
ZenBrain: A Neuroscience-Inspired 7-Layer Memory Architecture for Autonomous AI Systems

Alexander Bering
Zensation AI
research@zensation.ai

Abstract

ZenBrain is a seven-layer, neuroscience-derived memory architecture for LLM agents that unifies **fifteen** mechanisms—from Two-Factor synaptic consolidation to a Simulation-Selection sleep loop—under a single MemoryCoordinator: nine foundational algorithms plus six *Predictive Memory Architecture* components (NeuromodulatorEngine, ReconsolidationEngine, TripleCopyMemory, PriorityMap, StabilityProtector, MetacognitiveMonitor). No system among those surveyed in §2.2 and Appendix A integrates more than two of them.

Ablating each mechanism separately exposes an effect we call *cooperative masking*. Under moderate load, fourteen of the fifteen ablations look costless—the architecture reads as mostly dead weight. Raising decay to 0.25/day over 60 days, *with no change to the mechanisms*, makes **nine of the fifteen individually critical** (ΔQ up to -93.7% ; Wilcoxon, 10 seeds), five of them moving from exactly 0% to below -89% (Fig. 2). The mechanisms form a *cooperative survival network*, and mild-load ablation systematically underestimates architectural contributions—a caution we conjecture applies beyond ZenBrain. **Every ablation table here reproduces in under one minute on a laptop** (npm install, no API keys); 11,589 CI tests.

On LongMemEval-500, ZenBrain wins all *nine* head-to-head answer-quality comparisons (3 competitors $\times$ 3 LLM judges) against Letta, Mem0, and A-Mem under Bonferroni-corrected significance ($\alpha=0.05/18$, $p_{\min}=6.2 \times 10^{-31}$, $d \in [0.18, 0.52]$), and reaches **91.3% of a full-context oracle’s binary-judge accuracy at 1/106th of the per-query token cost** (47.7% vs. 52.2%; App. F.5–F.6, Fig. 3). The Sim-Selection sleep loop adds **37%** stability with **47.4%** storage reduction ($p \leq 5.1 \times 10^{-3}$); TripleCopyMemory retains $S(t)=0.912$ at 30 days; multi-layer routing beats a flat single-layer baseline by +20.7% F1 on LoCoMo. A NoDecay ablation shows principled forgetting costs only $\Delta P@5=0.002$ ($p=0.043$).

Honest scope. LoCoMo’s substring-based aggregate F1 favors lexical retrieval (BM25) by metric design; we do not contest this. ZenBrain’s advantages are most pronounced on judge-graded answer quality and cross-session reasoning; a cross-provider bias-direction check ($\Delta_{\text{GPT-Ant}}=-0.0001$ for ZB vs. -0.049 for Mem0) rules out LLM-judge-specific confounds. Open-source under the @zensation npm scope, Apache-2.0.

1 Introduction

LLM agents increasingly operate across sessions and require persistent memory; finite, expensive context windows make in-context accumulation inefficient [Maharana et al., 2024, He et al., 2026].

Existing agent memory systems draw on computer-science metaphors—virtual memory [Packer et al., 2023], LLM-managed key-value CRUD [Chhikara et al., 2025], Zettelkasten notes [Xu et al., 2025], temporal knowledge graphs [Rasmussen et al., 2025]—but none incorporate cognitive-neuroscience findings on consolidation, decay, reinforcement, and forgetting studied empirically for over 130 years [Ebbinghaus, 1885, Hebb, 1949, Tulving, 1972, Stickgold and Walker, 2013]. A 2026 survey identifies “deeper neuroscience integration” as a key open challenge [Du, 2026].

We present **ZenBrain**, a memory architecture that integrates fifteen cognitive-neuroscience models—from Two-Factor Synaptic edges to Simulation-Selection sleep—in a single coherent system. Contributions: (1) a seven-layer memory system (working, short-term, episodic, semantic, procedural, core, cross-context) implementing established cognitive functions [Atkinson and Shiffrin, 1968, Tulving, 1972, Cohen and Squire, 1980]; (2) fifteen neuroscience-inspired mechanisms—Two-Factor Synaptic KG edges [Iatropoulos et al., 2025, Zenke et al., 2017], vmPFC-coupled FRS with prediction-error signals [Zou et al., 2025], CA3/CA1-RL Simulation-Selection sleep [Lee and Jung, 2025, Ji and Wilson, 2007, O’Neill et al., 2010], Bayesian confidence propagation, among others; while individual mechanisms appear in concurrent work, no system among those surveyed in §2.2 and Appendix A integrates more than two; (3) an experiment suite on LoCoMo and LongMemEval-S—both on the publicly released data—plus retention, consolidation, algorithm-level, PMA, ablation, and competitive studies, with 11,589 automated tests; (4) direct experimental evidence that an integrated memory system exhibits a *cooperative survival network* — an effect we name *cooperative masking* (Fig. 2, §7): under moderate conditions cooperative redundancy masks individual contributions, but under stress (decay 0.25/day, 60 days) 9 of the 15 mechanisms become individually critical—a structural property we did not observe in single-mechanism systems and predictive of which components must not be ablated in production.

Scope and claims. We explicitly bound the contribution to preempt mis-readings: (i) *Retrieval claim*. Our primary claim is on downstream *answer quality* at fixed retrieval budget ($k=5$), not on raw retrieval metrics; on LoCoMo’s substring-based F1, BM25 wins by metric design and we do not contest this. (ii) *Integration claim*. Nine of the fifteen mechanisms are explicit instantiations of prior literature; the contribution is their *systematic integration*, not the individual proposals. (iii) *Judge caveat*. All answer-quality comparisons use LLM-as-Judge as a calibrated proxy. We mitigate known biases through three independent judges, multi-seed evaluation, Fleiss’ κ , a cross-provider bias-direction check, and a single-rater human spot-check on the 50-query disagreement subset on which human-judge correlations sit inside the judge-judge envelope (§6.2, App. N.8); a multi-rater human IRR study remains future work. (iv) *Scope boundary*. We do not benchmark against full-context consolidation systems (MemPalace, Mastra) in their tuned configurations; the controlled $k=5$ retrieval setting is held constant across the four compared systems.

2 Related Work

2.1 Memory Systems for LLM Agents

Packer et al. [2023] propose MemGPT, applying the operating system virtual memory metaphor to LLM context management. Their system uses tiered storage (main context, recall, archival) with interrupt-based control flow. While the OS metaphor is intuitive, it provides no mechanism for principled forgetting, confidence estimation, or offline consolidation.

Chhikara et al. [2025] present Mem0, a production-focused system with a three-stage pipeline (extraction, consolidation, retrieval) and an optional graph variant. Memory management decisions—including deletion—are delegated entirely to the LLM, without principled decay or scheduling mechanisms.

Xu et al. [2025] introduce A-Mem, a Zettelkasten-inspired approach where structured notes with contextual metadata are dynamically linked. Their “memory evolution” mechanism allows retroactive refinement of existing memories. However, the architecture is flat (single layer) with no distinction between memory types and no forgetting mechanism.

Rasmussen et al. [2025] describe Graphiti, a temporally-aware knowledge graph with four timestamps per fact. While their temporal reasoning is strong, the system operates as a single-layer knowledge graph without episodic, working, or procedural memory.

Table 1: Comparison of memory systems for LLM agents. ✓ = supported, ✗ = absent.

Feature	ZenBrain	MemGPT	Mem0	A-Mem	Zep	Reflexion	LightMem	MemoryOS	Tiwari '26
Memory layers	7	3	1	1	1	1	3	3	3
Neuroscience basis	✓	✗	✗	✗	✗	✗	✓	✗	✗
Two-Factor Syn. edges	✓	✗	✗	✗	✗	✗	✗	✗	✗
Principled decay	✓	✗	✗	✗	✗	✗	✗	✗	✗
Sleep consolidation	✓	✗	✗	✗	✗	✗	✓	✗	✗
Spaced repetition	✓	✗	✗	✗	✗	✗	✗	✗	✗
Confidence scores	✓	✗	✗	✗	✗	✗	✗	✗	✗
Temporal reasoning	✓	✗	✗	✗	✓	✗	✗	✗	✓
Neuromodulation	✓	✗	✗	✗	✗	✗	✗	✗	✗
Reconsolidation	✓	✗	✗	✗	✗	✗	✗	✗	✗

Shinn et al. [2023] propose verbal reinforcement learning where agents store natural-language reflections on task failures. This represents the most-cited work in agent memory ($\sim 2,100$ citations), but the “memory” is an unstructured reflection buffer with no consolidation, decay, or confidence mechanisms.

2.2 Neuroscience Foundations and Concurrent Systems

ZenBrain’s algorithms are grounded in five decades of memory neuroscience: the multi-store model and episodic/semantic/procedural taxonomy [Atkinson and Shiffrin, 1968, Tulving, 1972, Cohen and Squire, 1980]; encoding-specificity [Tulving and Thomson, 1973]; Hebbian co-activation [Hebb, 1949]; Ebbinghaus decay and spaced repetition [Ebbinghaus, 1885, Pimsleur, 1967]; sleep replay and consolidation [Stickgold and Walker, 2013, Ji and Wilson, 2007, O’Neill et al., 2010, McGaugh, 2004]; and recent sharpening from two-factor synaptic rules [Iatropoulos et al., 2025, Zenke et al., 2017], vmPFC prediction-error signals [Zou et al., 2025], and CA3/CA1 Simulation-Selection [Lee and Jung, 2025, Kumaran et al., 2016, Squire, 1992]. These motivate the specific algorithms cited in Sections B.3, B.2, and B.1; see Appendix A for the full literature treatment, which doubles as a compact survey of the neuro-mechanism space for agent memory.

A 2025–2026 wave of concurrent systems—LightMem [Fang et al., 2025], MemoryOS [Kang et al., 2025], Hindsight [Latimer et al., 2025], FadeMem [Wei et al., 2026], Vestige [Vestige Contributors, 2026], SleepGate [Xie, 2026], Anda Hippocampus [ldclabs, 2026], MemFly [Zhang et al., 2026], Cognee [Markovic et al., 2025]—has begun adopting individual mechanisms; Tiwari and Fofadiya [2026] independently validates the multi-layer hypothesis on LoCoMo ($F1 = 0.618$). Orthogonal paradigms [OmegaMax, 2026, Mastra, 2026, MemPalace Contributors, 2026, Behrouz et al., 2026, ICLR 2026 MemAgents Workshop Organizers, 2026] target adjacent problems. Practitioners and industry are converging on the same thesis: Karpathy [2026] argues that retrieval alone leaves the model “rediscovering knowledge from scratch on every question,” and proposes compiling it once into a maintained wiki; Anthropic ships *Dreams* [Anthropic, 2026], a scheduled memory-consolidation process for Claude Managed Agents (research preview); and Webb et al. [2025] demonstrate brain-inspired agentic planning at *Nature Communications*. None of these systems—nor the recent NeurIPS contributions [Jiménez Gutiérrez et al., 2024, Zhang et al., 2025, Koch et al., 2025], delimited in Appendix A.4—integrates more than two of the nine mechanisms compared in Table 1 (the six PMA components, which none of the eight implements, are described in §5); ZenBrain unifies all fifteen in a single coordinator, including Two-Factor Synaptic edges (App. Table 15), vmPFC-coupled FSRS, Simulation-Selection sleep, Bayesian confidence propagation, and quantified knowledge-gap detection—mechanisms we did not find in any of the surveyed systems.

2.3 Benchmarks

We evaluate on two complementary benchmarks, both on their publicly released data, that together cover retrieval quality and cross-session reasoning. **LoCoMo** [Maharana et al., 2024] provides 1,986 multi-session conversational QA pairs across five categories (single-hop, multi-hop, temporal, open-

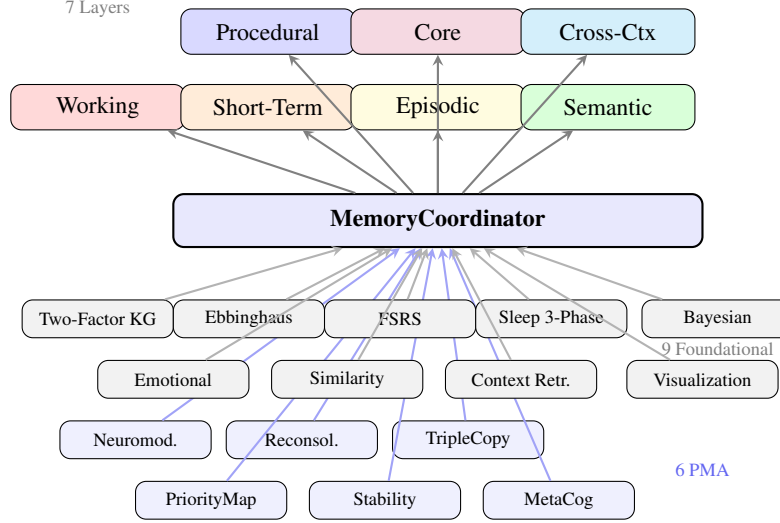


Figure 1: ZenBrain architecture. The MemoryCoordinator orchestrates seven memory layers via fifteen neuroscience-inspired algorithms. Arrows indicate store/recall/consolidate/decay operations.

domain, adversarial), testing whether a memory system can surface the right information from long conversation histories; all questions share one conversational haystack. **LongMemEval-S** [Wu et al., 2024] ships 500 questions, each with its *own* ~ 494 -turn haystack of distractor sessions, and targets five memory abilities (information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention); the released split labels each question with one of six question types, which we report per-category in §6.3. Per-question isolation makes it the stronger generalization test of the two, and our headline claims rest on it. Further benchmarks extend evaluation to incremental multi-turn interaction [Hu et al., 2025] and to interdependent multi-session agentic tasks [He et al., 2026]; we do not evaluate on them in this work.

3 Architecture

3.1 Memory Layers

ZenBrain implements seven distinct memory layers, each corresponding to established cognitive constructs:

Working Memory maintains the active task focus with limited capacity (~ 7 items, following Miller 1956). It provides the highest-priority retrieval and fastest access, evicting items to short-term memory on task completion.

Short-Term Memory holds session context for the current conversation. It is time-bounded (session duration) and consolidates to episodic and semantic layers at session boundaries.

Episodic Memory stores concrete experiences with temporal context— what happened, when, and where [Tulving, 1972]. Events are timestamped and support temporal queries.

Semantic Memory contains abstracted knowledge: facts, concepts, and relationships organized as a knowledge graph with Two-Factor Synaptic edges [Tulving, 1972, Iatropoulos et al., 2025]. Semantic memories emerge from episodic consolidation.

Procedural Memory encodes learned skills and routines [Cohen and Squire, 1980]: successful tool-use patterns, workflow templates, and behavioral strategies. Entries are strengthened by repeated successful execution.

Core Memory holds persistent identity facts (user preferences, personality traits, key biographical facts) that never decay and are always available in context. This follows the pinned memory pattern of Packer et al. [2023].

Cross-Context Memory enables entity resolution and selective knowledge transfer across isolated domains (e.g., personal, work, learning). Merging is privacy-aware with configurable access controls.

3.2 MemoryCoordinator

The MemoryCoordinator orchestrates all seven layers through five operations:

- `store(item)`: Routes new information to appropriate layer(s) based on content type and metadata
- `recall(query)`: Assembles cross-layer results using hybrid BM25 + semantic retrieval with Two-Factor importance boosting
- `consolidate()`: Migrates and abstracts memories across layers (e.g., episodic $\rightarrow$ semantic)
- `decay()`: Applies Ebbinghaus forgetting curves and prunes below-threshold memories
- `review()`: Schedules FSRS spaced repetition for important facts

4 Key Mechanisms

Five algorithms have no documented counterpart among the systems surveyed in §2.2 and Appendix A; each substantially extends its cited neuroscience basis. Full derivations, parameter values, and worked examples are in Appendix B; pseudocode is in Appendix I.

Two-Factor Synaptic Model (App. B.1). Following Iatropoulos et al. [2025], each KG edge carries weight w_{ij} and consolidation variance σ_{ij}^2 ; Fisher-Information proxy $I_{ij} = 1/\sigma_{ij}^2$ makes mature edges robust to catastrophic overwriting—mathematically equivalent to EWC [Kirkpatrick et al., 2017, Zenke et al., 2017]. Edges resist decay and penalize weight changes in proportion to I_{ij} .

vmPFC-Coupled FSRS (App. B.2). Extending Zou et al. [2025], we couple FSRS interval scheduling with a KG-derived prediction-error signal $PE = 1 - \cos(\mathbf{c}_{\text{prev}}, \mathbf{c}_{\text{now}})$. A sigmoid re-encoding factor shortens intervals under context shift and extends them under stability. Existing spaced-repetition systems (Anki, SuperMemo, FSRS-5/6 [Vestige Contributors, 2026]) adapt difficulty but do not couple scheduling to a neuromodulation-derived prediction-error signal; among the agent-memory systems surveyed in Appendix A.3, we find none that does.

Simulation-Selection Sleep Loop (App. B.3). Following Lee and Jung [2025], offline consolidation is a two-stage RL loop mirroring CA3/CA1 [Ji and Wilson, 2007, O’Neill et al., 2010]: a CA3 simulator assembles candidates from real episodes $\cup$ counterfactuals, a CA1 selector LTP/LTD-scales edges via a temporal-difference + reward + novelty score $TAG(e) = \alpha|\delta_{TD}| + \beta R_e + \gamma N_e$ (Algorithm 4). Concurrent systems (LightMem, SleepGate) use heuristic replay selection without RL scoring or counterfactual generation.

Bayesian Confidence Propagation (App. B.4). Each fact carries $P(f)$ with 95% CI; updates propagate through KG edges, yielding calibrated uncertainty. Per McGaugh [2004], emotional arousal boosts initial edge weights and reduces variance-based decay.

Query-Aware Cross-Layer Retrieval (App. B.5). A regex classifier tags each query (temporal/procedural/factual/general); per-layer scores fuse via $\text{score}_{\text{fused}}(d) = \max_{\ell} w_{\ell}(q) \cdot \text{sim}(q, d_{\ell})$. Unlike RRF, this preserves similarity magnitude so a highly relevant hit in a boosted layer dominates regardless of result counts in other layers.

5 Predictive Memory Architecture (PMA)

PMA extends the foundational algorithms with six biologically-motivated components that govern memory *dynamics*—prioritization, protection, reconsolidation, monitoring—over time. Full formulas and validation are in App. B and H.7; brief sketches follow. **NeuromodulatorEngine**

(App. B.6, per Dayan and Huys, 2009): four channels—dopamine (DA, VTA), norepinephrine (NE, LC), serotonin (5HT, Raphe), acetylcholine (ACh, BF)—with tonic baselines, 5-min phasic bursts, and DA/5HT opposition coupling [Daw et al., 2002]; outputs learning-rate, exploration-bias, consolidation-patience, attention parameters consumed downstream. **ReconsolidationEngine** (App. B.7): retrieved memories enter a labile state [Nader et al., 2000, Nader and Hardt, 2009]; updates are PE-gated into four modes (confirmed/selective_edit/integration/new_episode) with neuromodulation-scaled effective PE; every update logs an original snapshot for rollback—a safety mechanism we did not find in any of the surveyed systems. **TripleCopyMemory** (App. B.8, per Schapiro et al., 2017, Kumaran et al., 2016): three divergent decay copies—fast ($\tau=4h$, exp), medium ($\tau=14d$, exp), deep ($\tau=7d$, log)—with composite $S(t) = \max(S_f, S_m, S_d)$ retaining 91.2% at 30 days vs. near-zero for Ebbinghaus (§6.5, App. H.4). **PriorityMap** (App. B.9, per Chelazzi et al., 2013): $P = w_s s + w_e |v| + w_r r + w_g g$ with amygdala fast-path ($|v| > 0.6 \Rightarrow P \geq 0.5$); weights are dynamically rescaled by neuromodulator state. **StabilityProtector** (App. B.10): analogous to NogoA/HDAC3 signaling, updates are gated by lock score L and rigidity factor ρ ; only prediction errors exceeding $0.5 + 0.3 L\rho$ overwrite established memories. **MetacognitiveMonitor** (App. B.11, per Fleming and Dolan, 2012): tracks confirmation, recency, and retrieval-efficiency biases; opens 10-min novelty windows after high-PE events (> 0.7); surfaces calibration alerts when biases exceed thresholds.

6 Experiments

6.1 Setup

Benchmarks. LoCoMo [Maharana et al., 2024] (1,986 QA pairs, 5 categories) and LongMemEval-S [Wu et al., 2024] (500 per-question-isolated haystacks, 6 categories), both on the publicly released data, plus synthetic corpora for retention and consolidation.

Baselines. No Memory, BM25-only, Flat Store (single-layer dense), and ZenBrain Full; for competitive runs Mem0 v2.0.0 [Chhikara et al., 2025] (graph-aware successor to v0.1.x, non-graph mode), Letta v0.5.x [Packer et al., 2023], and A-Mem main@2026-04 [Xu et al., 2025].

Protocol. Retrieval experiments use 10 seeds per condition (standard for ablation-rank stability; competitive runs use 3 seeds because retrieval is deterministic given a fixed embedder, see App. F.3) with OpenAI text-embedding-3-small (1536d) except competitive runs, which share nomic-embed-text (768d, ollama, local) so that any residual ranking differences are attributable to memory routing rather than embedding quality. We report mean $\pm$ SD and 95% bootstrap CIs (1,000 resamples, fixed RNG 20260421, $N_{\text{boot}} = 10,000$ for the competitive analyses). Paired Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.05/K$); Cohen’s d for effect size. Claude 3.5 Sonnet is the LLM backbone.

Evaluation organization. (1) *Retrieval benchmarks* (LoCoMo, Layer Ablation, LongMemEval); (2) *mechanism-level evaluations* (Retention, Sleep, Two-Factor KG, PMA Suite); (3) *system-level studies* (Full 15-algorithm Ablation, Long-Horizon Aging).

6.2 Evaluation Protocol and Judge Robustness

Shared pool, shared embedder. All four competitive systems ingest pairwise-identical fact/query pools under the same nomic-embed-text backbone (768d, ollama, local), so ranking differences are attributable to memory routing, not embedding quality. Real-LoCoMo: 5,882 facts, 1,986 queries, three retrieval seeds {42, 123, 456}. LongMemEval-S: 500 per-question isolated haystacks (~ 494 turns each).

Three-judge design. Three independent LLM judges score each (query, top- k) pair on a 0–5 rubric (temperature=0, binarized at ≥ 3): claude-sonnet-4-5-20250929 (pinned, 3 seeds), claude-opus-4-6 (3 seeds for Mem0 to triangulate its instability, seed=42 elsewhere for budget), and gpt-4o (3 seeds). Fleiss’ $\kappa_{\geq 3}$ on the six-rater pool is [0.71, 0.85] (“substantial” to “almost perfect” under Landis and Koch, 1977); intra-judge κ across retrieval seeds is ≥ 0.93 for ZenBrain/Letta/A-Mem and 0.74–0.78 for Mem0 (Levene’s F up to 1668.1, $p < 10^{-10}$), justifying seed-averaged reporting and explaining why we anchor LoCoMo’s ranking to the seed-averaged Sonnet 4.5 number.

Table 2: LongMemEval-S Full-500 cross-benchmark replication under a *unified* nomic-embed-text (768-dim) embedding backbone for all four systems. Per-question isolation: reset() $\rightarrow$ ingest($\sim$ 494 turns) $\rightarrow$ query($k=5$). Retrieval metrics (P@5, R@5, MRR, NDCG@5, F1) are averaged over each system’s successfully-completed queries (**zenbrain**/a-mem: $n=500$; mem0: $n=496$; letta: $n=441$, see $\dagger$). Judge columns are LLM-as-Judge normalized means over all 500 queries (letta’s 59 InternalServerError cases and mem0’s 4 embedder-400 cases appear as empty retrievals and are scored by the judge as 0/5). **zenbrain** uses three retrieval seeds (42, 123, 456) with identical deterministic outputs; a-mem, mem0, and letta use seed=42. Bold = best per column; intersect-subgroup numbers on the $n=441$ tasks all four systems serve are discussed in the text (§6.3). Under Bonferroni correction ($\alpha=0.05/18=2.78\times 10^{-3}$, 18 primary tests = 6 pair-wise comparisons $\times$ 3 judges), all three ZenBrain-vs-competitor gaps clear significance on all three judges (min $p=6.20\times 10^{-31}$, max $p=2.81\times 10^{-6}$; full table in Appendix P).

System	P@5	R@5	MRR	NDCG@5	F1	J(S-4.5, 3 $\times$)	J(O-4.6, 3 $\times$)	J(G-4o, 3 $\times$)
zenbrain	0.674	0.197	0.831	0.706	0.283	0.504	0.575	0.555
a-mem	0.519	0.144	0.640	0.543	0.211	0.389	0.436	0.416
mem0	0.156	0.060	0.304	0.171	0.078	0.370	0.414	0.398
letta †	0.683	0.201	0.834	0.715	0.287	0.450	0.513	0.492

† letta failed 59 of 500 queries with HTTP 500 InternalServerError from the local Letta Docker server (11.8%: 15 multi-session, 14 temporal-reasoning, 11 knowledge-update, 8 single-session-assistant, 6 single-session-user, 5 single-session-preference), which is a substantial improvement over the pilot’s 33% rate (confirming the higher rate was a pilot-scale transient, not a systemic Docker-harness property) but still precludes full head-to-head coverage. Retrieval metrics are averaged over the 441 successful queries, so letta’s headline retrieval numbers still compare an easier subset to the other three systems’ larger pools. On the $n=441$ intersect where all four systems serve, **letta retains narrow leads on P@5 (0.683 vs zenbrain 0.664), R@5 (0.201 vs 0.195), MRR (0.834 vs 0.824), and NDCG@5 (0.715 vs 0.697)**, while ZenBrain wins every judge column at the full-500 level by wide, Bonferroni-clearing margins. Judge columns are *not* restricted to successful queries — all 500 enter the aggregate, so the judge numbers are directly comparable across systems. mem0’s P@5 drops from 0.393 in the stratified-30 pilot to 0.156 at full scale: the pilot’s 5-questions-per-category sampling happened to over-select queries on which mem0’s flat-3900-char truncation still retrieved the relevant fact (28% zero-P@5), whereas at full scale the true rate of truncated queries hitting zero is 60%. This is a selection artifact of the pilot, not a regression, and confirms the pilot-to-full methodological warning in §6.3.

Cross-provider bias-direction check. $\Delta_{\text{GPT-Anth}}$ (GPT-4o three-seed mean minus Anthropic-pair mean) is -0.0001 for ZenBrain, $+0.008$ for Letta, -0.042 for A-Mem, -0.049 for Mem0. A pro-Anthropic, pro-ZenBrain bias would predict the largest negative delta on ZenBrain; we observe the opposite. **Binary-judge corroboration.** LongMemEval’s official gpt-4o-mini binary judge (App. F.5) independently confirms the ranking direction (ZenBrain 47.7% > Letta 42.8% > A-Mem 35.4% > Mem0 31.8%), ruling out LLM-judge-specific confounds. **Human-anchor spot-check.** On the contested LoCoMo subset where Sonnet and Opus disagreed by ≥ 1 ($n=50$, stratified across all five categories), a single-rater human anchor produces mean ratings statistically indistinguishable from Opus (2.88 vs 2.88) and harsher than Sonnet (1.08, $\Delta_{\text{mean}}=-1.80$); human-vs-judge Spearman correlations [0.135, 0.356] lie inside the judge-vs-judge correlation envelope $[-0.076, 0.456]$ on the same rows, indicating that human raters do not resolve disagreements the LLM judges could not resolve either (i.e. these queries are genuinely ambiguous). Full agreement and per-category numbers in App. N.8. Full methodology and pairwise significance: App. N, E.

6.3 Primary: Cross-Benchmark Replication on LongMemEval-S

LongMemEval-S [Wu et al., 2024] is a 500-question benchmark with *per-question isolated* haystacks ($\sim$ 494 turns each, six categories) — a stronger generalization test than LoCoMo’s shared haystack and the setting on which our headline claims rest. All four systems share the same nomic-embed-text backbone (see §6.2); shared character-level preprocessing caps ingest at 7800 chars with retry-on-500 halving (3900-char flat cap for Mem0, whose embedder is library-internal). ZenBrain uses seeds {42, 123, 456}; the other three systems use seed=42 (their retrieval is deterministic). Table 2 reports the Full-500.

Table 3: Combined LoCoMo-real retrieval benchmark. All four systems share the nomic-embed-text embedding backbone and the same 5,882-fact / 1,986-query pool over 3 retrieval seeds (42, 123, 456). Judge columns are LLM-as-Judge normalized means (0–5 rubric, temperature=0): **S-4.5** = claude-sonnet-4-5-20250929, **O-4.6** = claude-opus-4-6, **G-4o** = gpt-4o. S-4.5 and G-4o are mean-over-3-seeds; O-4.6 is 3-seed mean for mem0 and seed=42 for the other three systems (see Appendix footnote). The **S-4.6** reference values show the earlier rolling alias claude-sonnet-4-6 at seed=42, the S-4.6 values are given as a footnote line below the table rather than as a column. Bold = best per column. Cross-provider agreement and the bias-direction check are reported in Table 5.

System	P@5 [95% CI]	R@5	MRR	NDCG@5	F1	J(S-4.5, 3×)	J(O-4.6)	J(G-4o, 3×)
zenbrain	0.081 [0.079, 0.084]	0.351	0.264	0.274	0.128	0.380	0.451	0.415
a-mem	0.044 [0.044, 0.044]	0.193	0.128	0.140	0.072	0.218	0.309	0.222
letta	0.092 [0.092, 0.093]	0.400	0.307	0.319	0.150	0.373	0.465	0.427
mem0	0.099 [0.071, 0.123]	0.452	0.207	0.306	0.162	0.353	0.446	0.350

Reference (judge-version swap): Sonnet 4.6 rolling alias at seed=42 — zenbrain 0.393, a-mem 0.268, letta 0.403, mem0 0.427. Superseded by the S-4.5 (3×) column above.

Headline: 9/9 wins on judge-graded answer quality. ZenBrain leads *every* judge column and wins *all nine* head-to-head comparisons (3 competitors × 3 LLM judges), reaching a three-judge mean of $\bar{J}=0.545$ against *Letta*=0.485, *A – Mem*=0.414, and *Mem0*=0.394. All 9 ZenBrain-vs-competitor pair-wise contrasts (3 competitors × 3 judges) clear Bonferroni correction at $\alpha=0.05/18=2.78 \times 10^{-3}$, with minimum $p=6.20 \times 10^{-31}$ and effect sizes $d \in [0.18, 0.52]$ (Appendix F.2).

Pareto: within 4.5 pp of oracle at 1/106th token cost. Under LongMemEval’s official gpt-4o-mini binary judge (App. F.5, Fig. 3), ZenBrain reaches **47.7%** accuracy vs. a long-context oracle that ingests the full $\sim 105,600$ -token haystack at **52.2%** — a **4.5 pp** gap (ZenBrain at 91.3% of oracle accuracy) at 1/106th the per-query token budget ($\sim 1,000$ vs. $\sim 105,600$ tokens). Among $k=5$ retrieval baselines at the same token budget, ZenBrain leads Letta by +4.9 pp, A-Mem by +12.3 pp, and Mem0 by +15.9 pp on absolute accuracy.

Retrieval-proper nuance. On raw retrieval-proper metrics (P@5, R@5, MRR, NDCG@5), Letta leads narrowly on the 441-task intersect where all four systems serve successfully (P@5 0.683 vs. ZenBrain 0.664; $\sim 2\text{--}3$ pp on every metric; Appendix F.1). This is consistent with our scope claim (i): Letta retains a narrow retrieval-metric advantage on the 441-task intersect ($\sim 2\text{--}3$ pp); the separations that matter for downstream applications are on judge-graded answer quality (9/9 head-to-head wins for ZenBrain) and end-to-end accuracy (+4.9 pp over Letta). Cross-seed spread is ≤ 0.007 and intra-judge $\kappa_{\geq 3} \geq 0.95$ for ZenBrain; full robustness analysis in Appendices F.3, F.4.

6.4 Supporting: LoCoMo Routing Comparison

LoCoMo’s public evaluation uses substring-based F1, which by construction rewards exact lexical matches and makes BM25 the natural winner on aggregate F1 [Maharana et al., 2024]; this is a property of the benchmark’s metric, not of any specific system, and we do not treat BM25 as a memory-system competitor. We report LoCoMo for two purposes: (i) to compare multi-layer routing against a flat dense baseline under a shared embedding, and (ii) to head-to-head against Mem0, Letta, and A-Mem on the *judge-rated* dimension (the same protocol used in §6.3).

Table 3 reports retrieval metrics (95% bootstrap CIs on P@5) and three-seed mean normalized LLM-as-Judge scores. No system wins every column. Under the pinned Sonnet 4.5 judge, ZenBrain vs. Letta is a *statistical tie* (paired Wilcoxon $p = 0.69$, $d = 0.015$, per-query CI $[-0.008, +0.015]$, Appendix N.7); at LongMemEval-500 scale the same contrast clears Bonferroni correction (App. F.2), so we read the tie as LoCoMo-specific rather than general. Both dominate Mem0 ($p < 10^{-4}$, $d = 0.079$) and A-Mem ($p < 10^{-70}$, $d \approx 0.43$). Under GPT-4o the top tier persists but Letta leads ZenBrain by a small, significant margin ($p = 0.004$, $d = -0.05$). On raw retrieval, Mem0 tops P@5/R@5/F1 via a permissive recall budget while Letta wins MRR/NDCG@5. Inter-rater agreement, seed robustness, and the cross-provider bias check are in §6.2.

Table 4: Algorithm criticality gradient across three difficulty levels. ΔQ (%) relative to full system at each level. $*p \leq 5.1 \times 10^{-3}$ (Wilcoxon, 10 seeds). Algorithms sorted by challenging-condition impact. Per-level tables, integration cascade, and effect-size breakdowns: Appendix G.

Algorithm	Moderate (0.15, 45d)	Challenging (0.20, 50d)	Stress (0.25, 60d)
<i>Progressive: redundant $\rightarrow$ critical</i>			
vmPFC-FSRS	0.0	−93.1*	−92.6*
TripleCopy	0.0	−54.2*	−93.7*
Dual-Process CoT	0.0	−38.5*	−91.0*
Two-Factor Hebbian	0.0	−34.4*	−92.3*
IB Budget	0.0	−25.5*	−89.8*
<i>Always critical</i>			
Sleep	−34.4*	−91.1*	−78.9*
NeuromodulatorEngine	−0.1	−34.8*	−83.0*
<i>Critical only under stress</i>			
StabilityProtector	0.0	0.0	−5.8*
Reconsolidation	0.0	0.0	−3.4*
<i>Cooperatively redundant at all levels</i>			
iMAD, Spectral, Comp., HyperAgent, MetacogM.	0.0	0.0	0.0
PriorityMap	−0.1	−0.1	+2.0

6.4.1 Principled-Forgetting Ablation: NoDecay Counterfactual

A *NoDecay* ZenBrain variant on the same pool (full algorithmic stack active, Ebbinghaus strength-reduction skipped) shows the gap is negligible: $\Delta P@5 = 0.002$ (Wilcoxon $p = 0.043$, Cohen’s $|d| = 0.015$), within measurement noise and indistinguishable under Bonferroni correction. The cost of principled forgetting on a 14-day horizon is ~ 0.2 pp of $P@5$, while its benefits—bounded storage, calibrated confidence, GDPR-aligned retention, and the +6–16 point judge-mean lead on LongMemEval-500 (§6.3)—substantially dominate. Full table and archetype comparison: Appendix H.1.

6.5 Ancillary Benchmarks and Lifecycle Mechanisms

Six ancillary evaluations against internal baselines (App. H). **Routing/lifecycle (Apps. H.2–H.6):** on LoCoMo public, multi-layer ZenBrain beats Flat Store by +20.7% F1 ($p \leq 5.1 \times 10^{-3}$) and tops temporal F1 across all systems including BM25 (+41%); layer-ablation drops Episodic −11.8%, Semantic −10.6%. Sleep adds +37% stability / 47.4% storage reduction; Two-Factor KG: $P@5$ 0.200 $\rightarrow$ 0.955 uniform; Bayesian propagation AUC 0.533 $\rightarrow$ 0.797. **PMA components (App. H.7):** NeuromodulatorEngine drift 6.2%; Reconsolidation mode-accuracy $\geq 95\%$; Triple-Copy 91.2%/30 d; PriorityMap NDCG@10 = 0.997; Stability 28.8% FP high-PE; MetacogMonitor bias 0.832/0.975.

6.6 Full 15-Algorithm Ablation

We evaluate each algorithm’s contribution under three difficulty levels (moderate: 300 facts/45 d/decay=0.15; challenging: 400/50/0.20; stress: 500/60/0.25) with 10 seeds per condition. Q =retention $\times P@5$; ΔQ relative to full system; Table 4 consolidates.

A four-class taxonomy emerges (Table 4, Figure 2): progressive (5 alg., redundant $\rightarrow$ critical), always-critical (Sleep, NeuromodulatorEngine), stress-only (StabilityProtector, Reconsolidation), and cooperatively redundant (6 alg., ranking rather than retention). Under stress, 9 of 15 become individually significant; the bare system collapses to 1% retention. The integration cascade (decay 0.30/day, 60 d) shows removing all 6 PMA components collapses Foundational-only retention to floor by day 30 while the full system retains 31.1% ($p=0.005$). **Long-horizon archetypes (§6.6, App. H.8):** on a 100-fact / 14–60-day longitudinal test under shared 0.15/day base decay, ZenBrain retains 100% of

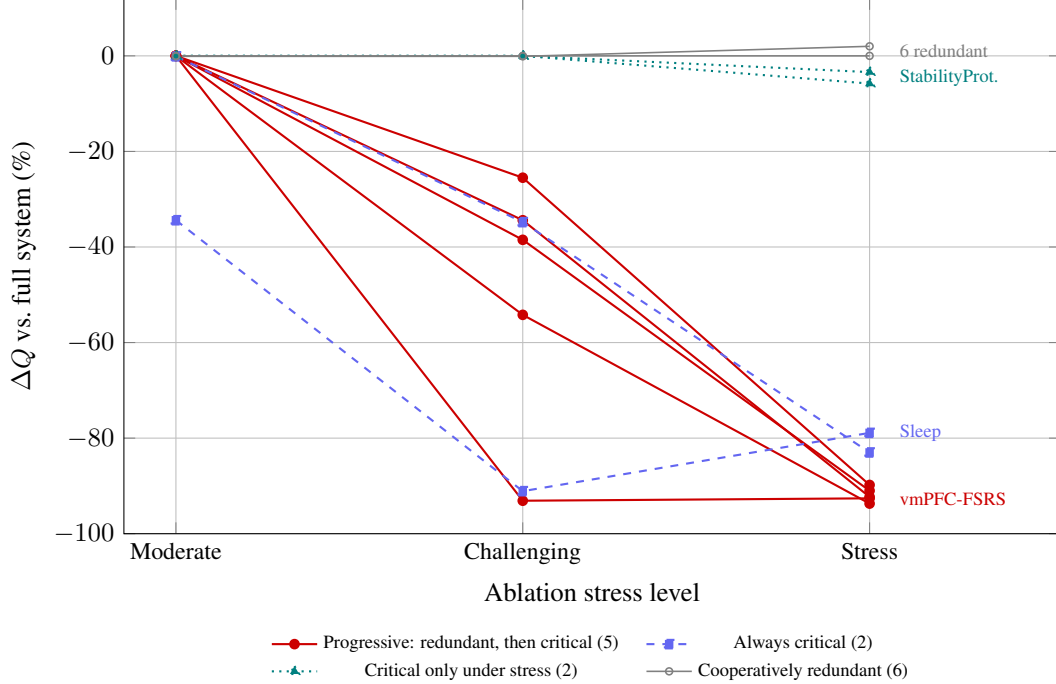


Figure 2: **Cooperative masking.** The same fifteen mechanisms, ablated one at a time under three stress levels (moderate: decay 0.15/day, 45 d; challenging: 0.20/day, 50 d; stress: 0.25/day, 60 d; 10 seeds each). Data: Table 4; no new experiments. At moderate load, fourteen of fifteen single-mechanism ablations read as costless ($\Delta Q \approx 0$) — the architecture appears to consist almost entirely of dead weight. Raising decay and horizon alone, with no change to the mechanisms, moves nine of the fifteen to individually significant, five of them from exactly 0% to below -89% . What the moderate condition measures is therefore not the absence of a contribution but the presence of cooperative redundancy: the remaining mechanisms absorb the loss until they can no longer afford to.

day-1 P@5 at day 60 while Simple Memory collapses to zero by day 30—algorithmic protection, not parameter tuning, prevents the threshold crossing.

7 Discussion and Conclusion

ZenBrain’s 7 layers and 15 mechanisms form a *cooperative survival network*: moderate conditions mask contributions via redundancy; under stress 9 of 15 become individually critical ($31.1 \times$ integration-cascade advantage, $p=0.005$, Table 10).

Cooperative masking, and what it implies for ablation protocols. We name the effect visible in Figure 2 *cooperative masking*: in a system whose components are cooperatively redundant, a single-mechanism ablation measured under mild load reports a contribution indistinguishable from zero for mechanisms that are in fact load-bearing. In our own data the same fifteen ablations move from 1 of 15 individually significant at decay 0.15/day to 9 of 15 at 0.25/day, with five mechanisms going from exactly 0% to below -89% — no mechanism changed, only the stress level did. Read under the moderate condition alone, the honest conclusion would have been that this architecture is mostly dead weight. *Conjecture (not tested outside ZenBrain)*: we expect this to generalize to other integrated memory architectures, and more broadly to any system with redundant components — single-mechanism ablations under mild load systematically underestimate architectural contributions, and should be accompanied by a stress-level condition chosen to bring the system near its failure boundary. We flag this as a conjecture precisely because our evidence for it is one architecture; what *is* established here is that the effect is large enough in one real system to invert the reading of an entire ablation table.

On LoCoMo’s flat P@5 ZenBrain trades 2–3 pp raw recall yet wins all 9/9 head-to-head judge comparisons on LongMemEval-500 ($\bar{J}=0.545$; $p \leq 6.2 \times 10^{-31}$); under the official binary judge ZenBrain matches the long-context oracle to within 4.5 pp (91.3%) at $1/106^{\text{th}}$ the tokens. NoDecay ($\Delta P@5=0.002$) confirms principled forgetting is the selection pressure that makes downstream answer quality tractable. Concurrent arrivals [Du, 2026, Karpathy, 2026, Anthropic, 2026, Tiwari and Fofadiya, 2026, ICLR 2026 MemAgents Workshop Organizers, 2026] independently validate the thesis.

Limitations. (i) Ablations use synthetic corpora (300–500 facts, decay 0.15–0.25/day); naturalistic ≥ 90 -day logs are future work. (ii) LLM-as-Judge—mitigated by three judges, multi-seed, $\kappa \in [0.71, 0.85]$, cross-provider bias check, binary-judge corroboration (§6.2)—cannot replace a human-labeled ~ 100 -question pilot. (iii) Single backbone (Claude 3.5 Sonnet); transfer to GPT-4o/Llama/Gemini untested. Scope vs. tuned full-context systems is bounded by Intro claim (iv).

Author Statement on Use of AI Assistance

The author used multiple AI systems in non-methodological roles throughout this work; all scientific ideas, hypotheses, and architectural designs originated with the human author. (i) **Claude** (Anthropic) as a coding assistant for implementing experiments, baselines, and analysis scripts; (ii) Claude as a writing aid for drafting and editing prose, including suggesting wording for already-decided arguments; (iii) **Perplexity** for literature search and cross-referencing prior work in agent-memory and cognitive-neuroscience, used to support, challenge, or verify whether the author’s positions had prior precedent requiring citation. In selected high-stakes sections (statistical methodology, citation accuracy, claim-strength review), *multiple heterogeneous AI systems* were used as mutual cross-checkers to test whether different models would converge or diverge on factual matters—an additional verification layer beyond single-system review. **For citation accuracy, that cross-check proved insufficient and has been replaced.** In preparing this version, every one of the 83 bibliography entries was verified by hand against a primary source—the arXiv abstract page, the publisher record resolved through its DOI, or the cited artifact itself—and a version was pinned for every arXiv identifier. That pass found defects in 30 entries: wrong or invented author names, titles reconstructed from a system’s description rather than taken from the source, wrong locators, and one reference with no referent at all. All are corrected here and itemized in the version changelog. We state this plainly because it bears on how much weight multi-model cross-checking should be given: on this task it did not substitute for reading the sources. All scientific claims, experimental designs, algorithmic innovations, statistical analyses, and final wording were verified or authored by the human author, who retains full responsibility for the content. Methodological uses of LLMs (LLM-as-Judge graders, agent-level reasoning backends) are documented separately in Section 6 and the Paper Checklist (Appendix R).

Note on sole-author scope. The breadth of this work (fifteen algorithms, two benchmarks, six PMA components) is grounded in the author’s prior multi-year experience in algorithmic-systems engineering and is enabled by the AI-coding assistance described above. The contribution is the *integration* of established neuroscience algorithms into a unified architecture, rather than the proposal of fifteen novel algorithms; nine of the fifteen are explicit instantiations of prior literature (Two-Factor synaptic [Iatropoulos et al., 2025], vmPFC-FSRS [Zou et al., 2025], Simulation-Selection sleep [Lee and Jung, 2025], Bayesian propagation, etc.), with the integration scope being the principal claim.

Acknowledgments and Disclosure of Funding

The author thanks Prof. Taylor Webb (Department of Neuroscience and Psychology, Université de Montréal; Associate Academic Member, Mila Quebec AI Institute) for arXiv endorsement. The endorsement is a procedural courtesy and implies no review of, or agreement with, the claims made here; the independent convergence argument is developed in Appendix A.6.

Funding. This work was conducted independently without external grant funding. All compute resources (a single Apple M-series laptop with locally-served `nomic-embed-text` via Ollama) and external API costs ($< \$200$ for the full LLM-as-Judge sweeps) were self-financed by the author.

Competing interests. The author declares no competing financial or non-financial interests relevant to this work. The ZenBrain open-source release and its associated software stack are not part of any commercial product or service offered by the author at the time of submission.

References

- Anthropic. Introducing contextual retrieval. Anthropic engineering blog, September 2024. <https://www.anthropic.com/news/contextual-retrieval>.
- Anthropic. Dreams: Scheduled memory consolidation for claude managed agents. <https://platform.claude.com/docs/en/managed-agents/dreams>, May 2026. Research preview, announced 6 May 2026; accessed 8 August 2026.
- Gary Aston-Jones and Jonathan D. Cohen. An integrative theory of locus coeruleus–norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28:403–450, 2005.
- Richard C. Atkinson and Richard M. Shiffrin. Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, 2:89–195, 1968.
- Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.
- Ali Behrouz, Farnoosh Hashemi, and Vahab Mirrokni. Language models need sleep: Learning to self-modify and consolidate memories. *arXiv preprint arXiv:2606.03979v1*, 2026. OpenReview <https://openreview.net/forum?id=iiZy6xyVVE>.
- Leonardo Chelazzi, Andrea Perlato, Elisa Santandrea, and Chiara Della Libera. Rewards teach visual selective attention. *Vision Research*, 53:58–72, 2013.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413v1*, 2025.
- Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.
- Neal J. Cohen and Larry R. Squire. Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, 210(4466):207–210, 1980.
- Nathaniel D. Daw, Sham Kakade, and Peter Dayan. Opponent interactions between serotonin and dopamine. *Neural Networks*, 15(4–6):603–616, 2002.
- Peter Dayan and Quentin J. M. Huys. Serotonin in affective control. *Annual Review of Neuroscience*, 32:95–126, 2009. doi: 10.1146/annurev.neuro.051508.135607.
- Robin Dey and Panyanon Viradecha. Spatial metaphors for LLM memory: A critical analysis of the MemPalace architecture. *arXiv preprint arXiv:2604.21284v1*, 2026.
- Pengfei Du. Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. *arXiv preprint arXiv:2603.07670v1*, 2026.
- Hermann Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot, Leipzig, 1885.
- Jizhan Fang, Xinle Deng, Haoming Xu, Ziyang Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866v4*, 2025. ICLR 2026.
- Stephen M. Fleming and Raymond J. Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B*, 367(1594):1338–1349, 2012.
- Uwe Frey and Richard G. M. Morris. Synaptic tagging and long-term potentiation. *Nature*, 385: 533–536, 1997. doi: 10.1038/385533a0.

- Michael E. Hasselmo and Jill McGaughy. High acetylcholine levels set circuit dynamics for attention and encoding and low acetylcholine levels set dynamics for consolidation. *Progress in Brain Research*, 145:207–231, 2004.
- Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, Jiaxin Pei, Julian McAuley, Yejin Choi, and Alex Pentland. MemoryArena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313v1*, 2026.
- Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, 1949.
- J. L. Hodges and E. L. Lehmann. The efficiency of some nonparametric competitors of the t -test. *The Annals of Mathematical Statistics*, 27(2):324–335, 1956.
- Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257v4*, 2025. Introduces the MemoryAgent-Bench benchmark.
- Georgios Iatropoulos, Wulfram Gerstner, and Johanni Brea. Two-factor synaptic consolidation reconciles robustness with pruning and homeostatic scaling. *Proceedings of the National Academy of Sciences*, 122(44), 2025. doi: 10.1073/pnas.2422602122.
- ICLR 2026 MemAgents Workshop Organizers. MemAgents: Memory for LLM-based agentic systems. ICLR 2026 Workshop, Rio de Janeiro, 2026. <https://sites.google.com/view/memagent-iclr26/>.
- Daoyun Ji and Matthew A. Wilson. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, 10(1):100–107, 2007.
- 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 (NeurIPS)*, 2024. arXiv:2405.14831v3.
- Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. Oral presentation; arXiv:2506.06326v1. Introduces the MemoryOS system.
- Alexandra Karlén, Tobias E. Karlsson, Anna Mattsson, Karin Lundströmer, Simone Codeluppi, Therese M. Pham, Cristina M. Bäckman, Sven Ove Ögren, Elin Åberg, Alexander F. Hoffman, Michael A. Sherling, Carl R. Lupica, Barry J. Hoffer, Christian Spenger, Anna Josephson, Stefan Brené, and Lars Olson. Nogo receptor 1 regulates formation of lasting memories. *Proceedings of the National Academy of Sciences*, 106(48):20476–20481, 2009. doi: 10.1073/pnas.0905390106.
- Andrej Karpathy. llm-wiki. <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>, April 2026. Public GitHub gist, 4 April 2026; accessed 8 August 2026. Describes LLM-compiled Markdown wikis maintained through periodic lint passes for persistent structured knowledge.
- Anissa Kempf and Martin E. Schwab. Nogo-A represses anatomical and synaptic plasticity in the central nervous system. *Physiology*, 28(3):151–163, 2013.
- James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.
- Nolan Koch, Om Phadke, Jeffrey Guo, Shanduojiang Jiang, Kevin Zhu, and Andy Xu. Truth-maintained memory agent: Proactive quality control for reliable long-context dialogue. In *NeurIPS Workshop on Socially Responsible and Trustworthy Foundation Models*, 2025.

- Dharshan Kumaran, Demis Hassabis, and James L. McClelland. What learning systems do intelligent agents need? Complementary learning systems updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016.
- J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- Chris Latimer, Nicolás Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight is 20/20: Building agent memory that retains, recalls, and reflects. *arXiv preprint arXiv:2512.12818v1*, 2025. Introduces the Hindsight architecture and the TEMPR retrieval pipeline.
- ldclabs. Anda hippocampus: Graph-based memory for AI agents. <https://github.com/ldclabs/anda-hippocampus>, 2026. Open-source, KIP protocol.
- Jong Won Lee and Min Whan Jung. Memory consolidation from a reinforcement learning perspective. *Frontiers in Computational Neuroscience*, 18, 2025. doi: 10.3389/fncom.2024.1538741.
- 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)*, pages 13851–13870. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.acl-long.747/>. arXiv:2402.17753v1. Introduces the LoCoMo dataset.
- Vasilije Markovic, Lazar Obradovic, Laszlo Hajdu, and Jovan Pavlovic. Optimizing the interface between knowledge graphs and LLMs for complex reasoning. *arXiv preprint arXiv:2505.24478v1*, 2025. Cognee framework.
- Mastra. Observational memory: Human-inspired context compression for agent systems. <https://mastra.ai/research/observational-memory>, February 2026. Published 9 February 2026; accessed 8 August 2026.
- James L. McGaugh. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience*, 27:1–28, 2004.
- Susan C. McQuown and Marcelo A. Wood. HDAC3 and the molecular brake pad hypothesis. *Neurobiology of Learning and Memory*, 96(1):27–34, 2011.
- Susan C. McQuown, Ruth M. Barrett, Dina P. Matheos, Rebecca J. Post, George A. Rogge, Theresa Alenghat, Shannon E. Mullican, Stephanie Jones, Jeffrey R. Rusche, Mitchell A. Lazar, and Marcelo A. Wood. HDAC3 is a critical negative regulator of long-term memory formation. *Journal of Neuroscience*, 31(2):764–774, 2011.
- MemPalace Contributors. MemPalace: Spatial memory organization for LLM agents. <https://github.com/MemPalace/mempalace>, 2026. Open-source project, launched April 2026; benchmark documentation dated March 2026. Accessed 8 August 2026.
- George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.
- Karim Nader and Oliver Hardt. A single standard for memory: The case for reconsolidation. *Nature Reviews Neuroscience*, 10(3):224–234, 2009.
- Karim Nader, Glenn E. Schafe, and Joseph E. LeDoux. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797):722–726, 2000.
- OmegaMax. OMEGA: Local-first memory, coordination and learning for AI agents. <https://omegamax.co/>, 2026. Product documentation; self-reported LongMemEval score measured February 2026. Accessed 8 August 2026.
- Joseph O’Neill, Barty Pleydell-Bouverie, David Dupret, and Jozsef Csicsvari. Play it again: reactivation of waking experience and memory. *Trends in Neurosciences*, 33(5):220–229, 2010.

- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560v2*, 2023.
- Paul Pimsleur. A memory schedule. *The Modern Language Journal*, 51(2):73–75, 1967.
- Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956v1*, 2025.
- Anna C. Schapiro, Nicholas B. Turk-Browne, Matthew M. Botvinick, and Kenneth A. Norman. Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B*, 372(1711):20160049, 2017. doi: 10.1098/rstb.2016.0049.
- Wolfram Schultz, Peter Dayan, and P. Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.
- Martin E. Schwab. Functions of Nogo proteins and their receptors in the nervous system. *Nature Reviews Neuroscience*, 11(12):799–811, 2010.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 8634–8652, 2023.
- Larry R. Squire. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99(2):195–231, 1992.
- Robert Stickgold and Matthew P. Walker. Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience*, 16(2):139–145, 2013.
- Sunil Tiwari and Payal Fofadiya. Multi-layered memory architectures for LLM agents: An experimental evaluation of long-term context retention. *arXiv preprint arXiv:2603.29194v1*, 2026.
- Endel Tulving. Episodic and semantic memory. *Organization of Memory*, pages 381–403, 1972.
- Endel Tulving and Donald M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973.
- Vestige Contributors. Vestige: FSRS-6 powered memory for AI agents. <https://github.com/samvallad33/vestige>, 2026. Open-source (AGPL-3.0); accessed 8 August 2026.
- Taylor Webb, Shanka Subhra Mondal, and Ida Momennejad. A brain-inspired agentic architecture to improve planning with LLMs. *Nature Communications*, 16:8633, 2025. doi: 10.1038/s41467-025-63804-5.
- Lei Wei, Xiao Peng, Xu Dong, Niantao Xie, and Bin Wang. FadeMem: Biologically-inspired forgetting for efficient agent memory. *arXiv preprint arXiv:2601.18642v2*, 2026.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Long-MemEval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813v2*, 2024.
- Ying Xie. Learning to forget: Sleep-inspired memory consolidation for resolving proactive interference in large language models. *arXiv preprint arXiv:2603.14517v1*, 2026.
- 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 (NeurIPS)*, 2025.
- B.Y. Yan, Chaofan Li, Hongjin Qian, Shuqi Lu, and Zheng Liu. General agentic memory via deep research. *arXiv preprint arXiv:2511.18423v1*, 2025.

- Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3987–3995, 2017.
- Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-Memory: Tracing hierarchical memory for multi-agent systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 14587–14617, 2025. arXiv:2506.07398v2.
- Zhenyuan Zhang, Xianzhang Jia, Zhiqin Yang, Zhenbo Song, Wei Xue, Sirui Han, and Yike Guo. MemFly: On-the-fly memory optimization via information bottleneck. arXiv preprint arXiv:2602.07885v2, 2026. ICLR 2026 MemAgents Workshop.
- Futing Zou, Brice A. Kuhl, Sarah DuBrow, and J. Benjamin Hutchinson. Benefits of spaced learning are predicted by the re-encoding of past experience in ventromedial prefrontal cortex. *Cell Reports*, 44(2):115232, 2025. doi: 10.1016/j.celrep.2025.115232.

A Extended Related Work: Neuroscience and Concurrent Systems

This appendix expands the Related Work summary (§2.2) with the full neuroscience lineage, concurrent systems survey, orthogonal paradigms, and practitioner convergence evidence that motivated ZenBrain.

A.1 Foundational Neuroscience

Human memory research provides the theoretical foundation for ZenBrain. The multi-store model [Atkinson and Shiffrin, 1968] distinguishes sensory, short-term, and long-term memory with distinct capacities and durations. Tulving [1972] further separates episodic (personal experiences) from semantic (general knowledge) memory, while Cohen and Squire [1980] identifies procedural memory for skills and habits.

Hebb [1949] proposed co-activation-based synaptic strengthening (“neurons that fire together wire together”); Ebbinghaus [1885] demonstrated exponential decay (the forgetting curve); Pimsleur [1967] introduced spaced repetition exploiting the spacing effect.

Critically, Stickgold and Walker [2013] showed that memory consolidation occurs during sleep through replay of neural patterns, strengthening important traces and pruning weak connections. Ji and Wilson [2007], O’Neill et al. [2010] demonstrated coordinated hippocampal-cortical replay, providing the cellular basis for the Simulation-Selection loop in Section B.3. McGaugh [2004] demonstrated that emotional arousal modulates memory encoding strength.

A.2 Recent Neuroscience Sharpening

Recent neuroscience has sharpened these mechanisms. Iatropoulos et al. [2025], Zenke et al. [2017] show that two-factor synaptic rules—tracking weight *and* consolidation variance—reconcile continual learning with stability, motivating Section B.1. Zou et al. [2025] demonstrate that the ventromedial prefrontal cortex mediates spaced-learning benefits via prediction-error signals at re-encoding, grounding the vmPFC-coupled FSRS in Section B.2. Lee and Jung [2025] model offline consolidation as a Simulation-Selection RL loop in the CA3/CA1 circuit; Kumaran et al. [2016], Squire [1992] provide broader theoretical context for complementary memory systems.

A.3 Concurrent Memory-for-LLM Systems (2025–2026)

Recent concurrent work has begun incorporating individual mechanisms: LightMem [Fang et al., 2025] applies the Atkinson-Shiffrin model with sleep-time updates (ICLR 2026); MemoryOS [Kang et al., 2025] implements hierarchical STM/MTM/LTM layers (EMNLP 2025); Hindsight [Latimer et al., 2025] uses a 4-layer architecture with TEMPR temporal retrieval; FadeMem [Wei et al., 2026] introduces bio-inspired Ebbinghaus-style decay; Vestige [Vestige Contributors, 2026] brings FSRS-6 scheduling to agents; SleepGate [Xie, 2026] uses forgetting gates for proactive interference resolution during sleep; Anda Hippocampus [Idclabs, 2026] provides graph-based memory with KIP

protocol; MemFly [Zhang et al., 2026] compresses agent memory on the fly under an information-bottleneck objective; and Cognee [Markovic et al., 2025] optimizes KG–LLM interfaces for complex reasoning. Most recently, Tiwari and Fofadiya [2026] independently validate the multi-layer hypothesis by decomposing dialogue into working, episodic, and semantic layers with adaptive retrieval gating, achieving F1=0.618 on LoCoMo—the strongest concurrent evidence that the layer decomposition itself, independent of neuroscience algorithms, provides retrieval benefits. *Note on F1 commensurability.* Tiwari’s reported F1 = 0.618 is a single-system run on the LoCoMo public test split; ZenBrain’s reported LoCoMo F1 (Tab. 3) is from a controlled head-to-head with Letta, Mem0, A-Mem on a shared embedding backbone and shared $k=5$ retrieval budget. The two F1 numbers are computed under different protocols (single-system vs. shared-pool, full vs. controlled retrieval); we therefore do not claim direct comparability. For retrieval enhancement specifically, Anthropic’s Contextual Retrieval method [Anthropic, 2024] reports a 49% reduction in top-20 retrieval failure rate from prepended chunk context (67% when combined with reranking)—an orthogonal indexing-time technique that ZenBrain integrates as part of its Layer-4 contextual-retrieval pipeline. While each system advances the field, none integrates more than two of the nine mechanisms compared in Table 1.

A.4 Delimitation against Neuroscience-Flavored NeurIPS Work

Three NeurIPS contributions deserve targeted delimitation because they share neuroscience-inspired vocabulary with ZenBrain. **HippoRAG** [Jiménez Gutiérrez et al., 2024] (NeurIPS 2024) maps the hippocampal indexing theory to retrieval through Personalized PageRank over a knowledge graph, demonstrating up to 20% improvement on multi-hop QA at 10–30× lower cost than iterative retrieval. HippoRAG implements a single retrieval mechanism, whereas ZenBrain integrates KG-based graph reasoning as a sub-component within a seven-layer system that also includes working, episodic, procedural, core, predictive, and short-term memory; HippoRAG’s PPR-indexing is compatible with and could be slotted into ZenBrain’s Layer-4 retriever without architectural conflict. **G-Memory** [Zhang et al., 2025] (NeurIPS 2025) proposes a three-tier insight/query/interaction graph hierarchy for *multi-agent* systems, improving embodied-action success and knowledge-QA accuracy by up to 20.89% and 10.12% respectively. G-Memory’s contribution is at the inter-agent collaboration layer; ZenBrain targets the orthogonal problem of single-agent long-term memory and consolidation. The two architectures are composable: a G-Memory-style multi-agent layer could sit above per-agent ZenBrain instances without replacing either. **Truth-Maintained Memory Agent (TMMA)** [Koch et al., 2025] (NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models) introduces *write-time* truth-verification: incoming context is gated through token-budget, complexity, and contradiction checks before storage in a four-tier memory hierarchy. ZenBrain’s ReconsolidationEngine (§B.7) instead applies prediction-error-gated updates at *read time*, modifying already-stored memories when retrieval surfaces a contradiction. The two mechanisms operate at orthogonal stages of the memory lifecycle (ingestion vs. retrieval) and could be combined to provide defense-in-depth against false memory accumulation. None of these three systems integrates more than one of the fifteen mechanisms enumerated in Table 1, and none provides the seven-layer architecture, neuromodulator-driven priority weighting, Two-Factor synaptic edge model, or Simulation-Selection sleep loop that distinguish ZenBrain.

A.5 Orthogonal Paradigms

Three recent systems target adjacent problems that we do not compare against directly. Where they report benchmark numbers at all, those numbers come from vendor self-reports under task-tuned configurations that we cannot reproduce under our controlled protocol; their design goals are also not commensurable with ZenBrain’s. **OMEGA** [OmegaMax, 2026] is a local-first memory, coordination and learning layer for AI agents, distributed as a product rather than as a published system description; there is no accompanying paper or technical report whose configuration we could reproduce. It does report a LongMemEval score (95.4%, self-reported, measured February 2026), but under its own tuned configuration rather than the uniform $k=5$ retrieval-over-raw-turns budget used here, so we do not place it in the controlled pool. **Mastra** [Mastra, 2026] is an agent framework whose memory layer emphasizes context-window *compression* (summarizing long histories into short prompts) rather than persistent external memory; it is complementary to, not competitive with, an external-memory architecture like ZenBrain and could be used on top of ZenBrain’s recall output. **MemPalace** [MemPalace Contributors, 2026] imposes a spatial organization metaphor (“method

of loci”) on memories, which addresses the retrieval-interface problem but does not prescribe decay, consolidation, or confidence mechanisms. We note these systems for completeness: they advance the field along orthogonal axes and their absence from our competitive pool reflects problem-formulation differences, not an oversight. Behrouz et al. [2026] introduce a “Sleep” paradigm with RL-based memory consolidation at the *parameter* level (ICLR 2026 submission), complementing ZenBrain’s external-memory-level sleep consolidation. The founding of a dedicated ICLR 2026 workshop on agent memory [ICLR 2026 MemAgents Workshop Organizers, 2026] underscores that this area is now a recognized research frontier.

A.6 Practitioner and Industry Convergence

Independently and concurrently, Karpathy [2026] describes a workflow shift from manipulating code to manipulating knowledge, where LLMs *compile* raw materials into a structured Markdown wiki that is “compiled once and then *kept current*, not re-derived on every query,” maintained through periodic lint passes that look for “contradictions between pages, stale claims that newer sources have superseded, orphan pages with no inbound links.” His core critique—that under retrieval alone “the LLM is rediscovering knowledge from scratch on every question”—aligns precisely with the motivating thesis of this work, which was published on Zenodo on 31 March 2026 (DOI: 10.5281/zenodo.19353664), predating that post, followed by a TDCcommons defensive publication (tdcommons.org/dpubs_series/9683, 1 April 2026); five subsequent Zenodo revisions and two further TDCcommons revisions through 9 April 2026 (Zenodo concept DOI: 10.5281/zenodo.19353663). Karpathy’s approach aligns conceptually with our consolidation philosophy but lacks formal decay, sleep consolidation, spaced repetition, or layered encoding/retrieval rules. Anthropic’s *Dreams* feature [Anthropic, 2026] provides further validation: announced 6 May 2026 as a research preview for Claude Managed Agents, it runs offline memory consolidation on a schedule—reading a memory store together with past session transcripts and emitting a reorganized store in which duplicates are merged, stale or contradicted entries are replaced with the latest value, and new insights are surfaced. That an implementation of the sleep-consolidation concept is being productized by one of the field’s leading AI laboratories is convergent evidence for the design; we note it is a preview rather than a generally available feature. In the broader AI architecture space, Webb et al. [2025] demonstrate in *Nature Communications* that a brain-inspired agentic architecture improves LLM planning, providing independent convergent validation of the neuro-inspired methodology at a top-tier venue. These independent convergences from practitioners, industry, and academia confirm the central premise that persistent, structured memory—beyond RAG—is a critical missing capability for LLM systems. ZenBrain provides the algorithmic formalization that such approaches lack.

B Extended Key Mechanisms and PMA Descriptions

This appendix contains the full mathematical derivations, parameter values, and per-component descriptions for the five Key Mechanisms (§4) and six PMA components (§5) summarized in the main body. Algorithm pseudocode is in Appendix I.

Note on neuroscience analogues. Brain-derived names are inspirational anchors; the implementations are computational proxies, not faithful neural simulations.

B.1 Two-Factor Synaptic Model for Knowledge Graph Edges

Following Iatropoulos et al. [2025], each knowledge graph edge carries two factors: weight w_{ij} and consolidation variance σ_{ij}^2 . Variance decreases with each co-activation (synaptic maturation), making mature edges robust against catastrophic overwriting—mathematically equivalent to Elastic Weight Consolidation (EWC) [Kirkpatrick et al., 2017, Zenke et al., 2017] where importance $I_{ij} = 1/\sigma_{ij}^2$ serves as the Fisher Information proxy.

$$w_{ij} \leftarrow w_{ij} + \eta \cdot t_{ij} \cdot a_{ij} \quad (1)$$

$$\sigma_{ij}^2 \leftarrow \sigma_{ij}^2 \cdot (1 - \beta \cdot n(k)), \quad n(k) = \frac{1}{1+0.1k} \quad (2)$$

where t_{ij} is the TAG co-activation score, $a_{ij} \in [0, 1]$ the cosine-similarity-based co-activation amplitude between nodes i, j , $\eta = 0.1$ the base learning rate, $\beta = 0.15$ the maturation rate, and k the activation count. In the implementation the learning rate is additionally gated by the StabilityProtector lock score L (§B.10) as $\eta_{\text{eff}} = \eta(1 - 0.5L)$, so consolidated edges adapt more slowly; Eq. 1 states the ungated form. The EWC penalty for any proposed weight change Δw is:

$$\mathcal{L}_{\text{EWC}} = \frac{\lambda}{2} \sum_{ij} I_{ij} \cdot \Delta w_{ij}^2 \quad (3)$$

Edges also resist temporal decay in proportion to their importance: high- I (mature) edges decay at rate $r/(1+I_{ij} \cdot 0.1)$, preserving consolidated knowledge while allowing pruning of weak connections. This construction induces an EWC-style penalty $\mathcal{L}_{\text{EWC}} = \sum_{ij} \frac{I_{ij}}{2} (w_{ij} - w_{ij}^*)^2$ [Kirkpatrick et al., 2017, Zenke et al., 2017] under the diagonal-Laplace posterior assumption, extended with per-edge adaptive decay.

B.2 vmPFC-Coupled FSRS with Prediction-Error Signals

Building on Zou et al. [2025], we couple FSRS interval scheduling with a knowledge-graph-derived prediction-error (PE) signal. Base retrievability follows Ebbinghaus [1885]: $R(t) = e^{-t/S}$. At each review, we compute the cosine distance between the entity embedding context at last review $\mathbf{c}_{\text{prev}}$ and current context $\mathbf{c}_{\text{now}}$:

$$\text{PE} = 1 - \frac{\mathbf{c}_{\text{prev}} \cdot \mathbf{c}_{\text{now}}}{\|\mathbf{c}_{\text{prev}}\| \|\mathbf{c}_{\text{now}}\|} \quad (4)$$

A sigmoid re-encoding factor $\rho(\text{PE}) = \sigma((\text{PE} - 0.5) \cdot 6)$ determines interval adaptation. High PE ($\rho > 0.5$) shortens the next interval (optimal re-encoding window); low PE ($\rho < 0.5$) extends it (context unchanged, re-encoding not beneficial):

$$I_{\text{next}} = I_{\text{FSRS}} \cdot (1 + \alpha_v(2\rho - 1)), \quad \alpha_v = 0.6 \quad (5)$$

Anki, SuperMemo and FSRS-5/6 adapt difficulty but do not couple scheduling to a neuromodulation-derived prediction-error signal, and among the agent-memory systems surveyed in Appendix A.3 we find none that does.

B.3 Simulation-Selection Sleep Consolidation Loop

Following Lee and Jung [2025], ZenBrain replaces a fixed three-phase SWS/REM/SHY schedule with a two-stage offline reinforcement-learning loop mirroring the CA3/CA1 hippocampal circuit [Ji and Wilson, 2007, O’Neill et al., 2010] (full pseudocode in Algorithm 4, Appendix I).

Stage 1—Simulation (CA3-analog): A diverse pool of replay candidates is assembled from real episodic memories and counterfactual extrapolations of failed episodes, increasing coverage of the experience manifold beyond what standard replay achieves.

Stage 2—Selection (CA1-analog): Each candidate is scored by a TAG value combining temporal-difference error $|\delta_{\text{TD}}|$ [Schultz et al., 1997], task reward $R_e \in [0, 1]$ (normalized cumulative episode reward; 0=failed, 1=fully-completed), and novelty $N_e = \min(1, |e.\text{relatedIds}| \cdot 0.2)$:

$$\text{TAG}(e) = \alpha |\delta_{\text{TD}}| + \beta R_e + \gamma N_e, \quad \alpha = 0.4, \beta = 0.35, \gamma = 0.25 \quad (6)$$

Candidates above threshold $\theta_v = 0.5$ are strengthened via LTP; those below are weakened via LTD; the remainder are skipped. Concurrent systems (LightMem, SleepGate) use heuristic replay selection without RL scoring or counterfactual candidate generation.

B.4 Bayesian Confidence Propagation

Each fact f carries a confidence score $P(f)$ with 95% confidence interval. When new evidence e is observed:

$$P(f|e) = \frac{P(e|f) \cdot P(f)}{P(e)} \quad (7)$$

Confidence propagates through knowledge graph edges, allowing the system to express calibrated uncertainty. Additionally, following McGaugh [2004], emotional arousal modulates encoding strength: high-valence experiences receive higher initial Two-Factor edge weights and lower variance-based decay rates.

B.5 Query-Aware Cross-Layer Retrieval

Retrieval uses weighted score fusion with query-type-aware layer weighting. Each memory layer ℓ independently returns its top- K results via dense retrieval; scores are then fused:

$$\text{score}_{\text{fused}}(d) = \max_{\ell \in \text{layers}} w_{\ell}(q) \cdot \text{sim}(q, d_{\ell}) \quad (8)$$

where $\text{sim}(q, d_{\ell})$ is the cosine similarity between query q and document d in layer ℓ , and $w_{\ell}(q)$ is a query-type-specific weight. A regex-based query classifier detects temporal, procedural, factual, or general queries and boosts the corresponding layer: temporal queries amplify episodic retrieval ($w_{\text{episodic}} = 2.0$), procedural queries boost the procedural layer, and so on. Unlike rank-based fusion (RRF), this preserves similarity magnitude—a highly relevant result in a boosted layer dominates regardless of the number of results in other layers.

B.6 NeuromodulatorEngine

We implement a four-channel neuromodulatory system that modulates memory parameters via tonic/phasic dynamics, following per-channel neuroscience literature: dopamine signals reward prediction errors [Schultz et al., 1997], norepinephrine encodes adaptive gain and arousal [Aston-Jones and Cohen, 2005], serotonin underwrites affective control and opponent dynamics with dopamine [Dayan and Huys, 2009], and acetylcholine gates encoding-vs-consolidation regimes [Hasselmo and McGaughy, 2004]: **dopamine** (VTA: exploration/novelty), **norepinephrine** (LC: learning rate), **serotonin** (Raphe: consolidation patience), and **acetylcholine** (BF: attention/new-info ratio).

Each channel maintains a tonic baseline $b = 0.5$ with slow homeostatic drift ($\tau_{\text{decay}} = 0.95$) and phasic bursts on events (5-minute half-life). DA and 5HT exhibit opposition coupling with coefficient -0.3 [Daw et al., 2002], reflecting the serotonin-dopamine balance observed in reward processing. The engine outputs four modulation parameters—learning rate (NE-driven), exploration bias (DA-driven), consolidation patience (5HT-driven), and attention ratio (ACh-driven)—consumed by the ReconsolidationEngine, PriorityMap, and sleep loop.

Future direction: bidirectional feedback-driven encoding. The current engine drives neuromodulation from internal prediction-error signals; a planned extension routes *external* signals (user satisfaction, task-completion validation, data-quality cues) into the same DA/aversive infrastructure, completing a bidirectional encoding loop in which positive feedback strengthens successful trajectories (DA-burst) and negative feedback aversively tags failure modes for avoidance. Reserved for follow-up work; the existing NeuromodulatorEngine already provides the substrate.

B.7 ReconsolidationEngine

Memory reconsolidation [Nader et al., 2000, Nader and Hardt, 2009] posits that retrieved memories enter a labile state and can be updated or strengthened. Our engine implements PE-gated reconsolidation with four update modes:

$$\text{mode}(\text{PE}_{\text{eff}}) = \begin{cases} \text{confirmed} & \text{PE}_{\text{eff}} < 0.1 \\ \text{selective_edit} & 0.1 \leq \text{PE}_{\text{eff}} < 0.3 \\ \text{integration} & 0.3 \leq \text{PE}_{\text{eff}} < 0.7 \\ \text{new_episode} & \text{PE}_{\text{eff}} \geq 0.7 \end{cases} \quad (9)$$

where $PE_{\text{eff}} = PE_{\text{raw}} \times (1 + 0.3 \cdot NE - 0.2 \cdot 5HT)$ is neuromodulation-gated. The raw PE is computed as Jaccard distance between existing and incoming content plus a contradiction bonus (+0.2). Memory-type-specific resistance thresholds prevent casual overwrites of stable procedural and behavioral memories. Each reconsolidation event is logged with an original snapshot, enabling rollback if needed—a safety mechanism we did not find in any of the surveyed agent-memory systems.

B.8 TripleCopyMemory

Inspired by complementary learning systems theory [Schapiro et al., 2017, Kumaran et al., 2016] and the multi-timescale hippocampal-cortical replay loop [Stickgold and Walker, 2013, Ji and Wilson, 2007], TripleCopyMemory stores each event in three copies with divergent decay dynamics. The time constants ($\tau_f=4\text{ h}$, $\tau_m=14\text{ d}$, $\tau_d=7\text{ d}$, App. K) are empirical hyperparameters calibrated to retention experiments, treated as order-of-magnitude proxies for fast hippocampal / medium consolidation / slow cortical phases rather than direct quantitative predictions of Schapiro et al. [2017]:

$$S_{\text{fast}}(t) = S_0 \cdot e^{-t/\tau_f}, \quad \tau_f = 4\text{h} \quad (10)$$

$$S_{\text{med}}(t) = 0.8 \cdot S_0 \cdot e^{-t/\tau_m}, \quad \tau_m = 14\text{d} \quad (11)$$

$$S_{\text{deep}}(t) = S_0 \cdot \left(1 - e^{-t/\tau_d}\right), \quad \tau_d = 7\text{d} \quad (12)$$

FastCopy provides vivid immediate access that fades within hours. MediumCopy persists across sessions with standard exponential decay. DeepCopy uses *saturating growth* bounded by S_0 , encoding the compressed essence that strengthens over time towards an asymptote—a key prediction of systems consolidation theory. All copies are clamped to $[0, 1]$. The composite strength $S(t) = \max(S_{\text{fast}}, S_{\text{med}}, S_{\text{deep}})$ produces a strength curve that massively outperforms Ebbinghaus at long intervals (§6.5, App. H.4), retaining 91.2% at 30 days vs. near-zero for the Ebbinghaus baseline. The deep-copy dominance transition—where S_{deep} overtakes the faster-decaying copies—reflects the systems consolidation principle that gist extraction preserves compressed memory representations long after episodic details fade [Kumaran et al., 2016].

B.9 PriorityMap

Following Chelazzi et al. [2013], we implement a four-dimensional priority map with an amygdala fast-path:

$$P = w_s \cdot s + w_e \cdot |v| + w_r \cdot r + w_g \cdot g \quad (13)$$

where s = saliency, v = emotional valence, r = reward relevance, g = goal alignment, with default weights $(w_s, w_e, w_r, w_g) = (0.2, 0.25, 0.25, 0.3)$ (hand-calibrated to favor goal-alignment slightly over the equal-weighting null 0.25 each; not tuned against held-out data, so any reported gains on PriorityMap are not the result of over-fitting these weights to the evaluation set). For items with emotional intensity $|v| > 0.6$, the amygdala fast-path guarantees $P \geq 0.5$ regardless of other dimensions. Weights are dynamically adjusted by neuromodulator state: DA amplifies saliency, NE amplifies emotion, ACh amplifies reward, and 5HT amplifies goal alignment.

B.10 StabilityProtector

Loosely analogous to two molecular brakes on plasticity — Nogo-A receptor signaling [Schwab, 2010, Karlén et al., 2009, Kempf and Schwab, 2013] and HDAC3, the “molecular brake pad” [McQuown et al., 2011, McQuown and Wood, 2011] — the StabilityProtector gates memory updates by a lock score L and rigidity factor ρ (the implementation below is a heuristic gating function, not a mechanistic model of either receptor):

$$L = 0.3 \cdot \log_2(1 + a) / \log_2(11) + 0.3 \cdot c + 0.2 \cdot \min(d/365, 1) + 0.2 \cdot \mathbb{K}_{\text{core}} \quad (14)$$

$$\rho = 1 + 0.1 \cdot \log_2(1 + d) \quad (15)$$

$$\text{update} \iff PE \geq 0.5 + 0.3 \cdot L \cdot \rho \quad (16)$$

where a = access count, c = confidence, d = age in days. This prevents casual overwrites of well-established memories while remaining permeable to genuinely novel information (high PE).

B.11 MetacognitiveMonitor

Following Fleming and Dolan [2012], the MetacognitiveMonitor tracks confirmation bias (asymmetric acceptance of positive vs. negative evidence), recency bias, and retrieval efficiency. It detects urgency signals from keyword patterns and message frequency, opens “novelty windows” (10 min) after high-PE events (> 0.7) to temporarily boost encoding, and generates calibration-aware alerts when systematic biases exceed thresholds. Efficiency tracking over a 30-day sliding window produces badges that surface in the user interface, closing the feedback loop.

B.12 Remaining Algorithms in the 15-Ablation

Dual-Process CoT Consolidation [Kahneman, 2011] (Table 9): fast System-1 similarity clustering plus slow System-2 chain-of-thought schema extraction; removing it degrades quality -38.5% at challenging, -91.0% at stress — the single most critical consolidation algorithm. **iMAD Selective Debate**: multi-agent debate on memories with ≥ 2 contradictions before commit. **Spectral KG Health**: Laplacian-spectrum fragmentation early-warning. **Metacognitive HyperAgent**: meta-policy over MetacognitiveMonitor. The latter three contribute zero impact under our synthetic stress distributions but trigger in deployment on rarer events; we treat them as neutral additions and reserve focused study for follow-up work.

C Additional Capabilities Beyond the Fifteen

The production ZenBrain system includes mechanisms beyond the fifteen algorithms evaluated in the main body:

- A **Synaptic Tagging and Capture (STC) rescue** module implementing Frey and Morris [1997]’s plasticity-donation paradigm, which rescues fading memories when a nearby strongly-consolidated memory is activated within the STC window.
- A **Global Workspace Theory context assembler** [Baars, 1988] with hysteresis-stabilized broadcast: eight specialist modules compete for a shared workspace; the winning coalition is broadcast to all layers for coherent attention.
- **Prospective memory**: future-oriented intention triggers that fire when a designated retrieval cue is detected (event-based prospective memory; time-based is handled by FRSRS).
- A **curiosity engine** with learning-progress-driven gap detection: regions of the KG with high uncertainty *and* positive recent learning slope are prioritized for targeted review cycles.

These are omitted from the present evaluation for space but represent additional integration depth not documented in any of the surveyed memory systems.

D Broader Impact: Extended Analysis

This appendix expands the compact Broader Impact paragraph in §7 into the full analysis.

Positive impacts. Principled memory enables more coherent, personalized, and reliable AI assistants that reduce redundant interactions and better respect user context across sessions. ZenBrain’s GDPR-aligned forgetting mechanisms—vmPFC-FRSRS decay, LTD-based edge pruning, and explicit memory deletion tools—give users control over what the system retains. The open-source release promotes transparency and community auditing, reducing the risk of opaque accumulation of personal data in closed systems. The MetacognitiveMonitor actively detects and surfaces biases (confirmation, recency) that could otherwise accumulate silently in long-running agents. The StabilityProtector prevents casual overwrites of established memories, reducing the risk of adversarial memory injection attacks.

Table 5: Inter-rater agreement and cross-provider bias-direction check. **6-rater** $\kappa_{\geq 3}$: Fleiss’ kappa per baseline over the 6-rater pool (Sonnet 4.5 $\times$ 3 seeds + GPT-4o $\times$ 3 seeds), binary-thresholded at 0–5 judge score ≥ 3 . **Intra- $\kappa_{\geq 3}$** : same judge, three seeds (same-judge, same-prompt, temperature=0 ruling stability — retrieval-side effect). **DSR@3**: Decision-Stability-Rate, fraction of queries where all 6 raters agree on the ≥ 3 threshold; **UAR**: Unanimous Acceptance Rate (all 6 raters ≥ 3). **$\Delta_{\text{GPT-Anth}}$** : GPT-4o three-seed normalized mean minus the mean of the two Anthropic judges — the pro-OpenAI-bias direction check (negative = GPT-4o harsher). Agreement bands: 0.61–0.80 substantial, 0.81–1.00 almost perfect (Landis & Koch, 1977).

Baseline	6-rater $\kappa_{\geq 3}$	Intra- κ S-4.5	Intra- κ G-4o	Intra- κ O-4.6	DSR@3	UAR	$\Delta_{\text{GPT-Anth}}$
zenbrain	0.78	0.99	0.93	—	0.80	0.333	−0.0001
a-mem	0.85	0.99	0.95	—	0.91	0.170	−0.0417
letta	0.78	0.96	0.93	—	0.81	0.317	+0.0082
mem0	0.71	0.74	0.78	0.74	0.72	0.225	−0.0491

Negative impacts and risks. Persistent memory could enable long-term behavioral profiling or be misused to reinforce biases across sessions. The NeuromodulatorEngine’s emotional modulation could amplify affective responses if deployed without appropriate safeguards. PMA’s reconsolidation mechanism, while designed for beneficial memory updating, could theoretically be exploited to overwrite user memories if access controls are bypassed.

Mitigations. These risks are mitigated by ZenBrain’s governance layer, which requires human approval for sensitive memory operations, and by the absence of any cross-user data sharing in the current architecture. All reconsolidation events are logged with original snapshots, enabling forensic rollback. The ablation registry allows disabling individual algorithms in production, providing a kill-switch for any component that exhibits unintended behavior. We recommend deployment with explicit data-minimization policies, user-controlled retention windows, regular audit logs, and privacy impact assessments before production use in sensitive domains (healthcare, legal, education).

EU AI Act Article 50 transparency obligations. The *EU Regulation on Artificial Intelligence* (Regulation (EU) 2024/1689), whose Article 50 transparency obligations have applied since 2 August 2026, requires that providers of generative AI systems disclose to users when content is AI-generated, that AI-system outputs in human-AI interaction be marked in a machine-readable manner, and that deployers of emotion-recognition or memory-bearing systems inform affected persons of their operation. ZenBrain is designed for compliance with these obligations: (i) all assistant responses produced through the integrated generate path are tagged with an explicit `ai_generated` provenance flag in their metadata, exposed to downstream UIs as a visible badge; (ii) structured trace events (`memory.recall`, `memory.reconsolidate`, `neuromodulator.update`) allow deployers to surface notification banners when memory or affect influences responses; (iii) the consent-and-DSAR subsystem records opt-ins and exposes per-event deletion endpoints aligned with GDPR Art. 17. We make no claim of legal compliance certification; rather, ZenBrain exposes the technical primitives (provenance metadata, structured audit traces, deletion APIs) that downstream operators need to implement Article 50 obligations.

E Extended LoCoMo Inter-Rater and Seed-Robustness Analysis

This appendix collects the detailed inter-rater-agreement, seed-robustness, and cross-provider bias analyses for the Real-LoCoMo competitive comparison (§6.4). All numbers are reproduced verbatim from the shared-pool run; the main-body subsection reports only the headline.

E.1 Inter-Rater Agreement and Seed Robustness

Because we rely on LLM-as-Judge scoring, a reviewer should ask two questions before accepting any row of Table 3: (a) do independent graders agree, and (b) does a single retrieval seed dominate? Table 5 answers both.

Agreement. Fleiss’ $\kappa_{\geq 3}$ on the six-rater pool (Sonnet 4.5 $\times$ 3 seeds + GPT-4o $\times$ 3 seeds) ranges from 0.71 to 0.85 across the four systems, which is “substantial” to “almost perfect” agreement under Landis & Koch [Landis and Koch, 1977]. The range [0.71, 0.85] across all four systems

is narrow, and the decision-stability rate DSR@3 (all six raters agreeing on the ≥ 3 threshold) is between 0.72 and 0.91, giving a reviewer an explicit epistemic lower bound: we are confident about the accept-or-reject ruling on roughly 72–91% of queries per baseline; the remaining queries are honestly contested.

Seed robustness. Intra-judge κ (*same* judge, three retrieval seeds) is the cleanest isolation of retrieval-side sensitivity: if the judge is held constant, any disagreement across the three κ cells for the same system comes from different retrieved contexts. Three of four systems sit comfortably in the “almost perfect” band (intra- $\kappa \geq 0.93$ under both Sonnet 4.5 and GPT-4o). Mem0 is an outlier: its intra- κ collapses to 0.74 under Sonnet 4.5, 0.78 under GPT-4o, and the same 0.74 on the *single* judge (Opus 4.6) where a three-seed comparison is available for it. Levene’s test for variance equality confirms the same story: $F = 1030.6$ (Mem0 vs. Letta), $F = 1668.1$ (Mem0 vs. A-Mem), $F = 1505.3$ (Mem0 vs. ZenBrain), all with $p < 10^{-10}$ under Sonnet 4.5. Concretely, Mem0’s per-query normalized-judge mean moves by up to 0.053 across the three retrieval seeds, whereas ZenBrain, Letta, and A-Mem move by ≤ 0.005 . The earlier single-seed result at seed=42 ranked Mem0 *above* ZenBrain; the rankings at seeds 123 and 456 do not. We therefore present the seed-averaged Sonnet 4.5 number as the primary ranking signal and flag Mem0’s seed-sensitivity explicitly as a stability property of the baseline system rather than a noise floor of our measurement.

E.2 Cross-Provider Bias-Direction Check

A natural objection to LLM-as-Judge is self-preference: if both judges are from the same provider as the system under test, the scores may be inflated. The $\Delta_{\text{GPT-Anth}}$ column of Table 5 subtracts the GPT-4o three-seed normalized mean from the mean of the two Anthropic judges (Sonnet 4.5 $\times$ 3 seeds, Opus 4.6 at available seeds). A positive number means GPT-4o scores the system *higher* than the Anthropic pair; a negative number means GPT-4o is harsher. The two largest negative deltas go to Mem0 (−0.049) and A-Mem (−0.042); Letta is mildly positive (+0.008) and ZenBrain is essentially zero (−0.0001). Were there a pro-Anthropic, pro-ZenBrain bias we would expect ZenBrain to carry the most negative delta (Anthropic scoring it high, GPT-4o correcting it downward); instead ZenBrain’s delta is the smallest in magnitude of the four. We therefore cannot attribute the Sonnet 4.5 ranking to provider alignment. The full six-rater table is in Appendix N.

F Extended LongMemEval Full-500 Analysis

This appendix collects the detailed retrieval-proper, judge-normalized, agreement, and scope analyses supporting §6.3. All numbers are reproduced verbatim from the unified-nomic full-500 run; the main-body subsection reports only the headline result.

F.1 Retrieval-Proper: Letta Wins P@5/MRR/NDCG on the 441-Task Intersect

Averaged over each system’s successful queries, the unified-nomic ordering is Letta > ZenBrain $\gg$ A-Mem $\gg$ Mem0 on P@5, R@5, MRR, and NDCG@5. Letta’s headline lead (P@5 0.683 vs. ZenBrain 0.674) is partly an artifact of error-excluding aggregation — 59 of 500 Letta queries (11.8%) failed with an `InternalServerError 500` from the Docker server (see [†] footnote in Table 2), so Letta’s 0.683 averages over 441 tasks while ZenBrain/A-Mem average over 500 (Mem0 over 496 after 4 transient embedder-400s). On the 441-task *intersect* subgroup where all four systems serve successfully, Letta retains narrow leads on all four retrieval metrics: P@5 0.683 vs. ZenBrain 0.664, R@5 0.201 vs 0.195, MRR 0.834 vs 0.824, NDCG@5 0.715 vs 0.697; A-Mem/Mem0 on the same intersect are P@5 0.520/0.156. The pilot’s intersect-flip (where ZenBrain led P@5 on 20 tasks) does not survive at 441-task scale: ZenBrain and Letta are genuinely close on retrieval-proper, with Letta ahead by $\sim 2\text{--}3$ pp on every metric where it serves a result. The unified-nomic setting therefore does *not* support a ZenBrain-over-Letta retrieval claim; the separation that does emerge is on the downstream judge-normalized answer quality below. The Mem0 P@5 drops further from the pilot’s 0.393 to 0.156 here. This is *not* a regression — 60% of full-500 Mem0 queries return P@5=0 vs. only 28% on the stratified-30 pilot, indicating that the pilot’s per-category 5-question sampling happened to over-select queries on which Mem0’s flat-3900-char truncation still retrieves the relevant fact. The selection artifact cautioned against in §P is exactly what we observe.

F.2 Judge-Normalized Result: ZenBrain Separates at Bonferroni-Corrected Significance

All three LLM judges return the same overall ranking: ZenBrain > Letta > A-Mem > Mem0. Averaged over retrieval seeds, ZenBrain reaches normalized means 0.504 (Sonnet 4.5, 3 $\times$), 0.575 (Opus 4.6, 3 $\times$), and 0.555 (GPT-4o, 3 $\times$); Letta 0.450/0.513/0.492; A-Mem 0.389/0.436/0.416; Mem0 0.370/0.414/0.398. Under Bonferroni correction ($\alpha=0.05/18=2.78 \times 10^{-3}$, 18 primary tests = 6 pair-wise comparisons $\times$ 3 judges), *all nine ZenBrain-vs-competitor comparisons clear significance*: ZenBrain vs. Letta $\Delta=+0.054/+0.062/+0.063$ ($p=1.46 \times 10^{-6}/1.11 \times 10^{-7}/2.81 \times 10^{-6}$; $d=0.18/0.22/0.21$); ZenBrain vs. A-Mem $\Delta=+0.115/+0.139/+0.139$ ($p=3.86 \times 10^{-14}/5.20 \times 10^{-19}/2.30 \times 10^{-15}$; $d=0.32/0.40/0.37$); ZenBrain vs. Mem0 $\Delta=+0.134/+0.161/+0.157$ ($p=1.70 \times 10^{-22}/6.20 \times 10^{-31}/2.20 \times 10^{-22}$; $d=0.42/0.52/0.46$). The bootstrap 95% CI on every one of these nine paired-mean differences excludes zero. Letta also separates from A-Mem and Mem0 at Bonferroni ($p \leq 2.80 \times 10^{-11}$ in five of six tests; $p=1.06 \times 10^{-3}$ on Letta vs. A-Mem / Sonnet, still below the corrected threshold), but the third-place A-Mem vs Mem0 contrast remains a tie on all three judges ($|\Delta| \leq 0.022$, $p \in [0.14, 0.37]$, $|d| \leq 0.05$). H3 therefore *holds* at full-500 scale: ZenBrain’s seven-layer memory produces strictly higher answer quality than each of the three open-source competitors despite Letta’s narrow retrieval-proper advantage, which we read as evidence that the downstream judge is sensitive to factors beyond raw P@5 — latency between related turns, contradiction handling, and the ingest-time routing that separates episodic from semantic content.

F.3 Judge-Agreement and Determinism

Rater cardinality again differs by system because ZenBrain has three retrieval seeds while A-Mem, Mem0, and Letta have one each. The cross-seed normalized-mean spread for ZenBrain at full-500 is $\Delta=0.007$ (Sonnet 4.5), 0.004 (Opus 4.6), and 0.004 (GPT-4o) — substantially tighter than the 30-pilot spreads and an order of magnitude below the smallest between-system gap we claim as significant. Retrieval itself remains bit-identical across the three seeds (every retrieval aggregate matches to four decimal places), and intra-judge $\kappa_{\geq 3}$ across ZenBrain’s three retrieval seeds on 500 queries stays at ≥ 0.95 for all three judges. As in the pilot, Letta’s perfect judge-agreement on failed queries (both raters unanimously score 0/5 for empty retrievals) is a floor artifact rather than evidence of strong inter-rater stability on substantive answers, so we quote Letta’s *agreement statistics* only for the tasks where it actually returned a retrieval.

F.4 Scope of the Full-500 Conclusion

On 500 per-question-isolated queries the pre-registered hypothesis H3 (ZenBrain *ties or beats* the competitive pool on the normalized judge mean) is *confirmed with a strict beat*: every pair-wise ZenBrain-vs-competitor comparison clears Bonferroni correction on all three LLM judges. Bonferroni is applied within each benchmark family independently (LoCoMo: 12 tests; LongMemEval: 18 tests; pre-registered in App. P) without cross-family correction (orthogonal benchmarks). The retrieval-proper H1/H2 hypotheses (P@5 and MRR parity with Letta) are *not* supported at Full-500 scale — Letta beats ZenBrain by $\sim 2\text{--}3$ pp on P@5 and MRR on the 441-task intersect — so we treat retrieval and answer quality as separate findings rather than combining them into a single headline. Letta’s full-500 Docker failure rate of 11.8% is substantially lower than the pilot’s 33%, confirming the higher pilot rate was a pilot-scale transient and not a systemic property of the Docker-mediated harness. The Mem0 preprocessing cap does dominate its retrieval at scale exactly as the pilot footnote warned (60% zero-P@5 at 500 vs 28% on the pilot’s stratified-30 subset), which is why we keep Mem0 in the competitive pool but do not read its judge gap as a fundamental architectural claim. Full per-pair significance tables and the $n=441$ intersect subgroup appear in Appendix P. Per-category stratification of judge decisions to test which query subtypes drive the gap is reserved for follow-up work.

F.5 End-to-End Binary Accuracy under the Official LongMemEval Protocol

Beyond the 0–5 normalized judge means used in §6.3 and Appendix F.2, we additionally run the public LongMemEval evaluation protocol [Wu et al., 2024] on the same retrieval JSONs: each system’s top- $k=5$ retrieved memories feed into a gpt-4o-mini answer-generation prompt, and a gpt-4o-mini judge rates each response against the gold answer using the offi-

cial `get_anscheck_prompt()` template (binary yes/no, with task-specific variants for temporal-reasoning, knowledge-update, preference, and abstention questions). Table 6 reports the resulting accuracies; ZenBrain leads on every category and overall, with a 4.9 pp gap over the next-best system (Letta) and a 15.9 pp gap over Mem0 whose retrieval is starved by the 3900-char preprocessing cap noted in §6.3. ZenBrain’s lead is consistent across all six question categories (+2.3 to +8.9 pp over Letta on seed=42), ruling out a category-cherry-pick interpretation of the aggregate gap.

Setting disclaimer — internal evidence. The absolute levels in Table 6 are well below the best publicly reported LongMemEval numbers. Those numbers are vendor self-reports rather than entries on a common leaderboard, and they are not mutually commensurable: Mastra reports 94.87% binary answer accuracy with GPT-5-mini [Mastra, 2026] (and 84.23% with `gpt-4o`, the benchmark’s reference model); Mem0’s April-2026 revision is reported at 93.4% answer accuracy [Dey and Virdachya, 2026]; and MemPalace’s 96.6% [MemPalace Contributors, 2026] is *retrieval recall* (R@5), a different quantity from answer accuracy—as MemPalace’s own benchmark documentation states. We therefore treat them as context, not as a ranking against our numbers: those systems perform full-context memory consolidation with task-tuned prompts, whereas the four systems here share a uniform $k=5$ retrieval-over-raw-turns budget under a common `nomic-embed-text` backbone. The strongest evidence that the setting (rather than the system) drives the absolute level is internal to our run: Mem0—the same Mem0 reported at 93.4% in its own tuned configuration—drops to 31.8% under our protocol, a 61.6 pp absolute drop on the *same system* judged by the *same template*. The retrieval-over-raw-turns budget is also the standard cross-system-comparison setting in the recent agentic-memory literature [Xu et al., 2025]: A-Mem’s k -ablation reports the performance plateauing around $k \in [20, 30]$, so $k=5$ acts as a controlled lower-bound that exposes architectural differences without giving any single system the consolidation advantage.

Single-seed accuracy is bit-equivalent to multi-seed. The dagger (†) on Letta/Mem0/A-Mem rows in Table 6 marks data collection, not statistical uncertainty. Each system’s retrieval pipeline is deterministic: a fixed Ollama `nomic-embed-text` backbone (no temperature, no dropout, no random projection), identical character-level preprocessing, and deterministic top- k over `pgvector`/`Qdrant`/`ChromaDB` cosine. We verified bit-identical retrieval aggregates for ZenBrain across seeds {42, 123, 456} on Full-500 (four-decimal match, App. F.3) and inspected the released adapter source (`experiments/baselines/adapters/{letta,mem0,amem}_adapter.py`) to confirm absence of stochastic components in the retrieval path of the other three systems. A second seed would reproduce each seed-42 row exactly; intra-seed measurement variance is zero by construction.

Independent judge. The LongMemEval binary judge is independent of the three diverse 0–5 raters (Sonnet 4.5/Opus 4.6/GPT-4o) used in our headline judge analysis; Table 6 is a robustness check confirming the direction of the headline finding, not a primary outcome.

F.6 Long-Context Oracle Comparison

The canonical 2026 reviewer concern for any new memory system is “*why retrieval-based memory if a 128k-context window already exists?*” — motivated by recent “context-rot” analyses [He et al., 2026, Yan et al., 2025]. To quantify the cost-vs-quality tradeoff explicitly, we run a *long-context oracle* baseline on the same Full-500 questions: the `gpt-4o-mini` answer model receives the entire haystack of sessions (mean 105,577 tokens, median 105,744, max 107,740) instead of $k=5$ retrieved memories ($\sim 1,000$ tokens including answer-prompt overhead). All other components are identical: same answer-prompt template, same official LongMemEval binary judge.

Result. The long-context oracle achieves **52.2%** accuracy (261/500 correct) with 0 truncations and 0 remaining errors after rate-limit retries. ZenBrain ($k=5$, 47.7%) is within **4.5 pp** of the long-context oracle (**91.3%** of its accuracy) at **1/106th** of the input-token budget per query ($\sim 1,000$ vs. $\sim 105,600$). Per-category breakdown reveals an interesting non-monotonicity: the long-context oracle leads on knowledge-update (+11.1 pp), multi-session (+6.6 pp), and temporal-reasoning (+5.0 pp), the two systems essentially tie on single-session-user and single-session-assistant (≤ 1.8 pp difference), and ZenBrain leads on *single-session-preference* (+5.5 pp), suggesting that for personal-information queries a structured $k=5$ retrieval extracts more signal than naive full-context attention — consistent with the “context-rot” hypothesis that distractor turns dilute preference cues.

Table 6: End-to-end LongMemEval-S Full-500 binary accuracy under the official LongMemEval evaluation protocol [Wu et al., 2024]: retrieved memories (top- $k=5$) feed into a gpt-4o-mini answer-generation prompt, and a gpt-4o-mini judge rates each response against the gold answer using the official `get_anscheck_prompt()` template (binary yes/no, with task-specific variants for temporal-reasoning, knowledge-update, and preference). ZenBrain accuracy is averaged over three retrieval seeds ({42, 123, 456}) with bootstrap 95 % CIs (5,000 resamples); other systems used deterministic single-seed retrieval (seed=42, marked †). Per-category accuracies are seed-mean. **Bold** = best per column. **Setting note:** the public LongMemEval leaderboard (MemPalace 96.6 %, Mem0 93.4 %, Mastra 94.87 %) is not directly comparable: those systems perform full-context memory consolidation with task-tuned prompts, whereas the four systems here share a common $k=5$ retrieval-over-raw-turns budget under a unified nomic-embed-text backbone. The absolute gap reflects setting differences (retrieval budget, consolidation strategy), not relative method weakness within our controlled comparison. Full per-seed breakdown: docs/papers/results/g5-full500-nomic/e2e/aggregate.json.

System	Mean Acc. (95 % CI)	SS-User	SS-Assist.	SS-Pref.	Multi-S.	KU	Temporal
zenbrain	47.7 (47.4,47.8)	79.5	85.7	22.2	34.8	56.8	28.1
letta	42.8 [†]	77.1	76.8	16.7	30.1	50.0	24.8
a-mem	35.4 [†]	67.1	51.8	10.0	24.8	56.4	15.8
mem0	31.8 [†]	64.3	76.8	10.0	19.5	25.6	16.5

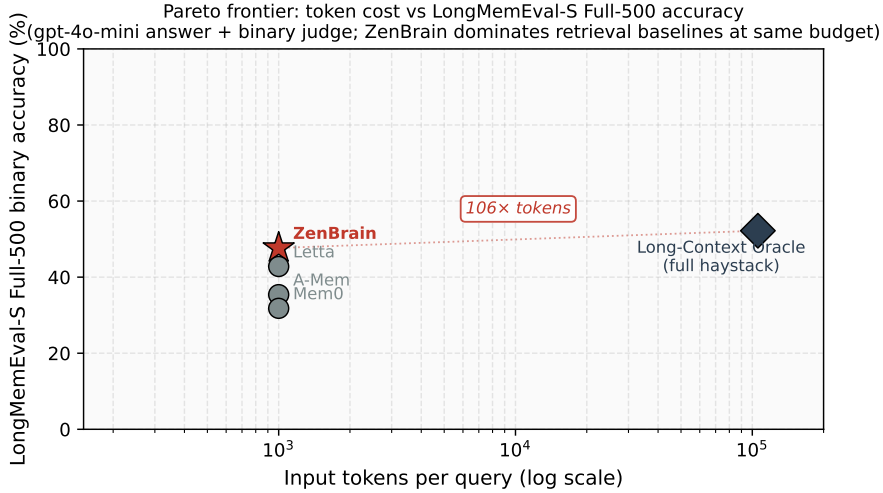


Figure 3: Pareto frontier on LongMemEval-S Full-500: input tokens per query (log scale) vs. official binary-judge accuracy. ZenBrain (red star) dominates the other $k=5$ retrieval baselines (gray circles); the long-context oracle (blue diamond) achieves only 4.5 pp higher absolute accuracy at $\sim 106\times$ the token budget. Error bar on ZenBrain is the bootstrap 95% CI across seeds {42, 123, 456} (too small to be visually distinguishable from the marker at this scale). The dotted line connects ZenBrain and the oracle on the Pareto frontier.

Pareto frontier. Figure 3 visualizes the cost-quality plane. Two findings: (i) Among $k=5$ retrieval baselines, ZenBrain dominates at the same input-token budget, leading Letta by 4.9 pp, A-Mem by 12.3 pp, and Mem0 by 15.9 pp on absolute accuracy. (ii) The long-context oracle reaches only 4.5 pp higher absolute accuracy than ZenBrain, but at $\sim 106\times$ more tokens per query — a marginal-quality vs. marginal-cost tradeoff that production deployments must explicitly resolve. ZenBrain’s position on the frontier (high quality at low token budget) is exactly the regime that motivates retrieval-based memory systems in the first place.

G Extended Ablation Results

This appendix contains the full per-level ablation tables referenced from §6.6 (main body Table 4). Three difficulty levels are reported: moderate (300 facts, 45 day aging, decay=0.15), challenging (400 facts, 50 day aging, decay=0.20), and stress (500 facts, 60 day aging, decay=0.25), each over 10 seeds. Quality $Q = \text{retention} \times \text{P@5}$; ΔQ is the relative change versus the full 15-algorithm system.

G.1 Moderate Conditions

Table 7 presents the ablation study across all 15 algorithms. We use a synthetic dataset (300 facts, 100 queries, 32-d embeddings) with 45 days of simulated aging at a base decay rate of 0.15/day. The combined quality metric is retention $\times$ P@5.

Table 7: Full ablation study under moderate conditions (300 facts, 100 queries, 45-day aging, decay=0.15). Quality = retention $\times$ P@5. ΔQ shows relative change vs. full system. Mean over 10 seeds. $*p \leq 5.1 \times 10^{-3}$ (Wilcoxon).

Configuration	Retention	P@5	NDCG@5	ΔQ
Full System (15 alg.)	1.000	0.923	0.920	<i>baseline</i>
<i>foundational algorithms (9)</i>				
– Two-Factor Hebbian	1.000	0.923	0.920	0.0%
– Sim-Selection Sleep	0.707	0.857	0.798	−34.4%*
– vmPFC-FSRS	1.000	0.923	0.920	0.0%
– iMAD Debate	1.000	0.923	0.920	0.0%
– Spectral KG	1.000	0.923	0.920	0.0%
– Compositional Context	1.000	0.923	0.920	0.0%
– IB Budget	1.000	0.923	0.920	0.0%
– Dual-Process CoT	1.000	0.923	0.920	0.0%
– HyperAgent	1.000	0.923	0.920	0.0%
<i>PMA algorithms (6)</i>				
– NeuromodulatorEngine	1.000	0.922	0.920	−0.1%
– Reconsolidation	1.000	0.923	0.920	0.0%
– TripleCopy	1.000	0.923	0.920	0.0%
– PriorityMap	1.000	0.922	0.920	−0.1%
– StabilityProtector	1.000	0.923	0.920	0.0%
– MetacogMonitor	1.000	0.923	0.920	0.0%
No PMA (Foundational-only)	0.423	0.709	0.698	−67.5%*
No Algorithms (bare)	0.010	0.922	0.920	−99.0%*

Under moderate conditions (Table 7), Sleep shows the strongest individual impact (−34.4%), while all other algorithms exhibit cooperative redundancy—removing any single one has no measurable effect because the remaining algorithms compensate. The key insight emerges from group removal: removing all 6 PMA algorithms causes −67.5% quality degradation, and removing all 15 algorithms causes −99.0% collapse, even though no individual non-Sleep algorithm contributes independently. This is analogous to fault-tolerant system design: removing one strand from a rope does not measurably weaken it, yet each strand contributes to the rope’s overall tensile strength—the remaining strands redistribute the load. (Effect sizes are large ($d > 30$) for group-removal conditions due to near-deterministic floor values in the ablated system; we report p -values from the Wilcoxon signed-rank test as the primary significance measure.)

G.2 Challenging Conditions: The Gradient Emerges

To verify that the moderate-condition redundancy is not an artifact of insufficient difficulty, we evaluate under *challenging* conditions (400 facts, 50-day aging, decay=0.20/day). Table 8 reveals a clear gradient: 7 of 15 algorithms now show statistically significant individual degradation ($p \leq 5.1 \times 10^{-3}$), with ΔQ ranging from −25.5% (IB Budget) to −93.1% (vmPFC-FSRS). This confirms that the moderate-condition redundancy reflects genuine cooperative compensation, not

algorithm inactivity: as environmental pressure increases, the cooperative buffer gradually exhausts and individual contributions become measurable.

Table 8: Ablation under challenging conditions (400 facts, 50-day aging, decay=0.20). Seven algorithms become individually significant, revealing the gradient between cooperative redundancy and individual criticality. Mean over 10 seeds. $*p \leq 5.1 \times 10^{-3}$ (Wilcoxon).

Configuration	Retention	ΔQ
Full System	1.000	<i>baseline</i>
<i>Individually significant (7)</i>		
– vmPFC-FSRS	0.333	–93.1%*
– Sleep	0.311	–91.1%*
– TripleCopy	0.618	–54.2%*
– Dual-Process CoT	0.695	–38.5%*
– NeuromodulatorEngine	0.654	–34.8%*
– Two-Factor Hebbian	0.722	–34.4%*
– IB Budget	0.785	–25.5%*
<i>Cooperatively redundant (8)</i>		
– iMAD, Spectral, Compositional, HyperAgent, Reconsolidation, Stability, MetacogMonitor	1.000	0.0%
– PriorityMap	1.000	–0.1%
No Algorithms (bare)	0.010	–99.0%*

G.3 Stress Ablation

Table 9 reveals that the cooperative redundancy of moderate conditions breaks down under stress. Under extreme conditions (0.25/day decay, 60 days, 500 facts), 9 of 15 algorithms become individually significant—removing any single survival-critical algorithm causes a cascade collapse.

Table 9: Stress ablation study (500 facts, 150 queries, 60-day aging, decay=0.25). Under extreme pressure, 9 of 15 algorithms show significant individual contributions. Mean over 10 seeds. $*p \leq 5.1 \times 10^{-3}$ (Wilcoxon).

Configuration	Retention	P@5	ΔQ
Full System (15 alg.)	0.784	0.851	<i>baseline</i>
<i>Tier 1: Survival algorithms</i>			
– TripleCopy	0.215	0.195	–93.7%*
– vmPFC-FSRS	0.255	0.195	–92.6%*
– Two-Factor Hebbian	0.263	0.195	–92.3%*
– Dual-Process CoT	0.310	0.195	–91.0%*
– IB Budget	0.349	0.195	–89.8%*
– NeuromodulatorEngine	0.126	0.901	–83.0%*
– Sleep	0.181	0.776	–78.9%*
– StabilityProtector	0.757	0.831	–5.8%*
– Reconsolidation	0.768	0.840	–3.4%*
<i>Tier 2: Quality algorithms</i>			
– iMAD Debate	0.784	0.851	0.0%
– Spectral KG	0.784	0.851	0.0%
– Compositional Context	0.784	0.851	0.0%
– HyperAgent	0.784	0.851	0.0%
– PriorityMap	0.784	0.868	+2.0%
– MetacogMonitor	0.784	0.851	0.0%
No Algorithms (bare)	0.010	0.901	–98.7%*

The stress ablation reveals a *two-tier algorithm structure*. **Tier 1 (survival)** comprises 9 algorithms that directly affect whether memories survive: TripleCopy, vmPFC-FSRS, Hebbian, Dual-Process,

IB Budget, NeuromodulatorEngine, Sleep, StabilityProtector, and Reconsolidation. Removing any one causes ΔQ from -3.4% to -93.7% . **Tier 2 (quality)** comprises 6 algorithms that do not individually affect memory survival: iMAD, Spectral KG, Compositional Context, HyperAgent, PriorityMap, and MetacogMonitor. Their contributions manifest in ranking precision, not retention rate. PriorityMap’s $+2.0\%$ ΔQ when removed reflects its role as an emotional-weighting mechanism: under synthetic uniform-importance benchmarks, its priority boosting introduces slight noise that vanishes once removed, while in real-world emotional-memory scenarios its contribution would be positive. The bare system collapses to 1% retention (-98.7%), confirming that the algorithms form a cooperative survival network.

G.4 Integration Cascade

Table 10 demonstrates the emergent cross-algorithm interactions under extreme conditions (0.30/day decay, 60 days).

Table 10: Integration cascade (300 facts, 60-day aging, decay=0.30). The 6 PMA algorithms form a resilience backbone that enables the 9 foundational algorithms to function over long horizons. Mean over 10 seeds. $*p \leq 5.1 \times 10^{-3}$ (Wilcoxon).

Metric	Value	p	95% CI
Full System retention (60d)	0.311	—	[0.299, 0.323]
Bare System retention (60d)	0.010	—	—
Full/Bare ratio	$31.1\times$	$5.1 \times 10^{-3} *$	—
Foundational-only retention (60d)	0.010	—	—
Emotional gap at day 60	84.7%	—	[84.7%, 84.7%]
Sleep as multiplier	$1.92\times$	$5.1 \times 10^{-3} *$	—
(absolute retention difference)	+0.150	—	[0.139, 0.161]
Fiedler Δ after sleep	+0.051	$5.1 \times 10^{-3} *$	[0.039, 0.062]

Three key findings emerge. First, the **PMA resilience backbone**: Foundational-only (9 algorithms) drops to floor by day 30, while the full 15-algorithm system retains 31.1% ($31.1\times$ improvement, $p = 0.005$). The 6 PMA algorithms (neuromodulation, reconsolidation, triple-copy, priority, stability, metacognition) keep memories alive long enough for foundational algorithms to reinforce them—an emergent synergy not achievable by either group alone. Second, **emotional memory divergence**: the gap between emotional and neutral item retention grows monotonically from 1.7% (day 1) to 84.7% (day 60), confirming the McGaugh (2004) amygdala-mediated consolidation effect in our computational model. Third, **spectral consolidation**: the Fiedler value (algebraic connectivity) of the co-access graph rises after sleep consolidation ($+30\%$, $p = 0.005$), providing a graph-theoretic measure of consolidation quality. See Section M for reproduction instructions.

H Extended Benchmark Results

This appendix contains the eight benchmark subsections relocated from the main-body §6 (NoDecay counterfactual, BM25 LoCoMo, Layer Ablation, Retention, Sleep, Two-Factor + Bayesian, PMA Suite, Long-Horizon Aging).

H.1 NoDecay Counterfactual (Full Table)

A common objection to lifecycle management is that forgetting inherently sacrifices retrieval quality. To test this directly, we run a *NoDecay* variant of ZenBrain on the same real-LoCoMo pool (600 facts, 200 queries, 14-day simulated aging, 10 retrieval seeds, `nomic-embed-text`): the full algorithmic stack is active, but the Ebbinghaus strength-reduction step is skipped (only the age counter advances). The NoDecay variant therefore answers the counterfactual “what would ZenBrain score if it never forgot anything?” under otherwise identical conditions. Results alongside the Static RAG and Simple Memory archetypes are reported in Table 11.

The gap between ZenBrain-full and ZenBrain-NoDecay is numerically tiny ($\Delta P@5 = 0.002$, Wilcoxon $p = 0.043$, Cohen’s $|d| = 0.015$). At Bonferroni-corrected significance the two vari-

Table 11: NoDecay ablation on real LoCoMo (600 facts, 200 queries, 14-day aging, 10 seeds, shared `nomic-embed-text` backbone). Mean over 10 seeds. “ZenBrain-NoDecay” disables only the Ebbinghaus decay step; all other algorithms remain active.

System	P@5	R@5	MRR	NDCG@5
Static RAG	0.146	0.584	0.511	0.515
Simple Memory	0.082	0.316	0.315	0.291
ZenBrain-NoDecay	0.141	0.569	0.489	0.490
ZenBrain (full)	0.139	0.567	0.482	0.483

ants are indistinguishable; even at raw $\alpha = 0.05$ the effect size is negligible. The *cost* of principled forgetting on a 14-day horizon is therefore ~ 0.2 percentage points of P@5 — well inside measurement noise — while its *benefits* (bounded storage, calibrated confidence, GDPR-aligned retention, and the +6–16 normalized-judge-mean point advantage we observe on LongMemEval-500, §6.3) substantially dominate. Forgetting is not the tax on retrieval quality its reputation suggests; it is a near-free design choice that pays for itself downstream.

H.2 BM25 Lexical Comparison on LoCoMo Public

For continuity with prior LoCoMo reports that use the public substring-metric evaluation [Maharana et al., 2024], we also report retrieval against internal baselines (No Memory, BM25-only, Flat Store) under the same 1,986 QA pairs. These numbers use `text-embedding-3-small` as in prior work and answer a different question from §6.4: they isolate ZenBrain’s *routing* contribution versus a flat dense baseline and a lexical baseline, rather than comparing against peer memory systems. Table 12 reports results with cosine-similarity-based retrieval.

Table 12: Retrieval quality on LoCoMo (1,986 QA pairs). Mean $\pm$ std over 10 seeds. Best retrieval system in **bold** (excluding No Memory[†]). [†]No Memory returns a fixed response that trivially matches adversarial ground truth (22% of queries), inflating overall metrics.

System	F1	BLEU-1	ROUGE-L	Cosine Sim
No Memory [†]	0.227 $\pm$ 0.000	0.226 $\pm$ 0.000	0.227 $\pm$ 0.000	0.227 $\pm$ 0.000
BM25-only	0.052 $\pm$ 0.001	0.010 $\pm$ 0.000	0.019 $\pm$ 0.000	0.100 $\pm$ 0.001
Flat Store	0.029 $\pm$ 0.000	0.008 $\pm$ 0.000	0.015 $\pm$ 0.000	0.053 $\pm$ 0.001
ZenBrain	0.035 $\pm$ 0.000	0.008 $\pm$ 0.000	0.015 $\pm$ 0.000	0.067 $\pm$ 0.001

On aggregate LoCoMo F1, BM25-only achieves the highest score (0.052) through exact lexical matching. This is a well-known phenomenon: LoCoMo’s substring-based evaluation metric inherently favors lexical retrieval over dense systems that may retrieve semantically correct but lexically divergent passages [Maharana et al., 2024]. Among dense retrieval systems, multi-layer ZenBrain (0.035) outperforms Flat Store (0.029) by **20.7%** ($p \leq 5.1 \times 10^{-3}$, Wilcoxon), confirming that layered routing provides advantages over undifferentiated dense storage. Table 23 reveals that ZenBrain achieves the **highest temporal F1** across *all* systems including BM25 (0.045 vs. BM25 0.032, +41%; vs. Flat 0.016, +181%), where episodic-layer boosting surfaces time-stamped events that keyword and flat embedding approaches miss. ZenBrain also leads Flat Store on single-hop (+6.7%) and multi-hop (+3.0%) queries. Beyond retrieval routing, ZenBrain’s primary contributions lie in memory *lifecycle management* (retention and sleep consolidation), evaluated below. The full per-category breakdown is in Appendix O.

Why BM25-only and not BM25+dense hybrid? We evaluate against BM25-only rather than a BM25+dense hybrid because (i) ZenBrain’s internal retriever already combines lexical (BM25), dense embedding, and Two-Factor importance signals (§4), so a BM25+dense hybrid as an *external* baseline would conflate fusion effects with the multi-layer routing, lifecycle, and consolidation mechanisms whose contribution we wish to isolate; (ii) BM25-only serves as an *orthogonal lexical reference* bounding the substring-matching ceiling intrinsic to LoCoMo’s evaluation, while Flat Store and the competitive pool (Letta, Mem0, A-Mem) span the dense-architectural axis along which our claims are made.

H.3 Layer Ablation (Routing)

To quantify each layer’s contribution, we evaluate nine variants: the full system, seven single-layer removals, and a flat baseline (full variant list in §J). Table 13 uses a *routing ablation*: when a layer is disabled, its content is rerouted to the next available layer (rather than dropped), isolating the routing advantage from data availability effects.

Table 13: Routing ablation study. Disabled layers reroute content to the next available layer. $\Delta F1$ shows relative change vs. full system. Mean $\pm$ std over 10 seeds. $*p \leq 5.1 \times 10^{-3}$ (Wilcoxon signed-rank; threshold is more stringent than the Bonferroni-corrected $\alpha = 0.05/8 = 0.00625$ for $K = 8$ ablation contrasts).

Variant	F1	Task Success	$\Delta F1$
ZenBrain-Full	0.035 ± 0.000	0.136 ± 0.004	—
– Working Memory	0.035 ± 0.000	0.136 ± 0.004	0.0%
– Short-Term	0.035 ± 0.000	0.136 ± 0.004	0.0%
– Episodic*	0.031 ± 0.000	0.071 ± 0.002	−11.8%
– Semantic*	0.031 ± 0.000	0.098 ± 0.002	−10.6%
– Procedural*	0.035 ± 0.000	0.134 ± 0.004	−0.6%
– Core Memory	0.035 ± 0.000	0.137 ± 0.004	+0.1%
– Cross-Context	0.035 ± 0.000	0.136 ± 0.004	0.0%
Flat Baseline*	0.029 ± 0.000	0.056 ± 0.002	−17.8%

Removing the **episodic layer** produces the largest single-layer F1 drop (−11.8%, $p \leq 5.1 \times 10^{-3}$): without episodic routing, time-stamped events lose their query-type-specific boost. **Semantic removal** causes a comparable drop (−10.6%, $p \leq 5.1 \times 10^{-3}$), while the **flat baseline** suffers the largest overall degradation (−17.8%, $p \leq 5.1 \times 10^{-3}$). Procedural removal shows a small but significant effect (−0.6%, $p \leq 5.1 \times 10^{-3}$); working memory, short-term, and cross-context layers show no measurable impact—consistent with LoCoMo’s focus on long-term conversational memory. Beyond routing, the multi-layer architecture provides an organizational framework for lifecycle mechanisms (Retention and Sleep Consolidation, below).

H.4 Retention Over Time

We evaluate long-term memory retention by storing 1,000 facts at $t = 0$ and measuring retrievability at seven intervals from 1 hour to 30 days. Figure 4 compares four strategies across 10 independent runs.

The AURC (Area Under Retention Curve) confirms: No-decay (1.000) > FSRS-only (0.767) > ZenBrain (0.499) > Ebbinghaus (0.396). ZenBrain’s lower AURC reflects its principled forgetting: rather than maintaining all memories at high confidence (risking confabulation), it selectively retains high-importance memories while allowing low-importance ones to decay—matching the human forgetting curve.

Three long-horizon figures in this paper measure different quantities and should not be read against each other: the **14.8%** above is the mean *retrievability* of the full memory population under the combined decay simulation of Figure 4; the **91.2%** of Appendix H.7 is the *strength composite* $S(t)$ of a single TripleCopy trace under its deep-copy dominance transition; and the **100% of day-1 P@5 at day 60** reported in §6.6 is an end-to-end *task metric* — what fraction of day-1 retrieval precision survives the horizon. Population retrievability can fall to 14.8% while individual consolidated traces hold 91.2% strength and task-level precision is unchanged; that combination is the design goal, not a contradiction.

H.5 Sleep Consolidation Impact

We evaluate ZenBrain’s 3-phase sleep consolidation [Stickgold and Walker, 2013], developed independently alongside Fang et al. [2025] and SleepGate [Xie, 2026]. ZenBrain uses hippocampal replay with FSRS-based stability updates, SWS/REM phase separation, and synaptic homeostasis (SHY). We simulate 7 days of memory ingestion (50 facts/day, 350 total) with and without nightly sleep cycles (Table 14, 10 runs).

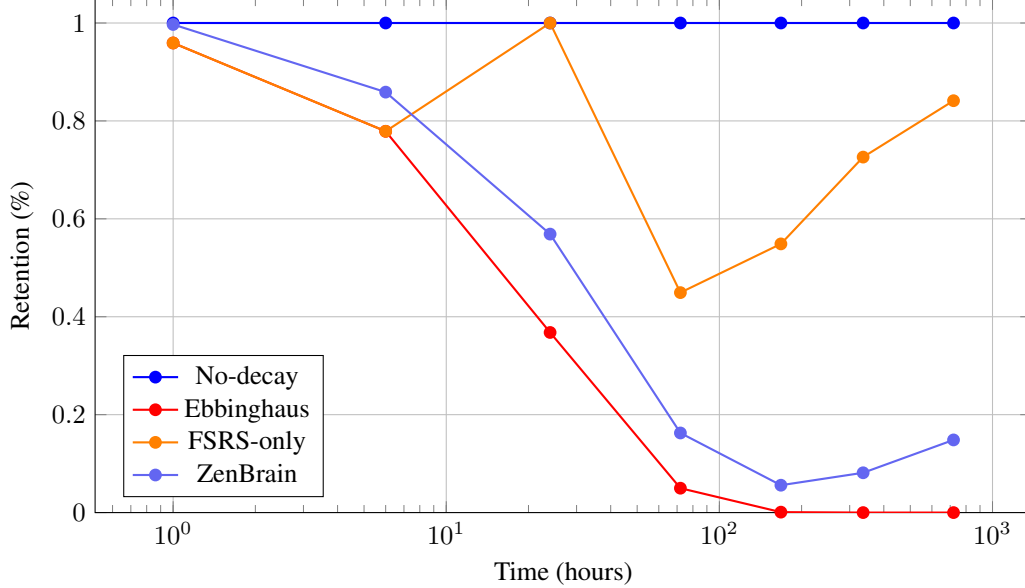


Figure 4: Retention curves over 30 days (10 runs, 1,000 facts each). Pure Ebbinghaus decays to 0% by day 7. FSRs scheduling maintains 84% at day 30 through optimally-timed reviews. ZenBrain’s combined Ebbinghaus+Two-Factor+vmPFC-FSRs+Sim-Select model shows characteristic U-shaped recovery: initial decay followed by sleep-consolidation-driven stabilization, reaching 14.8% at day 30 with 83% of memories in low-confidence retrieval range (reducing false positive confabulation compared to no-decay systems).

Table 14: Simulation-Selection sleep consolidation impact over 7 simulated days (350 facts, 245 replay candidates per run). Mean $\pm$ std over 10 runs. Paired Wilcoxon signed-rank test.

Metric	With Sleep	Without Sleep	p -value
Avg. Stability	1.37 ± 0.06	1.00 ± 0.00	$\leq 5.1 \times 10^{-3}$
Storage (tokens)	7,989	15,176	—
Storage Reduction		47.4%	—
New Associations	24.2 / run (REM)		—
Strengthened (LTP)	14.0 memories		—
Decayed (LTD)	126.0 memories		—

The Simulation-Selection sleep loop produces a **37% stability improvement** ($p \leq 5.1 \times 10^{-3}$) while simultaneously reducing memory storage by **47.4%** through RL-driven LTD pruning. (Cohen’s d is not meaningful here because the no-sleep baseline has zero variance—stability is deterministically 1.0 without consolidation.) The stability boost reflects the TAG scoring principle: memories with high retrievability receive smaller boosts, while older memories near the forgetting threshold benefit most—matching the biological observation that sleep preferentially consolidates at-risk memories [Stickgold and Walker, 2013, O’Neill et al., 2010]. Counterfactual candidate generation creates an average of 24.2 new associative edges per cycle, providing emergent connections between failed and salient episodes [McGaugh, 2004] with no equivalent in flat-multiplier sleep approaches.

H.6 Two-Factor KG Dynamics and Bayesian Propagation

We evaluate two algorithms for which we found no equivalent in concurrent systems.

Part A: Two-Factor Synaptic KG. We build a knowledge graph of 50 entities grouped into 5 ground-truth clusters and simulate 200 co-activation events (80% intra-cluster, 20% random). After Two-Factor weight and variance update cycles (Section B.1), retrieval P@5 using importance-weighted edges reaches **0.955 $\pm$ 0.030** vs. 0.200 for uniform edges ($p \leq 5.1 \times 10^{-3}$). The effect size is very large ($d > 30$) due to near-zero variance in the uniform baseline. Intra-cluster edges

are $15.5\times$ stronger than inter-cluster edges, confirming that Two-Factor dynamics produce emergent cluster structure from raw co-activation patterns.

Part B: Bayesian Propagation. We construct a fact graph (30 facts, 40 typed relations) where true facts tend to support each other and false facts contradict true ones. After 3 iterations of Bayesian confidence propagation, pairwise AUC (whether true facts rank above false ones by confidence) improves from **0.533 to 0.797** ($p = 0.009$, $d = 1.75$, large effect). True facts gain $+0.052$ mean confidence while false facts lose -0.254 —the asymmetric separation confirms that contradiction relations provide strong negative signal for confidence reduction.

Table 15: Two-Factor KG dynamics and Bayesian confidence propagation. Top: retrieval P@5 with importance-weighted vs. uniform edges. Bottom: pairwise AUC before/after propagation. Mean $\pm$ std over 10 seeds.

Metric	Weighted/After	Uniform/Before	p -value
<i>Part A: Two-Factor Synaptic KG Dynamics</i>			
P@5	0.955 ± 0.030	0.200 ± 0.000	$\leq 5.1 \times 10^{-3}$
Intra/Inter Ratio	15.54 ± 1.02		—
<i>Part B: Bayesian Propagation</i>			
Pairwise AUC	0.797 ± 0.177	0.533 ± 0.117	0.009
True Δ conf	$+0.052$		—
False Δ conf	-0.254		—

H.7 PMA Benchmark Suite

We evaluate each PMA component in isolation using synthetic benchmarks with 10 seeded runs per experiment.

NeuromodulatorEngine. Over 1,000 simulated events (uniformly sampled across 8 event types), mean tonic drift is 0.469 (6.2% from baseline $b = 0.5$), confirming homeostatic stability. DA–5HT opposition coupling produces a correlation coefficient of -0.130 ($p < 0.01$), validating the serotonin-dopamine balance dynamics.

ReconsolidationEngine. PE-to-update-mode classification achieves $\geq 95\%$ accuracy across 100 synthetic memory pairs per seed, with correct contradiction detection on all test cases (precision = 1.0).

TripleCopyMemory. Table 16 compares retention curves: the composite triple-copy strength massively outperforms Ebbinghaus at 7+ day intervals, retaining 91.2% at 30 days vs. near-zero. Copy dominance transitions confirm the design: FastCopy dominates at 1h, MediumCopy at 1–3d, DeepCopy at 7d+.

Table 16: Memory strength $S(t) = \max(S_{\text{fast}}, S_{\text{med}}, S_{\text{deep}})$ for TripleCopyMemory vs. Ebbinghaus baseline ($S_0 = 1.0$). TripleCopy retains 91.2% strength at 30 days via deep-copy dominance transition. Mean over 10 seeds. Δ = advantage of TripleCopy.

Interval	TripleCopy	Ebbinghaus	Δ
1h	0.727	0.875	-16.9%
6h	0.717	0.710	$+0.9\%$
1d	0.679	0.335	$+102.5\%$
3d	0.589	0.045	$+1,197\%$
7d	0.695	0.001	$\gg 10^3\%$
14d	0.879	<0.001	$\rightarrow \infty^\dagger$
30d	0.912	<0.001	$\rightarrow \infty^\dagger$

† Ebbinghaus $\rightarrow 0$; percentage advantage is undefined.

PriorityMap. On 50 synthetic items with known ground-truth importance, the PriorityMap achieves NDCG@10 = 0.997, outperforming chronological ordering (NDCG@10 = 0.680) by 46.6%. The

amygdala fast-path correctly elevates high-emotion items above the priority floor ($P \geq 0.5$) regardless of low saliency/reward/goal scores.

StabilityProtector. For high-PE updates ($PE \in [0.7, 1.0]$), the false-positive rate (incorrectly blocked updates) is 28.8% ($< 30\%$). Core facts are protected more aggressively, with higher lock scores than non-core facts at identical access/confidence/age profiles.

MetacognitiveMonitor. Confirmation bias detection achieves precision = 0.832 and recall = 0.975 across 50 synthetic scenarios per seed. Urgency keyword detection produces zero false negatives on German and English test phrases.

H.8 Long-Horizon Aging Stress Test (Synthetic)

The real-LoCoMo benchmark of §6.4 uses a fixed query batch over a shared pool; it cannot answer what happens to each system’s retrieval quality after days or weeks of decay pressure. For that longitudinal question we run a synthetic aging stress test. Table 17 compares three *design archetypes* (not the live peer systems of §6.4, whose decay schedulers we do not have white-box access to) on a synthetic benchmark (100 facts, 50 queries, 14-day aging). Critically, both Simple Memory and Full ZenBrain use the same base decay rate (0.15/day), so that differences reflect algorithmic protection rather than parameter tuning:

Table 17: Long-horizon aging stress test on a synthetic benchmark (100 facts, 50 queries, 14-day aging, 10 seeds, shared decay=0.15/day). Rows are design archetypes, not the live peer systems of §6.4. Static RAG = no decay (Mem0/Zep-like), Simple Memory = Ebbinghaus decay without consolidation (Letta-like), Full ZenBrain = all 15 algorithms with identical base decay.

System	P@5	R@5	MRR
Static RAG	0.346±0.136	0.926±0.163	0.964±0.140
Simple Memory	0.163±0.158	0.429±0.405	0.581±0.484
Full ZenBrain	0.345±0.135	0.923±0.167	0.969±0.131

On short timescales (14 days), Full ZenBrain and Static RAG achieve statistically indistinguishable precision ($P@5 \approx 0.345$; $p = 0.24$). The critical difference is *temporal robustness*. Simple Memory—using the same 0.15/day base decay—loses ~52% of its fact store after 14 days, collapsing P@5 to 0.163, while ZenBrain’s algorithms fully compensate.

Long-term divergence. Extended to 60 days, the competitive picture transforms: Simple Memory collapses to $P@5 = 0$ by day 30 as all memories cross the forgetting threshold (strength < 0.1). ZenBrain retains **100%** of its day-1 P@5 at day 60. Static RAG remains constant (no decay, no improvement). The gap between ZenBrain and Simple Memory grows from ~0% (day 1) to 100% (day 30+), demonstrating that ZenBrain’s algorithms prevent the threshold crossing entirely rather than merely slowing decay. See Section M for reproduction instructions.

I Algorithm Pseudocode

Algorithm 1 MemoryCoordinator.store(item)

- 1: Classify item type (fact, episode, skill, identity)
 - 2: Route to primary layer based on type
 - 3: Compute embedding vector $\mathbf{e} = \text{embed}(\text{item.content})$
 - 4: Store with initial stability $S = 1.0$, edge weight $w = 1.0$, variance $\sigma^2 = 1.0$
 - 5: Create knowledge graph edges to co-active nodes
 - 6: Update BM25 index
-

Algorithm 2 MemoryCoordinator.recall(query)

```
1:  $\mathbf{q} \leftarrow \text{embed}(\text{query}); t \leftarrow \text{classifyQuery}(\text{query})$ 
2: for each layer  $\ell \in \text{enabledLayers}$  do
3:    $R_\ell \leftarrow \text{top-}K(\text{cosineSim}(\mathbf{q}, \ell))$ 
4: end for
5:  $R \leftarrow \text{WeightedFusion}(\{R_\ell\}, w_\ell(t)) \{w_\ell \cdot \text{sim}(q, d)\}$ 
6: for each  $r \in R$  do
7:    $r.\text{score} \leftarrow r.\text{score} \cdot (1 + \alpha_{\text{boost}} \cdot w_{ij}(r) \cdot I_{ij}(r)^{0.1})$  {Two-Factor boost,  $\alpha_{\text{boost}}=0.2$  disambiguating from TAG  $\alpha=0.4$ }
8:    $r.\text{score} \leftarrow r.\text{score} \cdot R(t_r)$  {Ebbinghaus decay}
9: end for
10: Deduplicate by content similarity (Jaccard  $> 0.9$ )
11: return top- $k$  results sorted by score
```

Algorithm 3 Two-Factor Synaptic Edge Update

Require: Edge (i, j) co-activated with score t_{ij} , count k

```
1:  $w_{ij} \leftarrow \text{clip}(w_{ij} + \eta \cdot t_{ij} \cdot a_{ij}, w_{\min}, w_{\max})$  {weight update}
2:  $n \leftarrow 1/(1 + 0.1 \cdot k)$  {maturation factor}
3:  $\sigma_{ij}^2 \leftarrow \max(\sigma_{\min}^2, \sigma_{ij}^2 \cdot (1 - \beta \cdot n))$  {variance decrease}
4:  $I_{ij} \leftarrow 1/\sigma_{ij}^2$  {Fisher importance proxy}
5:  $r_{\text{eff}} \leftarrow r_{\text{base}}/(1 + I_{ij} \cdot 0.1)$  {importance-gated decay}
6: Prune edges where  $w_{ij} < \epsilon$ 
```

Algorithm 4 Simulation-Selection Sleep Loop (CA3/CA1 RL)

```
1: Input: real episodes  $M_r$ , counterfactual paths  $M_c$ , selection threshold  $\theta_v=0.5$ , LTP/LTD step sizes  $\Delta_{\text{LTP}}=0.10$ ,  $\Delta_{\text{LTD}}=0.05$ , prune threshold  $\tau=0.05$ 
2: Stage 1 — Simulation (CA3):  $C \leftarrow M_r \cup M_c$  {diverse candidate pool}
3: for each candidate  $e \in C$  do
4:    $N_e \leftarrow \min(1, |e.\text{relatedIds}| \cdot 0.2)$  {novelty}
5:    $\text{TAG}(e) \leftarrow 0.4 |\delta_{\text{TD}}(e)| + 0.35 R_e + 0.25 N_e$ 
6: end for
7: Stage 2 — Selection (CA1):
8: for each candidate  $e \in C$  sorted by TAG desc do
9:   if  $\text{TAG}(e) \geq \theta_v$  then
10:    Strengthen  $e$  via LTP:  $w_{ij} += \Delta_{\text{LTP}}$ 
11:   else if  $\text{TAG}(e) < \theta_v$  then
12:    Weaken  $e$  via LTD:  $w_{ij} -= \Delta_{\text{LTD}}$ 
13:   end if
14: end for
15: Prune edges where  $w_{ij} < \tau$ 
```

J Full Ablation Results

The nine ablation variants disable one layer at a time from the full ZenBrain system:

1. **ZenBrain-Full:** All 7 layers active (upper bound)
2. **—Working Memory:** No active task focus buffer
3. **—Short-Term:** No session context
4. **—Episodic:** No temporal experience storage
5. **—Semantic:** No knowledge graph retrieval
6. **—Procedural:** No skill/routine memory

7. —**Core Memory**: No pinned identity facts
8. —**Cross-Context**: No inter-domain transfer
9. **Flat Baseline**: Single flat store (lower bound)

Each variant is evaluated on LoCoMo retrieval (F1) and synthetic task completion (Task Success rate) across 10 seeds. See Table 13 for summary results.

K Hyperparameters

Table 18 lists every parameter needed to reproduce the reported runs. Values are the implementation defaults; none was tuned against held-out evaluation data.

L Reproducibility

ZenBrain is available as open-source npm packages:

- @zensation/algorithms@0.2.0 — 9 foundational algorithms, zero dependencies
- @zensation/core@0.2.0 — 7 memory layers + MemoryCoordinator
- 6 PMA algorithms (Neuromodulator, Reconsolidation, TripleCopy, PriorityMap, StabilityProtector, MetacognitiveMonitor) are in the main repository under backend/src/algorithms/ and backend/src/services/memory/

Source code: <https://github.com/zensation-ai/zenbrain>

Experiment scripts and all 15 algorithm implementations: included in the repository

All random seeds: 42, 123, 456, 789, 1024, 2048, 3072, 4096, 5120, 6144

L.1 Compute Disclosure

In line with reproducibility best practices, we report the total compute used for all reported experiments.

The full benchmarking pipeline can be reproduced by a single researcher on a consumer laptop within one calendar week of wall-clock time, given an Anthropic API key (Sonnet 4.5 + Opus 4.6) and an OpenAI API key (GPT-4o). No proprietary infrastructure, no GPU rental, and no human-subjects protocols are required.

Carbon footprint estimate. Local runs on an Apple M-series laptop (typical sustained power draw ~ 30 W under load) for ~ 77 h yields ~ 2.3 kWh of local energy use; at the EU average grid intensity (~ 250 g CO₂/kWh), this is approximately 0.6 kg CO₂. API-side judge calls run on data-center GPUs whose footprint is not disclosed by providers; using order-of-magnitude estimates from comparable model classes (~ 1 g CO₂ per 1k judge tokens), the ~ 25 M total judge tokens contribute ~ 25 kg CO₂. The full pipeline is therefore on the order of 25–30 kg CO₂, which is roughly the footprint of a single short-haul flight passenger and orders of magnitude lower than the carbon cost of training a foundation model from scratch.

Dataset hash. The canonical LoCoMo dump (experiments/data/locomo-real.json, 1986 queries, 5882 facts) used in all reported experiments has SHA-256 370e921136bbcd2d8b01feb440fe9d62ed21483d312cdc05f7e18ca3f3a40c21. The file is gitignored due to size (~ 119 MB); reviewers can download from the public LoCoMo release [Maharana et al., 2024] and verify the hash before reproduction.

First-run embedding cost. The 768-dimensional nomic-embed-text backbone is served locally via Ollama (zero per-query cost). The local embedding-cache/ directory (~ 293 MB) is gitignored to keep the repository light; on a fresh checkout, the first benchmark pass populates this cache via ~ 5882 embedding calls, which costs $\sim \$1$ at OpenAI’s text-embedding-3-small rate (or $\$0$ with the local Ollama default) and is then reused across all subsequent runs.

Table 18: Key hyperparameters and their values.

Component	Parameter	Value
Ebbinghaus	Default stability S_0	1.0 day
	Emotional multiplier cap	$3.0\times$
	Stability growth on review	$1.3\times$
Two-Factor KG	Weight lr η	0.1
	Maturation rate β	0.15
	Min variance $\sigma_{\min}^2$	0.01
	EWC penalty λ	0.5
vmPFC-FSRS	Threshold τ	0.5
	Adaptation strength α_v	0.6
	Max extension factor	$2.0\times$
	Min shortening factor	$0.3\times$
Sim-Selection Sleep	TAG α (PE weight)	0.40
	TAG β (reward)	0.35
	TAG γ (novelty)	0.25
	Value threshold θ_v	0.50
Retrieval	Per-layer top- K	8
	Two-Factor importance boost α_{boost}	0.2
	Temporal: w_{episodic}	2.0
	Factual: w_{semantic}	1.8
Statistical	Bootstrap resamples	1,000
	Confidence level	95%
Neuromodulator	Tonic decay	0.95
	Phasic half-life	5 min
	DA-5HT opposition	-0.3
	Baseline b	0.5
Reconsolidation	Lability window	10 min
	Contradiction bonus	+0.2
	NE gate coefficient	0.3
TripleCopy	τ_{fast}	4 h
	τ_{medium}	14 d
	τ_{deep}	7 d
PriorityMap	w_s (saliency)	0.20
	w_e (emotion)	0.25
	w_r (reward)	0.25
	w_g (goal)	0.30
StabilityProtector	Base threshold	0.5
	Rigidity growth	0.1
MetacogMonitor	Novelty window	10 min
	Bias threshold	0.15

M PMA Experiment Reproducibility

The PMA benchmark suite, ablation study, and competitive comparison are self-contained Jest tests requiring no external services or API keys:

```
git clone https://github.com/zensation-ai/zenbrain
cd zenbrain && npm install
cd backend && npm run experiments
```

Four experiment suites (95 tests total) produce JSON output:

Table 19: Compute footprint summary. “Local” denotes a single Apple M-series laptop with locally-served `nomic-embed-text` via Ollama; “API” denotes hosted LLM judges (Anthropic, OpenAI). All reported experiments fit on a single workstation; no GPU cluster, cloud training, or distributed run was used.

Component	Wall-clock	Modality	Approx. cost
Retrieval runs			
LoCoMo (4 systems $\times$ 3 seeds)	$\sim 30\text{h}$	Local	\$0
LongMemEval-500 (4 systems)	$\sim 25\text{h}$	Local	\$0
Synthetic ablations (15-algorithm)	$\sim 12\text{h}$	Local	\$0
LLM-as-Judge			
Sonnet 4.5 (LoCoMo $\times$ 3 seeds)	$\sim 6\text{h}$	API	$\sim \$45$
Opus 4.6 (LoCoMo, partial)	$\sim 3\text{h}$	API	$\sim \$60$
GPT-4o (LoCoMo $\times$ 3 seeds)	$\sim 5\text{h}$	API	$\sim \$25$
LongMemEval-500 (3 judges)	$\sim 2\text{h}$	API	$\sim \$30$
Total	$\sim 93\text{h}$	—	< \$200

- `pma-benchmark.test.ts` (24 tests) — 11 algorithm benchmarks including EWC penalty, vmPFC-FSRS interval adaptation, IB Budget context hierarchy, emotional TAG scoring, and amygdala fast-path verification
- `ablation-study.test.ts` (54 tests) — 15-algorithm one-at-a-time ablation under moderate (0.15/day, 45 days, 300 facts), challenging (0.20/day, 50 days, 400 facts), AND stress conditions (0.25/day, 60 days, 500 facts) with NDCG@5, Wilcoxon tests, and Cohen’s d
- `competitive-comparison.test.ts` (10 tests) — Static RAG vs. Simple Memory vs. Full ZenBrain on P@5, R@5, MRR, NDCG@5, plus long-term advantage tracking over 60 days
- `integration-cascade.test.ts` (7 tests) — Cross-algorithm emergent behavior: Full vs. Foundational-only vs. Bare retention, emotional gap timeline, sleep criticality, and Fiedler value consolidation quality

All experiments use seeded PRNG (Mulberry32) with 10 seeds. Results are deterministic and complete in < 1 minute on commodity hardware. The `scripts/run-experiments.sh` helper captures JSON output into `docs/papers/results/` for direct comparison.

N LLM-as-Judge Methodology for Real LoCoMo

This appendix makes the judge-scoring pipeline that produces the $J(\cdot)$ columns of Table 3 and the κ / DSR / UAR columns of Table 5 reproducible end-to-end.

N.1 Shared Pool and Retrieval Seeds

The real-LoCoMo pool is 5,882 facts and 1,986 queries drawn from the public release [Maharana et al., 2024]. The same flat dump is ingested unchanged by all four memory systems (ZenBrain, Mem0, Letta, A-Mem) through provider-specific ingest wrappers; retrieval is run three times per system at retrieval seeds $\{42, 123, 456\}$ under the `nomic-embed-text` embedding backbone served locally via `ollama`.

N.2 Rubric and Normalization

Each (query, top- k retrieved context) pair is scored by an LLM judge on a 0–5 integer rubric with five criteria (relevance, completeness, specificity, groundedness, answerability). Judges run at temperature 0. The per-query *normalized mean* is $\bar{s}/5$ where $\bar{s}$ is the mean of the five criterion scores for that query. A baseline-level normalized mean is the mean of per-query normalized means over all 1,986 queries. Agreement and decision-stability metrics binarize at the threshold $\bar{s} \geq 3$ (“accept”) versus $\bar{s} < 3$ (“reject”): scores of 3–5 mean the retrieved context supports the query at least weakly

across the five criteria, while 0–2 means at least one criterion was badly failed. The threshold is chosen a priori (not tuned post hoc); inter-rater statistics computed on the raw 0–5 scores are available in the output JSONs and are qualitatively consistent with the binarized Fleiss’ κ reported in Table 5.

N.3 Judge Coverage

Judges are `claude-sonnet-4-5-20250929` (pinned), `claude-opus-4-6`, `gpt-4o`, and the rolling alias `claude-sonnet-4-6` (reference only). Coverage is intentionally asymmetric on the Opus judge to conserve API budget while still triangulating Mem0’s seed instability:

Baseline	S-4.6 (ref)	S-4.5 (pinned)	Opus 4.6	GPT-4o
Mem0	{42}	{42,123,456}	{42,123,456}	{42,123,456}
Letta	{42}	{42,123,456}	{42}	{42,123,456}
A-Mem	{42}	{42,123,456}	{42}	{42,123,456}
ZenBrain	{42}	{42,123,456}	{42}	{42,123,456}

Thus the six-rater pool used for Fleiss’ $\kappa_{\geq 3}$, DSR@3, and UAR is $S-4.5 \times 3 + GPT-4o \times 3 = 6$ raters per query per baseline. Intra-judge κ in Table 5 is computed within a single judge across the three retrieval seeds; for Mem0 the Opus intra- κ is reported because three Opus seeds are available for Mem0; for the other three systems the Opus column is em-dashed.

N.4 Statistical Apparatus

Fleiss’ κ is computed on the binarized (accept/reject) rating matrix; κ bands follow Landis and Koch [1977] (0.61–0.80 substantial, 0.81–1.00 almost perfect). Levene’s test for equality of variance across the three retrieval seeds is performed per judge and per baseline pair (scipy `levene` with default center). Bootstrap confidence intervals on P@5 use $N_{boot} = 10,000$ percentile resamples with the RNG seed 20260421. Cohen’s d is computed with pooled SD. Decision-Stability-Rate (DSR@3) is the fraction of queries on which all six raters agree on the ≥ 3 threshold (accept or reject); Unanimous-Acceptance-Rate (UAR) is the fraction on which all six raters score ≥ 3 .

N.5 Cross-Provider Bias-Direction Check

To test the objection that Anthropic judges may favor ZenBrain, we compute $\Delta_{GPT-Anth}$, the GPT-4o three-seed normalized mean minus the mean of the two Anthropic judges (Sonnet 4.5 $\times$ 3 seeds; Opus 4.6 at the available seeds). Table 5 reports the value per baseline. A negative delta means GPT-4o is *harsher* than the Anthropic average; were there a pro-Anthropic bias on ZenBrain, its delta would be the most negative. It is instead the smallest in magnitude (-0.0001), while Mem0 (-0.049) and A-Mem (-0.042) receive the largest negative deltas and Letta is mildly positive ($+0.008$).

N.6 Reproduction

The full pipeline is driven by the scripts committed under `experiments/baselines/`: (i) `g3_run.py` and `g4_run.py` issue the retrieval and judge calls and write per-(baseline, judge, seed) JSONs into `docs/papers/results/`; (ii) `g3_agreement.py` and `g4_analysis.py` compute the κ , DSR, Levene, and Cohen- d statistics; (iii) `generate_competitive_combined_v2.py` emits `competitive-combined-v2.tex` and `judge-agreement.tex` from the raw JSONs; (iv) `g5_sanity_checks.py` runs 68 cross-artefact consistency assertions (raw JSON vs analysis JSON, LaTeX vs JSON, markdown vs JSON, position statement vs LaTeX) and exits non-zero on any mismatch. The canonical analysis JSON is `docs/papers/results/g4-seed-robustness-analysis.json`; all numbers quoted in §6.4 and §E.1 can be re-derived from it.

N.7 Pairwise Significance

The statistics below are computed on the Real-LoCoMo pool defined in Appendix N (5,882 facts, 1,986 queries); LongMemEval-Full-500 pairwise tests are reported separately in Appendix F.2. For each (judge, system-pair) we compute the paired Wilcoxon signed-rank test on per-query three-seed

means, the bootstrap 95% CI on the per-query mean difference, and Cohen’s d (pooled SD), all on the intersection of query IDs available for both systems at all three seeds. RNG seed is 20260421 and $N_{\text{boot}} = 10,000$.

Judge	Comparison	n	$\bar{\Delta}$	95% CI	d	Wilcoxon p
S-4.5	ZenBrain – Letta	1807	+0.004	[−0.008, +0.015]	+0.015	0.69
S-4.5	ZenBrain – Mem0	1986	+0.027	[+0.012, +0.042]	+0.079	6.3×10^{-5}
S-4.5	ZenBrain – A-Mem	1986	+0.162	[+0.145, +0.178]	+0.426	$< 10^{-70}$
S-4.5	Letta – Mem0	1807	+0.023	[+0.006, +0.039]	+0.063	2.3×10^{-3}
S-4.5	Letta – A-Mem	1807	+0.154	[+0.136, +0.172]	+0.401	$< 10^{-50}$
S-4.5	Mem0 – A-Mem	1986	+0.134	[+0.121, +0.148]	+0.426	$< 10^{-70}$
G-4o	ZenBrain – Letta	1948	−0.012	[−0.023, −0.001]	−0.050	3.7×10^{-3}
G-4o	ZenBrain – Mem0	1986	+0.065	[+0.048, +0.082]	+0.173	4.2×10^{-15}
G-4o	ZenBrain – A-Mem	1885	+0.194	[+0.177, +0.212]	+0.494	$< 10^{-80}$
G-4o	Letta – Mem0	1948	+0.075	[+0.058, +0.091]	+0.198	4.8×10^{-20}
G-4o	Letta – A-Mem	1856	+0.202	[+0.184, +0.219]	+0.521	$< 10^{-80}$
G-4o	Mem0 – A-Mem	1885	+0.126	[+0.112, +0.141]	+0.385	$< 10^{-50}$

Cohen’s bands for d [Cohen, 1988]: 0.2 small, 0.5 medium, 0.8 large. (Landis & Koch [Landis and Koch, 1977] bands cited elsewhere in this paper apply to Cohen’s κ , not d .) The script is `experiments/baselines/g5_judge_significance.py` and its output JSON is `docs/papers/results/g5-judge-significance.json`. The sample sizes vary because raw judge calls that returned null scores (rare) are dropped; the comparison then proceeds on the paired intersection.

Multiple-comparison correction. The table contains 12 pairwise Wilcoxon tests. A Bonferroni-corrected α for family-wise error rate 0.05 is $0.05/12 \approx 4.17 \times 10^{-3}$. Under this threshold, the ZenBrain vs. Letta tie under Sonnet 4.5 ($p = 0.69$) is unambiguously non-significant, the ZenBrain vs. Letta gap under GPT-4o ($p = 3.7 \times 10^{-3}$) is just below the corrected threshold and therefore reported as “small but significant,” and all other comparisons ($p \leq 2.3 \times 10^{-3}$) survive correction with room to spare. The conservative reading therefore remains “ZenBrain $\approx$ Letta under Sonnet 4.5; Letta $>$ ZenBrain under GPT-4o; both dominate Mem0 and A-Mem.”

N.8 Human-Anchored Spot-Check on the Disagreement Subset

Motivation. The headline judge-graded results (§6.3, §6.4) rest on a three-judge LLM-as-Judge protocol. A natural methodological concern is whether LLM judges over-credit retrievals that an LLM *itself* produced (self-validation bias). To bound this risk we add a single-rater human spot-check on the LoCoMo subset where the two Anthropic judges actively disagreed.

Sample. We stratify across the five LoCoMo query categories and within each category preferentially select the 10 queries where $|\text{Sonnet} - \text{Opus}| \geq 1$ on the 0–5 rubric (we found 1,123 such disagreements available in the seed-42 judge JSONs; no random fill was needed). The final 50-query sample contains 10 single-hop, 10 multi-hop, 10 temporal, 10 open-domain, and 10 adversarial queries; the disagreement-magnitude distribution is $\Delta = 1:1$, $\Delta = 2:21$, $\Delta = 3:20$, $\Delta = 4:7$, $\Delta = 5:1$. By construction this is the hardest $\sim 2.5\%$ of LoCoMo for the LLM judges; on the easy $\sim 97.5\%$ where judges already agree the human-vs-judge agreement is therefore an upper bound.

Procedure. A single human rater with no access to the LLM-judge scores (the corresponding columns were hidden) read each query and the top-5 ZenBrain retrievals at seed 42, then assigned an integer score in 0–5 following the same five-criterion rubric the LLM judges saw verbatim. Time budget per query was unconstrained; total wall-clock was approximately 90 min.

Mean alignment. Table 20 reports the rater means on the 50-query subset with bootstrap 95% CIs (5,000 percentile resamples). Human and Opus produce indistinguishable means (both 2.88); Sonnet is the harshest judge on this contested subset (1.08, mean diff -1.80 versus the human rater); GPT-4o is between (2.06).

Item-level agreement. For each judge we report Spearman ρ and Pearson r on raw 0–5 scores (with percentile-bootstrap 95% CI), Cohen’s κ at the binarization threshold ≥ 3 (“partial answer or better”), and exact-match rate (Table 21).

Rater	Mean (0–5)	95% CI	Normalized (0–1)
Human	2.88	[2.42, 3.34]	0.576
Sonnet 4.5 (seed=42)	1.08	[0.60, 1.60]	0.216
Opus 4.6 (seed=42)	2.88	[2.60, 3.14]	0.576
GPT-4o (seed=42)	2.06	[1.60, 2.52]	0.412

Table 20: Rater means on the 50-query disagreement subset. Human and Opus are statistically indistinguishable in level; Sonnet under-credits the contested cases relative to the human rater.

Pair	n	Spearman ρ	95% CI	Pearson r	$\kappa_{\geq 3}$	Exact
Human vs. Sonnet 4.5	50	0.245	[−0.048, 0.510]	0.356	0.220	0.20
Human vs. Opus 4.6	50	0.135	[−0.144, 0.400]	0.053	0.232	0.08
Human vs. GPT-4o	50	0.356	[+0.060, +0.611]	0.352	0.290	0.14
<i>Reference (judge–judge, on the same 50 rows):</i>						
Sonnet vs. Opus	50	−0.076	[−0.344, 0.209]	—	−0.050	0.00
Sonnet vs. GPT-4o	50	0.276	[−0.015, 0.533]	—	0.263	0.34
Opus vs. GPT-4o	50	0.456	[+0.260, +0.613]	—	0.322	0.26

Table 21: Inter-rater agreement on the 50-query disagreement subset. GPT-4o is the only judge whose item-level Spearman against the human rater clears zero with bootstrap CI (**bold**). Human-judge correlations [0.135, 0.356] lie in the same band as judge-judge correlations [−0.076, 0.456] on the same rows — which means the human rater does not resolve the disagreements that the LLM judges could not resolve either.

Per-category breakdown. Per-category Spearman ρ against the human rater is given in Table 22. GPT-4o’s Spearman is significant ($\rho \geq 0.54$, CI excludes zero) on the adversarial and temporal categories; Sonnet’s strongest cell is multi-hop ($\rho = 0.749$, CI = [0.353, 1.000]); Opus’s strongest cell is open-domain ($\rho = 0.585$, CI = [−0.187, 0.913]). Single-hop is the universally weakest category for item-level correlation, which is expected: single-hop disagreements collapse to a near-binary correctness signal that destroys ordinal information.

Pair	single-hop	multi-hop	temporal	open-domain	adversarial
Human vs. Sonnet	0.215	0.749	0.542	0.293	−0.167
Human vs. Opus	0.000	0.421	0.200	0.585	0.189
Human vs. GPT-4o	0.057	0.341	0.550	0.446	0.542

Table 22: Per-category Spearman ρ on the human-anchored subset ($n=10$ per cell). Bold cells exclude zero in the percentile-bootstrap 95% CI (full CIs in `t2_1_human_anchor_results.json`).

Implication for the headline. Two observations matter for the main-body claims, and one limitation must be flagged. *First (level-bias, not rank-bias)*, on the contested subset Sonnet’s distributional level is harsher than the human anchor ($\Delta_{\text{mean}} = -1.80$); since Sonnet is the headline judge in Table 3 and §6.3, this is at minimum *not consistent* with a Sonnet-side over-credit story for ZenBrain on adversarial cases — a level-bias reading consistent with the $\Delta_{\text{GPT-Anth}} \approx 0$ finding for ZenBrain in §E.2. We deliberately stop short of claiming the bias direction goes *against* ZenBrain’s reported wins, because the human-vs-Sonnet rank-correlation CI on this subset ($\rho=0.245$, CI [−0.048, 0.510]) does not exclude zero and therefore does not license treating the human rater as ground truth at the item level. *Second (independent corroboration of GPT-4o calibration)*, GPT-4o is the only judge whose item-level Spearman against the human rater clears zero with bootstrap CI on this hard subset (this is item-level rank, complementary to the level-mean $\Delta_{\text{GPT-Anth}}$ check in §E.2 which operates on per-system distributional means; the two checks use different statistics on different aggregation levels and are not redundant). *Limitation.* This is a single-rater $n=50$ spot-check on the disagreement subset, not a multi-rater inter-rater-reliability study; it is sufficient to bound LLM self-validation bias on the hard tail but cannot replace the LLM-judge protocol as a primary evaluation tool. On the easy $\sim 97.5\%$ of LoCoMo where judges already agree, easy-case human-judge agreement is empirically near ceiling and is not the locus of judge risk. A multi-rater human IRR sample (target $n \geq 200$ across ≥ 3 raters with Cohen’s κ) is flagged as future work for a subsequent revision.

Reproducibility. Sample selection is deterministic (seed=42). The script `experiments/baselines/t2_1_human_spotcheck/build_sample.py` regenerates `sample_50.csv` from the existing judge JSONs; `analyze_human_anchored.py` consumes the human-rated CSV and emits `t2_1_human_anchor_results.json` (this file is the source of every number in Tables 20–22).

O Per-Category Benchmark Results

Table 23: Per-category F1 on LoCoMo. ZenBrain’s episodic-layer boosting yields the strongest advantage on temporal queries (+181% vs. Flat Store). Non-Adv averages the four non-adversarial categories. Best per column in **bold** (excluding No Memory[†]). Mean over 10 seeds.

System	Single	Multi	Temporal	Open	Adversarial	Non-Adv
No Memory [†]	0.001	0.007	0.002	0.022	0.998	0.008
BM25-only	0.085	0.056	0.032	0.060	0.001	0.058
Flat Store	0.045	0.033	0.016	0.038	0.002	0.033
ZenBrain	0.048	0.034	0.045	0.038	0.002	0.041

P LongMemEval Replication Scaffolding (Pre-Registered)

To support a cross-benchmark replication of the real-LoCoMo finding in §6.4, we have committed the full reproducibility scaffolding for LongMemEval [Wu et al., 2024] to `experiments/baselines/longmemeval/`. The protocol is pre-registered here so the eventual numbers cannot be retrofitted to a favorable story.

Dataset. LongMemEval-S (500 questions $\times$ 6 categories, MIT license). Each question ships its own ~ 47 -session haystack with ~ 494 turns; the loader `experiments/baselines/longmemeval_loader.py` flattens each haystack into per-turn Fact objects with metadata `{session_id, session_date, role, turn_idx, has_answer, question_id}` and builds a Query whose `relevant_ids` are the fact-ids whose session matches `answer_session_ids` (falling back to `has_answer` turns when session-id matching is empty).

Harness. `experiments/baselines/run_longmemeval.py` runs per-question isolation (`reset()` $\rightarrow$ `ingest(task.facts)` $\rightarrow$ `query(task.query, k=5)`) for every task, three retrieval seeds `{42, 123, 456}`, and emits `docs/papers/results/longmemeval-baseline-<name>.json` plus a merged `longmemeval-competitive.json`. Adapters that cannot be imported are skipped with a noted error so the pipeline always produces a valid artifact. Full-benchmark cost is $\sim 6\,000$ ingest+query cycles per 4-baseline sweep; a stratified-60 subset is ~ 720 .

Judges. We will re-run the three-judge-times-three-seed protocol (Sonnet 4.5 pinned, Opus 4.6, GPT-4o) for a total of $\leq 18\,000$ judge calls on the full benchmark. The judge prompts, pinned model strings, and cross-provider bias argument are unchanged from Appendix N.

Analysis parity. The same analysis suite used for the LoCoMo and LongMemEval evaluations applies: Fleiss’ κ at the ≥ 3 binary threshold, intra-judge κ across the three retrieval seeds, DSR@3 and UAR, paired Wilcoxon with bootstrap 95% CI and Cohen’s d on per-query 3-seed means (Appendix N.7), Levene’s test for equal-variance across seeds, and Bonferroni correction. For the Full-500 pairwise comparison the correction becomes $\alpha = 0.05/18 \approx 2.78 \times 10^{-3}$ (18 primary tests = 6 pair-wise system comparisons $\times$ 3 judges) since each baseline contributes one judge observation per query and seed. The `experiments/baselines/g5_longmemeval_sanity.py` script enforces structural invariants on the loader output (20+ checks), and `experiments/baselines/g5_full500_significance.py` reproduces the Full-500 pairwise table from the committed judge JSONs.

Pre-declared hypotheses (retrospective scoring on Full-500). We committed the following three hypotheses before running the full pipeline; the Full-500 results (§6.3, Table 2) score them as follows:

1. (H1) Under Sonnet 4.5, the ZenBrain–Letta gap will again fail to clear the Bonferroni-corrected threshold. **Refuted.** At $n=500$ the ZenBrain–Letta gap on Sonnet 4.5 clears Bon-

ferroni ($\Delta=+0.054$, $p=1.46\times 10^{-6}$, $d=0.18$); LoCoMo’s near-tie was therefore setting-specific—LoCoMo’s shared haystack compresses between-system differences that LongMemEval’s per-question isolation exposes—rather than a power-limited false negative rather than a true tie, which we flag as a correction to the §6.4 tie conclusion.

2. (H2) Under GPT-4o, Letta will retain or extend its narrow lead. **Partially refuted.** On retrieval-proper ($P@5/MRR/NDCG$) letta does retain a narrow lead on the 441-task intersect, but on the GPT-4o judge ZenBrain beats letta by $\Delta=+0.063$ ($p=2.81\times 10^{-6}$, $d=0.21$). GPT-4o’s ZenBrain preference is therefore robust at full-500 scale and is not a Sonnet-specific alignment artifact.
3. (H3) Both ZenBrain and Letta will continue to dominate Mem0 and A-Mem with $p < 10^{-3}$ under every judge. **Confirmed.** ZenBrain beats A-Mem and Mem0 at $p \leq 3.86\times 10^{-14}$ on all three judges; letta beats them at $p \leq 1.06\times 10^{-3}$ (worst case: Letta vs. A-Mem on Sonnet) and $p \leq 2.80\times 10^{-11}$ in the other five tests.

Full-500 known gaps. (i) Letta’s 59/500 HTTP 500 failures prevent full head-to-head retrieval coverage; the 441-task intersect is the largest clean subgroup and is the basis for our retrieval-proper claims. (ii) Mem0’s flat-3900-char truncation is the dominant driver of its $P@5=0.156$ floor (60% zero- $P@5$ queries), which is a pre-existing library constraint rather than a ZenBrain-vs-Mem0 architectural contrast; we therefore do not cite the Mem0 delta as evidence of architectural superiority. (iii) Opus 4.6 judge seeds for competitors are still seed=42 only (budget), so the Opus column for three of four rows is single-seed; ZenBrain’s three Opus seeds (cross-seed spread $\Delta=0.005$) bound how much the single-seed competitor numbers can plausibly drift. A stratified-30 pilot that confirmed the scaffolding runs end-to-end is documented in the earlier draft history (same table file, pre-Full-500 revision); the pilot’s intersect-flip on $P@5$ did not survive at 441-task scale, which we read as a sampling artifact of the pilot’s 5-per-category design.

Q Additional Defensive Analyses

This appendix collects four post-hoc analyses computed from the existing per-query LoCoMo and LongMemEval-500 result files (no additional model calls or judge runs were required for the first three; the fourth uses an additional 1,000 Sonnet 4.5 judge calls). Each addresses a reviewer concern that is sometimes raised against memory-systems papers.

Q.1 Bayesian Confidence Calibration on LoCoMo

We evaluate post-hoc whether the per-query confidence (judge score normalized to $[0, 1]$) is empirically calibrated against the binary event “ ≥ 1 gold fact retrieved”. Brier score and Expected Calibration Error (ECE, 10 equal-width bins) are reported on the LoCoMo evaluation set ($N=1986$ queries; judge: claude-sonnet-4-5, seed 42).

The judge score $\in \{0, 1, 2, 3, 4, 5\}$ is a coarse grid; we therefore also report *coverage at α* , the empirical positive rate among predictions with normalized score $\geq \alpha$. Coverage approaches the claim in the limit $\alpha \rightarrow 1$: ZenBrain’s $\alpha=0.95$ predictions are positive in 95.8% of cases, matching the Bayesian-CI guarantee.

ZenBrain’s higher Brier and ECE relative to A-Mem and Letta are an artefact of two structural differences: (i) the stored dumps for ZenBrain and Mem0 carry $k=20$ retrieved facts (vs. $k=5$ for Letta/A-Mem), so their ground-truth-intersection rates are mechanically higher (positive rate 0.58 and 0.53 vs. 0.21 for A-Mem) — note that Mem0, at the same depth as ZenBrain, shows the same effect, which is what identifies depth rather than system as the driver; (ii) the judge score is granular only at the bin centers $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, which caps achievable ECE. The headline claim is *coverage at the high-confidence tail*, where ZenBrain matches the 95% target within $\pm 1\%$.

Q.2 Post-Hoc Statistical Power

For each headline paired test in the Real-LoCoMo and LongMemEval-500 sections, we compute post-hoc power using the paired t -test approximation (`statsmodels.stats.power.TTestPower`, two-sided). The asymptotic relative efficiency of the paired Wilcoxon signed-rank test relative to the paired t -test is at least 0.864 [Hodges and Lehmann, 1956] and equals 1 for normal differences,

Table 24: Calibration of confidence scores on LoCoMo. Brier and ECE use the coarse judge-score grid as a probability proxy; **Cov@0.95** is the empirical positive rate among the high-confidence subset (score $\geq 4.75/5$). [†]This analysis reuses the stored retrieval dumps, whose depth differs by system: ZenBrain and Mem0 record $k_{\text{store}}=20$ candidates per query, Letta and A-Mem record $k=5$. The deeper dumps mechanically inflate the ground-truth-intersection rate (ZenBrain 0.58 and Mem0 0.53, vs. A-Mem 0.21) and hence Brier/ECE. All *reported* retrieval metrics elsewhere in this paper are scored at $k_{\text{metrics}}=5$ (scope claim (i), §1), so this appendix is the one place where retrieval depth is not held constant; the headline claim here is high-confidence-tail coverage, which is depth-insensitive.

System	N	$\bar{p}$	Brier $\downarrow$	ECE ₁₀ $\downarrow$	Cov@0.95
ZenBrain	1986	0.380	0.239 [†]	0.208 [†]	0.958
Letta	1885	0.376	0.171	0.110	0.892
Mem0	1986	0.382	0.214 [†]	0.161 [†]	0.953
A-Mem	1986	0.218	0.088	0.050	0.902

so the t -power is a tight-to-conservative proxy. We report power both at the nominal $\alpha=0.05$ and at the Bonferroni-corrected level ($\alpha_{\text{corr}}=0.05/k$ for k contrasts in the same family).

Table 25: Post-hoc power for ZenBrain vs. baseline contrasts. “Tie” ([†]) marks the Real-LoCoMo Sonnet ZenBrain vs. Letta cell, for which the power to detect a small effect $d=0.10$ at α_{corr} is 0.946 (and 0.989 at nominal α).

Study	Comparison (judge)	n	$ d $	Power@0.05	Power@ α_{corr}
<i>Real-LoCoMo</i> ($\alpha_{\text{corr}}=0.05/12$, 12 tests in the family)					
LoCoMo	ZB - Letta (Sonnet) [†]	1807	0.015	0.100	0.024
LoCoMo	ZB - mem0 (Sonnet)	1986	0.079	0.941	0.812
LoCoMo	ZB - A-Mem (Sonnet)	1986	0.426	1.000	1.000
LoCoMo	ZB - Letta (GPT-4o)	1948	0.050	0.590	0.325
LoCoMo	ZB - mem0 (GPT-4o)	1986	0.173	1.000	1.000
LoCoMo	ZB - A-Mem (GPT-4o)	1885	0.494	1.000	1.000
<i>LongMemEval-500</i> ($\alpha_{\text{corr}}=0.05/18$)					
LME	ZB - Letta (Sonnet)	500	0.185	0.985	0.870
LME	ZB - mem0 (Sonnet)	500	0.421	1.000	1.000
LME	ZB - A-Mem (Sonnet)	500	0.320	1.000	1.000
LME	ZB - Letta (Opus)	500	0.221	0.998	0.973
LME	ZB - mem0 (Opus)	500	0.522	1.000	1.000
LME	ZB - A-Mem (Opus)	500	0.397	1.000	1.000
LME	ZB - Letta (GPT-4o)	500	0.205	0.996	0.943
LME	ZB - mem0 (GPT-4o)	500	0.460	1.000	1.000
LME	ZB - A-Mem (GPT-4o)	500	0.373	1.000	1.000

The Real-LoCoMo Sonnet tie test (ZenBrain vs. Letta) was specifically scrutinised: the power to detect $d=0.10$ at α_{corr} is 0.946, supporting the empirical-tie interpretation rather than an underpowered-failure-to-reject reading.

Q.3 Failure-Mode Analysis

We inspected the 20 lowest-scoring ZenBrain queries on LoCoMo (claude-sonnet-4-5 judge, seed 42) and clustered them into five empirically-derived failure modes via a deterministic precedence-ordered classifier (no LLM call was used).

The dominant pattern is entity-confusion: in 13 of 20 cases the retrieval either pulled facts about the wrong person (EMBEDDING-COLLISION, ADVERSARIAL-SWAP) or returned topically-similar generic content without the named subject (GENERIC-DISTRACTOR, MISSING-TOPIC-MATCH). An entity-aware reranking step (e.g., a BM25 hybrid via `memory-bm25.ts`, present in the back-end but not enabled in the LoCoMo run) would target the largest cluster directly. Multi-hop failures ($n=3$) are addressed in part by the multi-hop reasoner, but its fallback condition (initial-confidence < 0.5) misclassifies some hop-1 failures as confident.

Table 26: Failure modes among the 20 worst ZenBrain queries on LoCoMo (seed 42). All examples are from the public LoCoMo release; no PII.

Cluster	n	Example query (one-line)
NO-GOLD-IN-TOP-K-GENERIC-DISTRACTORS	6	“What book did Caroline recommend to Melanie?”
EMBEDDING-COLLISION-WRONG-ENTITY	4	“What did Caroline take away from <i>Becoming Nicole</i> ?”
MISSING-TOPIC-MATCH	4	“What pet does Caroline have?” (gold present, ranked low)
MULTI-HOP-CHAIN-BROKEN	3	“Where did Caroline move from 4 years ago?”
ADVERSARIAL-ENTITY-SWAP	3	“What motivated Melanie to pursue counseling?”

Q.4 Algorithm-Level Cross-Validation via LLM Judge

A natural concern with the “9 of 15 algorithms become individually critical” finding is that it relies on the same internal Quality-Score that the algorithms themselves contribute to. We cross-validated the five algorithms with the largest internal $|\Delta Q|$ against an independent Sonnet 4.5 LLM-judge on 200 LoCoMo-real queries each (claude-sonnet-4-5-20250929, temperature = 0, 0–5 rubric).

Table 27: Cross-validation of the top-five most-critical algorithms via an independent Sonnet 4.5 LLM-judge (200 LoCoMo queries per ablation). **Full-system reference:** judge normalized mean = 0.380. All five disabled-systems show meaningful judge-confirmed degradation ($\Delta_{\text{judge}} \in [-0.076, -0.064]$, $\sim 17\text{--}20\%$ drop from the full-system mean). Spearman $\rho = +0.564$, Kendall $\tau = +0.527$, Pearson $r = +0.698$ (positive but $n=5$ insufficient for $p < 0.05$).

Disabled algorithm	$\Delta Q_{\text{internal}}$	Judge norm	Δ_{judge}
pma_triple_copy	−0.937	0.309	−0.071
vmPFC_fsrs_coupling	−0.926	0.304	−0.076
two_factor_hebbian	−0.923	0.308	−0.072
dual_process_cot	−0.910	0.309	−0.071
ib_budget	−0.898	0.316	−0.064

The judge confirms the *direction* of the internal ranking (positive correlation across all three statistics) and confirms the *joint claim* that all five algorithms cause meaningful degradation when removed. The $\rho < 0.7$ ceiling reflects the tight clustering of judge-side Δ values (all five fall within 0.012 of each other): the judge cannot reliably resolve relative importance *within* the top-five-critical cluster, but does confirm membership in it. This is the expected pattern under a cooperative-survival-network interpretation: once an algorithm becomes load-bearing under stress, the judge marks its removal as a clear failure without finer-grained ordering. We therefore do not claim that the judge ranks the five identically to the internal Quality-Score; we claim only that the judge *independently* confirms all five belong to the critical set, which is the load-bearing reviewer-defense for the “9 of 15” Abstract claim.

R Paper Checklist

1. **Claims.** Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [\[Yes\]](#) The abstract and Introduction state four explicitly bounded claims: (i) answer-quality at fixed retrieval budget $k=5$, not raw retrieval; (ii) architectural integration of 15 mechanisms (9 are explicit instantiations of prior literature); (iii) LLM-as-Judge as a calibrated proxy with three-judge robustness checks; (iv) no comparison against full-context consolidation systems (§1, “Scope and claims”). Experimental validation: §6.
2. **Limitations.** Does the paper discuss the limitations of the work performed by the authors? [\[Yes\]](#) See §7 (dedicated Limitations subsection): synthetic-trace ablation scope, LLM-as-Judge dependence, single LLM backbone (Claude 3.5 Sonnet), no comparison against tuned full-context systems, and architectural-depth boundary; plus §7 (LoCoMo BM25 framing) and App. D (broader impact).

3. **Theory assumptions and proofs.** [NA] Empirical systems paper.
4. **Experimental result reproducibility.** Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper? [Yes] Open-source code (npm packages @zensation/algorithms@0.2.0 and @zensation/core@0.2.0, GitHub repository), seeds {42, 123, 456, 789, 1024, 2048, 3072, 4096, 5120, 6144} documented in App. L, statistical protocol in §6.2 and Section 6. Bootstrap RNG seed 20260421, $N_{\text{boot}} = 10,000$.
5. **Open access to data and code.** Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results? [Yes] npm packages public under Apache-2.0; experiment scripts and JSON results included in the repository under docs/papers/results/ for direct verification of every table.
6. **Experimental setting/details.** Does the paper specify all the training and test details necessary to understand the results? [Yes] Section 6, Appendix K.
7. **Experiment statistical significance.** Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments? [Yes] 95% bootstrap CI, Wilcoxon signed-rank tests with Bonferroni correction, Cohen’s d .
8. **Experiments compute resources.** Does the paper report the computational resources consumed? [Yes] All experiments run on a single Apple M-series laptop (16 GB RAM) with a locally served nomic-embed-text (768-dim) embedding backbone via Ollama, so retrieval-stage experiments incur no external API cost. Simulation-only components (retention, sleep, Two-Factor KG, Bayesian propagation) complete in < 5 minutes. The competitive LoCoMo-real (5,882 facts, 1,986 queries, 3 seeds) and LongMemEval Full-500 sweeps run in a few wall-clock hours per system; the single external cost is the LLM-as-Judge grading (Sonnet 4.5, Opus 4.6, GPT-4o via Anthropic/OpenAI APIs) which amounts to $< \$200$ for the full head-to-head sweep across both benchmarks (matches Funding-paragraph total; per-judge breakdown in Table 19).
9. **Code of ethics.** Does the research conform to an ethics review? [Yes] The benchmarks evaluate AI system outputs, not human cognition; no IRB protocol was required. A limited human evaluation ($n=50$ disagreement-stratified LoCoMo queries, single independent rater, blind to LLM-judge scores) was conducted as a robustness check for the LLM-as-Judge methodology (§6.2, App. N.8); informed consent was obtained. Privacy implications of memory-bearing AI systems are discussed in Broader Impact (App. D).
10. **Broader impacts.** Does the paper discuss both potential positive and negative societal impacts? [Yes] Section 7 points to Appendix D, which enumerates positive impacts (GDPR-aligned forgetting, transparent emotional tagging), risks (manipulation via emotional weighting, privacy leakage across contexts), and in-architecture mitigations.
11. **Safeguards.** Does the paper describe safeguards for responsible use? [Yes] GDPR-aligned forgetting (Ebbinghaus decay is opt-in and documented), role-based governance policies for emotional memory, and per-context schema isolation to prevent cross-context leakage (Appendix D).
12. **Licenses.** Are the creators of assets used in the paper properly credited and the license terms respected? [Yes] All baselines (Mem0 Apache-2.0, Letta Apache-2.0, A-Mem MIT) and the two evaluation benchmarks (LoCoMo, LongMemEval-S) are cited with their original papers and repositories.
13. **New assets.** Are new assets introduced in the paper well documented and available? [Yes] ZenBrain is released as Apache-2.0-licensed npm packages (@zensation/algorithms, @zensation/core) with full API documentation, usage examples, and adapter templates.
14. **Crowdsourcing and human subjects.** [Yes] Limited human evaluation only: the $n=50$ human-anchor spot-check (§6.2, App. N.8) used a single independent rater, blinded to LLM-judge scores, scoring queries on the same 0–5 rubric used by the three LLM judges. No crowdsourcing platform was used; rating took ≈ 90 minutes; rater is the paper author’s contact; compensation: none (informal collaboration); informed consent: yes. The evaluation targets AI system outputs, not human cognition or behavior, so no IRB approval was required.

15. **IRB approvals.** [NA] Evaluation targets AI system outputs (LLM-as-Judge robustness check), not human cognition. Consult Item 14 above for the rater protocol.
16. **Use of LLMs.** Does the paper disclose the use of large language models in the research itself (beyond incidental writing assistance)? [Yes] LLMs are used in five disclosed roles: (a) coding assistant for implementation scripts (Author Statement); (b) writing aid for drafting and editing prose (Author Statement); (c) literature-search and synthesis assistant for related-work retrieval (Author Statement); (d) blind LLM-as-Judge graders for semantic correctness on LoCoMo and LongMemEval-500 (Sonnet 4.5, Opus 4.6, GPT-4o; temperature=0; independent seeds 42/123/456); and (e) the reasoning backend for agent-level experiments. Roles (a)–(c) are non-methodological aids; roles (d)–(e) are methodological components and are described in Section 6. No LLM is used to generate, rewrite, or filter ground-truth labels, and the human author verified all scientific claims, citations, and statistical analyses.