Value-Driven Memory-Augmented Generation for Agentic LLMs: Towards Structured and Adaptive Knowledge Utilization

Cassandra Hui-Ming Tan

Singapore Management University
hm.tan.2023@phdcs.smu.edu.sg

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in reasoning, yet their efficacy is constrained by a fundamental memory limitation: a static context window that resets with each interaction. This prevents them from accumulating experience and adapting to dynamic, long-term tasks. To address the limitations of long-term memory in agentic LLMs, this work introduces a neuro-inspired framework with two key contributions. First, we propose **ARTEM** (Agentic Retrieval with Temporal-Episodic Memory), a system that organizes experiences into structured events and manages utility-based memory consolidation. Second, we extend this framework with a distinct governance component, **Value-driven ARTEM**, that validates candidate outputs against core principles before finalization. Together, these components equip LLM agents with continual learning, adaptive reasoning, and robust value-aligned decision-making. Looking forward, we outline future directions including dynamic memory adaptation, memory decay mechanisms, and applications in interactive multi-agent environments.

Introduction

Large Language Models (LLMs) are increasingly deployed as autonomous agents, yet their capacity for memory remains limited. Current models operate with a fixed-size context window that resets with each session (OpenAI 2023), preventing them from developing persistent histories, learning from past interactions, or adapting behavior across time. True agency requires more than one-shot reasoning: it requires *episodic memory*, the ability to encode events, retrieve them by multiple cues, and update behavior in light of prior outcomes.

This work draws inspiration from two key sources. First, from cognitive science, Tulving (2002) describes human episodic memory as the capacity to recall specific events situated in time and space. Second, from computational neuroscience, Adaptive Resonance Theory (ART) (Carpenter and Grossberg 1987) demonstrates how vigilance-controlled retrieval balances generalization and specificity, a principle well-suited to dynamic memory management.

Building on these foundations, we propose **ARTEM** (Agentic Retrieval with Temporal-Episodic Memory), a framework that organizes agent experiences into structured events and manages their consolidation by utility. We further introduce **Value-driven ARTEM**, extending the framework with a governance component that validates all candidate outputs against explicit principles before execution. This ensures that not only memory but also action is value-aligned.

The long-term vision of this work is to enable adaptive, trustworthy LLM agents. In addition to the current framework, we plan to explore dynamic memory, memory decay, and interactive multi-agent environment as future directions.

Research Questions

This work is guided by the following research questions:

1. What are effective ways to structure long-term memory in LLM agents to support temporal reasoning, multi-cue recall, and continual learning?
2. How can we balance precision and generalization in retrieval to enable adaptive, context-sensitive memory use?
3. How can memory-augmented LLM agents be governed to remain trustworthy, transparent, and aligned with human values as their memories evolve over time?

Related Work

Research on augmenting LLMs with knowledge and memory falls into two major paradigms.

Retrieval-Augmented Generation (RAG). RAG grounds outputs in large, static knowledge sources at inference time. RAG (Lewis et al. 2020) use dense retrieval to provide factual evidence to the model. While effective for grounding, these approaches rely on external corpora and do not accumulate an agent's lived history.

Memory-Augmented Generation. This line of work builds dynamic memory from interaction history. Early systems such as *MemGPT* (Packer et al. 2023) introduced hierarchical structures for long contexts, while recent efforts like *LongMem* (Wu et al. 2024), *M+* (Wang et al. 2025), and *MemOS* (Li et al. 2025) focus on scalability and system-level coordination. Despite progress, most methods still store unstructured text and rely on semantic similarity for retrieval, limiting temporal reasoning and multi-cue recall.

Adaptive Resonance Theory (ART). From the cognitive modeling community, ART (Carpenter and Grossberg 1987)

provide a biologically inspired model of learning and retrieval. ART introduces vigilance-controlled resonance, balancing specificity and generalization. These principles have influenced incremental learning architectures and motivate the resonance-based retrieval mechanism explored here.

Positioning ARTEM in the Memory Landscape

ARTEM extends memory-augmented generation by addressing the key limitations of structure, and governance.

While RAG and prior memory-augmented approaches store **unstructured text chunks**, ARTEM encodes experiences as **structured episodic events** with time, space, entity, and content dimensions. This supports richer, cue-based recall (e.g., "What happened yesterday in the lab?").

Besides, ARTEM incorporates a **value-driven governance component** that validates generated responses against explicit principles before execution or storage. This layer is absent in prior work, ensuring that both actions and memories are aligned with core values.

ARTEM retrieves structured event memories through vigilance-modulated resonance. These retrieved events guide the generation process. A governance component then evaluates candidate outputs against core values, defined by the ground principle and operational checklists. If violations are detected, a corrective regeneration loop is triggered. Only responses that pass validation are executed. If violations remain unresolved, the system suppresses the output and instead returns an explicit error message detailing the value violations.

Progress and Timeline

Progress as of September 30, 2025

- I extended the Spatial-Temporal Episodic Memory (STEM) model (Chang and Tan 2017) to support textual input
- I designed and implemented the ARTEM pipeline, consisting of event extraction from long text, event encoding via STEM, and LLM response generation with agentic retrieval; this work has been submitted to the AAAI main track, currently under Phase 2 review
- I evaluated ARTEM on the Episodic Memory Benchmark, a dataset of books and QA pairs
- I developed a governance component to guide LLM responses, ensuring alignment with human values

Anticipated Progress by the Workshop Date

- October 2025: Finalize integration of the governance component into ARTEM
- November 2025: Implement ARTEM within an interactive multi-agent environment
- December 2025: Adapt and extend the governance component for multi-agent interaction

Anticipated Contributions

This research advances memory-augmented LLMs from static buffers toward adaptive, value-governed systems. Anticipated contributions include: (1) a structured episodic memory framework for LLM agents, (2) a principled governance component ensuring value alignment, (3) empirical evaluation on episodic QA and interactive agent tasks, and (4) design insights for interpretable, trustworthy agentic AI.

Conclusion

This thesis proposed ARTEM, a framework that advances LLM memory from static buffers to structured, event-based representations with adaptive retrieval. It further introduced a governance component that enforces transparency and value alignment as memories evolve. Together, these contributions lay the foundation for LLM agents that can learn continually, adapt behavior over time, and act as trustworthy collaborators in dynamic environments.

References

Carpenter, G. A.; and Grossberg, S. 1987. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics, and Image Processing*, 37(1): 54–115.

Chang, P.-H.; and Tan, A.-H. 2017. Encoding and recall of spatio-temporal episodic memory in real time. In *Proceedings of the Twenty-sixth International Joint Conference on Artificial Intelligence*, 1490–1496. Melbourne, Australia. ISBN 978-0-9992411-0-3.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.

Li, Z.; Song, S.; Wang, H.; Niu, S.; Chen, D.; Yang, J.; Xi, C.; Lai, H.; Zhao, J.; Wang, Y.; Ren, J.; Lin, Z.; Huo, J.; Chen, T.; Chen, K.; Li, K.; Yin, Z.; Yu, Q.; Tang, B.; Yang, H.; Xu, Z.-Q. J.; and Xiong, F. 2025. MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models. arXiv:2505.22101.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Packer, C.; Fang, V.; Lin, S. G.; Zhu, S.; Gonzalez, J. E.; Stoica, I.; and Jordan, M. I. 2023. MemGPT: Towards LLMs as Operating Systems. In *Advances in Neural Information Processing Systems*, volume 36.

Tulving, E. 2002. Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, 53(1): 1–25.

Wang, Y.; Krotov, D.; Hu, Y.; Gao, Y.; Zhou, W.; McAuley, J.; Gutfreund, D.; Feris, R.; and He, Z. 2025. M+: Extending MemoryLLM with Scalable Long-Term Memory. In *Forty-second International Conference on Machine Learning*.

Wu, L.; Xiong, W.; Yih, W.-t.; and Lewis, M. 2024. LongMem: Scaling LLM Memory with Retrieval. *arXiv preprint arXiv:2402.11410*.