Generation-Based and Emotion-Reflected Memory Update: Creating the KEEM Dataset for Better Long-Term Conversation

Jeonghyun Kang* **Hongjin Kim*** **Harksoo Kim†**
Konkuk University
{jeonghyun97, jin3430, nlpdrkim}@konkuk.ac.kr

Abstract

In this work, we introduce the Keep Emotional and Essential Memory (KEEM) dataset, a novel generation-based dataset designed to enhance memory updates in long-term conversational systems. Unlike existing approaches that rely on simple accumulation or operation-based methods, which often result in information conflicts and difficulties in accurately tracking a user's current state, KEEM dynamically generates integrative memories. This process not only preserves essential factual information but also incorporates emotional context and causal relationships, enabling a more nuanced understanding of user interactions. By seamlessly updating a system's memory with both emotional and essential data, our approach promotes deeper empathy and enhances the system's ability to respond meaningfully in open-domain conversations.

1 Introduction

Open-domain conversational systems have increasingly achieved human-like interaction capabilities. Recent advancements have led to the development of long-term conversation chatbots and datasets (Xu et al., 2022a). Despite these significant strides, the study of memory management in long-term conversations has not been adequately addressed. In real-life conversations, a user's state or conditions frequently change, necessitating dynamic memory adjustments. Previous methods have primarily relied on accumulating new memories while retaining previous ones, a strategy that inadequately addresses the need for revising or deleting outdated information. Furthermore, in multi-session conversations, as dialogues extend, the requirement for expanded memory capacity intensifies. Although some studies (Shuster et al., 2022; Ma et al., 2021; Xu et al., 2022b) have suggested a simplistic approach of deleting old memories to make room for new ones, this method fails to recognize the necessity of updating previous memories based on new information, highlighting the need for a more sophisticated memory management solution.

Updating memory based on information from new dialogues, through comparison with previous memories, can significantly enhance user engagement. This approach allows conversational systems to inquire about recent conditions and reference past conversations accurately, based on updated memories. Moreover, this method contributes to efficient memory capacity management by deleting or revising outdated memories to reflect the latest user state. Such dynamic memory management not only supports maintaining a concise and relevant memory database but also facilitates accurate tracking of the user's current conditions, thereby improving the conversational system's responsiveness and personalization. Despite the importance of memory management, it has been relatively understudied, possibly due to the scarcity of relevant datasets. Bae et al. (2022) introduced a method and dataset for memory updating named *CareCallmem*, which classifies the relationships between sentence pairs into four operations: PASS, APPEND, DELETE, and REPLACE. These operations enable the identification and removal of information that could cause confusion or redundancy in subsequent conversations. The DELETE operation eliminates both the previous and new memory when the information is no longer needed, while REPLACE retains only the new memory in cases of conflict between the old and new information. However, these operations present significant challenges in memory management. For instance, determining the appropriate operation can be ambiguous, as illustrated when comparing sentences like "I have a cold and am taking medicine" with "I

* Equal contribution
† Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 9260–9277
January 19–24, 2025. ©2025 Association for Computational Linguistics

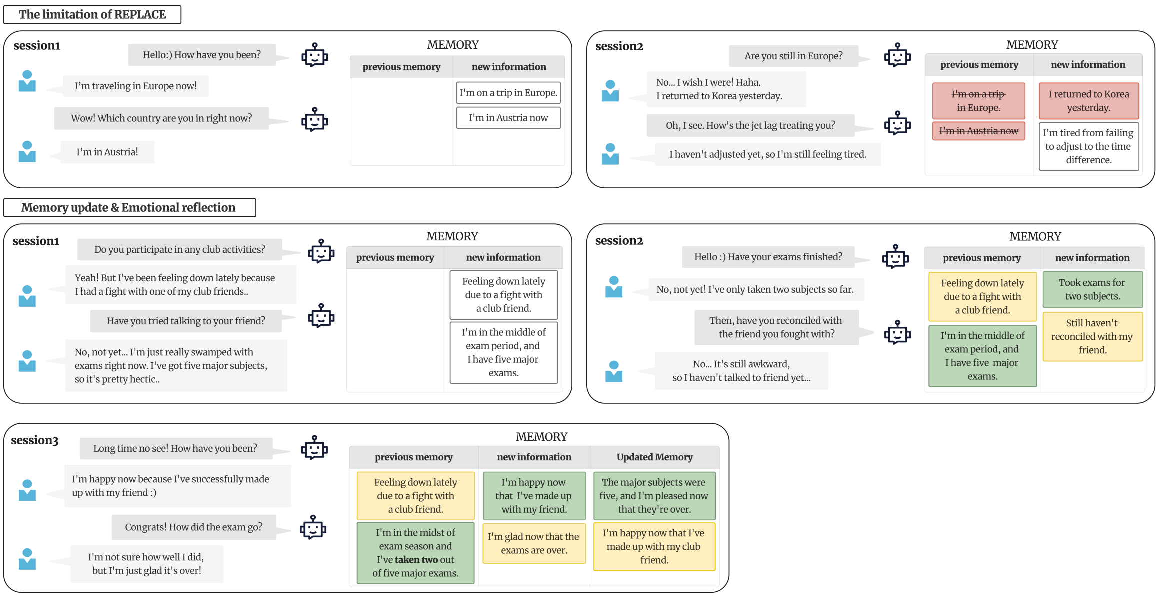

Figure 1: Limitations of previous method and example of memory update.

am completely cured of my cold," leading to confusion about the correct operation to apply. Furthermore, using DELETE in situations where both sentences, such as "I have COVID" and "I am completely cured of COVID," are eliminated, the system loses the ability to remember that the user had COVID. Similarly, with REPLACE, if the system updates from "I am traveling in Europe" to "I am back in Korea," it fails to retain the information that the user was previously in Europe (See Figure 1). In daily conversations, user experiences and episodes unfold over time. For instance, processing two memories, "I have 4 subjects left for the exam" and "I took the exam for 1 subject," using DELETE or REPLACE operations proves inadequate. To address these challenges, we introduce a generation-based memory update dataset designed to preserve essential information for long-term conversation systems. Our approach aims to generate a new sentence that retains critical information when confronted with sequential memories, such as transitioning from "I am traveling in Europe" to "I am back in Korea." Instead of merely selecting or removing, our system generates an integrative memory, such as "I traveled to Europe," ensuring the preservation of essential user information.

Existing multi-session chat (MSC) datasets often focus on summarizing events or episodes, neglecting the emotions related to these events or failing to account for the causes of the user's emotions. Emotion plays a pivotal role in human-like conversational systems. Systems that reference emotions without their underlying causes offer only superficial empathy. In contrast, acknowledging both emotions and their causes facilitates a deeper understanding and empathy (Gao et al., 2021). This becomes particularly valuable when users express negative emotions, enabling the system to offer solutions or advice, with potential applications in mental health care (Wang et al., 2023a). To address this gap, our memory update dataset incorporates both emotion and its causes from dialogues, as shown in Figure 1. Finally, we introduce the Keep Emotion and Essential Memory (KEEM) dataset, which is generation-based and uniquely integrates emotional context and causality. To sum up, our main contributions are as follows:

1. We propose a novel generation-based memory update dataset that enhances long-term conversational systems by dynamically generating integrative memories to retain essential information, addressing the limitations of existing DELETE and REPLACE operations.

2. Our work introduces the Keep Emotion and Essential Memory (KEEM) dataset, the first of its kind to integrate both emotion and causality into memory updates facilitating deeper

empathy and understanding in conversational AI.

3. We demonstrate the quality and applicability of our dataset in improving user engagement and system responsiveness in multi-session conversations.

2 Related Work

To support the study of long-term conversations, Xu et al. (2022a) developed an MSC dataset, which comprises multiple dialogue sessions. Each session in the MSC dataset is accompanied by a summary that serves as a memory for subsequent dialogue sessions. However, while the MSC dataset facilitates the continuation of dialogues across sessions, it does not explicitly address the process of memory updates between these sessions.

CareCallmem (Bae et al., 2022) represents the first Korean memory update dataset, developed from the *CareCall* dataset, which was designed for a role-specific open-domain dialogue system targeting elderly individuals living alone. The *CareCall* system's dialogues predominantly center on the users' well-being, emphasizing health and meals, which narrows the diversity of dialogues within the dataset. Moreover, the summaries within the *CareCall* dataset often omit significant details, tending to highlight events at the expense of the users' emotions or thoughts about these events. This approach to summarization leads to brief summaries that overlook the nuanced points of events and fail to capture the full context, potentially resulting in inaccurate memory updates in subsequent dialogues (See example in Appendix A.2).

Recent advancements in empathetic dialogue (ED) research have expanded on the ED dataset initially introduced by Rashkin et al. (2019). Sabour et al. (2022) have developed a model that boosts cognitive empathy by using common sense to better understand and empathize with users' situations. Furthermore, Zhou et al. (2022) introduces a technique that merges cognitive insights with emotional reactions through their CASE model, highlighting the synergy between cognitive and emotional elements. Nonetheless, these studies primarily concentrate on crafting empathetic responses using the dataset that is based on English. Recently, the Korean cultural dialogue dataset was constructed (Jin et al., 2024); however, this dataset does not focus on memory updates for long-term dialogue.

3 KEEM Dataset Creation

To diversify dialogue topics, we opt for the Korean MSC (KMSC) dataset from AI Hub[1] instead of the CareCall dataset. The KMSC dataset encompasses 13 themes, such as 'individuals and relationships', 'education', 'climate', and 'beauty and health', among others (for more details on the KMSC dataset, see Appendix A.1). It comprises four sessions, each with a summary. However, it is not designed with memory updates in mind; summaries from each session are merely appended and used as the basis for subsequent dialogues' memories. Furthermore, similar to other datasets, the KMSC summaries often lack information on the user's emotions or their causes, with some containing emotions but omitting their origins(utterance: "I work as a designer. Recently, I sometimes feel embarrassed because of past work mistakes." -> summary: "I'm a designer and I sometimes feel embarrassed these days.").

To construct our KEEM dataset, which integrates both emotion and causality into generation-based memory updates, we create new summaries using dialogues from the KMSC dataset and ChatGPT 4.0 (OpenAI, 2023). ChatGPT has demonstrated significant capabilities in various NLP tasks, including dataset creation (Wang et al., 2023b), with multiple studies validating its effectiveness and efficiency (Cegin et al., 2023). We leverage ChatGPT to first incorporate users' emotions and their causes into the summaries. Subsequently, ChatGPT is instructed to generate an updated summary that combines the previous and current sessions' summaries. Figure 2 shows the framework of our KEEM dataset creation. Table 1 shows the statistics of the KMSC and our KEEM datasets. In Table 1, note that 'Session 1-n' refers to data comprising n sessions. For example, 'Session 1-4' indicates data encompassing four sessions, not the fourth session.

3.1 Emotions & Causes Reflection

Given a session's dialogue D and the session's summary S, we instruct ChatGPT to incorporate the user's emotions and their causes into S, drawing from any cues present in D. We explore three prompting methods to effectively reflect emotions and their causes: 1) Initially, we test instructions in both English and Korean, given the dialogues are in Korean. We apply these instructions to 100 samples and conduct a manual evaluation of the

[1] https://www.aihub.or.kr/

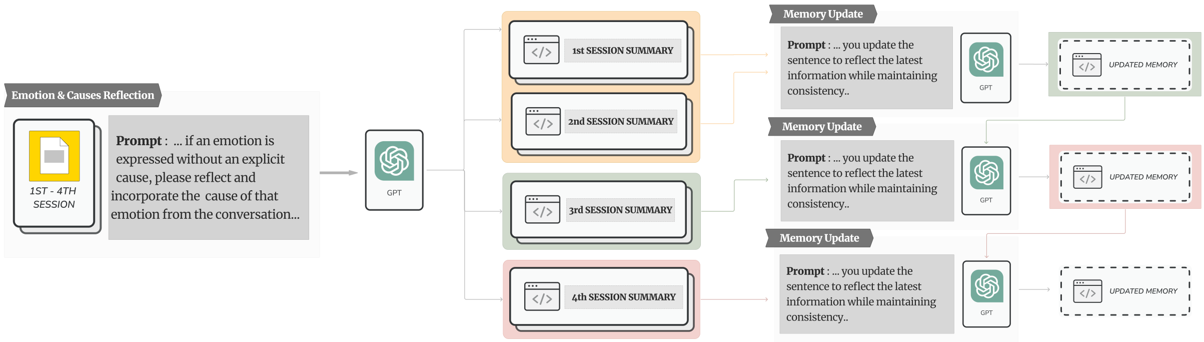

Figure 2: Process of KEEM dataset creation.

outcomes from each language setting (For examples of the results, see Appendix A.3). The Korean instructions yield superior results in reflecting dialogue content and updating whole dialogue contexts; thus, we have adopted them for the entire dataset. 2) We then experiment with a reasoning prompt that include potential emotions observed in the dialogues, such as pleasure, sadness, and happiness, among others. We assess 100 examples comparing the effectiveness of prompts with and without specified emotion candidates. The prompt without emotion candidates generate summaries that more accurately reflected emotions and their causes. 3) Lastly, we employ a few-shot approach, providing ChatGPT with specific examples to enhance its performance. This method demonstrates improved outcomes in capturing the nuances of emotions and their causes. Consequently, we select the Korean instructions without emotion candidates, incorporating few-shot examples, as our final prompting strategy.

3.2 Memory Update

In the context of memory updates, given the existing memory (i.e., previous sessions' summaries) M and the current session's summary S (or the current session's dialogue D), our objective is to generate an updated memory M^U. Consistent with Section 3.1, we employ Korean instructions and a few-shot approach. We explore two main prompting strategies for effective memory updating: 1) We initiate by comparing the outcomes of providing ChatGPT with either the current session's summary S or dialogue D, applying these instructions to 100 samples for a manual evaluation. Given the similarity in results and to minimize costs, we opt for using the summary S, as its brevity translates to lower ChatGPT API usage costs. 2) We experiment with a reasoning prompt aimed at facilitating ChatGPT's handling of multiple sentences in both M and S, which can be challenging due to the need to correlate summary sentences with the corresponding memory. Initially, we direct ChatGPT to match summary sentences with related memory sentences and then generate the updated memory M^U. This approach is tested both as a single integrated instruction and through separate instructions. However, neither method outperformed direct instructions for memory updating (for examples, see Appendix A.4). Therefore, we proceed by simply providing ChatGPT with M and S, instructing it to update the memory. To ensure the relevance and accuracy of updated memory, we conduct a final verification step using ChatGPT. We input all session dialogues up to the current time step alongside updated memory, instructing the model to verify whether the updated memory accurately reflects the content of the dialogues. If updated memory does not pass this final verification, it is excluded from the KEEM dataset. Post-verification, the refined memory is designated as M^U.

4 Experiments

4.1 Implementation Details

We utilize the ChatGPT 4.0 (gpt-4-1106-preview) API to create the KEEM dataset, setting the temperature parameter to 0.0. Both top-p and top-n settings are configured to 1 to ensure deterministic output. The evaluation of our KEEM datasets is conducted on one NVIDIA A100 80GB GPU, which supports all experiments involving long-term chatbot models.

4.2 Human Evaluator Details

For the manual evaluation process, we engage five native speakers familiar with chatbot systems as human evaluators. These individuals are not in-

Attribute	KMSC			KEEM		
	Session 1-2	Session 1-3	Session 1-4	Session 1-2	Session 1-3	Session 1-4
Total Episodes	40,000	20,000	20,000	2,006	1,560	1,005
Total Utterances	980,919	731,705	953,405	61,354	70,974	59,847
Total Memory sentences	567,176	395,808	498,497	33,972	31,846	23,148
Avg. Length of Utterances	16.98	17.13	17.22	17.02	17.26	17.32
Avg. # of Utterances	30.65	45.73	59.59	30.58	45.49	59.54
Avg. # of Memory Sentences	17.72	24.73	31.15	16.93	20.41	23.03

Table 1: Statistics of the original KMSC dataset and KEEM dataset. Avg. denotes the average.

Memory	E.&C. Ref. rate	Score
KMSC memory	35%	1.18
KEEM memory	93%	1.90

Table 2: Results of manual evaluation of emotion and cause reflection. *E.&C. Ref. rate* denotes the percentage in which emotions and their causes appear in memory.

Session	Update Need Rate	Score
Session 1-2	12%	1.75
Session 1-3	31%	1.64
Session 1-4	21%	1.62

Table 3: Results of manual evaluation of updated memory. Note that 'Session 1-n' refers to data comprising n sessions, rather than indicating the session number.

Updated Memory	Score	Voting	
		Human	ChatGPT
KEEM	1.86	76%	81%
CareCallmem	1.60	24%	19%

Table 4: Comparative results of memory update methods between the KEEM dataset and the *CareCallmem* dataset.

formed about the development processes of either the KEEM or the *CareCallmem* datasets.

4.3 Evaluation of Emotion and Cause

To validate the effectiveness and quality of our KEEM dataset, we conduct comprehensive evaluations and experiments focusing on the reflection of emotion's causes in the updated summaries. We manually assess whether the causes behind emotions are accurately integrated into the memory updates (Details for the human evaluator are in Section 4.2). The evaluation is based on a scale ranging from 0 to 2, where a score of 0 indicates no reflection of emotions or their causes, a score of 1 signifies that only emotions are reflected without their causes, and a score of 2 denotes that both emotions and their causes are effectively captured. We analyze 50 sessions and their corresponding memories to evaluate the accuracy of emotion and cause reflection. The results, as detailed in Table 2, demonstrate that our updated memory significantly improves the retention of users' emotions and their causes compared to the raw memory. This enhancement suggests that using our updated memory in conversational AI systems could facilitate deeper empathy in subsequent interactions with users.

4.4 Evalution of Memory Update

To further validate the accuracy of our memory updates, we conduct a manual evaluation based on a scale from 0 to 2. A score of 0 indicates that no update is performed on the memory, a score of 1 suggests that a partial update is made, and a score of 2 denotes a fully accurate update. This assessment is carried out on 100 samples for each session. Even in long-term conversations, memory updates are not always required after each session. Therefore, we measured the rate at which memory updates are needed across the entire conversation dataset. As depicted in Table 3, the necessity for memory updates tends to escalate as the number of sessions increases. The results reveal that our updated memory achieves highly accurate updates, indicating that employing our updated memory in long-term conversational systems could effectively maintain and track the user's evolving state or conditions.

4.5 Comparative Evaluations of Memory Update Methodologies

4.5.1 Automatic & Human Evaluation

To benchmark our KEEM dataset against the existing update method, especially the *CareCallmem*

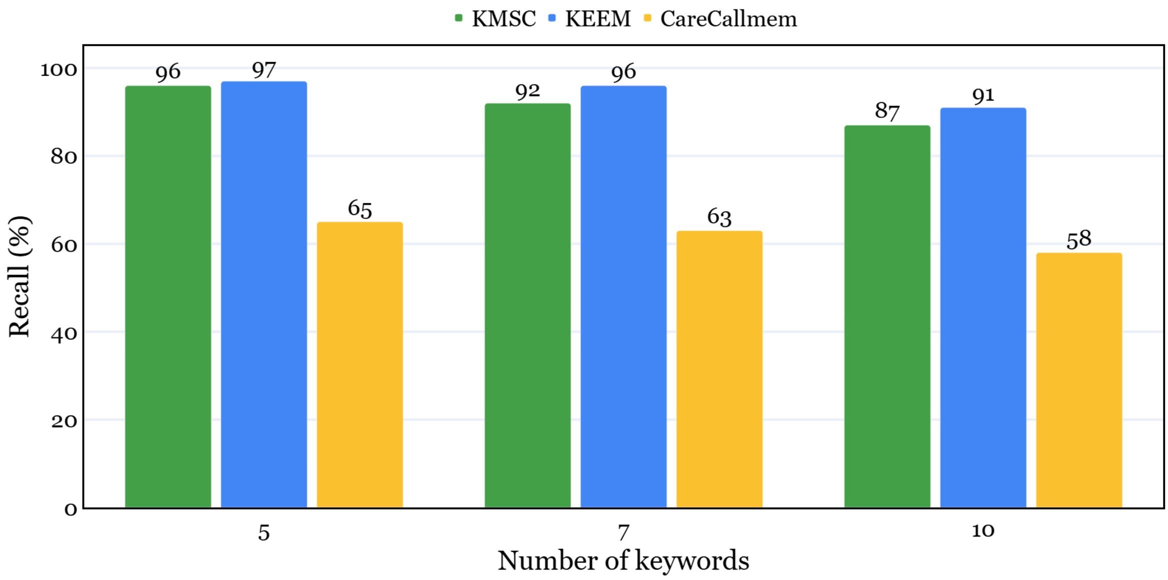

Figure 3: Results of keyword coverage across the different memory update methodologies.

dataset, we apply *CareCallmem*'s methods to update 50 memories in the KMSC dataset. For a fair comparison, we utilize ChatGPT 4.0 API to execute the DELETE, REPLACE, APPEND, and PASS operations for memory updates within the KMSC dataset. Detailed instruction is provided in Appendix A.5. Subsequently, a manual evaluation is conducted using a scale from 0 to 2, mirroring the assessment criteria used for our memory updates. Moreover, we ask evaluators and ChatGPT to determine which method produces better memory updates: our approach or the *CareCallmem* method. Table 4 presents the results of evaluations conducted by both human evaluators and ChatGPT. As indicated in Table 4, both our KEEM dataset and the *CareCallmem* dataset demonstrate strong performance in human evaluations, with our KEEM dataset achieving superior results compared to the *CareCallmem* dataset. Furthermore, in the combined voting evaluation conducted by humans and ChatGPT, the KEEM dataset significantly outperforms the *CareCallmem* dataset. These results underscore the efficacy of our memory update methodology, showcasing its ability to provide more comprehensive and informative updates to users.

4.5.2 Informativeness & Conflicts Evaluation

In this section, we discuss the informativeness of updated memory and conflicts between memory sentences.

Informativeness of Updated Memory To compare the informativeness of summaries and updated memories across different methodologies—accumulation (i.e., KMSC), generation (i.e., our KEEM), and operation (i.e., *CareCallmem*)—we extract keywords from the entire session dialogues, denoted as $K^D = [keyword_1^d, keyword_2^d, \dots, keyword_I^d]$, and from updated memory, denoted as $K^M = [keyword_1^m, keyword_2^m, \dots, keyword_J^m]$. Here, I and J represent the number of keywords from the whole dialogue and the last updated memory, respectively. We then analyze the extent to which K^M covers K^D, as follows:

$$\text{Recall} = \frac{\text{Number of correct keywords in } K^M}{\text{Number of keywords in } K^D} \tag{1}$$

We utilize TextRank (Mihalcea and Tarau, 2004) along with the TF-IDF method to extract keywords from the entire session dialogues and the last updated memory. Initially, we calculate word co-occurrence across the entire dialogue. Subsequently, we extract different sets of keywords, specifically 5, 7, and 10, and calculate their recall. Figure 3 presents the results of keyword recall for different memory update methods. The results demonstrate that the updated memories by our generation method, KEEM, outperform those from both accumulation and operation-based methods in keyword recall. Although we anticipate that the accumulation method, KMSC, would show the highest keyword recall due to its comprehensive

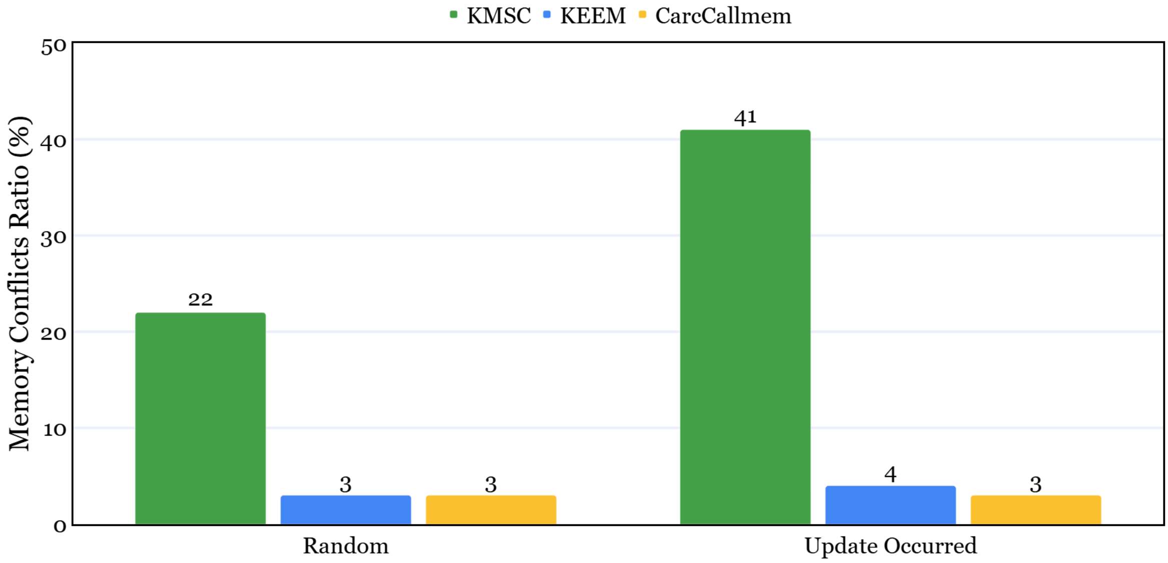

Figure 4: Ratio of conflicts between memory sentences across the different memory update methodologies.

memory storage, our KEEM method exhibits superior recall. Upon further analysis, the underperformance of the KMSC method is attributed to the quality of its summaries. Despite the KMSC dataset accumulating session summaries, it appears that important information is missing from these summaries. Meanwhile, the operation-based method, *CareCallmem*, exhibits poor keyword recall performance. This deficiency can be attributed to the REPLACE and DELETE operations, which tend to eliminate essential information during the memory updating process. Consequently, the higher keyword recall scores of our KEEM dataset confirm that our updated memory is highly effective and useful.

Conflicts between Memory Sentences To analyze conflicts between memory sentences according to the update methodologies, we employ the natural language inference (NLI) task. Specifically, we instruct ChatGPT 4.0 to determine whether a hypothesis is an entailment, neutral, or contradiction for all possible sentence pairs in the updated memory, with one memory sentence as the premise and another as the hypothesis. For a thorough analysis of conflicts, we utilize two data sampling methods: first, we randomly sample 50 updated memories; second, we specifically select 50 samples where a memory update occurs. Figure 4 presents the results regarding the sampling method for all datasets. The accumulation method (i.e., KMSC) exhibits the highest ratio of conflicts between memory sentences, likely due to the method's practice of merely accumulating session summaries as conversations progress. In samples with memory updated, the ratio of conflicts significantly increases under the accumulation method, indicating that merely accumulating session summaries is not effective for long-term conversation tasks. For the generation (i.e., KEEM) and operation-based (i.e., *CareCallmem*) methods, both show a significantly lower ratio of conflicts. The operation-based method, due to its REPLACE and DELETE operations, already eliminates sentence pairs that could lead to contradictions. Although this characteristic of the *CareCallmem* method reduces conflicts, it also tends to remove essential information from dialogues. Conversely, our generation-based update method, KEEM, not only effectively retains crucial information in dialogues but also ensures minimal conflicts between memory sentences.

4.5.3 Evaluation with Long-term Conversation Models

To assess the application of updated memories across different methodologies, we utilize various long-term conversation models. Previous research of multi-session chat by Xu et al. (2022a) employs models such as RAG (Retrieval-Augmented Generation) (Lewis et al., 2020), FiD (Fusion-in-Decoder) (Izacard and Grave, 2021), and FiD-RAG (Shuster et al., 2021) to model multi-session chats, evaluating their performance on the MSC dataset. However, the challenge remains in assessing the

Model	Perplexity			Voting		
	KMSC	KEEM	CareCallmem	KMSC	KEEM	CareCallmem
RAG	9.74	8.10	10.55	26%	67%	7%
FiD	9.51	7.88	10.49	22%	68%	10%
FiD-RAG	9.66	7.90	10.52	19%	77%	4%
Llama2 7B	8.70	6.99	10.50	16%	81%	3%
Llama2 13B	8.53	6.89	10.51	15%	82%	3%
Llama10.8B tuned Korean	5.47	4.56	8.13	22%	74%	4%
Phi2.8B tuned Korean	5.83	4.61	8.72	18%	79%	3%

Table 5: Perplexity results on various models

application of updated memories, as merely conversing using these models does not ensure that information has been updated, even in multi-session conversations. To address this, we design a fifth session dialogue specifically to include at least one turn discussing updated information, ensuring the dialogue directly engages with the updated memories. We then input this fifth session dialogue into the models and evaluate their ability to respond appropriately to utterances mentioning the updated information. Following the previous work (Xu et al., 2022a), we use RAG, FiD, and FiD-RAG. Additionally, we use large language models (LLMs) such as Llama2 [2] [3] (META, 2023) to assess the application of memories. We also use models tuned with Korean based on Llama[4] or Phi-2[5] (Li et al., 2023). For the automatic evaluation metric, we utilize perplexity, anticipating that the most effective memory would exhibit the lowest perplexity, indicating lower generation difficulty for the given input. Initially, we input each updated memory from different methodologies (Note: memories are generated using the methodology of *CareCallmem*, as mentioned in Section 4.5.1) along with a fifth session dialogue into the models. We then evaluate their perplexity when responding to utterances that mentioned the updated information. Subsequently, we assess the responses generated by the models by instructing ChatGPT 4.0 to vote for the most appropriate response given the context of the entire session's dialogue. Table 5 presents the perplexity performance of various models, showing that models tuned with Korean exhibit better perplexity compared to others. Memories updated using the *CareCallmem* method exhibit the highest perplexity, suggesting that the operation-based method is inadequate for retaining essential information, leading to poor response generation. In contrast, the accumulation method (KMSC) displays better perplexity than the operation-based method, despite potential contradictions between memory sentences, as it does not remove memorable information. Our generation-based method, KEEM, outperforms all baselines across various models. These results suggest that our updated memories are more useful and that the KEEM dataset is more effective in managing the recent state or information of the user while retaining memorable information.

5 Analysis

Because we utilize the ChatGPT API to create the KEEM dataset, we analyze whether hallucinations occurred in the updated memory. In the memory updating process, we instruct ChatGPT to generate updated memory by integrating the previous memory with a new session summary. Consequently, hallucinations—generated content not present in the dialogue—might not occur. We also manually assess these hallucinations and confirm their absence in the KEEM dataset. Therefore, we report the generation errors—such as incorrect integration of unrelated information or contradictions during the integration of related information—that occurred during the memory update process. Table 6 presents examples of these generation errors. The first example in Table 6 illustrates a contradiction in the updated memory, while the second example highlights the integration of unrelated information.

To assess generation errors, we manually evalu-

[2]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[3]https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[4]https://huggingface.co/yanolja/EEVE-Korean-10.8B-v1.0

[5]https://huggingface.co/yanolja/EEVE-Korean-2.8B-v1.0

Previous Memory	Subsequent Session's Summary	Updated Memory
My son is a high school student and his grades were good until his first year.	My son entered the physics department at a university in Seoul.	My son is a high school student; his grades were good until his first year, and now he has entered the physics department at a university in Seoul.
I recently visited Yeosu and have chosen Damyang as my next travel destination.	I have been to Jeonju.	I recently visited Yeosu and have chosen Jeonju as my next travel destination.

Table 6: Examples of memory update errors

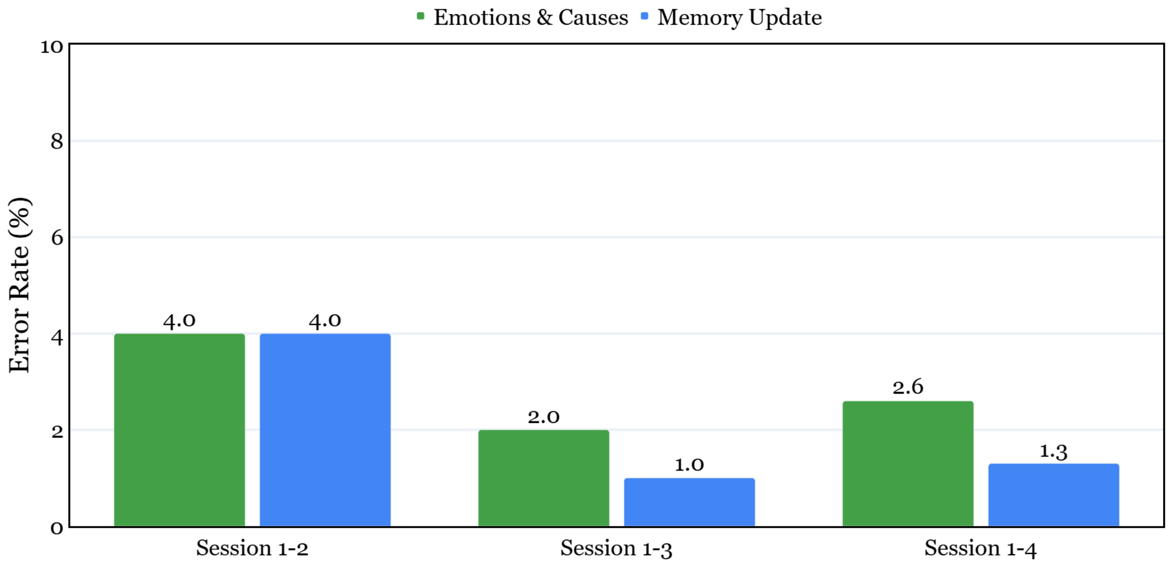

Figure 5: Results of generation error analysis.

ate updated memories in our KEEM dataset across 150 samples, distributed evenly with 50 memories from each of Session 2, Session 3, and Session 4 data. The KEEM dataset creation involves two phases: 1) Emotion & Causes Reflection (Section 3.1); and 2) Memory Update (Section 3.2). Generation errors can occur in either phase. To comprehensively analyze these errors, we identify the specific phase in which each error occurs. Figure 5 displays the generation error rates, where Session n indicates data comprising n sessions. The result reveals significantly low generation errors in both phases. Additionally, we sample 70 datasets where information is updated (note that information updates do not always occur) and manually evaluate the occurrence of generation errors during the update process. Only 6 samples exhibited errors, approximately 8.5%. While this rate of generation error in our KEEM dataset is low, a comparison with other datasets is needed to determine its relative performance. Wang et al. (2023b) reported error rates of their created dataset in the task description, input, and output generation using GPT-3 (Brown et al., 2020) as 8%, 21%, and 42%, respectively, resulting in only 54% valid data across all fields in their dataset. This underscores the effectiveness of ChatGPT 4.0 in creating new datasets, as evidenced by the notably lower error rate in our KEEM dataset. Moreover, studies on automatic summarization have reported generation error rates around 30%. Compared to these levels, our KEEM dataset demonstrates a remarkably low error incidence, suggesting a high level of data quality.

6 Conclusion

We introduced the KEEM dataset, which integrates emotion and causality into memory updates for long-term conversations. Our evaluations showed that KEEM effectively captures emotions and their causes, outperforming traditional operation-based methods in updating user information. Future work will explore models that consider temporal gaps between sessions, utilizing large language models to improve memory understanding.

Limitations

1) In our analysis, we noted instances where ChatGPT inadvertently deleted content unrelated to the targeted memory updates, which contradicts the intended behavior of preserving relevant information. To address this issue, we explicitly instructed ChatGPT to avoid deleting any content not directly related to the updates. Although this strategy somewhat mitigated the problem, it did not fully resolve the inadvertent deletion of unrelated content. 2) Furthermore, to reduce costs in the memory update process, we opted to provide ChatGPT with the previous memory and the current session's summary instead of the full dialogue from the current session. This strategy, though cost-effective, risks omitting crucial information if the current session's summary lacks comprehensiveness. 3) Additionally, since ChatGPT is a generative model, there is a risk of generation error being introduced into the updated memory. To address this, we implemented a verification step, re-evaluating the updated memory with ChatGPT to ensure its accuracy and relevance.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190004, Development of semisupervised learning language intelligence technology and Korean tutoring service for foreigners). This work was also supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI).

References

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905. Association for Computational Linguistics.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 555–564.

GenAI META. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11229–11237.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023a. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3952–3963.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. *arXiv preprint arXiv:2208.08845*.

A Appendix

A.1 Details for Topics

Table 7 shows the various topics used in our KMSC dataset and their statistics.

A.2 *CareCallmem* Examples

Table 8 displays the examples of *CareCallmem* dataset.

A.3 Emotions & Causes Reflection Result Examples

Table 10, 11 show the examples of emotion and its cause reflection.

Topic	Count	Ratio
Individuals & Relationships	12,783	15.98%
Entertainment	11,063	13.83%
Beauty & Health	9,038	11.30%
Society	7,328	9.17%
Work & Job	6,997	8.75%
Arts & Culture	6,508	8.13%
Education	6,288	7.86%
Food	4,130	5.15%
Climate	3,975	4.97%
Traffic	3,795	4.75%
House	3,489	4.37%
Fashion	235	0.29%

Table 7: Topic statistics of KMSC dataset

A.4 Memory Update Result Examples

Table 12 shows our memory update instruction and its result. Table 13, 14 show the used few-shot examples for memory updates.

A.5 Examples of *CareCallmem* Instruction

Table 15 shows the instruction of *CareCallmem* method.

Input dialogue and summary (translated into English)
Dialogue: system: Hello, this is the Administration Welfare Center calling to check in with you. user: Yes. system: Alright, is it okay to talk now? user: Yes. system: How have you been lately? user: I've been tending to an additional houseplant that I got. system: Wow, what kind of plant is it? user: I'm not sure. My husband brought it as a gift. system: I see. Take good care of it. user: Thank you. system: Alright, is there anything bothering you by any chance? user: Nope, nothing at all. system: That's good to hear. I hope you stay healthy in the future too. user: Okay. system: How are you feeling today? user: I'm feeling really good. system: Oh, I see. What were you doing before our call? user: I was cleaning the new houseplant that arrived. system: I see. Having a new plant must make you feel good. user: Yes, it does. system: Keep taking good care of it. user: Yes, thank you. system: Alright, can we end the call here today? user: Sure. system: Take care. Thank you for taking the call. **Memory:** · Health condition is good. · Frequently drinks warm water and tea for health. · Sleeps well · Eats rice cakes · Eats meals well **Summary:** · Has a husband

Table 8: The example of CareCallmem.

Instruction (translated into English)
If there are any inaccuracies or deficiencies in the sentences of [{speaker}'s summary] after understanding the overall flow and content of the [conversation], please correct those sentences. In particular, If an emotion is expressed without an explicit cause, please reflect and incorporate the cause of that emotion from the conversation when making corrections. Do not add emotions if the speaker did not explicitly express them. Please note that you should not generate any new content not present in the conversation, and during this process, no new sentences should be added to [{speaker}'s summary], nor should any existing sentences be deleted. Additionally, [{speaker1}'s summary] should not contain the speaker2's speech. Output the corrected sentences along with the remaining sentences in [{speaker}'s summary] in the same format as [Modified {speaker}'s summary].

Table 9: Our prompts that allow the addition of context and causes when only emotions are expressed in summaries of each speaker, enabling understanding of the conversation's content and context.

Input dialogue and summary (translated into English)
Dialogue: Speaker1: Hello, I'm a female CEO in my 40s at a mid-sized company. Speaker2: Hi there, I'm a 40-something woman working in a professional field. Speaker1: Nice to meet you. I play VR games for work every day. Speaker2: Oh, I see. I often do yoga after work. Speaker1: Is yoga your hobby? Mine is occasionally playing the oboe. It's nice not having neighbors around for this. Speaker2: That's a cool! I sometimes listen to K-pop music, and these days, idols are really talented, especially in singing. Speaker1: Absolutely! Nowadays, idols are so versatile. Lately, I've been eating fruit to boost my vitamins, but I don't like persimmons. Do you like fruit? Speaker2: I'm not really into fruits. Lately, I've been feeling a bit restless at times. I've heard that kiwi can help with sleep, so I've been making sure to eat plenty of kiwis as I've been having trouble sleeping. Speaker1: I didn't know kiwi could help with sleep! What's been going on with you lately? Speaker2: Well, my brother is a hairstylist, and he keeps talking about expanding his salon, especially in times like these...you know, during this economic downturn. Speaker1: It's probably best not to expand recklessly at times like this, you know. Speaker2: I agree. It's important to manage what he have well, but I'm worried. Speaker1: I hope your brother's salon continues to do well! **Summary:** [speaker2's summary] · I am a woman in my 40s working in a professional field. · I often do yoga after work. · I listen to K-pop music occasionally. · I don't particularly like fruit, but I make sure to eat kiwi. · I've been feeling anxious occasionally lately. · My brother is a hairdresser. · I'm worried about my brother. [modified speaker2's summary] · I am a woman in my 40s working in a professional field. · I often do yoga after work. · I listen to K-pop music occasionally. · I don't particularly like fruit, but I make sure to eat kiwi because it helps with sleep. · I sometimes feel anxious and have trouble sleeping because my brother is considering expanding his hair salon during the recession, which worries me. · My brother is a hairdresser. · I'm worried about my brother.

Table 10: The example #1 of input dialogue and its summary. The modified summary reflects the emotion and its cause.

Input dialogue and summary (translated into English)
Dialogue: Speaker1: It's been 4 weeks! Have you been doing well? I wonder if you went on a trip Speaker2: Yeah, I had a great trip. I had a blast after a long time! How have you been for the past month? Speaker1: I've been good. How was your trip to Gangneung? Was it as good as you expected? Speaker2: Yeah! It was so much fun! I even had lobster, and it was both reasonably priced and delicious. How's your pet doing? Speaker1: I'm doing well! I'm planning to go back to the vet today after about a month. Feeling a bit worried. Speaker2: Oh, if you're worried, now I'm worried too... Was it usual for you to go once a month? Speaker1: I went because I've been feeling sick recently. I got treated for it, but I still want to make the visit. Is it alright? Speaker2: I'm worried to hear that you've been feeling sick. Since you've got treated, everything should be fine. Don't worry too much. Speaker1: It's not good to worry too much, right? I guess I'll just try to clear my mind while having a meal.. Have you had your meal? Speaker2: I haven't eaten yet! I'm about to eat now. What kind of food are you thinking of having? Speaker1: I'm planning to have some soft beef soup! Do you happen to like beef soup? Speaker2: Yeah, I like various types of soups. Today, I suddenly felt like having pork cutlet, so I'm planning to have that! Speaker1: Pork cutlet sounds good! Well then, enjoy your meal, and let's talk again next time Speaker2: Sure thing! Enjoy your meal, and have a good visit to the hospital! See you next time!
Summary: [speaker1's summary] · I've been doing well. · I'm going to go back to the animal hospital again. · I'm going to try to clear my mind while having a meal. · I'm going to have some soft beef soup. [modified speaker1's summary] · I'm planning to go back to the animal hospital because I'm worried that my pet might be sick. · I think it's not good to worry too much, so I'm going to try to clear my mind while having a meal. · I'm going to have some soft beef soup.

Table 11: Example of input for Dialogue and Summary. The modified summary reflects the emotion and its cause.

Input dialogue and summary (translated into English)
Instruction: You are a language model that operates in a multi-session chatbot. In cases where the information in [{speaker}'s summary] necessitates the modification or deletion of a sentence in [Memory], you update the sentence to reflect the latest information while maintaining consistency. The sentences in [Memory] are accumulated from past conversations, whereas [{speaker}'s summary] reflects the current conversation. Therefore, I will acknowledge that [{speaker}'s summary] contains more recent information compared to [Memory] and update accordingly. You won't merge or modify sentences in [Memory] just because they're on the same topic. You'll only update sentences in [Memory] when they need to be changed based on the content of [{speaker}'s summary] or when the content continues. When updating, ensure that no new content is generated, and existing content is not lost. Referencing examples 1 and 2, merge the updated [Memory] and [{speaker}'s summary] to produce the output in example 3. ### Input: Example1: {Example 1} Example2: {Example 2} ### Output: Example3: [Memory] I am a woman in my forties. I live in Seoul. I am a traditional Korean medicine doctor. I am currently writing a paper on Oriental acupuncture. It's been just under a month since I started writing the paper. I'm ambitious and want to make my name known. I feel jealous when my friends are more successful than me. I have been commuting by public transportation for 10 years, and it's exhausting. [{speaker}'s summary] I've been busy writing my paper lately, but I've just finished writing it. [Updated memory] I am a woman in my forties. I live in Seoul. I am a traditional Korean medicine doctor. I've been busy writing a paper on Oriental acupuncture lately, and I've just finished writing it. I am ambitious and want to make my name more known. I feel jealous when my friends are more successful than me. I have been commuting by public transportation for 10 years, and commuting is tough.

Table 12: Example of Memory update instruction

Few-shot examples used for memory updates(translated into English)
[Memory] I am waiting for my car to be repaired. I work at a trading company. I am a high school student living in Yangcheon-gu. I definitely want to visit Universal Studios if I go to Japan. I haven't booked a flight to Japan yet. I haven't reconciled with my boyfriend yet. I've developed an interest in interior design and have been looking at furniture lately. I'm working overtime. I'm having doubts about marrying my boyfriend. [Summary] My car was ultimately declared a total loss. I quit my job and became a housewife after having a child. I find it hard to concentrate when taking video classes. My overtime work is finished. I moved to Guro-gu because of my father's company. I think the public transportation transfer system is good. I eventually broke up with my boyfriend. I've booked a flight and am now looking for accommodation. [Updated Memory] My car, which I had left for repairs, was ultimately declared a total loss. I worked at a trading company but became a housewife after having a child. I've developed an interest in interior design and have been looking at furniture lately. I definitely want to visit Universal Studios if I go to Japan. I used to live in Yangcheon-gu but moved to Gwangjin-gu because of my father's company. I find it hard to concentrate when taking video classes. I eventually broke up with my boyfriend, with whom I had been contemplating marriage. I think the public transportation transfer system is good. I've booked a flight to Japan and am currently looking for accommodation.

Table 13: Few-shot examples used for memory updates (Part 1)

Few-shot examples used for memory updates(translated into English)
[Memory] I am looking for a part-time job near my home. I buy a diary every year but have never used one properly. I am a woman in my 40s living in Seocho-gu. I've always been abroad on my birthday. I am a university student attending Sejong University. I tested positive for COVID-19 and am currently in quarantine. I am a speech therapist. I am planning to do volunteer work next weekend. I plan to volunteer at an animal shelter. I have been volunteering at the same animal shelter for two years. [Summary] I majored in physical education but work as a speech therapist. Yesterday was my graduation ceremony. I completed my volunteer work. I recently moved to Yangcheon-gu. I decided to work morning shifts only at a restaurant near my home. I will start my part-time job next week. I have recovered from COVID-19. [Updated Memory] I graduated from University. I buy a diary every year but have never used one properly. I am a woman in my 40s who moved from Seocho-gu to Yangcheon-gu. I've always been abroad on my birthday. Although I majored in physical education, I work as a speech therapist. I did volunteer work at an animal shelter over the weekend. I have been volunteering at the same animal shelter for two years. Starting next week, I will work part-time only in the mornings at a restaurant near my home. Although I had COVID-19, I have fully recovered.

Table 14: Few-shot examples used for memory updates (Part 2)

Input dialogue and summary (translated into English)
Instruction: You are a model that updates information by combining existing information sentences from [Memory] with new information sentences from [speaker's summary]. In this process, you must ensure that the updated information is consistent, non-redundant, and complete. The update process follows these steps 1. For each sentence in [Memory], classify its relationship with each sentence in [speaker's summary] using one of four labels: "PASS" label: The sentence from [Memory] is implied by a sentence in [speaker's summary]. "REPLACE" label: The sentence from [speaker's summary] contradicts or implies the sentence in [Memory]. "APPEND" label: The sentences from [Memory] and [speaker's summary] are neutral towards each other. "DELETE" label: Both the sentence from [Memory] and [speaker's summary] are no longer true or no longer need to be remembered. 2. If a sentence is classified as "REPLACE," delete the sentence from [Memory]. If classified as "DELETE," delete both the sentence from [Memory] and [speaker's summary]. 3. For each sentence in [speaker's summary], classify its relationship with each sentence in [Memory] using one of four labels. 4. If a sentence is classified as "PASS," delete the sentence from [speaker's summary]. 5. Merge the final versions of [Memory] and [speaker's summary]. Refer to examples 1 and 2 for guidance, and output the updated result as shown in example 3. ### Input: Example1: {Example 1} Example2: {Example 2} ### Output: Example3: [Memory] I am looking for a nearby place to work part-time. I buy a planner every year but never use it properly. I am a woman in my 40s living in Seocho-gu. I have always been abroad on my birthday. I am a university student attending Sejong University. I tested positive for COVID-19 and am currently in isolation. I am a speech therapist. I am planning to volunteer next weekend. I volunteer at an animal shelter. I have been volunteering at the same animal shelter for two years. [{speaker}'s summary] I majored in physical education but work as a speech therapist. I had my graduation ceremony yesterday. I recently moved to Yangcheon-gu.decided to work part-time in the mornings at a nearby restaurant. I start my part-time job next week.have recovered from COVID-19. [Updated] I graduated from Sejong University. I buy a planner every year but never use it properly. I used to live in Seocho-gu but recently moved to Yangcheon-gu. I have always been abroad on my birthday. I majored in physical education but now work as a speech therapist. I volunteered at an animal shelter last weekend. I have been volunteering at the same animal shelter for two years. I will start working part-time in the mornings at a nearby restaurant next week. I had COVID-19 but have recovered.

Table 15: Example of memory update instruction using CareCallmem's methodology