Volume 11, Issue 1, January – 2026 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan155

Dynamic Memory Updating in RAG: Lifelong Learning and Adaptation

Sivarama Krishna Akhil Koduri[1]

[1]Independent Researcher / PhD Student, Information Technology (Artificial Intelligence), University of the Cumberlands, Florence, KY, USA

[1]ORCID: 0009-0009-9663-5740

Publication Date: 2026/01/15

Abstract: Retrieval-Augmented Generation (RAG) has established itself as the standard for reducing hallucinations in Large Language Models (LLMs) by grounding generation in external knowledge. However, conventional RAG implementations rely on static vector stores, limiting their utility in dynamic environments where information evolves rapidly. This reliance on fixed knowledge bases restricts adaptability and long-term scalability. This paper synthesizes recent literature on RAG system design, specifically focusing on mechanisms for continuous learning. Building on frameworks by Zheng et al. and Zhang et al., we analyze architectures that support continuous memory addition, deletion, consolidation, and re-weighting. These mechanisms transition RAG from static retrieval to incremental learning, mirroring biological memory processes. Our analysis demonstrates that dynamic memory architectures outperform static systems in adaptability, robustness to distribution shifts, and long-term retention. We conclude that dynamic memory updating is not merely an optimization but a fundamental architectural requirement for sustaining lifelong learning in RAG systems.

Keywords: *Retrieval-Augmented Generation, Dynamic Memory Updating, Lifelong Learning, Continual Learning, Large Language Models, Memory-Augmented Systems, Adaptive AI Agents.*

How to Cite: Sivarama Krishna Akhil Koduri (2026) Dynamic Memory Updating in RAG: Lifelong Learning and Adaptation. *International Journal of Innovative Science and Research Technology*, 11(1), 724-727. https://doi.org/10.38124/ijisrt/26jan155

I. INTRODUCTION

➢ *Background and Motivation*

Despite their fluency, Large Language Models (LLMs) remain prone to hallucinations and are constrained by the static nature of their pre-training data. To address this, Retrieval-Augmented Generation (RAG) integrates parametric memory with non-parametric external knowledge [22].

However, many deployed RAG systems persist in using static information retrieval pipelines. While these pipelines minimize the need to update model parameters, they fail to account for the evolving nature of user needs and real-world data. As noted by Mohammed [14] and Fan et al. [13], static systems suffer from knowledge staleness, domain drift, and the accumulation of redundant content. This degradation over time erodes the initial advantages of the RAG architecture.

The integration of LLMs into continuous applications—such as autonomous agents and personalized assistants—exacerbates these challenges. These systems must retrieve accurate information and adapt autonomously without frequent retraining. Current static architectures are ill-equipped for this reality, highlighting a critical need for rigorous design principles regarding memory management and updates.

➢ *The Relevance of RAG and Dynamic Memory*

Lifelong learning requires systems to acquire, store, and refine knowledge perpetually within non-stationary environments. For RAG systems to achieve this, memory components must adapt to incoming data. Jiang et al. [2] and Zheng et al. [8] argue that static memory fundamentally contradicts continual learning principles by assuming a closed-world knowledge distribution.

Dynamic memory in RAG redefines external knowledge repositories as active, mutable components rather than permanent archives. This involves mechanisms for adding, updating, and re-sequencing structures to reflect revised information. By leveraging temporal dynamics, systems can prioritize recent or frequently accessed data while pruning obsolete content.

Furthermore, dynamic memory is essential for self-regulation in agentic architectures. Liang et al. [9] emphasize that agents performing multi-step reasoning require memory that records and updates itself to reflect interaction history. Similarly, Hu et al. [12] posit that without flexible, adaptive

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan155

memory, agents are restricted to reactive behaviors and cannot achieve true autonomy.

➢ *Current Knowledge and Gaps*

Prior research has largely prioritized retrieval quality, focusing on embedding models, re-ranking, and query reformulation. While studies by Mao et al. [10] and Jeong et al. [16] demonstrate improvements in document merging and generation fusion, they often overlook the necessity of mutable memory systems.

Consequently, the evolution of memory structures remains under-explored. Engineering approaches often bypass critical questions regarding update frequency, retention policies, and the stability-plasticity dilemma. Zhang et al. [7] identify a lack of unifying principles for memory updates in RAG, noting that existing solutions are fragmented across disparate domains. Gruia and Ionescu [19] further argue that this fragmentation hinders meaningful comparison and long-term system reasoning.

➢ *Purpose and Objectives*

This research addresses these gaps by analyzing dynamic memory updating mechanisms in RAG systems. We aim to map the design space of adaptive architectures, moving beyond single-algorithm evaluations to a broader conceptual analysis. Specifically, we assess techniques for memory addition, updating, consolidation, and elimination.

Our objectives are twofold: first, to critique dynamic updating systems in recent literature, identifying their architectural strengths and assumptions; and second, to synthesize these findings into foundational design principles for scalable, enduring RAG systems, drawing on the roadmaps proposed by Zheng et al. [3] and Lei et al. [15].

II. METHODOLOGY

➢ *Research Design*

We employ a qualitative meta-analysis and architectural comparison to evaluate dynamic memory mechanisms. Rather than benchmarking a single model, we synthesize insights across the literature to identify emerging best practices and conceptual frameworks. This approach aligns with methodologies used in recent surveys on memory and RAG [25, 8], bridging the systemic gap between retrieval, generation, and continual learning.

➢ *Selection Criteria*

We filtered the literature based on scope, relevance, and recency, focusing primarily on works published between 2023 and 2025—the period marking significant advancements in RAG and agentic systems. We included peer-reviewed publications (e.g., EMNLP, ACL, SIGIR) and high-impact preprints.

Selected works address one of three core themes: (i) RAG system design, (ii) memory mechanisms in LLMs, or (iii) continual learning in neural systems. We excluded studies focused solely on static retrieval optimization, ensuring our analysis remains centered on memory evolution.

➢ *Analysis Framework*

To systematize our review, we utilized an analytical framework based on three dimensions [4, 20]:

- Memory Persistence: The duration of information retention and the presence of explicit forgetting mechanisms.
- Update Frequency: How often memory is modified (e.g., batch vs. real-time).
- Integration Depth: The degree to which memory influences the broader RAG pipeline.

III. LITERATURE REVIEW

➢ *From Static to Memory-Augmented RAG*

Early RAG architectures were designed to mitigate knowledge cutoffs using static vector stores [22]. While effective for fixed domains, these models oversimplify knowledge distribution. As Mohammed [14] observes, static stores inevitably accumulate noise and stale data, degrading retrieval precision.

Memory-augmented RAG represents a paradigm shift, treating external knowledge as an adaptive resource. Research in this vein prioritizes the lifecycle of knowledge—encoding, retention, and elimination—transforming retrieval from a passive lookup into an active cognitive process.

➢ *Foundations of Continual Learning*

Dynamic RAG draws heavily from continual learning principles. Jiang et al. [2] identify the stability-plasticity paradox and memory consolidation as central challenges. Wang et al.'s Lifespan Cognitive Systems framework [11] is particularly relevant, suggesting that intelligent systems must integrate perception and memory over extended timescales.

➢ *Dynamic Memory Architectures*

Technical implementations of dynamic memory vary. Gutiérrez et al. [5] propose adaptive stores that refresh content based on usage and relevance decay. Qin et al. [4] demonstrate that selective retention of frequently accessed items improves robustness during distribution shifts.

More advanced architectures introduce hierarchy. Memorage [18] distinguishes between local (task-specific) and global (long-term) memory, while Comorag [20] integrates this hierarchy with reasoning control. These designs balance short-term flexibility with long-term stability.

➢ *Agent-Centric Models*

In autonomous agents, memory supports planning and reflection. Liang et al. [9] introduce agents with reflective memory logs, while Hu et al. [12] argue that self-updating memory is a prerequisite for autonomy. In robotics, systems like RoboMemory [15] demonstrate how memory updates allow agents to adapt to physical environments without retraining [1].

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan155

➢ *Multimodal and Domain Extensions*

Recent work extends dynamic RAG to new domains. Multi-RAG [21] integrates video and image retrieval, addressing spatio-temporal challenges. In industrial settings, Choi and Jeong [24] and Shan [17] emphasize the necessity of real-time memory updates to maintain safety and operational efficiency.

➢ *Problem Statement*

While RAG is pivotal for grounding LLMs, operational implementations largely rely on static vector stores. This creates a "memory problem": systems cannot learn or adapt in fluid environments. Static memories enforce a closed knowledge distribution [7, 13], leading to performance degradation due to concept drift [14].

The core issue is the failure to integrate lifelong learning principles into RAG design. Without mechanisms for adaptive memory, selective forgetting, and incremental integration, systems cannot achieve true autonomy.

IV. RESULTS

➢ *Key Observations*

Our analysis confirms distinct advantages for memory-evolving architectures.

Improved Task Adaptation Dynamic systems demonstrate superior adaptation to new tasks and data distributions. Qin et al. [4] and Long et al. [1] show that unlike static systems, which degrade under drift, dynamic architectures maintain performance by updating retrieval relevance.

Mitigation of Catastrophic Forgetting Gutiérrez et al. [5] highlight the efficacy of "memory refreshing" and "selective forgetting." These strategies prevent information overload, whereas static systems suffer from "retrieval forgetting" due to accumulated noise.

➢ *Comparative Summary*

Table 1 Comparative Summary

Memory Type	Update Strategy	Adaptability	Forgetting Resistance
Static Vector Store	Offline embedding; periodic re-indexing	Low (degrades under domain shift)	Low (retrieval noise accumulates)
Dynamic Flat Memory	Online insertion and selective refresh	Medium (adapts to new data)	Medium (partial retention via re-weighting)
Hierarchical Memory	Multi-level (local + global) updates	High (supports task/domain shifts)	High (separation of short/long-term)
Cognitive / Agent Memory	Event-driven, reflective updates	Very High (context-aware)	High (consolidation via reflection)

➢ *Performance Metrics*

Quantitative evidence supports these architectural shifts. Qin et al. [4] report a 10–18% improvement in retention on continual learning benchmarks with dynamic updating. Conversely, Fan et al. [13] observe significant performance drops in static systems. In long-horizon evaluations, Gutiérrez et al. [5] note that retrieval error grows by 30% in systems lacking memory updates.

V. LIMITATIONS

Despite these findings, dynamic RAG faces practical constraints. Much of the evidence is derived from highly regulated environments [3], potentially limiting generalizability. Furthermore, dynamic updates introduce computational overhead. As Mohammed [14] warns, frequent pruning and re-indexing can increase latency. Finally, the field lacks unified standards for measuring memory behavior [6], resulting in fragmented evaluation metrics.

VI. DISCUSSION

➢ *Interpretation of Findings*

The data suggests that update frequency and quality outweigh raw memory size. Large, static memories often dilute relevance with noise. In contrast, frequent, selective updates improve the signal-to-noise ratio [4]. Wang et al. [20] further demonstrate that frequent updates deepen system integration, capturing transient context more effectively.

➢ *Comparison with Prior Studies*

Our results validate Wang et al.'s Lifespan Cognitive Systems Theory [11], which posits that adaptive systems must revise rather than merely accumulate memory. This contrasts with earlier works [16, 10] that optimized retrieval algorithms while assuming static memory structures.

➢ *Discrepancies*

Interestingly, some dynamic systems underperform static baselines in short-term tasks. Fan et al. [13] and Shan [17] attribute this to instability caused by overly aggressive update frequencies, which can induce transient accuracy drops.

➢ *Implications*

For autonomous agents, flexible memory is a non-negotiable requirement for autonomy [12]. In enterprise AI, static RAG poses operational risks by surfacing outdated organizational knowledge [23].

VII. CONCLUSION

This review underscores the necessity of dynamic memory updating in RAG frameworks. The static vector store model, while useful for establishing baselines, is insufficient for lifelong learning. To achieve resilience against distributional shift and knowledge obsolescence, RAG systems must adopt dynamic, hierarchical memory architectures. We recommend that future research focuses on

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan155

establishing standardized benchmarks and evaluation procedures to operationalize these concepts [19, 25].

ACKNOWLEDGMENT

This manuscript represents original research and has not been previously published. A version of this work is prepared for submission to the IJISRT Conference 2026. This work received no specific grant funding from any agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1]. Y. Long, K. Chen, L. Jin, and M. Shang, "DRAE: Dynamic Retrieval-Augmented Expert Networks for Lifelong Learning and Task Adaptation in Robotics," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2025, pp. 23098--23141.

[2]. X. Jiang et al., "Long term memory: The foundation of AI self-evolution," *arXiv preprint arXiv:2410.15665*, 2024.

[3]. J. Zheng et al., "Lifelong learning of large language model based agents: A roadmap," *arXiv preprint arXiv:2501.07278*, 2025.

[4]. Q. Qin et al., "Towards adaptive memory-based optimization for enhanced retrieval-augmented generation," in *Findings of the Assoc. for Comput. Linguistics: ACL 2025*, Jul. 2025, pp. 7991--8004.

[5]. B. J. Gutiérrez et al., "From RAG to memory: Non-parametric continual learning for large language models," *arXiv preprint arXiv:2502.14802*, 2025.

[6]. D. Zhang et al., "Memory in Large Language Models: Mechanisms, Evaluation and Evolution," *arXiv preprint arXiv:2509.18868*, 2025.

[7]. D. Zhang et al., "Conversational Agents: From RAG to LTM," in *Proc. 2025 Annu. Int. ACM SIGIR Conf. on R&D in Inf. Retr. in the Asia Pacific Region*, Dec. 2025, pp. 447--452.

[8]. J. Zheng, S. Qiu, C. Shi, and Q. Ma, "Towards lifelong learning of large language models: A survey," *ACM Comput. Surveys*, vol. 57, no. 8, pp. 1--35, 2025.

[9]. X. Liang et al., "SAGE: Self-evolving Agents with Reflective and Memory-augmented Abilities," *Neurocomputing*, p. 130470, 2025.

[10]. K. Mao et al., "RAG-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment," in *Findings of the Assoc. for Comput. Linguistics: EMNLP 2024*, Nov. 2024, pp. 725--735.

[11]. Y. Wang et al., "Towards lifespan cognitive systems," *arXiv preprint arXiv:2409.13265*, 2024.

[12]. Y. Hu et al., "Memory in the Age of AI Agents," *arXiv preprint arXiv:2512.13564*, 2025.

[13]. Y. Fan et al., "Research on the online update method for retrieval-augmented generation (RAG) model with incremental learning," *arXiv preprint arXiv:2501.07063*, 2025.

[14]. A. S. Mohammed, "Dynamic Data: Achieving Timely Updates in Vector Stores," *Libertatem Media Private Limited*, 2024.

[15]. M. Lei et al., "RoboMemory: A Brain-inspired Multi-memory Agentic Framework for Lifelong Learning in Physical Embodied Systems," in *NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI*, 2025.

[16]. S. Jeong et al., "Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity," *arXiv preprint arXiv:2403.14403*, 2024.

[17]. R. Shan, "LearnRAG: Implementing Retrieval-Augmented Generation for Adaptive Learning Systems," in *2025 Int. Conf. on AI in Inf. and Commun. (ICAIIC)*, Feb. 2025, pp. 0224--0229.

[18]. H. Qian et al., "Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation," in *Proc. ACM Web Conf. 2025*, Apr. 2025, pp. 2366--2377.

[19]. L. Gruia and B. Ionescu, "Continual Learning for Generative AI Systems: Retrieval-Augmentation, Graph Reasoning, and Multimodal Integration," 2025.

[20]. J. Wang et al., "Comorag: A cognitive-inspired memory-organized RAG for stateful long narrative reasoning," *arXiv preprint arXiv:2508.10419*, 2025.

[21]. M. Mao et al., "Multi-RAG: A Multimodal Retrieval-Augmented Generation System for Adaptive Video Understanding," *arXiv preprint arXiv:2505.23990*, 2025.

[22]. S. Siriwardhana et al., "Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 1--17, 2023.

[23]. M. Walker, "SCMS Whitepaper-Complete Document Sparse Contextual Memory Scaffolding," *SCMS Whitepaper*, Oct. 2025.

[24]. H. Choi and J. Jeong, "A Conceptual Framework for a Latest Information-Maintaining Method Using Retrieval-Augmented Generation and a Large Language Model in Smart Manufacturing," *Machines*, vol. 13, no. 2, p. 94, 2025.

[25]. Z. Ke, Y. Ming, and S. Joty, "Adaptation of Large Language Models," in *Proc. 2025 Annu. Conf. NAACL: HLT*, May 2025, pp. 30--37.