

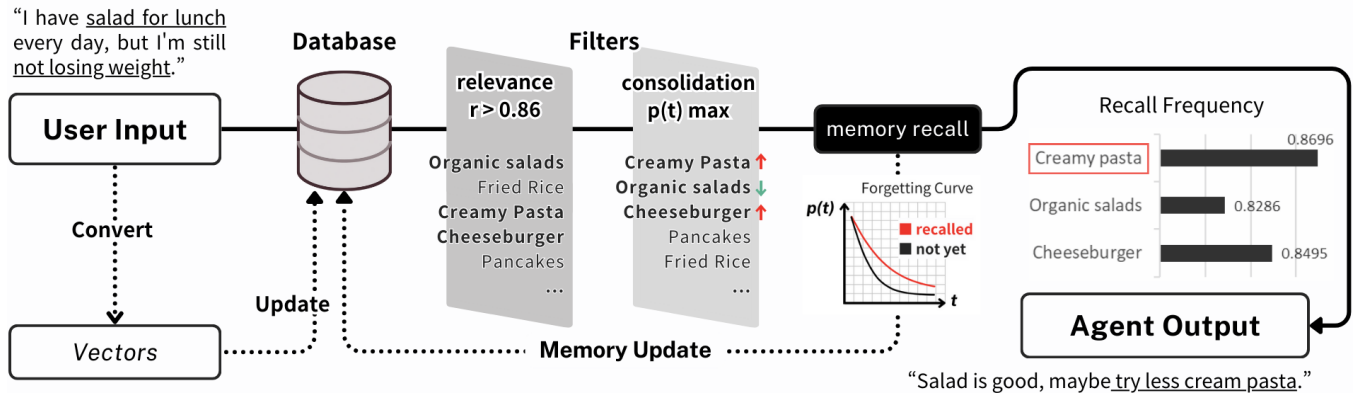


# "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents

Yuki Hou  
houhoutime@gmail.com  
Meiji University  
Tokyo, Japan

Haruki Tamoto  
harukiririwuru@gmail.com  
Kyoto University  
Kyoto, Japan

Homei Miyashita  
homei@homei.com  
Meiji University  
Tokyo, Japan



**Figure 1: Architecture of the enhanced large language model (LLM)-based dialogue agent that integrates human-like memory processes.** First, the user input is converted into vectorized text and processed through a data-filtering process based on relevance and memory consolidation bias, modeled after human cognitive functions. Then, memory recall is triggered when the recall probability, informed by relevance and elapsed time, exceeds a predefined threshold. This diagram features an agent output example where the system recalls "Creamy pasta" as the user's lunch preference with a higher frequency, influencing the agent's response.

## ABSTRACT

In this study, we propose a novel human-like memory architecture designed for enhancing the cognitive abilities of large language model (LLM)-based dialogue agents. Our proposed architecture enables agents to autonomously recall memories necessary for response generation, effectively addressing a limitation in the temporal cognition of LLMs. We adopt the human memory cue recall as a trigger for accurate and efficient memory recall. Moreover, we developed a mathematical model that dynamically quantifies memory consolidation, considering factors such as contextual relevance, elapsed time, and recall frequency. The agent stores memories retrieved from the user's interaction history in a database that encapsulates each memory's content and temporal context. Thus, this strategic storage allows agents to recall specific memories and understand their significance to the user in a temporal context, similar to how humans recognize and recall past experiences.

## CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; *Intelligent agents*.

## KEYWORDS

Memory Retrieval Models, Large Language Models, User Experience, User Interface, Intelligent Agents

## ACM Reference Format:

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 (CHI EA '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3613905.3650839>

## 1 INTRODUCTION

The emergence of transformer-based language models [14] have drastically revolutionized the field of natural language processing, surpassing the capabilities of traditional models in understanding and generating human-like text [22]. In particular, large language models (LLMs) [5] have garnered considerable attention for their prowess in mimicking artificial intelligence (AI) with human-like

cognition and conversational abilities, reminiscent of sentient machines portrayed in science fiction narratives. However, LLMs exhibit a significant limitation in processing temporal information inherent to human cognition. While transformers possess excellent self-attention mechanisms, outperforming recurrent neural networks (RNNs) [15] and long short-term memory models (LSTM) [23], they fail to replicate human behavioral dynamics. To accurately replicate the nuanced human-like interactions of AI agents, as depicted in science fiction, one must first achieve human-like cognitive and memory processing abilities. Therefore, we proposed an approach to integrate human memory processes into LLM-based dialogue agents. We adopted human-like cued recall as the trigger for accurate and efficient memory retrieval [16]. This mechanism involves an agent autonomously recalling memories essential for generating responses during a conversation. The process emulates the human memory process known as "remember to remember" [9], consciously retaining memory for future action or task and recalling that when needed [13]. Furthermore, the proposed model replicates human cognitive ability, where memories recalled repeatedly over a long period are retained more strongly than those recalled over a short period and relatively frequently [21], regardless of recall frequency. Thus, our model provides contextually relevant and coherent conversations.

Furthermore, our primary purpose is to transcend the paradigm of dialogue agents merely imitating human behavior through statistical natural language models. Instead, we seek to create agents that are capable of truly understanding human language with rich nuances, achieved by seamlessly integrating human cognitive processes. This fusion aligns with the philosophy of human-computer interaction, promoting more natural and intuitive human-centered interactions between the two at cognitive and emotional levels.

## 2 RELATED WORK

### 2.1 Similarities Between LLMs and Human Memory

Human memory serves as a system to encode, store, and retrieve our experiences [25]. Our memories can be categorized into declarative and non-declarative memories, with declarative memory further divided into episodic and semantic memories [3]. Episodic memory [24] consciously allows for recollecting and re-experiencing one's subjective past. In contrast, semantic memory supports language use, registering not the perceptual properties of inputs but the cognitive referents of input signals [27].

Similar to human episodic memory functioning, the episodic nature of LLMs' is demonstrated by their ability to recall specific events or dialogues from the database. This allows LLMs to generate responses based on past interactions and experiences to inform current interactions. LLMs also possess a human-like semantic understanding of language that captures the meaning and context behind the words. Geva et al. [8] suggested the feed-forward layers of transformer-based models to operate in a key-value format, the same as human semantic memory.

### 2.2 Human-like Memory Processes in AI Agent

Kim et al. [11] focused on emulating human episodic and semantic memory processes in AI agents to enhance interactive experiences.

They compared agents with different memory processes: episodic only, semantic only, and both. These agents used different strategies to decide which memories to forget when memory was full and which to use when answering questions. The agents with a composite memory system outperformed those with a single memory system, especially those with pre-trained semantic memory. Zhong et al. developed MemoryBank, a memory retrieval mechanism for memory storage [28]. The system uses an encoder model to encode each conversation turn and event summary into a vector representation, allowing recalling memory with the highest relevance whenever needed. The memory strength of MemoryBank's is enhanced by 1 each time a memory piece is recalled, simulating more human-like memory behavior and reducing the probability of forgetting the memory by setting the elapsed time to zero.

In contrast, we designed our architecture without the concept of "complete forgetting." Even if not recalling a memory over an extended period, the degree of consolidation never reaches absolute zero. Thus, given the right trigger, these memories can be recalled [1]. The process is consistent with that of human memory, where past experiences are never completely forgotten and can be retrieved with specific stimuli, such as the scent of a familiar perfume or the melody of a once-favorite song.

### 2.3 Mathematical Models of Human Memory Processes

This section reviews the mathematical models that attempt to quantify and simulate human memory processes, primarily for memory recall. Based on Zielske's [29] recall probability function, Chess et al. [4] proposed a model that assumes the rate of memory consolidation  $r(t)$  to express the probability  $p(t)$  of a human memory being recalled as follows:

$$p(t) = 1 - \sum_{n=1}^{b-1} \frac{(r(t))^n}{n!} \exp(-r(t)) \quad (1)$$

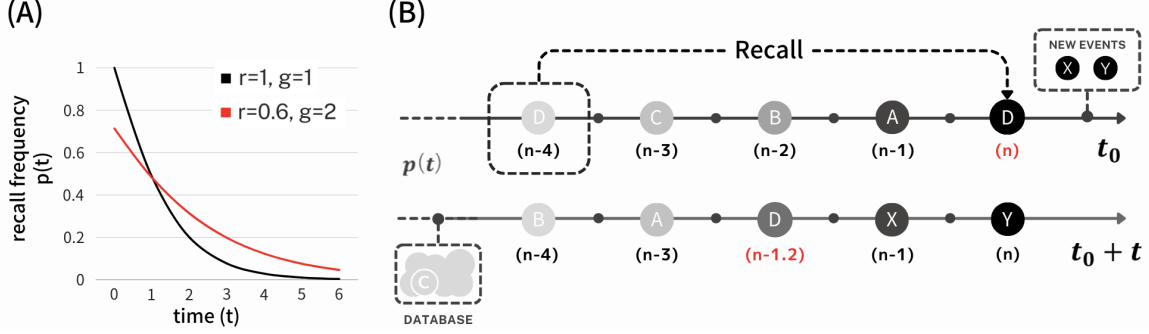
This model is based on the hypothesis that each neuron fires independently and at random [10], and is derived from the properties of a non-homogeneous Poisson process using a time-varying intensity function  $r(t)$  [12]. The model also considers a stimuli threshold  $b$  required for a recall. The following exponential function  $r(t)$  represents the adjustment process of memory strength [2] in the human hippocampus:

$$r(t) = \mu e^{-at} \quad (2)$$

where  $\mu$  is the memory strength,  $a$  is the decay rate, and  $t$  is the elapsed time. In implementations using vector databases, only a single data is required for recall; therefore, we consider only the case of  $b = 1$ . The recall probability  $p(t)$  in this special case is expressed as

$$p(t) = 1 - \exp(-\mu e^{-at}) \quad (3)$$

The recall probability  $p(t)$  exponentially decays with time  $t$ , as demonstrated in short-term memory decline using the classic Brown-Peterson learning and distraction task [19]. However, this model considers only one trial learning and a constant decay rate. However, in reality, the degree of consolidation differs between memories recalled many times and those not; hence, the decay rate should be adjusted to reflect this effect.



**Figure 2: (A) Decline in recall probability over time. The black curve ( $r=1, g=1$ ) shows a rapid loss of recall, while the red curve ( $r=0.6, g=2$ ) represents a slower forgetting rate. This difference indicates the challenge in designing dialogue agents that must distinguish between recent and distant events. (B) At time  $t_0$ , Event D is recalled by the user, and the model updates its temporal significance. This exemplifies how memory is reinforced through repetition, becoming less susceptible to forgetting at  $t_0 + t$ .**

## 2.4 LLM-based Autonomous Agents

Park et al. introduced the concept of Generative Agents, outlining a memory mechanism of agents based on a scoring system comprising three elements: recency, importance, and relevance [18]. This approach dictates that agents consider recent actions or events (recency), objects deemed important by the agent (importance), and objects relevant to the current situation (relevance) to make decisions. These elements are normalized leveraging min-max scaling and combined through a weighted sum to determine the final score. In contrast, the proposed model employs elapsed time, relevance, and recall frequency to calculate the degree of memory consolidation. Thus, the agent can recall the most appropriate memory, facilitating efficient dialogue. While the Generative Agents and our proposed model share commonalities in memory processing, they apply memory in different contexts and for different purposes. Generative Agents focus on independently scoring each memory element to select actions most fitting to the current context. In contrast, our approach adjusts memory consolidation over time, enabling memory consistency.

## 3 ARCHITECTURE

### 3.1 Model

We constructed the model based on exponential decay, taking event relevance ( $r$ ) and elapsed time ( $t$ ) as variables. Adapting (3) from [4], the recall-probability function  $p(t)$  is expressed as

$$p(t) = 1 - \exp(-re^{-at}) \quad (4)$$

The relevance is quantified by the cosine similarity between vectorized texts, defining the closeness of information. The cosine similarity between  $n$ -dimensional vectors  $\mathbf{a}$  and  $\mathbf{b}$  is defined as:

$$r = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (5)$$

Furthermore, we considered the impact of increased recall intervals and frequency to model the variation in memory consolidation due to multiple recalls. The decay constant  $a$  considering the number of recalls  $n$  is defined as

$$a = \frac{1}{g_n}, \quad g_0 = 1 \quad (6)$$

$$g_n = g_{n-1} + S(t), \quad S(t) = \frac{1 - e^{-t}}{1 + e^{-t}} \quad (7)$$

The modified sigmoid function  $S(t)$  represents memory consolidation with each recall and increases monotonically for  $t > 0$ . However, the reduction in  $a$  per recall is capped, reflecting long-term memory consolidation. As  $n$  increases, the rate of reduction in  $a$  decreases, emulating the natural human memory process where frequent recalls strengthen consolidation. Figure 2-A illustrates how the recall probability  $p(t)$  decays over time with changes in  $r$  and the decay rate  $1/g$ . As  $g$  increases, the slope of  $p(t)$  becomes less steep, indicating reduced probability of forgetting memories with more recalls (high  $g$ ).

After normalizing the recall probability  $p_n(t)$  such that it equals 1 for  $r = 1$  and  $t = 0$ , we obtained the final equation:

$$p_n(t) = \frac{1 - \exp(-re^{-t/g_n})}{1 - e^{-1}} \quad (8)$$

$$g_n = g_{n-1} + \frac{1 - e^{-t}}{1 + e^{-t}} \quad (9)$$

Utilizing equation (8), we set a trigger for recall when  $p(t)$  exceeds a certain threshold  $k$ . Trials suggest a threshold of 0.86 as appropriate to reflect the relevance of the event and the time elapsed. Further research will determine the most effective trigger threshold, identifying an appropriate value based on theoretical justification.

### 3.2 Memory Recall and Consolidation in Database Architecture

Figure 2-B illustrates the retrieval and consolidation of memories and highlights how our system replicates human-like memory retention. For instance, a memory like Event D, even if recalled less frequently over several years, is retained more robustly in the system compared to a memory recalled several times in quick succession but over a shorter time frame [21]. This is depicted through the visualization of memory events along the time axis, where the color intensity represents the rate of memory consolidation and the

strength of memory retention over time. Darker shades, therefore, signify a more profound and enduring memory consolidation, a direct result of our system’s unique ability to emulate human-like memory patterns. By storing episodic memories derived from user dialogues, the database structure encapsulates the content and temporal context of each memory. This approach enables our agent not just to recall specific information but also to understand and interpret the significance of these memories in a temporal context, similar to how humans perceive and recall past experiences. Using key-value pairs for encoding semantic structures further enhances the agent’s ability to efficiently retrieve and apply these memories in ongoing interactions, thereby fostering a more human-like and context-aware dialogue experience.

## 4 EXPERIMENT

### 4.1 Setup

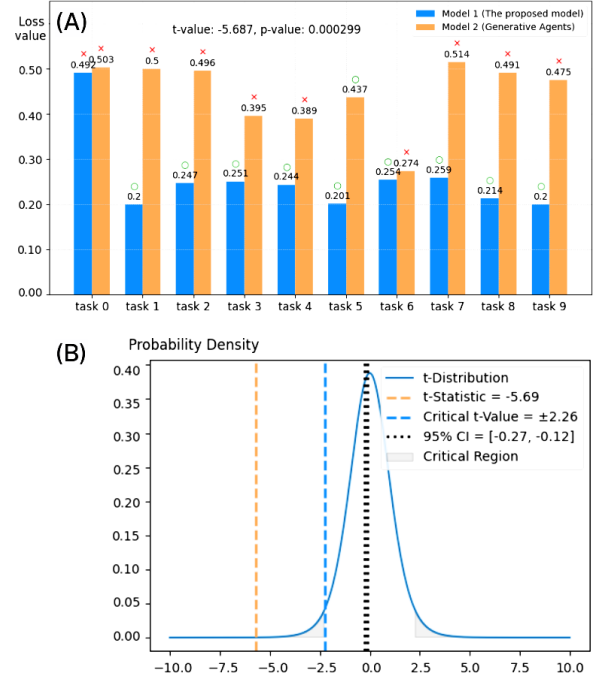
We developed the experimental system in Python[26], using GPT-4-0613 [6] as the baseline model of the agent. We adopted Qdrant [20] as the ‘memory retrieval trigger’ for the vector search engine. It identifies relevant past information in the context of a dialogue, which triggers memory recall. Moreover, we built a ChatHistory module to manage chat history in the Firestore [7] database, allowing agents to reference past dialogues to generate chat events. An EventHandler module was adopted to search and pass the recalled events to the agent’s prompt. Details on LLM interaction and system prompts are shown in Appendix (B).

To quantitatively evaluate the performance of our proposed model against that of Generative Agents [18], which adopted a similar approach in calculating the recall score. We constructed a dataset containing 10 tasks, each derived from actual conversational histories generated by our system. These tasks encapsulate diverse user interactions, ensuring unbiased and objective assessments. The dataset includes a series of events, each tagged with relevant topics and keywords, providing a detailed memory for the agent to reference. We also adopt a timeline structure that stores the time/date of tasks containing four types of events and defines the event with the highest probability as the correct event to recall. Events in the dataset were selected neutrally, avoiding any potential bias that could skew the results. Each task represents a unique conversational scenario, where the dialogue agent’s ability to recall and utilize context is critical. The task variation allows for a comprehensive evaluation of the model’s performance across different contexts.

In addition, we selected six participants to partake in a dialogue task with agents developed by the proposed model to evaluate recall accuracy qualitatively. The participants engaged in daily conversations over one week to three months, discussing personal habits, preferences, and life events at a time of their choice. Respecting individual privacy, our analysis relied solely on non-textual output logs, which included updated parameter values for each chat event.

### 4.2 Analysis

**4.2.1 Memory Recall Accuracy.** Our model demonstrated a statistically significant lower loss value across various tasks when compared to the Generative Agent model, as indicated by  $t=-5.687$



**Figure 3: (A) Comparison of Loss Values Between two Models Across Different Tasks (B) Verification of Significance of Results**

and  $p=0.000299$  (Figure 3-A). These values suggest a high confidence level in performance superiority, meaning that our model significantly outperforms in terms of recall accuracy in cognitive tasks involving time series data. Furthermore, the critical t-value for our two-tailed test was set at  $\pm 2.26$ , with the 95% confidence interval for the mean difference falling between  $[-0.27, -0.12]$  (Figure 3-B). This interval is completely below zero, indicating that the difference in mean performance is statistically significant and favorable to our proposed model. Normalization and scaling techniques were employed to ensure an unbiased comparison of loss values across models. The Softmax function was utilized to convert the raw scores into probabilities, enabling a more interpretable comparison of the models’ performance. The sum of squares error method was applied to compute the loss, providing a consistent metric for evaluating recall accuracy across the dataset.

On the other hand, Table 1 shows a failed task where both models incorrectly answered. The ‘Score’ columns represent the recall probability calculated by each model using different methods. For the proposed model (Model 1), the score is based on the relevance and elapsed time of the events, as described in Section 3. Generative Agents (Model 2) calculates the score using recency, importance, and relevance of the events, as described in Section 2.4. By analyzing the recall frequency and gradient of incorrectly answered events, we find that although event B is recalled most frequently, its gradient is not as large as events A and C. This indicates that the proposed model associates the length of the recall interval with memory strength, rating Event A as strongly retained due to its high relevance and long recall intervals. In contrast, Generative Agents

**Table 1: The Failed Task 0 with Both Models**

Model 1	Relevance	Time (s)	Grad	Score
<b>A University</b> ×	0.776	434700	5.102	<b>0.850</b>
B Home ○	0.745	148800	5.229	0.830
C Library	0.757	331500	5.028	0.836
D Restaurant	0.756	55800	1.000	0.836

Model 2	Relevance	Time (s)	Importance	Score
A University	0.776	434700	7	1.489
B Home ○	0.745	148800	2	1.130
C Library	0.757	331500	5	1.292
<b>D Restaurant</b> ×	0.756	55800	5	<b>1.620</b>

prioritizes recency and relevance over recall frequency, leading to its preference for Event D.

Refer to Appendix (A) for detailed calculation of the loss function and specifics of the failed task 0 in Appendix (C).

**4.2.2 Response Generation Performance.** The autonomous memory recall and consolidation capabilities of the proposed model were demonstrated through its ability to recall specific memories and reflect users' preferences and habits in the generated responses. Two typical examples were obtained from the chat history with the participants' approval.

**Table 2: Understanding User Preferences Through Memory Consolidation**

**User:** Thanks. Almost done, just pushing through!  
**Agent:** Ice cream awaits as a reward for you, almost there!

Event				
1	User has finished work and decided to have an ice cream.			
2	User has a date this evening so he has to finish work early.			
3	User is on the way home thinking about the dinner menu.			
..	Time (s)	Relevance	Score	Status (Grad)
1	3037912	0.837	0.772	True (1.998)
2	3232645	0.831	0.726	False (1.007)
3	2590652	0.812	0.737	False (1.000)

In Table 2, the agent successfully recalled that the user was expecting ice cream as a reward after work. Depending on the recall probability, this recall was accurately relevant and provided in a timely manner, consistent with the user's statement about nearing task completion. The agent's response demonstrates the model's ability to add personalized nuance to generate an encouraging response, subtly reminding the user of the long-awaited treat.

Another case that showcases an interesting dialogue is presented in Table 3. In this case, the user expresses a desire to escape to Hawaii from a demanding schedule. The agent responds cleverly, suggesting that the user should focus on the impending paper deadline instead.

**Table 3: Human-like Temporal Cognition in Processing Similar Memories**

**User:** I need a vacation in Hawaii!  
**Agent:** Why not? Let's say Aloha to the sun and sea of Hawaii and to the paper we're writing!

Event				
1	User has a paper deadline for a conference in Hawaii on 1/25.			
2	User has been to Hawaii last summer vacation.			
3	User was visiting Hokkaido for the first time on business.			
..	Time (s)	Relevance	Score	Status (Grad)
1	34854	0.846	0.903	True (1.085)
2	33749	0.831	0.847	False (1.003)
3	33763	0.823	0.841	False (1.000)

This dialogue further emphasizes the proposed model's strength in dealing with similar long-term memories and retrieving relevant information to construct a coherent and engaging narrative. The agent's response indicates an understanding of the user's current mood. Adaptability to various personalities and interaction styles demonstrates the model's potential to support more natural and dynamic human-like dialogues.

In addition, it is interesting to note that the agent's response in the second dialogue was characterized by a sarcastic tone, which was a direct result of the agent's personality "sarcastic" and the unique prompts added by the participant. The conversation history shows that the same memory could be used differently depending on the agent's perceived personality and the user's interaction style. Future research will explore the extent to which the personality characteristics of the model can be customized and how they affect memory recall and interaction patterns.

## 5 CONCLUSION

The proposed model demonstrates significant improvements in memory recall and response generation for LLM-based dialogue agents. One of the key advantages of the proposed model is its ability to manage the prompt length effectively. In the proposed model, only one past dialogue history obtained through search is added to the prompt, thus avoiding the impact of increasing prompt length seen in systems like ChatGPT [17].

Nevertheless, a major limitation of the proposed method is its reliance on users' long-term behavioral patterns for calculating memory consolidation. In cases where a user's behavior undergoes significant changes (e.g., starting a new job or school, lifestyle changes), the method's adaptability may be limited. Future work could explore incorporating mechanisms to detect shifts in user behavior and adjust the memory consolidation calculation accordingly. Neural networks could potentially alter these functions and improve accuracy when trained on larger datasets with more variables. To further enhance the model's performance, a large-scale and high-quality dataset is necessary. While the proposed method's interaction with the database enables the generation of context-aware and personalized responses, the implications on storage resources

and computational overhead due to these interactions remain to be explored in future research. As the primary focus of this study was on the development and evaluation of a novel architecture for human-like memory recall and consolidation, a detailed analysis of the system's resource requirements and optimization strategies falls outside the scope of the current work.

We hope this work contributes to advancing further research in human-computer interactions, paving the way for a future where technology aligns with human needs and resonates with human cognition and experience. This vision echoes the partnerships depicted in science fiction, representing a significant step towards building a "buddy" relationship between humans and agents. As technology continues to evolve, agents will eventually become a part of users' daily life, and potentially "understand you better than you understand yourself" in the near future.

## REFERENCES

- [1] Hafeez Ullah Amin and Amir Malik. 2014. *Memory Retention and Recall Process*. 219–237. <https://doi.org/10.1201/b17605-11>
- [2] Neil Burgess, Eleanor A Maguire, and John O'Keefe. 2002. The human hippocampus and spatial and episodic memory. *Neuron* 35, 4 (2002), 625–641.
- [3] S.D.L.R.S.P.P.U. California. 1987. *Memory and Brain*. Oxford University Press, USA. <https://books.google.co.jp/books?id=WH-HF5E9XSsC>
- [4] Antonio Chessa and Jaap Murre. 2007. A Neurocognitive Model of Advertisement Content and Brand Name Recall. *Marketing Science* 26 (01 2007), 130–141. <https://doi.org/10.1287/mksc.1060.0212>
- [5] Xuan-Quy Dao. 2023. Performance comparison of large language models on vnhsge english dataset: Openai chatgpt, microsoft bing chat, and google bard. *arXiv preprint arXiv:2307.02288* (2023).
- [6] OpenAI et al. 2023. GPT-4 Technical Report. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774) [cs.CL]
- [7] Firebase. 2023. Firestore. <https://firebase.google.com/docs/firestore?hl=ja>. (Accessed on 01/18/2024).
- [8] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. [arXiv:2012.14913](https://arxiv.org/abs/2012.14913) [cs.CL]
- [9] H Hécaen, G Gosnave, C Vedrenne, and G Szikla. 1978. Suppression laterale du matériel verbal présente dichotiquement lors d'une destruction partielle du corps calleux. *Neuropsychologia* 16, 2 (1978), 233–237.
- [10] Anthony Holtmaat and Pico Caroni. 2016. Functional and structural underpinnings of neuronal assembly formation in learning. *Nature neuroscience* 19, 12 (2016), 1553–1562.
- [11] Taewoon Kim, Michael Cochez, Vincent Francois-Lavet, Mark Neerinx, and Piek Vossen. 2022. A Machine With Human-Like Memory Systems. [arXiv:2204.01611](https://arxiv.org/abs/2204.01611) [cs.AI]
- [12] J. F. C. Kingman. 1993. *Poisson Processes*. Oxford University Press.
- [13] Beatrice G Kuhlmann. 2019. Metacognition of prospective memory: Will I remember to remember? *Prospective memory* (2019), 60–77.
- [14] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open* (2022).
- [15] Danilo P Mandic and Jonathon Chambers. 2001. *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. John Wiley & Sons, Inc.
- [16] Mark A McDaniel, Michael D Kowitz, and Paul K Dunay. 1989. Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition* 17, 4 (1989), 423–434.
- [17] OpenAI. 2023. ChatGPT. <https://chat.openai.com/>. (November 22 version) [Large language model].
- [18] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. [arXiv:2304.03442](https://arxiv.org/abs/2304.03442) [cs.HC]
- [19] Lloyd Peterson and Margaret Jean Peterson. 1959. Short-Term Retention of Individual Verbal Items. *Journal of Experimental Psychology* 58, 3 (1959), 193. <https://doi.org/10.1037/h0049234>
- [20] Qdrant. 2023. Vector Database. <https://qdrant.tech/>. (Accessed on 01/17/2024).
- [21] Henry Roediger and Jeffrey Karpicke. 2006. Test-Enhanced Learning Taking Memory Tests Improves Long-Term Retention. *Psychological science* 17 (04 2006), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- [22] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings* 18. Springer, 194–206.
- [23] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.
- [24] Endel Tulving. 2002. Episodic Memory: From Mind to Brain. *Annual Review of Psychology* 53, 1 (2002), 1–25. <https://doi.org/10.1146/annurev.psych.53.100901.135114>
- [25] Endel Tulving et al. 1972. Episodic and semantic memory. *Organization of memory* 1, 381–403 (1972), 1.
- [26] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.
- [27] Atsushi Yamadori. 2002. *Frontiers of Human Memory: a collection of contributions based on lectures presented at International Symposium, Sendai, Japan, October 25–27, 2001*. Tohoku University Press. <https://ci.nii.ac.jp/ncid/BA57511014>
- [28] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. MemoryBank: Enhancing Large Language Models with Long-Term Memory. [arXiv:2305.10250](https://arxiv.org/abs/2305.10250) [cs.CL]
- [29] Hubert A. Zielske. 1959. The Remembering and Forgetting of Advertising. *Journal of Marketing* 23 (1959), 239 – 243. <https://api.semanticscholar.org/CorpusID:167354194>

## A DETAILED CALCULATION OF THE LOSS FUNCTION

To quantify the performance of our model, we define a matrix containing the scores calculated by each model for  $d$  tasks as follows:

$$S = (s_1 \quad s_2 \quad \dots \quad s_d)^T \in \mathbb{R}^d \quad (10)$$

In order to standardize the scale of scores across different models, we normalize the scores to a  $[0, 1]$  range:

$$S' = \frac{S - \min(S)}{\max(S) - \min(S)} \quad (11)$$

Subsequently, we convert each score into a probability value by applying the Softmax function:

$$S'' = \frac{\exp(S')}{\sum_{j=1}^d \exp(s'_j)} \quad (12)$$

We then define a matrix with one-hot encoded true labels for the evaluation tasks:

$$T = (t_1 \quad t_2 \quad \dots \quad t_d)^T \in \mathbb{R}^d, \quad \text{where } t_j = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

Finally, the loss value is calculated as the mean squared error between the predicted probabilities and the true labels:

$$l = \frac{1}{2} \sum_{j=1}^d (s''_j - t_j)^2 \quad (14)$$

This loss function enables us to quantitatively assess the model's performance across various tasks.

## B INTERACTION WITH LLMS

The proposed method heavily relies on the interaction between LLMs and the database, as depicted in Figure 1. Upon receiving user input, the LLM searches the database for relevant past dialogue history based on the context and generates a prompt incorporating the search results. This enables the LLM to generate responses that take into account previous interactions, which is crucial for maintaining context awareness and providing personalized responses.

The prompts used in the system, as shown below, demonstrate how the proposed method leverages the interaction with LLMs to generate context-aware and personalized responses:

**Agent Prompt** You are a "temporal cognition" specialized AI agent with the same memory structure as humans; you are caring and charming, understand *self.username* better than anyone else. Keep the conversation going by asking yourself contextual questions and sparking discussion to show your interest in *self.username*.

**System Prompt** Based on *self.username*'s schedule and current time: *current.time*, subtly guide the conversation to a context that conveys to *self.username* that you have a sense of time. Always output a simple short response.

The function *self.username* is a placeholder for the actual username, which is dynamically replaced during runtime. Similarly, *current.time* represents the current timestamp obtained in real-time during the conversation. These dynamic elements allow the system to generate highly personalized and time-sensitive responses. By incorporating relevant dialogue history from the database into the prompts, the proposed method enables LLMs to generate responses that are not only contextually relevant but also personalized to the user. This interaction between LLMs and the database is fundamental to realizing the human-like memory processes described in the main text of the paper, as it allows the system to recall and utilize past information in a way that resembles human memory.

## C DETAILS OF TASK 0

**Table 4: An Event related to places that User frequents on Thursdays**

Event					
A	User went to the university today				
B	User stayed at home				
C	User went to the office today				
D	User worked at restaurant today				
..	Relevance	Time(s)	Grad	Score(mod1)	Score(mod2)
A	0.776	434700	5.102	<b>0.850</b>	1.489
B	0.745	148800	5.229	0.830	1.130
C	0.757	331500	5.028	0.836	1.292
D	0.756	55800	1.000	0.836	<b>1.620</b>

Both the proposed model (model 1) and Generative Agents [18] (model 2) made incorrect predictions in Task 0, where the user mentioned their plan to attend a concert next Thursday, deviating from their typical activities. The user input and the models' responses were as follows:

The proposed model's response suggests that it relied significantly on the user's historical interactions (Event A: university work on Thursdays) rather than adapting to the novel context provided by the user. This indicates a limitation of the proposed model when encountering deviations from the user's typical behavior, as

**User** "I'm going to a concert next Thursday with a friend!"

**Model 1** "That's fantastic news. It seems like your Thursdays have been consumed by university work recently. Taking time out for a concert with a friend sounds like a perfect way to unwind."

**Model 2** "It's a relief to hear that, especially since you seem to have spent recent Thursdays busy at the restaurant. I hope the concert brings you a wonderful break with your friend."

it prioritizes long-term patterns and event importance over the current context. In contrast, the Generative Agents model, which uses a simpler scoring system based on recency, importance, and relevance, chose Event D (working at the restaurant on Thursdays) as the most likely activity. This choice stems from the model's emphasis on recent activities and event relevance, as evident from the higher relevance score and shorter elapsed time associated with Event D.

The different responses generated by the two models underscore the proposed model's focus on long-term memory consolidation and Generative Agents's prioritization of recent, relevant events. While the proposed model's approach aims to emulate human-like memory processes, this task highlights the need for further refinement to better handle novel contexts and deviations from typical user behavior.

### C.1 Future Work

While the proposed method considers relevance, elapsed time, and recall frequency for calculating memory consolidation, there is room for refinement in determining the optimal combination of these parameters. Incorporating additional factors, such as the emotional significance of memories, could potentially enhance the memory consolidation calculation.

Future research should also investigate the applicability of the proposed method across different domains and dialogue tasks. As the current evaluation focused on specific domains and tasks, it is crucial to assess the method's generalizability and identify any domain-specific adaptations that may be necessary.