstable scenario	Section 3.1, Table S3	pre-registered matrix, post-hoc one-sided statistic (Appendix D, item 2)
Compact dense controls reverse the selected SDM exception	representation sweep: wins 6/3/1/0 by compression	Section 3.2	post hoc; this system; deterministic
MemoryBank substrate swap favors dense on its fixture	MemoryBank: dense 0.7714 bit-exact, SDM 0.17–0.19 at 6$\times$ bytes	Section 3.3	post hoc; second system; deterministic (n=35)
Tested dense fixture has output identity at a fixed retention clock	1,200 paired rows bit-identical	Section 4.1	post hoc; this system; deterministic
Same replacement decisive over 730 days	zombie fixture, 20 seeds, 0.00% vs 100.00%	Section 4.2	post hoc; this system; deterministic
Decay entire benefit; reconsolidation entire cost (bounded)	Table S7 + 3,600-run sweep; holds where updates **exceed** the salience of the superseded rule (gap $\geq$ +0.1)	Section 4.2, D	post hoc; this system; deterministic
Immediate-QA with a fixed retention clock cannot credit elapsed-time retention	constructive proof + third-party replication	Section 4.1–4.1	post hoc; second system; deterministic
Re-clocking to in-corpus time does credit dynamics	LME-KU n=78: wins 16/0, p = 3.1e-5; wall clock zero discordant	Section 4.1, D	pre-registered prediction; real data; deterministic
In-corpus clocks are live in deployed designs	GA probe: removing the recency term reshuffles 28–32% of top-30	App. D	post hoc; third system; deterministic
Examined MemoryBank release disables forgetting; tested nonzero-time settings lower hit@2	repo forensics + its own fixture, 20 seeds	Section 4.1, F	post hoc; second system; deterministic
Routing avoids cold search at answer parity	removal 27–75 pp; cold-search substitution arm reaches 1.000	Section 5	pre-registered arms (Appendix A.1), post-hoc framing; this system; deterministic
Routing carries a real benefit on a second system	HippoRAG 2: +3.86/+9.81 pp F1 vs. budget-matched dense; one-step arm retains most of it	Section 5	pre-registered; second system; deterministic
Zero cross-agent leakage (cue-prefix boundary noted)	9,943-window audit	Section 5, B	post hoc; this system; deterministic
0 of 510 corpora inside the SDM-viable region	placement, two encoders; region re-derived from independent impl.	Section 3.3, G	pre-registered region; independent implementation; deterministic
Lower zombie rates can coincide with unrelated retrievals	Section 4.2 grid: SDM 3.75% vs dense 100.00%	Section 4.2	post hoc; this system; deterministic
Formal conditions and limits	Controller identities, conditional separation, deletion counterexample, PPR score bound	App. K	post hoc; deterministic
HippoRAG store results depend on both representation and byte budget	byte-matched binary tracks float32 within a point; 1/32 bytes costs 4–9 pp R@5	App. L	pre-registered; second system; deterministic

Table J1 states registration status row by row. The closed-loop confirmation, held-out matrix, synthetic phase diagram, HippoRAG routing study, and second-system representation sweep were frozen before their outcomes (Appendix A); the one-sided reinterpretation of the earlier held-out

matrix remains post hoc. All remaining analyses are post hoc unless their provenance cell says otherwise.

APPENDIX K — FORMAL STATEMENTS AND PROOFS

The sweep's boundary is the theory's phase condition. With Eq. (1) and α strictly decreasing in salience,

$$\log \frac{w_{\text{old}}(t)}{w_{\text{new}}(t)} = \big(\alpha(s_{\text{new}}) - \alpha(s_{\text{old}})\big) \log(t/\tau) + o(1), \tag{K1}$$

Equation (K1) shows that the latent weight ratio tends to zero when $s_{\text{new}} > s_{\text{old}}$ and diverges in the opposite case. At equal salience, the newer trace retains a strictly higher weight at finite times, although the ratio tends to one. Score saturation, content similarity, tie-breaking, and dormancy determine whether these weight differences affect retrieval (Corollary 1; Appendix D). Explicit update semantics should not rely on a salience advantage alone.

Notation, from the shipped implementation (Appendix A.6): every trace carries a base weight w_0 (initial value 1.0), a salience $s \in [0, 1]$, and a decay exponent $\alpha(s) = \ln 2 / \ln(1 + h(s)/\tau)$ with $\tau = 30$ d and half-life $h(s)$ log-interpolated from 7 d at $s = 0$ to 15 y at $s = 1$; α is strictly decreasing in s. The instantaneous weight follows Eq. (1), with Δt measured from the last clock reset. Retrieval scores are content $\times\ \varphi(w)$, using Eq. (2). A recall reconsolidation applies $w_0 \mapsto \min(w_{\max},\, b \cdot \max(w_0, w))$ with $b = 1.5$, $w_{\max} = 4$, resets the decay clock, and applies the salience ratchet $s \mapsto \min(1, s + 0.05)$. Dormancy removes a trace from recall when $w < \theta_d = 0.05$. The original controller analysis followed the measurements it organizes (Appendix D, item 6). Archived numerical checks report saturation identities on 1,184,400 sweep rows, agreement of the weight curve with 795,144 checkpoints to four decimals, and a boundary tally matching 144 of 144 reconsolidation-off cells. These finite checks do not prove unrestricted ranking claims. The statements below distinguish exact state identities, sufficient ranking conditions, and limits on what follows from them.

Proposition 1 (projection identity). *Run two substrates S_1, S_2 against the same event schedule E with controller state $w(\cdot)$ computed once from E (readout → reconsolidation feedback severed). The controller trajectories are identical for every paired unit. The paired readout contrast therefore identifies the substrate's controlled direct effect for this fixed trajectory; it does not identify effects mediated by substrate-dependent future controller states.*

Proof. The coupling puts both arms on one probability space with one w-path. w is a deterministic function of E, which is shared; hence $w^{S_1}(t) = w^{S_2}(t)$ for all t, not merely in distribution. The paired difference $F(S_1, w) - F(S_2, w)$ is then a functional of (S_1, S_2) evaluated at a common w; any additive term defined as a function of the controller-state difference vanishes identically. □

Proposition 2 ($\Delta t = 0$ dense-readout identity). *Consider the dense-readout fixture with identical writes, candidate content, and tie-breaking in both arms. Let all traces start at $w_0 = 1$, let the retention clock remain fixed, and let controller state affect ranking only through $\varphi(w)$ and dormancy. With no interference or additional state-dependent candidate changes, the full controller and append-only storage return identical readouts on every query.*

Proof. At $\Delta t = 0$ the decay factor is $(1 + 0)^{-\alpha} = 1$, so $w = w_0$. Initially $w_0 = 1$; reconsolidation maps $w_0 \mapsto \min(4, 1.5\, w_0) \geq w_0$, so $w_0 \geq 1$ is invariant. On $[1, \infty)$, $\varphi \equiv 1$ and dormancy ($\theta_d = 0.05 < 1$) never fires. Scores in both arms therefore equal the content term for every trace at every query, and the dense readouts coincide exactly. This does not assert identity for an SDM operator that also consumes unsaturated weights. □

Theorem 1 (joint-state saturation). *Consider an old/current pair with initial base weights 1, saliences $s_o < s_n$, and synchronized co-recalls under the specified controller. Assume both traces remain eligible, with no intervening unequal reinforcement or interference. Base weights reach $w_{\max}$ after $k^* = \lceil \ln(w_{\max}) / \ln b \rceil = 4$ co-recalls. Saliences reach 1 after $\lceil (1 - s_i)/\delta \rceil$ co-recalls. After the maximum of these counts, both traces have equal base weights, exponents, and reset times. Their subsequent weight trajectories agree until an asymmetric update; with equal content scores, deterministic tie-breaking determines their ordering while both remain eligible.*

Proof. With nonnegative elapsed time and positive decay exponent, $w \leq w_0$, so each co-recall applies $w_0 \mapsto \min(w_{\max}, bw_0)$. The recursion in Eq. (3) reaches the cap after four recalls. The salience update adds δ until its cap. Equal capped saliences give equal exponents; the shared last co-recall gives equal reset times. Substitution into the weight function gives equality. At $s_o = 0.7$, $s_n = 0.9$, and $\delta = 0.05$, the salience counts are six and two, so four recalls alone do not guarantee equality of subsequent decay curves. Equal weights give equal score factors; content and tie-breaking then determine the ordering. □

Corollary 1 (latent weight ratio under decay alone). *Disable reconsolidation and interference, set equal initial base weights, and let the old/current write times be* 0 *and* $d_u > 0$*. For the weight trajectories before eligibility filtering,*

$$\log \frac{w_o(t)}{w_n(t)} = (\alpha(s_n) - \alpha(s_o)) \log(t/\tau) + o(1). \tag{K2}$$

Thus the ratio tends to zero for $s_n > s_o$*, diverges for* $s_n < s_o$*, and tends to one from below for equal salience. At equal salience,* $w_n(t) > w_o(t)$ *at every finite* $t \geq d_u$*.*

Proof. Subtract the logarithms of the two power laws. Both logarithmic age terms equal $\log(t/\tau) + o(1)$, giving Eq. (K2). Since α decreases with salience, its coefficient determines the unequal-salience limits. At equal exponent, the newer trace has smaller age and hence strictly larger weight; the age ratio tends to one. □

This is a statement about latent weights, not universal retrieval order. The score factor is flat above $w = 1$; content can differ; and dormancy may remove traces before an asymptotic ordering is observed. Consequently neither weight dominance nor an asymptotically vanishing gap alone establishes a finite-horizon zombie rate.

Theorem 2 (sufficient conditions for top-1 separation). *Suppose two eligible traces have equal, fixed, positive content scores. After a current-trace win under top-1 reconsolidation, let their states satisfy* $w_{0,n} \geq w_{0,o}$*,* $t_n \geq t_o$*, and* $\alpha_n \leq \alpha_o$*, where* t_i *is the last reset time. Assume positive exponents, no other writes, replay, or interference, and tie-breaking that favors the current trace. Then the current trace wins every subsequent query while either remains eligible. If the old trace is never reinforced, it becomes dormant at times satisfying*

$$t - t_o > \tau \left[(w_{0,o}/\theta_d)^{1/\alpha_o} - 1 \right]. \tag{K3}$$

Proof. For $t \geq t_n$, the current trace has at least as large a base weight, no greater age, and no larger exponent. Each comparison favors $w_n(t) \geq w_o(t)$. The nondecreasing score factor preserves weak dominance, with the stated tie rule deciding equality. Reinforcing only the winner increases its base weight up to the cap, resets its age, and does not increase its exponent; the three inequalities are preserved. Induction gives the subsequent wins. Solving the unreinforced old trace's decay inequality gives Eq. (K3). □

A single win alone does not establish these state inequalities or the tie condition. This theorem supplies sufficient conditions for a mitigation, not a general guarantee against stale winners.

Proposition 3 (candidate deletion is not monotone for hit@k). *Even with a fixed similarity ordering and independent Bernoulli retention, deleting candidates can either increase or decrease expected hit@*k*.*

Proof by counterexample. For $k = 1$, rank one non-gold item before one gold item. Without deletion, hit@1 is zero. If each item is retained independently with probability $p \in (0, 1)$, retaining only the gold item has probability $p(1-p) > 0$ and yields a hit. Conversely, a one-item gold corpus has hit@1 equal to one before deletion and expectation $p < 1$ afterward. Thus neither monotonic direction holds generally. □

MemoryBank's reported reductions are empirical findings for its measured configurations. A separate condition, such as preservation of every non-gold candidate while only gold candidates can be deleted, would be needed for the proposed non-increase argument; that condition is not assumed here.

Proposition 4 (PPR one-step envelope). *Personalized PageRank with damping* β *is the Neumann series* $p = (1 - \beta) \sum_{k \geq 0} \beta^k M^k r$*, where* M *is the column-stochastic transition matrix the harness*

builds and r a reset distribution. The deployed one-step arm applies the PPR operator once at $x = r$, giving $x_1 = (1 - \beta)r + \beta Mr$ — not the K=1 Neumann partial sum — and satisfies

$$\|p - x_1\|_1 \leq 2\beta^2. \tag{K4}$$

At the deployed $\beta = 0.5$ this bounds score-vector L_1 distance by 0.50. Without additional score-margin and answer-stability assumptions it supplies no bound on top-k changes, answer accuracy, F1, or the fraction of an observed answer gain.

Proof. For column-stochastic M and non-negative r of unit mass, $\|M^k r\|_1 = 1$. Writing $p - x_1 = -\beta^2 Mr + (1 - \beta)\sum_{k\geq 2} \beta^k M^k r$ and applying the triangle inequality gives $\|p - x_1\|_1 \leq \beta^2 + (1 - \beta)\sum_{k\geq 2} \beta^k = 2\beta^2$. (The K=1 Neumann partial sum obeys the tighter β^2, but the deployed arm does not compute it, so we quote the bound that applies.) The empirical answer contrasts are reported separately; they are not consequences or numerical validations of this score-vector bound. □

Lemma 1 (SRP-LSH bridge). *For unit vectors at angle θ, one signed random Gaussian projection flips with probability θ/π; over D independent projections the normalized Hamming distance concentrates as*

$$\Pr(|d_H/D - \theta/\pi| > \varepsilon) \leq 2e^{-2\varepsilon^2 D}. \tag{K5}$$

The encoder therefore carries cosine geometry to address geometry monotonically, with $O(D^{-1/2})$ fluctuation.

Proof. The flip probability is the classical hyperplane argument; bits are i.i.d. Bernoulli(θ/π) across projections, and Hoeffding gives Eq. (K5). Monotonicity follows since $\theta \mapsto \theta/\pi$ is increasing. □

APPENDIX L — REPRESENTATION BUDGET ON THE SECOND SYSTEM'S STORE

We repeat the representation sweep of Section 3.2 on the store carrying dense retrieval in the HippoRAG 2 testbed of Section 5: the float32 bge-m3 passage matrix behind the matched-budget dense arm, swapped in place with queries treated symmetrically, full-corpus re-ranking, and the campaign's paired protocol (doc-cluster bootstrap, 316/886 clusters; no LLM in the confirmatory loop; design and predictions pre-registered before execution, with one logged deviation — a query-projection bug in an exploratory arm, fixed before any interpretation). The float32 control reproduces the archived dense arm exactly (identical top-5 on 1000/1000 questions, both datasets). int8 at one quarter of the vector bytes produces small recall changes (MuSiQue ΔR@5 +0.17 pp, 95% CI [−0.04, +0.39]; 2Wiki +0.05 pp, [+0.00, +0.15]). A byte-matched binary store (32,768 SRP sign bits = 4,096 B per row, Hamming ranking) tracks float32 within a point: −0.61 pp ([−1.18, −0.05]) and +0.05 pp ([−0.33, +0.41]); the pre-registered −1 pp non-inferiority bound holds on 2Wiki and misses by 0.09 pp on MuSiQue. Compressing to 1,024 bits (1/32 of the bytes) costs real recall — −9.03 pp ([−10.51, −7.63]) and −4.15 pp ([−5.16, −3.14]), retention 0.83/0.94 — and on 2Wiki the retrieval loss propagates one-for-one to answers (F1 −4.12 pp, [−6.11, −2.19]). These results show that byte allocation materially changes the representation comparison. They do not establish universal format interchangeability: the matched-byte binary arm meets the registered non-inferiority criterion on 2Wiki but not MuSiQue. Scope, disclosed: the sweep covers the dense-retrieval store (the StandardRAG channel); the graph-internal fact and entity embeddings feed an LLM filter and PPR reset weights whose caches a swap would invalidate, so they were pre-registered as out of scope.

APPENDIX M — THE MULTIMODAL EXTENSION (DETAIL FOR SECTION 3.3, SECTION 4.1, SECTION 5)

A reviewer may object that every verdict above is a fact about text. We repeat all three layers on image–text data, on the system we built, under the protocols already stated. No layer required a new instrument.

Substrate: placement. The Section 3.3 coordinates are computed for image–text corpora — MSCOCO 2017 val (5,000 images / 25,014 captions) and a 5,000-image Flickr30k subsample

— under two vision–language encoders (CLIP ViT-B/32, SigLIP base), three memory scenarios (cross-modal: stored images cued by captions; intra-image: stored images cued by augmented re-observations; a BGE-m3 caption arm as text-consistency anchor), and four store sizes from the deployment-aligned 0.125× to a forced 4× overload — 1,510 corpus points, the pipeline verified bit-for-bit against Section 3.3's locator (12 archived coordinates reproduce at $d = 0$). **Zero of 1,510 fall inside the SDM-viable region**, under either encoder and under the independently re-derived region. The exclusion is stronger than for text: every point is nearer clustered than iid (correlation 0.37–0.53, the lowest 2.9× above threshold), and cross-modal cue noise is 0.39–0.47 — worse than the harshest synthetic condition and worse than the text spectrum's 0.31–0.37, a direct measurement of the cross-modal gap. Sixteen forced-overload intra-image points do enter the loosest bracket band, the zone the win-cell derivation left untested; a pre-registered readout probe on those points' own hypervectors settles it: dense scores 0.989–1.000 while SDM top-k collapses to $\leq$0.0012 across 20 address seeds.

Dynamics: a prediction, then the run. Theorem 1's controller reads no content, so the 730-day fixture was rebuilt with image rule pairs — two views of one MSCOCO scene (CLIP ViT-B/32; stale view day 0, $s = 0.70$; corrected view day 90, $s = 0.90$), cued by a held-out caption embedding through the production encoder seam. A prediction function frozen pre-run (closed-form decay plus the saturation collapse, fed only a pre-registered static content measurement, zero free parameters) fixed every verdict in advance and **matched 21,120 of 21,120 pair × checkpoint × arm cells**, with all saturation identities and the $\Delta t = 0$ identity bit-exact. Decay alone retires the stale view everywhere (0%, as in text). Under reuse the pair collapses to content tie-breaking, and the zombie rate equals the measured fraction of caption cues sitting nearer the stale view — **59.2%, not the text fixture's 100%**. That gap is the finding: text reaches 100% because its cue quotes the stale rule verbatim, pinning the tie; the near-tied image pairs instead yield the reported 59.2% under their measured content geometry. **In this paired-content construction, modality changes content-score comparisons while the controller equations remain fixed.** A nearest-neighbour control that breaks the near-tie premise leaves 23.3% of stale scenes retrieved under decay in the measured fixture, reinforcing the case for explicit supersession handling.

Routing: retrieval-level, judge-free. Section 5's operational verdict is extended to cross-agent *image* memories: 5,000 COCO images (CLIP and SigLIP embeddings as traces, one episode per image) distributed over four agents, queried by a fifth with 400 held-out caption embeddings per encoder; lexical channels are provably zeroed, so retrieval is embedding-only. On the deployed dense substrate the handle index matches direct archive search **query for query** (gold recall@1/@5 identical on all 400 pairs, zero discordant, both encoders) at **0 versus 400 cold searches per encoder** and lower latency; removing the index without a cold fallback collapses recall to 0.000. The hybrid arm loses 16.0–23.5 pp at recall@1 through SDM verification alone — reproducing the placement verdict by a third, deterministic instrument, and confirming routing is orthogonal to the substrate.

Scope: the placement is embedding geometry, not end-to-end QA; the dynamics and routing arms are our own system, with the multimodal second-system evidence in Appendix N.

APPENDIX N — FOUR SECOND SYSTEMS AT SUBSTITUTION LEVEL (DETAIL FOR SECTION 3.3, SECTION 4.1, SECTION 5)

HippoRAG 2, MG^2-RAG, and HippoMM include protocol-frozen campaigns scored without an LLM judge. MemoryBank is post hoc, and individual exploratory analyses are identified below and in the main intervention table. Deviations, including two quality gates that failed and were recorded rather than relaxed, ship with the artifacts.

HippoRAG 2 (text; routing, substrate). Its published gaps reproduce here: on the official MuSiQue-1000 and 2Wiki-1000 fixtures, with corpus, embedder, reader and deterministic EM/F1 shared and read exposure meeting the stated signed-median criterion, full HippoRAG beats the corresponding dense control by **+3.86 and +9.81 pp F1** (Section 5). The store obeys Section 3.2: int8 changes R@5 by +0.17 / +0.05 pp, a byte-matched binary store tracks float32 within a point (−0.61 / +0.05), and a 32× budget cut costs 4–9 pp, propagating one-for-one to answers on 2Wiki (F1 −4.12 pp). Its dynamics layer is anatomically empty: the released code contains no decay, consolidation, or retention clock.

Table N1: Interventions and evidence across systems. Each entry states its own resource constraint; these are not pooled as one lifecycle budget. KB denotes knowledge base.

System	Intervention, resource constraint, endpoint, and study status
MAM	SDM/hybrid versus dense; controller variants; index versus cold search. Operational bytes and shared writes/queries. Exact-match answers in the frozen 1,280-stream substrate confirmation; other arms have separately stated status. Authors' system.
MemoryBank	SDM versus dense; forgetting on versus off. Matched bytes or shipped candidate pool, as appropriate. Retrieval hit@2. Post-hoc study; source instrumentation.
HippoRAG 2	Full graph versus dense and one-step; graph rebuild versus native dense update. Stated read criterion or prospective shared ceiling; lifecycle cost measured separately. EM/F1 and resource costs. Frozen predictions and amended descriptive cost study; disclosed source/interface repairs.
MG^2-RAG	Full PPR versus dense and one step; store randomization. Read exposure controlled. Recall and answer scores. Pre-registered; KB and grounding repairs disclosed.
HippoMM	Separate substitutions for segmentation, summaries, filters, and gates. Segment, token, and frame budgets. Strict answer match. Pre-registered, with one exploratory filter result; source repairs disclosed.

MG^2-RAG (multimodal KG; routing, substrate). 1,200 InfoSeek questions over a 5,000-document multimodal KB, official scorer. The demo case reproduces bit-identically and the paper's reported ordering holds in our build; absolute accuracy sits below the published figure for reasons we chose, not defects we found — a 5k rather than 100k knowledge base and a different reader — and the arms share all of it. Full PPR versus budget-matched dense shows **no positive answer-layer evidence under the stated criterion** across three shared-query campaigns of increasing fidelity (+0.17, +1.17, **+0.42 pp**, one-sided LB -0.85 at full fidelity), while one-step and full PPR have identical per-query R@5 and R@10: **ΔR@5 = ΔR@10 = 0.00 with degenerate CIs**. Two findings are specific to this system. **56.3% of its store bytes are never read** — randomising all 53,739 entity embeddings leaves the output bit-identical — so deleting the store nobody reads dominates compressing the store everybody reads (0.00 versus -13.9 pp). And the graph arm retains more recall under this compression intervention: at 1/32 bytes, retention is 0.851 under PPR against 0.811 under matched-budget dense. Its dynamics layer is likewise anatomically empty.

On the fidelity of that reproduction. The first run covered 53.5% of KB images and disabled the licence-gated grounding model. Fetcher throttling and inadequate status logging contributed to the missing images; a revised single-stream fetch recovers 4,801/4,872, with 70 genuine 404 responses and one decode failure. Restoring grounding adds 1,940 sub-image vectors and 1,576 image–entity edges. Measured recall@5 and recall@10 remain unchanged. Answer differences coexist with repeat variation under invariant dense retrieval; this does not establish a reader-noise floor or an isolated grounding effect. Grounding changes 13.0% of top-10 orderings, with no detected directional imbalance (exploratory exact binomial $p = 0.503$) and **no gold document crossing the top-5 or top-10 boundary**. Object crops cover 19.45% of indexed images and are prompted by text-extracted entity strings, including abstract concepts. These results therefore apply to partial activation in this corpus. The earlier image-present subgroup *R@1* gain (+2.55 pp) is **withdrawn as robust positive evidence**: the grounded campaign gives +1.51 pp with LB -0.09, while subgroup membership grows from 785 to 1,189. This instability warrants caution about nonrandom image availability but does not identify selection as its unique cause. The corresponding grounded subgroup *answer-accuracy* contrast is +0.42 pp with LB -0.85.

HippoMM (multimodal video; all three layers). HippoVlog, 960 questions, 24 video clusters, strict binary matching. The system reproduces as published and then some: our full arm scores 81.98% against the reported 78.2%, and its headline removal-level increment comes back almost exactly (+6.98 pp against +7.1 pp). What follows questions the attribution of that increment, not its existence. Substitution then probes which of the named mechanisms improves on a simpler control. *Routing:* removing CA3 entirely costs +0.21 pp (CI $[-0.83, +1.35]$); removing CA1 costs -0.42 pp; the confidence gate never fires, because self-reported confidence collapses onto $\{0.9, 0.95, 1.0\}$, leaving the published $\tau = 0.75$ in a structural dead zone — deleting the gate *improves* accuracy (+1.15 pp). A single flat full-context call has a small observed difference from the whole system

(Δ = 0.31 pp, p = 0.86), consistent with a read-budget explanation but not establishing equivalence; the lean retrieval arm's 7 pp loss is audio-evidence starvation of merged cross-modal ranking (805 of 960 questions retrieve zero audio seeds), an embedding-space failure. *Substrate:* replacing the dentate-gyrus-style episodic segmentation with a uniform grid at matched segment count gives a small observed difference (-0.31 pp, $p = 0.74$); *seeded random* boundaries give +0.31 pp ($p = 0.70$). Both controls fail to detect a quality difference; these tests alone do not establish equivalence. This is mechanistically legible: the published segmenter already degenerates to a near-fixed 30 s grid (median 29.03 s, 30.5% at one value) and provably cannot reach the visual pathway (key frames bit-identical under any segmentation, Jaccard 1.000 on 24/24). *Dynamics:* replacing the consolidation summary with an extractive excerpt under the token-matching protocol yields +0.52 pp ($p = 0.56$), with 5.4% more realized reading. This is not proof of zero cost or exact exposure equality. Only consolidation's redundancy filter shows a positive point estimate against uniform subsampling at matched key-frame budget (+1.98 pp, CI [+0.31, +3.65], $p = 0.049$) — with a post-hoc MDE_{80} of 2.18 pp. The tightest exposure-matched subset gives +0.89 pp (CI [-0.55, +2.36]), leaving robustness to this restriction unresolved. The MDE alone does not negate the full-sample estimate. Every substitution changed 6.6–14.8% of detailed-path answers: the manipulations reached behavior even where aggregate accuracy differences were unresolved.

MemoryBank (text; substrate, dynamics). Reported in Section 3.3 and Section 4.1.

Reproducibility note. Two of these systems required source-level repair before their released code would run at all; in one case the batch path did not persist the audio side of the index. We record this as an obstacle we met, not as a claim about the work.

APPENDIX O — CHRONOLOGICAL ANSWER-LEVEL TRANSFER

Population and schedule. We use all ten simulated persona files in RealMemBench, repository revision `67afd0891d603adcc4458ff0449df306ef296b7a`. Sessions are sorted by calendar date, retaining original order within a day. Eligible user queries have a following reference response and nonempty gold source-session annotations contained in earlier written sessions. The first 50 eligible queries per persona in a fixed hash order yield 500 queries. Gold annotations and target/later responses are withheld from the adapter, encoder, and reader. Earlier observed dialogue remains public history. This is chronological replay with retrieval-induced reconsolidation, not a rollout in which generated answers alter future user interactions.

Interventions and endpoint. Hybrid-all is the reference hybrid readout with all returned hits reconsolidated. Dense-all changes only readout; hybrid-top1 changes only reinforcement; dense-top1 changes both. The timestamp-aware int8 arm is a practical compound baseline, not a quantization-only intervention. Each arm follows its own retrieval/update path. Qwen is the answerer; the independent judge is identified by the provider as a DeepSeek model family, whose exact weights were not independently verified. The primary endpoint averages a 0–3 rubric after division by three, equally weighting personas. It is neither exact-match accuracy nor human ground truth.

Table O1: RealMem scores and paired contrasts. Intervals are descriptive 95% paired-persona bootstrap intervals over ten personas (20,000 fixed-seed draws).

Method	Normalized rubric score (%)
Hybrid-all	29.07
Dense-all	31.93
Hybrid-top1	29.40
Dense-top1	31.40
Timestamp-aware int8	35.40

Contrast	Effect (pp)	95% interval
Dense-top1 − hybrid-all (primary)	+2.33	[−2.13, +7.07]
Dense-all − hybrid-all	+2.87	[−0.33, +6.60]
Hybrid-top1 − hybrid-all	+0.33	[−2.13, +3.13]
Dense-top1 − timestamp-aware int8	−4.00	[−9.13, +0.93]

What transfers. Every interval spans zero. The experiment establishes neither superiority nor equivalence of the combined intervention. This uncertainty limits transfer of the synthetic mechanism diagnosis; it does not reverse its conditional result. The ten simulated personas and selected query population do not represent independent real users.

Termination amendment and retained failures. The initial execution treated reaching the fixed 1,024-token output cap as a retryable failure. Before quality scoring, version 2 retained nonempty capped outputs without retrying them, reusing 338 completed answers and all existing vectors. Samples, prompts and output limits were unchanged; this was nevertheless an execution amendment after outputs existed. All 13 capped candidates and the superseded attempts are retained. Setting capped candidates to zero leaves the primary point estimate and interval unchanged. The original generator/selection, amendment, and scoring freezes are separate records.

Audit and accounting. There are 2,500 candidate answers, 2,288 distinct formal scoring payloads, and two scoring fixtures. The 2,290 external judge requests report 3,093,854 tokens. Currency charges were not verified. Reader generation separately used 14,432,989 local-model tokens including failures. Local replay verifies blinded-candidate joins, first-valid judgments, source/input hashes, all scores, and the frozen bootstrap; it makes no model calls. Full generator vectors and model weights are not part of the public saved-results replay.

APPENDIX P — PROSPECTIVE HIPPORAG MECHANISM PREDICTIONS

Freeze and population. Before formal extraction or answers, we froze 1,000 HotpotQA questions and 9,811 passages from the existing HippoRAG repository base revision `c617143f01477243992a63b2e2151cc003dd3b21`, including the vendored local patches. There are 811 bridge and 189 comparison questions. Connected components of shared supporting-document titles define 988 bootstrap clusters. A bounded project-history audit found no matching earlier question hashes; it does not establish absence from model pretraining or global novelty. Earlier MuSiQue/2Wiki results informed the predictions.

Mechanism and read ceiling. Full graph PPR and one-step propagation share the actual graph, fact filter, linking and normalized reset; the one-step operator uses damping 0.5 and the same dangling-mass and self-loop conventions. Dense uses the identical passage matrix with the upstream dense-ranking method. Full/dense compares the entire graph/filter/linking retrieval bundle. All use the same Qwen reader, CPU float32 BGE-m3 encoder, top-five passages and 6,000 evidence-token ceiling. Actual mean evidence tokens are 701.37/688.78/676.78 for full/dense/one-step. Common ceilings do not give identical exposure or equal lifecycle costs.

Table P1: Prospective HotpotQA outcomes and the three frozen directional criteria. Adjusted lower bounds use Bonferroni correction; descriptive intervals do not replace these tests.

Method	F1 (%)	EM (%)
Full graph PPR	68.14	56.4
Dense	60.95	50.0
One-step graph	67.16	55.9

Prespecified contrast	Effect (pp)	Adjusted lower	Decision
Full − dense, bridge	+9.24	+6.79	> 0: pass
Bridge gain − comparison gain	+10.86	+6.41	> 0: pass
One-step − full, all questions	-0.98	-2.38	> -2: not established

The primary bridge effect has descriptive 95% interval [6.98, 11.56] pp; the interaction interval is [6.81, 15.03]. Comparison alone is -1.62 pp [-5.07, +1.75], so harm is not established. The all-question full-minus-dense contrast is +7.19 pp [5.23, 9.17], a supplementary endpoint rather than a replacement for the bridge primary. One-step minus full has interval [-2.28, +0.29]; failing noninferiority does not prove that further propagation improves F1. All statistics use 20,000 draws at the frozen seed.

Activation and preserved repairs. Graph routing activates on 855 questions. Full/dense top-five order differs on 821 and set on 677; one-step/full order differs on 668. All extraction records are retained, including 66 empty entity/triple outputs. Formal answers have no truncation. Before freezing, served-model alias correction, complete string-array NER parsing, and a zero-weight diagnostic correction were tested without rerolling successful model outputs. Original failures and patched sources are preserved.

Evidence and cost boundary. The 3,000 saved answers use 2,330 distinct exact-input reader requests. Formal work totals 22,952 physical requests and 18,675,035 reported tokens, including 19,622 initial-extraction requests. Eight successful synthetic-interface requests add 5,700 tokens; four early HTTP-404 attempts have unknown usage. All use the existing local model service. Local audit verifies source/input/tokenizer hashes, exact prompts, first-valid outputs, official EM/F1, accounting, and bootstrap summaries. It does not rebuild the numerical graph locally. Appendix Q measures fresh-build lifecycle costs separately; its live filter calls are not the fixed-reset quality comparison here.

APPENDIX Q — REPLICATED LIFECYCLE MEASUREMENT AND APPROVED VALIDATION AMENDMENT

Design. Three repetitions each contain ten workers: separate graph and dense construction; three initial 1,000-query runs (full, dense, one-step); two maintenance runs; and three 50-query post-update probes. The 30 completed workers yield 9,450 saved answers/probes. Build and read order rotate according to the frozen manifest. Application indexes and response caches start empty, but OS, model and shared inference-service caches are not claimed cold. The same CPU float32 BGE-m3 encoder uses eight threads, batch eight and maximum length 512; the local reader service uses the protocol's Qwen weights. No external paid API is used in this study.

One physical graph build supplies the full and one-step alternatives. It is charged in full to each logical deployment, while physical requests are counted once. Dense constructs only its own StandardRAG store. Each read worker uses live filter/reader calls and executes only the selected propagation operator, including one-step transition preparation and resident bytes. These runs measure costs; they do not reproduce Appendix P's shared-filter/reset intervention. We do not assign that experiment's F1 to the fresh lifecycle runs.

Mixed maintenance policy. A fixed hash rule selects 20 inserted synthetic copies, ten replaced passages and ten deletions, leaving 9,821 passages. This is a synthetic stress batch, not an estimate of natural update frequency. The graph policy rebuilds the complete index with retained exact-input LLM responses and full re-encoding; dense uses native delete/insert on a task-owned copy. The copy cost is separate. The batch cannot identify per-operation insertion or deletion costs. Native graph fixture discrepancies are preserved; the measured rebuild policy is a conservative implementation choice, not a lower bound for all graph maintenance methods.

Table Q1: Lifecycle phase-wall time in three fixed repetitions. Every row includes its worker's timed setup and selected phases; process elapsed time and the separately measured physical dense-index copy are reported in the supplement. Hours and seconds are explicitly distinguished.

Stage	Repeat 1	Repeat 2	Repeat 3	Mean
Graph initial build (h)	3.30	3.32	3.30	3.31
Dense initial build (h)	0.45	0.45	0.45	0.45
Full: 1,000 initial reads (s)	1144.34	1178.38	1150.05	1157.59
Dense: 1,000 initial reads (s)	295.87	291.87	292.37	293.37
One-step: 1,000 initial reads (s)	1094.09	1105.17	1098.30	1099.19
Graph mixed maintenance (h)	4.33	4.16	3.57	4.02
Dense mixed maintenance (s)	9.41	9.65	11.86	10.31
Full: 50 post-update probes (s)	92.72	64.78	64.31	73.94
Dense: 50 post-update probes (s)	15.47	17.72	16.10	16.43
One-step: 50 post-update probes (s)	63.23	65.16	63.07	63.82

Interpretation. At every measured initial-query prefix (0, 1, 10, 50, 100, 250, 500, 1,000), full graph retrieval has greater cumulative phase-wall time than dense in all three repetitions. Initial construction averages 3.31 h for graph versus 0.45 h for dense; the corresponding 1,000-read phases average 1,157.59 s and 293.37 s. One-step averages 1,099.19 s and retains the entire graph construction cost. Under a constant-mean recurring-cost assumption, neither graph alternative has a finite wall-time break-even against dense because both its build and recurring reads cost more in these runs. This does not compare quality-adjusted utility or equal lifecycle budgets.

The graph rebuild maintenance policy averages 4.02 h versus 10.31 s for the dense native batch. The gap is specific to these policies and batch; it does not establish unavoidable graph-update cost. Initial active indexes occupy approximately 916–918 MB for graph and 44.45 MB for dense, excluding response caches. OS client-process peak RSS during construction is 13.77–13.86 GB for graph and 3.14–3.18 GB for dense. Full per-repeat client CPU, sampled/OS RSS, index and response-cache bytes, process elapsed times, and all prefixes are supplied separately. Timed worker setup must not be interpreted as steady-state per-query latency. Shared inference-server resources, energy and billed currency were not measured.

Post-observation validation amendment. The original pipeline stopped after seven successful workers because post-maintenance passage vectors failed exact equality. Initial graph/dense vec-

tors were bit-identical. For repeat 1, the maintained stores have identical document IDs/text, finite 9,821×1,024 arrays, but 574 rows differ (maximum component difference 1.112×10^{-6}; maximum row L2 difference 8.925×10^{-6}). An isolated replay reproduces one differing retained passage under its two batch contexts. It diagnoses that case, not invariance of all rankings or answers.

A separately approved and frozen continuation retains initial exact equality as a gate and reports maintained-vector inequality as a failed diagnostic. It still requires expected document identity/content, shape, finite values and frozen sources, and reports discrepancies for every repetition. Seven original receipts are reused unchanged; only the remaining 23 workers execute. No post-hoc epsilon relabels the failed test as equivalence. The original failed status, amendment, and completion of the amended cost audit remain separate records. This is an amended descriptive policy-cost study, not a maintained-representation quality comparison.

Independent local audit and reproducibility boundary. Local replay verifies all 30 receipts, the seven reused hashes, 9,450 prompts/first-valid outputs, 133,542 invocation events, 74,796 physical requests, and 108 cost-prefix rows. It independently checks all twelve initial/updated passage-vector stores and reproduces the three maintenance diagnostics. Total reported usage is 56,402,342 prompt and 11,812,105 completion tokens, with no missing usage or length-terminated requests. The graph, entity and fact numerical construction was not rerun locally; original model execution and shared-service costs are outside this audit. Public replay preserves the scientific inputs, outputs, all attempts and cost measurements, with documented removal of identifying operational fields. Large intermediate graph/vector files and model weights remain explicit prerequisites for numerical reconstruction rather than hidden dependencies of the saved-results analysis.

APPENDIX R — EXTENDED RELATED WORK

Associative memory and explicit retrieval. Kanerva's SDM provides a distributed associative store (Kanerva, 1988). Subsequent work connects attention to SDM (Bricken & Pehlevan, 2021) and develops SDM-derived continual learners (Bricken et al., 2023). These connections motivate studying memory mechanisms, but they concern different objects: an explicit agent trace store is not an SDM-derived multilayer perceptron trained for continual learning. Our byte-matched retrieval comparisons concern the former. They do not test, or refute, the continual-learning properties of the latter. Likewise, a favorable associative-memory result with a different representation, workload, or learning objective is not contradicted by our tested readout failures.

Universal Hopfield Networks place associative memories in a framework of similarity, separation, and projection operations (Millidge et al., 2022). This is close in spirit to asking which operation produces a result. Our constructive projection is a different use of the term: it maps one recorded controller trajectory into two storage implementations so that state equality is guaranteed by construction. We then release that coupling in a separate closed-loop experiment. The empirical question is whether the storage contrast persists under a specified byte budget and controller, rather than which associative update rule has the largest capacity in general. Dense Passage Retrieval establishes a strong learned dense-retrieval alternative for open-domain QA (Karpukhin et al., 2020); our matched controls additionally vary how dense rows are stored under the operational-byte constraint.

Temporal memory mechanisms. MemoryBank combines long-term retrieval with a forgetting-curve-inspired update mechanism and reinforcement (Zhong et al., 2024). Generative Agents store experiences, synthesize reflections, and retrieve memories to support behavior in an interactive environment (Park et al., 2023). Both motivate testing what changes over time, but an elapsed-time weight, a newly written reflection, and a reorganized index need not be activated by the same intervention. Our zero-time result is about the tested clock-dependent controller under its initial-state and score-plateau conditions. It does not imply that reflection or other active transformations cannot help on a static conversation corpus. The corpus-time replay and MemoryBank source inspection address this distinction directly.

Graph retrieval and propagation. HippoRAG introduces neurobiologically inspired graph retrieval, and HippoRAG 2 develops passage integration and online model use alongside personalized PageRank (Gutiérrez et al., 2024; 2025). Their system-level results motivate our routing experiments. A full-versus-dense comparison changes graph access, linking, filtering, and propagation together; it supports a claim about that retrieval bundle. Keeping the graph and reset fixed while applying the propagation operator once instead isolates further iteration. The prospective bridge-question improvement supports the former claim, whereas failed one-step noninferiority leaves the latter unresolved. Our lifecycle repetitions add costs for the measured implementation policies; they do not substitute those policies for all possible graph-update algorithms.

Memory taxonomies and benchmarks. Hu et al. (2025b) organize agent memory by forms, functions, and dynamics. Our storage/dynamics/routing split is an operational choice of intervention sites, not a claim to exhaust the survey's design space. LoCoMo evaluates memory over long conversations (Maharana et al., 2024); LongMemEval separates abilities including information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention (Wu et al., 2025). MemoryAgentBench uses incremental interactions to evaluate retrieval, test-time learning, long-range understanding, and selective forgetting (Hu et al., 2025a). These benchmarks establish useful tasks, while component attribution additionally requires checking how their events reach the implementation. A corpus timestamp need not become the controller's retention time, and a hot-store intervention need not be exercised by a cold-retrieval adapter. Our chronological replay and activation checks make those execution paths part of the evidence.