Continual Knowledge Updating in LLM Systems: Learning Through Multi-Timescale Memory Dynamics

Andreas Pattichis [1] **Constantine Dovrolis** [1 2]

Abstract

LLMs are trained once, then deployed into a world that never stops changing. External memory compensates for this, but most systems manage it explicitly rather than letting it adapt on its own. Biological memory works differently: coupled multi-timescale dynamics make new associations immediately usable, strengthen what repetition confirms, and let the rest fade. We argue that external memory should follow a similar principle. In *Memini*, this view takes the form of an associative memory that organizes knowledge as a directed graph. Each edge carries two coupled internal variables, one fast and one slow, following the Benna-Fusi model of synaptic consolidation. From this coupling, episodic sensitivity, gradual consolidation, and selective forgetting are expected to emerge as facets of a single mechanism, reframing external memory as a learning substrate that reorganizes through its own dynamics. This workshop article describes an early-stage conceptual design without experimental evaluation.

1. Introduction

Large language models (LLMs) are increasingly deployed in settings where factual knowledge does not remain static (Dhingra et al., 2022; Lazaridou et al., 2021). New facts emerge, old ones become obsolete, and associations between concepts strengthen, weaken, or disappear. A model frozen at deployment cannot keep up with any of this. The challenge is not merely one of access to new information. It is one of selective adaptation: strengthening what ongoing evidence confirms, letting go of what it no longer supports, and doing so continuously as new experience arrives.

[1]The Cyprus Institute, Nicosia, Cyprus [2]School of Computer Science, Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Andreas Pattichis <a.pattichis@cyi.ac.cy>, Constantine Dovrolis <c.dovrolis@cyi.ac.cy>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI".

Existing work addresses this through two main paradigms, and neither resolves the problem at the level it requires. The first is parametric continual learning, which adapts the model by updating its weights (Shi et al., 2025). This changes the language processing substrate directly, but at significant cost: parameter updates risk catastrophic forgetting, are computationally expensive, and are often unavailable to downstream developers who access the model as a service (De Lange et al., 2022; Sun et al., 2022). These difficulties are compounded by a more fundamental limitation. Different kinds of knowledge change at different rates, yet a single parameter-update mechanism treats all evidence as equally worth absorbing. This makes forgetting both more likely and harder to control (Wu et al., 2024). The second paradigm keeps the backbone frozen and augments it with external memory (Mialon et al., 2023; Zaharia et al., 2024). This improves access to prior information, but the mechanisms that store, organize, and retrieve do not themselves change over time. The memory may grow or be pruned, but it does not reorganize: associations are not strengthened by repeated evidence, not weakened by absence, and retrieval does not shift as new information arrives.

We therefore argue for a sharper distinction. A system whose memory merely grows is accumulating. A system whose memory consolidates, forgets, and reshapes its organization in response to experience is learning (Kumaran et al., 2016). This distinction motivates a different paradigm, one in which learning happens through structured memory dynamics. **The central claim is that continual learning in LLM systems can be implemented as the reorganization of associative memory through multi-timescale dynamics.** Prior memory-augmented LLM systems store or retrieve information; they do not learn by reshaping their association structure over time. Three principles define this paradigm. First, dynamics: associations evolve through reinforcement and decay, not just by being appended to. Second, multi-timescale learning: fast processes capture recent evidence while slow processes consolidate what repetition confirms, and these coexist on every association (Benna & Fusi, 2016). Third, selectivity: the system does not store everything or retain everything, and what it forgets is as consequential as what it keeps (Nørby, 2015).

We instantiate this paradigm in *Memini*[1], a persistent directed association graph in which each edge carries two coupled internal variables, one fast and one slow, following the Benna-Fusi synaptic consolidation model (Benna & Fusi, 2016). The interaction between these two variables is intended to produce episodic sensitivity, gradual consolidation, and selective forgetting as emergent behaviors of the same dynamics, without separate stores, explicit rules, or gating modules. Retrieval operates through spreading activation over these evolving edge weights, so the same query issued at different times follows different paths and recovers different context. The novelty is not an external memory, a graph, or a retrieval procedure. It is that the memory itself is a dynamical state, driven directly by the incoming document stream rather than by retrieval activity or an external manager such as an agent or LLM revising it from outside. Its current organization reflects prior experience and shapes the assimilation of future experience.

2. Related Work

Continual learning for LLMs has been studied extensively as a parametric problem (Shi et al., 2025; Wu et al., 2024), with dominant approaches including replay, regularization, architectural expansion, and targeted knowledge editing. In all of these approaches, learning occurs in the model's parameters, and the central challenge is managing the stability-plasticity tradeoff so that new knowledge can be absorbed without overwriting what came before (De Lange et al., 2022).

Other approaches keep the backbone frozen and augment it with external memory (Mialon et al., 2023). These range from vector-store retrieval-augmented generation (Lewis et al., 2020) to agent memory with tiered hierarchies (Packer et al., 2024) or memory streams with periodic reflection and summarization (Park et al., 2023). But in each case, the management is external, applied by fixed rules rather than driven by ongoing evidence.

Within this paradigm, three lines of work push back against this rigidity, each in a different direction. The first adds structural organization to memory through entity graphs, PageRank retrieval, or LLM-revised note links (Edge et al., 2025; Gutiérrez et al., 2025; Xu et al., 2025), but the edges remain static as new documents arrive. The second makes retrieval adaptive over a graph through spreading activation or query-aware reweighting (Jiang et al., 2026; Pavlović et al., 2025; Lau et al., 2026), yet the weights (fixed at indexing, set by time decay, or modulated only per query) are never shaped by repeated evidence. The third targets forgetting through single-timescale decay with recall-driven reinforcement (Zhong et al., 2024; Honda et al., 2026), but a single timescale cannot be both fast enough for recent evidence and slow enough to preserve what repetition confirms.

[1]*Meminī* is Latin for "I remember" or "I hold in mind." Grammatically perfect in form but present in meaning, it denotes a present state of retention shaped by prior experience.

Table 1. Memory-augmented LLM systems compared on the three properties that together define Memini: *evolving* edge weights driven autonomously by incoming evidence, *multi-timescale* dynamics coupling fast access with slow consolidation, and *selective forgetting* of unsupported associations. Explicit deletion, per-query reweighting, threshold archival, and recall-driven reinforcement do not qualify.

System	Evolving weights	Multi-timescale	Selective forgetting
Standard RAG	✗	✗	✗
MemGPT	✗	✗	✗
GraphRAG	✗	✗	✗
HippoRAG 2	✗	✗	✗
A-MEM	✗	✗	✗
SYNAPSE, SA-RAG, CatRAG	✗	✗	✗
MemoryBank, ACT-R-LLM	✗	✗	✓
Memini	✓	✓	✓

Table 1 summarizes the gap: no system combines evolving edge dynamics, multi-timescale consolidation, and emergent selective forgetting. The memory may be structured, traversed, or decayed, but it does not learn.

3. System Design

3.1. Memory Architecture

The memory is a directed graph $G = (V, E)$ in which nodes represent entities or concepts identified in the incoming document stream and edges represent directed associations between them. Three structural properties distinguish these edges from those in prior graph-based systems (Table 1). They are unlabeled, with meaning encoded in the pattern, strength, and direction of connectivity rather than in explicit relation types. They are directed, with $A \to B$ distinct from $B \to A$, reflecting the asymmetry of associative strength found in spike-timing-dependent plasticity (Bi & Poo, 1998). They carry persistent, evolving weights that change as new documents arrive.

Each directed edge (A, B) carries an internal state consisting of two coupled variables, w_{fast} and w_{slow}. Together, the tuple $(A, B, w_{\text{fast}}, w_{\text{slow}})$ constitutes the minimal unit of memory in this system. The fast variable w_{fast} is the accessible component. It responds directly to co-occurrence events, decays quickly without reinforcement, and is the sole variable read during retrieval. The slow variable w_{slow} is the hidden consolidation component. It receives no direct external input, changes only through bidirectional coupling with w_{fast}, and decays much more slowly. Although w_{slow} is never read directly, it influences retrieval by sustaining w_{fast} at non-zero levels for associations that have been consolidated through repeated reinforcement. The

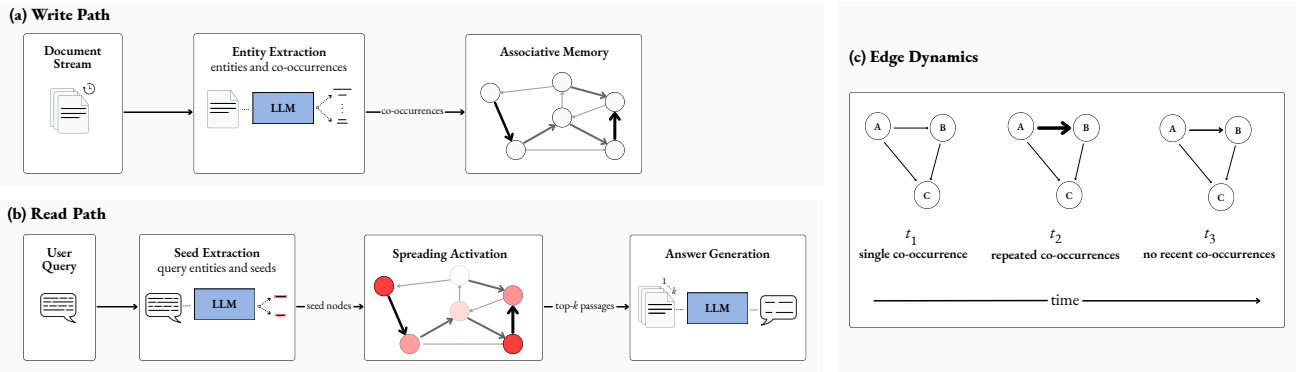

Figure 1. Overview of *Memini*. **(a) Write path.** The LLM extracts entities and co-occurrences from arriving documents. Each co-occurrence updates an edge of the association graph through coupled fast and slow variables. Edge thickness reflects w_{fast}. **(b) Read path.** Entities from the user's query activate seed nodes (red). Activation propagates along w_{fast}-weighted edges for a bounded number of steps, and the passages associated with the highest-activation nodes are returned as context for answer generation. **(c) Edge dynamics.** A single edge $(A \to B)$ tracked across three moments while the rest of the graph is held fixed. A single co-occurrence creates a transient trace (t_1). Repeated co-occurrences consolidate the association (t_2). Without recent reinforcement it fades but persists (t_3).

full state of the memory at any time t is therefore the graph $G(t) = (V(t), E(t))$ together with the edge states $\{(w_{\text{fast}}(e,t),\, w_{\text{slow}}(e,t))\}_{e \in E(t)}$.

The backbone LLM and the memory system occupy different layers (Figure 1). The LLM reads and writes text; the memory does the learning. This separation has two consequences. The backbone can be replaced or upgraded without disturbing what the memory has learned, and the memory can evolve continuously without modifying the backbone.

3.2. Multi-Timescale Edge Dynamics

Without dynamics, the graph can only grow by appending new nodes and edges, and associations set at creation remain fixed indefinitely. A single-timescale update rule cannot resolve the resulting plasticity-stability tension. Learning fast enough to incorporate new evidence risks overwriting what came before, while learning slowly enough to retain prior structure prevents timely adaptation. Work on synaptic consolidation has shown that memory systems require multiple internal timescales to remain both plastic and stable over extended experience (Benna & Fusi, 2016; Zenke & Laborieux, 2025), a principle that generalizes the cascade model of Fusi et al. (2005). The same multi-timescale principle has been validated independently in optimization theory (Behrouz et al., 2025a) and in empirical neural architectures (Behrouz et al., 2025b).

We adopt the simplest instantiation of the Benna-Fusi chain model, with two coupled variables per edge ($n = 2$) following Kaplanis et al. (2018). The dynamics on each directed edge (A, B) are:

$$\frac{dw_{\text{fast}}}{dt} = -\frac{w_{\text{fast}}}{\tau_{\text{fast}}} + C(w_{\text{slow}} - w_{\text{fast}}) + I(t), \qquad (1)$$

$$\frac{dw_{\text{slow}}}{dt} = -\frac{w_{\text{slow}}}{\tau_{\text{slow}}} + C(w_{\text{fast}} - w_{\text{slow}})\,, \qquad (2)$$

where τ_{fast} and τ_{slow} are the decay time constants ($\tau_{\text{slow}} \gg \tau_{\text{fast}}$), C is the coupling strength, and $I(t)$ is the co-occurrence-driven input defined as $I(t) = b$ when concepts A and B co-occur in a document at time t, and $I(t) = 0$ otherwise.

Three forces act on w_{fast}: decay toward zero at rate $1/\tau_{\text{fast}}$, coupling that pulls it toward w_{slow}, and input from co-occurrence events. Only two forces act on w_{slow}: a much slower decay, and coupling that pulls it toward w_{fast}. The key structural property is that w_{slow} receives no direct external input. It accumulates only indirectly, through coupling, when w_{fast} is repeatedly elevated by co-occurrence events, and once accumulated, it sustains w_{fast} through the reverse direction of the same coupling. There is no explicit negative input when A appears without B; weakening is handled through the decay terms, so that prolonged absence of reinforcement is functionally equivalent to gradual depression.

During retrieval, the system reads w_{fast} alone. This variable reflects both recency, through direct boosts from recent co-occurrence events, and consolidation, through the sustained level provided by coupling from w_{slow}. If w_{slow} were included in the retrieval weight, a recently mentioned association and a deeply consolidated one could appear similar, and the recency signal would be lost. Two edges with identical current w_{fast} but different w_{slow} values will diverge over subsequent time steps. The edge with high w_{slow} remains retrievable as coupling sustains it, while the edge with low w_{slow} decays. The system thereby differentiates episodic from consolidated associations through its own dynamics.

These dynamics make retention conditional on repeated evidence. Associations that the evidence stream continues to

support consolidate and persist, while those that no longer receive reinforcement weaken and eventually disappear. This selective retention is not imposed by an explicit rule, a threshold, or a gating module, but emerges from the dynamics themselves. The role of forgetting is thereby reframed. Continual learning has conventionally treated it as uniformly harmful, a degradation to be minimized (De Lange et al., 2022). In a system where knowledge changes over time, however, selective decay is what keeps the memory aligned with the current state of the world (Nørby, 2015). A system that retains every association indefinitely accumulates stale and contradictory evidence that degrades retrieval quality. Under this framing, what the system forgets becomes as consequential as what it retains.

3.3. Emergent Memory Properties

Three characteristic behaviors are expected to follow from the dynamics without requiring separate stores, explicit rules, or additional modules. First, a single co-occurrence creates an immediately retrievable but transient episodic trace. w_{fast} rises sharply but decays if no further evidence arrives, while w_{slow} barely changes. Second, repeated co-occurrence produces consolidation. The coupling term gradually pulls w_{slow} upward, and once w_{slow} has accumulated sufficiently, it sustains w_{fast} between reinforcement events, so that the association remains retrievable without depending on recent input. This transition from episodic trace to stable association parallels the episodic-to-semantic transition described by CLS theory (McClelland et al., 1995) and Tulving's memory taxonomy (Tulving, 1984), but here it emerges from dynamics on a single structure rather than from explicit dual systems. Third, when a consolidated association stops being reinforced, both variables decay jointly but more slowly than either would alone. The coupled decay departs from the simple exponential profile of single-timescale systems and, in the general Benna-Fusi chain model, approaches the power-law form established empirically for human memory (Wixted & Ebbesen, 1997).

These dynamics also compound. Each co-occurrence event lands on an edge whose current state reflects its entire prior history, so earlier consolidation shapes how later evidence is absorbed. The resulting memory state is trajectory-dependent rather than merely data-dependent, since the same set of documents arriving in a different order produces a different associative landscape.

3.4. Retrieval via Spreading Activation

Retrieval follows the same associative logic as the memory structure (Collins & Loftus, 1975). Entities mentioned in a query are matched to graph nodes and activated simultaneously as seed nodes, each receiving initial activation $u_i^{(0)} = 1$. Activation then propagates outward along directed, weighted edges according to

$$u_i^{(t+1)} = (1-\delta)\, u_i^{(t)} + \sum_{j \in \mathcal{N}_i^-} S \cdot \frac{w_{\text{fast}}(j \to i)}{\deg_{\text{out}}(j)}\, u_j^{(t)}, \quad (3)$$

where δ is a per-iteration retention decay, S is a global spreading factor, $w_{\text{fast}}(j \to i)$ is the persistent dynamic edge weight, and $1/\deg_{\text{out}}(j)$ is the fan effect that penalizes high-degree hub nodes (Anderson, 1974). This propagation equation is adapted from Jiang et al. (2026), with the key difference that w_{fast} values here are persistent and history-shaped rather than computed per query.

The process is a deterministic, constrained propagation procedure that runs for a fixed number of iterations T, respects edge direction, and terminates without stochastic sampling. After T iterations, nodes are ranked by final activation score and the top-k associated passages are retrieved. Retrieval reads the current memory state but does not modify it; w_{fast} and w_{slow} remain unchanged.

Because retrieval operates over experience-shaped weights, the system's effective retrieval behavior changes with experience. The same query issued at different times traverses different routes through the graph and recovers different context, since consolidated pathways carry more activation while decayed ones carry less or none. This is a structural departure from systems in which retrieval is a fixed function applied to a growing store. The effect is amplified by multi-cue convergence. All query entities seed simultaneously, each spreading its own activation wave through the graph, so retrieval naturally surfaces nodes where several independent signals converge through the association structure.

4. Discussion

Memini argues that as deployed LLM systems operate in a world that keeps shifting, the continual learning this requires can happen in the memory itself. Rather than a store managed from outside, memory becomes a substrate that evolves through coupled multi-timescale dynamics. These dynamics draw on synaptic consolidation models that resolve the same plasticity-stability tension found in biological memory. From this single mechanism, episodic sensitivity, gradual consolidation, and selective forgetting are expected to emerge together, without separate stores, explicit thresholds, or gating modules. What this opens up is a research direction in which memory is not the place where past information sits, but the place where the system learns.

Appendix A reports an initial check on a Wikipedia document stream, where the expected regimes emerge and the slow variable is shown to be necessary rather than incidental. Scaling this up to full empirical validation, including retrieval and benchmarks suited to memory that adapts to an evolving stream of evidence, is the immediate next step.

Acknowledgements

This work is funded by the European Union Horizon MSCA DN programme FINALITY (G.A. 101168816).

Impact Statement

This paper presents a theoretical design for continual knowledge updating in LLM systems through memory dynamics. We do not foresee specific negative societal consequences of this work that require further discussion here.

References

Anderson, J. R. Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 6(4): 451–474, October 1974. ISSN 0010-0285. doi: 10.1016/0010-0285(74)90021-8.

Behrouz, A., Razaviyayn, M., Zhong, P., and Mirrokni, V. Nested Learning: The Illusion of Deep Learning Architectures. In *Advances in Neural Information Processing Systems 38*. Curran Associates, Inc., 2025a.

Behrouz, A., Zhong, P., and Mirrokni, V. Titans: Learning to Memorize at Test Time. In *Advances in Neural Information Processing Systems 38*. Curran Associates, Inc., 2025b.

Benna, M. K. and Fusi, S. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19 (12):1697–1706, 2016. ISSN 1097-6256. doi: 10.1038/nn.4401.

Bi, G.-q. and Poo, M.-m. Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. *The Journal of Neuroscience*, 18(24):10464–10472, 1998. doi: 10.1523/jneurosci.18-24-10464.1998.

Collins, A. M. and Loftus, E. F. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975. ISSN 0033-295X. doi: 10.1037/0033-295X.82.6.407.

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. ISSN 0162-8828. doi: 10.1109/TPAMI.2021.3057446.

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., and Cohen, W. W. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, March 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00459.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., and Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv preprint arXiv:2404.16130, February 2025. URL https://arxiv.org/abs/2404.16130.

Fusi, S., Drew, P. J., and Abbott, L. F. Cascade Models of Synaptically Stored Memories. *Neuron*, 45(4):599–611, 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.02.001.

Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., and Su, Y. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025.

Honda, Y., Fujita, Y., Zempo, K., and Fukushima, S. Human-Like Remembering and Forgetting in LLM Agents: An ACT-R-Inspired Memory Architecture. In *Proceedings of the 13th International Conference on Human-Agent Interaction*, HAI '25, pp. 229–237, New York, NY, USA, January 2026. Association for Computing Machinery. ISBN 979-8-4007-2178-6. doi: 10.1145/3765766.3765803.

Jiang, H., Chen, J., Pan, Y., Chen, L., You, W., Zhou, Y., Zhang, R., Sikora, A., Zhao, L., Abate, Y., and Liu, T. SYNAPSE: Empowering LLM Agents with Episodic-Semantic Memory via Spreading Activation. arXiv preprint arXiv:2601.02744, 2026. URL https://arxiv.org/abs/2601.02744.

Kaplanis, C., Shanahan, M., and Clopath, C. Continual Reinforcement Learning with Complex Synapses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2497–2506. PMLR, 2018.

Kumaran, D., Hassabis, D., and McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016. doi: 10.1016/j.tics.2016.05.004.

Lau, K. H., Zhang, F., Ruan, B., Zhou, Y., Guo, Q., Zhang, R., and Zhou, X. Breaking the Static Graph: Context-Aware Traversal for Robust Retrieval-Augmented Generation. arXiv preprint arXiv:2602.01965, 2026. URL https://arxiv.org/abs/2602.01965.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liška, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kociský, T., Ruder, S., Yogatama, D., Cao, K., Young, S., and Blunsom, P. Mind the gap: Assessing temporal generalization in neural language models. In *Advances*

in Neural Information Processing Systems 34, pp. 29348–29363. Curran Associates, Inc., 2021.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33*, pp. 9459–9474. Curran Associates, Inc., 2020.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995. doi: 10.1037/0033-295x.102.3.419.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., and Scialom, T. Augmented Language Models: A Survey. *Transactions on Machine Learning Research*, March 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=jh7wH2AzKK.

Nørby, S. Why Forget? On the Adaptive Value of Memory Loss. *Perspectives on Psychological Science*, 10(5):551–578, September 2015. ISSN 1745-6916. doi: 10.1177/1745691615596787.

Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., and Gonzalez, J. E. MemGPT: Towards LLMs as Operating Systems. arXiv preprint arXiv:2310.08560, February 2024. URL https://arxiv.org/abs/2310.08560.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pp. 1–22, New York, NY, USA, October 2023. Association for Computing Machinery. ISBN 979-8-4007-0132-0. doi: 10.1145/3586183.3606763.

Pavlović, J., Krész, M., and Hajdu, L. Leveraging Spreading Activation for Improved Document Retrieval in Knowledge-Graph-Based RAG Systems. arXiv preprint arXiv:2512.15922, December 2025. URL https://arxiv.org/abs/2512.15922.

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, Z., Ebrahimi, S., and Wang, H. Continual Learning of Large Language Models: A Comprehensive Survey. *ACM Computing Surveys*, 58(5):1–42, 2025. ISSN 0360-0300. doi: 10.1145/3735633.

Sun, T., Shao, Y., Qian, H., Huang, X., and Qiu, X. Black-Box Tuning for Language-Model-as-a-Service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20841–20855. PMLR, June 2022.

Tulving, E. Précis of Elements of episodic memory. *Behavioral and Brain Sciences*, 7(2):223–238, June 1984. ISSN 0140-525X. doi: 10.1017/S0140525X0004440X.

Wixted, J. T. and Ebbesen, E. B. Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, 25(5):731–739, 1997. doi: 10.3758/bf03211316.

Wu, T., Luo, L., Li, Y.-F., Pan, S., Vu, T.-T., and Haffari, G. Continual Learning for Large Language Models: A Survey. arXiv preprint arXiv:2402.01364, February 2024. URL https://arxiv.org/abs/2402.01364.

Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., and Zhang, Y. A-MEM: Agentic Memory for LLM Agents. In *Advances in Neural Information Processing Systems 38*. Curran Associates, Inc., 2025.

Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., Zou, J., Carbin, M., Frankle, J., Rao, N., and Ghodsi, A. The Shift from Models to Compound AI Systems. Berkeley Artificial Intelligence Research Blog, February 2024. URL https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

Zenke, F. and Laborieux, A. Theories of synaptic memory consolidation and intelligent plasticity for continual learning. In *Learning and Memory: A Comprehensive Reference*, pp. 169–186. Academic Press, 2025. doi: 10.1016/b978-0-443-15754-7.00070-5.

Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y. MemoryBank: Enhancing Large Language Models with Long-Term Memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

A. Testing the Dynamics on a Wikipedia Document Stream

This appendix supplements Section 3.3 by examining the coupled dynamics defined in Equations (1) and (2) on a stream of co-occurrence events extracted from real, temporally ordered text. The aim is twofold. First, we check that the three regimes derived analytically, namely episodic sensitivity, gradual consolidation, and selective forgetting, also emerge when the input is drawn from a real document stream. Second, we test whether the slow variable is necessary, by comparing Memini against a matched single-timescale ablation across all entity pairs in the stream. We use Wikipedia articles tracking the COVID-19 pandemic because the topic has a clear, widely understood phase structure, with associations that should plausibly consolidate within a phase such as *vaccine* and *mRNA* during the vaccine rollout, and others that should fade across phases such as *bat* and *SARS-CoV-2* once the discourse moves beyond origin. The pandemic's well-known timeline therefore provides an external check on whether the dynamics behave sensibly.

A.1. Document Stream and Versioning

The document stream consists of 13 English Wikipedia articles on COVID-19 topics, ordered by the period in which each topic became prominent during the pandemic. It spans four phases, namely origin (Phase 1, step 0), containment (Phase 2, steps 1–4), vaccines (Phase 3, steps 5–7), and variants and endemic transition (Phase 4, steps 8–12).

Wikipedia articles are continuously edited, so using current revisions would contaminate early time steps with content written years after the events they describe. To prevent this, every article is fetched through the MediaWiki revision API at the last edit on or before a target date corresponding to its phase, ensuring that every sentence used for co-occurrence extraction existed in Wikipedia at that date. Table 2 lists the resulting articles, revision identifiers, and timestamps, and any reported event can be audited by visiting the corresponding revision on Wikipedia.

Table 2. Wikipedia articles used as the document stream, with the revision identifier and timestamp at which each article was fetched. Each article corresponds to one time step in the input sequence. Each revision identifier links to the exact historical revision used for content extraction, allowing any reported event to be independently verified.

Step	Article	Phase	Revision ID	Revision date
0	SARS-CoV-2	1	943272842	2020-02-29
1	COVID-19 lockdowns	2	954085088	2020-04-30
2	Face masks during the COVID-19 pandemic	2	960065515	2020-05-31
3	COVID-19 testing	2	965344908	2020-06-30
4	Hydroxychloroquine	2	970150834	2020-07-29
5	COVID-19 vaccine	3	991568447	2020-11-30
6	Moderna COVID-19 vaccine	3	1003792446	2021-01-30
7	mRNA vaccine	3	1009464972	2021-02-28
8	SARS-CoV-2 Delta variant	4	1036472511	2021-07-31
9	SARS-CoV-2 Omicron variant	4	1058010984	2021-11-30
10	Booster dose	4	1062765134	2021-12-30
11	Long COVID	4	1095797409	2022-06-30
12	Endemic COVID-19	4	1130409555	2022-12-30

A.2. Extracting the Event Stream

We track 20 entities spanning all four pandemic phases, with five entities per phase covering origin, containment, vaccines, and variants. These entities act as the nodes of the association graph. They are detected in text by case-insensitive word-boundary string matching against the 20 predefined terms and a small set of hand-curated aliases, with abbreviations such as WHO and PCR matched case-sensitively to avoid spurious hits. Co-occurrence is detected at the sentence level, with each entity pair generating at most one event per document regardless of how many sentences contain both. This choice favors stronger signals of semantic association and makes the temporal pattern across documents, rather than within them, the driver of consolidation. Across the 13 documents, the procedure produces 124 events covering 68 unique entity pairs.

Each pair is integrated independently using the same parameters. We use the simplest two-variable instantiation of the Benna-Fusi chain ($n = 2$), with $\tau_{\text{fast}} = 2$, $\tau_{\text{slow}} = 10$, coupling $C = 0.2$, binary input $b = 1$, and a forward Euler step $\Delta t = 1$ corresponding to one document per step. Both variables are clamped to be non-negative. The timescale ratio $\tau_{\text{slow}}/\tau_{\text{fast}} = 5$ is chosen so that the three expected regimes are clearly distinguishable within the 13-step window. The remaining parameters are not tuned per pair and are held fixed across all events.

A.3. Visualizing the Dynamics across Regimes

Figure 2 shows the w_{fast} and w_{slow} trajectories for four entity pairs across the 13-document stream under identical parameters. The four were selected so that each illustrates one of the event-pattern groups later defined in Table 3, with all differences between panels arising entirely from the temporal pattern of co-occurrence events.

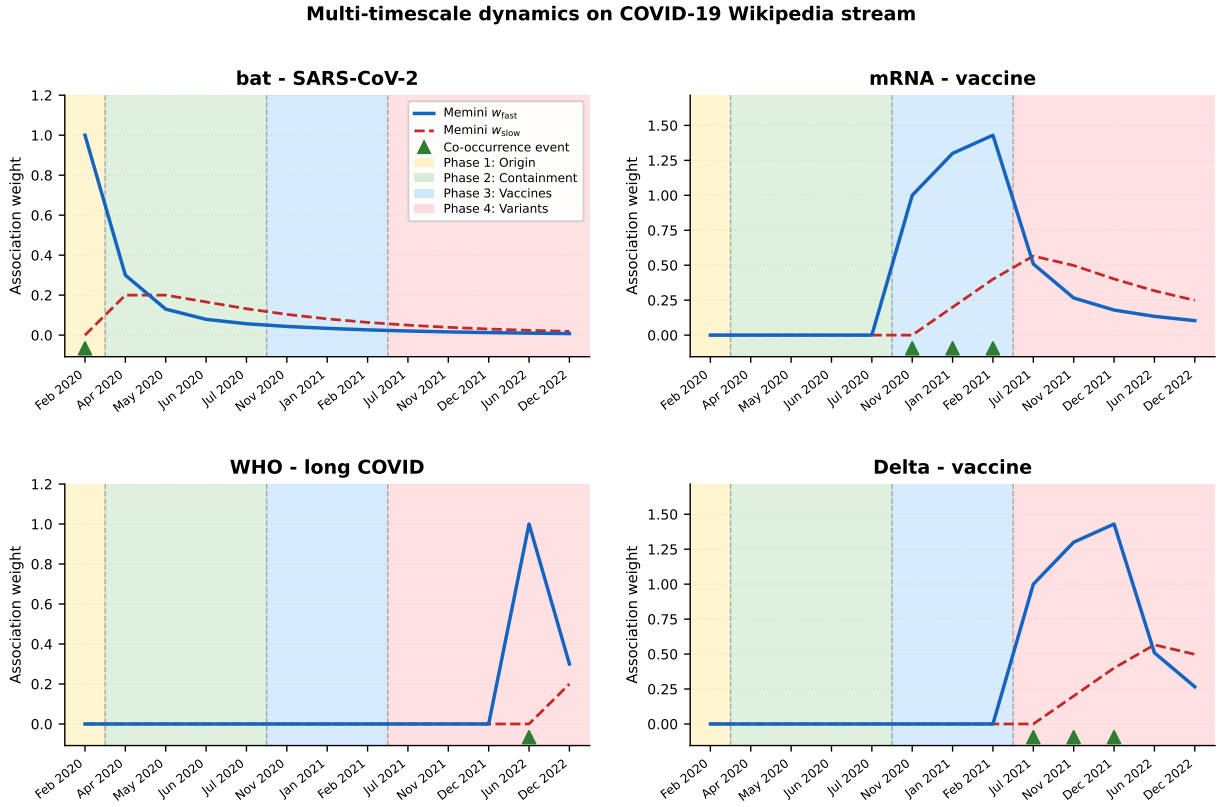

Figure 2. Multi-timescale association dynamics on the COVID-19 Wikipedia stream. Solid blue: w_{fast}. Dashed red: w_{slow}. Green triangles: co-occurrence events, each corresponding to a specific sentence in a versioned revision. Background colors indicate pandemic phase. Each panel illustrates one of the regimes analyzed in Table 3.

The panels illustrate the three regimes from Section 3.3. *Bat – SARS-CoV-2* appears once at step 0 in the description of the virus's suspected origin in bats and never again. This is the episodic regime, where w_{fast} spikes and decays within a few steps while w_{slow} rises briefly through coupling but never accumulates. *mRNA – vaccine* appears in three consecutive Phase 3 articles tracking the vaccine rollout. The clustered events drive w_{slow} upward, and the elevated w_{slow} then sustains w_{fast} above the episodic baseline well into 2022 after direct input stops, illustrating consolidation. *WHO – long COVID* appears once near the end of the stream, reflecting the late institutional formalization of long COVID as a recognized condition, and produces a sharp w_{fast} response with no accumulated w_{slow} to sustain it. *Delta – vaccine* appears in three consecutive Phase 4 articles on Delta, Omicron, and booster doses, showing ongoing consolidation, with w_{slow} still rising near the end of the stream because the events occurred too recently to have decayed.

Recency itself emerges from the dynamics rather than being prescribed. The *mRNA – vaccine* and *Delta – vaccine* pairs receive the same number of events with the same spacing, yet *Delta – vaccine* retains a higher final state because its events arrive later in the stream and have had less time to decay. This recency sensitivity follows directly from Equations (1) and (2).

A.4. Is the Slow Variable Necessary? An Ablation across All Pairs

Figure 2 illustrates that the dynamics behave qualitatively as expected on a small set of representative pairs, but it cannot establish that the slow variable is necessary. To test this, we compare Memini against two alternatives on the full set of 68 entity pairs in the stream. The first is a single-timescale ablation in which w_{slow} is removed and the remaining

variable evolves as $dw/dt = -w/\tau + I$. We choose τ to match Memini's effective decay rate when $w_{\text{slow}} = 0$, namely $\tau = 1/(1/\tau_{\text{fast}} + C) \approx 1.43$, so that the two systems are mathematically identical in the absence of consolidation. Any divergence between them must therefore originate from w_{slow}. The second is a uniform-retention baseline that simply accumulates event counts without any decay, representing a memory that never forgets.

We classify each of the 68 pairs into one of four groups using only two quantities. The first is the number of co-occurrence events n that the pair receives across the stream. The second is the document index at which the most recent of those events occurs, where documents are indexed 0 to 12, so a higher index means more recent. *Repeated, no longer mentioned* pairs ($n \geq 3$, last index ≤ 7) are associations that should require multi-timescale dynamics to be retained. *Few mentions, no longer mentioned* pairs ($n \in \{1, 2\}$, last index ≤ 7) are weakly supported associations that should be forgotten. *Repeated, recently mentioned* pairs ($n \geq 3$, last index ≥ 10) are strongly supported and current. *Few mentions, recently mentioned* pairs ($n \in \{1, 2\}$, last index ≥ 10) are weakly supported but current. Pairs whose last event falls at index 8 or 9 are excluded so that the old and recent regions remain cleanly separated. The four-way classification covers 51 of the 68 pairs and is determined entirely by event metadata.

Table 3 reports the mean final association weight in each group under each of the three configurations, and the four rows together produce the expected pattern. The headline result is in the first row. Repeated associations that are no longer mentioned are exactly the case where multi-timescale architecture should matter most, and Memini retains a mean weight roughly thirty times larger than the single-timescale ablation. This gap cannot be explained by faster decay, since the two systems share the same effective decay rate when $w_{\text{slow}} = 0$, and it isolates the contribution of the slow variable. The remaining three rows act as controls. Few-mention associations that are no longer mentioned collapse to near zero in both Memini and the ablation, while uniform retention still carries them at a mean of 1.28, showing the cost of having no decay at all. Repeated associations that remain recently mentioned look similar across Memini and the ablation, since recent input dominates and consolidation has had less time to make its effect visible. Few-mention associations that are recently mentioned are tied within numerical precision, confirming that Memini gains no spurious advantage where consolidation has not had time to occur.

Table 3. Mean final association weight across the 51 entity pairs classified by event pattern, with N denoting the number of pairs in each group. Memini denotes the full coupled dynamics. The single-timescale ablation uses τ matched to Memini's effective early-decay rate when $w_{\text{slow}} = 0$, so any divergence between Memini and the ablation is attributable to w_{slow}. Uniform denotes cumulative event count with no decay. The pattern labels follow the conventions defined in the main text, where "repeated" means three or more events, "few mentions" means one or two, "no longer mentioned" means the last event occurred at least five documents before the end of the stream, and "recently mentioned" means the last event occurred in the final three documents. Parameters are $\tau_{\text{fast}} = 2$, $\tau_{\text{slow}} = 10$, and $C = 0.2$.

Pattern	N	Memini w_{fast}	Single-timescale	Uniform
Repeated, no longer mentioned	5	0.104	0.003	3.000
Few mentions, no longer mentioned	25	0.026	0.000	1.280
Repeated, recently mentioned	8	0.426	0.260	4.625
Few mentions, recently mentioned	13	0.463	0.451	1.000

The pattern across all four rows supports a sharper claim than any single comparison. Memini retains old associations only when those associations were repeatedly confirmed, fades old associations that received sparse support, and treats recent associations consistently with whatever evidence has accumulated for them. Selective retention here is not produced by an explicit deletion rule, a threshold, or a gating module, but emerges from the coupled dynamics applied uniformly to every pair. This positions forgetting not as a failure mode to be avoided, but as a functional property of the substrate, in line with views of memory in which selective decay supports adaptation rather than degrading it (Nørby, 2015). In Memini, the same mechanism that allows associations to persist also allows them to fade, and the architecture itself decides which.

A.5. Scope

This experiment is intentionally limited in scope. The corpus contains 13 documents, the four classified groups range from 5 to 25 pairs, and we report no retrieval metrics or comparison against published systems. What carries the result is therefore not any single group but the consistent pattern across all four, with Memini diverging from the matched ablation only where expected. The broader empirical questions, including validation on larger and more varied document streams, retrieval quality on temporal question-answering corpora, and comparison against existing memory-augmented systems, remain the appropriate targets for rigorous evaluation and are left to future work, as outlined in Section 4.