© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722568

Layer-Conditional Emotional Weighting: Does Emotion-Aware Memory Consolidation Need to Be Layer-Specific, and Does It Transfer from Dialogue Personalization to Agentic Task Memory?

A Stage 1 Registered Report: Introduction, Related Work, and Pre-Registered Experimental Protocol

V. PRANAV ABHINAY GUPTA
Independent Researcher, Bengaluru, Karnataka, India

Abstract*—Personalized LLM agents increasingly rely on structured long-term memory, and a growing line of work conditions memory consolidation and retrieval on a computational emotional-significance signal rather than semantic relevance alone. Existing emotion-aware memory systems, however, treat memory as a single undifferentiated store and are evaluated only on conversational personalization. We ask a narrower, previously untested question: should the influence of emotional significance on memory consolidation and retrieval be conditioned on the target memory layer (e.g., episodic vs. semantic), rather than applied as a single global weight, and do gains from emotion-aware memory observed on dialogue-personalization benchmarks transfer to long-horizon, multi-session agentic task memory? We propose Layer-Conditional Emotional Weighting (LCEW), a memory architecture built on the standard CoALA memory taxonomy in which every consolidation and retrieval weight is generated by a shared, low-parameter hypernetwork conditioned on a learned layer embedding, rather than looked up from an independent per-layer table, so that global-weight scoring (Park et al.) and flat emotion-weighted scoring (Dynamic Affective Memory Management) are recovered as provable limiting cases of a single ridge-regularized objective rather than as separate architectures, and in which memory-layer transitions are modeled as calibrated probabilities rather than hardcoded rules. We specify the full mathematical formulation, algorithm, and a pre-registered experimental protocol comparing LCEW against seven baselines, including a reimplementation of the closest existing system, on both an existing affective-dialogue benchmark and a new agentic-task transfer benchmark. This manuscript is submitted as a Stage 1 Registered Report: no results are reported, and none should be expected at this stage; the object of review is the formal specification, the isolation of layer-conditioning as the untested variable, and the protocol required to test it honestly, including a mandatory replication of a documented personalization-bias risk before any positive claim is made. In-principle acceptance of this Stage 1 manuscript would commit the experiments in Sections XII-XIII to be run and reported, without further alteration to the hypotheses or protocol, in a Stage 2 submission.

***Index Terms*—agent memory, long-term memory, emotional memory, memory consolidation, affective computing, LLM agents, personalization, memory forgetting, registered report, pre-registration**

I. INTRODUCTION

Long-term memory has become a standard component of large language model (LLM) based agents, motivated by the fixed size of the context window and by the goal of personalization across sessions [1], [2]. A parallel line of work argues that emotional significance, not just semantic relevance, recency, or frequency, should shape which memories are retained and retrieved, on the grounds that human episodic memory is itself mood- and salience-dependent [3], [4]. This has produced systems that tag memories with an affective state and bias consolidation or retrieval toward emotionally significant content.

This project began as an attempt to design a new "Adaptive Multi-Layer Emotional Memory Architecture." A structured literature audit, summarized in Section III, showed that essentially every individual component of that proposal already exists: the layered memory taxonomy is the Cognitive Architectures for Language Agents (CoALA) framework [5]; the weighted multi-factor consolidation and retrieval score is structurally the Generative Agents formulation of Park et al. [2]; Ebbinghaus-style decay is MemoryBank [3]; and, most directly, emotion-conditioned memory consolidation with a dedicated benchmark already exists as Dynamic Affective Memory Management [4], which uses a Bayesian, entropy-minimizing update rule over a memory vector database, evaluated

on DABench, a benchmark of emotional expression and emotional change toward objects.

Rather than re-deriving this system under a new name, this paper isolates the one structural choice that, to our knowledge, has not been tested in the published literature: whether the weight given to emotional significance should vary by destination memory layer, and whether any resulting gain generalizes beyond conversational personalization to agentic, tool-using, multi-session tasks — precisely where current computer-use agents are known to fail, not on tool selection, but on losing track of state and constraints across a long horizon [6]. If emotional salience is a useful memory-retention signal, it should plausibly help there — but this has not been shown.

II. PROBLEM DEFINITION

Given a stream of interaction events produced by a user and an agent across one or more sessions, a memory system must decide, for each event: (i) whether to retain it at all; (ii) which memory layer(s) it should occupy; (iii) how its retention priority should change over time through decay, reinforcement, or forgetting; and (iv) which memories to surface for a given query or task state. Prior work parameterizes (ii)–(iv) using a single, global scoring function over signals such as recency, importance, and relevance [2], optionally including an emotional-significance term [4]. We study whether replacing the global emotional-significance weight with a layer-conditional weight function improves consolidation calibration and retrieval quality, and whether such improvements, if present, hold on agentic task-memory benchmarks and not only on dialogue-personalization benchmarks.

III. RELATED WORK

A. Layered and Hierarchical Agent Memory

CoALA [5] formalizes the working, episodic, semantic, and procedural memory taxonomy this paper adopts without modification. MemGPT [1] implements a two-tier main-context/external-context system with LLM-driven self-editing, later evolved into Letta. Subsequent systems, including A-MEM, MemoryBank [3], Mem0, Zep/Graphiti, and MemMachine, extend this with graph structure, temporal decay, and profile/episodic separation. None of these introduces an emotional signal as a first-class consolidation input; emotion, where present in the broader literature, is typically treated as an application-layer concern rather than a memory-architecture parameter.

B. Memory Scoring Functions

Park et al. [2] define the retrieval score as a weighted sum of recency, importance, and relevance, with hand-set, typically equal weights. This is the direct structural ancestor of the multi-factor weighted-score pattern used throughout the field. Recent critiques argue that hand-set weights are a known weakness: importance in the Park et al. formulation is static, assigned once at write time, and governs retrieval only, not encoding depth or forgetting; neither their weights nor comparable systems' weights are learned from outcomes [7], [8]. We adopt this critique directly: any weight in our formulation that plausibly benefits from data, in particular task-usefulness, is specified as a quantity to be fit, not asserted.

C. Emotion-Aware Memory

The closest prior system is Dynamic Affective Memory Management [4], which introduces a Bayesian-inspired update minimizing memory entropy over a single memory vector database, evaluated on DABench with reported ablations validating the update rule against memory bloat. Related systems include an emotion-aware AR companion combining recognition, memory compression, and multi-agent modularity [9], appraisal-based chain-of-emotion architectures operating in pleasure-arousal-dominance (PAD) space [10], and a benchmark for proactive memory retrieval specifically in emotional-support settings [11]. All of these operate over an undifferentiated memory store; none conditions the emotional weight on memory-layer identity.

D. A Documented Risk

Fang et al. [12] show that injecting user-memory profiles into emotional-reasoning tasks produces systematically divergent, and demographically biased, emotional interpretations. Because LCEW explicitly increases the influence of emotional signal on what an agent remembers and retrieves, this finding is a direct risk to the system proposed here and is treated as a mandatory evaluation, not an afterthought (Section XII).

E. Forgetting and Agentic Long-Horizon Memory

MemoryBank's Ebbinghaus-curve decay [3] and outcome-driven memory-governance work [8] establish forgetting as a first-class, principled mechanism rather than simple deletion. Separately,

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722568

computer-use agent benchmarks report that current state-of-the-art agents fail on long-horizon, multi-step tasks primarily because they lose track of constraints and miss information introduced mid-task — a memory problem more than a tool-use problem [6]. Agent S [13] already augments a computer-control loop with episodic and narrative memory, but without any emotional-salience component. This is the applied setting in which we test transfer (Section X).

TABLE I. NOVELTY POSITIONING VS. PUBLISHED SYSTEMS

System	Lay.	Emo.	L-C	Cal.	Lrn.	Trf.
MemGPT/Letta [1]	Y	N	N	N	N	N
CoALA [5]	Y	N	N	N	N	N
Park et al. [2]	N	N	N	N	N	N
MemoryBank [3]	P	N	N	N	N	N
DAMM [4]	N	Y	N	N	N	N
Agent S [13]	P	N	N	N	N	Y
LCEW (ours)	Y	Y	Y	Y	Y	Y

Lay. = layered memory; Emo. = emotion-aware; L-C = layer-conditional emotion weight; Cal. = calibrated layer transitions; Lrn. = learned (not hand-set) weights; Trf. = evaluated on agentic task-memory transfer. Y = yes, P = partial, N = no.

IV. RESEARCH GAP

No existing published system tests layer-conditional emotional weighting, as opposed to a single global emotional weight, and no existing emotion-aware memory system has been evaluated on agentic, tool-using, long-horizon task memory rather than solely on dialogue-personalization benchmarks such as DABench. Both are narrow, falsifiable questions; neither justifies inventing a new architecture, and both can be tested by modifying an existing, standard memory pipeline.

V. LCEW ARCHITECTURE

LCEW instantiates the CoALA taxonomy with four concrete stores: short-term context, working memory, semantic memory, and episodic memory, and adds a single new architectural element relative to prior work: the weight vector applied to each consolidation and retrieval signal is a function of the destination memory layer, $w(\ell')$, rather than a constant. A secondary, smaller change is that layer transitions are emitted as calibrated probabilities rather than deterministic threshold rules, which makes the promotion policy itself measurable via calibration error, rather than merely inspectable by outcome. Fig. 1 shows the overall system.

Fig. 1. LCEW system architecture. The shaded tool-router block and downstream computer-control/robotics stack are explicitly out of scope for the v1 study (Section IX).

The reasoning core, tool use, computer control, robotics, and any multimodal perception beyond text embedding and emotion appraisal are explicitly out of scope for this study (Section IX). Including them would confound the single variable under test.

VI. MATHEMATICAL FORMULATION

A. Memory Item

A memory item is a tuple $m = (x, \tau_c, \tau_a, \ell, v, e, n)$ where x is natural-language content, τ_c and τ_a are creation and last-access timestamps, $\ell \in$ {short-term, working, semantic, episodic} is the current layer, $v \in \mathbb{R}^d$ is an embedding, $e \in \mathbb{R}^3$ is an emotion vector restricted to (valence, arousal, confidence), and n is an access count.

B. Layer-Conditional Consolidation Score

For candidate promotion of memory m into target layer ℓ', the score is:

$$M_{\text{score}}(m, \ell') = w_r(\ell')\cdot R(\text{m}) + w_e(\ell')\cdot E(\text{m}) + w_f(\ell')\cdot F(\text{m}) \quad (1)$$

$$+ w_t(\ell')\cdot T(\text{m}) + w_u(\ell')\cdot U(\text{m}) + w_s(\ell')\cdot S(\text{m}) \quad (2)$$

where each weight $w\cdot(\ell')$ is layer-specific — the structural difference from Park et al. [2] and from Dynamic Affective Memory Management [4], both of

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722568

which use a single global weight set — and all component scores lie in [0,1] via min-max or logistic normalization, following the normalization convention of [2]. R(m) = exp(−λ(t − τa)) is recency, with λ following MemoryBank's Ebbinghaus-inspired decay [3] rather than an arbitrary constant; E(m) = ‖e‖ is emotional significance; F(m) is normalized reinforcement/access count; T(m) is task usefulness, fit from outcome labels per the learned-value-function critique [7] rather than hand-set, defaulting to zero when outcome labels are unavailable; U(m) is a user-correction signal; and S(m) is a source-reliability prior (user-stated > tool-retrieved > model-inferred).

To keep this claim falsifiable rather than a relabeling of six additional free parameters, w(ℓ′) is not stored as an independent per-layer table but generated by a shared hypernetwork: w(ℓ′) = ψ(zℓ′; Φ), where zℓ′ ∈ ℝ^k is a learned embedding of layer ℓ′ and ψ is a small MLP with parameters Φ shared across all layers. Training minimizes L(Φ, {zℓ′}) = L_task + λ·Σℓ′ ‖w(ℓ′) − w̄‖², where w̄ is the mean weight vector across layers and λ ≥ 0 is a single ridge coefficient controlling how strongly layers are permitted to diverge.

Proposition 1 (Nesting). As λ → ∞, the penalty forces w(ℓ′) → w̄ for every ℓ′, and Eqs. (1)–(2) collapse to a single global weight vector applied uniformly across layers — the Park et al. [2] scoring rule when E(m) is dropped, and the flat emotion-weighted rule of Dynamic Affective Memory Management [4] when E(m) is retained. As λ → 0, w(ℓ′) is unconstrained per layer, recovering full layer-conditional weighting. Whether layer-conditioning helps therefore reduces to a single scalar sweep over λ on one model rather than a discrete comparison between structurally different systems (Section XV), and a new memory layer added at deployment time needs only a new embedding zℓ′ rather than six freshly-fit weights, since ψ already encodes how layer identity maps to weight.

C. Calibrated Layer Transition

$$P(\ell \to \ell' \mid m) = \sigma(M_{\text{score}}(m, \ell') - \theta_{\ell \to \ell'}) \quad (3)$$

with a layer-pair-specific threshold θ, calibrated post hoc against held-out promotion outcomes (Section XII), rather than asserted.

D. Emotional State Update

$$E(t+1) = (1-\beta)\cdot E(t) + \beta \cdot \Delta(\text{interaction}_t, \text{appraisal}(m_{\text{ret}}), \text{state}_t) \quad (4)$$

an exponential moving average with a bounded, specified appraisal function Δ(·), chosen over an unconstrained f(·) so that the update rule is falsifiable and β can be ablated (Section XV).

E. Layer-Gated Retrieval Score

$$\text{Retrieval}(m \mid q) = \gamma_{\text{sim}}\cdot\cos(v_q, v_m) + \gamma_{\text{layer}}(\ell)\cdot 1[\ell \in \text{layers}(q)] \quad (5)$$

$$+ \gamma_e \cdot \text{align}(e_q, e_m) + \gamma_{\text{task}}\cdot\text{match}(m, q) \quad (6)$$

with γlayer(·) implementing adaptive layer selection as a query-conditioned gate rather than a fixed constant, illustrated in Fig. 3.

VII. MEMORY CONSOLIDATION ALGORITHM

Algorithm 1 formalizes the layer-conditional consolidation pass, also shown as a flowchart in Fig. 2.

```
Algorithm 1: consolidate(m, L, ψ, Φ, z, θ)
  best_layer, best_score ← NULL, −∞
  for ℓ' in L:
    w ← ψ(z[ℓ'], Φ)                  // hypernetwork, not a flat lookup
    score ← w.r*R(m) + w.e*E(m) + w.f*F(m)
            + w.t*T(m) + w.u*U(m) + w.s*S(m)
    if score > best_score:
      best_layer, best_score ← ℓ', score
  p ← sigmoid(best_score − θ[m.layer, best_layer])
  if random() < p:
    move(m, to=best_layer)
  log_promotion_event(m, best_layer, p)
  return m
```

Fig. 2. Layer-conditional consolidation algorithm (Sec. 7).

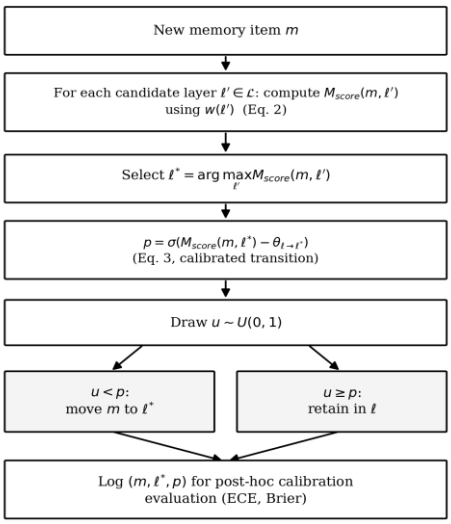

Fig. 2. Layer-conditional consolidation algorithm (Section VII).

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722568

VIII. EMOTIONAL STATE MODEL

```
Algorithm 2: update_emotion(E_t, x,
mem, state, β)
  appraisal ← appraise(x, mem, state)
  delta ← f_delta(appraisal)   //
bounded (val., arousal, conf.)
  return (1 − β) * E_t + β * delta
```

appraise(·) is grounded in appraisal-theory literature [10] rather than an unconstrained free-form emotion generator, so its output space is fixed and auditable.

IX. ADAPTIVE RETRIEVAL

```
Algorithm 3: retrieve(q, store, γ, k)
  qv, qe ← encode(q)
  candidates ← []
  for layer in select_layers(q,
γ.layer_selector):
    for m in store[layer]:
      s ← γ.sim*cos(qv, m.v) +
γ.layer(layer)
            + γ.e*align(qe, m.e) +
γ.task*match(m, q)
      candidates.append((s, m))
  sort candidates descending by s
  return top k memories
```

Fig. 3. Layer-gated adaptive retrieval pipeline (Sec. 9).

Fig. 3. Layer-gated adaptive retrieval pipeline (Section IX).

Complexity. Consolidation (Algorithm 1) is O(|L|·d) per candidate memory, where |L| is the number of layers (fixed and small, |L| = 4 here) and d is the embedding dimension used inside sim/align terms, so the hypernetwork adds a constant O(|L|·|Φ|) overhead over a flat-weight scorer and does not change the asymptotic order. Retrieval (Algorithm 3) is O(|L|·|store|) in the worst case if select_layers(q, γ.layer_selector) is not restrictive, and O(|store_selected_layers|) when the query-conditioned gate γ.layer(·) prunes to a subset of layers before scoring, which is the common case; an approximate nearest-neighbor index over v within each layer reduces this further to O(|L|·log|store|) and is used in the reference implementation for the 500-session horizon. Space overhead relative to a single global-weight system is O(|L|·k) for the layer embeddings plus O(|Φ|) for the shared hypernetwork, both independent of the number of stored memories, versus O(|L|·|w|) for an unshared per-layer weight table — for the six-term score in Eq. (1)–(2), this is a constant, not a scaling, difference, but it is the constant that Proposition 1 exploits to make λ a single interpretable knob rather than six independent ones per layer.

X. FORGETTING MECHANISM

Forgetting follows the established lever taxonomy of importance, merge, decay, and eviction [8], applied per layer using the layer-specific λ from Section VI-B. No new forgetting theory is claimed here; the only layer-conditional element is that decay thresholds are layer-specific rather than global.

XI. SYSTEM IMPLEMENTATION

The reference implementation targets two open-weight local LLMs, drawn from the Qwen and Llama instruct-tuned families, with exact versions to be pinned at experiment time, to test whether LCEW's gains, if any, are model-agnostic. Embeddings use a fixed sentence-embedding model shared across all baselines to avoid confounding retrieval quality with embedding-model choice. All weight vectors w(ℓ′), thresholds θ, and gate parameters γ are logged in full per run for reproducibility.

XII. EXPERIMENTAL SETUP

Datasets: DABench [4] is reused directly for affective-dialogue evaluation, enabling a same-benchmark comparison against the closest existing system. LongMemEval/LoCoMo-style long-term-conversation benchmarks are used for general recall and personalization. A new, narrowly-scoped agentic-transfer benchmark, consisting of multi-session tool-use tasks with emotionally salient events interleaved among neutral task steps, is constructed specifically to test the transfer question in Section IV; this is the only new dataset introduced in this study. Session horizons of 1, 10, 50, 100, and 500 are used to study degradation and promotion behavior over time. A mandatory bias check replicates the Personalization Trap protocol [12] on LCEW's own outputs before any personalization-improvement claim is made public.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722568

XIII. BASELINES

- No memory (context window only).
- Full raw conversation history.
- Standard flat vector RAG.
- Recency-only decay memory.
- Park et al. [2] global-weight recency/importance/relevance score, no emotion term (isolates the layering effect).
- Dynamic Affective Memory Management [4], reimplemented as closely as license/availability permits (isolates whether layer-conditioning beats the existing flat emotional approach; the single most important comparison in this paper).
- CoALA-style layered memory, no emotion term (isolates the emotion contribution from the layering contribution).

XIV. STAGE 2 REPORTING PLAN

Consistent with the Stage 1 registered-report format, no experimental results are reported in this manuscript, and reviewers are asked to evaluate the protocol rather than outcomes not yet collected. Following in-principle acceptance, the Stage 2 submission will report, per baseline and per benchmark, without deviation from the metrics specified here: Recall@K, Precision@K, and mean reciprocal rank; preference-consistency and contradiction rate; promotion-probability calibration via Brier score and expected calibration error; token consumption, latency, and memory footprint across the five session horizons; and task-completion rate on the agentic-transfer benchmark. All numbers will be reported only after the protocol in Sections XII–XIII has actually been run, with pre-registered primary hypotheses fixed in advance (Section XVI).

XV. ABLATION STUDY

Planned ablations, results pending per Section XIV: a continuous sweep of the ridge coefficient λ from Proposition 1 (Section VI-B) across at least five values spanning $\lambda \to 0$ (full LCEW, unconstrained per-layer weights) to $\lambda \to \infty$ (weights collapse to a single global vector, recovering baseline 6 exactly rather than approximately); minus the emotion term entirely, approximating baseline 5/7; minus learned T(m) in favor of a fixed constant, testing whether the learned-weight critique matters in practice here; minus the forgetting mechanism; minus calibrated transition probabilities in favor of a deterministic threshold rule; and a dimensionality ablation on e (3-D versus a larger emotion space) to test whether the reduced dimensionality in Section VI-A costs anything measurable.

XVI. DISCUSSION

The contribution here is deliberately narrow. Emotional weighting of memory is not new; layered memory is not new; what is untested is whether the same weight for emotional salience should apply uniformly across memory types, or whether, consistent with the intuition that episodic memory is more mood-dependent than semantic memory in cognitive-science accounts, the weight should vary by destination layer. If results show no measurable difference between layer-conditional and global weighting, that is a valid and useful negative result: it would suggest the added complexity of per-layer weighting is not justified, and that simpler global-weight systems such as [4] are sufficient. Equally, if LCEW improves dialogue-personalization metrics but shows no transfer to the agentic benchmark, that result is informative on its own: it would indicate that emotion-aware memory research to date, evaluated almost exclusively on conversational benchmarks, has been answering a narrower question than the "personalized AI agent" framing typically implies. Paired comparisons against each baseline will use bootstrapped confidence intervals with Holm–Bonferroni correction for multiple comparisons, and the primary hypothesis, that layer-conditioning beats baseline 6 on the transfer benchmark, will be pre-registered before any run.

XVII. LIMITATIONS

The proposed appraisal function and emotion dimensionality are design choices grounded in cited prior work but not uniquely determined by it. The agentic-transfer benchmark introduced in this study has not been externally validated; its results should be treated as preliminary evidence rather than a general claim about agentic memory. The reimplementation of the closest competing system may not exactly reproduce its reported numbers if code or weights are not fully available, which will be disclosed explicitly wherever it applies. Results obtained from open-weight local LLMs may not generalize to larger proprietary models.

XVIII. ETHICS AND SAFETY

Emotion-conditioned memory retention and retrieval directly increases a system's sensitivity to user affect, which is precisely the mechanism flagged as bias-prone in [12]. This paper treats the bias-replication check in Section XII as a required gate before any deployment-oriented claim, not an optional robustness check. Systems that remember emotionally significant events for longer also raise standard long-term-memory privacy concerns; any deployed instantiation should implement user-controlled deletion and should not use emotion-weighted retention as a justification for retaining sensitive personal data longer than a user would reasonably expect.

XIX. CONCLUSION

Adaptive multi-layer emotional memory, as originally conceived, restates existing work: CoALA's layering, Park et al.'s weighted scoring, and existing emotion-conditioned consolidation are each already published. This paper narrows the contribution to a single testable question, whether emotional weighting should be layer-conditional rather than global, and whether any resulting gain transfers from dialogue personalization to agentic task memory, and specifies the full formulation, algorithm, and protocol needed to answer it honestly, including comparison against the closest existing system and a mandatory bias-replication check. As a Stage 1 registered report, acceptance of this manuscript is a commitment to run exactly this protocol and report whatever the results show, including a null result on layer-conditioning or a failure to transfer to agentic memory; either outcome would be reported as the Stage 2 contribution, since both answer the falsifiable question posed in Section IV.

XX. FUTURE WORK

If layer-conditional weighting shows a measurable, transferring benefit, natural extensions include learning $w(\ell')$ end-to-end rather than fitting components separately, extending the retrieval gate to a full multimodal setting including vision and screen state, and only then integrating the memory system into a full computer-use or robotics agent loop, evaluated on long-horizon benchmarks where memory-driven state tracking is the documented bottleneck [6]. The multimodal, computer-control, robotics, and gesture-interface components considered in the original project scope remain legitimate systems and demonstration contributions in their own right, but should be pursued as a separate paper once the underlying memory claim in this paper has been validated or falsified.

ACKNOWLEDGMENT

The author thanks the maintainers of the open-source agent-memory frameworks cited in this work for making their systems and benchmarks available for comparison.

REFERENCES

[1] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, “MemGPT: Towards LLMs as operating systems,” arXiv:2310.08560, 2023.

[2] J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative agents: Interactive simulacra of human behavior,” in Proc. 36th Annu. ACM Symp. User Interface Softw. Technol. (UIST), 2023, arXiv:2304.03442.

[3] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, “MemoryBank: Enhancing large language models with long-term memory,” in Proc. AAAI Conf. Artif. Intell., 2024.

[4] “Dynamic affective memory management for personalized LLM agents,” arXiv:2510.27418, 2025.

[5] T. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths, “Cognitive architectures for language agents,” arXiv:2309.02427, 2023.

[6] “OSWorld 2.0: Benchmarking computer use agents on long-horizon real-world tasks,” arXiv:2606.29537, 2026.

[7] “Learning what to remember: A cognitively grounded multi-factor value model for agentic memory,” arXiv:2606.12945, 2026.

[8] “When to forget: A memory governance primitive,” arXiv:2604.12007, 2026.

[9] “Livia: An emotion-aware AR companion powered by modular AI agents and progressive memory compression,” arXiv:2509.05298, 2025.

[10] “An appraisal-based chain-of-emotion architecture for affective language model game agents,” PMC, 2024.

[11] “ENPMR-Bench: Benchmarking proactive memory retrieval for emotional support agents,” arXiv:2605.27240, 2026.

[12] X. Fang, W. Xu, Y. Zhang, S. Eckman, S. Nickleach, and C. K. Reddy, "The personalization trap: How user memory alters emotional reasoning in LLMs," arXiv:2510.09905, 2025.

[13] S. Agashe et al., "Agent S: An experience-augmented hierarchical planning framework for computer-use agents," 2024.