VariaMem: A New Benchmark for Memory Planning, Retrieval, and Incorporation of Large Language Models

Anonymous ACL submission

Abstract

Long-term memory is crucial in various scenarios such as secretarial tasks, agent planning, and decision-making. The complexity and diversity of human memory are evident in the different memory types, individual variabilities, and social memory associations. However, most existing research oversimplifies memory by focusing solely on dialogue or profile, neglecting its rich and diverse nature. To address this limitation, we introduce a memory dataset called VariaMem [1], consisting of two parts: 1) various types of memory, such as profiles (e.g., user characteristics), semantic memory (e.g., intersocial relationships), episodic memory (e.g. events), and dialogues; and 2) an evaluation benchmark, including 10,388 evaluation questions centered around 30 characters. Specifically, we propose a novel personal memory integration generation method to generate different types of memory and explore the utilization of diverse long-term memories. We then design three sub-tasks: ***Memory Planning***, ***Memory Retrieval*** and ***Memory Incorporation*** to assess the long-term memory utilization of five LLMs and three retrievers. Experimental results demonstrate that BERT-based models significantly outperform LLMs such as ChatGLM3 and ChatGPT in the memory planning task. Furthermore, when furnished with accurately retrieved memories, various LLMs exhibit differing capabilities in generating memory-based responses.

1 Introduction

Long-term memory serves as a pivotal element of human cognition, finding application in diverse scenarios such as secretarial tasks and agent planning (Sumers et al., 2023). Within administrative contexts, long-term memory facilitates the proficient handling of extensive information, enabling the effortless recall of previous correspondences, details of meetings, and contact data. In the realm of agent planning (Sumers et al., 2023), long-term memory transforms into a reservoir of amassed knowledge that significantly impacts decision-making processes. For instance, it steers the strategic decisions of business executives by leveraging historical market trends and the results of past initiatives. Therefore, it is evident that long-term memory plays an essential part of daily human life.

In cognitive science (Atkinson and Shiffrin, 1968), long-term memory (Cowan, 2008) can be divided into episodic memory and semantic memory (Eysenck and Keane, 2020; Bruck, 2022), both of which play a vital role in communication. Similarly, in the realm of computer science, the concept of long-term memory also holds significant importance, where personal information and distant dialogue are considered as two crucial factors of long-term memory (Xu et al., 2021, 2022).

To enhance the storage and utilization of long-term memory in large language models (LLMs) (Wang et al., 2023b), we categorize long-term memory into four types, inspired by the memory classifications in cognitive science and computer science: ***profiles***, ***semantic memory***, ***episodic memory***, and ***dialogues***. Specifically, the profile is fixed memory information that does not change with interpersonal interaction (Xu et al., 2022). Semantic memory includes relationships between people, etc., which may change with communication. Episodic memory refers to individual past experiences and events that can be abstracted and reflected from the dialogue. Dialogue is a more fragmented memory, which contains more language fragments or emotional attitude memories, etc. As indicated in Table.1, we are the pioneers in considering these four types of memory concurrently.

Within the context of these four types of memory, we explore their utilization in LLMs. Drawing on our understanding of memory characteristics and previous research on long-term memory (Zhong

[1] Our code and dataset will be publicly released once accepted.

Memory Type	Profile	Semantic Memory	Episodic Memory	Dialogues
Persona-Chat (Zhang et al., 2018)	✓	✗	✗	✓
Personal-Dialogue (Zheng et al., 2019)	✓	✗	✗	✓
DialogueSum (Chen et al., 2021)	✓	✗	✗	✓*
MSC (Xu et al., 2021)	✓	✗	✗	✓
DuLeMon (Xu et al., 2022)	✓	✗	✗	✓
MemoryBank (Zhong et al., 2023)	✓	✗	✗	✓*
Zhang et al. (2023b)	✓	✗	✗	✓*
VariaMem (ours)	✓	✓	✓	✓

Table 1: A long-term memory typology in existing dialogue datasets with $*$ indicating the presence of a dialogue summary as a long-term memory analogous to episodic memory

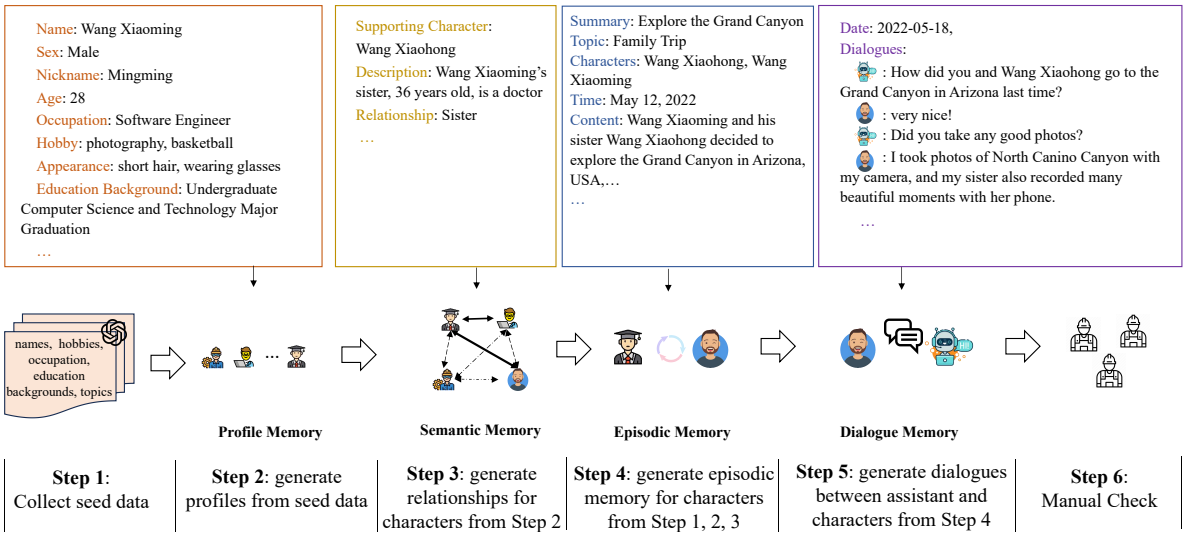

Figure 1: The process of VariaMem memory database generation. A six-step process: Step 1. Seed data collection. Step 2. Profile generation. Step 3. Semantic memory generation. Step 4. Episodic memory generation. Step5. dialogue generation and Step 6. manual check.

et al., 2023; Wang et al., 2023a,c,d; Xu et al., 2022; Lu et al., 2023), we employ a three-step approach: ***memory planning***, ***memory retrieval***, and ***memory incorporation*** for response generation. Initially, when asked a question, we must determine the appropriate type of memory to use, a process that we refer to as the Memory Planning task. Subsequently, within the context of the identified memory type, we retrieve the most pertinent memory information to the question, a step known as the Memory Retrieval task. Finally, upon retrieval of the relevant memory, we must articulate our response accurately, necessitating the generation of a memory-based response, a process we term as the memory incorporation for response generation task. By following these steps, we can accurately identify the type of memory and the information based on the user's question, enabling us to provide a correct response.

To explore the capabilities of LLMs in utilizing long-term memory, we have developed the VariaMem dataset. This dataset includes a memory database, generated through a personal memory integration in-context generation method, as depicted in Fig.1. The memory database stores various types of long-term memory, serving as data sources for retrieval models during conversation. In addition, we have also established a memory evaluation benchmark. This benchmark is used to assess how LLMs generate responses that mirror the accurate memory.

In summary, our research makes three main contributions:

- We introduced the VariaMem dataset that includes a memory database integrating 141 profiles, 1,339 semantic memories, 3,445 episodic memories and 3,409 dialogues, and a memory benchmark including 10388 memory-related evaluation questions.
- We propose three subtasks ***memory planning***,

memory retrieval, and ***memory incorporation*** response generation to evaluate the memory utilization capabilities of LLMs. We carry out experiments using five LLMs(Wang et al., 2023b) and three retrieval models.

- Our experiment results indicate that BERT-based models excel in memory planning tasks, surpassing LLMs like ChatGLM3 and ChatGPT, and LLMs show varied proficiency in generating memory-based responses when provided with accurately retrieved memories.

2 Related Work

2.1 Long-term Memory Dialogue Datasets

Numerous datasets have been developed with an emphasis on long-term memory in the context of dialogue systems (Zhang et al., 2018; Zheng et al., 2019; Chen et al., 2021; Zhong et al., 2023). Specifically, DialogSUM (Chen et al., 2021) offers dialogue content and summaries, but it does not include other memory-related information. PersonalChat (Zhang et al., 2018) and PersonalDialog (Zheng et al., 2019) are dialogue datasets focused on personal information. The evaluation of PersonalChat does not primarily consider memory, while PersonalDialog strives to fully integrate personal data. The MSC dataset (Xu et al., 2021) captures dialogue content to augment memory capacity, but it is limited in its scale and scope. The Dulemon dataset (Xu et al., 2022), a Chinese resource, includes personal preference memory information. MemoryBank (Zhong et al., 2023) introduces a bilingual memory dataset by utilizing GPT4 to summarize dialogues and personal information. Despite these advancements, there remains a deficiency in long-term memory datasets that encompass well-annotated long-term memory and a diverse range of memory types, which are essential for the comprehensive exploration of memory incorporation in LLMs.

2.2 Long-term Memory Utilization

Long-term memory, an essential aspect of human life, is categorized into episodic and semantic memory within cognitive psychology (Eysenck and Keane, 2020). This type of memory plays a significant role in conversations. In dialogue systems, long-term memory is often leveraged to enhance dialogue models through retrieval from an external knowledge base (Xu et al., 2022). For instance, (Zhao et al., 2023) presents a unified dialogue framework that bolsters long-term memory by reinforcing the connection between different stages of relevant representation. (Lee et al., 2023) promotes the integration of pre-trained models with few-shot prompting and CoT techniques to use dialogue summarization as a form of long-term memory. Despite (Wang et al., 2023c)'s efforts to preserve long-term memory by summarizing dialogue data, the challenge of catastrophic forgetting remains. MemoryBank (Zhong et al., 2023) utilizes LLM to integrate long-term memory into dialogues and introduces a forgetting mechanism. The study (Sumers et al., 2023) enhances LLMs by incorporating memory components, specifically episodic and semantic memory. Despite this, the focus is not on concrete memory types, but rather on the agent itself. Nonetheless, accurately classifying memory types and utilizing them remains a significant challenge.

3 Dataset Collection

In this section, we introduce how to construct the VariaMem dataset. The VariaMem includes two parts: memory database and memory benchmark. we propose a personal memory integration in-context generation method to generate a memory database and use a similar method to collect the memory benchmark.

3.1 Memory Database

As shown in Fig.1, the memory database generation is decomposed into six steps:

Step 1. Diverse Seed Data Collection. In our work, the seed data is defined as occupations, educational backgrounds, hobbies, and topics. These data are derived from a variety of sources, including information gathered from ChatGPT [2] and Wikipedia. It comprises professional backgrounds that span across 10 categories and 299 specialties, hobbies that are categorized into 7 groups with 140 items, and a comprehensive range of topics structured into 49 categories with 2442 subtopics. Complementing this approach, `gpt-3.5-turbo` is employed to generate 141 virtual names. This step ensures the richness and diversity of our dataset.

Step 2. Profile Generation. Our objective is to generate comprehensive profiles using the provided seed data. Using the capabilities of ChatGPT (`gpt-3.5-turbo`), we generate fundamental facts of personal attributes that span 14 cate-

[2] https://chat.openai.com.

gories. These categories encompass various aspects such as name, gender, nickname, age, occupation, nationality, appearance, hobbies, achievements, education, profession, employer, awards, and role models.

Step 3. Semantic Memory Generation. To establish various interpersonal relationships, we utilize prompts to generate more than ten distinct types of relationships. These include family, friends, romantic partners, acquaintances, colleagues, mentors/students, neighbors, community members, and strangers, each tailored to every character.

Step 4. Episodic Memory Generation. Each character is enriched with a multitude of events, intricately woven into their episodic memory and interconnected with other characters. The generation of these events begins with the creation of detailed descriptions of background events on randomly chosen topics. Following this, prompts are utilized to generate the episodic memory that is intimately linked to these foundational events. To ensure a harmonious alignment between character relationships and background events, we employ few-shot instructions (Brown et al., 2020). This methodology assists ChatGPT (`gpt-3.5-turbo`) in preserving the required consistency.

Step 5. Dialogues Generation. Utilizing the episodic memory produced in Step 4, we facilitate dialogues between the AI assistant and the characters. In our directives, we encourage LLMs to refer to the recollected events from episodic memory and the corresponding interpersonal relationships from semantic memory, thus enriching and broadening the conversations. This methodology ensures that our dialogues are deeply anchored in memory, promoting a more genuine and contextually pertinent narrative.

Step 6. Manual Check. We employ a team of three individuals to perform data post-processing. Their primary tasks include reducing the noise introduced by large language models, adjusting formatting, eliminating duplicate data, and ensuring ethical compliance. These collective measures significantly improve the quality of the memory database generated for character.

To complete the above data generation process and establish connections between memories, we have carefully designed the prompts as shown in the Appendix.A

	Memory Database				
	profile	**semantic**	**episodic**	**dialogues**	**total**
Numbers	141	1,339	3,445	3,409	8334
	Memory Benchmark				
QA	370	949	5,516	3,553	10,388
MA	370	949	6,216	3,880	11,415

Table 2: VariaMem dataset statistics. QA indicates question and answer pairs and MA indicates memory anchors.

3.2 Memory Benchmark

To comprehensively evaluate each type of memory for a character, we collect four evaluation items (*evaluation question*, *answer*, *background memory*, and *memory anchor*) to evaluate the memory. The memory benchmark collection process is delineated into three steps:

Question and answer construction. Utilizing `gpt-3.5-turbo`, we generate evaluation questions prompted by the memory items stored in our database. Corresponding answers are also produced, designed to align with the provided background memories, adhering to the prompts we crafted.

Memory Anchor Annotation. We utilize an exact match to identify the memory anchor, a crucial text segment within the answer that correlates with the background memory. The memory anchor positions are also labeled reference by the background memory.

Manual Check on QA pairs and Memory Anchor. Following the procedures described, we perform a detailed human review to ensure alignment between the questions and the memory database. This is followed by a verification process to confirm the accuracy of the answers. Moreover, a hands-on examination is undertaken to correct any inconsistencies in the memory anchor labels and their corresponding positions.

3.3 Dataset Statistics

The VariaMem dataset consists of a memory database and a memory benchmark. The statistics for this dataset are presented in Table.2. The memory database includes 141 characters, each associated with profiles, semantic memories, episodic memories, and dialogues. This amounts to a total of 8,834 memory items. The memory evaluation benchmark encompasses corresponding questions, answers, and memory anchors for each memory item. This results in a comprehensive set of 10,388 question-and-answer pairs and 11,415 memory an-

chors. It is worth noting that the VariaMem dataset generates an abundance of social attributes for a variety of characters, encompassing more than 80 types of hobbies and over 100 different educational backgrounds. Furthermore, it includes 18 different types of social relationships. Social interactions among these characters have led to more than 30 types of topical events in more than 100 occupational behaviors, spanning fields such as law, biology, and environmental protection, among others. In particular, these events, which are centered on the characters and their social relationships, extend over up to 4 years, thereby providing a comprehensive representation of character information.

4 Evaluation Tasks

4.1 Task Definition

The Memory dababase is formulated as $M = \{m_i = (P_i, S_i, E_i, D_i)\}_{i=1}^{n}$, where each m_i represents the memory of the i-th character. Here, $P_i = \{p_{ik}\}_{k=1}^{l_1}$ denotes the profile memory information with l_1 being the length of the profile, $S_i = \{s_{ik}\}_{k=1}^{l_2}$ represents semantic memory with l_2 being the length of semantic memory, $E_i = \{e_{ik}\}_{k=1}^{l_3}$ signifies the episodic memory with l_3 being the length of episodic_memory, $D_i = \{d_{ik}\}_{k=1}^{l_4}$ stands for the dialogue with l_4 being the length of dialogue turns. The variable n is the number of characters.

The memory benchmark concludes to $T = \{t_j^{\pi}\}_{j=1}^{n}$, where π encompasses four types of memory data: profile, semantic memory, episodic memory, and dialogues. Each t_j^{π} has four main elements, namely $t_j^{\pi} = (q_j^{\pi}, r_j^{\pi}, m_j^{\pi}, a_j^{\pi})$, where $q_j^{\pi}, r_j^{\pi}, m_j^{\pi}, a_j^{\pi}$ denotes the evaluation question, answer, background memory, and corresponding memory anchors, respectively. As shown in Fig.2, we decompose the evaluation framework into three sub-tasks: *memory planning*, *memory retrieval* and *memory incorporation* for response generation.

Memory Planning can be formulated as a multi-class classification task to identify memory type, as illustrated in Fig.2. We devise classification instructions, denoted as G, (the details are provided in A) which are subsequently concatenated with the evaluation question q_i for the i-th character. These concatenated inputs are then processed by a memory planning model (MP), which outputs the memory type π. The task is defined as Eq.1.

$$\pi = MP(G, q_i) \tag{1}$$

Memory Retrieval is a classic information retrieval task. It involves extracting the memory associated with the evaluation question from the memory benchmark. We define this memory retrieval task as Eq.2.

$$V_i = RM(q_i, \pi, M_i^{\pi}) \tag{2}$$

where M_i^{π} denotes the memory sentences from the memory type π, and V_i refers to the retrieved memory sentences set and the length of V_i depends on the experiment settings by retrieval model RM.

Memory Incorporation is the response generation task using LLMs. Given the retrieved memory V_i and evaluation question q_i, we define this task as Eq.3.

$$r_i' = LLM(q_i, V_i) \tag{3}$$

where r' is the predicted response generated by LLM with q_i and V_i as inputs.

4.2 Evaluation Metrics

For the memory planning task, we use precision (P), recall (R), F1, and Accuracy to serve as key indicators. For the memory retrieval task, we utilize Recall@K (Manning et al., 2008) as our primary metric, defined as Eq.4.

$$\text{Recall@k} = \frac{1}{|T|}\sum_{t=1}^{|T|} I(\#retr_{m_t^k} > 0) \tag{4}$$

where $I(\#retr_{m_t^k} > 0)$ is the indicator function returning 1 if the relevant memory sentence of the top k is retrieved from the memories.

To evaluate the memory incorporation for the response generation task, we measure the correctness and coherence of responses with a ChatGPT (`gpt-3.5-turbo`)-based evaluation method (Zhong et al., 2023). We use Memory Anchor Points (MMAP), as adapted from (Xu et al., 2021) to evaluate the memory incorporation ability defined by Eq. 5.

$$\text{MMAP} = \frac{1}{|T|}\sum_{t=1}^{|T|}\frac{1}{|a_t|}\sum_{j=1}^{|a_t|} I(\#res_{a_{tj}} > 0) \tag{5}$$

where $|T|$ is the number of evaluation questions, $I(\#res_{a_{tj}} > 0)$ is an indicator function that returns 1 only when the memory anchor $a_t j$ in the response. We finally use the average of the memory anchor points as the metric to evaluate the memory incorporation capacity in response.

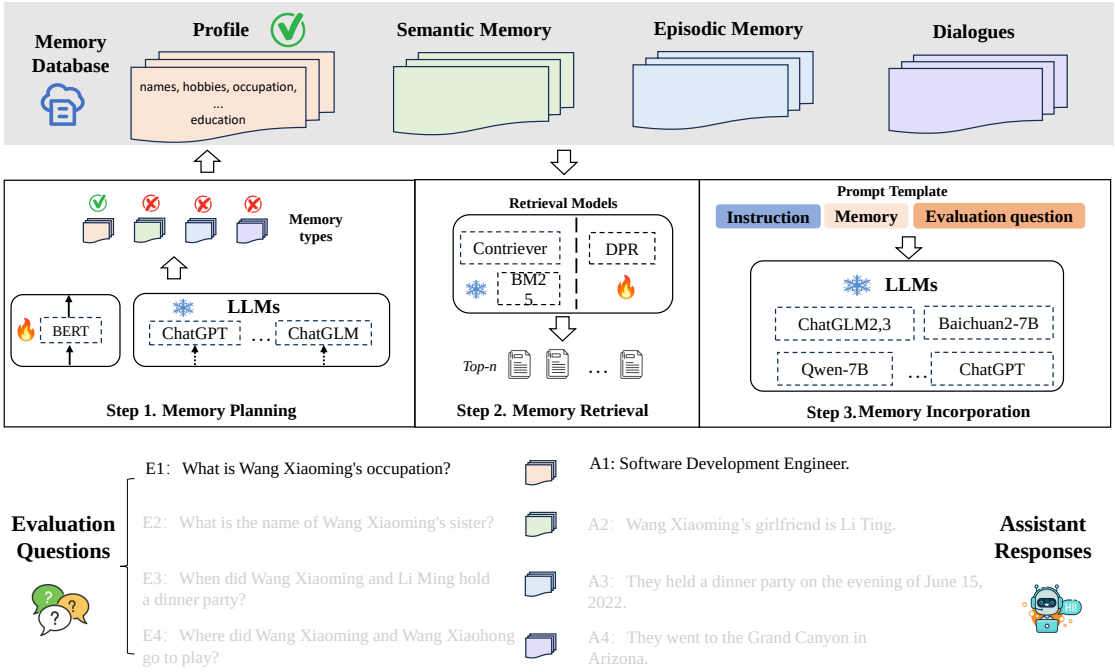

Figure 2: Illustration of using memory in LLMs.

5 Experiments

5.1 Implementation details

Memory Planning settings. We conduct multi-class classification experiments on memory types using BERT (Devlin et al., 2018), ChatGLM3 (Zhang et al., 2023a), and ChatGPT. For BERT, we employ fine-tuning with the evaluation question to predict the memory type. For ChatGLM3 and ChatGPT, we use prompts to guide LLMs to output the memory type. We also conduct boosting experiments on these models. Specifically, we train BERT-boost models with 7,516 benchmark questions, treating questions with the correct memory types as positive examples and questions with incorrect memory types as negative examples. We train separate binary classification BERT models for different memory types. For LLMs, we use prompts to ascertain if the evaluation question pertains to a specific memory type. If the response is affirmative, we cease further inquiry and return the corresponding answer. Finally, we evaluate the performance of memory type planning on a test set of 2,872 evaluation questions.

Memory Retrieval settings. Our experiment is designed to emulate practical scenarios of user data isolation, with each character being treated separately. We employ three retrieval models - DPR (Karpukhin et al., 2020), BM25 (Doostmohammadi et al., 2023), and Contriever (Mueller et al., 2023) - to gather character memories. For the BM25 model, we created unique memory banks for each character. In the case of DPR, we train the DPR model using additional 7516 evaluation questions. Contriever used the text2vec model (Xu, 2023) from Hugging Face to calculate the similarity between memory sentences and evaluate questions.

Memory Incorporation settings. LLMs have emerged as a popular trend in response generation tasks. We have five LLMs that have demonstrated superior performance on CEval (Huang et al., 2023). It includes models such as ChatGLM2, ChatGLM3 (Zhang et al., 2023a), Baichuan2-7B-Chat (Yang et al., 2023), Qwen-7B-Chat (Bai et al., 2023), and ChatGPT (Wang et al., 2023b). While there are other exceptional LLMs, such as InternLM (Zhang et al., 2023c), and larger-scale LLMs like the 13B and 30B models (Zeng et al., 2022), our resources are limited, and we have only utilized LLMs with fewer than 10 billion parameters. These models are prompted by retrieved memories. To ensure smooth operation on an Nvidia-3090 GPU with 24GB of memory, we have implemented a semi-precision inference setting.

5.2 Memory Planner

BERT-based models provide better performance than LLMs for memory planning. As shown in Table.3, BERT demonstrates superior performance in direct multi-class classification experiments compared to other LLMs. Even under few-shot settings, where we provide an evaluation ques-

Metrics	P	R	F1	Acc
ChatGLM3-M	0.86	0.56	0.65	0.56
ChatGPT-M	0.92	0.71	0.77	0.71
BERT-M	0.95	0.88	0.90	0.88
few-shot ChatGPT	0.91	0.57	0.66	0.57
ChatGLM3-Boost	0.81	0.16	0.20	0.16
ChatGPT-Boost	0.91	0.48	0.56	0.64
BERT-Boost	**0.95**	**0.93**	**0.93**	**0.93**

Table 3: Memory Planning performance in ChatGLM3, ChatGPT, and BERT for direct multi-class classification, few-shot multi-class classification, and boost settings.

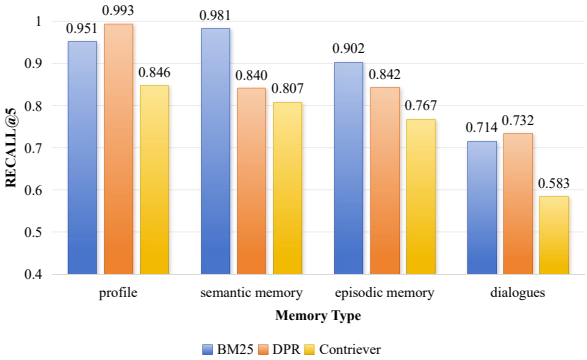

Figure 3: Memory Retrieval RECALL@5 results across profile, semantic memory, episodic memory, and dialogues utilizing Contriever, BM25, and DPR retrieval models.

tion along with corresponding samples for each memory type, ChatGPT's performance decreases compared to the direct multi-class classification method. Our boosting experiments further highlight that LLMs fall short when compared to direct multi-class classification. However, the BERT-boost model stands out, achieving state-of-the-art results. In summary, across three different settings and base models, the BERT-boost model consistently exhibits superior performance, achieving the highest precision (95%), recall (93%), F1 score (93%), and accuracy (93%).

5.3 Memory Retrieval

Different retrieval models perform variably across various memory types. In the context of the memory retrieval task, we assess the efficacy of retrieval models (BM25, Contriever, and DPR) in retrieving memory. The accompanying bar chart illustrates the retrieval performance in the metric of Recall@k of each memory type. As demonstrated in Figure 3, DPR exhibits superior Recall@k performance for profile memory and dialogue retrieval. Conversely, for semantic and episodic memory retrieval, BM25 demonstrates optimal performance at the metric of Recall@5. In comparison, Contriever exhibits comparatively inferior performance.

5.4 Memory Incorporation

The study evaluated four scenarios: without retrieval, retrieval, correct retrieval, and incorrect retrieval. Without a retrieval model, LLMs directly generate responses. With retrieval, k candidate memories are retrieved with each evaluation question. Correct retrieval is defined as the scenario where at least one candidate's memory aligns with the background memory, while incorrect retrieval refers to instances where none of the candidates matches.

Qwen excels in memory incorporation over other LLMs with retrieval. As illustrated in Table.4, the integration of a retrieval model significantly enhances the performance of all LLMs. In particular, correct retrieval outperforms general retrieval, and, intriguingly, even incorrect retrieval surpasses the scenario without retrieval. This suggests that accurate responses can still be generated even when the retrieved memories do not match. Most importantly, correct retrieval allows all LLMs to achieve a significant improvement in the MMAP and response correctness metric. It is worth noting that Qwen outperforms other LLMs, such as ChatGPT, as measured by MMAP.

Baichuan2 outperforms other LLMs by achieving superior response quality with noise memory. In scenarios where memory retrieval settings are incorrect, our retrieval model can retrieve noise memory. However, our model may inadvertently retrieve noise memory. Despite this, the potential to respond accurately to user queries is maintained, attributable to the uniformity of memory across various memory types associated with a single character. The results demonstrate that Baichuan2 achieves the highest MMAP performance compared to other LLMs.

6 Analysis and Case Study

6.1 Ablation Study

Memory planning boosts memory retrieval and incorporation. Table.5 demonstrates the enhanced capabilities of the dialogue system when the memory planning mechanism is integrated. This improvement is evident in the RMCC metric (average score of Recall@K, MMAP, correctness, and coherency), indicating better memory retrieval and incorporation. Case studies 4 provide further insight. Specifically, models with the memory planning

Models	W/o Retrieval		W/ Retrieval		Incorrect Retrieval		Correct Retrieval	
	MMAP	Correctness	MMAP	Correctness	MMAP	Correctness	MMAP	Correctness
Baichuan2-7B	0.0592	0.1967	0.5927	0.785	0.4969	0.5975	0.6832	0.7921
Qwen-7B	0.0576	0.0743	0.6621	0.8121	0.3904	0.6103	0.7402	0.8243
ChatGLM2	0.0439	0.1396	0.5736	0.773	0.2765	0.5841	0.6452	0.7642
ChatGLM3	0.0568	0.1616	0.6247	0.7804	0.3195	0.7843	0.6853	0.8431
ChatGPT	0.0321	0.1141	0.6482	0.7781	0.0213	0.5844	0.7156	0.8423

Table 4: Comparison of MMAP and Correctness Results under different retrieval settings: withoutRetrieval, with Retrieval, Incorrect Retrieval, and Correct Retrieval

Evaluate Question: 王伟的职业是什么？(What is Wang Wei's occupation?)
Ground Truth: 王伟是一名摄像师 (Wang Wei is a cameraman)
Memory Anchor : 摄影师 (cameraman)

Without Retrival #case 1

R-1:王伟是一位老师。(Wang Wei is a teacher.)
Memory Anchor Score: 0/1

With Incorrect Retrieval and without Memory Guiding #case 2

Retrieved Memory: AI助手:听说你和王伟在电影里的合作非常成功，得到了很高的评价。(AI Assistant: I heard that your cooperation with Wang Wei in the movie was very successful and received high praise.)(episodic memory)
R-2 : 王伟是一位演员。(Wang Wei is an actor.)
Memory Anchor Score: 0/1

With Correct Retrieval and Memory Guiding #case 3

MMG-Retrieved Memory: 王伟是徐佳的电影制作公司的同事，年龄30岁，是一名摄像师。他们经常合作拍摄电影和电视剧，默契度很高。徐佳和王伟的关系是同事。(Wang Wei is a colleague of Xu Jia's film production company. He is 30 years old and a cameraman. They often work together on movies and TV series and have a very good rapport. Xu Jia and Wang Wei are colleagues.) (semantic memory)
R-3 : 王伟是一名摄像师。(Wang Wei is a cameraman.)
Memory Anchor Score: 1/1

Figure 4: Comparative analysis of response performance without Retrieval, Incorrect Retrieval and Correct Retrieval with MP.

	W MP	W/o MP
Baichuan2-7B	0.7626	0.7847
ChatGLM2	0.7496	0.7683
ChatGLM3	0.7615	0.7919
ChatGPT	0.7824	0.7999
Qwen-7B	0.7800	0.8117

Table 5: Comparison of LLMs performance with and without Memory Planning on the metric of RMCC, which is an average of Recall@5, MMAP, Correctness, and Coherency.

mechanism can accurately identify memory types, preventing confusion, and improving retrieval performance. This accurate memory retrieval subsequently enhances response correctness and coherence.

6.2 Case Study

We present specific cases in Figure 4 to evaluate the question 'What is Wang Wei's occupation?' with the verifiable answer 'cameraman'. Without memory retrieval, Qwen generates a speculative response 'Wang Wei is a teacher', a common hallucination in most LLMs, or provides context-less responses. Introducing memory retrieval, we observe two cases. In case 2, the model responds 'Wang Wei is an actor' based on the dialogues retrieved. Despite higher accuracy due to analogous character experiences, case 2 still provides an incorrect answer. The key difference between cases 2 and 3 is the memory planning mechanism. While case 2 retrieves relevant dialogues, it fails to retrieve essential background memory, relying on semantic memory as in case 3. With memory planning, our models retrieve accurate background memory, yielding correct responses. In this evaluation, with 'cameraman' as memory anchor, only case 3 correctly incorporates the pertinent memory, as indicated by MMAP.

7 Conclusion

In summary, our research presents the VariaMem dataset, integrating a comprehensive memory database and evaluation benchmark for assessing memory-based LLMs. We introduce three subtasks - memory planning, retrieval, and incorporation - and conduct baseline experiments with various LLMs and retrieval models. The experimental results demonstrate the effectiveness of the Bert-boost memory planning mechanism outperforms LLMs in memory type classification and further improves memory retrieval accuracy. Importantly, we discern that different types of LLMs exhibit significant variations in delivering accurate memory-based responses. This work significantly contributes to advancing the understanding and evaluation of long-term memory-based LLMs.

Limitations

In this work, we utilize `gpt-3.5-turbo` to generate a memory-based dataset and assess its ability to generate responses based on memory across three distinct subtasks. However, we would like to acknowledge the following four limitations: 1. The process of generating memory data in the VariaMem memory database could be varied. We have only implemented a step-by-step generation method based on memory types. Furthermore, the prompts used during the generation process still have room for optimization. 2. All the content is fictional. Despite our thorough manual screening, there may still be instances where the common knowledge presented does not align with reality. 3. Our evaluations are limited to four open-source LLMs that are less than 10B in size, including ChatGPT. We do not evaluate other LLMs of varying scales and types. 4. For the evaluation of the correctness and coherence of response generation, we adopted the evaluation methods of LLMs. However, this metric may still have uncertainties in accurately measuring the quality of responses.

Ethics Statement

The work presented in this paper introduces the VariaMem dataset, which is generated from ChatGPT (`gpt-3.5-turbo`). This dataset does not violate any licenses or policies, nor does it infringe on privacy. The dataset can be utilized for academic exploration in memory-based QA, dialogue, and other related fields. To ensure the quality of the data, we have employed three researchers in the field of natural language who are proficient in both Chinese and English and possess excellent communication skills. Each researcher is paid $20 per hour (above the average local payment of similar jobs). The design, annotation, and review of the entire dataset took a total of four months, costing approximately an average of about 200 hours per annotator. The annotators have no affiliation with any of the companies that are used as targets in the dataset, eliminating any potential bias due to conflict of interest.

References

Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier.

Jinze Bai, Shuai Bai, and et al. 2023. Qwen technical report.

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.

Jason N Bruck. 2022. Long-term memory. In *Encyclopedia of animal cognition and behavior*, pages 4005–4009. Springer.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. Dialogsum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313.

Nelson Cowan. 2008. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ehsan Doostmohammadi, Tobias Norlund, Marco Kuhlmann, and Richard Johansson. 2023. Surface-based retrieval reduces perplexity of retrieval-augmented language models. *arXiv preprint arXiv:2305.16243*.

Michael W Eysenck and Mark T Keane. 2020. *Cognitive psychology: A student's handbook*. Psychology press.

Tongwen Huang, Qingyun She, and Junlin Zhang. 2020. Boostingbert: Integrating multi-class boosting into bert for nlp tasks. *arXiv preprint arXiv:2009.05959*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

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

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted LLMs as Chatbot Modules for Long Open-domain Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554. ArXiv:2305.04533 [cs].

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. MemoChat: Tuning LLMs to Use Memos for

Consistent Long-Range Open-Domain Conversation. ArXiv:2308.08239 [cs].

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Aaron Mueller, Kanika Narang, Lambert Mathias, Qifan Wang, and Hamed Firooz. 2023. Meta-training with Demonstration Retrieval for Efficient Few-shot Learning. ArXiv:2307.00119 [cs].

Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023a. Large language models as source planner for personalized knowledge-grounded dialogue.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023b. A survey of the evolution of language model-based dialogue systems. *arXiv preprint arXiv:2311.16789*.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2023c. Recursively Summarizing Enables Long-Term Dialogue Memory in Large Language Models. ArXiv:2308.15022 [cs].

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023d. Augmenting language models with long-term memory. *arXiv preprint arXiv:2306.07174*.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. ArXiv:2107.07567 [cs].

Ming Xu. 2023. Text2vec: Text to vector toolkit. `https://github.com/shibing624/text2vec`.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long Time No See! Open-Domain Conversation with Long-Term Persona Memory. ArXiv:2203.05797 [cs].

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

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

Jing Zhang, Xiaokang Zhang, Daniel Zhang-Li, Jifan Yu, Zijun Yao, Zeyao Ma, Yiqi Xu, Haohua Wang, Xiaohan Zhang, Nianyi Lin, et al. 2023a. Glm-dialog: Noise-tolerant pre-training for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2302.14401*.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. 2023b. Memory-augmented llm personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023c. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Kang Zhao, Wei Liu, Jian Luan, Minglei Gao, Li Qian, Hanlin Teng, and Bin Wang. 2023. UniMC: A Unified Framework for Long-Term Memory Conversation via Relevance Representation Learning. ArXiv:2306.10543 [cs].

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*.

A Appendix

A.1 Memory Database Generation Prompts

The design of the VariaMem dataset prompts is illustrated in Fig.5. The variable within the profile prompt is intended to be populated with the seed data. Meanwhile, the variables in the remaining prompts should be filled using the results generated from previous iterations.

A.2 Memory Planning Model

Within memory planning, we construct three memory planning (MP) instructions as shown in Fig.6 tailored to the memory type π, enhancing the models by prioritizing them based on the confidence in the validation performance of F1. Three distinct instructions $G = \{G_1, G_2, G_3\}$ are meticulously crafted and seamlessly integrated with the evalua-

Table 6: Comparison of Recall@5 and MMAP Results: Evaluation Between Settings Without Memory Planning Mechanism (w/o MP) and with Memory Planning Mechanism (MP)

Models	w/o MP				MP			
	Recall@5	MMAP	Correctness	Coherency	Recall@5	MMAP	Correctness	Coherency
Baichuan2-7B		0.5843	0.7650	0.8675		0.5927	0.7850	0.8875
Qwen-7B		0.6422	0.7775	0.8666		0.6621	0.8121	0.8993
ChatGLM2	0.8335	0.5647	0.7510	0.8492	0.8734	0.5736	0.7730	0.8531
ChatGLM3		0.5936	0.7618	0.8573		0.6247	0.7804	0.8891
ChatGPT		0.6348	0.7688	0.8923		0.6482	0.7781	0.9001

Profile Prompt: Please help me create a random profile for the above user? Include the following details: **[name]**, gender, nickname, title, age, **[occupation],** nationality, physical features, **[hobbies]**, achievements, ethnic background, **[educational background],** occupation, employer, awards and role models?

Semantic Memory Prompt: Relationships between individuals include family, friends, romantic partners, acquaintances, colleagues, mentors/mentees, neighbors, community members, and strangers. Based on **[profile description]**, can you help me randomly create relationships for **[name]** and provide their names? The answer should be in the JSON format like {relationship: {name:, description}})

Dialogue Prompt: (Please integrate **[episodic memory]** to generate a multi-turn, temporally related dialogue between **[name]** and the AI assistant. Requirements: Please note that the speakers are the AI assistant and [**name**] . Please use the appropriate titles. The dialogue should include entities such as time, characters, locations, and specific plot details. Please generate the JSON response in the following format:\n[{\"date\":,\"dialogue\":[**[name]** :, AI Assistant:, ...]}]

Episodic Memory Prompt: **(**Given **[profile description]**, please integrate **[relationship description]**, and the relationship between **[name]** and **[s_name]** is **[relationship]**. Generate episodic memories related to the events with **[name]** and **[s_name]** , as much as possible while retaining the entity names. **[topic cases])** The generated response should conform to the following JSON format: {date | topic | supporting character name | relationship | event | detailed description}

Figure 5: Prompts for profile, semantic memory, episodic memory, and dialogues generator.

tion questions q^{π}[3]. To ensure substantial margin loss between various memory types, we established a positive example $P = [G_w; q^{\pi+}]$ $(w = 1, 2, 3)$ by pairing each memory-type question q^{π} with an accurately corresponding instruction G_i and evaluation question. Simultaneously, we incorporated a memory-type question different from the one in focus as a negative example $N = [G_w; q^{\pi-}]$, thus ensuring a comprehensive approach to finding memory type accurately.

Following fine-tuning with the bert (Devlin et al., 2018) model, we employ the boost model (Huang et al., 2020) to facilitate the memory type identification process. We align each evaluation question with its corresponding instruction sentence, allowing real-life adjustments to the instruction based on task requirements. Adherent to the specified guiding instruction, we assign priority rankings: profile as 1, semantic memory as 2, episodic memory as 3, and dialogues as 4.

A.3 Comparative Performance of LLMs with and without Memory Planning Mechanism (MP)

We have compared the results of different large models under the conditions of whether the memory planning mechanism was used, focusing on Recall@5, MMAP, Correctness, and Coherency for additional illustration for Sec.5.2. From the experimental results, it can be seen that with the addition of the memory planning mechanism, the retrieval model achieved an improvement of approximately 4% under the setting of Recall@5. At the same time, there were varying degrees of improvement in MMAP and correctness. Notably, Qwen and ChatGLM achieved significant improvements. As for coherency, all models demonstrated considerable performance.

[3]The missing dialogue types are due to the richness of dialogue content. After confirming that the first three types are not suitable, we will then proceed to search for dialogues

Prompt under Direct Multi-Class Classification Settings

Please follow the judgment rules below to determine which type of memory information is suitable for answering the given question:

1.Determine if the question asks about personal information such as gender, nickname, title, age, occupation, nationality, physical characteristics, hobbies, achievements, racial background, education, employer, awards, and role models. If so, return 'profile'.

2.Determine if the question asks about information related to social relationships. If so, return 'semantic_memory'.

3.Determine if the question asks about personal experiences or events. If so, return 'episodic_memory'.

4.If none of the above, return 'dialogues'.

Memory Planning Instructions under Boost Settings

G1：Determine if the question asks about personal information such as gender, nickname, title, age, occupation, nationality, physical characteristics, hobbies, achievements, racial background, education, employer, awards, and role models. If so, return 1, else, return 0.

G2：Determine if the question asks about information related to social relationships. If so, return 1, else, return 0.

G3：Determine if the question asks about personal experiences or events. If so, If so, return 1, else, return 0.

Figure 6: Prompts and Instructions under memory planning task.