ARTEM: Enhancing Large Language Model Agents with Spatial-Temporal Episodic Memory

Cassandra Hui-Ming Tan, Budhitama Subagdja, Ah-Hwee Tan

School of Computing and Information Systems, Singapore Management University, Singapore
hm.tan.2023@phdcs.smu.edu.sg, {budhitamas,ahtan}@smu.edu.sg

Abstract

Current large language models (LLMs) exhibit significant deficiencies in episodic memory tasks including encoding, storing, and retrieving specific information from temporally dependent events over a long period of time. Recent approaches to handle memory tasks in LLMs, such as in-context learning, retrieval-augmented generation (RAG), and fine-tuning, may resolve the long-term retention issues, but are still inadequate to handle tasks requiring chronological awareness of the stored information. We introduce Agentic Retrieval with Temporal-Episodic Memory (ARTEM), a hybrid LLM-based agent architecture integrating LLMs with a self-organizing neural network named Spatial-Temporal Episodic Memory (STEM), designed to handle episodic memory tasks. Our approach employs LLMs for event extraction from the inputs to represent temporal, spatial, entitative, and semantic information that may facilitate future retrieval, aside from generating outputs or direct responses. The extracted events can then be encoded vectorially and stored in a fast and stable manner in the episodic memory through an instance-based incremental learning in STEM. STEM supports precise episodes retrieval and helps reduce computational overhead in generating the appropriate responses by LLMs. Evaluation on standardized episodic memory benchmarks across four tasks—partial cue retrieval, epistemic uncertainty detection, recent event identification, and chronological recall—demonstrates superior performance of ARTEM compared to in-context learning, RAG, and fine-tuning in various popular LLMs.

Code — https://github.com/cassthm/ARTEM

Introduction

Large language models (LLMs) have demonstrated impressive capabilities in language understanding and generation (Brown et al. 2020; OpenAI 2023). However, their memory is largely static and context-limited (Liu et al. 2024), raising concerns about their ability to retain and recall episodic knowledge over long durations. Episodic memory—a structured representation of personal past experiences (Tulving 2002)—is essential for coherent reasoning, especially in tasks requiring temporal and contextual grounding.

Traditional transformer-based architectures face fundamental limitations in episodic memory tasks. The attention mechanism, while powerful for capturing contextual relationships, struggles with precise temporal sequencing and often produces hallucinated content when presented with partial cues that do not correspond to stored experiences. These limitations become particularly pronounced in scenarios requiring chronological reasoning, where models must maintain coherent temporal relationships across extended event sequences.

Recent advances in memory-augmented neural architectures have attempted to address these limitations through various approaches, including external memory banks (Packer et al. 2023; Zhong et al. 2024), retrieval-augmented generation (Gao et al. 2023; Ram et al. 2023), and specialized attention mechanisms (Bulatov, Kuratov, and Burtsev 2022). However, these solutions often suffer from computational overhead, lack interpretability (Zhao et al. 2024), or fail to capture the nuanced temporal dynamics essential for robust episodic memory (Wang et al. 2023).

In view of the existing limitations, in this work, we introduce a biologically-inspired memory retrieval pipeline built on fusion ART (Tan et al. 2019), a generalized model of Adaptive Resonance Theory (ART) networks (Grossberg 2013), to support structured storage and cue-driven recall of event sequences. We propose an architecture termed Agentic Retrieval with Temporal-Episodic Memory (ARTEM), designed to enhance LLMs' capacity for retrieving events across extended temporal spans and long-range dependencies while improving chronological awareness of the LLMs for more effective temporal inference.

The ARTEM framework integrates a specialized variant of fusion ART—the Spatial-Temporal Episodic Memory (STEM) model—with LLM in a hybrid architecture (Chang and Tan 2017). Within this framework, LLMs serve dual roles as event extraction and question-answering agents, while STEM functions as an efficient memory and event retrieval agent that provides extracted episodic information to LLMs for memory-augmented prompting.

ARTEM's primary contribution lies in adapting STEM for textual data processing through a structured embedding methodology that preserves semantic relationships while enabling ART-based similarity computations. The architecture employs STEM to incorporate four distinct encoding channels—temporal, spatial, entity-based, and content-based—to represent events from diverse knowledge sources

including books and dialogues. This multi-channel approach enables the encoding of episodic memories as discrete nodes within the STEM network, facilitating targeted event retrieval through a configurable mechanism. By leveraging STEM's tunable vigilance parameter, ARTEM's memory-augmented prompting mechanism optimizes the trade-off between confabulation mitigation and precision-recall performance in LLMs through controlled retrieval selectivity.

For evaluation, we refer to the Episodic Memory Benchmark (Huet, Houidi, and Rossi 2025), which provides a principled framework for evaluating episodic memory capabilities. We utilize the synthetic dataset from this benchmark to conduct a comprehensive evaluation of our proposed models. The evaluation is structured into two complementary components: first, assessing the performance of STEM on event retrieval tasks, and second, evaluating the integrated ARTEM system incorporating large language model (LLM) components.

This bifurcated evaluation approach addresses the inherent trade-offs within the system architecture. While LLM integration introduces potential error propagation through event extraction, question answering, and 'LLM-as-a-Judge' evaluation processes, it illustrates the LLM's semantic understanding capabilities to identify and select the most relevant responses from STEM's retrieved event candidates. This dual assessment methodology enables a nuanced understanding of both the core retrieval mechanism's performance and the overall system's effectiveness when augmented with language model capabilities.

The novelty and contributions of our work are summarized as follows. First, we introduce ARTEM that combines the language understanding capabilities of LLMs with the structured memory representation of Spatial-Temporal Episodic Memory (STEM) model. Second, we demonstrate that biologically-inspired memory models with explicit memory encoding capabilities can enhance LLM performance on episodic tasks, especially on the chronological awareness capabilities, while maintaining computational efficiency. Third, we provide a comprehensive evaluation across multiple episodic memory dimensions, establishing a new benchmark for temporal reasoning in neural systems.

Related Work

Memory-augmented models span several research directions, including Key-Value Memory Networks (Miller et al. 2016), Differentiable Neural Computers (Graves et al. 2016), and retrieval-augmented generation (RAG) (Lewis et al. 2020). While Key-Value Memory Networks provide learned associations, they struggle with temporal dependencies and require extensive supervision. Differentiable Neural Computers offer differentiable memory operations but suffer from poor computational scaling. RAG approaches (Liu et al. 2024; Rempe et al. 2023) typically employ dense vector similarity but lose fine-grained temporal information and provide limited interpretability.

In-context learning (ICL) (Brown et al. 2020; Dong et al. 2024) and fine-tuning represent the predominant approaches for integrating new knowledge or memory into LLMs. However, these methods face significant limitations: ICL is constrained by context window length and lacks persistent memory across sessions, while fine-tuning requires substantial computational resources and risks catastrophic forgetting of previously learned knowledge (Kirkpatrick et al. 2017).

Adaptive Resonance Theory (ART) (Grossberg 2013) is a class of neural network architectures that enables stable learning of arbitrary input sequences through competitive learning and vigilance-controlled pattern matching mechanisms. Fusion ART (Tan, Carpenter, and Grossberg 2007; Tan et al. 2019) extends this framework by integrating multiple information channels through complementary learning, allowing simultaneous processing of heterogeneous data types within a unified resonance-based system. Building upon fusion ART, the Spatial-Temporal Episodic Memory (STEM) model (Chang and Tan 2017) specifically adapts these multi-channel capabilities for episodic memory representation, incorporating temporal, spatial, entity-based, and content-based encoding channels to capture the multifaceted nature of event sequences.

The Episodic Memory Benchmark (Huet, Houidi, and Rossi 2025) establishes standardized evaluation for episodic memory capabilities, revealing significant limitations in current LLMs regarding temporal coherence and confabulation avoidance. This work follows the episodic question answering task structure from this benchmark.

Methodology

We propose a hybrid agent architecture called ARTEM (Agentic Retrieval with Temporal-Episodic Memory), integrating structured episodic memory with temporal reasoning capabilities. ARTEM consists of an episodic memory model (STEM) and a large language model (LLM) serving as a dual role agent that operates through four sequential processes: (1) LLM-based event extraction from input sources, (2) multi-channel episodic memory encoding within STEM, (3) vigilance-guided memory retrieval using configurable similarity thresholds, and (4) memory-augmented response generation by the LLM.

By leveraging STEM's structured memory organization alongside LLMs' natural language processing capabilities, our approach addresses the critical limitations of existing memory-augmented systems in maintaining long-span memory recall and temporal reasoning coherence.

Problem Formulation

The episodic memory task is formulated as a spatial-temporal pattern learning, where each event e is represented as:

$$e = (t, s, ent, c)$$

where $t \in [0, 1]$ is normalized time, and $s, ent, c \in \mathbb{R}^{384}$ are SentenceTransformer embeddings for space, entities, and content respectively.

Given an event sequence $E = \{e_1, e_2, \ldots, e_n\}$, the system must: (1) encode events preserving spatio-temporal relationships, and (2) retrieve events using partial cues $q = (t_q, s_q, ent_q, c_q)$ with potentially missing attributes.

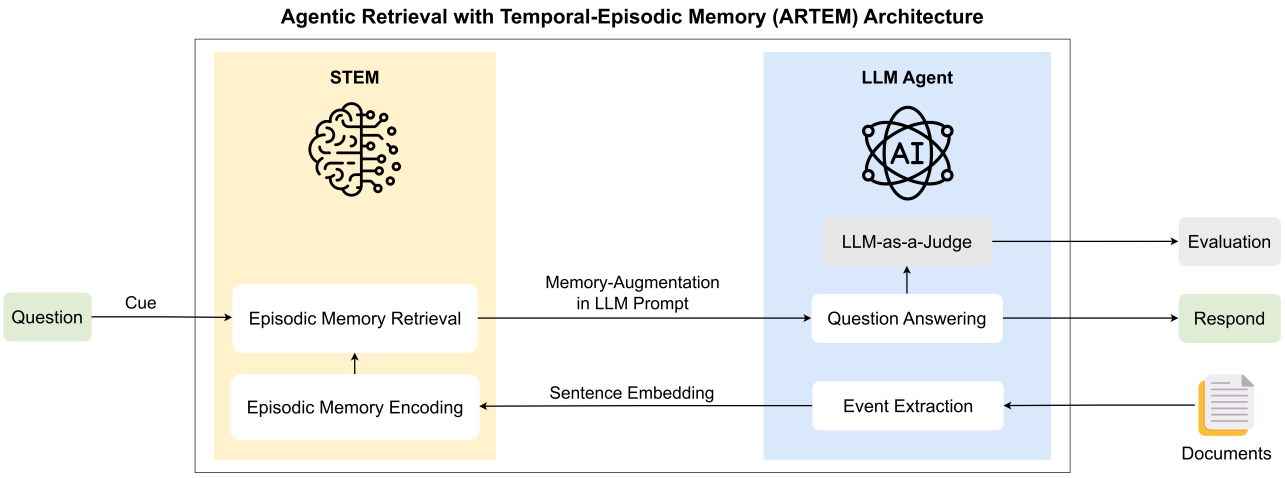

Figure 1: ARTEM Architecture showing event extraction from documents via LLM agent, multi-channel memory encoding through STEM network, vigilance-guided retrieval for episodic memory access, and memory-augmented question answering. LLM-as-a-Judge evaluation is included for experimental validation but is not a core component of the ARTEM framework.

The task requires four critical capabilities: **partial cue retrieval**, **epistemic uncertainty** recognition when no relevant memories exist, **recent event identification** for the identification of the most recent event, and **chronological recall** for chronological ordering, from incomplete information.

STEM Memory Architecture

STEM is a specific instantiation of fusion ART with four specialized channels for encoding episodic information. The four channel are Channel 0 (Time), Channel 1 (Spatial), Channel 2 (Entity), and Channel 3 (Content).

Channel 0 processes temporal information through normalized timestamps that preserve chronological relationships between events. **Channel 1** handles spatial information using 384-dimensional embeddings to maintain spatial coherence across different locations. **Channel 2** manages entity-based representations by encoding named entities through transformer-based embeddings. Finally, **Channel 3** processes semantic content using dense vector representations to capture the contextual meaning of events. Detailed configuration setting is provided in the Appendix.

Each channel maintains independent vigilance parameters enabling fine-grained retrieval control. Events are represented as:

$$e_i = \langle t_i, s_i, ent_i, c_i, \rho_i \rangle \tag{1}$$

where ρ_i includes activation scores and match confidence for interpretability.

Event Encoding and Normalization in STEM STEM requires normalized event features for memory encoding. The normalization strategy addresses the heterogeneous nature of episodic information while ensuring compatibility with STEM's matching functions.

Temporal Normalization: Channel 0 captures temporal information using global min-max scaling across the entire event corpus, preserving relative temporal relationships while constraining values to $[0, 1]$. Raw timestamps are converted as follows:

$$\tau_{norm} = \frac{\tau - \tau_{min}}{\tau_{max} - \tau_{min}} \tag{2}$$

where τ_{min} and τ_{max} represent corpus-wide temporal bounds.

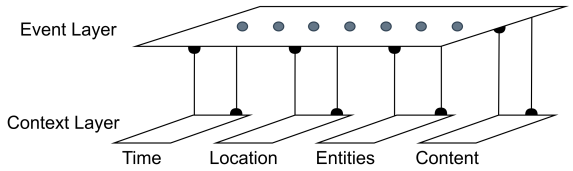

Figure 2: The STEM network model enabling parallel encoding of temporal, spatial, entity, and content information across multiple channels with independent vigilance control.

Semantic Normalization: Channels 1-3 encode spatial, entity, and content information as 384-dimensional embeddings generated using SentenceTransformer models. Unlike temporal normalization, semantic channels employ per-vector min-max normalization to address the variable dynamic ranges of different semantic embeddings. For each embedding vector $\mathbf{v}$, normalization follows:

$$\mathbf{v}_{norm} = \frac{\mathbf{v} - \mathbf{v}_{min}}{\mathbf{v}_{max} - \mathbf{v}_{min}} \tag{3}$$

where $\mathbf{v}_{max}$ and $\mathbf{v}_{min}$ are the maximum and minimum values across all dimensions within the specific embedding vector $\mathbf{v}$.

Per-vector min-max normalization ensures each dimension falls within $[0, 1]$ while preserving semantic relationships within individual embeddings.

This hybrid approach of global scaling for temporal data and per-vector scaling for semantic embeddings balances temporal consistency with semantic preservation, enabling effective cross-channel matching during retrieval while maintaining episodic integrity for cue-based memory access.

Events are stored in the STEM network through adaptive resonance learning, where vigilance parameters control memory granularity. Given these normalized representations, the event field learns recognition nodes in response to the presented temporal, spatial, entity, and content information. The detailed encoding algorithm is described in (Chang and Tan 2017).

Temporal Access Strategies ARTEM supports three distinct temporal access strategies for episodic memory retrieval, as described below.

Complete Access (`all`) returns all events associated with a query, enabling comprehensive temporal reasoning:

$$\mathcal{R}_{\text{all}}(q_i) = \{e_j \mid e_j \in r_i\} \tag{4}$$

This strategy provides complete information on all relevant events, supporting complex reasoning tasks that require understanding of event sequences and relationships.

Latest Access (`latest`) retrieves only the most recent event, simulating recency-biased memory access:

$$\mathcal{R}_{\text{latest}}(q_i) = \{\arg\max_{e_j \in r_i} \tau(e_j)\} \tag{5}$$

where $\tau(e_j)$ represents the temporal ordering of the event e_j.

This strategy mimics human memory tendencies to prioritize recent events and is particularly useful for questions about current states or the latest occurrences.

Chronological Access (`chronological`) returns all events with explicit temporal ordering emphasis:

$$\mathcal{R}_{\text{chronological}}(q_i) = \text{sort}(\{e_j \mid e_j \in r_i\}, \tau) \tag{6}$$

This strategy emphasizes temporal sequencing and is essential for tasks requiring understanding of event progression and causal relationships.

Vigilance-Guided Retrieval STEM employs vigilance-guided matching to control precision and prevent hallucination. A match score is computed for each channel k as:

$$m_j^k = \frac{|I^k \wedge w_j^k|}{|I^k|} \tag{7}$$

where I^k is the input pattern, w_j^k is the retrieved stored pattern for channel k, and the $\wedge$ symbol represents the element-wise minimum (fuzzy intersection) operator.

Events are retrieved only if $m_j^k \geq \rho^k$ for all active channels, ensuring quality control and reducing false positives.

LLM Integration

Large language models (LLMs) are employed in ARTEM for three primary functions: event extraction, answer generation, and automated evaluation of generated responses. The implementation details of these functions are described in this section.

Event Extraction Event extraction in ARTEM employs an LLM-based agent that processes textual sources to identify and structure episodic information. Given a text segment T_i from chapter i formatted according to the prompt template π_{extract}, the extraction process is formalized as:

$$e_i = \mathcal{E}(T_i, \pi_{\text{extract}}) \tag{8}$$

where $\mathcal{E}$ represents the LLM extraction agent and π_{extract} denotes the structured extraction prompt template.

The agent extracts four key components through structured prompting: temporal markers, spatial information, entity references, and semantic content summaries, formatted as standardized JSON output for seamless STEM integration. The extraction employs deterministic sampling (temperature = 0.1) with robust post-processing mechanisms to handle parsing errors and ensure consistent output formatting.

Answer Generation Given a query q_i and a set of retrieved events $\mathcal{R}(q_i)$ from STEM, formatted according to the prompt template π_{context}, the system generates responses as:

$$a_i = \mathcal{A}(q_i, \mathcal{R}(q_i), \pi_{\text{context}}) \tag{9}$$

using a structured prompt emphasizing factual accuracy and temporal coherence.

Automated Evaluation For evaluation, we adopt the LLM-as-a-Judge method as formulated in (Huet, Houidi, and Rossi 2025). This method identifies the final answers from the LLM-generated responses and computes precision and recall as follows:

$$\text{Precision} = \frac{\sum_{j=1}^{|GT|} s_j}{\min(|P|, |GT|)} \tag{10}$$

$$\text{Recall} = \frac{\sum_{j=1}^{|GT|} s_j}{|GT|} \tag{11}$$

where $s_j \in [0, 1]$ are the matching scores for each ground truth item, $|P|$ is the number of predictions, and $|GT|$ is the ground truth count. F1 scores are computed as the harmonic mean of precision and recall.

Experimental Setup

This section describes the benchmark datasets, baseline comparisons, and implementation details for comprehensive evaluation of ARTEM's episodic memory capabilities.

Dataset and Benchmarks

We evaluate ARTEM following the Episodic Memory Benchmark (Huet, Houidi, and Rossi 2025) which provides standardized tasks with narrative sequences and temporally ordered events. The evaluation encompasses:

- **Partial Cue Retrieval**: Retrieve events from incomplete information
- **Epistemic Uncertainty Detection**: Recognize absent memories, avoiding confabulation
- **Recent Event Identification**: Identify most recent matching events
- **Chronological Recall**: Order events temporally

Implementation Configuration

ARTEM is implemented in PyTorch with customized STEM modules and SentenceTransformers (all-MiniLM-L6-v2) for 384-dimensional embeddings. The system operates on four NVIDIA L40S GPUs with 48GB memory for efficient large-scale processing.

ARTEM uses DeepSeek-R1-Distill-Qwen-14B with deterministic generation ($T = 0.1$). STEM employs operation-specific parameters: uniform vigilance ($\rho = 1.0$) during encoding, and relaxed thresholds during retrieval ([1.0, 0.99, 0.99, 0.98] for the time, space, entity, content channels, respectively). Channel weights (γ) activate selectively based on the available query information. Complete specifications are provided in the Appendix.

Experimental Results

This section presents comprehensive performance evaluations across two primary dimensions: (1) comparative analysis of memory architectures and model variants, and (2) benchmarking against state-of-the-art language models on episodic recall and temporal reasoning tasks. All reported statistics are evaluated on the synthesized long book named 'Synaptic Echoes', with 200 chapters consisting of $102,870$ tokens.

STEM and ARTEM Performance Analysis

Tables 1 and 2 present comprehensive performance breakdowns for both models across ground-truth answer bins, temporal instruction types, and question cue categories.

Category	Count	Mean F1	Std F1
Performance by Ground Truth Answers			
Bin 0	150	1.000	0.000
Bin 1	150	0.550	0.490
Bin 2	82	0.690	0.370
Bin 3-5	128	0.590	0.330
Bin 6+	90	0.660	0.280
Performance by Time Instruction			
All	487	0.730	0.360
Latest	54	0.540	0.500
Chronological	59	0.640	0.350
Performance by Question Cue			
Entities	119	0.770	0.340
Event contents	152	0.730	0.380
Full event details	10	0.500	0.500
Other entities	10	0.500	0.500
Spaces	155	0.730	0.380
Times	154	0.640	0.380
Overall Performance Metrics			
Total Dataset	600	0.707	0.379
Mean Precision		0.714	
Mean Recall		0.705	

Table 1: Detailed performance results of the STEM model in the event retrieval task on 'Synaptic Echoes'.

Both models demonstrate exceptional confabulation detection capabilities, with STEM achieving perfect F1 scores

Category	Count	Mean F1	Std F1
Performance by Ground Truth Answers			
Bin 0	150	0.950	0.130
Bin 1	150	0.680	0.440
Bin 2	90	0.560	0.410
Bin 3-5	98	0.540	0.410
Bin 6+	60	0.510	0.380
Performance by Time Instruction			
All	437	0.710	0.390
Latest	55	0.660	0.390
Chronological	56	0.580	0.440
Performance by Question Cue			
Entities	97	0.690	0.400
Event contents	139	0.680	0.400
Full event details	9	0.670	0.500
Other entities	9	0.890	0.330
Spaces	139	0.660	0.430
Times	155	0.730	0.360
Overall Performance Metrics			
Total Dataset	548	0.691	0.399
Mean Precision		0.778	
Mean Recall		0.622	

Table 2: Detailed performance results of the ARTEM model in the event retrieval task on 'Synaptic Echoes'.

(1.00) and ARTEM near-perfect scores (0.95) for queries with zero ground truth answers (Bin 0). This represents a critical advancement in controlling hallucination in memory retrieval systems.

ARTEM's hybrid architecture addresses STEM's vigilance parameter constraints by leveraging LLM reasoning capabilities for final answer synthesis. Unlike STEM's rigid vigilance thresholds, ARTEM retrieves broader sets of potentially relevant events with relaxed vigilance settings, then utilizes contextual understanding to filter and synthesize appropriate responses. This enables ARTEM to achieve enhanced single-event precision ($F1 = 0.68$ vs 0.55) while maintaining competitive overall performance.

However, ARTEM evaluated fewer queries than STEM (548 vs. 600) due to LLM processing and token-length constraints. Specifically, 50 queries were discarded for exceeding the 512-token limit, which was deliberately enforced to balance computational efficiency and response quality. As reflected in Tables 1 and 2, these discarded cases predominantly fall within the higher complexity bins (3–5 and above), corresponding to longer event spans. An additional two queries failed during event extraction due to malformed outputs. Further details on these cases are provided in the Appendix. This controlled truncation strategy highlights the trade-off between ARTEM's enhanced flexibility in handling unstructured input and its reliance on the underlying LLM's token and output stability.

Cross-Architecture Performance Comparison

Table 3 reveals distinct performance patterns across memory paradigms, demonstrating the superiority of our episodic memory architectures.

Memory	Model	Number of events matching the cues				
		0 (150)	1 (150)	2 (90)	3-5 (98)	6+ (60)
In-context	gpt-4o-mini	0.51±0.50	0.54±0.46	0.44±0.36	0.47±0.27	0.50±0.17
	gpt-4o	0.84±0.37	0.81±0.38	0.60±0.31	0.57±0.21	0.53±0.14
	claude-3-haiku	0.84±0.37	0.39±0.48	0.37±0.30	0.37±0.28	0.38±0.19
	claude-3-5-sonnet	0.92±0.27	0.35±0.48	0.35±0.33	0.32±0.25	0.41±0.20
	o1-mini	0.97±0.16	0.05±0.19	0.12±0.24	0.12±0.19	0.24±0.19
	llama-3.1-405b	0.80±0.40	0.49±0.47	0.38±0.33	0.40±0.25	0.45±0.20
RAG	gpt-4o-mini	0.63±0.49	0.60±0.46	0.60±0.34	**0.59±0.26**	0.62±0.22
	gpt-4o	0.82±0.59	0.60±0.46	0.55±0.33	0.55±0.28	0.59±0.21
	claude-3-haiku	0.71±0.45	0.57±0.47	0.59±0.33	0.58±0.26	0.59±0.25
	claude-3-5-sonnet	0.91±0.28	0.59±0.47	0.59±0.35	**0.59±0.27**	0.62±0.25
Fine-tuning	gpt-4o-mini	0.00±0.00	**0.83±0.35**	0.37±0.32	0.28±0.21	0.19±0.07
Episodic Memory	STEM (Ours)	**1.00±0.00**	0.55±0.49	**0.69±0.37**	**0.59±0.33**	**0.66±0.28**
	ARTEM (Ours)	0.95±0.13	0.68±0.44	0.56±0.41	0.54±0.41	0.51±0.38

Table 3: Performance comparison ($F1$) across memory architectures and model variants. STEM and ARTEM demonstrate superior performance in multi-event scenarios, with STEM achieving perfect accuracy on null queries and maintaining robust performance across complex retrieval tasks.

Model	Simple Recall	Chronological
STEM (Ours)	**0.733**	**0.644**
ARTEM (Ours)	0.709	0.585
gemini-2-pro	0.708	0.290
gemini-2-flash-thinking	0.708	0.288
gpt-4o	0.670	0.204
deepseek-v3	0.600	0.103
gemini-2-flash	0.596	0.173
deepseek-r1	0.572	0.147
llama-3.1-405b	0.504	0.129
gpt-4o-mini	0.492	0.077
claude-3-haiku	0.470	0.109
claude-3-5-sonnet	0.470	0.090
o3-mini	0.424	0.044
o1	0.384	0.052
o1-mini	0.300	0.033

Table 4: Performance comparison ($F1$) on simple recall and chronological tasks. The models are evaluated using in-context memory. Our models demonstrate state-of-the-art performance, with STEM achieving the highest scores in both categories.

Episodic Memory Architectures: Our models demonstrate superior overall robustness, with STEM achieving perfect null-query accuracy ($F1 = 1.00$) while maintaining strong multi-event retrieval capabilities across all bins. ARTEM enhances single-event precision through intelligent LLM-guided synthesis.

In-context Approaches: Exhibit significant variability, with reasoning-optimized models (o1-mini) achieving near-perfect null-query performance ($F1 = 0.97$) but failing catastrophically on single-event retrieval ($F1 = 0.05$). Standard LLMs show balanced but degrading performance as event complexity increases.

RAG Systems: Demonstrate improved consistency across event bins, achieving stable performance around $F1 = 0.59 - 0.62$ for multi-event scenarios, but lack unified behavior on confabulation detection ($F1 \in [0.63, 0.91]$).

Fine-tuning: Shows task-specific optimization potential with highest single-event performance ($F1 = 0.83$) but severe degradation in multi-event scenarios and complete failure on confabulation detection ($F1 = 0.00$), indicating overfitting and critical hallucination control shortcomings.

Temporal Reasoning Benchmark

Table 4 positions our approaches within the broader landscape of contemporary language models, revealing our architectures' distinctive advantages in temporal reasoning tasks.

STEM achieves state-of-the-art performance in both simple recall (0.733) and chronological tasks (0.644), while ARTEM maintains competitive recall performance (0.709) and strong temporal reasoning capabilities (0.585). The performance gap is particularly pronounced in chronological tasks, where our models achieve $2-19\times$ improvements over commercial alternatives.

Notably, reasoning-optimized models (o1, o3-mini) show counterintuitive underperformance on chronological tasks, with o1 achieving only 0.052 and o3-mini reaching 0.044. This suggests that general reasoning capabilities do not directly translate to temporal sequence proficiency, supporting the necessity of specialized architectures for episodic memory applications.

Discussion

This section summarizes the performance of our proposed models across the four main tasks to be evaluated: partial cue retrieval, epistemic uncertainty detection, recent event identification, and chronological recall.

Partial Cue Retrieval Both STEM and ARTEM demonstrate exceptional robustness across different question cue categories, effectively handling incomplete or fragmented query information. STEM achieves particularly strong performance for entity-based queries ($F1 = 0.77$) and spatial

information retrieval ($F1 = 0.73$), demonstrating highly effective feature representation and similarity matching capabilities. ARTEM exhibits superior balance across cue types, with exceptionally strong results for temporal queries ($F1 = 0.73$). The systems' remarkable ability to match partial cues across multiple information dimensions (entities, events, spaces, times) validates the multi-channel architecture's effectiveness in capturing diverse aspects of episodic memory. Notably, ARTEM shows substantial improvement in handling complex "full event details" queries (from STEM $F1 = 0.5$ to ARTEM $F1 = 0.67$), demonstrating the architecture's enhanced capacity for comprehensive information integration and constraint satisfaction.

Epistemic Uncertainty Detection The most significant achievement of both architectures lies in their exceptional confabulation detection capabilities. STEM achieves perfect performance ($F1 = 1.00$) on null queries (Bin 0), while ARTEM maintains near-perfect accuracy ($F1 = 0.95$). This represents a critical advancement in addressing the hallucination problem prevalent in contemporary LLM-based systems. The stark contrast with traditional approaches—where reasoning-optimized models like o1-mini achieve high confabulation detection ($F1 = 0.97$) but catastrophically fail on actual retrieval tasks ($F1 \in [0.05, 0.24]$)—demonstrates that our architectures successfully integrate uncertainty estimation with functional memory retrieval. The vigilance threshold mechanism effectively serves as an epistemic uncertainty estimator, enabling the system to recognize when insufficient evidence exists to support a confident answer. This capability is particularly crucial for real-world applications where false positives can have significant consequences.

Recent Event Identification Performance on "Latest" event queries reveals distinct capabilities between the two architectures. ARTEM demonstrates enhanced recent event detection ($F1 = 0.66$) compared to STEM ($F1 = 0.54$), representing a notable improvement in this specific temporal reasoning task. This performance difference indicates that the integrated LLM component in ARTEM provides enhanced capabilities for recency determination and temporal ranking. STEM maintains consistent baseline performance for latest event detection, establishing the foundational capability of the similarity-based retrieval mechanism. The comparative results suggest that while both architectures can handle recent event identification, ARTEM's hybrid approach offers advantages for tasks requiring nuanced temporal discrimination and recency-based prioritization of retrieved information.

Chronological Recall The exceptional performance on chronological ordering tasks represents the most transformative achievement of our architectures, establishing a new paradigm for temporal reasoning in AI systems. STEM achieves outstanding chronological recall performance ($F1 = 0.644$), while ARTEM demonstrates strong capabilities ($F1 = 0.585$), both significantly outperforming all contemporary language models. The benchmark comparison reveals a striking performance hierarchy: our models achieve $2 - 19$ improvements over state-of-the-art alternatives, with the closest competitor (gemini-2-pro) reaching only 0.290. This dramatic performance gap highlights a fundamental limitation in current LLM architectures for temporal sequence understanding. Notably, even reasoning-optimized models (o1: 0.052, o3-mini: 0.044) struggle with chronological tasks, demonstrating that advanced reasoning capabilities do not inherently translate to temporal sequence proficiency. Our architectures' success stems from explicit temporal modeling and biologically-inspired similarity matching that preserves chronological relationships during encoding and retrieval. This research suggests that chronological awareness is a distinct capability that benefits from specialized architectural design. Our models, which incorporate this approach, demonstrate strong performance on applications requiring robust temporal sequence understanding.

The integration of structured memory retrieval with adaptive LLM processing represents a promising direction for developing more reliable and contextually aware episodic memory systems, addressing fundamental limitations in current approaches while maintaining the flexibility required for real-world applications.

Conclusion

We introduced ARTEM, a hybrid architecture integrating LLMs with biologically-inspired Adaptive Resonance Theory networks to enhance episodic memory. Our approach addresses critical LLM limitations in temporal reasoning, cue-based retrieval, and hallucination mitigation.

The STEM component provides a structured memory that preserves multi-modal event information and enables efficient retrieval via vigilance-guided matching. This architecture achieves competitive recall with superior transparency and interpretability over traditional vector-based methods. Its vigilance-guided rejection mechanism is a crucial advance in hallucination control, enabling the system to abstain from answering when information is insufficient.

ARTEM's enhanced performance on chronological reasoning reveals significant potential for advancing AI personalization. It maintains sustained contextual awareness over extended periods, unlike models limited by conversational windows, thus reducing user burden by autonomously retrieving historical context.

Despite these strengths, ARTEM has limitations, including high sensitivity to vigilance parameter calibration and dependencies on the LLM's capabilities. Performance windows are particularly narrow in single-event scenarios, where precise parameter tuning becomes critical. Additionally, the lack of structured real-world data has resulted in a reliance on synthetic benchmarks. This necessitates validation in broader real-world contexts as suitable datasets become available.

The architecture presents significant opportunities for multimodal extensions, such as visual and audio processing. While constrained by underlying LLM limitations, ARTEM's structured memory and adaptive retrieval establish a promising foundation for more multimodal contextually aware and temporally sophisticated AI systems.

Acknowledgments

This research was supported in part by the SMU Research Scholarship awarded to Cassandra Hui-Ming Tan and the Lee Kong Chian Professorship awarded to Ah-Hwee Tan, both by Singapore Management University.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Girish, S.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Bulatov, A.; Kuratov, Y.; and Burtsev, M. 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35: 11079–11091.

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: International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-0-3.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.; et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, 1107–1128.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwińska, A.; Colmenarejo, S. G.; Grefenstette, E.; Ramalho, T.; Agapiou, J.; et al. 2016. Hybrid Computing Using a Neural Network with Dynamic External Memory. *Nature*, 538(7626): 471–476.

Grossberg, S. 2013. Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural networks*, 37: 1–47.

Huet, A.; Houidi, Z. B.; and Rossi, D. 2025. Episodic Memories Generation and Evaluation Benchmark for Large Language Models. *International Conference on Learning Representations*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

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.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Miller, A.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-Value Memory Networks for Directly Reading Documents. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1400–1409.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Packer, C.; Wooders, S.; Lin, K.; Fang, V.; Patil, S. G.; Stoica, I.; and Gonzalez, J. E. 2023. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560*.

Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331.

Rempe, D.; Luo, Z.; Bin Peng, X.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; and Litany, O. 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13756–13766.

Tan, A.-H.; Carpenter, G. A.; and Grossberg, S. 2007. Intelligence through interaction: towards a unified theory for learning. In Liu, D.; Fei, S.; Hou, Z.-G.; Zhang, H.; and Sun, C., eds., *Advances in Neural Networks – ISNN 2007*, volume 4491, 1094–1103. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-72382-0 978-3-540-72383-7. Series Title: Lecture Notes in Computer Science.

Tan, A.-H.; Subagdja, B.; Wang, D.; and Meng, L. 2019. Self-organizing neural networks for universal learning and multimodal memory encoding. *Neural Networks*, 120: 58–73.

Tulving, E. 2002. Episodic memory: From mind to brain. *Annual Review of Psychology*, 53(1): 1–25.

Wang, W.; Dong, L.; Cheng, H.; Liu, X.; Yan, X.; Gao, J.; and Wei, F. 2023. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36: 74530–74543.

Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; and Cui, B. 2024. Retrieval-augmented generation for AI-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. MemoryBank: Enhancing Large Language Models with Long-Term Memory. 38(17): 19724–19731.