MemoryART: Enhancing LLMs via Multi-Memory Models with Adaptive Resonance Theory for Healthcare Agents

Renke Dai[1*], **Hebin Hu**[1*], **Jiahui Zhang**[1], **Yilin Kang**[1†], **Ah-Hwee Tan**[2†]

[1]South-Central Minzu University
[2]Singapore Management University
{2024110293, ylkang}@mail.scuec.edu.cn, ahtan@smu.edu.sg

Abstract

Though promising in healthcare consultation applications, large language models (LLMs) face critical limitations in retaining and utilizing long-term memory across multi-turn interactions. In particular, existing memory enhancing paradigms are constrained by limited context windows and embedding-based retrieval, often failing to maintain task relevance and still suffering from memory prototype collapse in multi-turn healthcare consultation. To address these challenges, we propose a cognitively-inspired memory framework named MemoryART, which is grounded in Adaptive Resonance Theory (ART)—a cognitive and learning theory of how humans and animals adapt to dynamic environments. MemoryART employs three memory modules—working memory, episodic memory, and semantic memory to support task-aware memory organization and dynamic retrieval. Specifically, episodic memory provides the storage of specific experiences along with contextual clues, which is crucial for managing patient-specific information and perfect for multi-turn healthcare consultation interactions. Building upon this concept, MemoryART leverages multi-channel competitive learning and resonance matching to enable efficient and interpretable episodic memory encoding, alleviating issues of prototype collapse and noisy memory associations. For evaluation, we construct a long-term medical dialogue benchmark called MediLongChat using a LLM-based generation pipeline. The resulting dataset features realistic, multi-disease chat histories, each exceeding 100K tokens across 20–30 dialogues, simulating real-world healthcare interaction patterns. Our experimental results show that MemoryART outperforms mainstream approaches in memory-intensive tasks, achieving SOTA results and significantly reducing token consumption across five popular LLMs, confirming its effectiveness and efficiency in providing scalable, reliable memory for LLMs in healthcare.

Code — https://github.com/dairkkriad/MemoryART

Introduction

Medical dialogue agents based on large language models (LLMs) show promise for clinical support (Lee, Zhang, and Zhao 2021; Ren et al. 2024; Alhur 2024) but still face diagnostic accuracy challenges (Wei, Zhao, and Miao 2018; Dixit et al. 2023; Newman-Toker et al. 2024). Effective diagnosis requires models to accurately interpret clinical symptoms and precisely retain and retrieve detailed information about lifestyle habits and medical histories of individual patients. As a result, significant demands are placed on the memory and retrieval capabilities of LLM-based medical dialogue agents (Zhang et al. 2024; Zhao et al. 2025).

Traditional LLM memory mechanisms mainly include two paradigms: in-context learning (ICL) (Wei et al. 2022) and retrieval-augmented generation (RAG) (Lewis et al. 2020). For ICL, LLMs typically rely on fixed-length context windows to simulate memory, allowing them to reason over recent input tokens using self-attention (Vaswani et al. 2017). While effective for short-term tasks, this mechanism is inherently limited by the context window size and cannot retain information across interactions. Approaches like RAG and external memory systems (Borgeaud et al. 2022) attempt to address this limitation by integrating memory beyond the immediate context. However, they introduce new challenges in memory-access efficiency and contextual consistency, especially in continual interaction settings. Overall, traditional memory mechanisms in LLMs remain basic, session-bound, and suffer from the following key limitations:

1. **Over-reliance on embedding-based retrieval:** Most systems (Packer et al. 2023; Zhong et al. 2024) depend heavily on embedding similarity, which often retrieves content that is semantically similar but irrelevant to the specific task, leading to poor alignment with user goals.
2. **Inefficient memory organization:** Some systems (Xu et al. 2025; Kang et al. 2025) adopt either costly LLM-based linking strategies or rigid memory structures, which may lead to segments containing a large number of semantically diverse pages. This, in turn, causes instability in segment summaries and increases retrieval difficulty. Moreover, such an organization may lead to prototype collapse, where heterogeneous content is compressed into overly similar representations, thereby reducing memory interpretability and overall system performance.
3. **Lack of task-awareness:** Memory linking and retrieval are typically task-agnostic, resulting in irrelevant associ-

[*]The first two authors contributed equally to this work.
[†]Corresponding author

ations and weak support for task-driven reasoning.

Working towards a more persistent and adaptive memory design, in this paper, we propose a novel memory system inspired by Adaptive Resonance Theory (ART) (Carpenter and Grossberg 1998; Grossberg 2013), named **MemoryART**, to address the limitations discussed above. This system is designed to support multi-turn healthcare memory tasks, where effective information organization and retrieval are critical. Consistent with theories from cognitive science, we structure the memory into three components: **semantic memory**, **episodic memory** and **working memory** (Wang et al. 2010, 2012). In particular, for the episodic memory module, we introduce a strategy analogous to fusion ART networks (Tan, Carpenter, and Grossberg 2007; Tan et al. 2019), leveraging multi-channel competitive learning and resonance condition matching for encoding events and episodic memories. This approach effectively mitigates the shortcomings of traditional embedding-based memory systems, such as prototype collapse, unstable segment summaries, and inefficient retrieval.

To evaluate MemoryART and other memory systems, we created a long-term medical dialogue dataset named **MediLongChat**. MediLongChat consists of multi-disease chat histories, each exceeding 100K tokens across 20–30 dialogues, simulating real-world healthcare interaction patterns. Additionally, we have designed three evaluation tasks—In-dialogue, Cross-dialogue, and Synthesis Reasoning—to assess the memory capabilities of MemoryART under various conditions. Our results highlight the overall superiority of MemoryART in handling complex, long-context medical reasoning tasks. MemoryART achieves SOTA results, improving F1-score by +14.88 compared to the best previous memory systems, demonstrating its effectiveness in handling complex tasks.

The main contributions of this paper are summarized as follows:

- We propose a cognitively inspired multi-memory framework for LLMs. It enables dynamic recall and reasoning with explicit memory representation, tailored to the challenges of healthcare-related multi-turn sessions.
- We present a self-organizing episodic memory mechanism based on Adaptive Resonance Theory to address the memory limitations of LLMs in continuous interaction scenarios.
- We build a long-term medical dialogue dataset called MediLongChat and design three distinct reasoning tasks to evaluate the memory systems of LLMs.
- Our method MemoryART achieves state-of-the-art performance on both the MediLongChat and LoCoMo benchmarks, improving over the previous best method by +14.88 F1-score on CDR and +6.01 F1-score on multi-hop QA with a relatively small number of tokens.

Related Work

Moving beyond in-context learning and retrieval-augmented generation, there has been new research focusing on enhancing LLMs with memory capabilities. MemGPT introduces a hybrid memory model that separates short-term and long-term memory inspired by operating system design. It supports persistent interaction across sessions but relies heavily on embedding-based retrieval, which may retrieve semantically similar yet task-irrelevant content, limiting its reliability in complex or multi-turn tasks.

A-Mem proposes an agentic memory framework that dynamically organizes memory through structured note-taking, semantic linking, and memory evolution. Compared to MemGPT, A-Mem enriches each memory entry with attributes such as keywords, tags, and contextual descriptions, and builds an evolving network of interconnections to support long-term reasoning. However, its linking process relies on LLMs to automatically construct associations, which is computationally expensive, hard to control, and often task-agnostic—potentially leading to irrelevant or distracting memory connections in goal-specific applications.

MemoryOS extends MemGPT by introducing a more structured and dynamic memory system. Each dialogue turn is encoded as a page, which serves as the basic memory unit. Pages are then grouped into segments based on embedding similarity and the Jaccard similarity of associated keywords. This approach enhances memory relevance compared to MemGPT's purely embedding-based mechanism.

However, due to noise in the dataset, semantically dissimilar pages may be incorrectly grouped into the same segment, increasing retrieval complexity and resulting in fragmented or misleading memory representations. This segmentation strategy directly leads to prototype collapse, where diverse content is compressed into overly similar representations (Govindarajan et al. 2024). Consequently, each segment may contain a large number of heterogeneous pages, making it difficult to maintain a coherent and stable summary. Frequent updates to segment-level summaries become necessary to accommodate conflicting information. Moreover, the retrieval process becomes less efficient, as the model must search through semantically overloaded segments, reducing precision and increasing computational cost.

Model Details

As shown in Fig 1, MemoryART is a comprehensive memory framework comprising three core components: episodic memory, working memory, and semantic memory. It dynamically updates memory and retrieves relevant memory, ensuring coherent and personalized interactions, while avoiding forgetting and hallucination in long conversations and multi-turn conversations. In particular, episodic memory, based on ART, is vital within this framework.

Episodic Memory

In cognitive science, episodic memory refers to the memory system responsible for encoding and retrieving context-rich personal experiences. It enables individuals to recall what happened, in what order, and under what conditions, forming the foundation for experiential learning and long-term reasoning. ART is a cognitive and neural theory that explains how the brain learns to categorize, recognize, and

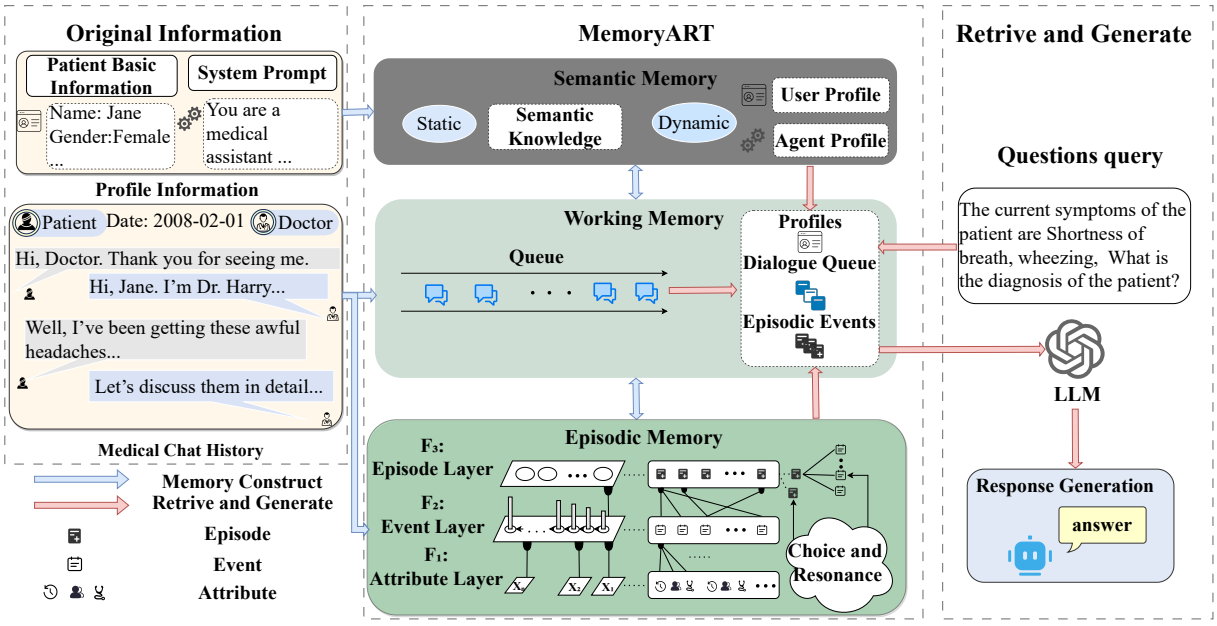

Figure 1: The overall framework of MemoryART consisting of three core components: semantic memory, working memory, and episodic memory. The semantic memory comprises dynamically updated user and agent profiles. The working memory maintains the active dialogue queue and recent episodic event information. The episodic memory module utilizes an ART network to aggregate key information from various events, forming a coherent episode. During interactions, MemoryART dynamically retrieves relevant episodic and semantic memories, enabling LLMs to generate context-aware responses.

predict objects and events in a changing world. It provides a biologically plausible framework explaining how neural systems achieve stable category learning while adapting to novel stimuli in changing environments. Leveraging the self-organizing and incremental characteristics of ART, we can use key information from conversations to build contextual awareness of the current dialogue, which we refer to as episodic memory. To emulate these mechanisms, our MemoryART implements episodic memory through a multi-channel Fusion ART network, which constructs hierarchical memory traces via competitive resonance matching. This module extends ART by fusing multiple channels to dynamically manage a growing collection of past interaction segments. In MemoryART, each episodic memory $\mathcal{M}_E^t$ at time step t is composed of a set of episodes, where each episode E_j is a collection of multiple events, and each event e_i is a multi-channel attribute vector:

$$\mathcal{M}_E^t = \{E_j \mid j = 1, 2, \ldots, M\},$$
$$E_j = \{\mathbf{x}_i \mid i = 1, 2, \ldots, N_j\},$$
$$\mathbf{x}_i = \{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)}\}$$

Here: M is the number of episodes maintained; N_j is the number of events within episode E_j; $\mathbf{x}_i$ is the i-th event represented across K channels. All episodes and their constituent events are dynamically managed by a fusion ART network, which incrementally encodes new events, assigns them to matching episodes, or creates new ones when no suitable match is found. This structure supports continual learning, pattern abstraction, and memory consolidation, while preserving temporal and contextual memory diversity avoiding prototype collapse. In MemoryART, each memory trace is grounded in a set of multi-channel events organized into episodes. These units are dynamically managed by a fusion ART network, enabling adaptive memory growth and structured recall.

Memory Encoding. All episodes in MemoryART are managed by a hierarchical self-organizing fusion ART network composed of three layers: F_3, F_2, F_1. The F_3 layer encodes episodes $E_j = \{\mathbf{x}_i \mid i = 1, 2, \ldots, N_j\}$ composed of several events. The F_2 layer encodes the current input event $\mathbf{x}_t = \{x_t^{(1)}, \ldots, x_t^{(K)}\}$ from recent dialogues across multiple semantic channels. The F_1 layer encodes the semantic channel attributes like summary and people of the event. The dynamics of a fusion ART network can be considered as a system of continuous choice and resonance search processes comprising the basic operations as follows. Given an input event vector $\mathbf{x} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\}$, the Fusion ART network selects or creates a category node in F_3 (episode) and F_2 (event) through a process of competitive matching and adaptive learning. We define the following mechanisms:

- **Choice Function.** For each stored event $\mathbf{x}_i$ across all episodes, we compute a choice score that balances semantic similarity and event salience. This allows the system to identify which events are most relevant to the current input. For each channel k, we define:

$$T_i = \frac{\sum_{k=1}^{K} \gamma^{(k)} \cdot \mathrm{Sim}^{(k)}(\mathbf{x}^{(k)}, \mathbf{x}_i^{(k)})}{\alpha + \sum_{k=1}^{K} \|\mathbf{x}_i^{(k)}\|_1}$$

where $\gamma^{(k)}$ is a channel-specific weighting hyperparameter, and $\alpha > 0$ is the choice regularization parameter. The top-k events with the highest T_i are selected for retrieval.

- **Resonance Condition.** To determine whether an input event $\mathbf{x} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\}$ matches an existing episode node j, the system computes similarity scores across all semantic channels. For each channel k, a channel-specific similarity function $\mathrm{Sim}^{(k)}(\cdot, \cdot)$ is applied between the input $\mathbf{x}^{(k)}$ and the corresponding episode prototype $\mathbf{w}_j^{(k)}$:

$$\mathrm{Match}_j^{(k)} = \mathrm{Sim}^{(k)}(\mathbf{x}^{(k)}, \mathbf{w}_j^{(k)})$$

  Each channel may employ a distinct similarity strategy depending on its channel. A node j is considered to satisfy the resonance condition if the overall similarity exceeds a vigilance threshold $\rho \in [0, 1]$. If no episode meets the threshold, a new one is created.

Based on these mechanisms, the encoding process proceeds as follows. Given a new input event $\mathbf{x}_t$, the system first computes choice scores T_i between $\mathbf{x}_t$ and all stored events across episodes. The top-k events with the highest choice scores are identified, and their corresponding episodes are shortlisted as candidates for resonance matching. For each candidate episode, the system evaluates the resonance condition by comparing $\mathbf{x}_t$ with the episode prototype $\mathbf{w}_j$. If an episode satisfies the vigilance criterion, $\mathbf{x}_t$ is incorporated into the matched episode, and its prototype $\mathbf{w}_j$ is updated. If none of the candidates meet the resonance threshold, a new episode is created. The input event $\mathbf{x}_t$ becomes both the initial member and the prototype $\mathbf{w}_{\text{new}}$ of this newly formed episode. This hierarchical encoding scheme allows MemoryART to grow adaptively while maintaining semantic coherence within each episode avoiding prototype collapse.

Memory Recall. Similar to encoding, recall begins by computing choice scores and selecting top-k episodes satisfied the resonance condition, then choose top-k events satisfied the resonance condition. Unlike encoding, it performs no prototype update or episode creation. Instead, it retrieves relevant contextual information from matched episodes and events as a return. The whole encoding and recall process follows Algorithm 1.

Working Memory

Working memory refers to the temporary storage and manipulation of information necessary for reasoning in cognitive science. Notably, it is limited in both capacity and duration. Inspired by this cognitive function, in MemoryART, we design it as a limited-capacity buffer that supports short-term reasoning during ongoing interactions. In our implementation, working memory stores real-time conversational data with a fixed capacity of N dialogue turns. Each turn is stored as a tuple (Q_t, R_t, T_t), where Q_t is the user's query, R_t is the model's response, and T_t denotes the timestamp. Formally, the working memory at time step t is defined as:

$$\mathcal{M}_W^t = \{(Q_{t-i}, R_{t-i}, T_{t-i}) \mid i = 0, 1, \ldots, N-1\}$$

Algorithm 1: Encoding and Recall in Episodic Memory

Require: Input event $\mathbf{x}_t = \{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(K)}\}$
Initialize candidate set $\mathcal{C} \leftarrow \emptyset$
for each episode E_j **do**
  Compute choice score between $\mathbf{x}_t$ and prototype $\mathbf{w}_j$
  if resonance between $\mathbf{x}_t$ and $\mathbf{w}_j$ is satisfied **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{E_j\}$
  end if
end for
if $\mathcal{C} = \emptyset$ **then**
  Create new episode with prototype $\mathbf{w}_{\text{new}} \leftarrow \mathbf{x}_t$
  return $\emptyset$
end if
Collect all events from episodes in $\mathcal{C}$
Compute choice scores T_i between $\mathbf{x}_t$ and collected events
Select top-k events $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}\}$
for each selected event $\mathbf{x}_{i_j}$ **do**
  Let E_j be the episode containing $\mathbf{x}_{i_j}$
  if resonance between $\mathbf{x}_t$ and prototype $\mathbf{w}_j$ is satisfied **then**
    Append $\mathbf{x}_t$ to episode E_j
    Update prototype $\mathbf{w}_j$
    return $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}\}$
  end if
end for

As the conversation progresses, new entries are appended while the oldest ones are evicted, following a sliding-window policy. This allows the model to always access the most recent and relevant context with minimal computational overhead. So we employ a First-In-First-Out update policy for information migration to the episodic memory.

Semantic Memory

Semantic memory in cognitive science refers to the memory system responsible for storing general knowledge, facts, concepts, and language meanings that are not tied to specific experiences. In MemoryART, we define semantic memory as the repository of factual and conceptual knowledge. We further divide semantic memory into two subcomponents: static knowledge and dynamic knowledge.

- Static Knowledge (Parameter Memory): This component includes all knowledge embedded within the pre-trained parameters of the LLM.
- Dynamic Knowledge (Editable Profiles): Dynamic semantic memory includes editable, structured information evolving over time, comprising two key submodules:
  - User Profile: personalized information about the user, such as identity, preferences, goals, and history.
  - Agent Profile: system-level role definitions, behavior policies, domain expertise, and task capabilities.

These profiles are stored explicitly and can be updated during interaction, allowing the system to adapt its behavior and reasoning strategy to the user and task context.

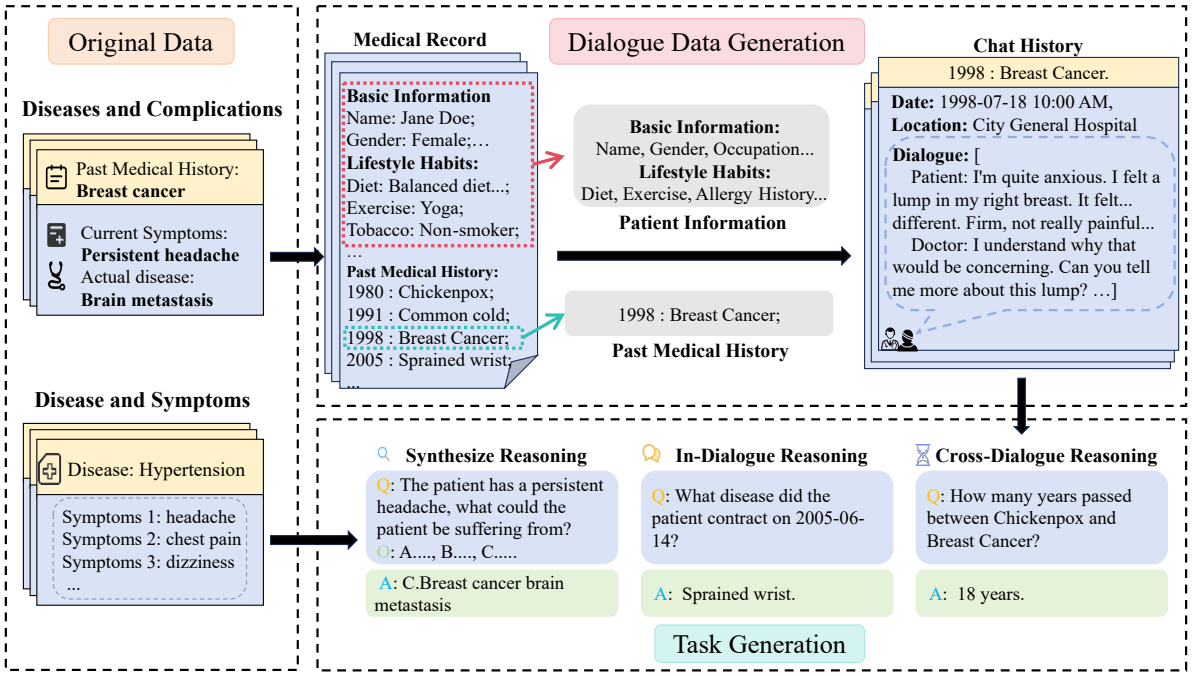

Figure 2: Overview of our dataset generation pipeline.

Dataset—MediLongChat

To evaluate the performance of MemoryART, we create a new dataset containing long-term medical dialogues generated by LLMs within the healthcare domain.

Generative Pipeline for Our Dataset

Figure 2 illustrates our data generation pipeline.

Original Data. The original data comprises two parts:

- **Disease and Complication**. This dataset records complications and secondary diseases associated with primary illnesses. The dataset is initially generated by LLMs and subsequently manually reviewed and corrected.
- **Disease and Symptom**. This dataset records various sets of symptoms associated with diseases, with a single disease potentially having multiple distinct symptom sets. This dataset is primarily used to generate distractor options for subsequent synthesis reasoning tasks.

Based on the original domain data sets, the synthetic datasets consisting of medical records and chat histories are generated using LLMs.

Medical Record. A medical record contains complete information about a specific patient. Each medical record consists of four components: personal information, lifestyle habits, past medical history, and additional information. Generated by LLMs based on original data, medical records represent fictitious yet realistic lifelong patient profiles, thus avoiding privacy concerns. Notably, medical records serve as intermediate data during the generation process and are not included explicitly in the final dataset. They also remain hidden from the LLMs during evaluation tasks.

Chat History. Chat history constitutes the core component of our dataset, encapsulating all medical consultations related to diseases experienced by a patient throughout their lifetime. It consists of multiple dialogues, with each dialogue representing a consultation regarding a particular disease, simulating realistic patient-physician interactions. Each dialogue captures details such as date and location, greetings, symptom inquiry, routine examinations, and treatment discussions of the consultation. Typically, each dialogue is about 40 turns (around 5K tokens), and a full chat history (20-30 dialogues) has approximately 100K tokens.

Evaluation Tasks

Based on the generated medical records and chat histories, we design three distinct tasks to evaluate how accurately MemoryART recalls patient consultation information. The three tasks include: In-dialogue Reasoning, Cross-dialogue Reasoning, and Synthesis Reasoning.

In-dialogue Reasoning (IDR). The In-dialogue Reasoning task involves questions exclusively derived from a single dialogue, focusing on extracting and summarizing key details such as consultation date, location, disease type, and treatment plan. This task assesses the ability of LLMs to extract and succinctly summarize simple information from a single consultation dialogue.

Cross-dialogue Reasoning (CDR). The Cross-dialogue Reasoning tasks contain questions involving multiple dialogues. This task assesses the ability of LLMs to identify relationships between multiple diseases, manage temporal sequencing (e.g., duration between two illnesses), or answer adversarial questions (e.g., whether a patient ever contracted

Model	Method	SR Accuracy	IDR		CDR		Ranking		Token Count
			F1	BLEU	F1	BLEU	F1	BLEU	
Deepseek-R1	Baseline	80.00	33.49	11.12	20.36	2.41	4	3	-
	MemoryOS	66.67	38.11	4.36	52.17	10.99	2	2	13503
	MemoryBank	48.75	15.24	1.51	28.58	3.46	5	5	4272
	A-Mem	47.50	26.36	5.03	36.44	5.33	3	4	16680
	MemoryART	**85.00**	**45.73**	**15.29**	**69.61**	**35.35**	**1**	**1**	6101
Qwen3-235 B	Baseline	80.00	27.19	6.31	19.61	2.13	4	4	-
	MemoryOS	66.67	36.63	4.32	53.69	12.14	2	3	13503
	MemoryBank	45.00	17.96	1.57	31.49	3.66	5	5	4272
	A-Mem	52.38	30.46	5.97	50.98	8.44	3	2	16680
	MemoryART	**83.75**	**41.29**	**13.11**	**55.96**	**29.66**	**1**	**1**	6101
Ernie-4.5-turbo	Baseline	80.00	31.74	8.73	16.46	1.33	3	4	-
	MemoryOS	50.00	22.78	2.04	**50.58**	10.94	2	2	13503
	MemoryBank	38.75	7.78	0.60	5.79	0.57	5	5	4272
	A-Mem	37.50	18.77	3.84	10.98	1.54	4	3	16680
	MemoryART	**85.00**	**33.85**	**9.80**	47.71	**27.81**	**1**	**1**	6101
GPT-4o mini	Baseline	80.00	23.30	3.28	24.25	2.46	5	4	-
	MemoryOS	75.00	29.29	2.12	48.76	4.04	2	3	13503
	MemoryBank	46.81	16.64	1.70	18.29	2.62	4	5	4272
	A-Mem	48.15	25.10	4.41	20.34	2.35	3	2	16680
	MemoryART	**86.25**	**38.79**	**11.40**	**74.62**	**36.40**	**1**	**1**	6101
GPT-4.1 mini	Baseline	83.75	27.15	6.18	18.37	1.66	4	4	-
	MemoryOS	75.00	30.32	2.28	36.59	2.59	2	3	13503
	MemoryBank	47.50	16.24	1.77	20.22	2.76	5	5	4272
	A-Mem	40.00	25.68	4.66	22.04	3.03	3	2	16680
	MemoryART	**85.00**	**41.93**	**13.46**	**55.81**	**31.47**	**1**	**1**	6101

Table 1: Performance comparison on the MediLongChat benchmark. F1 and BLEU are reported for IDR and CDR. Rankings are based on average performance in IDR and CDR. Typically boldfaced values indicate the best performance.

a particular disease). A complete chat history may exceed the context window of the LLM, so strong memory capabilities become essential for success.

Synthesis Reasoning (SR). The Synthesis Reasoning task requires the LLM to diagnose a secondary disease or complication based on provided symptoms, given a complete medical chat history. This challenging task demands that the model not only recall the entire disease history of the patient but also accurately link current symptoms with past diseases. Strong performance on Synthesis Reasoning is essential for LLMs to effectively serve as medical assistants. Due to the complexity of this task, we design it as a multiple-choice format, with distractor options selected based on symptom similarity from the Disease and Symptom dataset.

Experiments

Experiment Settings

Dataset. We evaluate MemoryART using two datasets: MediLongChat and LoCoMo (Maharana et al. 2024). MediLongChat is the dataset containing long-text medical dialogues mentioned above. The LoCoMo benchmark is explicitly constructed to evaluate the long-term conversational memory of LLMs.

Evaluation Metrics. On the MediLongChat benchmark, we evaluate performance using accuracy for the SR. For the IDR and CDR, we report both F1 score and BLEU-1 as evaluation metrics. On the LoCoMo benchmark, standard F1 and BLEU-1 are employed to evaluate the model's performance.

Compared Methods. In order to explore the performance of the MemoryART, we compare the MemoryART results against several other models, including TiM (Liu et al. 2023), MemGPT, MemoryBank, A-Mem, and MemoryOS. More detailed introduction is provided in the appendix.

Results

We report comprehensive results in Table 1 and Table 2, comparing the proposed MemoryART with existing memory-augmented methods across two benchmarks.

Results on MediLongChat. Table 1 summarizes the performance on three tasks (SR, IDR, and CDR) of the MediLongChat benchmark. Across all models—GPT, Deepseek, Qwen, and Ernie—MemoryART achieves the top performance in almost all metrics, while maintaining a low memory footprint, indicating its efficiency in long-context retention.

Interestingly, the baseline (no memory, split dialogue input) performs strongly, especially on the SR task. We hypothesize this is because memory mechanisms, while retrieving relevant information, can also introduce irrelevant or incorrect data. In the accuracy-critical healthcare context, this noise misleads the LLM, whereas the baseline

Model	Method	Single Hop		Multi Hop		Temporal		Open Domain	
		F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1
GPT-4o-mini	Tim	16.25	13.12	18.43	17.35	8.35	7.32	23.74	22.05
	MemoryBank	5.00	4.77	9.68	6.99	5.56	5.94	6.61	5.16
	MemGPT	26.65	17.72	25.52	19.44	9.15	7.44	41.04	34.34
	A-Mem	27.02	20.09	45.85	36.67	12.14	12.00	44.65	37.06
	MemoryOS	**35.27**	25.22	41.15	30.76	20.02	16.52	48.62	42.99
	MemoryART	31.43	**36.05**	**47.04**	**36.85**	**27.46**	**22.07**	**48.86**	**43.33**

Table 2: Performance comparison on LoCoMo. Questions are classified into four categories—Single-hop, Multi-hop, Temporal, and Open-domain—to evaluate the long-term conversational memory of LLMs. Boldfaced values indicate the best performance.

avoids retrieval-induced noise. Additionally, all memory-based methods underperform on IDR tasks compared to both the baseline and their own CDR performance. This counter-intuitive result occurs because the baseline's approach preserves the local in-dialogue context crucial for IDR. Conversely, the encode-retrieve process disrupts this local context, degrading IDR performance, even while it benefits CDR tasks by aggregating cross-dialogue information.

Results on LoCoMo. Table 2 presents the evaluation results on the LoCoMo benchmark, focusing on fine-grained memory tasks: Single-hop, Multi-hop, Temporal, and Open-domain reasoning. Compared to strong memory baselines such as MemoryOS and A-Mem, MemoryART achieves state-of-the-art performance in all tasks.

Ablation Study on Retrieved Events

To investigate the influence of memory retrieval granularity, we conduct an ablation study by varying the number of retrieved events k in the MemoryART framework. As illustrated in Figure 3, we evaluate model performance on both IDR and CDR tasks using F1 and BLEU-1.

In the IDR setting, performance consistently improves as more memory events are retrieved, indicating that expanding the memory scope enhances in-dialogue reasoning. However, beyond a certain retrieval size, the improvement plateaus or slightly declines. These findings highlight the importance of calibrating the number of retrieved events to balance relevance in complex reasoning scenarios.

We also examine the impact of semantic channels on the resonance mechanism by varying the number of similarity channels m involved in episode matching. As shown in Figure 4, incorporating additional channels—such as timestamp proximity, symptom overlap, and structural cues—progressively improves model performance.

In both IDR and CDR tasks, we observe that leveraging multi-channel similarity consistently enhances retrieval accuracy and reasoning quality. This demonstrates that heterogeneous semantic cues offer complementary perspectives, allowing models to better align with salient memory traces.

Prior methods rely on chunk-level summarization over noisy dialogues, often capturing superficial content such as greetings or small talk. This results in prototype collapse, where distinct clinical events are reduced to similar, non-discriminative summaries—making accurate memory retrieval difficult. In contrast, built on ART, MemoryART incorporates a Choice Function and Resonance Condition that adaptively select semantically aligned memory traces. This mechanism ensures fine-grained, context-sensitive retrieval, even under noisy or sparse clue settings, leading to more faithful and case-specific reasoning. Hence, MemoryART constructs structured, user-specific memory entries and applies a dual retrieval strategy: clue-based symbolic matching and dense embedding similarity.

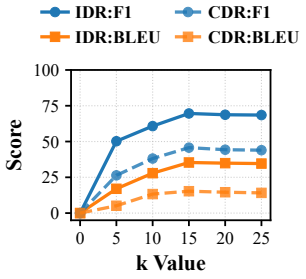

Figure 3: Ablation results on k for IDR and CDR.

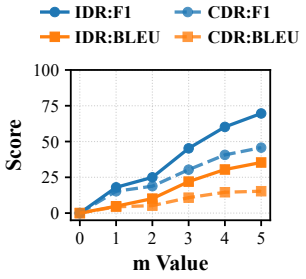

Figure 4: Ablation results on m for IDR and CDR.

Conclusion

To address the challenges faced by current memory systems in medical agents, we propose MemoryART—a novel memory system that uses fusion ART to establish episodic memory, dynamically updating and retrieving event-related memories while integrating working and semantic memory. We also build MediLongChat, a long-term medical dialogue dataset with three distinct tasks for evaluating MemoryART against baselines. Evaluations on Locomo and MediLongChat show that MemoryART outperforms existing methods in accuracy while significantly reducing token usage. Although MemoryART effectively mitigates shortcomings of traditional embedding-based memory systems, its task-relevance episodic memory partially relies on manual design, limiting the mechanism's generalizability to more diverse and complex tasks. For future work, we aim to leverage MemoryART to encode multimodal data such as medical images and physiological signals, aligning these with multimodal LLMs to enhance the capabilities of healthcare agents.

Acknowledgments

This work was partly supported by the General Project of the Central Universities of China (No. CZY23007), Hubei Province Key Research and Development Special Project of Science and Technology Innovation Plan (2023BAB087), Wuhan Key Research and Development Projects (2023010402010614), and Lee Kong Chian Professorship awarded to Ah-Hwee Tan by Singapore Management University.

References

Alhur, A. 2024. Redefining healthcare with artificial intelligence (AI): the contributions of ChatGPT, Gemini, and Copilot. *Cureus*, 16(4).

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Carpenter, G. A.; and Grossberg, S. 1998. Adaptive resonance theory (ART). In *The handbook of brain theory and neural networks*, 79–82.

Dixit, R. A.; Boxley, C. L.; Samuel, S.; Mohan, V.; Ratwani, R. M.; and Gold, J. A. 2023. Electronic health record use issues and diagnostic error: a scoping review and framework. *Journal of Patient Safety*, 19(1): e25–e30.

Govindarajan, H.; Sidén, P.; Roll, J.; and Lindsten, F. 2024. On partial prototype collapse in clustering-based self-supervised learning.

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

Kang, J.; Ji, M.; Zhao, Z.; and Bai, T. 2025. Memory OS of AI Agent. *arXiv preprint arXiv:2506.06326*.

Lee, C. H.; Zhang, Z.; and Zhao, X. 2021. A survey of smart healthcare for the elderly based on user requirements and supply accessibility. In *5th International Conference on Crowd Science and Engineering*, 108–112.

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. *Advances in neural information processing systems*, 33: 9459–9474.

Liu, L.; Yang, X.; Shen, Y.; Hu, B.; Zhang, Z.; Gu, J.; and Zhang, G. 2023. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*.

Maharana, A.; Lee, D.-H.; Tulyakov, S.; Bansal, M.; Barbieri, F.; and Fang, Y. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13851–13870.

Newman-Toker, D. E.; Nassery, N.; Schaffer, A. C.; Yu-Moe, C. W.; Clemens, G. D.; Wang, Z.; Zhu, Y.; Tehrani, A. S. S.; Fanai, M.; Hassoon, A.; et al. 2024. Burden of serious harms from diagnostic error in the USA. *BMJ Quality & Safety*, 33(2): 109–120.

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

Ren, Z.; Zhan, Y.; Yu, B.; Ding, L.; and Tao, D. 2024. Healthcare copilot: Eliciting the power of general llms for medical consultation. *CoRR*.

Tan, A.-H.; Carpenter, G. A.; and Grossberg, S. 2007. Intelligence through interaction: Towards a unified theory for learning. In *International symposium on neural networks*, 1094–1103. Springer.

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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Subagdja, B.; Tan, A.-H.; and Starzyk, J. A. 2010. A self-organizing approach to episodic memory modeling. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Wang, W.; Subagdja, B.; Tan, A.-H.; and Starzyk, J. A. 2012. Neural modeling of episodic memory: Encoding, retrieval, and forgetting. *IEEE transactions on neural networks and learning systems*, 23(10): 1574–1586.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wei, S.; Zhao, X.; and Miao, C. 2018. A comprehensive exploration to the machine learning techniques for diabetes identification. In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, 291–295. IEEE.

Xu, W.; Mei, K.; Gao, H.; Tan, J.; Liang, Z.; and Zhang, Y. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Zhang, K.; Kang, Y.; Zhao, F.; and Liu, X. 2024. LLM-based Medical Assistant Personalization with Short-and Long-Term Memory Coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2386–2398.

Zhao, X.; Liu, S.; Yang, S.-Y.; and Miao, C. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, 4442–4457.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19724–19731.