

# Towards a personalized LLM-based daily edge memory aid

Sangil Ha\*

*AI Lab*

*Tutorus Labs Inc.*

Deajeon, Republic of Korea

siha@tutoruslabs.com

Dongbeom Ko\*

*AI Computing System SW Research Section*

*Electronics and Telecommunications Research Institute (ETRI)*

Deajeon, Republic of Korea

dbko112@etri.re.kr

Joon-Ho Lim

*AI Lab*

*Tutorus Labs Inc.*

Deajeon, Republic of Korea

jhlim@tutoruslabs.com

Sungjoo Kang

*AI Computing System SW Research Section*

*Electronics and Telecommunications Research Institute (ETRI)*

Deajeon, Republic of Korea

sjkang@etri.re.kr

Hyeon Soo Kim<sup>†</sup>

*Department of Computer Science and Engineering*

*Chungnam National University (CNU)*

Deajeon, Republic of Korea

hskim401@cnu.ac.kr

**Abstract**—This paper introduces a personalized memory aid system that combines a high-performance cloud-based large language model (LLM) with a low-power edge device running a small language model (sLM) to enhance user productivity by processing and summarizing daily conversations in real-time. The system optimizes efficiency and user experience while ensuring privacy through localized data handling. By combining cloud and edge resources, the system provides scalable, real-time memory support without compromising data security. The research highlights the system’s architecture, its role in protecting user data, and its potential to seamlessly integrate into daily life, offering a sustainable and efficient approach to personalized artificial intelligence applications.

**Index Terms**—Personalized Large Language Model, Cloud-computing, Large Language Model, Conversational AI, Artificial Intelligence

## I. INTRODUCTION

With the recent advancements in artificial intelligence(AI), the impact of generative AI on our society has been increasing [1]. According to several reports, the personalized service market, particularly those based on language models tailored to meet specific user needs, is experiencing rapid growth. Data from various analytics firms indicate that this market is expected to continue its expansion in the future [2] [3]. In this context, this paper explores how conversational AI can be utilized as a daily memory aid for individuals. Specifically, it focuses on providing personalized information based on user data while ensuring security and proposes a method to integrate this into everyday devices.

Traditional personalized language model services have been provided by fine-tuning large language models (LLMs) using personalized data. However, this process presents two significant issues: 1) The need to allocate resource-intensive LLMs for each individual, and 2) The exposure of personal data to the cloud during the fine-tuning process. To address

these issues, this paper proposes a solution by combining edge devices with small language model (sLM) and cloud-based LLMs [4] [5]. In this system, the user’s personal data is analyzed and summarized in real-time on the edge device using the low-spec language model such as sLM, and this refined data is then used for personalization. The cloud-based generalized LLM generates conversations based on the refined personal data from the edge device, thereby compensating for the limitations of the low-spec language model. This edge-cloud collaborative personalized model service mitigates the drawbacks of existing systems, optimizing system efficiency and enhancing user experience.

This paper aims to present the architectural design and implementation methods of the proposed technology, as well as a sustainable and reliable model for technology integration.

## II. RELATED WORK

The development of AI systems as memory aids is increasingly adopting a more personalized approach. Recent studies explore the potential of personalized memory aid systems that leverage LLMs. Zhang et al. [6] proposed a method for coordinating users’ short-term and long-term memory in the medical field by implementing personalized AI using the Dual Process Enhanced Memory (DPeM) mechanism and Partial Fine-Tuning (PEFT) [7]. This research provides insights into how LLMs can reflect individual data and needs to deliver tailored medical information, highlighting current trends in personalized AI.

Traditional research has primarily employed PEFT to deliver personalized LLMs to users [7]. However, this approach necessitates allocating a customized LLM for each user, leading to inefficient use of expensive GPU resources. This issue is particularly problematic when supporting a large number of users simultaneously, as it results in significant hardware resource wastage. Additionally, in the case of these personalized

\*These authors contributed equally to this work.

<sup>†</sup>Corresponding author.

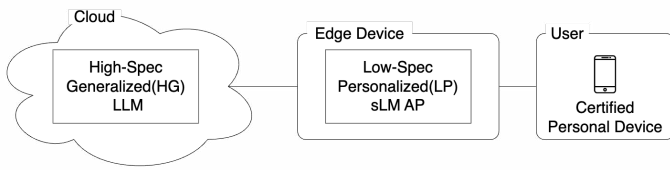


Fig. 1. Deployment Diagram.

LLMs, the provision of expensive GPU resources from the cloud could pose a risk of personal information leakage.

One notable aspect of these studies is the importance of privacy protection and hardware efficiency in personalized LLM-based systems. Ensuring the security of personal data while delivering tailored information to users remains a technical challenge on the cloud. Deploying high-spec LLM for each user can cause a waste of resources. Against this backdrop, this research proposes a new approach to designing and implementing personalized memory aid systems that can provide practical assistance to users.

To address these challenges, this study proposes an approach that utilizes a combination of an edge-based low-spec personalized sLM (LP-sLM) and a cloud-based high-spec generalized LLM (HG-LLM) to efficiently serve multiple users. Instead of customizing LLMs for each individual, this method leverages a HG-LLM served from the cloud, maximizing hardware resource efficiency. And by locating LP-sLM to the edge, we can reduce the risk of personal information leakage. The goal of this research is to build a system that minimizes the use of costly GPU memory and the risk of personal information leakage while still providing personalized memory aid functions tailored to user needs.

### III. WHAT IS PERSONALIZED MEMORY AID

The personalized memory aid system proposed in this study collects raw conversation data from user interactions and converts it into long-term memory using a relatively small-scale, LP-sLM such as Llama3-8b [8]. When the user later inquires about a specific memory or asks a question, the system generates a detailed response based on previous long-term memory by utilizing a HG-LLM like Llama3.1-405b located in the cloud [8]. In this process, the user’s personal conversation data is managed on the edge device rather than in the cloud, and it is temporarily transferred to the cloud only when needed to generate a response for a specific “Memory”. The data is immediately deleted after the response is generated.

This section discusses the components of the personalized memory aid system, the roles of each component, the integration between cloud and edge devices, and strategies for ensuring security and privacy protection.

#### A. System Composition

The system proposed in this study is composed of a combination of a cloud-based HG-LLM and a LP-sLM installed on the user’s personal device. As shown in Fig. 1, this system

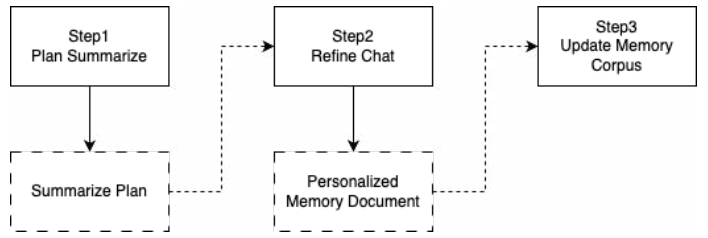


Fig. 2. Summarization Process

consists of the cloud, edge devices, and the user’s personal device, with each component performing the following roles:

- **Cloud (HG-LLM):** HG-LLM is located in the cloud, providing complex conversation analysis and response generation services to multiple users. This model has the capability to generate personalized responses using the user’s personalized memory, aiming for efficient use of hardware resources.
- **Edge Device (LP-sLM AP):** The edge device performs roles such as collecting user conversation data, converting it into long-term memory through summarization, and retrieving memory. Equipped with a LP-sLM, it processes the user’s conversation data locally, ensuring privacy protection and guaranteeing fast responses. And this module similarly to a network access point (AP) within the memory aid system.
- **Certified Personal Device (User Device):** To protect user data and privacy, the user interacts with the system through a pre-certified personal device. The personal device allows the collection of conversation data that the user has approved during interactions.

#### B. Memory Aid Process

The operation of the system can be divided into two main flows: the memory summarization flow and the memory retrieval flow. Each flow is explained with reference to the attached diagrams. Fig. 2 illustrates the memory summarization flow, while Fig. 3 shows the memory retrieval flow. The memory summarization flow operates automatically in the background when the chat history accumulates beyond a certain number of entries. The memory retrieval Process activates with each user input.

##### 1) Memory Summarize Flow (Summarization Process):

- **Step 1 - Plan Summarize:** Once a sufficient amount of user conversation data has been collected, the Summarize Planning Module formulates a plan to summarize the conversation content.
- **Step 2 - Refine Chat:** The Refine Module refines the conversation content according to the summarization plan. During this process, unimportant conversations are filtered out, leaving only key information such as the “5Ws and 1H” (who, what, when, where, why, and how) and future plans.
- **Step 3 - Update Memory Corpus:** The refined conversation content is stored in the Memory Corpus, preserving

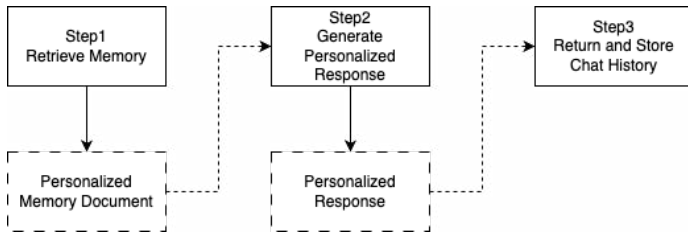


Fig. 3. Memory Retrieval Process

it as part of the user’s long-term memory. This allows for the retrieval of relevant information in future conversations.

### 2) Memory Search Flow (Memory Retrieval Process):

- **Step 1 - Retrieve Memory:** When a new conversation begins, the retrieve module of the memory aider searches the Memory Corpus for relevant information.
- **Step 2 - Generate Personalized Response:** Based on the retrieved information, the conversation module generates a personalized response. This response includes customized information related to the user’s past conversation history.
- **Step 3 - Return and Store Chat History:** The generated response is delivered to the user, and simultaneously, the conversation history is stored by the chat history module. This stored data is later used for summarization.

### C. System Architecture

The core modules of this system consist of the chat-manager, memory aider, and LP-sLM, with each module performing the roles described below. Refer to the Fig. 4 diagram for a clearer understanding of the system architecture.

1) *Chat-Manager*: It acts as the main class of the entire system, controlling the overall flow of the conversation. It provides a chat interface for interacting with the user and manages the chat history.

- **Conversation Module:** Analyzes the user’s real-time conversations and generates appropriate responses. It processes complex conversation content with the support of the HG-LLM in the cloud.
- **Chat History Module:** Manages the user’s conversation history. This module operates on the personal device, enhancing privacy protection.

2) *Memory Aider*: In the system proposed by this study, this module plays a crucial role by either converting simple conversation logs into long-term memory or retrieving and returning the converted long-term memory. For details on summary prompting, examples of raw conversation data, and results of long-term memory please refer to the Appendix.

- **Retrieve Module:** Searches the Memory Corpus to provide information related to the current conversation. This plays a crucial role in understanding the context of the conversation and generating appropriate responses.
- **Summarize Planning Module:** Summarizes the conversation content and memory data to select key information

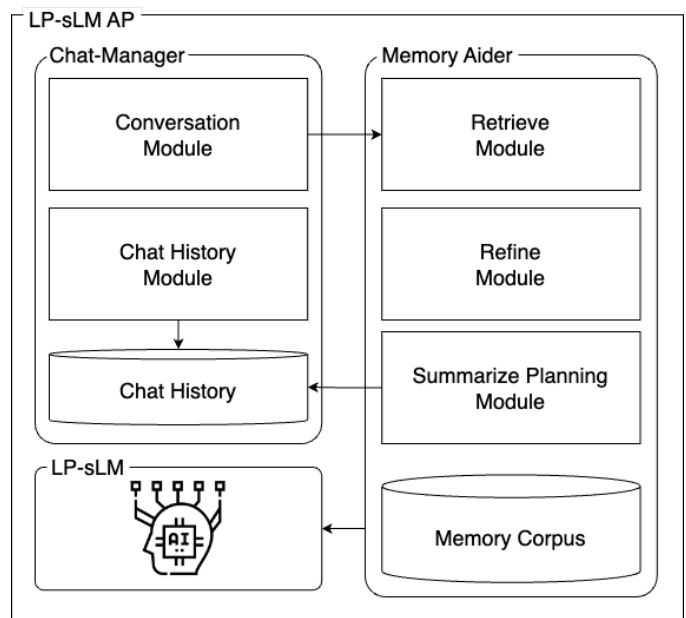


Fig. 4. System Architecture

for long-term memory storage. During this process, a personalized summary is created based on the user’s interests and needs.

- **Refine Module:** Aligns with the summary plan generated by the Summarize Planning Module to summarize the user’s conversation history, making it easier to understand, and removing unnecessary information for documentation.

3) *LP-sLM*: Installed on the personal device, the LP-sLM can summarize the user’s conversation history in conjunction with the Refine Module, even without the support of the cloud-based HG-LLM. In this study, the Llama3-8b-Instruct model is used [8].

### D. Integration of Cloud and Edge Device

The system proposed in this study efficiently integrates a cloud-based high-spec LLM with a sLM on edge devices to provide personalized memory aid services and personalized memory-based assistance to multiple users in real-time. In the cloud, the HG-LLM generates tailored responses for users by leveraging the augmented memory provided by the memory aider.

On the other hand, the LP-sLM installed on the personal device handles local processing of user data, directly reviewing the user’s conversation history and converting it into refined data.

This structure enables the robust functionality of the HG-LLM to support users’ daily memory needs while securely protecting their personal data. Furthermore, efficient data communication between the cloud and edge devices optimizes overall system resource usage and minimizes hardware waste.

### E. Security and Privacy Protection

In personalized memory aid systems, security and privacy protection are among the most important factors. This system manages user data locally on edge devices, reducing the risk of data breaches and gaining user trust. Moreover, as most data processing occurs on personal devices (Edge devices), the likelihood of sensitive user information being leaked externally is minimized. Additionally, when generating requests using HG-LLM, various preprocessing methods, such as masking, can be applied to protect users' personal information.

## IV. IMPLEMENTATION

In this section, we focus on the implementation details related to the summary process within the proposed system. As specified in the system architecture, the summary flow operates collaboratively between the refine module and the LP-sLM within the LP-sLM AP. For this study, the Llama3-8b-Instruct model was selected as the LP-sLM [8].

To enable the Llama3-8b-Instruct model to perform the summary task effectively, we crafted a prompt based on the prompt construction guidelines provided by the Llama3 model documentation and Bsharat et al.(2023) [9] [11]. The detailed prompt is available in the appendix. The primary goal of this prompt is to summarize conversations in alignment with the "5Ws and 1H" (who, what, when, where, why, and how) principles, condensing them into 3 to 4 sentences while including essential information about subsequent actions that need to be taken. Additionally, since the purpose of the summary task is focused on documentation, the output is structured in JSON format.

### A. Summarize and Cost Efficiency

When the summary task was executed using the prompt crafted for the aforementioned objectives, the token count was reduced by an average of 26.97%, as shown in Fig. 5.

Given the pricing model provided by OpenAI for token usage [10], the cost for input tokens is \$5.00 per million tokens, and the cost for output tokens is \$15.00 per million tokens. If the token count is reduced by 26.97%, significant cost savings can be achieved in both input and output token usage, thereby enhancing the overall cost efficiency of the system.

### B. Memory Aid Generation

Table I presents data retrieved from the memory corpus based on the questions, along with the contrasting generation results produced by GPT-4o as a HG-LLM and Llama3-8b-Instruct as a LP-sLM. The generation prompt used for this table was originally designed for Llama3, with special tokens excluded when using GPT. As demonstrated, the responses generated by GPT-4o are more specific and structured compared to those generated by Llama3-8b-Instruct, despite the latter being a low-parameter LLM.

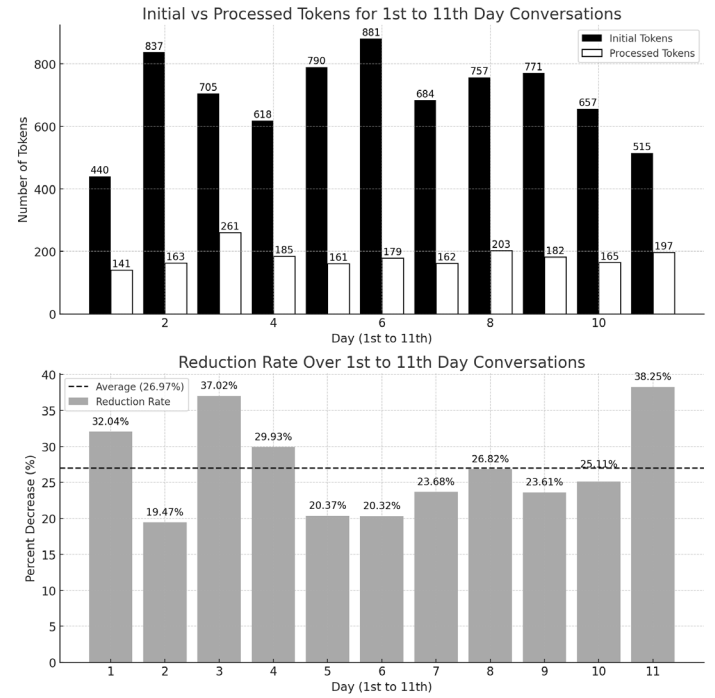


Fig. 5. Token Reduction for Daily Conversations Summarization

## V. CONCLUSION

In this study, a combination of a cloud-based HG-LLM and a LP-sLM installed on edge devices is proposed to implement a personalized memory aid system. This system utilizes the HG-LLM in the cloud to perform complex personalized response generation using the Memory Corpus, while the LP-sLM on the edge device handles personalized data processing to provide real-time customized memory aid services to users.

The proposed system is designed to optimize the use of hardware resources, ensuring high performance even when multiple users are simultaneously using the system. A significant advantage of this approach is the ability to provide high-quality, personalized services while protecting user data privacy through the collaboration between the cloud and edge devices.

The key contributions of this study are as follows:

- **Efficient Resource Utilization:** By combining the cloud-based high-spec LLM with the low-spec LLM on edge devices, an efficient system was designed that reduces the waste of expensive hardware resources and can support multiple users simultaneously. Additionally, by actively leveraging summarization features, we reduced the number of tokens stored in the Memory Corpus by an average of 26.97%, significantly enhancing the overall cost-efficiency of the system.
- **Enhanced Security and Privacy:** By conducting most data processing on edge devices, the system architecture was implemented to protect user privacy by preventing external exposure of user data.

TABLE I  
COMPARISON OF GENERATED RESPONSES FOR MEMORY AID

<b>Question</b>	I feel nervous what can I do for this feeling?
<b>Generation Prompt</b>	<pre>&lt; begin_of_text &gt;&lt; start_header_id &gt;system&lt; end_header_id &gt; Your personal assistant. Generate an answer to the given question based on the provided document.It is important to carefully read and understand the criteria for generation.&lt; eot_id &gt; &lt; start_header_id &gt;user&lt; end_header_id &gt;  Generation Criteria: 1 - You are a personal assistant that is well-versed in the user’s life records. 2 - The answer should be generated in a summarized form based on the document, with supporting evidence. 3 - The output should be generated in the format '### Answer:'. 4 - It is very important to provide clear evidence to support the answer.  Document: {document}  Question: {question}  }&lt; eot_id &gt;</pre>
<b>Retrieved Memory Corpus</b>	<pre>{'when': '2024-02-01, 10AM, Thu', 'how': 'Through conversation and guidance from doctor', 'who': 'Me and Doctor', 'where': 'Therapy room', 'why': 'To discuss and improve mental health and well-being', 'what': 'Initial conversation about mental health and well-being', 'summary': 'The user mentioned feeling depressed and unmotivated, and the doctor asked about stressors and sleep patterns. The user reported poor sleep and lack of exercise, and the doctor suggested trying therapy and medication. The user expressed concerns about medication but agreed to try therapy.', 'futureWork': 'The next therapy session is scheduled for 2024-02-05, 10AM, Monday.'}</pre>
<b>Chat GPT-4o Response</b>	<p>Based on your recent conversation with the doctor on 2024-02-01, where you discussed your mental health and well-being, here are a few suggestions to manage your nervousness:</p> <ol style="list-style-type: none"> <li><b>Therapy</b>: Since you have agreed to try therapy and have a session scheduled for 2024-02-05, you can bring up your feelings of nervousness with your therapist. They can provide strategies and tools tailored to help you manage these feelings.</li> <li><b>Sleep and Exercise</b>: The doctor noted that you reported poor sleep and lack of exercise. Improving your sleep patterns and incorporating regular exercise can significantly reduce feelings of nervousness. Try to establish a bedtime routine and find a physical activity you enjoy.</li> <li><b>Stress Management Techniques</b>: During your session, the doctor asked about stressors. Identifying and addressing these stressors through relaxation techniques like deep breathing, meditation, or journaling can help alleviate nervousness.</li> </ol> <p>These steps were suggested by your doctor during your initial conversation about your mental health, providing a clear path forward based on professional guidance.</p>
<b>Llama3-8b-Instructed Response</b>	<p>Based on your previous therapy session on 2024-02-01, you mentioned feeling depressed and unmotivated. During the session, you reported poor sleep and lack of exercise, which may be contributing to your nervous feelings. The doctor suggested trying therapy and medication to help improve your mental health and well-being. Although you expressed concerns about medication, you agreed to try therapy.</p>

- Real-time Personalized Memory Aid: A personalized memory aid function was proposed that analyzes user conversations in real time and provides timely information through summarization and search features.

The system proposed in this study suggests the potential for effectively managing data generated from everyday conversations, thereby maximizing user experience. Future research will focus on applying this system to various application domains and further optimizing its performance. Additionally, the goal is to develop a memory aid system that more users can easily access and use by improving the user interface and interaction methods.

#### ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program-(Alchemist Project, 20025013) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

#### REFERENCES

- [1] Fortune Business Insights. (2023). Generative AI Market Size, Share & Trends Analysis Report. Fortune Business Insights. Retrieved from <https://www.fortunebusinessinsights.com/generative-ai-market-107837>
- [2] SkyQuest Technology Consulting Pvt. Ltd. (2023). Artificial intelligence-based personalization market. SkyQuest. <https://www.skyquestt.com/report/artificial-intelligence-based-personalization-market>
- [3] Future Data Stats. (2023). Artificial intelligence in personalized marketing market. Future Data Stats. <https://www.futuredatastats.com/artificial-intelligence-in-personalized-marketing-market>
- [4] Zapier. (2024). What are small language models?. Zapier. <https://zapier.com/blog/what-are-small-language-models>
- [5] SuperAnnotate. (2024). Small Language Models (SLMs) [2024 overview]. SuperAnnotate. <https://www.superannotate.com/blog/small-language-models-2024-overview>
- [6] Zhang, K., Kang, Y., Zhao, F., Liu, X. (2024, June). LLM-based Medical Assistant Personalization with Short-and Long-Term Memory Coordination. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 2386-2398).
- [7] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C. A. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35, 1950-1965.
- [8] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- [9] Meta. (n.d.). Model cards and prompt formats - Meta LLaMA-3. Retrieved August 29, 2024, from <https://llama.meta.com/docs/model-cards-and-prompt-formats/llama-3>
- [10] OpenAI. (n.d.). Pricing. Retrieved August 29, 2024, from <https://openai.com/api/pricing/>
- [11] Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2023). Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*.

## APPENDIX

**TABLE II**  
**MEMORY AID SUMMARIZATION**

<b>Summarization Prompt</b>	<pre> &lt; begin_of_text &gt;&lt; start_header_id &gt;system&lt; end_header_id &gt; Your personal assistant. summarize upper messages for feature conversation.&lt; eot_id &gt; &lt; start_header_id &gt;user&lt; end_header_id &gt;  messages = {chat_history}  ## Generation Criteria: 1 - "human" is me. 2 - The summary must include when, how, who, where, why, what. 3 - And the summary must include a summary of the conversation within 3~4 sentences. 4 - Future plans should also be included. 5 - You must summarize each day. 6 - Output should be in the same Json Format as below. 7 - There should be no output other than Json.  ## Principle of generation: summary: Summarize within 3~4 sentences without any additional comments futureWork: Write what to do or Next therapy schedule.  Json Format: [ {   "when": "2024-02-23, 1PM, Friday", "how": "Through conversation and guidance from doctor", "who": "Me and Doctor",   "where": "Online", "why": "To continue discussing and improving mental health and well-being",   "what": "Follow-up conversation about mental health and well-being",   "summary": "The user mentioned that they have continued with mindfulness exercises and light workouts, and   their sleep patterns have slightly improved. They are trying to manage their workload better and plan to   attend a support group. The user expressed nervousness but aimed to approach it with a positive attitude.   They received advice to participate with an open mind.",   "futureWork": "The next reservation schedule is 2024-02-28, 10AM, Wednesday." } ]&lt; eot_id &gt; &lt; start_header_id &gt;assistant&lt; end_header_id &gt; </pre>
<b>Raw Chat Data</b>	<pre> ('system', 'Date: 2024-02-01, Time: 10AM, Day: Thu') ('human', "I've been feeling really depressed and unmotivated lately.") ('doctor', "I'm sorry to hear that. Have you experienced any significant stressors or changes in your routine recently?") ('human', "Nothing major, but I've been really tired because of work. I haven't been sleeping well.") ('doctor', "Poor sleep can greatly affect your mood. What is your current sleep pattern like? How many hours are you sleeping?") ('human', "I only sleep about 4-5 hours a night. Even when I sleep, I don't feel rested.") ('doctor', "That's concerning. Lack of sleep can worsen feelings of depression. What activities do you usually engage in? Do you exercise or have any hobbies?") ('human', "I don't really have any hobbies, and I don't exercise. After work, I just rest at home.") ('doctor', "Engaging in exercise or hobbies can be very beneficial for improving your mood. How about trying to incorporate some light exercise or activities you enjoy?") ('human', "I'll try, but it's hard to find the motivation right now.") ('doctor', "I understand. When motivation is low, it's helpful to start with small goals. Have you considered therapy or medication?") ('human', "I've never tried therapy. I'm a bit scared of taking medication.") ('doctor', "Therapy has helped many people, and medication can be very effective if needed. Don't be afraid to explore these options. How about starting with therapy?") ('human', "Okay, I'll give therapy a try. Thank you, doctor.") ('doctor', "You're welcome. Let's schedule your first therapy session. How does next Monday at 10 AM sound?") ('human', "That works for me.") ('doctor', "Great. I have you scheduled for Monday at 10 AM. See you then.") </pre>
<b>Retrieved Memory Corpus</b>	<pre> {'when': '2024-02-01, 10AM, Thu', 'how': 'Through conversation and guidance from doctor', 'who': 'Me and Doctor', 'where': 'Therapy room', 'why': 'To discuss and improve mental health and well-being', 'what': 'Initial conversation about mental health and well-being', 'summary': 'The user mentioned feeling depressed and unmotivated, and the doctor asked about stressors and sleep patterns. The user reported poor sleep and lack of exercise, and the doctor suggested trying therapy and medication. The user expressed concerns about medication but agreed to try therapy.', 'futureWork': 'The next therapy session is scheduled for 2024-02-05, 10AM, Monday.'} </pre>