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

**Anonymous ACL submission**

## Abstract

In conversational AI, effectively employing long-term memory improves personalized and consistent response generation. Existing work only concentrated on a single type of long-term memory, such as preferences, dialogue history, or social relationships, overlooking their interaction in real-world contexts. To this end, inspired by the concept of semantic memory and episodic memory from (Eysenck and Keane, 2020), we create a new and more comprehensive dataset, coined as PerLTQA, in which world knowledge, profiles, social relationships, events, and dialogues are considered to leverage the interaction between different types of long-term memory for question answering (QA) in conversation. Further, based on PerLTQA, we propose a novel framework for memory integration in QA, consisting of three subtasks: **Memory Classification**, **Memory Retrieval**, and **Memory Fusion**, which provides a comprehensive paradigm for memory modeling, enabling consistent and personalized memory utilization. This essentially allows the exploitation of more accurate memory information for better responses in QA. We evaluate this framework using five LLMs and three retrievers. Experimental results demonstrate the importance of personal long-term memory in the QA task[1].

## 1 Introduction

Long-term memory is a crucial element in conversational communication, facilitate the consistent and personalized response generation(Xu et al., 2021b; Zhong et al., 2024). Previous studies, as shown in Table 1, have explored its various aspects, such as world knowledge(Kwiatkowski et al., 2019; Reddy et al., 2019; Chen et al., 2020b), profiles (Zhang et al., 2018; Zheng et al., 2019; Xu et al., 2022), social relationships, events (Jang et al., 2023), and dialogue history (Zhong et al., 2024; Maharana et al., 2024; Xu et al., 2021a; Chen et al., 2021).

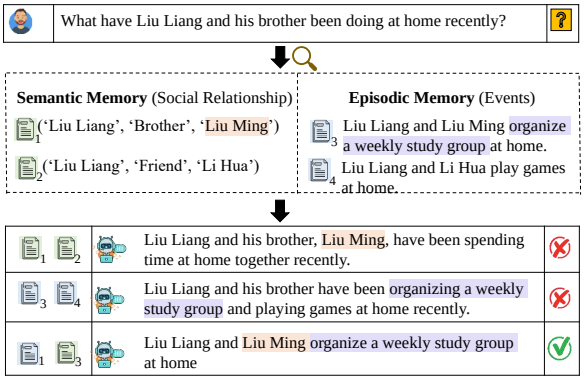


Figure 1: Example of external semantic and episodic memory used for QA in conversation.

However, existing research largely focused on a single type of long-term memory, ignoring the interaction of different types of memory, which are crucial for consistent and personalized response generation. As illustrated in Figure 1, with only event memory, the model cannot recognize social relationship *brother* in the query and fails to distinguish the event involving *LiuMing*. However, when integrating semantic and episodic memory, not only does it enhance the retrieval model (Izacard et al., 2021) to recall social relationships *LiuMing* but also aids generation model to accurately fuse the event *organize a weekly study group*. Based on the definition provided by cognitive psychology (Eysenck and Keane, 2020), long-term memory is categorized into semantic memory and episodic memory. Semantic memory encompasses structured data, including world knowledge, profiles, and relationships. In addition, episodic memory captures personal histories such as events and dialogues, typically represented as lengthy text. Combining these types of memory can enhance the retrieval of accurate memory, thus improving responses to user queries.

To establish a unified long-term memory bank, we leverage the in-context generation capabilities

[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., 2024) | ✗ | ✓ | ✗ | ✓ | ✗ | Personal long-term memory dialogue |
| CONVERSATION CHRONICLES (Jang et al., 2023) | ✗ | ✗ | ✓ | ✓ | ✓ | Long-term multi-session open domain conversation |
| 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).

of large language models (LLMs) to generate various memory categories: world knowledge, profiles, social relationships, events, and dialogue history, as illustrated in Figure 2. The dataset consists of 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.

In the realm of long-term memory research (Zhong et al., 2024; Stacey et al., 2024; Packer et al., 2023), retrieval models (Karpukhin et al., 2020; Izacard et al., 2021; Robertson et al., 1995) and generative models (Yang et al., 2023; Bai et al., 2023; Touvron et al., 2023; Zhang et al., 2023a; Jiang et al., 2023) are the two most commonly used modules to integrate external long-term memory. Furthermore, considering the variety of memory types examined in PerLTQA, classification models provide an effective means to refine the scope of retrieval and improve response consistency. Therefore, we propose three subtasks memory classification, memory retrieval, and memory fusion to evaluate the memory utilization capabilities of LLMs. We carry out experiments using five LLMs and three retrieval models.

The main contributions of this work are summarised as follows:

- We introduce a new personal long-term memory dataset, coined as PerLTQA, for QA. The PerLTQA provides a new research paradigm for the modeling of interaction between different memory types, paving the way for personalized question-answering systems and lifelong companion agents.
- We propose a new framework consisting of three subtasks memory classification, memory retrieval, and memory fusion to evaluate the memory utilization capabilities of LLMs.
- We carry out experiments using five LLMs and three retrieval models. The results demonstrate that a classification-based re-ranking mechanism improves the consistency of responses generated by LLMs when accessing unified long-term memory.

## 2 Related Work

The long-term memory differentiation is mirrored in the datasets like (Kwiatkowski et al., 2019; Chen et al., 2021; Zhong et al., 2024). 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., 2024) 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 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 fusion.

Efficient retrieval methods for external memory in dialogue system fall into two main cate-

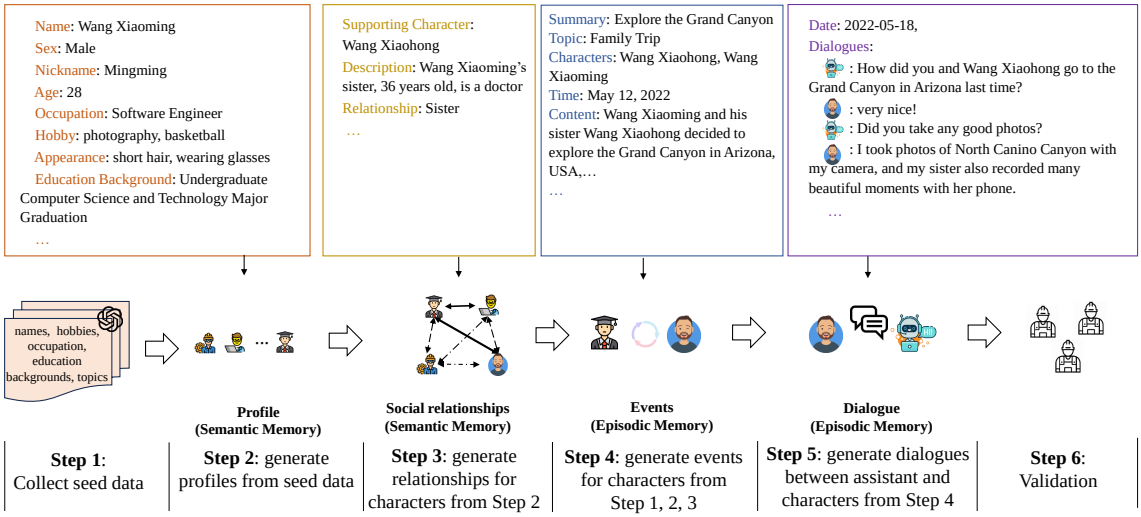


Figure 2: The process of PerLT Memory generation. A six-step process: Step 1. Seed data collection. Step 2. PRO generation. Step 3. SR generation. Step 4. EVT generation. Step 5. DLG generation and Step 6. Validation.

gories: sparse retrieval method like BM25 (Robertson et al., 1995) and vector-based retrieval method like DPR (Karpukhin et al., 2020), Contriever (Izacard et al., 2021). The use of Retrieval-Augmented Generation (RAG) is increasingly enhancing retrieval tasks within large language models (LLMs). Within this framework, fine-tuned embeddings are employed for text similarity searches, such as REPLUG (Shi et al., 2023), OpenAI Embeddings [2]. This integration helps generate context-aware responses that consider personal memory, thereby improving the interaction quality in systems like those documented in recent studies and platforms like LangChain [3] and LlamaIndex [4].

With the aim of integrating the memories recovered in the responses, LLMs provide the consistent response generation method based on prompts (Zhang et al., 2023a; Yang et al., 2023; Bai et al., 2023; Zhang et al., 2023c; Touvron et al., 2023). In dialogue systems, this approach incorporates memory directly into prompts, generating tailored responses that reflect individual memory (Zhao et al., 2023; Lee et al., 2023; Zhong et al., 2024).

## 3 Dataset Collection

We detail the creation of the PerLTQA dataset, which involves collecting PerLT memories and generating and annotating PerLT QA pairs. Using an in-context technique, we build a memory database that encompasses profiles, social relationships, world knowledge, events, and dialogues. We then semi-automatically annotate components of memory-based Q&A, including questions, answers, reference memories, and memory anchors that connect answers to their respective memories.

### 3.1 PerLT Memory Generation

As shown in Figure 2, the generation of PerLT memories is decomposed into six steps:

**Step 1. Diverse Seed Data Collection.** We select ChatGPT and Wikipedia as initial world knowledge source for our seed dataset due to their comprehensive coverage of a wide range of occupations, educational backgrounds, hobbies, and event topics, essential for foundational world 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 process, 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 charac-

[2] https://platform.openai.com/docs/api-reference/embeddings

[3] https://www.langchain.com/

[4] https://docs.llamaindex.ai/en/latest/index.html

ter 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, colleagues 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 social networks. 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 events. 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. Even so, there still be some biases as shown in Limitations.

### 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 start with unbiased random sampling and a detailed error categorization in QA, references, and memory anchors, alongside pronominal reference checks for accuracy, with all errors cataloged in the Appendix.A. We employ LLMs to score QA pairs on a scale from 0 to 10, automatically accepting those scoring 10, reviewing scores between 6 and 9, and discarding scores below 6. This process includes automated validation to verify reference memory accuracy and remove irrelevant stopwords, followed by thorough manual corrections and alignment checks to guarantee the highest quality 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 vivid 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

| 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.

words, respectively. This rich compilation of data supports the development of dialogue QA system with a profound understanding of human-like memory recall and fusion 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 3, to explore the integration of memory information in QA, we propose three subtasks: *memory classification*, *memory retrieval* and *memory fusion* for response generation. In particular, memory fusion 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 $\pi$ denotes the classification result, $MC$ is the classification model, and $q$ is the input query. The outputs from our classification model improve memory retrieval by assisting in the post-ranking of various types of retrieved memories, thereby reducing the over-reliance on memory classification. Further details are elaborated in Appendix.A.4.

**Memory Retrieval.** For each character, we perform memory retrieval for a given evaluation question from the PerLT memory database $M$ separately, 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 $\pi$. The top $k$ memories are then selected based on these final scores. $\alpha$ and $\beta$ represent the weight of each term, and we set both to 0.5 to balance their contributions.

**Memory Fusion.** Memory fusion leverages $LLM$ for response generation. This task uses a prompt template $z$ (as illustrated in Appendix.8), 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 fusion for the response generation

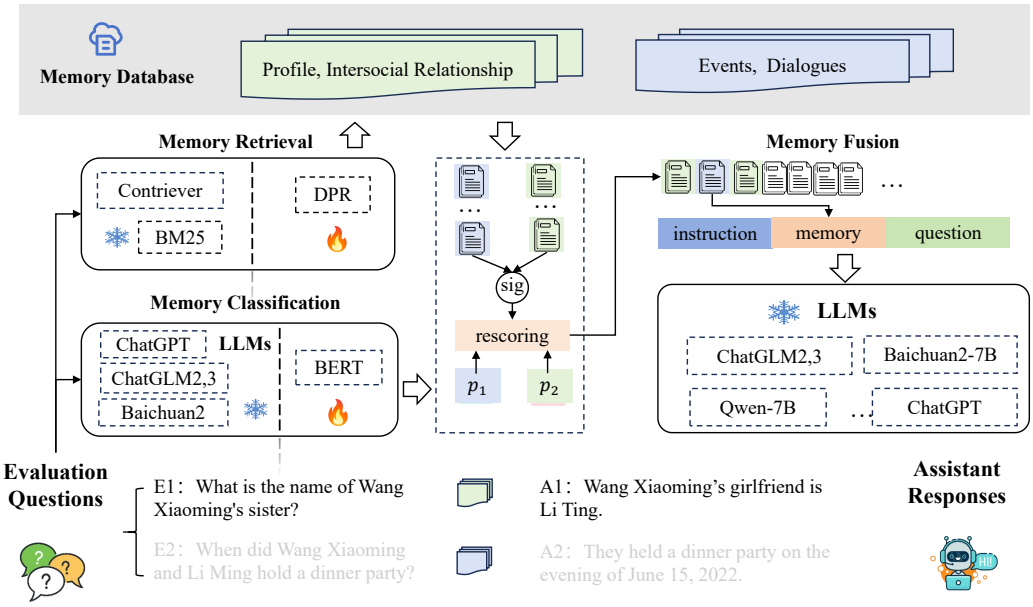


Figure 3: The framework of memory classification, memory retrieval and memory fusion in QA.

task, we measure the correctness and coherence of responses with `gpt-3.5-turbo`-based evaluation method (Zhong et al., 2024) and use MAP (mean average precision) of memory anchors as shown in Eq.(5) to evaluate memory fusion ability (Nakamura et al., 2022).

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

where N represents the total number of questions in the evaluation dataset. $mar$ denotes memory anchors, EM represents the tally of exact matches between answers 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 performance 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 fusion task, we use the above five LLMs to generate responses of no more than 50 words, given reranked retrieved memories, employing in-context learning methods.

The memory fusion 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 (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 3.

| | 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).

| **Models** | **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 |
| P+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 prompt-based training (P).

| **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.

### 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.

| **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.

### 4.4 Memory Fusion

**Memory classification and retrieval significantly improve LLMs to integrate memory into responses.** The results in Table 3 indicate LLMs enhanced with memory classification and retrieval models significantly improve the generation of personally consistent responses, with notable increases in precision (MAP peaking at 0.756) and correctness (up to 0.573). Without memory classification, robust scores decrease (MAP 0.688-0.745), underscoring the vital role of memory classification. Coherency remains consistently high across configurations, never falling below 0.953, highlighting the ability of LLMs to produce coherent text. Additionally, smaller-scale LLMs can achieve performance similar to ChatGPT, demonstrating that even less complex models can be optimized to deliver comparable output 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 accurate memory based responses. This consistency underscores that LLMs experience a substantial improvement when they

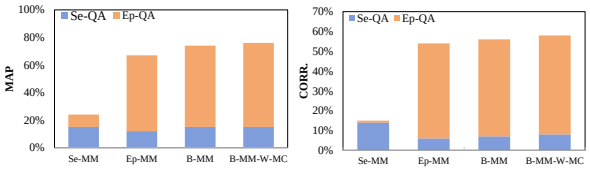


Figure 4: Evaluation results by memory type in Corr. and MAP metrics with different external memory configurations: Semantic Memory Only (Se-MM), Episodic Memory Only (Ep-MM), and Both (B-MM), Both with memory classifier (B-MM-W-MC).

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.

**Episodic and semantic memories enhance each other and improve memory fusion performance.** As shown in Figure 4, the results demonstrate that lacking any memory type significantly compromises the evaluation performance. Notably, even with only one memory type present like semantic memory, the system could still correctly address some questions related to the missing episodic memory, suggesting possible mutual enhancement between memory types. However, while including all memory types improves overall correctness and MAP, performance for individual memory types decreases compared to when only one memory type is used. This indicates that mixing memory types introduces additional noise, a prevalent issue with mixed interference. Compared to the mix retrieval, our soft classification mechanism improve performance for both memory types, emphasizing the importance of distinguishing memory features for better integration.

### 5.2 Case Study

We present specific cases in Figure 5 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

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

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.

## 6 Conclusion

Our study introduces the PerLTQA dataset, which includes a memory database and memory-based question-answer pairs, covering personal long-term memory such as profiles, social relationships, events, and dialogues, categorized into semantic and episodic types. We outline three subtasks—memory classification, retrieval, and fusion—and report baseline experiments involving five large language models (LLMs) and three retrievers. Our findings indicate that Bert-based classifiers excel at categorizing memory types compared to other LLMs. Additionally, we observe significant variances among LLMs in producing accurate memory-based responses. We also discover that enhancing personalization and consistency in responses requires integrating the unique characteristics of various memory types with those of different retrieval models. Future research should focus on refining retrieval models to better manage complex memory structures and on minimizing irrelevant noise in the context, thus improving the quality of responses generated by LLMs.

## 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. This dataset may exhibit certain biases, which are evident in several key aspects. Firstly, the range of names and nationalities included in the dataset is relatively limited, which may lead to potential discrepancies between the generated character events and the actual era, cultural background, and professional experiences of the characters. Secondly, due to the step-by-step generation process and the use of relatively uniform prompts, the diversity of the generated data remains constrained. Consequently, these biases make the dataset more suited for simulating personal narratives and science fiction scenarios, rather than accurately reflecting real-life situations. When utilizing this dataset, it is important to consider these limitations to avoid misinterpretations or inappropriate applications.

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


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.

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*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. *arXiv preprint arXiv:2310.13420*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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].

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*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

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*.

Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.

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*.

Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor, and Anders Johannsen. 2024. Lucid: Llm-generated utterances for complex and interesting dialogues. *arXiv preprint arXiv:2403.00462*.

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.

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, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

# A Appendix

## A.1 Memory Database Generation Prompts

The design of the PerLT memory dataset prompts are illustrated in Figure 7. 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 6: 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 6. 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 7: 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 7. 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 7: 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 8: Prompts for answer 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 Fusion 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.