RESEARCH-ARTICLE

Human-Like Remembering and Forgetting in LLM Agents: An ACT-R-Inspired Memory Architecture

YUDAI HONDA, Kyushu University, Fukuoka, Fukuoka, Japan

YUKI FUJITA, Tokyo National University of Fine Arts and Music, Tokyo, Tokyo, Japan

KEIICHI ZEMPO, University of Tsukuba, Tsukuba, Ibaraki Prefecture, Japan

SHOGO FUKUSHIMA, Kyushu University, Fukuoka, Fukuoka, Japan

Open Access Support provided by:

University of Tsukuba

Kyushu University

Tokyo National University of Fine Arts and Music

Published: 10 November 2025

HAI '25: International Conference on Human-Agent Interaction
November 10 - 13, 2025
Yokohama, Japan

HAI '25: Proceedings of the 13th International Conference on Human-Agent Interaction (November 2025)
https://doi.org/10.1145/3765766.3765803
ISBN: 9798400721786

Human-Like Remembering and Forgetting in LLM Agents: An ACT-R-Inspired Memory Architecture

Yudai Honda
Graduate School of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
honda.yudai.771@s.kyushu-u.ac.jp

Yuki Fujita
Art Media Center
Tokyo University of the Arts
Tokyo, Japan
fujita.yuki@noc.geidai.ac.jp

Keiichi Zempo
Institute of Systems and Information Engineering
University of Tsukuba
Tsukuba, Japan
zempo@iit.tsukuba.ac.jp

Shogo Fukushima
Faculty of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
shogo@ait.kyushu-u.ac.jp

Abstract

This study explores the implementation of human-like memory behavior in language agents by integrating the ACT-R cognitive architecture with large language models (LLMs). We designed a dialogue agent capable of dynamically retrieving and forgetting memories based on context, time, and usage frequency. The proposed system utilizes a vector-based activation mechanism, incorporating temporal decay, semantic similarity, and probabilistic noise to mimic natural memory dynamics. Simulation experiments confirmed the model's ability to reproduce memory reinforcement through repeated topics, as well as stochastic variability in memory retrieval, reflecting human memory behavior. We also identified optimal parameters for balancing contextual sensitivity and memory stability. This work contributes to the development of more human-compatible AI dialogue systems by modeling memory not as mere storage, but as a strategic, context-sensitive process.

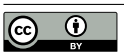
CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; *Knowledge representation and reasoning*; Natural language processing.

Keywords

Cognitive architecture, Memory modeling, ACT-R, Large Language Models, Forgetting, Dialogue systems, Natural language processing

ACM Reference Format:
Yudai Honda, Yuki Fujita, Keiichi Zempo, and Shogo Fukushima. 2025. Human-Like Remembering and Forgetting in LLM Agents: An ACT-R-Inspired Memory Architecture. In *13th International Conference on Human-Agent Interaction (HAI '25), November 10–13, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3765766.3765803



HAI '25, Yokohama, Japan

ACM ISBN 979-8-4007-2178-6/25/11
https://doi.org/10.1145/3765766.3765803

1 Introduction

Human memory is not a static repository for storing and retrieving information but a dynamic, context-dependent process that evolves over time and experience. Phenomena such as spontaneous recollection and natural forgetting are not merely noise but reflect the brain's strategic regulation of memory to support coherent, adaptive behavior. Forgetting, in particular, is not memory failure but a functionally beneficial suppression process mediated by the prefrontal cortex, allowing emotional regulation, cognitive load reduction, and social harmony [2]. In this sense, humans optimize memory not by remembering everything, but by choosing what not to recall, a process tightly related to the principle of bounded rationality [6] —making "good enough" decisions under cognitive constraints.

In contrast, large language models (LLMs) excel at generating human-like language but lack the ability to strategically manage memory. Their memory systems depend on external retrieval or embedding similarity, without mechanisms for context-sensitive remembering or forgetting. This often leads to incoherent dialogues or inappropriate responses despite referencing previous utterances. These limitations suggest the need to shift from "accumulation" to "control" in memory design for conversational agents.

To explore this, we draw inspiration from ACT-R (Adaptive Control of Thought–Rational), a cognitive architecture that models human reasoning and memory processes. While recent studies have integrated ACT-R with LLMs—typically for improving task consistency or explainability [14]—they have rarely addressed the dynamic nature of long-term memory, including probabilistic retrieval or strategic forgetting.

In this study, we propose a language agent that implements a human-like memory mechanism based on ACT-R's memory model. The agent stores past utterances and experiences as memory chunks and selectively recalls them by computing cosine similarity with current inputs. This enables context-sensitive and temporally dynamic retrieval, simulating variations in recallability over time and usage frequency.

By embedding retrieved memory chunks into LLM prompts, the agent can generate responses grounded in prior experiences,

making memory behaviors such as bias and forgetting both visible and controllable. The contributions of this work are twofold:

(1) To construct and evaluate a dialogue agent that dynamically recalls and forgets memories in a human-like manner based on context and time;
(2) To analyze how parameter configurations in the ACT-R-based activation model affect memory retrieval, and to identify design guidelines for future implementations.

Unlike conventional memory-augmented LLMs (e.g., RAG or retrieval-based history systems), our approach explicitly models the structure and dynamics of memory. This work explores the potential of integrating cognitive models and LLMs to realize more human-like conversational agents.

2 Related Work

2.1 Memory-Augmented LLMs and Forgetting Mechanisms

Among approaches to augment the memory capabilities of large language models (LLMs), Retrieval-Augmented Generation (RAG) is one of the most representative. RAG enhances output consistency by retrieving contextually relevant information from an external knowledge base and appending it to the prompt. Models such as LongMem [13] and Titans [4] introduce more persistent long-term memory modules, enabling flexible memory access independent of the attention mechanism.

In reinforcement learning, techniques like Prioritized Experience Replay [10] and Elastic Weight Consolidation (EWC) [8] have been developed to regulate forgetting during training. More recently, there have been attempts to incorporate forgetting as a strategic function, mimicking human memory. For instance, Hou et al. (2024) proposed an agent that assigns "trigger words" to each utterance, selectively reactivating memories based on matches [7]. Additionally, they introduced a mechanism for dynamically adjusting memory retrieval priority based on lifespan and frequency of access.

Furthermore, Sumida et al. (2024) proposed a model that dynamically adjusts memory retention and forgetting based on psychological indicators such as emotional arousal and surprise [12]. Derbinsky et al. (2012) introduced forgetting policies within the Soar architecture that automatically discard knowledge deemed unlikely to be useful—based on low base-level activation and high reconstructability—and demonstrated improved scalability and reactivity in both working and procedural memory systems [5]. While these studies show promise in capturing the ambiguity and selectivity of human memory, many existing methods still rely on static scoring functions and thresholds. As a result, they fall short in modeling dynamic memory changes influenced by time, usage frequency, and contextual cues.

This study aims to address these limitations by implementing a memory agent that supports mathematically controlled, probabilistic, and context-sensitive memory activation and strategic forgetting.

2.2 Memory Activation and Language Processing in ACT-R

ACT-R is a cognitive architecture that mathematically models mechanisms such as memory and reasoning [1]. In this framework, memory is organized into units called *chunks*, and the probability of recall is determined by an activation value $A(m)$ for a chunk m:

$$A(m) = B(m) + \sum_{j \in C} W_j \cdot S_j + \epsilon. \tag{1}$$

Here, $B(m)$ is the base-level activation reflecting the frequency and recency of prior use; C is the set of contextually relevant chunks; W_j is the attentional weight assigned to chunk j in C; S_j is the semantic similarity between chunk j and m; and ϵ is a Gaussian noise term introducing stochastic variability in retrieval.

The base-level activation $B(m)$ is defined as:

$$B(m) = \ln\left(\sum_{i=1}^{n} (t - t_i)^{-d}\right). \tag{2}$$

where t is the current time, t_i is the time of the i-th retrieval of chunk m, n is the total number of past retrievals, and d is a decay parameter (typically around 0.5). This formulation ensures that recently and frequently used memories remain highly accessible, while others naturally decay over time.

Combined with spreading activation based on semantic similarity, this mechanism enables flexible, context-sensitive retrieval that resembles human-like memory recall. Applications include syntactic parsing, semantic inference, and context-aware robotic response [3, 9, 11].

2.3 Toward ACT-R × LLM Integration

In recent years, there has been a growing interest in integrating ACT-R with LLMs. Most existing work, however, treats ACT-R as a downstream component used to evaluate decision-making or supplement prompt design. For example, Wu et al. (2024) injected ACT-R-based reasoning into LLMs to improve decision quality in design tasks [14]. Other studies have explored the automatic generation of ACT-R models using LLMs [15].

In contrast, this study directly integrates the ACT-R memory activation model into the generation process of LLMs. In doing so, memory recall and forgetting become directly linked to the content generation itself. This represents a novel approach to memory control, introducing transparency and controllability into the otherwise opaque LLM generation process.

Moreover, the proposed framework bridges symbolic processing (ACT-R) and distributed representations (LLMs) via the common concept of *memory activation*, thereby providing a mathematical connection between symbolic reasoning and probabilistic generation, which have traditionally been separated.

3 Methodology

3.1 System Architecture

The proposed system is a memory-augmented dialogue agent designed to autonomously recall and forget memory chunks based on user interaction, context, and time. Rather than accumulating memory indefinitely, the agent dynamically reactivates relevant information and gradually suppresses or discards low-activation

memories, emulating the adaptive and bounded nature of human memory. Figure 1 illustrates the overall architecture.

The system comprises five core components:

(1) **User Interface (UI):** Handles real-time communication with the user, including input reception and response display.
(2) **ACT-R Cognitive Core:** Implements the core processing pipeline using the `pyactr` library, based on the ACT-R architecture. It is responsible for input parsing, memory retrieval, and output generation, and governs the overall cognitive flow of the system.
(3) **Extended Declarative Memory Module:** A novel module introduced in this study, it extends the standard ACT-R declarative memory to support natural language-based interactions. It computes activation scores dynamically by integrating semantic similarity in a vector space, frequency of use, temporal decay, and contextual relevance. This enables the reproduction of human-like memory phenomena such as probabilistic recall and forgetting.
(4) **Memory Database:** A persistent storage component that retains all memory chunks, including dialogue history, episodic facts, and metadata such as timestamps and usage counts.
(5) **Control Parameters:** A configurable set of parameters that govern the behavior of the memory system, including decay rate, noise level, spreading activation weights, and activation thresholds.

Together, these components enable the agent to engage in interactive conversations while maintaining a bounded, dynamic memory system that adapts to user context and dialogue flow.

3.2 User Interface and Input/Output Processing

User utterances are entered into the system through the Dialog Environment and stored in ACT-R's Speech Input Buffer. After response generation, the output is placed in the Speech Output Buffer and returned to the user via the same environment as the System Response.

This layer serves as an I/O gate linking ACT-R's internal processes with the external world, responsible for receiving natural language input and delivering responses. The Dialog Environment is a module that extends the standard ACT-R environment class to handle interactive input/output. It captures user utterances as speech chunks and synchronizes response generation and delivery with simulation time.

In this architecture, the Dialog Environment functions as a conduit between ACT-R's internal mechanisms and the outside world, facilitating the input of natural language into the model and the delivery of generated output back to the user.

3.3 ACT-R Cognitive Core

The ACT-R module forms the cognitive core of the system, handling information processing based on human cognitive mechanisms. It consists of three main components:

First, the **Working Memory Buffers** include the Speech Input Buffer (holding user input), the Retrieval Buffer (storing memory queries and results), and the Speech Output Buffer (holding generated responses). These buffers temporarily store and transfer information necessary for dialogue.

Second, the **Production Rules** constitute the central engine of ACT-R's operations. Organized into rule sets for Input Processing, Memory Retrieval, and Response Generation, they execute sequential processing based on the current state of buffers. These rules monitor buffer contents and perform actions (e.g., chunk creation, transfer, output) based on specific conditions.

Third, in the **Memory Retrieval Flow**, a user utterance is entered into the Speech Input Buffer and structured into a chunk (e.g., `isa = fact, content = "How are you?"`) that is placed into the Retrieval Buffer as a search query. Triggered by this, relevant production rules fire and initiate the memory retrieval process.

In this system, the retrieval process is enhanced beyond simple attribute matching. It evaluates *semantic similarity* between the query and memory content using vector-based spreading activation, incorporates *base-level activation* (BLA) derived from historical access patterns and temporal decay, and adds *stochastic noise* to simulate variability. The total activation value is computed, and only chunks whose activation exceeds a preset threshold are probabilistically retrieved and returned to the Retrieval Buffer.

This process allows the model to recall memory not through exact keyword matching, but based on contextual and semantic relevance—enabling more flexible, human-like dialogue responses.

3.4 Extended Declarative Memory Retrieval Module

This memory retrieval module extends ACT-R's base-level learning mechanism by incorporating context-based search using natural language vectors. The retrieval process proceeds as follows:

First, the search query (extracted from the `content` field of the user utterance) and all memory chunks are converted into semantic vector representations. The cosine similarity between the query and each chunk is calculated, serving as the measure of contextual relevance (i.e., spreading activation).

Each memory chunk maintains a base-level activation value B based on its history of access (see Chapter 2).

During retrieval, a uniform context weight w is applied to all chunks, and the total activation A is computed by adding the contextual similarity to the base-level activation as follows:

$$A = B + w \cdot \text{cosine\_similarity} + \epsilon. \tag{3}$$

Here, $\epsilon \sim \mathcal{N}(0, \sigma)$ is a Gaussian noise term that reflects the probabilistic nature of memory retrieval.

A memory chunk is recalled only if the total activation A exceeds a predefined retrieval threshold. Even if a chunk is semantically relevant, it may not be retrieved if its activation is low due to infrequent or distant past usage—thus naturally modeling memory failure or forgetting.

3.5 Parameter Design and Control

The memory retrieval process in this system is governed by several key parameters, each corresponding to a component of the activation function (see Equation $A = B + wS + \epsilon$):

- **decay (parameter d):** Controls the rate at which base-level activation declines over time. A larger value leads to faster forgetting of older memories.

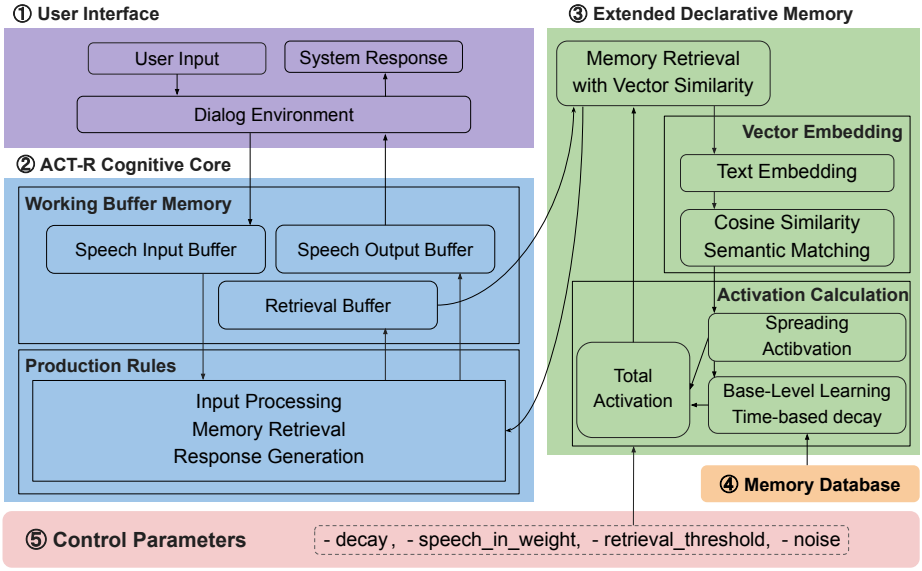


Figure 1: Overall system architecture.

- **speech_in_weight (parameter w):** Scales the influence of semantic similarity (cosine similarity) in the total activation. Higher values increase the chance of retrieving semantically similar memories.
- **noise (standard deviation σ):** Determines the magnitude of Gaussian noise added to activation values, introducing stochasticity and variability into the recall process.
- **retrieval_threshold:** Defines the activation threshold required for memory recall. Chunks below this threshold are not retrieved, effectively representing forgotten memories.

These parameters enable memory retrieval to reflect not just relevance or similarity, but also temporal decay, probabilistic uncertainty, and contextual weighting—creating a memory system that captures human-like imperfection and adaptivity.

3.6 Comparison of Embedding Methods

In this system, the semantic relevance between memory chunks and retrieval queries is evaluated using cosine similarity of natural language vector embeddings, which serves as the basis for activation computation. In this study, we compare two types of language embedding models:

- **OpenAI:** `text-embedding-3-small`
- **Sentence-BERT (SBERT):** `all-MiniLM-L6-v2`

These models are based on different architectural principles and training methodologies, leading to distinct distributional characteristics in their embeddings.

For a concrete comparison, we extracted 50 memory chunks and 10 retrieval queries from the dataset used in the experiments (described later). We computed cosine similarity scores between all query–memory pairs using each embedding model, and visualized the results as heatmaps to illustrate the similarity distributions (see Figure 2).

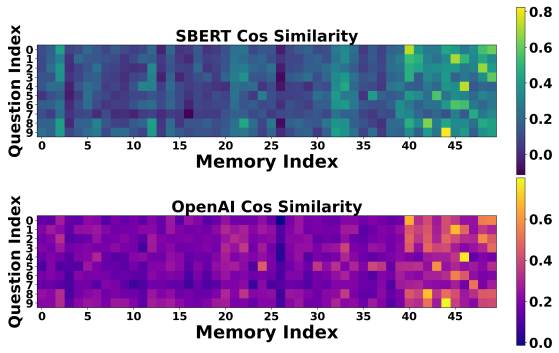


Figure 2: Cosine similarity heatmaps: SBERT shows sparser, high-contrast matches; OpenAI yields broader, smoother distributions.

As shown in the heatmaps in Figure 2, the outputs of the two models exhibited the following differences:

SBERT (left panel): For each query, only a limited number of memory chunks exhibited high similarity scores, resulting in a sparse and high-contrast distribution. This makes relevant memories stand out clearly, which is advantageous for analyzing selective recall behavior.

OpenAI Embedding (right panel): Each query yielded moderately high scores across a wide range of memory chunks, producing a more diffuse distribution with less distinct boundaries of relevance. While this characteristic is suited for comprehensive recall, it tends to blur contextual distinctions.

The primary objective of this study is to model the probabilistic recall and non-recall of memory in accordance with temporal and contextual factors, and to observe and analyze variability in this process. From this perspective, SBERT, which produces more

clearly differentiated similarity distributions, was judged to be more suitable as a component in the calculation of memory activation values.

Based on these findings, we adopted SBERT (`all-MiniLM-L6-v2`) as the embedding model in the implementation of this system.

4 Simulation Experiments

4.1 Objectives and Overview

This study aims to evaluate whether the proposed dialogue agent can replicate behaviors characteristic of human memory. Specifically, it focuses on phenomena such as the increased likelihood of certain topics becoming consolidated in memory over time, and the probabilistic fluctuation in recall—where the content remembered varies even when the same input is provided.

Based on the assumption that human memory is not a static repository of information but a dynamic process influenced by context and experience, this experiment centers on examining the dynamic aspects of memory behavior.

Beyond merely reproducing these phenomena, the study also seeks to establish design principles for future applications. To this end, we conduct a quantitative analysis of how various parameters within the ACT-R-based memory model affect memory recall performance.

4.2 Research Questions (RQ)

To achieve these objectives, the evaluation framework is structured around two main research questions (RQ). The first research question (RQ1) focuses on the core validity of the proposed model—namely, whether it can replicate "human-like" memory behaviors. In order to empirically examine this multifaceted concept of "human-likeness," RQ1 is further divided into the following two sub-questions:

- **RQ1-1: Memory Consolidation of Specific Topics** Does continued interaction selectively strengthen memory for specific topics, making them more likely to be recalled?
- **RQ1-2: Probabilistic Variation in Memory Recall** Does the content recalled vary probabilistically, depending on the current memory state and noise fluctuations, even for identical inputs?

Having validated the basic effectiveness of the model through these questions, the second research question (RQ2) addresses more practical concerns by aiming to clarify design guidelines for future implementation. Specifically, we quantitatively analyze how the contextual activation weight w affects the content and frequency of memory recall.

This two-layered approach allows us to first establish the model's ability to replicate key cognitive phenomena, and then explore its potential for practical application—thus demonstrating the research's value from both theoretical and applied perspectives.

4.3 Experiment Setup

To address the research questions outlined above, we designed and conducted a series of simulation experiments, described in detail below. All experiments in this study were conducted through simulations, and no data from human participants were used.

4.3.1 RQ1: Evaluating Human-like Memory Behavior. To examine RQ1-1 (memory consolidation) and RQ1-2 (recall fluctuation), we conducted a simulation replicating a long-term dialogue scenario.

- **Agent Initialization:** Five distinct topics were predefined—`engineer`, `doctor`, `activist`, `teacher`, and `artist`. For each topic, 10 memory chunks were prepared, resulting in a total of 50 chunks being embedded in the agent's initial memory state.
- **Simulation Procedure:** The agent was presented with queries exclusively related to a single target topic—`artist` in this case—at a frequency of 10 times per day over a simulated period of 30 days.
- **Data Logging:** Throughout the simulation period, we recorded the temporal evolution of the base-level activation $B(m)$ for each memory chunk, both in the target topic (`artist`) and non-target topic groups. These values were used for subsequent analysis.

4.3.2 RQ2: Evaluating the Effect of Parameter Design. To validate the design principles proposed in RQ2, we systematically varied key parameters of the memory model and analyzed their impact on recall behavior.

- **Fixed Decay Rate:** The temporal decay rate parameter (`decay`) was fixed at 0.5, a value widely accepted as the default in standard ACT-R settings.
- **Parameter Manipulation:** Under the above condition, the contextual activation weight (w) was varied from 0 to 25. The noise variance (σ) was held constant at 1.2, and all other settings were also kept fixed. Recall patterns were quantitatively evaluated for each parameter setting.

5 Results

5.1 Basic Recall and Forgetting Behavior

This section examines the core mechanisms of the proposed model —*selective recall* and *time-based forgetting*—through a minimal illustrative scenario.

The agent is initialized with two distinct memory chunks: (1) "Exhibited paintings at the art museum." (artist-related) and (2) "Conducted a code review in the office." (engineer-related). Both chunks are activated on Day 0. On Day 7, a single query—"Do you know anything about paintings?"—is presented, which is semantically relevant only to Chunk (1).

Figure 3 plots the base-level activation $B(m)$ of each chunk over time, revealing three distinct behavioral phases:

Initial Decay (Day 0–7). In the absence of further input, both memory chunks decay gradually, following ACT-R's time-based forgetting curve. This reflects natural decay of unused information.

Selective Recall (Day 7). The Day 7 query closely matches the artist memory. As a result, its total activation ($B(m) + S(m) + \text{noise}$) exceeds the retrieval threshold, triggering recall. This retrieval acts as a reinforcing experience, leading to a boost in $B(m)$. In contrast, the engineer memory remains unaccessed and continues to decay.

Post-Recall Divergence (After Day 7). The recalled artist memory now decays from a higher baseline, making it more resilient. Meanwhile, the engineer memory—never reactivated—continues to decline, becoming increasingly inaccessible.

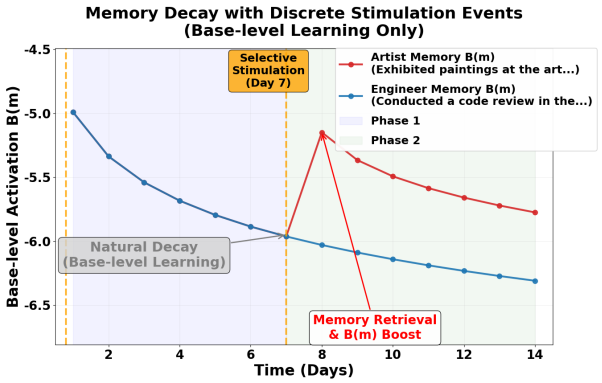


Figure 3: Selective recall and decay of memory chunks over time.

This scenario demonstrates that the model captures key human-like memory dynamics: relevant memories are selectively reinforced through use, while irrelevant ones fade over time. These dynamics form the basis for long-term topic-specific memory behaviors discussed in RQ1-1.

5.2 Evaluation of Human-like Memory Behavior

Building on the previous section, which confirmed that the model exhibits fundamental recall and forgetting behavior in response to a single stimulus, we now evaluate whether the model can replicate human-like memory characteristics as posed in RQ1—namely, topic-specific memory consolidation (RQ1-1) and probabilistic variability in recall (RQ1-2)—via long-term simulation.

5.2.1 RQ1-1: Topic-Specific Memory Consolidation. The results of the long-term simulation show that memories associated with a specific topic are selectively reinforced, while others decay. Figure 4 displays the average base-level activation $B(m)$ for each topic over a 30-day period.

As illustrated in the top-left graph, the *artist* topic (in purple), which was repeatedly queried, consistently maintained a high level of activation and diverged from the other topics. This is due to repeated recall boosting its activation, effectively overcoming temporal decay. In contrast, other topics—never retrieved—exhibited continuous decay and eventual forgetting.

We also performed statistical analysis on the final $B(m)$ values. The Shapiro–Wilk test indicated non-normal distributions for engineer ($W = 0.7075$, $p = .0011$), doctor ($W = 0.3657$, $p < .0001$), activist ($W = 0.3657$, $p < .0001$), and teacher ($W = 0.7872$, $p = .0101$), while the artist group satisfied the assumption of normality ($W = 0.8767$, $p = .1196$).

A Kruskal–Wallis test, used as a non-parametric alternative to one-way ANOVA, revealed a significant difference between groups ($H = 17.56$, $df = 4$, $p < .01$). Although most Steel–Dwass post-hoc comparisons showed no significant differences, the artist group exhibited the highest mean rank ($M_{\text{rank}} = 38.1$), and was significantly higher than both the doctor ($p < .05$) and activist ($p < .05$) groups.

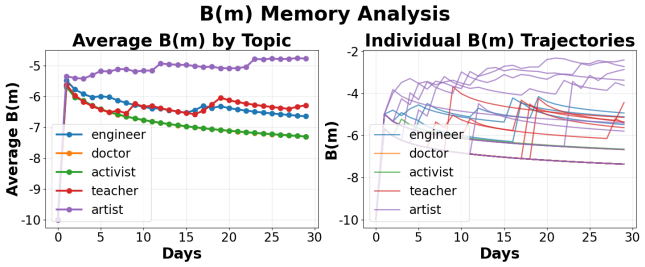


Figure 4: Memory activation changes by topic (left) and individual chunk (right) over time.

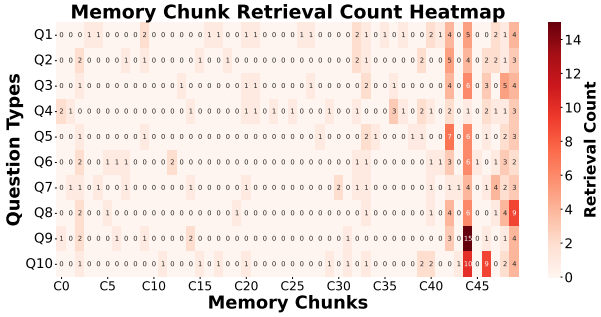


Figure 5: Probabilistic memory retrieval patterns across artist-related queries.

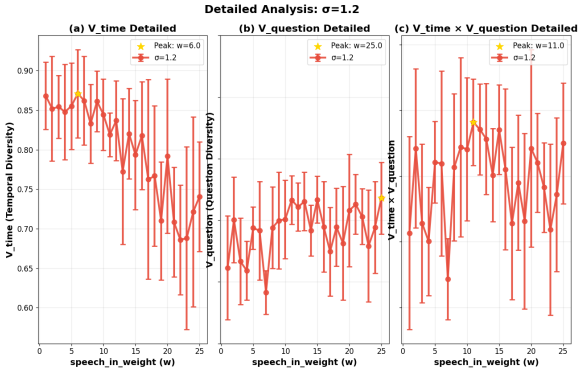


Figure 6: Effect of w on (a) V_{event}, (b) V_{question}, (c) $Score = V_{\text{event}} \cdot V_{\text{question}}$.

These visual and statistical results strongly support RQ1-1: that frequently referenced topics in dialogue become more consolidated and are more likely to be recalled.

5.2.2 RQ1-2: Probabilistic Fluctuation in Memory Recall. Next, we examined whether the proposed model could replicate the inherently variable nature of human memory—specifically, the phenomenon in which the recalled content fluctuates even in response to identical queries. To test this, we analyzed the recall history of the agent across a 30-day simulation period, focusing on which memory chunks were recalled in response to each of ten distinct "artist"-related queries.

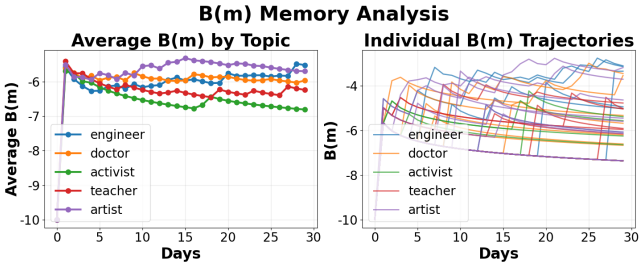


Figure 7: Average base-level activation by topic ($w = 7.0$).

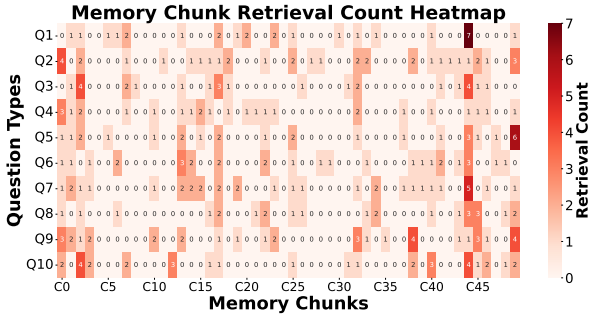


Figure 8: Recall heatmap over chunks and time ($w = 7.0$).

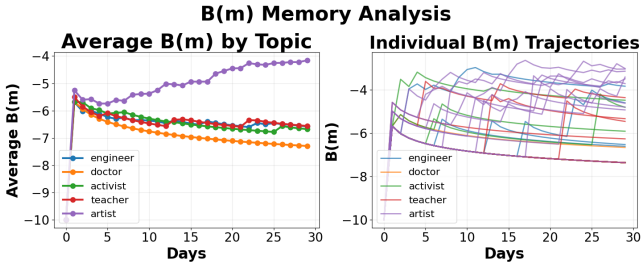


Figure 9: Average base-level activation by topic ($w = 25.0$).

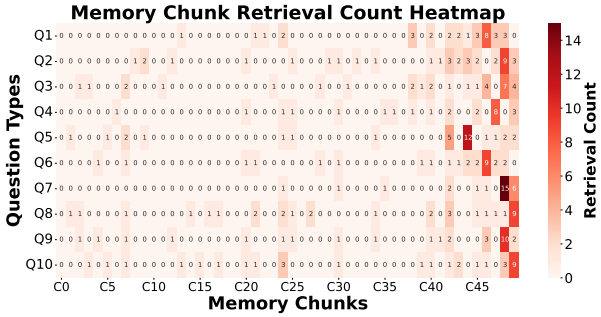


Figure 10: Recall heatmap over chunks and time ($w = 25.0$).

The results are presented as a heatmap in Figure 5. The vertical axis represents the ten queries, while the horizontal axis lists the ten memory chunks related to the "artist" topic. The intensity of each cell reflects the frequency (i.e., probability) with which a given chunk was recalled in response to a given query.

If recall were purely deterministic, each row would contain only a single dark cell indicating a unique response. However, as seen in the figure, each query does most frequently recall a specific, semantically related chunk (indicated by the darkest cell in each row), but other related chunks are also occasionally recalled with non-zero probability.

This fluctuation arises from the probabilistic noise term in the activation formula: $Activation = B(m) + S(m) + noise$. The presence of this noise introduces stochastic variation into the recall process, allowing for the possibility that, even when one chunk deterministically has the highest activation, another may occasionally exceed the retrieval threshold due to random fluctuation.

These results indicate that the model is capable of replicating the inherently probabilistic and contextually ambiguous nature of human memory recall, thereby supporting RQ1-2.

5.3 RQ2: Validity and Impact of Parameter Design

This section analyzes how the contextual relevance weight w affects memory retrieval dynamics, aiming to identify optimal design guidelines (RQ2).

5.3.1 Role of Parameter w and Basic Hypothesis. The weight parameter w (as defined in Eq. 3) controls the contribution of semantic similarity in total activation, balancing between past usage ($B(m)$) and current context. We hypothesize that varying w qualitatively changes recall behavior:

- **Low** w: Activation is dominated by $B(m)$; recall favors frequently accessed chunks and ignores the query context.
- **High** w: Activation is dominated by $S(m)$; recall becomes highly context-sensitive but may neglect stable past knowledge.

These tendencies suggest a trade-off between *memory stability* and *context sensitivity*, with w as a key control parameter.

5.3.2 Definition of Evaluation Metrics. To evaluate this trade-off, we define two diversity metrics:

Temporal Diversity (V_{event}): Measures how many new chunks are recalled daily.

$$V_{\text{event}} = \frac{1}{D-1} \sum_{i=2}^{D} |M_i \setminus M_{i-1}| \,. \tag{4}$$

Response Diversity (V_{question}): Measures how evenly different chunks are recalled per query using normalized entropy.

$$V_{\text{question}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{H(q)}{\log_2 |U_q|}, \quad H(q) = -\sum_{c \in U_q} p_c \log_2 p_c. \tag{5}$$

We define the overall score as $Score = V_{\text{event}} \cdot V_{\text{question}}$ and consider the w that maximizes it as optimal.

5.3.3 Trade-off in Recall Diversity Induced by Parameter w. The weight parameter w plays a crucial role in balancing the influence of past memory usage $B(m)$ and contextual relevance $S(m)$. Figure 6 shows how the diversity metrics behave under different values of w (with noise strength $\sigma = 1.2$ held constant).

Figure 6(a) shows that V_{event} peaks at low w and declines as w increases, reflecting reduced stochasticity and more fixed recall. Figure 6(b) shows the opposite trend for V_{question}, which increases with w due to stronger context sensitivity. The integrated score in Figure 6(c) peaks around $w = 11.0$, indicating a well-balanced regime.

5.3.4 Qualitative Differences in Recall Patterns by w. To illustrate this trade-off, we compare two conditions:

$w = 7.0$: As shown in Figure 7, base-level activations are spread across topics; Figure 8 shows dispersed recall patterns with no strong topic preference.

$w = 25.0$: Figure 9 shows high activation concentrated in the *artist* topic; Figure 10 shows deterministic recall centered on chunks c40–c49.

These results confirm that w qualitatively alters recall behavior. Too small w leads to rigid memory; too large w causes over-reactive recall. The proposed score identifies $w = 11.0$ as a balanced setting, offering useful guidance for designing agents with human-like memory traits and personality.

6 Discussion and Conclusion

This study examined the feasibility of a conversational agent that exhibits human-like memory behavior by integrating the ACT-R memory model with a large language model (LLM). This section summarizes the key findings, academic contributions, limitations, and future directions of the research.

6.1 Key Findings

The experimental results confirmed that the proposed model successfully reproduced core characteristics of human memory:

- **Memory Consolidation (RQ1-1):** Repeated references to a specific topic selectively maintained and reinforced its associated memories, while others naturally faded over time.
- **Fluctuation in Recall (RQ1-2):** The same question yielded different recalled content across instances, with semantically related memory chunks retrieved probabilistically.
- **Effect of Parameter w (RQ2):** The contextual weight w controlled the trade-off between temporal stability and contextual sensitivity, revealing an optimal balance that shapes the model's memory behavior.

6.2 Contributions and Novelty

This study moves beyond static information retrieval to model the human-like process of *remembering*.

Unlike conventional methods such as Retrieval-Augmented Generation (RAG), which avoid forgetting by relying on static knowledge bases, our approach integrates the ACT-R framework to treat *forgetting* as a functional and adaptive feature. The agent dynamically adjusts memory importance based on conversational context, learning what to retain or let go.

This fusion of cognitive memory modeling with LLMs introduces a dynamic memory structure where forgetting plays a key role, enhancing human-AI alignment in long-term interaction.

6.3 Limitations

Despite its promising results, this study has several limitations:

- The validation was conducted entirely through simulations. Evaluation based on real human dialogue data has not yet been performed. Given the complexity and unpredictability of real-world interactions, further testing is required.
- Memory chunks were treated as independent sentences. The model does not yet support more complex episodic memory or hierarchical/semantic relationships between memories.
- The current activation function considers only usage frequency, time decay, and contextual similarity. Human memory is also affected by emotional salience and subjective importance, which are not modeled here.

6.4 Future Work

Based on these limitations, we propose the following directions for future research:

- **Long-term evaluation in real-world settings:** Deploy the model in actual environments and conduct longitudinal evaluations of subjective impressions such as naturalness, consistency, and perceived "personality."
- **Refinement of memory structure:** Extend the model to incorporate graph-based representations of relationships among memory chunks, enabling more complex associations and inference.
- **Incorporation of psychological factors:** Introduce emotional salience and importance scores into the activation calculation to better reflect human memory dynamics.
- **Adaptive parameter control:** Implement meta-learning mechanisms to dynamically adjust parameters such as w based on dialogue context or user behavior.

This study presents a computational foundation for memory modeling aimed at fostering deeper and more natural relationships between humans and AI. We hope that by addressing these future directions, conversational agents can evolve beyond mere information tools into true dialogue partners.

References

[1] John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. An Integrated Theory of the Mind. *Psychological Review* 111, 4 (2004), 1036–1060. doi:10.1037/0033-295X.111.4.1036

[2] Michael C. Anderson and Justin C. Hulbert. 2021. Active Forgetting: Adaptation of Memory by Prefrontal Control. *Annual Review of Psychology* 72 (2021), 1–36. doi:10.1146/annurev-psych-072720-094140 Epub 2020 Sep 14.

[3] Jerry Ball. 2004. A Cognitively Plausible Model of Language Comprehension. (05 2004).

[4] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. 2024. Titans: Learning to Memorize at Test Time. arXiv:2501.00663 [cs.CL] https://arxiv.org/abs/2501.00663 Introduces neural long-term memory module parallel to attention.

[5] Derbinsky2012. 2012. Computationally Efficient Forgetting via Base-Level Activation. *Proceedings of the 11th International Conference on Cognitive Modeling, ICCM 2012* (01 2012).

[6] Ralph Hertwig and Peter Todd. 2005. More Is Not Always Better: The Benefits of Cognitive Limits. *Thinking : Psychological perspectives on reasoning, judgment and decision making, 213-231 (2003)* (01 2005). doi:10.1002/047001332X.ch11

[7] Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 7, 7 pages. doi:10.1145/3613905.3650839

[8] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, Tiago Quan, Tiago Kumar, Raia Hadsell, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (2017), 3521–3526. doi:10.1073/pnas.1611835114

[9] Richard L. Lewis and Shravan Vasishth. 2005. An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science* 29, 3 (2005), 375–419. doi:10.1207/s15516709cog0000_25

[10] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized Experience Replay. In *Proceedings of ICLR 2016.* https://arxiv.org/abs/1511.05952 Originally arXiv preprint 1511.05952.

[11] Thomas Sievers and Nele Russwinkel. 2025. Retrieving Memory Content from a Cognitive Architecture by Impressions from Language Models for Use in a Social Robot. *Applied Sciences* 15, 10 (2025). doi:10.3390/app15105778

[12] Ryuichi Sumida, Koji Inoue, and Tatsuya Kawahara. 2024. Should RAG Chatbots Forget Unimportant Conversations? Exploring Importance and Forgetting with Psychological Insights. arXiv preprint. arXiv:2409.12524 [cs.CL] https://arxiv.org/abs/2409.12524 Proposes LUFY: focusing on emotional arousal and surprise to drive selective memory retention.

[13] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting Language Models with Long-Term Memory. arXiv:2306.07174 [cs.CL] https://arxiv.org/abs/2306.07174

[14] Siyu Wu, Alessandro Oltramari, Jonathan Francis, C. Lee Giles, and Frank E. Ritter. 2024. Cognitive LLMs: Towards Integrating Cognitive Architectures and Large Language Models for Manufacturing Decision-making. arXiv preprint. arXiv:2408.09176 [cs.AI] Introduces the LLM-ACTR framework integrating ACT-R and LLMs.

[15] Siyu Wu, Rafael F. Souza, Frank E. Ritter, and Wilson T. Lima Jr. 2024. Comparing LLMs for Prompt-Enhanced ACT-R and Soar Model Development: A Case Study in Cognitive Simulation. *Proceedings of the AAAI Symposium Series* 2, 1 (2024), 422–427. doi:10.1609/aaaiss.v2i1.27710