MemoryBank: Enhancing Large Language Models with Long-Term Memory

Wanjun Zhong[1], **Lianghong Guo**[1], **Qiqi Gao**[2], **He Ye**[3], **Yanlin Wang**[1]
[1] Sun Yat-Sen University [2] Harbin Institute of Technology
[3] KTH Royal Institute of Technology
{zhongwj25@mail2, wangylin36@mail}.sysu.edu.com
2231612405@qq.com, 18b903026@stu.hit.edu.cn
heye@kth.se

Abstract

Revolutionary advancements in Large Language Models (LLMs) have drastically reshaped our interactions with artificial intelligence (AI) systems, showcasing impressive performance across an extensive array of tasks. Despite this, a notable hindrance remains—the deficiency of a long-term memory mechanism within these models. This shortfall becomes increasingly evident in situations demanding sustained interaction, such as personal companion systems, psychological counseling, and secretarial assistance. Recognizing the necessity for long-term memory, we propose MemoryBank, a novel memory mechanism tailored for LLMs. MemoryBank enables the models to summon relevant memories, continually evolve through continuous memory updates, comprehend, and adapt to a user's personality over time by synthesizing information from previous interactions. To mimic anthropomorphic behaviors and selectively preserve memory, MemoryBank incorporates a memory updating mechanism, inspired by the Ebbinghaus Forgetting Curve theory. This mechanism permits the AI to forget and reinforce memory based on time elapsed and the relative significance of the memory, thereby offering a more human-like memory mechanism and enriched user experience. MemoryBank is versatile in accommodating both closed-source models like ChatGPT and open-source models such as ChatGLM. To validate MemoryBank's effectiveness, we exemplify its application through the creation of an LLM-based chatbot named SiliconFriend in a long-term AI Companion scenario. Further tuned with psychological dialog data, SiliconFriend displays heightened empathy and discernment in its interactions. Experiment involves both qualitative analysis with real-world user dialogs and quantitative analysis with simulated dialogs. In the latter, ChatGPT acts as multiple users with diverse characteristics and generates long-term dialog contexts covering a wide array of topics. The results of our analysis reveal that SiliconFriend, equipped with MemoryBank, exhibits a strong capability for long-term companionship as it can provide emphatic response, recall relevant memories and understand user personality. This underscores the effectiveness of MemoryBank[1].

1 Introduction

The advent of Large Language Models (LLMs) such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) has led to increasing influence across various sectors, from education and healthcare to customer service and entertainment. These powerful AI systems have demonstrated a remarkable ability to

[1]The materials are released in `https://github.com/zhongwanjun/MemoryBank-SiliconFriend`. The corresponding author is Yanlin Wang.

understand and generate human-like responses. Despite the remarkable capabilities of LLMs, a key limitation is their lack of long-term memory, an essential aspect of human-like communication, particularly noticeable in scenarios requiring sustained interactions like personal companionship, psychological counseling, and secretarial tasks. Long-term memory in AI is vital to maintain contextual understanding, ensure meaningful interactions and understand user behaviors over time. For instance, personal AI companions need to recall past conversations for rapport building. In psychological counseling, an AI can provide more effective support with knowledge of the user's history and past emotional states. Similarly, secretarial AI requires memory for task management and preference recognition. The absence of long-term memory in LLMs hinders their performance and user experience. Therefore, it is essential to develop AI systems with improved memory capabilities for a more seamless and personalized interaction.

Therefore, we introduce MemoryBank, a novel mechanism designed to provide LLMs with the ability to retain long-term memory and draw user portraits. MemoryBank enables LLMs to recall historical interactions, continually evolve their understanding of context, and adapt to a user's personality based on past interactions, thereby enhancing their performance in long-term interaction scenarios. Inspired by the Ebbinghaus Forgetting Curve theory, a well-established psychological principle that describes how the strength of memory decreases over time, MemoryBank further incorporates a dynamic memory mechanism closely mirroring human cognitive process. This mechanism empowers the AI to remember, selectively forget, and strengthen memories based on time elapsed, offering more natural and engaging user experience. Specifically, MemoryBank is built on a memory storage with memory retrieval and updating mechanism, and ability to summarize past events and users' personality.

MemoryBank is versatile as it can accommodate both closed-source LLMs like ChatGPT and open-source LLMs like ChatGLM (Zeng et al., 2022) or BELLE (Yunjie Ji & Li, 2023).

To exemplify the practical implications of MemoryBank, we develop SiliconFriend, an LLM-based AI Companion chatbot integrated with this innovative memory mechanism. SiliconFriend is designed to retain and reference past interactions, reinforcing the transformative influence of MemoryBank in crafting a more personable AI companion. A distinctive features of SiliconFriend is its tuning with 38k psychological conversations, collected from various online sources, which enables it to exhibit empathy, carefulness, and provide useful guidance, making it adept at handling emotionally charged dialogues. Moreover, one of the standout capabilities of SiliconFriend is to understand a user's personality by summarizing from past interactions, which empowers it to tailor responses to the user's individual traits, thereby enhancing user experience. Additionally, SiliconFriend supports bilingual functionality, catering to users who communicate in English and Chinese. This multi-language support broadens its accessibility and usability across different user groups. SiliconFriend is implemented with two open-source models, ChatGLM and BELLE, along with one closed-source model, ChatGPT, showcasing the versatility of MemoryBank in accommodating different LLMs.

To evaluate the effectiveness of MemoryBank, we conduct evaluations covering both qualitative and quantitative analyses, where the former involves real-world user dialogs and the latter employs simulated dialogs. For the quantitative analysis, we create a memory storage consisting of 10 days of conversations encompassing a diverse range of topics. These conversations involve 15 virtual users with diverse personalities, for which ChatGPT plays the role of users and generates dialog contexts according to their personalities. Based on this memory storage, we design 194 probing questions to assess whether the model could successfully recall pertinent memories and provide appropriate responses. Experiment results showcase the capabilities of SiliconFriend in memory recall, provision of empathetic companionship, and understanding of user portraits. These findings corroborate the potential of MemoryBank to significantly improve the performance of LLMs in long-term interaction scenarios. In this paper, we summarize the key contributions as follows:

- We introduce MemoryBank, a novel human-like long-term memory mechanism, which enables LLMs to store, recall, update memory, and draw user portrait.
- We demonstrate the practical applicability of MemoryBank through SiliconFriend, an LLM-based AI companion equipped with MemoryBank and tuned with psychological dialogs. It can recall past memories, provide empathetic companionship, and understand user behaviors.
- We show the generalizability of MemoryBank in three key aspects: (1) Accommodation of both open-source and closed-source LLMs; (2) Bilingual ability in both Chinese and English; (3) Applicability with and without memory forgetting mechanism.

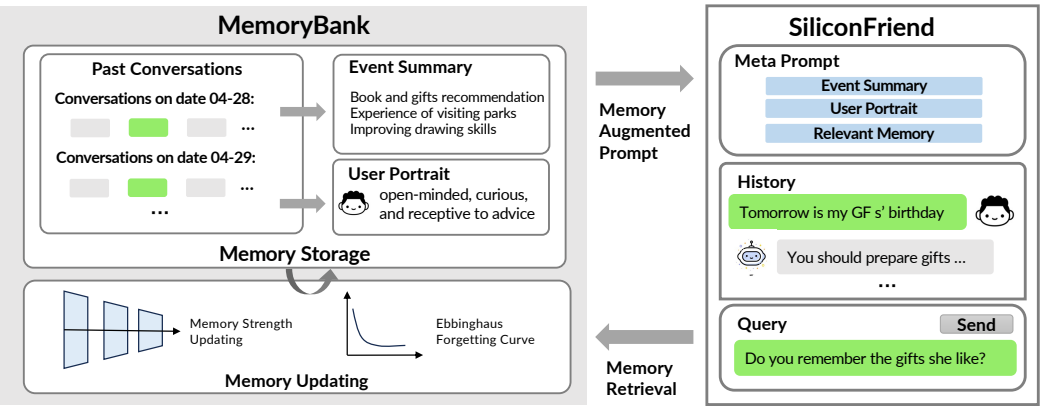

Figure 1: Overview of MemoryBank. The memory storage (§ 2.1) stores past conversations, summarized events and user portraits, while the memory updating mechanism (§ 2.3) updates the memory storage. Memory retrieval (§ 2.2) recall relevant memory. SiliconFriend (§ 3) serves as an LLM-based AI companion augmented with MemoryBank.

2 MemoryBank: A Novel Memory Mechanism Tailored for LLMs

In this section, we provide a detailed description of MemoryBank, our novel memory mechanism designed for LLMs. As shown in Fig. 1, MemoryBank is a unified mechanism structured around three central pillars: (1) a memory storage (§ 2.1) serving as the primary data repository, (2) a memory retriever (§ 2.2) for context-specific memory recollection, and (3) a memory updater (§ 2.3) drawing inspiration from the Ebbinghaus Forgetting Curve theory, a time-tested psychological principle pertaining to memory retention and forgetting.

2.1 Memory Storage: The Warehouse of MemoryBank

Memory storage, the warehouse of MemoryBank, is a robust data repository holding a meticulous array of information. As shown in Fig. 1, it stores daily conversations records, summaries of past events, and evolving assessments of user personalities, thereby constructing a dynamic and multi-layered memory landscape.

In-Depth Memory Storage: MemoryBank's storage system captures the richness of AI-user interactions by recording multi-turn conversations in a detailed, chronological fashion. Each piece of dialogue is stored with timestamps, creating an ordered narrative of past interactions. This detailed record not only aids in precise memory retrieval but also facilitates the memory updating process afterwards, offering a detailed index of conversational history.

Hierarchical Event Summary: Reflecting the intricacies of human memory, MemoryBank goes beyond mere detailed storage. It processes and distills conversations into a high-level summary of daily events, much like how humans remember key aspects of their experiences. We condense verbose dialogues into a concise daily event summary, which is further synthesized into a global summary. This process results in a hierarchical memory structure, providing a bird's eye view of past interactions and significant events. Specifically, taken previous daily conversations or daily events as input, we ask the LLMs to summarize daily events or global events with the prompt "*Summarize the events and key information in the content* `[dialog/events]`".

Dynamic Personality Understanding: MemoryBank focuses on user personality understanding. It continuously assesses and updates these understandings with the long-term interactions and creates daily personality insights. These insights are further aggregated to form a global understanding of the user's personality. This multi-tiered approach results in an AI companion that learns, adapts, and tailors its responses to the unique traits of each user, enhancing user experience. Specially, taken the daily conversations or personality analysis, we ask the LLM to deduce with prompts: "*Based on the following dialogue, please summarize the user's personality traits and emotions.* `[dialog]`" or "*The*

following are the user's exhibited personality traits and emotions throughout multiple days. Please provide a highly concise and general summary of the user's personality`[daily Personalities]`".

2.2 Memory Retrieval

Built on the robust infrastructure of memory storage, our memory retrieval mechanism operates akin to a knowledge retrieval task. In this context, we adopt a dual-tower dense retrieval model similar to Dense Passage Retrieval (Karpukhin et al., 2020). In this paradigm, every turn of conversations and event summaries is considered as a memory piece m, which is pre-encoded into a contextual representation h_m using the encoder model $E(\cdot)$. Consequently, the entire memory storage M is pre-encoded into $\boldsymbol{M} = \{\boldsymbol{h}_m^0, \boldsymbol{h}_m^1, ...\boldsymbol{h}_m^{|M|}\}$, where each h_m is a vector representation of a memory piece. These vector representations are then indexed using FAISS (Johnson et al., 2019) for efficient retrieval. Parallel to this, the current context of conversation c is encoded by $E(\cdot)$ into $\boldsymbol{h}_c$, which serves as the query to search M for the most relevant memory. In practice, the encoder $E(\cdot)$ can be interchanged to any suitable model.

2.3 Memory Updating Mechanism

With the persistent memory storage and the memory retrieval mechanism discussed in § 2.1 and § 2.2, the memorization capability of LLMs can be greatly enhanced. However, for scenarios that expect more anthropopathic memory behavior, memory updating is needed. These scenarios include AI companion, virtual IP, etc. Forgetting less important memory pieces that are long time ago and have not been recalled much can make the AI companion more natural.

Our memory forgetting mechanism is inspired from Ebbinghaus Forgetting Curve theory and follow the following principle rules[2]:

- **Rate of Forgetting.** Ebbinghaus found that memory retention decreases over time. He quantified this in his forgetting curve, showing that information is lost rapidly after learning unless it is consciously reviewed.
- **Time and Memory Decay.** The curve is steep at the beginning, indicating that a significant amount of learned information is forgotten within the first few hours or days after learning. After this initial period, the rate of memory loss slows down.
- **Spacing Effect.** Ebbinghaus discovered that relearning information is easier than learning it for the first time. Regularly revisiting and repeating the learned material can reset the forgetting curve, making it less steep and thereby improving memory retention.

The Ebbinghaus forgetting curve is expressed using an exponential decay model: $R = e^{-\frac{t}{S}}$, where R is the memory retention, or what fraction of the information can be retained. t is the time elapsed since learning the information. e is approximately equal to 2.71828. S is the memory strength, which changes based on factors such as the depth of learning and the amount of repetition. To simply memory updating process, we model S as a discrete value and initialize it with 1 upon its first mention in a conversation. When a memory item is recalled during conversations, it will persist longer in memory. We increase S by 1 and reset t to 0, hence forget it with a lower probability.

It is important to note that this is an exploratory and highly simplified memory updating model. Real-life memory processes are more complex and can be influenced by a variety of factors. The forgetting curve will look different for different people and different types of information. In summary, MemoryBank weaves together these critical components to form a more comprehensive memory management system for LLMs. It enhances their ability to provide meaningful and personalized interactions over extended periods, opening up new possibilities for AI applications.

3 SiliconFriend: An AI Chatbot Companion Powered by MemoryBank

To demonstrate the practicality of MemoryBank in the field of long-term personal AI companionship, we create an AI chatbot named SiliconFriend. It is designed to serve as an emotional companion for users, recalling pertinent user memories, and understanding users' personalities and emotional states.

[2]While Ebbinghaus Forgetting Curve theory includes additional features such as *overlearning* and *meaningful material effect* , our paper focuses on simulating the listed three principle rules.

Our implementation demonstrates adaptability by integrating three powerful LLMs that originally lack long-term memory and specific adaptation to the psychology domain. **1) ChatGPT** (OpenAI, 2022), a closed-source conversation model built by OpenAI, is a proprietary conversational AI model known for its ability to facilitate dynamic and interactive conversations. This model is trained on vast amount of data and further fine-tuned with reinforcement learning from human feedback. This approach enables ChatGPT to generate responses that are not only contextually appropriate but also closely align with human conversational expectations. **2) ChatGLM** (Zeng et al., 2022): ChatGLM is an open-source bilingual language model founded on the General Language Model (GLM) framework. This model is characterized by its 6.2 billion parameters and its specific optimization for Chinese dialogue data. The model's training involves processing approximately one trillion tokens of Chinese and English text, supplemented by supervised fine-tuning, feedback bootstrap, and reinforcement learning with human feedback. **3) BELLE** (Yunjie Ji & Li, 2023): BELLE is an open-source bilingual language model that is continuously fine-tuned from 7B LLaMA (Touvron et al., 2023). BELLE's feature is its automated instruction data synthesis using ChatGPT, which enhances its Chinese conversation ability.

The development of SiliconFriend is divided into two stages. The first stage (only for open-source LLMs) involves parameter-efficient tuning of the LLM with psychological dialogue data. This step is crucial as it allows SiliconFriend to offer useful and empathetic emotional support to users, mirroring the understanding and compassionate responses one would expect from a human companion. The second stage is to integrate MemoryBank into SiliconFriend, thereby instilling it with a robust memory system. MemoryBank allows the chatbot to retain, recall, and leverage past interactions and user portrait, providing a richer, more personalized user experience.

Parameter-efficient Tuning with Psychological Dialogue Data: The initial stage of SiliconFriend's development involves tuning the LLMs using a dataset of 38k psychological dialogues. This data, parsed from online sources, comprises a range of conversations that cover an array of emotional states and responses. This tuning process enables SiliconFriend to understand and respond to emotional cues effectively, mimicking the empathy, understanding, and support of a human companion. It equips the AI with the ability to handle emotionally guided conversations with psychological knowledge, provide meaningful emotional support to users based on their emotional state.

To adapt LLMs to scenarios with limited computational resources, we utilize a computation-efficient tuning approach, known as the Low-Rank Adaptation (LoRA) method (Hu et al., 2021). LoRA significantly reduces the quantity of trainable parameters by learning pairs of rank-decomposition matrices, while keeping the original weights frozen. Formally, consider a linear layer defined as $y = Wx$ with weight W. LoRA modifies this into $y = Wx + BAx$, where $W \in \mathcal{R}^{d\times k}$, $B \in \mathcal{R}^{d\times r}$, $A \in \mathcal{R}^{r\times k}$, and $r \ll \min(d, k)$. This method greatly reduce amount of parameters need to be learned, which is crucial for efficiency in resource-limited scenarios. We set LoRA rank r as 16 and train the model for 3 epochs with an A100 GPU.

Noting that this stage is only conducted for open-source LLMs like ChatGLM and BELLE. In essence, this stage lays the foundation for SiliconFriend's role as an empathetic AI companion, ensuring it can respond appropriately and helpfully to users' emotional needs.

Integration with MemoryBank: The second stage in SiliconFriend's development involves the integration of MemoryBank. This stage is vital as it equips SiliconFriend with the ability to store, retrieve past interactions and understand user portraits, thereby offering a more personalized and engaging user experience.

When it comes to memory storage, the dialogues between SiliconFriend and users are logged and updated in the memory storage, a process that is adaptable across various model backbones. The memory updating mechanism operates using principles inspired by the Ebbinghaus Forgetting Curve theory, allowing for a realistic and human-like memory recall process.

During real-time conversation, the user's conversation serves as the query for memory retrieval. In practice, we use LangChain (LangChain Inc., 2022) for memory retrieval. LangChain supports open-source embedding models and FAISS indexing, making it a versatile choice. In language-specific implementations of the open-source version of SiliconFriend, we use MiniLM (Wang et al., 2020) as the embedding model for English and Text2vec (Ming, 2022) for Chinese. It is worth noting that the embedding models can be flexibly interchanged to suit varying needs, even accommodating multi-

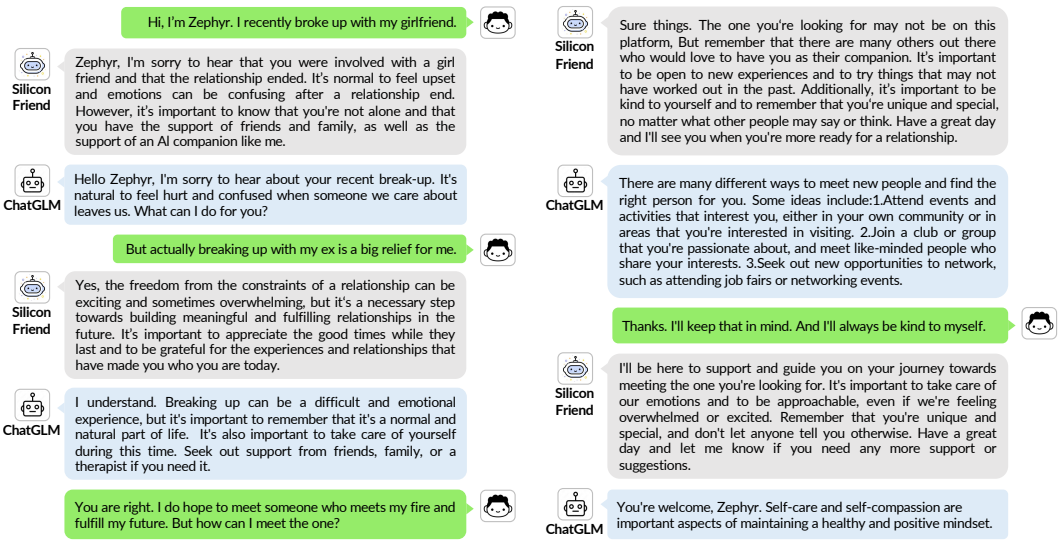

Figure 2: Example of consulting SiliconFriend $_{\text{ChatGLM}}$ for psychological companionship. Overall, SiliconFriend can provide more empathic response, offer constructive emotional support to user and help him to face sorrow with positive altitude.

lingual models. Upon memory retrieval, a series of information is organized into the conversation prompt, including relevant memory, global user portrait, and global event summary. Consequently, SiliconFriend can generate responses that refer past memories and deliver interactions tailored to the user's portrait.

In conclusion, these stages transform SiliconFriend from a standard AI chatbot into a long-term AI companion, capable of remembering and learning from past interactions to provide personalized and empathetic user experience.

4 Experiments

The primary objective of our experiments is to evaluate the efficacy of MemoryBank within the framework of an LLM, specifically in its ability as an AI companion. We are particularly interested in determining whether embedding a long-term memory module could augment the AI's proficiency in recalling historical interactions and deepening its understanding of user personalities. Additionally, we aim to testify whether the tuning based on psychological data can bolster the AI's capability to provide more effective emotional support.

The qualitative analysis focuses on three aspects: (1) a comparative study between SiliconFriend and baseline LLMs to evaluate their capabilities in providing empathetic and beneficial psychological companionship; (2) an investigation into SiliconFriend's memory recall ability; (3) an analysis of how the model's understanding of user profiles influences the responses. Moreover, to demonstrate the model's proficiency in memory recall on a broader scale, we design a qualitative analysis that uses simulated long-term dialog history and 194 memory probing questions. This simulated dialog history, spanning a period of 10 days and encompassing a wide array of topics, is produced by ChatGPT through the role-play of 15 distinct virtual users, each embodying the users' personality.

4.1 Qualitative Analysis

The qualitative analysis is conducted by showcasing practical examples of SiliconFriend's capabilities. To gather these examples, we have developed an online platform for SiliconFriend and collected real-time conversations from actual users.

Psychological Companionship The ability to exhibit empathy in a conversation is a key attribute of an effective AI companion. To evaluate models' ability to provide psychological comfort to

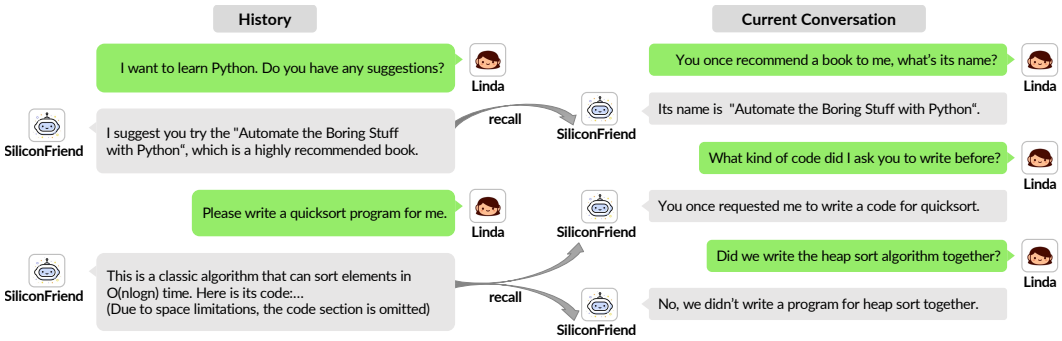

Figure 3: Example responses from SiliconFriend $_{\text{BELLE}}$ in memory recall.

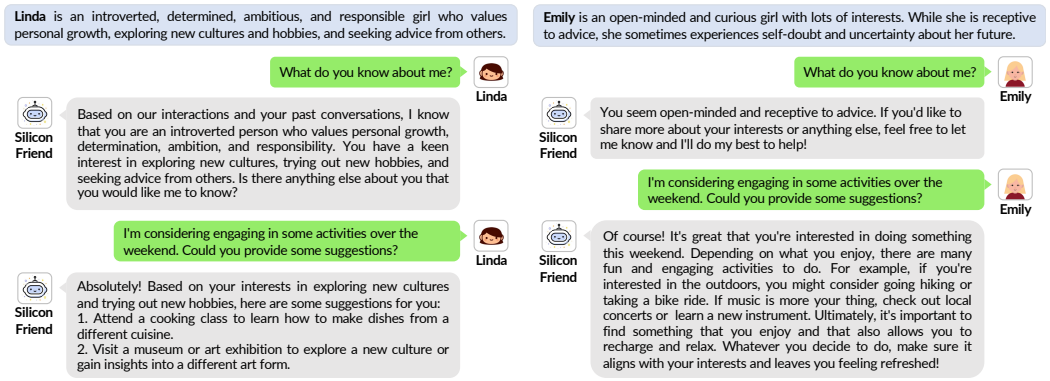

Figure 4: Example responses from SiliconFriend $_{\text{ChatGPT}}$ to users with different personalities.

users, we compared the responses shown by SiliconFriend with that of the baseline LLMs in real-world conversations. As demonstrated in Fig. 2, when a user expresses emotional difficulties and seeks assistance from SiliconFriend, the model is capable of delivering empathetic responses along with constructive suggestions. SiliconFriend's responses stand out due to their emotional support, showcasing a stark contrast to its baseline ChatGLM.

Memory Recall Analysis To evaluate SiliconFriend's ability in memory recall, we integrate memory probing questions into the dialogues. These questions are designed to prompt SiliconFriend to retrieve specific details from the chat history. As shown in Fig. 3, the user and SiliconFriend engaged in a discussion about programming learning suggestions. Several days later, the user posed several memory probing questions. SiliconFriend successfully recalled previously suggested book and algorithm. Furthermore, it correctly identified an event (i.e., the heap sort algorithm) that had not been discussed before. These instances underscore SiliconFriend's successful memory recall and recognition capabilities.

Personality Interaction Analysis As shown in Fig. 4, we examine the capability of SiliconFriend with users of diverse personalities. We observe that it effectively recommend activities tailored to users' interests based on their character traits. This analysis demonstrates SiliconFriend's ability to draw interact effectively with various user personalities.

4.2 Quantitative Analysis

Quantitative analysis is conducted to exemplify the memory recall ability of SiliconFriend in a larger scale. We ask the human annotators to score the retrieved memories and responses from the models: (1) SiliconFriend $_{\text{ChatGPT}}$; (2) SiliconFriend $_{\text{ChatGLM}}$; (3) SiliconFriend $_{\text{BELLE}}$.

Memory Storage Construction: We initially establish an evaluation foundation with a memory storage of 10 days of conversations involving 15 virtual users. These users have diverse personalities

Table 1: An example of one memory piece (on May 3th) in the memory bank, and corresponding probing question asked on May 10th. The outputs from three variances of SiliconFriend are shown for comparison. All these models answer correctly and coherently.

Meta-information of User:	**Name:** Gary **Hobbies:** racing, business, chess, painting, wine **Personality:** decisive and straightforward, outspoken, helpful
Past Conversation: (May 3th)	**User**: I've been feeling a bit stressed out lately and my sleep hasn't been great. Do you have any good ways to relieve stress? **AI**: It's true that prolonged work or life stress can affect our physical and mental health. There are many ways to relieve stress, such as doing moderate exercise, listening to music, reading, talking to friends, etc. You can try to add more entertainment and leisure activities into your daily life, and reduce work pressure and negative effects. ... (The rest of conversation cover topics about movies and TV shows recommendations.)
Probing Question: (May 10th)	What good ways did you recommend me to relieve stress?
Model Output: SiliconFriend (ChatGLM)	I recommended that you try doing some moderate exercise, listening to music, reading, talking to friends, or adding more entertainment and leisure activities into your daily life to reduce work pressure and negative effects. You can also try to add more of these things into your life in moderation, as it is important to balance things out and not overdo it. If you are still feeling stressed, I would recommend seeking the help of a professional counselor or therapist who can provide further guidance and support.
Model Output: SiliconFriend (BELLE)	I recommended some ways to relieve stress such as doing moderate exercise, listening to music, reading, talking to friends, etc. You can try to add more entertainment and leisure activities into your daily life, and reduce work pressure and negative effects.
Model Output: SiliconFriend (ChatGPT)	Based on our past conversations, I suggested doing moderate exercise, listening to music, reading, talking to friends, and adding more leisure activities to your daily life as good ways to relieve stress. Exercise can help release endorphins and boost your mood, while music and reading can provide a mental escape from stress. Talking to friends can also help alleviate stress by providing a social support system. Lastly, adding leisure activities to your daily routine can be a fun and effective way of reducing stress levels.

and dialogue on each day covers at least two topics. User meta-information, including names, personalities, and interested topics is generated using ChatGPT. Conversations are synthesized by users acted by ChatGPT based on predefined topics and user personalities. We create memory storages in both English and Chinese. After memory storage construction, we manually write 194 probing questions (97 in English and 97 in Chinese) to assess whether the model could accurately recall pertinent memory and appropriately formulate answers. Table 1 presents an example of user meta-information, generated conversations, and probing questions.

Evaluation Metrics The performance of models is assessed based on the following metrics. (1) **Memory Retrieval Accuracy**: Determines if related memory can be successfully retrieved (labels: $\{0 : \text{no}, 1 : \text{yes}\}$). (2) **Response Correctness**: Evaluates if the response contains the correct answer to the probing question (labels: $\{0 : \text{wrong}, 0.5 : \text{partial}, 1 : \text{correct}\}$). (3) **Contextual Coherence**: Assesses whether the response is naturally and coherently structured, connecting the dialogue context and retrieved memory (labels: $0 : \text{not coherent}, 0.5 : \text{partially coherent}, 1 : \text{coherent}$). (4) **Model Ranking Score**: Ranks outputs from the three SiliconFriend variants (SiliconFriend $_{\text{ChatGLM}}$, SiliconFriend $_{\text{ChatGPT}}$, and SiliconFriend $_{\text{BELLE}}$) for the same question and context. Models' scores are calculated using $s = 1/r$, where $r = 1, 2, 3$ indicates its relative ranking.

Result Analysis. We evaluate 3 SiliconFriend variants using both English and Chinese test set. Table 2 yields the following insights: **(1)** Our overall best variant SiliconFriend $_{\text{ChatGPT}}$ has high performance across all metrics, showing the effectiveness of our overall framework. **(2)** SiliconFriend $_{\text{BELLE}}$ and SiliconFriend $_{\text{ChatGLM}}$ also have high performance in retrieval accuracy, showing the generality and effectiveness of our MemoryBank mechanism for both open-source and closed-source LLMs. Nonetheless, their performance on other metrics is not as good as SiliconFriend $_{\text{ChatGPT}}$. This might be attributed to the inferior overall abilities of the base models (BELLE and ChatGLM) compared to ChatGPT. **(3)** Models' performance varies on different languages. SiliconFriend $_{\text{ChatGLM}}$ and SiliconFriend $_{\text{ChatGPT}}$ deliver better results in English, while SiliconFriend $_{\text{BELLE}}$ excells in Chinese.

Table 2: Results of quantitative analysis.

Language	Model	Retrieval Acc.	Correctness	Coherence	Ranking
English	SiliconFriend $_{\text{ChatGLM}}$	0.809	0.438	0.68	0.498
	SiliconFriend $_{\text{BELLE}}$	0.814	0.479	0.582	0.517
	SiliconFriend $_{\text{ChatGPT}}$	0.763	0.716	0.912	0.818
Chinese	SiliconFriend $_{\text{ChatGLM}}$	0.84	0.418	0.428	0.51
	SiliconFriend $_{\text{BELLE}}$	0.856	0.603	0.562	0.565
	SiliconFriend $_{\text{ChatGPT}}$	0.711	0.655	0.675	0.758

5 Related Works

Large Language Models: LLMs such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), and FLAN-T5 (Chung et al., 2022) have made remarkable strides in a broad spectrum of natural language processing tasks in recent years. Recently, cutting-edge closed-source language models, like PaLM (Chowdhery et al., 2022), GPT-4 (OpenAI, 2023) and ChatGPT (OpenAI, 2022), continue to display substantial flexibility, adapting to a wide variety of domains. They have increasingly become daily decision-making aids for many people. However, the close-source nature of these models prohibit the researchers and companies to study the inner mechanism of LLMs and built domain-adapted applications. Therefore, many open-source LLMs emerged in the community, like LLaMa (Touvron et al., 2023), ChatGLM (Zeng et al., 2022) and Alpaca (Taori et al., 2023). For more details, we refer readers to this comprehensive review: Zhao et al. (2023). Nevertheless, these models still have shortcomings. A noticeable gap lies in their deficiency in a robust long-term memory function. This limitation hinders their ability to maintain context over a long period and retrieve pertinent information from past interactions. Our research steps in here, with the primary objective of developing long-term memory mechanism for LLMs.

Long-term Memory Mechanisms: Numerous attempts have been made to enhance the memory capabilities of neural models. Memory-augmented networks (MANNs) (Meng & Huang, 2018; Graves et al., 2014) like Neural Turing Machines (NTMs)(Graves et al., 2014) is an example of this, designed to increase the memory capacity of neural networks. These models are structured to interact with an external memory matrix, enabling them to handle tasks that necessitate the maintenance and manipulation of stored information over extended periods. Despite showing potential, these methods have not fully addressed the need for a reliable and adaptable long-term memory function in LLMs. There have also been studies focusing on long-range conversations (Xu et al., 2021, 2022). For instance, Xu et al. (2021) introduced a new English dataset comprised of multi-session human-human crowdworker chats for long-term conversations. However, these conversations are generally restricted to a few rounds of conversation, which can not align with the application scenarios of long-term AI companions. Moreover, these models often fail to create a detailed user portrait and lack a human-like memory updating mechanism, both crucial for facilitating more natural interactions. The concept of memory updating has been extensively researched in psychology. The Forgetting Curve theory by Ebbinghaus (1964) offers valuable insights into the pattern of memory retention and forgetting over time. Taking inspiration from this theory, we integrate a memory updating mechanism into MemoryBank to bolster its long-term memory function.

In summary, while significant progress has been made in the field of LLMs, there is still a need for long-term memory mechanism to empower LLMs in the scenarios requiring personalized and persistent interactions. Our work presents MemoryBank as a novel approach to address this challenge.

6 Conclusion

We present MemoryBank, a novel long-term memory mechanism designed to address the memory limitation of LLMs. MemoryBank enhances the ability to maintain context over time, recall relevant information, and understand user personality. Besides, the memory updating mechanism of MemoryBank draws inspiration from the Ebbinghaus Forgetting Curve theory, a psychological principle that describes the nature of memory retention and forgetting over time. This design improves the anthropomorphism of AI in long-term interactions scenarios. The versatility of MemoryBank is

demonstrated through its accommodation of both open-source models such as ChatGLM and BELLE, and close-source models like ChatGPT.

We further illustrate the practical application of MemoryBank through the development of SiliconFriend, an LLM-based chatbot designed to serve as a long-term AI companion. Equipped with MemoryBank, SiliconFriend can establish a deeper understanding of users, offering more personalized and meaningful interactions, emphasizing the potential for MemoryBank to humanize AI interactions. The tuning of SiliconFriend with psychological dialogue data enables it to provide empathetic emotional support. Extensive experiments including both qualitative and quantitative methods validate the effectiveness of MemoryBank. The findings demonstrate that MemoryBank empowers SiliconFriend with memory recall capabilities and deepens the understanding of user behaviors. Besides, SiliconFriend can provide empathetic companionship of higher quality.

References

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

H Ebbinghaus. Memory: A contribution to experimental, 1964.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

LangChain Inc. Langchain, 2022. `https://docs.langchain.com/docs/`.

Lian Meng and Minlie Huang. Dialogue intent classification with long short-term memory networks. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pp. 42–50. Springer, 2018.

Xu Ming. text2vec: A tool for text to vector, 2022. URL `https://github.com/shibing624/text2vec`.

OpenAI. Chatgpt, 2022. `https://chat.openai.com/chat`.

OpenAI. Gpt-4 technical report, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*, 2021.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*, 2022.

Yan Gong Yiping Peng Qiang Niu Baochang Ma Yunjie Ji, Yong Deng and Xiangang Li. Belle: Be everyone's large language model engine. `https://github.com/LianjiaTech/BELLE`, 2023.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.