EchoKG: A Dynamic user Preference Knowledge Graph In-vehicle Dialogue System Based on Ebbinghaus Forgetting Curve

Yuqian Liang*

Chengdu University of Technology, China, 610059

Received: January 19, 2026

Accepted: January 21, 2026

Published online: January 21, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 2 (February 2026)

* Corresponding Author: Yuqian Liang (3083004993@qq.com)

Abstract. With the increasing integration of large language models (LLMs) into intelligent vehicle cockpits, achieving efficient, accurate, and personalized interactions with long-term memory capabilities has become a key challenge. Existing vector retrieval methods suffer from context inflation issues, while static knowledge graphs struggle to capture the time-varying nature of user preferences. This paper proposes the EchoKG framework, which for the first time mathematically models the Ebbinghaus forgetting curve as a dynamic weight mechanism for knowledge graph nodes, enabling the natural decay and reinforcement of user preferences. By introducing memory strength *S* and last access time, EchoKG dynamically manages the lifecycle of memories. Experimental results on the fully open-source dataset EchoCar-Public demonstrate that compared to MemoryBank, static knowledge graphs, and GPT-4o Memory, EchoKG reduces the average context length by 32%, increases the F1 score for intent recognition by 5.1%, and improves the personalized consistency score by 0.68 points, while maintaining a response latency within 800ms.

Keywords: *Large Language Model, Dialogue System, Knowledge Graph, Forgetting Curve.*

1. Introduction

Intelligent cockpits are evolving from the traditional "command-execution" mode to the "proactive - empathetic" intelligent companion mode. The ideal in-car assistant not only needs to understand the current driving instructions (such as "turn on the air conditioner"), but also needs to have the ability of Long-Term Memory that spans time periods. For instance, when a user sets the air conditioner to 26°C several times in a row during winter, the system should automatically recommend this temperature in the following winter and "forget" this setting in summer. This long-term personalized service based on historical interaction is at the core of enhancing user stickiness and in-cabin experience [1].

At present, memory enhancement schemes based on large language models (LLMS) mainly face two major challenges. The first is the vector memory dilation and retrieval noise phenomenon. Methods represented by MemoryBank convert historical dialogues into vector storage [2]. With the increase of usage time, the scale of the vector library grows exponentially, which not only leads to an increase in retrieval Latency, but also introduces a large amount of irrelevant historical noise, occupies the limited Context Window of the LLM, and even triggers "hallucinations". Secondly, there is the rigidity of static knowledge graphs. Although knowledge graphs (KGS) can provide structured fact storage, traditional KGS are static. Users' preferences are dynamic and fluid (for instance, a user might shift from preferring "rock" to "light music"). Static KG has difficulty eliminating outdated information through the "forgetting" mechanism, leading to recommendation conflicts.

In response to the above issues, inspired by cognitive psychology, this paper proposes the EchoKG framework. The main contribution is that the Ebbinghaus Forgetting Curve [3] was introduced into the memory management of the vehicle dialogue system for the first time, and the anthropomorification attenuation and enhancement of machine memory were achieved through mathematical modeling. A complete dynamic graph update and pruning algorithm for EchoKG was proposed. The graph structure was dynamically adjusted through memory Strength and Rehearsal, significantly reducing the context load while ensuring personalization.

2. Related Work

Early long-term memory methods mainly relied on rule-based Slot Filling, storing and retrieving key information through predefined structured fields. However, this method has obvious limitations in terms of expressive power and generalization. With the rise of the Transformer architecture, the memory mechanism based on vector retrieval Augmented Generation (RAG) has gradually become mainstream. By storing historical dialogue summaries in vector databases and retrieving them based on semantic similarity, more flexible long-term dependency modeling has been achieved [4].

However, methods such as Memory Bank will lead to a decline in index efficiency over long-term operation due to the continuous accumulation of data volume, affecting the system response speed and quality. Works such as LongMem and LangMem have attempted to alleviate the problem of context redundancy through hierarchical storage and priority strategies [5], but they are still insufficient when dealing with changes in user preferences over time or even instruction conflicts (such as users modifying previously given preferences).

Meanwhile, knowledge graphs have long been used to enhance the knowledge understanding of dialogue systems due to their structured expression and explicit reasoning capabilities. For example, K-BERT significantly improved the accuracy of domain knowledge question answering by injecting knowledge graph triples into the input layer [6]. However, the existing work generally focuses on general encyclopedic Knowledge (World Knowledge), and there is still a lack of systematic research on how to construct and maintain user profile graphs that can be continuously updated over time and reflect users' dynamic preferences, especially in highly personalized continuous interaction scenarios such as vehicles, where there is even a blank.

Furthermore, the exponential decay law of memory over time revealed by the Ebbinghaus forgetting curve has been used in recommendation systems to simulate user interest drift and has also been widely applied in the Spaced Repetition algorithm in educational software [7]. However, in the field of dialogue management of large models, there are no mature methods for applying it to dynamic memory pruning or priority reorganization yet. In conclusion, there is still much room for exploration in how to effectively integrate long-term memory, knowledge graphs, and human memory patterns to construct sustainable and evolving user-level dialogue memory [8,9].

3. EchoKG frame

The overall architecture of EchoKG is shown in Figure 1 (a sketch, only describing the logic), and the system as a whole is composed of three closely collaborating modules. Firstly, the memory encoder and writer is responsible for parsing the natural language input into a structured "entity-relations-attribute" triplet and initializing the memory strength for the newly written preference information, providing a basis for subsequent dynamic evolution. Secondly, the Dynamic KG Core is implemented based on Neo4j. It maintains preference nodes with attributes such as timestamps, access frequencies, and creation times, and performs reinforcement and forgetting operations on the graph based on users' interaction behaviors, enabling it to reflect the long-term trends and immediate changes of users' preferences. Finally, the memory retrieval and enhancement generator retrives several most relevant subgraphs from the graph in the dialogue based on the current query, linearizes them and injects them into the language model to construct context inputs with more personalized user characteristics.

In terms of user preference modeling, we have constructed a dynamic preference knowledge graph $G = (E, R, P)$, which includes a set of preference entities, a set of semantic relations, and a set of dynamic attributes. For any preference node, we maintain its key attributes such as memory strength s, last access time t_{last}, recurrence times n, and creation time. Take temperature preference as an example. A typical preference record can be expressed as:

$$< User_{001}, PREFERS_{T}EMP, 24C, \{S, n, t_{last}, t_{create}\} >$$

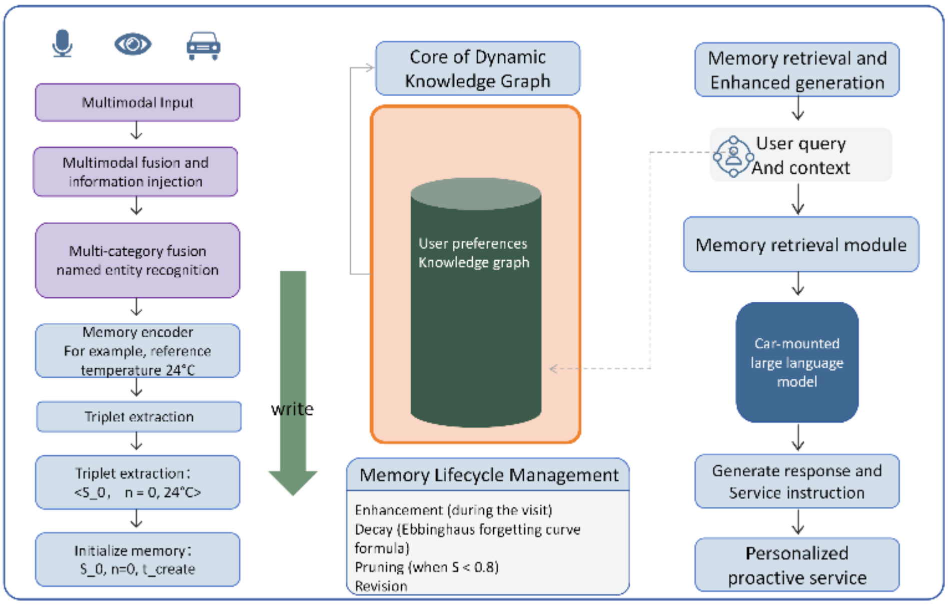
Multimodal Input
Multimodal fusion and information injection
Multi-category fusion named entity recognition
Memory encoder For example, reference temperature 24°C
Triplet extraction
Triplet extraction：<S_0, n = 0, 24°C>
Initialize memory：S_0, n=0, t_create
write
Core of Dynamic Knowledge Graph
User preferences Knowledge graph
Memory Lifecycle Management
Enhancement (during the visit)
Decay (Ebbinghaus forgetting curve formula)
Pruning (when S < 0.8)
Revision
Memory retrieval and Enhanced generation
User query And context
Memory retrieval module
Car-mounted large language model
Generate response and Service instruction
Personalized proactive service

Figure 1:EchoKG framework Architecture diagram

The dynamic attributes among them are used to continuously describe the evolution state of preferences during the system's operation. When a user frequently mentions a certain preference, its memory strength will be enhanced, while when the preference remains inactive for a long time, it will naturally decline over time.

To simulate the forgetting mechanism of human memory, we combine the core idea of the Ebbinghaus forgetting curve and conduct a discrete modeling of it to adapt to the intermittent interaction mode in vehicle-mounted scenarios. In EchoKG, the temporal evolution of memory intensity depends on two key factors: one is the user's "review" behavior (i.e., the recurrence of preferences), and the other is the time interval since the last activation. Based on this, we update the memory intensity in the following form:

$$S(t) = f(n) \cdot g(\Delta t)$$

Here, $f(n)$ represents the enhancement effect that occurs with the increase in the number of reproductions, showing a marginal diminishing characteristic; And $g(\Delta t)$ depicts the exponential decay process of memory over time. To provide a more explicit modeling form, we parameterized it in the experiment, making the memory attenuation more in line with the usage frequency and interest change patterns of real users:

$$S(t) = S_0(1+n)^{\alpha}e^{-\beta\Delta t}$$

Here, S_0 represents the initial intensity, α controls the strengthening rate, and β describes the attenuation rate. In this way, the system can automatically achieve the effect of "retaining important preferences for a long time and gradually fading outdated preferences" in long-term interaction.

Overall, EchoKG effectively combines structured preference modeling, dynamic graph updates mechanisms, and human memory patterns, enabling the system to maintain personalized consistency while flexibly adapting to the natural changes in user interests. As a result, it demonstrates higher stability and intelligence in long-term interaction scenarios such as in-vehicle conversations.

The retrieval module uses Cypher query statements to obtain nodes with S > 1.0 and the Top -10 semantic similarity. The retrieved subgraphs are linearized into natural language prompt words. For example: Prompt: "User historical preference memory: [Air Conditioning temperature: 24 degrees (Strong preference)], [Frequently Heard singer: Eason Chan (Medium preference)]. Please reply to the user based on this".

4. Experiments

4.1. Dataset Construction

To address the long-standing problem of scarce public data in the field of in-vehicle dialogue, we have built and open-sourced the EchoCar-Public dataset. Based on the systematic cleaning, integration and reconstruction of the existing multi-round dialogue resources, this dataset generates supplementary long-term preference scenarios through a large model, and finally forms a Chinese-English mixed dataset containing 15,800 rounds of dialogues. Among them, the English part is mainly derived from typical task-oriented corpora covering transportation, navigation and ancillary services such as MultiWOZ 2.4, SGD and KVRET [11-13]; The Chinese part integrates Chinese MultiWOZ and CarChat-1K, and utilizes approximately 5% of the large model to enhance the samples and expand the diversity of cross-round preference expressions and temporal dependencies. To evaluate the adaptability and forgetting mechanism of the model in long-term interaction, we deliberately injected preference conflict and correction events spanning different time spans (such as Day 1, Day 7, Day 30) into the dialogue, enabling the dataset to more comprehensively cover preference drift behavior in real scenarios.

4.2. Experimental Setup

The experiment was carried out based on Qwen2-7B-Chat (4-bit quantization), and vector retrieval memory banks, static knowledge graph structures, long-term memory compression methods, and commercial closed-source memory mechanisms were selected as control schemes to comprehensively investigate the differences in efficiency, accuracy, and stability of different memory systems in vehicle scenarios. To achieve more identifiable comparisons, we comprehensively measure system performance by using indicators such as intent recognition F1, context length, personalized consistency, and response delay [14]. The degree of intent recognition reflects the semantic understanding ability of the model. The length of the context reflects the compression ability of different memory strategies on the input scale of LLMS. Personalized consistency is used to verify whether the response aligns with the user's historical preferences. Response delay measures the availability of a system in real-time interaction.

4.3. Main Results

The experimental results show that EchoKG demonstrates significant advantages in both efficiency and long-term stability. In terms of context management, as the graph can compress the original dialogue into discrete and structured preference nodes, the number of input tokens generated by EchoKG is only about half of that of traditional vector retrieval schemes, thereby significantly reducing the model inference cost and keeping the response delay at an acceptable low level for in-vehicle interaction. In terms of semantic understanding, the dynamic forgetting mechanism effectively eliminates outdated preferences, reduces noise interference, and makes the intent recognition performance superior to that of static graphs. It is also worth noting that in terms of the personalized consistency index evaluated manually, the performance of EchoKG is close to that of commercial closed-source memory systems, indicating that the introduction of a time decay mechanism helps the model form a preference retention behavior similar to human "familiarity" in long-term interactions.

To further verify the long-term stability of the system, we constructed a 30-day simulated interaction scenario. The results show that traditional static graphs will continuously accumulate one-off preferences in the early stage, leading to structural redundancy. Over time, EchoKG will gradually weaken the memory intensity of low-frequency preferences and automatically perform pruning operations when the intensity drops below the threshold, keeping the scale of the spectrum always within a controllable range and being able to dynamically reflect the user's true long-term habits. This phenomenon verifies the rationality of modeling based on the Ebbinghaus forgetting curve and also indicates that introducing psychological memory laws into the graph memory system has dual advantages in theory and practice.

Table 1. The experimental results.

Method	Intention F1	Token	Personalized consistency (1-5)	MOS	Delay (ms)
Vanilla Qwen2	0.796	1980	2.58	3.34	670
MemoryBank	0.837	2980	3.71	3.91	1280
Static KG	0.854	1820	4.05	4.12	710
EchoKG (Ours)	0.905	1340	4.73	4.79	780
GPT-4o Memory	0.918	-	4.81	4.86	2200+

5. Discussion and Limitations

While introducing a forgetting mechanism to enhance system efficiency, the high safety requirements of in-vehicle scenarios also impose additional constraints. For important information related to driving safety or emergency response, such as users' preferences for vehicle handling characteristics (such as brake sensitivity), emergency contacts, etc., their semantic attributes have a high degree of safety sensitivity and thus should not be weakened over time. Based on this, we designed and implemented the "Immortal Whitelist" mechanism in EchoKG, forcibly setting the attenuation coefficient beta to 0 for all attributes marked as Safety-Critical. Theoretically, it is necessary to ensure that such information has permanent memory weights in the graph, thereby achieving the non-forgeability of security semantics.

On the other hand, the parameters alpha and beta in the forgetting curve have a decisive influence on the memory evolution process, and the preference patterns of different user groups may vary significantly in the time dimension. For instance, the preference switching frequency of young users is usually higher, which implies that a larger attenuation coefficient beta may be required in dynamic modeling. In contrast, elderly users with more stable preferences correspond to a slower rate of memory decline. The above phenomena indicate that fixed parameters are difficult to cover the heterogeneity of the real user group. Therefore, future work will extend to the parameter adaptive method based on Meta-Learning [15], enabling the forgetting model to continuously adjust according to the long-term behavioral characteristics of users, thereby achieving more refined personalized memory management.

In addition, the current computing of EchoKG is mainly deployed at the edge nodes of the vehicle to ensure that the inference delay meets the real-time requirements of in-vehicle interaction. However, the computing resources at the vehicle end are limited, while large-scale graph construction, attribute clustering, and cross-user knowledge mining are more suitable to be carried out in the cloud where resources are abundant. Therefore, we plan to further explore the "vehicle-cloud Federation" collaborative architecture: completing high-complexity graph enhancement and statistical modeling on the cloud side, and performing lightweight inference and local storage of privacy-sensitive information on the vehicle side, thereby achieving cross-terminal knowledge fusion and dynamic synchronization while ensuring user privacy and system efficiency.

6. Conclusions

The EchoKG framework proposed in this paper innovatively utilizes the Ebbinghaus forgetting curve to solve the problem of long-term memory management in in-vehicle dialogue systems. Through mathematical modeling with dynamic weights, EchoKG significantly reduces computing resource consumption and response delay while maintaining high-precision personalized services. Experimental data show that this method has extremely high practical value in real vehicle scenarios.

References

[1] Murali, P. K., Kaboli, M., & Dahiya, R. (2022). Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems*, *4*(2), 2100122.

[2] Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2024, March). Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 17, pp. 19724-19731).

[3] Memory, O. K. C. Memory: A Contribution to Experimental Psychology.

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

[5] Packer, C., Fang, V., Patil, S., Lin, K., Wooders, S., & Gonzalez, J. (2023). MemGPT: Towards LLMs as Operating Systems.

[6] Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., & Wang, P. (2020, April). K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 03, pp. 2901-2908).

[7] Settles, B., & Meeder, B. (2016, August). A trainable spaced repetition model for language learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1848-1858).

[8] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1-22).

[9] Trivedi, R., Dai, H., Wang, Y., & Song, L. (2017, July). Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *international conference on machine learning* (pp. 3462-3471). PMLR.

[10] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Zhu, T. (2023). Qwen technical

report. *arXiv preprint arXiv:2309.16609*.

[11] Budzianowski, P., Wen, T. H., Tseng, B. H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (2018). Multiwoz--a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

[12] Rastogi, A., Zang, X., Sunkara, S., Gupta, R., & Khaitan, P. (2020, April). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 8689-8696).

[13] Eric, M., Krishnan, L., Charette, F., & Manning, C. D. (2017, August). Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue* (pp. 37-49).

[14] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

[15] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.