PerLTQA: A Personal Long-Term Memory Dataset for Memory Classification, Retrieval, and Synthesis in Question Answering

Anonymous ACL submission

Abstract

Long-term memory plays a critical role in personal interaction, considering long-term memory can better leverage world knowledge, historical information, and preferences in dialogues. Our research introduces PerLTQA [1], an innovative QA dataset that combines semantic and episodic memories, including world knowledge, profiles, social relationships, events, and dialogues. This dataset is collected to investigate the use of personalized memories, focusing on social interactions and events in the QA task. PerLTQA features two types of memory and a comprehensive benchmark of 8,593 questions for 30 characters, facilitating the exploration and application of personalized memories in Large Language Models (LLMs). Based on PerLTQA, we propose a novel framework for memory integration and generation, consisting of three main components: **Memory Classification**, **Memory Retrieval**, and **Memory Synthesis**. We evaluate this framework using five LLMs and three retrievers. Experimental results demonstrate that BERT-based classification models significantly outperform LLMs such as ChatGLM3 and ChatGPT in the memory classification task. Furthermore, our study highlights the importance of effective memory integration in the QA task.

1 Introduction

Long-term memory is an essential component of the human memory system, characterized by the ability to store extensive information and retrieve it when necessary. Spanning durations from mere minutes to an entire lifetime, this type of memory is fundamental to cognitive functions, as noted by (Tulving and Craik, 2000). Within the conversational domain, as explored by (Xu et al., 2021b; Zhong et al., 2023), integrating personal long-term memories can yield more personalized responses. Consequently, simulating the mechanisms of long-term memory is important for improving current dialogue systems.

Cognitive science classifies long-term memory into episodic and semantic types (Atkinson and Shiffrin, 1968). **Semantic memory** (Eysenck and Keane, 2020) is a mental representation that involves personal facts and world knowledge, such as profiles and relationships. It does not depend on the specific experience of the individual. While **episodic memory** (Eysenck and Keane, 2020) is about personal histories, specifically in events and dialogues. Previous research indirectly employs data aligning with episodic and semantic memory categories of cognitive psychology, despite not explicitly adopting its memory frameworks (Eysenck and Keane, 2020). Traditional QA systems (Kwiatkowski et al., 2019; Reddy et al., 2019; Chen et al., 2020b) and dialogue systems (Nakamura et al., 2022; Gopalakrishnan et al., 2023; Hu et al., 2023) initialize semantic memory as world knowledge from an external database. Existing dialogue systems consider dialogue history (Budzianowski et al., 2018; Zhang et al., 2018; Li et al., 2017; Zheng et al., 2019; Xu et al., 2021a) or dialogue summaries (Xu et al., 2021a; Chen et al., 2021) as episodic memory. Research on personalization within dialogue systems (Zhang et al., 2018; Zheng et al., 2019; Xu et al., 2022) has increasingly concentrated on the preferences in dialogues, uncovering their substantial influence on personal memory. We consolidate these findings into Table 1 and identify a notable research gap in integrating personal semantic memory concerning profiles, social relationships and episodic memory concerning events into the QA system.

To explore personal long-term memory in question answering (Nakamura et al., 2022), we propose the PerLTQA dataset, as depicted in Figure 1. The PerLTQA dataset is designed to capture the essence of semantic and episodic memories. It is

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

Dataset	Semantic Memory			Episodic Memory		Goal
	WK	PRO	SR	DLG	EVT	
Natural-QA (Kwiatkowski et al., 2019)	✓	✗	✗	✗	✗	QA on Wikipedia
CoQA (Reddy et al., 2019)	✓	✗	✗	✗	✗	Dialogue QA on world knowledge
HybridQA (Chen et al., 2020b)	✓	✗	✗	✗	✗	Multi-Hop QA on world knowledge
OTT-QA (Chen et al., 2020a)	✓	✗	✗	✗	✗	QA on tables and text
Multi-Woz (Budzianowski et al., 2018)	✗	✗	✗	✓	✗	Task-oriented Dialogue
Persona-Chat (Zhang et al., 2018)	✗	✓	✗	✓	✗	Consistent personality dialogue
DailyDialog (Li et al., 2017)	✗	✗	✗	✓	✗	Multi-turn dialogues on daily life
Personal-Dialogue (Zheng et al., 2019)	✗	✓	✗	✓	✗	Multi-turn personalized dialogues
MSC (Xu et al., 2021a)	✗	✓	✗	✓	✗	Long-Term open-domain conversation
DialogueSum (Chen et al., 2021)	✗	✗	✗	✓	✗	Dialogue summarization
Dulemon (Xu et al., 2022)	✗	✓	✗	✓	✗	Personal long-term Chinese conversation
HybridDialogue (Nakamura et al., 2022)	✓	✗	✗	✗	✗	Dialogue QA on tables and text
Topical-Chat (Gopalakrishnan et al., 2023)	✓	✗	✗	✗	✗	Knowledge-grounded open-domain conversations
ChatDB (Hu et al., 2023)	✓	✗	✗	✗	✗	Question answering with structured memory
MemoryBank (Zhong et al., 2023)	✗	✓	✗	✓	✗	Personal long-term memory dialogue
PerLTQA	✓	✓	✓	✓	✓	Question answering on personal long-term memory including semantic and episodic memory

Table 1: Typology of memories in QA/Dialogue datasets: Analysis of World Knowledge (WK), Profiles (PRO), Social Relationships (SR), Dialogues (DLG), and Events (EVT).

constructed using an in-context approach (Brown et al., 2020) that merges personal memory integration with contextual generation. This process produces an extensive representation of character-specific long-term memory, encompassing profiles (semantic memory), social relationships (semantic memory), events (episodic memory), and dialogues (episodic memory) as shown in Figure 1. PerLTQA features various QA items, including questions, answers, reference memories, and memory anchors (the key memory fragments in the answer are targeted to the question). The memory anchors and reference memory are used for aligning answers with their corresponding memories during evaluations. Our methodology involves three steps: **memory classification**, **memory retrieval**, and **memory synthesis**. Initially, we classify the given question to determine its memory type, ensuring an accurate categorization of memory. Next, we proceed to memory retrieval, where we utilize the probabilities derived from the classification stage to re-rank and prioritize the retrieved memories related to the question. Finally, in the memory synthesis stage, we generate the answers in LLMs by integrating the re-ranked memory information.

In summary, our research makes three main contributions:

- Our research contributes to introducing the PerLTQA dataset, encompassing a memory database with 141 profiles, 1,339 semantic social relationships, 4,501 events and 3,409 dialogues, and 8,593 memory-related evaluation questions.
- We propose three subtasks memory classification, memory retrieval, and memory synthesis to evaluate the memory utilization capabilities of LLMs. We carry out experiments using five LLMs and three retrieval models.
- Our experiment results indicate that the BERT (Devlin et al., 2018) model excels in memory classification tasks, surpassing LLMs like ChatGLM3 (Zhang et al., 2023a) and ChatGPT [2], and LLMs show varied proficiency in generating memory-based responses when provided with accurately retrieved memories.

2 Related Work

In cognitive psychology, semantic memory relates to world knowledge and social relationships, whereas episodic memory involves events. This differentiation is mirrored in the datasets like (Kwiatkowski et al., 2019; Chen et al., 2021; Zhong et al., 2023). In the realm of question answering (Kwiatkowski et al., 2019; Reddy et al., 2019; Chen et al., 2020b,a), Natural-QA (Kwiatkowski et al., 2019) and CoQA (Reddy et al., 2019) both target Wikipedia-based knowledge, exemplifying the use of world knowledge as semantic memory. Within dialogue tasks (Wang et al., 2023), MSC (Xu et al., 2021a) and Dulemon (Xu et al., 2022) consider dialogues as episodic memory. MemoryBank (Zhong et al., 2023) introduces a bilingual dataset using GPT-4 to summarize dialogues and personal data, effectively simulating episodic memory in multi-turn dialogues. However, existing datasets (Hu

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

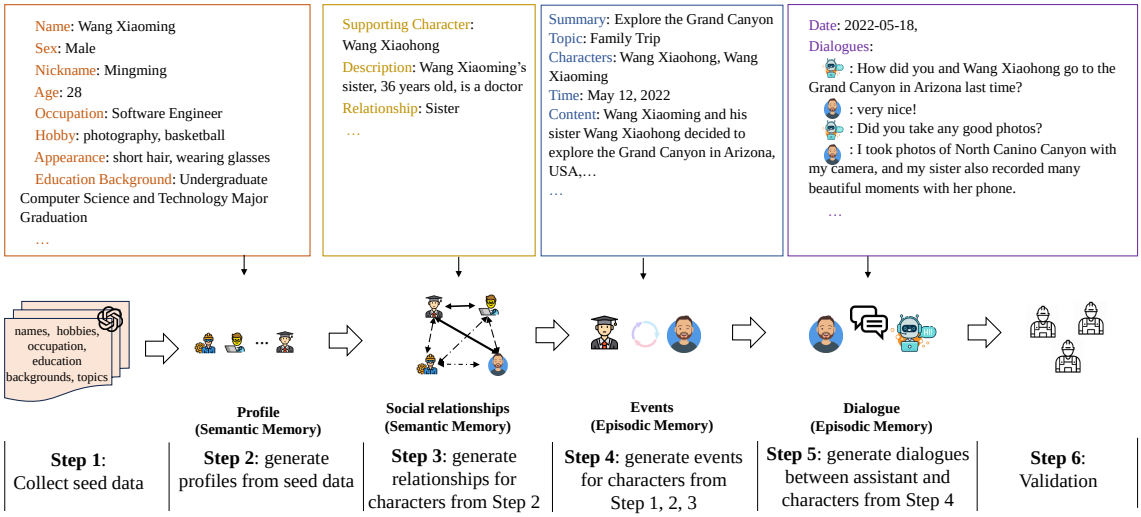

Figure 1: The process of PerLT Memory generation. A six-step process: Step 1. Seed data collection. Step 2. PRO generation. Step 3. SR generation. Step 4. EVT generation. Step5. DLG generation and Step 6. Validation.

et al., 2023; Zhang et al., 2023b) lack comprehensive coverage of both memory types with detailed annotations on social relationships and events, highlighting a research gap for LLMs in personal long-term memory synthesis.

Memory retrieval is a crucial component for QA and dialogue system to successfully generate memory-based response. The existing retrieval methods fall into three main categories: BM25 (Robertson et al., 1995), employing a statistical approach for document ranking based on query terms; DPR (Karpukhin et al., 2020), demonstrating supervised retrieval capabilities; and Contriever (Izacard et al., 2021), showcasing unsupervised retrieval techniques. With the rise of LLMs (Wang et al., 2023), an increasing number of works (Zhang et al., 2023b; Zhong et al., 2023) are utilizing the RAG (Retrieval-Augmented Generation) (Lewis et al., 2020) for retrieval-enhanced tasks. Within this framework, fine-tuned embeddings are employed for text similarity searches, such as REPLUG (Shi et al., 2023), OpenAI Embeddings [3]. This concept has been implemented in frameworks such as LlamaIndex [4], LangChain [5]. This approach leverages the refined embeddings to efficiently identify and retrieve content that is most relevant to the given query.

Aiming to integrate retrieved memories into responses, LLMs provide a prompt-based generation method facilitating the generation of memory-informed responses (Zhang et al., 2023a; Yang et al., 2023; Bai et al., 2023; Zhang et al., 2023c; Touvron et al., 2023). In the dialogue system (Zhao et al., 2023; Lee et al., 2023; Zhong et al., 2023), they incorporate memory into the prompt and generate the memory-related response. In this way, we can improve the relevance and specificity of the response, leading to context-aware responses of personal memory.

3 Dataset Collection

We detail the creation of the PerLTQA dataset, which involves collecting PerLT memories and generating and annotating PerLT QA pairs. Using in-context technique, we build a memory database and semi-automatic annotate memory-based Q&A pairs.

3.1 PerLT Memory Generation

As shown in Figure 1, the generation of PerLT memories is decomposed into six steps:

Step 1. Diverse Seed Data Collection. We select ChatGPT and Wikipedia as initial sources for our seed dataset due to their comprehensive coverage of a wide range of occupations, educational backgrounds, hobbies, and event topics, essential for foundational knowledge. 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. We implement a manual review pro-

[3] https://platform.openai.com/docs/api-reference/embeddings
[4] https://docs.llamaindex.ai/en/latest/index.html
[5] https://www.langchain.com/

cess, allowing us to avoid the unrealistic use for data generation.

Step 2. Profile (Semantic Memory) Generation. To study personalized memories, generating character profiles is essential. We leverage seed data, particularly occupations, educational backgrounds, hobbies inputs, within prompt templates that include descriptions of other attributes (gender, nickname, age, nationality, appearance, achievements, education, profession, employer, awards, and role models). By utilizing ChatGPT (`gpt-3.5-turbo`), we generate random character profiles. The detailed prompts for this process is available in Appendix.A.1.

Step 3. Social Relationship (Semantic Memory) Generation. For the development of diverse social connections, we utilize structured prompts shown in Appendix.A.1 to craft 50 distinct categories of relationships. These categories span a wide array, including but not limited to family, friends, romantic partners, acquaintances, colleagues, mentor/student dynamics, and neighbors, aiming to comprehensively cover social interactions.

Step 4. Event (Episodic Memory) Generation. Each character includes a series of narrative events, deeply embedded in their episodic memory and linked to interactions with others. The event generation starts by generating descriptions of background events chosen at random from the seed topics highlighted in Step 1. Following this step, we use prompts to help create detailed accounts of events that are deeply tied to these initial occurrences and the web of social connections. To ensure coherence between the dynamics of character interactions and the backdrop of events, few-shot learning techniques, as outlined by (Brown et al., 2020), are employed. This strategy aids ChatGPT (`gpt-3.5-turbo`) in achieving narrative consistency, weaving together individual events and relationships into a cohesive story for each character.

Step 5. Dialogues (Episodic Memory) Generation. Building on the events generated in Step 4, we craft historical dialogues between the AI assistant and the character. This process, anchored in historical events, ensures that conversations maintain relevance to past occurrences. We utilize prompt templates that merge character profiles and event details to help dialogue generation, as detailed in Appendix.A.1. Furthermore, embedding the dialogues maintains a profound connection to the shared histories and relationships.

Step 6. Validation. We start with small batches for quality checks and scale up after ensuring error-free outputs. We conduct random sampling of the generated memory data, identifying types of issues as detailed in Appendix A.3, and then manually refine the memories. This refinement includes removing anomalies in profiles, discriminatory content, inconsistencies in character memories, and brief event narratives, enhancing the accuracy and consistency of the memory.

3.2 PerLT Question Answering

To thoroughly assess each memory type for a character, we gather four QA-related metrics (*question*, *answer*, *reference memory*, and *memory anchor*) for evaluating the memory-based QA. The process of collecting PerLT QA items unfolds in three phases:

Question and answer generating. Utilizing ChatGPT, we generate questions and answers prompted by the memory sentences stored in PerLT Memory database. The answers are designed to align with the reference memories provided, adhering to the prompts we created, as shown in the Appendix.A.2.

Memory Anchor Annotation. The memory anchor, a key text segment in the answer that aligns with the referenced memory and question, is essential for memory evaluation in response generation. We employ exact match techniques and human verification to annotate the start and end positions of memory anchors, guided by the reference memory. Given the intensive labor involved in manual adjustments, we have annotated memory anchors for a limited set of 30 characters.

Validation on QA pairs and Memory Anchor. To ensure the integrity of PerLT QA pairs, we initiate our quality control with an unbiased random sampling. This is followed by a detailed categorization of errors in QA, references, and memory anchors, supplemented by thorough pronominal reference checks for accuracy. All error types are meticulously cataloged in the Appendix.A. We use LLMs as a scoring mechanism, evaluating on a scale from 0 to 10, directly accepting QA pairs with a score of 10, reviewing those scored between 7 and 9, and eliminating those below 6. Automated validation is conducted to ensure the accuracy of reference memories and to remove irrelevant stop words. This is followed by meticulous manual corrections and alignment checks between memory anchors and references, ensuring the highest qual-

Dataset Statistics		
Profiles	# Character profiles	141
	# Jobs	98
Semantic Memory	# Relationship Descriptions	1,339
	# Relationship Categories	50
	# Average Social Relationships per Character	9.5
Episodic Memory	# Topics	49
	# Events	4,501
	# Average Words per Events	313
	# Event-related Historical Dialogs	3,409
	# Utterances	25,256
	# Average Words per Utterance	43.7
Memory QA	# Question Answer Pairs	8,593
	# Average Words per Question	16.7
	# Average Words per Answer	27.4
	# Memory Anchors	23,697
	# Average Anchors	2.8

Table 2: PerLTQA dataset statistics.

ity of QA items.

3.3 Dataset Statistics

The PerLTQA dataset, presented in Table 2, includes 141 character profiles with detailed occupations and relationships. With 50 relationship categories, an average of 9.5 social relationships per character, the dataset provides a vavid social relationship for semantic memory. Furthermore, PerLT Memory features 4,501 events, averaging 313 words each, which fuel 3,409 event-related historical dialogues, totaling 25,256 utterances. In the QA section, 8,593 question-answer pairs and 23,697 memory anchors average 16.7 and 27.4 words, respectively. This rich compilation of data supports the development of dialogue QA system with a profound understanding of human-like memory recall and synthesis within a concise framework.

3.4 Task Definition

The PerLT memory database is formulated as $M = \{(S_i(l_1), E_i(l_2)) \mid i = 1, 2, \ldots, p\}$, where each tuple consists of semantic memories including profiles and social relationship and episodic memories including events and dialogs. Each $S_i(l_1)$ and $E_i(l_2)$ are defined to have l_1, l_2 elements, respectively, which are specific to the i-th character memory representation.

The PerLT QA dataset comprises a set of items $T = \{t_j\}_{j=1}^{N}$, where each item t_j is a tuple consisting of four elements: $t_j = (q_j, r_j, m_j, a_j)$. Here, q_j denotes the question, r_j the reference memory, m_j the memory anchor, and a_j the answer. The dataset spans various data types including semantic memory, and episodic memory, which are implicitly reflected in the construction of each t_j. The variable N represents the total number of QA items in the dataset.

As shown in Figure 2, to explore the integration of memory information in QA, we propose three subtasks: *memory classification*, *memory retrieval* and *memory synthesis* for response generation. In particular, memory synthesis is our ultimate goal.

Memory Classification. We introduce a classification model designed to assist queries in finding semantic memory or episodic memory. This model can operate through an instruction-based LLM, few-shot-based LLM, or BERT-based classifier. The classification model conforms to a unified formula as Eq.(1)

$$\pi = MC(q) \tag{1}$$

where π represents the classification result, MC denotes the classification model and q is the input query. The outputs from our classification model enhance memory retrieval by facilitating post-ranking of the retrieved memories, thereby reducing the excessive dependence on memory classification within the framework. Further details are elaborated in Appendix.A.4.

Memory Retrieval. We aim to perform memory retrieval by extracting relevant character memories for a given evaluation question from the PerLT memory database M, formalized as Eq.(2).

$$m, s = R(q, M, k) \tag{2}$$

where m is the retrieved memory with size k, s is the corresponding scores, R is the retrieval model.

Our method distinguishes itself by initially retrieving k memories from each category within the memory database, amassing $2k$ potential memory candidates. These candidates undergo a re-ranking process influenced by their classification scores, culminating in a composite score for each memory m_i, which is computed as follows:

$$s'_i = \alpha \cdot P(\pi|m_i) + \beta \cdot \text{sigmoid}(s_i) \tag{3}$$

where $P(\pi|m_i)$ is the probability given by the classification model that the memory item m_i belongs to π. The top k memories are then selected based on these final scores. α and β represent the weight of each term, and we set both to 0.5 to balance their contributions.

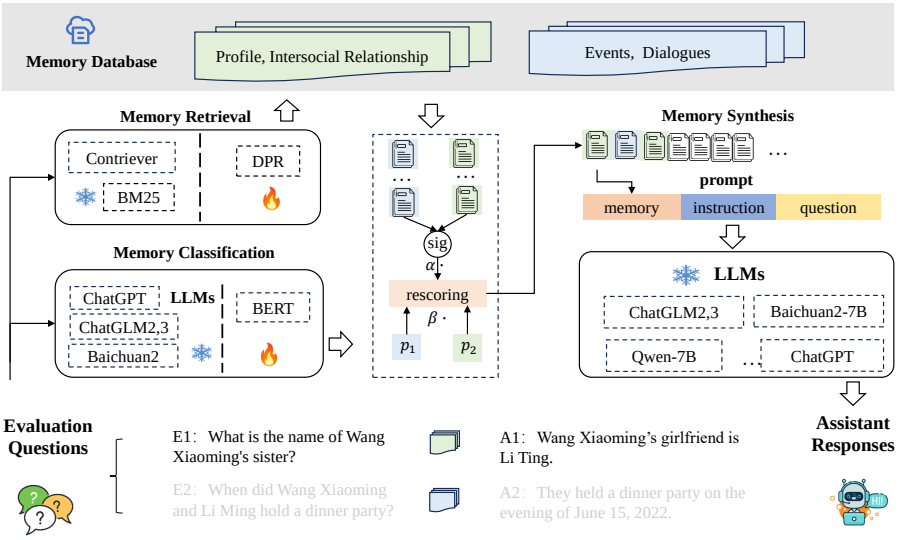

Figure 2: The framework of memory classification, memory retrieval and memory synthesis in QA.

Memory Synthesis. Memory synthesis leverages LLM for response generation. This task uses a prompt template z (as illustrated in Appendix.6), an evaluation question q, and retrieved memories m as Eq.(4).

$$r' = LLM(z, q, m) \tag{4}$$

3.5 Evaluation Metrics

For the memory classification task, we use precision (P), recall (R), F1, and Accuracy to serve as metrics. For the memory retrieval task, we utilize Recall@K (Manning et al., 2008) as our metric. To evaluate memory synthesis for the response generation task, we measure the correctness and coherence of responses with `gpt-3.5-turbo`-based evaluation method (Zhong et al., 2023) and use MAP (mean average precision) of memory anchors as shown in Eq.(5) to evaluate memory synthesis ability (Nakamura et al., 2022).

$$\text{MAP} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{EM}(q_i, mar_i)}{\text{NUM}(mar_i)} \tag{5}$$

where N represents the total number of questions in the evaluation dataset. mar denotes memory anchors, EM represents the tally of exact matches between queries and memory anchors, and $\text{NUM}(mar_i)$ is the count of memory anchors per question.

4 Experiments

4.1 Implementation details

In our work, we divide the data from the PerLT QA dataset into training (5155), validation (1719), and test sets (1719) for model training and evaluation. In the memory classification task, we fine-tune BERT-base model and compare the sentence classification results on the test dataset with ChatGLM2, ChatGLM3 (Zhang et al., 2023a), Baichuan2-7B-Chat (Yang et al., 2023), Qwen-7B-Chat (Bai et al., 2023), and ChatGPT under instructional and few-shot settings. For the memory retrieval task, we employ three retrieval models - DPR (Karpukhin et al., 2020), BM25 (Robertson et al., 1995), and Contriever (Izacard et al., 2021) - to collect character memories. In the memory synthesis task, we use the above five LLMs to generate responses of no more than 50 words, given re-ranked retrieved memories, employing in-context learning methods.

The memory synthesis task is evaluated across three scenarios: with memory classification and retrieval (W-MC+R), without memory classification but with retrieval (W/o-MC+W+R), and without both classification and retrieval (W/o-MC+R). Experiment details are shown in the appendix.A.5

4.2 Memory Classification

BERT-based model provides better performance than LLMs for memory classification. As shown in Table 4, BERT demonstrates superior performance compared to other LLMs under instruction and few-shot settings. Specifically, in few-shot scenarios where an evaluation question is paired with corresponding examples for each type of memory, the performance of `gpt-3.5-turbo` declines in comparison to methods that rely solely on instruction-based classification. In summary, the BERT-base model achieves the highest weighted precision

	W-MC+R			W/o-MC+W-R			W/o-MC+R		
	MAP	**Corr.**	**Coh.**	**MAP**	**Corr.**	**Coh.**	**MAP**	**Corr.**	**Coh.**
ChatGLM2	0.688	0.483	0.963	0.688	0.481	0.962	0.128	0.054	0.960
ChatGLM3	0.704	0.517	0.971	0.695	0.517	0.969	0.130	0.060	0.962
Qwen-7B	0.729	0.535	0.960	0.720	0.532	0.959	0.131	0.057	0.957
Baichuan2-7B	0.736	0.535	0.966	0.728	0.522	0.968	0.132	0.051	0.953
gpt-3.5-turbo	0.756	0.573	0.969	0.745	0.562	0.969	0.156	0.088	0.961

Table 3: Comparison of MAP, Correctness (Corr.), Coherency (Coh.) across three settings: With memory classification and retrieval (W-MC+R), without memory classification but with retrieval (W/o-MC+W-R), and without memory classification and without retrieval (W/o-MC+R).

Metrics	P	R	F1	Acc
ChatGLM2-6B	0.749	0.712	0.729	0.712
ChatGLM3-6B	0.864	0.485	0.538	0.485
Qwen-7B	0.730	0.631	0.673	0.631
Baichuan2-7B	0.848	0.602	0.657	0.602
gpt-3.5-turbo	0.868	0.668	0.715	0.668
F+ChatGLM2-6B	0.770	0.806	0.785	0.806
F+ChatGLM3-6B	0.778	0.445	0.508	0.445
F+Qwen-7B	0.804	0.402	0.452	0.402
F+Baichuan2-7B	0.860	0.324	0.337	0.324
F+gpt-3.5-turbo	0.864	0.511	0.566	0.511
I+BERT-base	0.720	0.849	0.779	0.849
BERT-base	**0.960**	**0.956**	**0.957**	**0.956**

Table 4: Comparative performance of five LLMs and BERT in memory classification tasks under few-shot settings (F) and instruction-based training (I).

RM	R@1	R@2	R@3	R@5	T(s)
Contriever	0.486	0.674	0.737	0.792	0.070
DPR	0.602	0.803	0.862	**0.919**	2.960
BM25	**0.705**	**0.847**	**0.871**	0.895	**0.030**

Table 5: Performance of Recall@K (R@K) and average retrieval time (T) in memory retrieval using Contriever, BM25, and DPR models.

(95.96%), weighted recall (95.64%), weighted F1 score (95.74%), and accuracy (95.64%). Moreover, the high performance in memory classification reinforces confidence in the rescoring mechanism, as illustrated in Figure 2.

4.3 Memory Retrieval

Different retrieval models show variable Recall@K and time performance. In the memory retrieval task, Table 5 reveals that the unsupervised retrieval model Contriever significantly lags behind the statistic-based BM25 and the supervised DPR model. Moreover, as the top k values increase, DPR notably improves Recall@K performance, surpassing BM25 after k equals 3. However, the retrieval time cost of DPR is substantially higher than BM25 retrieval. This suggests that we need to balance the retrieval performance and time cost when deployment in dialogue QA tasks.

4.4 Memory Synthesis

Memory classification and retrieval significantly improve LLMs to integrate memory into responses. The results shown in Table 3 indicate that LLMs augmented with memory classification and retrieval model show a marked improvement in generating memory-informed responses over those relying solely on LLMs itself, with notable increases in precision (MAP peaking at 0.756) and correctness (correctness reaching up to 0.573). The absence of memory classification (W/o-MC+W-R) has a minimal impact on improving MAP and Correctness, maintaining robust scores (0.688-0.745 for MAP), which emphasizes the importance of retrieval mechanisms in sustaining performance. Moreover, coherency levels are impressively stable across all configurations, never falling below 0.953, reflecting the strength of LLMs in producing coherent text. These results highlight the crucial importance of the retrieval model and maintaining coherence throughout the process. Additionally, models with fewer than 10 billion parameters have shown memory synthesis capabilities similar to those of ChatGPT, indicating that smaller models can also be optimized to produce outputs of comparable quality.

5 Analysis and Case Study

5.1 Ablation Study

Correct memory retrieval significantly enhances the accuracy of responses across various LLMs. The experimental results, as shown in Table 6, demonstrate the consistent ability of different LLMs to generate responses that are both relevant and accurate. This consistency underscores that

Models	NR		IR		CR	
	MAP	Corr.	MAP	Corr	MAP	Corr.
Baichuan2-7B	0.132	0.051	0.396	0.225	0.782	0.581
Qwen-7B	0.131	0.057	0.390	0.221	0.786	0.574
ChatGLM2	0.128	0.054	0.396	0.248	0.738	0.523
ChatGLM3	0.130	0.060	0.365	0.216	0.754	0.561
ChatGPT	0.156	0.088	0.375	0.252	0.842	0.609

Table 6: Performance of LLMs on MAP and Correctness (Corr.) under No Retrieval (NR), Incorrect Retrieval (IR) and correct retrieval (CR) settings.

	Semantic Memory			Episodic Memory		
	MAP	Corr.	Coh.	MAP	Corr.	Coh.
gpt-3.5-turbo	0.242	0.150	0.834	0.721	0.543	0.966

Table 7: The results of gpt-3.5-turbo performance solely on episodic or semantic memory on the metric of MAP, Correctness, and Coherency

LLMs experience a substantial improvement when they have access to accurate external memory. The findings further indicate that LLMs possess a degree of tolerance towards misinformation and are capable of leveraging accurate memory information to some extent. Despite incorrect memory retrieval, all models manage to sustain a reasonable degree of precision, with MAP scores from 0.365 to 0.396, underlining their robustness in less-than-ideal information conditions.

Semantic and episodic memory contribute to improving memory synthesis. As shown in Table 7, the results illustrate the comparative performance of gpt-3.5-turbo when integrating only with semantic or episodic memory. The provision of only one type of memory—either episodic or semantic, leads to varying levels of performance degradation. This variation can be attributed to the different proportions of episodic and semantic memory data within the test set, which in turn affects the response accuracy of the Large Language Models (LLMs). However, integrating any form of personal long-term memory, whether episodic or semantic, into LLMs proves beneficial for generating responses that are informed by past personal history. This suggests that the presence of personal long-term memory, regardless of its type, help LLMs to produce relevant and accurate responses.

5.2 Case Study

We present specific cases in Figure 3 to evaluate the question 'What is Wang Wei's occupation?' with the verifiable answer 'cameraman'. Without memory retrieval, *gpt-3.5-turbo* 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 response '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 classification mechanism. While case 2 retrieves relevant dialogues, it fails to retrieve essential semantic memory as in case 3. With memory classification, our models retrieve accurate social relationship memory, yielding correct responses. In this evaluation, with 'cameraman' as the memory anchor, only case 3 correctly incorporates the pertinent memory.

Question: What is Wang Wei's occupation?
Ground Truth Answer: Wang Wei is a cameraman.
Memory Anchor : cameraman

NR (W/o-MC+R) #case 1

R-1: Wang Wei is a teacher.
Memory Anchor Score: 0/1

IR (W/o-MC+WR) #case 2

Retrieved Memory:
AI Assistant: I heard that your cooperation with Wang Wei in the movie was very successful and received high praise. (episodic memory)
R-2 : Wang Wei is an actor.
Memory Anchor Score: 0/1

CR (W-MC+R) #case 3

Retrieved Memory: 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 : Wang Wei is a cameraman.
Memory Anchor Score: 1/1

Figure 3: Comparative analysis of response performance without retrieval (NR), incorrect retrieval (IR), and Correct Retrieval (CR).

6 Conclusion

Our study introduces the PerLTQA dataset, which features a memory database and memory-based QA pairs. This dataset encompasses personal long-term memory, including profiles, social relationships, events, and dialogues, divided into semantic and episodic memory categories. We also outline three subtasks: memory classification, retrieval, and synthesis, and report on baseline experiments with five LLMs and three retrievers. Our findings reveal that Bert-based memory classification surpasses other LLMs in categorizing memory types. We also note considerable differences among LLMs in generating accurate memory-based answers. This research significantly deepens the understanding and evaluation of LLMs in the context of personal long-term memory.

Limitations

In this work, we utilize `gpt-3.5-turbo` to generate a memory-based dataset and evaluate its ability to generate responses based on memory in three distinct subtasks. However, we acknowledge the following limitations: 1. The process of generating memory data in the PerLTQA 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 and 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 PerLTQA 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 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.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020a. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*.

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.

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*.

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

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

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*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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].

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. Hybridialogue: An information-seeking dialogue dataset grounded on tabular and textual data. *arXiv preprint arXiv:2204.13243*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Endel Tulving and Fergus IM Craik. 2000. *The Oxford handbook of memory*. Oxford University Press.

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

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

Jing Xu, Arthur Szlam, and Jason Weston. 2021b. 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 preprint arXiv:2203.05797*.

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*.

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 PerLT memory dataset prompts are illustrated in Figure 5. The "Profile Generation" prompt creates character profiles using specified seed data and a prompt template. Following this, the "SR (Social Relationship) Generator" prompt produces social relationships based on ten provided seed relationships. Additionally, the "EVT (Event) Generator" prompt is employed to create events that align with the established social relationships between characters. Lastly, the "DLG (Dialogue) Generator" prompt facilitates the generation of event-based dialogues between a character and an AI assistant. Collectively, these prompts enable our model to generate raw memory data effectively.

Profile Generation 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?

ISR Generation 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}})

EVT Generation 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}

DLG Generation 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:, ...]}]

Figure 4: Prompts for PRO, SR, EVT, and DLG memory generator.

A.2 Memory QA items Generation Prompts

The design of the PerLT QA generation prompts are illustrated in Figure 4. The "Question and Answer Generation" prompt is designed to create questions and answers based on a provided reference memory and character name. Additionally, the "Memory Anchor Candidates Searching" prompt is utilized to identify key fragments that are crucial for crafting questions. These fragments are specifically chosen because they are present both in the generated answer and in the reference answer, ensuring relevance and coherence.

Question and Answering Generation Prompt

Based on the provided memory information, construct question-answer pairs and return them as a JSON array [{Q, A}], where Q and A are the keys that represent question and answering respectively.

Memory Anchor Candidates Searching Prompt

Based on the provided question-and-answer pair, identify the correct key answer word(s) from the response. Here is the given example:
Question: When Zhou Ting's family was planning their summer vacation, who took the initiative to help arrange the itinerary?
Answer: Zhang Tao took the initiative to help with the planning.
Memory Anchor Candidates: ["Zhang Tao"]

Question: [question]
Answer: [answer]
Memory Anchor Candidates:

Figure 5: Prompts for question answering generation, and memory anchor candidate searching.

A.3 Dataset Generation Error Types

In the dataset generation process for PerLT Memory and PerLT QA, several categories of errors are identified and corrected as shown in Table 8. Anomalies, such as missing information in profiles, are rectified by removing or emptying the faulty fields. Incorrect character relationships that do not provide sufficient event data are excluded from the dataset. Instances of brief event narratives without detailed information are eliminated. Referent errors, which include incorrect or ambiguous references, are replaced with accurate information to ensure clarity. Redundant answers are streamlined to avoid unnecessary repetition, ensuring concise and relevant data. Finally, blurred memory anchor boundaries are corrected to precisely reflect the intended memory cues. These steps are taken to enhance the accuracy and reliability of the dataset.

A.4 Optimizing Memory Retrieval with Memory Classification Re-Ranking

We devise a method in which the output probabilities of the classification model are utilized to furnish the retrieval model with classification insights, allowing for the re-ranking of candidate memories. This strategy minimizes the risks associated with memory retrieval based on specific memory bank classification results. Such risks primarily stem from potential classification inaccuracies that could lead to memory retrieval from an incorrect memory type, thereby unduly influencing the reliance on classification model precision within the framework. The introduction of a re-ranking strategy ensures the retrieval of a predefined number

Error Type	Source	Error Example	Operation	Revision
Anomalies in profiles	PerLT Memory	{hobbies: "Not Provided"}	Remove	{hobbies: ""}
Invalid character relationship	PerLT Memory	Zheng Yong has a wife and girlfriend at the same time.	Remove	Remove the relationship wife or girlfriend which not provide enough events data.
Brief event narratives	PerLT Memory	Xiaoming's father used to participate in the activities.	Remove	-
Referent error	PerLT QA	When will Wang Xiaoming and the AI assistant plan to visit the exhibition?	Replace	When will Wang Xiaoming and Wang Xiaohong plan to visit the exhibition?
Redundant answer	PerLT QA	Who is the mentor of Wangxiaoming? Wangxiaoming's mentor is Zhangwen.	Reduce	Zhangwen.
Blurred Memory anchor boundaries	PerLT QA	Answer: They met at Bali Memory Anchor:["At Bali"]	Correct	Answer: They met at Bali Memory Anchor:["Bali"]

Table 8: The error types observed in PerLT Memory and QA items generation and revision by human.

of memories across all memory types, regardless of the initial confidence levels of classification results. This is achieved through a weighted score re-ranking mechanism that effectively reduces the influence of classification inaccuracies on the ultimate ranking. For those instances with high classification confidence, revising their scores and reordering them accentuates their relevance, thereby optimizing the retrieval process.

Answer Generation Prompt:

Please answer the following question based on the provided memory information, ignoring any irrelevant memories. Keep the response under fifty words.

Memory Information: [memories]
Question: [question]
Answer:

Figure 6: Prompts for answering generation.

A.5 Experiment Settings

Memory Classification settings. We conduct binary-class classification experiments on semantic memory, and episodic memory using BERT, Baichuan, ChatGLM2, ChatGLM3, and ChatGPT. For BERT, we employ fine-tuning with the evaluation questions to predict the memory type. For LLMs, we use instructions to guide LLMs in predicting the memory type. We also conduct instruction augmentation BERT experiments. Specifically, we train BERT-base classification models with 7,516 QA pairs. We finally evaluate the performance of memory type classification on a test set of 1,719 evaluation questions.

Memory Retrieval settings. We create unique memory banks for each character. In the case of DPR, we train the DPR model using 7516 evaluation questions. Contriever uses the text2vec model (Xu, 2023) from Hugging Face to calculate the similarity between memory sentences and questions.

Memory Synthesis settings. In the W-MC+R setting, responses are generated using retrieved memories that are post-ranked based on memory classification outcomes. Conversely, in the W/o-MC+W+R scenario, responses are produced solely through memory retrieval, without the aid of memory classification for re-ranking. Meanwhile, in the W/o-MC+R framework, responses are generated directly without utilizing any external memory, relying solely on the inherent knowledge in LLMs. These configurations not only validate the effectiveness of each component but also underscore the importance of external memory. Due to limited resources, we only evaluated 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.