Mem-PAL: Towards Memory-based Personalized Dialogue Assistants for Long-term User-Agent Interaction

Zhaopei Huang[1], Qifeng Dai[2], Guozheng Wu[1], Xiaopeng Wu[2], Xubin Li[2], Tiezheng Ge[2], Wenxuan Wang[1], Qin Jin[1]*

[1]Renmin University of China
[2]Taobao & Tmall Group of Alibaba
huangzhaopei@ruc.edu.cn, wuguozheng@ruc.edu.cn, qjin@ruc.edu.cn

Abstract

With the rise of smart personal devices, service-oriented human-agent interactions have become increasingly prevalent. This trend highlights the need for personalized dialogue assistants that can understand user-specific traits to accurately interpret requirements and tailor responses to individual preferences. However, existing approaches often overlook the complexities of long-term interactions and fail to capture users' subjective characteristics. To address these gaps, we present **PAL-Bench**, a new benchmark designed to evaluate the personalization capabilities of service-oriented assistants in long-term user-agent interactions. In the absence of available real-world data, we develop a multi-step LLM-based synthesis pipeline, which is further verified and refined by human annotators. This process yields **PAL-Set**, the first Chinese dataset comprising multi-session user logs and dialogue histories, which serves as the foundation for PAL-Bench. Furthermore, to improve personalized service-oriented interactions, we propose **H^{2}Memory**, a hierarchical and heterogeneous memory framework that incorporates retrieval-augmented generation to improve personalized response generation. Comprehensive experiments on both our PAL-Bench and an external dataset demonstrate the effectiveness of the proposed memory framework.

Code, Dataset and Appendix —
https://github.com/hzp3517/Mem-PAL

Introduction

The development of mobile internet has significantly enhanced interactions between users and their personal smart devices, such as using smart bands to monitor health, sending messages via smartphones, or conversing with virtual assistants. We refer to this pattern of communication as *user-agent interaction*, where the agent can access both the user's behavioral history and prior dialogue context. In such scenarios, an ideal service-oriented assistant should effectively leverage this interaction history for personalized modeling. This allows the assistant to deliver tailored solutions that align with individual user preferences and needs, without requiring cumbersome and repetitive explanations (Ha et al. 2024), much like seeking help from a familiar pal.

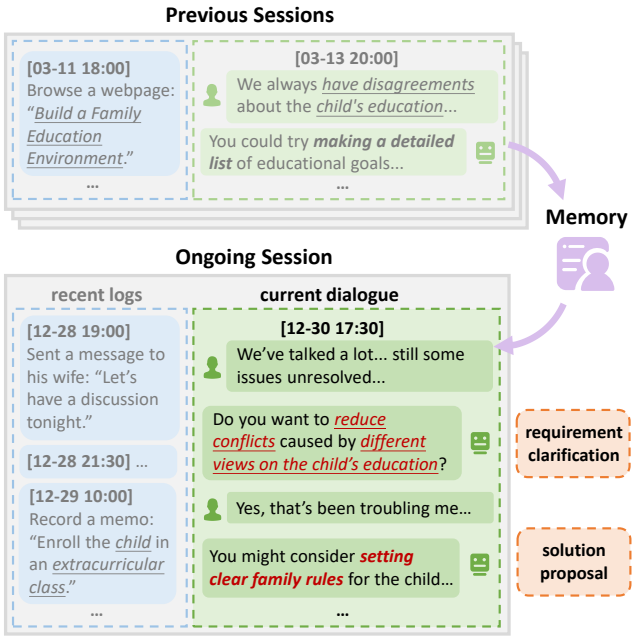

Figure 1: An example of our long-term, multi-session user-agent interaction data. The assistant is expected to leverage the historical interaction data (shown in lighter color) for memory modeling, enabling a more accurate understanding of user requirements and delivery of preference-aligned responses in the current dialogue.

Despite growing interest in service-oriented dialogue systems (e.g., medical assistants (Zhang et al. 2023), psychological counseling (Zhang et al. 2024c,a), and mobile virtual assistants (Guan et al. 2024), etc.), most existing approaches treat users uniformly and lack the capacity to generate truly personalized responses. On the other hand, some recent studies have begun exploring long-term dialogue scenarios (Xu, Szlam, and Weston 2022; Maharana et al. 2024; Wu et al. 2025b), introducing benchmarks that focus on retrieving personal facts from long-term interactions. However, these efforts often overlook the more subjective and nuanced task of modeling user preferences and individualized requirements. Moreover, their histories are typically limited

*Qin Jin is the corresponding author.

to dialogues, ignoring user behavioral records—an essential component of real-world user-agent interactions. To the best of our knowledge, only Wang et al. (2024b) have considered app screenshots as a form of behavioral history, but their dataset is not publicly available due to privacy concerns, which limits progress in developing and evaluating personalized service-oriented systems.

To address these challenges, we propose the first benchmark for personalized user-agent interaction, **PAL-Bench**. It introduces three evaluation tasks—*Requirement Restatement*, *Solution Proposal*, and *Multi-turn Dialogue Interaction*—designed to evaluate the capability of service-oriented dialogue assistants to understand and adapt to users' personalized requirements and preferences. Since collecting real-world long-term interaction data is costly and constrained by privacy concerns, we design a multi-stage, LLM-based data synthesis pipeline that incorporates verification and refinement procedures, resulting in a **PAL-Set**. This Chinese-language dataset contains both user behavior logs and user-assistant dialogues. As shown in Figure 1, PAL-Set captures realistic long-term interaction patterns: it features 100 users, each with an average of 29 sessions, 996 behavioral logs, and 401 dialogue turns. We also perform a human evaluation on the dataset, which confirms the high quality of the generated data.

To further support personalized modeling in these scenarios, we propose **H^{2}Memory**, a hierarchical and heterogeneous memory framework. In contrast to previous long-term memory modeling approaches (Wang et al. 2025; Yuan et al. 2025; Zhong et al. 2024; Ong et al. 2024), our approach explicitly models different forms of user history (e.g., behaviors versus dialogues) and introduces a two-level memory storage tailored to capture subjective user requirements and preferences, with update mechanisms for persona dynamics. Equipped with H^{2}Memory and a retrieval-augmented generation (RAG) strategy, assistants are better positioned to serve personalized, context-aware responses, as required by PAL-Bench. We also validate the generalizability of our method on an external dataset.

Our contributions are threefold: (1) We present PAL-Bench, the first Chinese benchmark for long-term user-agent interactions supported by a scalable LLM-based data synthesis and human refinement pipeline. It features three tasks focused on modeling user personalized requirements and preferences. (2) We propose H^{2}Memory, a hierarchical and heterogeneous memory framework that supports effective modeling, retrieval, and updating of diverse interaction histories for personalized service delivery. (3) We demonstrate the quality of our PAL-Bench dataset through human evaluation, and validate the effectiveness of H^{2}Memory via comprehensive experimental analysis on both PAL-Bench and an external dataset, advancing research in personalized dialogue systems.

PAL-Bench

Dataset Construction

User interaction records over the long term often reflect stable user profiles and traits at the macro level, while exhibiting dynamic changes at finer granularity. To simulate such patterns, we design a multi-stage generation pipeline (Figure 2). We first define each user's overall profile and create a corresponding persona. The persona is then expanded into multiple session-specific scenarios, which serve as fine-grained control signals for synthesizing interaction records. All synthesis steps are automatically performed using Qwen2.5-Max (Team 2024), followed by verification and refinement steps to ensure data quality.

Profile and Persona We begin by generating a basic profile for each user, including gender, age, personality, and brief descriptions across four aspects: work, health, family, and leisure. Building on this profile, we further synthesize a persona that includes: (i) a personal timeline spanning several months, outlining monthly objective events to guarantee temporal coherence; and (ii) a set of user traits, covering multiple general requirement types with corresponding preference descriptions, to support subjective consistency over long-term interactions.

Session-specific Information Since our ultimate goal is to synthesize multi-session interaction records, we need to expand the initial profile and persona into multiple session-specific pieces of information, which serve as references for the subsequent interaction records synthesis. We first expand the timeline for each month into multiple requirement-oriented situation entries. Each situation entry is a brief description composed of a few sentences, revolving around several requirement types predefined in the user's traits. We then expand each situation entry into a diary-style experience description that captures rich, detailed, and behaviorally grounded events over the same period. These experiences will help synthesize subsequent objective, time-stamped logs. For the guidance of synthesizing dialogues, we further construct a dialogue framework based on each situation.

Specifically, each dialogue framework may consist of multiple *topics*, with each topic containing two parts: "Requirements" and "Solutions". The "Requirements" part includes three components: *user query*, *implicit needs*, and *requirement*. The *user query* represents the user's initial question for each topic, typically brief and underspecified to reflect real-world user behavior. The *implicit needs* section includes two entries that are not explicitly stated in the query but are relevant to the user's background or experiences and are expected to be inferred by the assistant. The *requirement* section combines the user query and implicit needs described above to offer a complete description of the user's intent. For the "Solutions" part, we first generate 8 diverse candidate solutions that are all considered generally reasonable. Then we identify 2 positive solutions (highly aligned with the user's preferences) and 2 negative solutions (not aligned) among these candidates based on the predefined user profile and persona.

Interaction Records The long-term interaction records are the final output of our synthesis pipeline, comprising multi-session logs and dialogues that are accessible to the assistant. For log synthesis, we prompt the LLM to simulate a

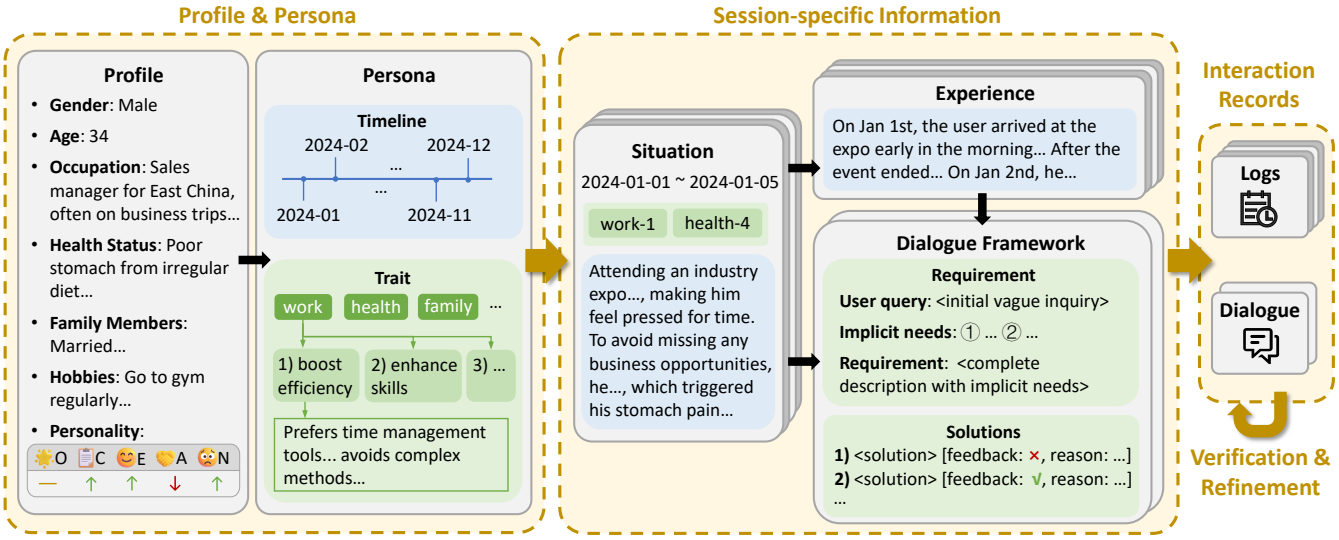

Figure 2: Overview of the generative pipeline for PAL-Set. We design a multi-stage LLM-based synthesis process to progressively specify the control information for interaction record generation. Additional verification and refinement steps are employed to ensure the final data quality.

variety of platforms on a user's personal device and generate a series of records that indirectly reflect the diverse events experienced by the user. For dialogue synthesis, we ensure that the generated dialogues adhere to the corresponding dialogue framework and reflect realistic flows commonly seen in service-oriented interactions. To achieve this, we define a set of dialogue actions for both user and assistant utterances. We also construct utterance-level dialogue templates that specify the action of each utterance and its linkage to specific components of the dialogue framework. These templates can ensure that the synthesized dialogues are consistent with our intended dialogue flows and content settings.

Data Verification and Refinement To ensure the quality of data synthesis, we employ a series of data verification and refinement methods. For each LLM-based generation step, we define a specific output format and set up validation rules. The output is automatically checked after each generation, and a regeneration process would be triggered if the output does not meet the rules. For the final logs and dialogues, we further perform validation and correction, targeting the following issues: (1) Log entries that do not match the specified type; (2) Dialogue utterances that are inconsistent with their assigned action; (3) Content that contradicts the user's profile or persona.

Dataset Analysis

Our PAL-Set contains 100 synthetic users, each associated with long-term, multi-session interaction records. For each user, sessions are divided into a history set and a query set. The query set includes all sessions from the last month of interaction and serves as the evaluation samples, while the history set comprises all earlier sessions and is used solely as contextual input. Note that a query set may contain multiple sessions, and earlier sessions within the query set can also serve as part of the history when evaluating later sessions. Table 1 summarizes the key statistics of PAL-Set. Our dataset features a long interaction span (9.4 months on average) and a relatively large number of sessions (29 sessions) for each user, offering a rich long-term context.

	History	Query
Avg. # sessions	25.7	3.3
Avg. # logs	888.7	107.5
Avg. # dialogue turns	361.7	39.3
Avg. # dialogue topics	62.5	8.3
Avg. # months	8.4	1.0

Table 1: Statistics of PAL-Set. Each value represents the average across 100 users.

In addition, we conduct human evaluations to verify whether the generated logs and dialogues align with the predefined user profiles and personas. Annotators rated each sample on a scale from 1 (non-matching) to 3 (completely matching). The average scores were 2.75 for logs and 2.67 for dialogues, indicating that the synthetic data is highly consistent with the intended user characteristics, demonstrating the high quality of our dataset.

Evaluation Tasks

Based on PAL-Set, we design three evaluation tasks as part of PAL-Bench: two single-turn question-answering tasks and one multi-turn dialogue interaction task. These tasks assess the assistant's ability to understand user requirements and generate personalized responses.

1) Requirement Restatement. This task evaluates the assistant's ability to accurately infer the user's complete requirement from user histories in a single-turn QA setting. The

input is the initial *user query* for each topic in the query set, while the expected output is the corresponding complete *requirement* description. We use BLEU score (Papineni et al. 2002) as the objective metric. Since the reference *requirement* descriptions are relatively abstract, we additionally introduce a GPT-4-based evaluation to specifically assess whether the generated content successfully captures the *implicit needs* that are not explicitly stated in the *initial query*, forming GPT-4 Scores for this task.

2) Solution Proposal. This task evaluates the assistant's ability to understand user-specific preferences based on user histories and provide responses that meet these preferences in a single-turn QA format. It consists of two subtasks: "solution generation" and "solution selection". In "solution generation", the assistant is required to generate a solution description based on the given complete *requirement*. We evaluate the output with the BLEU score as an objective metric, taking the 2 predefined positive solutions in the dialogue framework as references. In "solution selection", the assistant is additionally presented with the 8 candidate solutions and asked to identify the 2 positive ones. A Selection Score is then calculated based on the preference labels of the selected solutions.

3) Multi-turn Dialogue Interaction. Since our ultimate goal is to enhance the assistant's performance in dialogue, we design evaluation tasks specifically targeting multi-turn interactions. Considering the high cost of interacting with real users, we take advantage of the strong role-playing abilities of LLMs (Chen et al. 2024) to construct a User-LLM that simulates the predefined users in PAL-Set and interacts with the assistant. We also introduce an Evaluation-LLM responsible for automatically assessing the interaction quality of different memory modeling methods, which focuses on two key dimensions: requirement understanding (the assistant's ability to accurately clarify the user's actual needs) and preference understanding (the extent to which the assistant's proposed solutions align with user preferences). We conduct pairwise comparative evaluations on both these two dimensions, and report the Win–Tie–Lose counts for our method over all evaluation samples. To reduce the effects of LLM randomness and positional bias on the evaluation, we follow the FairEval framework (Wang et al. 2024a) to evaluate each pair 6 times with different input orders. In addition, we conduct human evaluations and analyze their correlation with LLM-based evaluations, as detailed in the Appendix.

H^{2}Memory Framework

Problem Formulation

In our user-agent interaction scenario, each session S consists of a series of logs $L = \{l_1, l_2, ..., l_m\}$ and a dialogue $D = \{u_1, a_1, ..., u_t, a_t\}$, where l denotes a log record, u and a denote user and assistant utterances, respectively. Given the current session S_c, the logs collected between the end of D_{c-1} and the beginning of the D_c are taken as L_c. We define the interaction history $\mathbb{H}$ as all previous sessions along with the latest logs L_c: $\mathbb{H} = \{\{L_1, D_1\}, ..., \{L_{c-1}, D_{c-1}\}, L_c\}$. Let Q denote the user's inquiry (along with preceding dialogue context, if any) in the current dialogue D_c, and R denote the assistant's generated response. The task is formalized as: $R = \mathrm{Assistant}(\mathbb{H}, Q)$.

However, due to the long-term nature of interactions, the original $\mathbb{H}$ can be prohibitively large and redundant, making it unsuitable as direct input. Therefore, we encode $\mathbb{H}$ into a concise and structured memory $\mathbb{M}$, from which relevant entries are retrieved to support personalized response generation. As shown in Figure 3, our method integrates a hierarchical heterogeneous memory structure with retrieval-augmented generation (RAG).

Memory Construction

In user-agent interaction scenarios, both logs and dialogues exhibit substantial heterogeneity—in their information formats and in the user persona perspectives they convey. A single memory structure cannot effectively capture both aspects. To address this, we adopt a differentiated memory organization tailored to long-term histories. For fragmented logs, we explicitly construct relational edges to form coherent and contextually complete situation descriptions. For service-oriented dialogues, which often follow dynamic topic flows, we define topic-level schemas and extract structured outlines to reflect the evolving conversation patterns. Building on these specific memory entries, we further summarize and abstract the user's overarching background and traits, resulting in a hierarchical two-tier memory structure.

Log Graph Logs are fragmented observations reflecting user behavior. To construct a coherent understanding of user experiences over time, we explicitly model the relationships between log entries. Inspired by the commonsense knowledge graphs that connect knowledge nodes via various relation edges (Hwang et al. 2021), we instruct the LLM to identify relational edges (including the *Caused_by* and *Follows* types) between logs in each session, forming relation set $\Upsilon = \mathrm{LLM}(L)$. These relations connect log entries chronologically, resulting in connected subgraphs $\{G_1, ..., G_g\}$.

We then instruct the LLM to generate a situation description s_i for each subgraph G_i, integrating the logs through their relations. The situation entries corresponding to each subgraph constitute the first part of the memory in the j-th session, denoted $M_G^j = \{s_1^j, ..., s_g^j\}$.

Background Since a single situation entry cannot comprehensively reflect all aspects of the user's overall background, we construct a long-term background memory M_B by summarizing across all situation entries in $\mathbb{H}$. We define several fixed aspects (e.g., work, family) and maintain a paragraph summary per aspect. Given that the user's experiences continuously evolves over time, to ensure the background remains up-to-date, we adopt a recursive memory updating mechanism (Wang et al. 2025) for M_B. Given the M_G^j of the j-th session, the updating process is: $M_B^{(j)} = \mathrm{LLM}(M_B^{(j-1)}, M_G^j)$.

Additionally, we use the LLM to associate each situation entry s_i in M_G with one or more aspects in M_B, forming a multi-valued mapping $\mathcal{F}_{GB} : M_G \rightarrow \mathcal{P}(M_B)$, where $\mathcal{P}(M_B)$ is the power set of M_B.

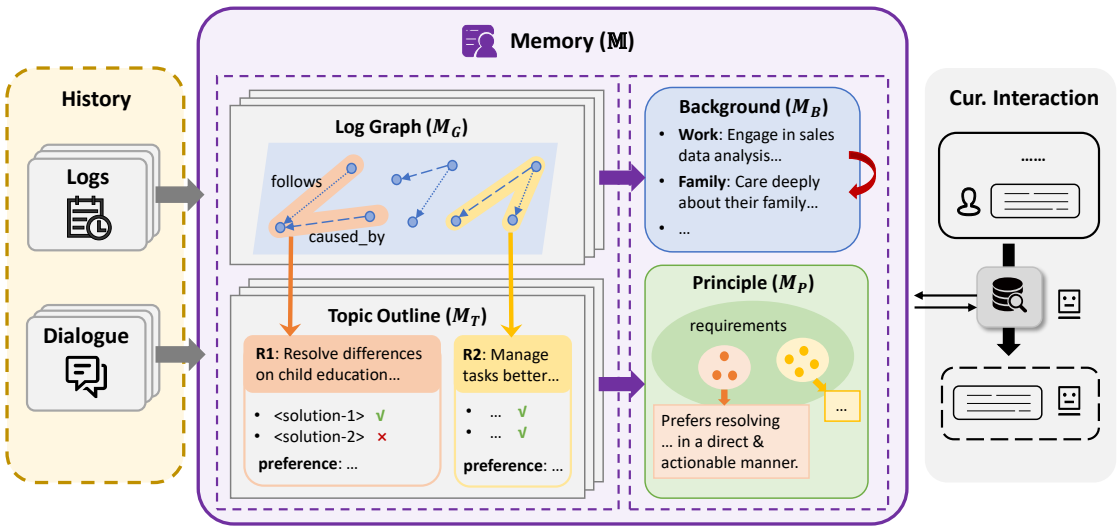

Figure 3: Overview of our method. We propose a hierarchical and heterogeneous memory mechanism (H^{2}Memory) to model user characteristics in user–agent interactions. Information from different sources is separately encoded into concrete- and abstract-level memory entries. The most relevant entries from each part are retrieved to enable personalized, retrieval-augmented response generation.

Topic Outline A service-oriented dialogue may involve discussions on multiple topics, with each topic containing both the clarification of user requirements and the discussion of solutions. For each topic t in the dialogue, we define an information schema $\{r, \mathbf{o}, p\}$. Here, r represents the user requirements reflected in the topic, $\mathbf{o}$ includes the multiple solutions provided by the assistant along with the user's feedback on each solution, and p summarizes the user's preference of solutions for the topic. The topic outline formed based on this schema can extract the key points of relevant topics from the original dialogue. Therefore, we guide the LLM to segment each dialogue D into τ topic-specific parts and generate a set of such outlines T, where $T = \{t_1, ..., t_\tau\}$. However, users often express their requirements in dialogues quite briefly, lacking the background or situation, which are often reflected in the logs from the corresponding period. To effectively integrate information from both sources, we retrieve k situation entries $\mathbf{s}$ from M_G^j of the corresponding j-th session that are most relevant to the requirement r. Based on this, we rewrite the r into a more detailed description $\hat{r}$, i.e., $\hat{r} = \text{LLM}(r, \mathcal{R}(\mathbf{s}))$, where $\mathcal{R}(\mathbf{s})$ denotes the retrieved situation entries related to r. We perform requirement rewriting for each r in T to get $\hat{T}$, and add these outlines to the memory bank, i.e., $M_T^{(j)} = M_T^{(j-1)} \oplus \hat{T}$.

Principle We further abstract several overall requirement types and corresponding preference principles from numerous specific topic outlines to form the memory M_P. To be specific, we first initialize M_P using all dialogues in the history set. All requirements (denoted as $\mathbf{r}$) extracted from these dialogues are encoded as features, and then we apply the KMeans clustering algorithm to obtain n clusters. For each cluster C_i, we instruct the LLM to extract the requirement type γ_i and preference principle ρ_i from bunch of specific descriptions ($\mathbf{r}_i$ & $\mathbf{p}_i$) belonging to the C_i. Thus, we obtain $M_P = \{\gamma, \rho\}$. Note that we retain the cluster assignments of all specific requirements, thereby forming a mapping $\mathcal{F}_{TP} : M_T \to M_P$.

Considering that the query set may contain multiple sessions, M_P needs to be continuously updated for new query sessions. For a requirement entry r, we first find the closest cluster center C_i, then update C_i with the features of r, and send both the previous γ_i and the current r to the LLM, prompting it to update the γ_i if necessary. We perform a similar update for the corresponding ρ_i as well.

Memory-based RAG

The final hierarchical and heterogeneous memory structure is $\mathbb{M} = \{M_G, M_B, M_T, M_P\}$, where M_G and M_T encode concrete-level information extracted from logs and dialogue context, respectively, while M_B and M_P capture abstract-level information. For an inquiry Q in the current j-th session, we retrieve relevant entries from each part of $\mathbb{M}$ as personalized context. Specifically, we first retrieve the k most relevant entries from the recent situation memory M_G^j, and then find the corresponding long-term background from M_B for the aspects associated with these situations. Next, we retrieve the k most relevant entries from the memory bank of the entire M_T part, since we consider that user traits are more stable factors in the long term compared to situations, and we further find the abstract principles in M_P corresponding to the specific topic entries. The whole retrieval process can be represented as:

$$\mathbf{m} = \{\mathcal{R}(M_G^j), \mathcal{F}_{GB}(\mathcal{R}(M_G^j)), \mathcal{R}(M_T^{(j)}), \mathcal{F}_{TP}(\mathcal{R}(M_T^{(j)}))\}$$

Methods	Requirement Restatement					Solution Proposal				
	B-1	B-2	B-3	B-4	G-Score	B-1	B-2	B-3	B-4	S-Score
Vanilla (w/o log)	13.59	5.76	2.58	1.41	17.50	18.85	6.96	3.52	2.06	18.95
Vanilla (with log)	19.71	8.85	4.10	2.29	23.00	19.76	7.68	4.00	2.32	22.88
Turn-level RAG	22.74	10.54	4.94	2.69	26.85	19.15	7.43	3.89	2.29	24.09
Session-level RAG	23.81	11.24	5.42	3.06	29.33	19.66	7.80	4.00	2.33	33.78
RecurSum (Wang et al. 2025)	23.29	10.64	4.95	2.75	28.36	19.89	7.59	3.96	2.26	25.61
ConditionMem (Yuan et al. 2025)	23.31	10.42	4.86	2.66	27.78	19.42	7.42	3.85	2.22	25.49
MemoryBank (Zhong et al. 2024)	23.89	11.11	5.23	2.91	28.57	20.49	8.12	4.07	2.34	29.85
H^{2}Memory (ours)	**26.67**	**12.18**	**5.68**	**3.09**	**32.54**	**22.24**	**8.38**	**4.39**	**2.65**	**38.32**
H^{2}Memory (w/o M_G)	26.32	11.93	5.42	2.93	30.90	22.19	8.24	4.35	2.62	37.71
H^{2}Memory (w/o M_B)	26.22	11.95	5.53	3.05	31.30	21.84	8.17	4.19	2.46	36.69
H^{2}Memory (w/o M_T)	24.04	11.01	5.20	2.82	28.00	18.23	6.49	3.21	1.86	28.09
H^{2}Memory (w/o M_P)	26.33	11.90	5.44	2.90	31.51	21.97	8.16	4.23	2.49	36.26

Table 2: Performance of the requirement restatement and solution proposal tasks. B-1 to B-4 represent BLEU scores. G-Score and S-Score denote the GPT-4 Score and the Selection Score described in "Evaluation Tasks", respectively.

where $\mathbf{m}$ denotes the retrieved personalized information, enabling us to realize a personalized augmented response generation process $R = \mathrm{LLM}(\mathbf{m}, Q)$.

Experiments

Experimental Setting

Implementation Details For retrieval, we use the "paraphrase-multilingual-mpnet-base-v2" (Reimers and Gurevych 2019; Song et al. 2020) as the encoder to extract text features from memory entries and queries, retrieving the top $k = 3$ most similar memory entries each time by cosine similarity. Qwen-Max-0428 is the base model (i.e., the Assistant-LLM) for memory construction and response generation in all experiments. In the multi-turn dialogue interaction task, User-LLM and Evaluation-LLM are Qwen2.5-Max and GPT-4-turbo, respectively.

Compared Baselines We compare our method with several baseline approaches using the same base model. *Vanilla (w/o log)* refers to using only the current query as input without any interaction history, while *Vanilla (with log)* includes the logs from the current session but still excludes previous sessions. *Turn-level RAG* and *Session-level RAG* retrieve historical dialogues at different granularities (utterance-level and session-level), respectively, while also incorporating logs from the current session. We also reimplement three prior memory-based methods: *RecurSum*(Wang et al. 2025), *ConditionMem*(Yuan et al. 2025), and *MemoryBank*(Zhong et al. 2024). As these methods did not originally account for logs in PAL-Bench, we add basic log processing for a fairer comparison. Furthermore, all retrieval-based baselines follow the same retrieval settings as our method.

Experimental Results on PAL-Bench

1) Requirement Restatement The left part of Table 2 presents the experimental results for the requirement restatement task. Incorporating logs under the vanilla setting improves performance, confirming that the logs in our PAL-Set are meaningfully related to user requirements. Further gains are observed when incorporating long-term historical information, suggesting that PAL-Set effectively captures the consistency of users' intrinsic characteristics over time. Our proposed H^{2}Memory framework achieves the best performance among all baselines, demonstrating its effectiveness in understanding requirements.

We also perform an ablation study on the four components of our memory structure. Among them, the relevant requirement descriptions in M_T extracted from historical dialogues provide the most significant contribution to understanding the current user needs. Nonetheless, the other memory components also yield measurable benefits, highlighting the complementary nature of the full memory design.

2) Solution Proposal The right part of Table 2 presents the experimental results for the solution proposal task. The overall trends mirror those observed in the requirement restatement task: the constructed long-term interaction records in our PAL-Set can effectively reflect specific user preferences, and our proposed method excels at modeling such preferences. Notably, our approach achieves a significant advantage over all baselines, particularly in the selection score. This indicates that user preferences tend to be abstract and nuanced, requiring more sophisticated modeling strategies beyond simple fact extraction from interaction history. Additionally, the ablation results reveal: removing M_B (background memory) leads to a larger performance drop than removing M_G (situation memory), which is the reverse of the trend observed in the requirement restatement task. This highlights that user preferences are more stable, high-level traits that accumulate over time, whereas requirements are more contextually grounded in recent events.

3) Multi-turn Dialogue Interaction Table 3 presents the "Win/Tie/Lose" statistics of our method in comparison with multiple baseline methods on the multi-turn dialogue interaction task. The results indicate that our method exhibits the most significant advantage over the vanilla baseline, which does not incorporate any interaction history. Moreover, our method consistently outperforms other memory-

Ours vs Baseline	Requirement	Preference
Vanilla (w/o log)	**478** / 29 / 319	**480** / 18 / 328
Vanilla (with log)	**447** / 33 / 346	**452** / 22 / 352
RecurSum	**421** / 29 / 376	**439** / 18 / 369
ConditionMem	**396** / 42 / 388	**413** / 19 / 394
MemoryBank	**449** / 33 / 344	**452** / 25 / 349

Table 3: Performance of the multi-turn dialogue interaction task. We report the "Win/Tie/Lose" numbers of our method compared to other baseline methods across all query topics.

Methods	Accuracy
Vanilla	10.00
RecurSum (Wang et al. 2025)	10.00
ConditionMem (Yuan et al. 2025)	40.00
MemoryBank (Zhong et al. 2024)	23.33
Ours ($M_T + M_P$)	**50.00**
Ours (M_T)	40.00
Ours (M_P)	46.67

Table 4: Experimental results on the "single-session-preference" subset in LongMemEval (Wu et al. 2025b).

based methods across both requirement and preference understanding dimensions. These findings suggest that our proposed dual-level heterogeneous memory modeling mechanism is effective in multi-turn dialogue settings and holds promise for improving user satisfaction in interactions. A qualitative case study is provided in the Appendix to further illustrate these outcomes.

While our method achieves strong overall performance, it still incurs a non-negligible number of "Lose" cases in the comparison with the vanilla baseline. We consider the primary reason is that, although the User-LLM follows predefined personas and dialogue actions, its utterances retain some randomness, which may also be a factor that guides Assistant-LLM responses toward varying content and affects their evaluation. Nonetheless, despite the existence of this factor, the overall trend clearly supports the effectiveness of our method in improving multi-turn dialogue interaction.

Validation on External Dataset

To verify the generalizability of our method, we also conduct experiments on the "single-session-preference" subset of LongMemEval (Wu et al. 2025b) since its scenario is most similar to our PAL-Bench, with a focus on understanding user preferences. Besides, this dataset is in English, which can evaluate the applicability of our method to different languages. However, their data only contains dialogues without logs. Therefore, we only employ the dialogue modeling component (M_T + M_P) of our proposed H^{2}Memory framework in this experiment.

As shown in Table 4, combining both specific-level memory (M_T) and abstract-level memory (M_P) achieves the best performance and outperforms other approaches, demonstrating the generalizability of our method on external data beyond our PAL-Bench.

Related Work

Long-term Dialogue Benchmarks. Recently, several long-term dialogue benchmarks have been proposed to facilitate research on personalized dialogue systems. Xu, Szlam, and Weston (2022) construct a multi-session dialogue dataset MSC by extending the Persona-Chat (Zhang et al. 2018). Jang, Boo, and Kim (2023) build a multi-session dialogue dataset CC, which includes relationships between dialogue roles. In addition, SHARE (Kim, Park, and Chang 2024) and LoCoMo (Maharana et al. 2024) construct long-term dialogue datasets featuring shared memory and multi-modal histories, respectively. While the above works mainly focus on human-human interactions, some benchmarks also address user-assistant interactions and evaluate different aspects of assistant capabilities. LongMemEval (Wu et al. 2025b) mainly focuses on the extraction and recall of user facts from dialogue history, with only a small subset addressing user preferences. ImplexConv (Li et al. 2025) centers on implicit reasoning over subtle information, and MapDia (Wu et al. 2025a) aims to enable proactive topic-shifting by assistants. However, these works overlook the subjective characteristics of users, and they do not consider user behavior histories in user-agent interaction settings. These are the main differences from our proposed PAL-Bench.

Personalized Response Generation Methods. Personalized response generation methods fall into three categories. The first directly includes all user history in the prompt, but this only works for models with long-context support and often misses key personalized details. The second type is memory parameterization (Ma et al. 2021; Zhang, Kim, and Liu 2024; Zhang et al. 2024b; Liu et al. 2024), but such methods cannot explicitly organize memory entries to handle complex personalized scenarios, and they are unsuitable for API-based LLMs since parameter fine-tuning is needed. The third type constructs external memories and employs retrieval-augmented generation (RAG) methods to enhance generation (Lu et al. 2023; Zhong et al. 2024; Wang et al. 2024b; Yuan et al. 2025). Our work also falls into this type, but unlike previous studies, we design a hierarchical and heterogeneous memory structure for service-oriented user-agent interaction scenarios.

Conclusion

In this work, we focus on personalized, long-term, service-oriented interactions. To support research in this area, we design a multi-stage data synthesis pipeline and construct the first Chinese user-agent long-term interaction dataset, PAL-Set. Building on this, we introduce a new evaluation benchmark, PAL-Bench, aimed at assessing assistants' abilities to understand user requirements and preferences based on long-term interaction histories. Additionally, we propose a hierarchical and heterogeneous memory modeling framework, H^{2}Memory, to improve response generation in personalized dialogue settings. Experimental results demonstrate the effectiveness of our proposed memory framework.

Acknowledgments

This work was sponsored by CCF-ALIMAMA TECH Kangaroo Fund (NO. CCF-ALIMAMA OF 2024007).

References

Chen, J.; Wang, X.; Xu, R.; Yuan, S.; Zhang, Y.; Shi, W.; Xie, J.; Li, S.; Yang, R.; Zhu, T.; Chen, A.; Li, N.; Chen, L.; Hu, C.; Wu, S.; Ren, S.; Fu, Z.; and Xiao, Y. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. *Transactions on Machine Learning Research*. Survey Certification.

Guan, Y.; Wang, D.; Chu, Z.; Wang, S.; Ni, F.; Song, R.; and Zhuang, C. 2024. Intelligent agents with llm-based process automation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5018–5027.

Ha, J.; Jeon, H.; Han, D.; Seo, J.; and Oh, C. 2024. CloChat: Understanding how people customize, interact, and experience personas in large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–24.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6384–6392.

Jang, J.; Boo, M.; and Kim, H. 2023. Conversation Chronicles: Towards Diverse Temporal and Relational Dynamics in Multi-Session Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13584–13606.

Kim, E.; Park, C.; and Chang, B. 2024. SHARE: Shared Memory-Aware Open-Domain Long-Term Dialogue Dataset Constructed from Movie Script. *arXiv preprint arXiv:2410.20682*.

Li, X.; Bantupalli, J.; Dharmani, R.; Zhang, Y.; and Shang, J. 2025. Toward Multi-Session Personalized Conversation: A Large-Scale Dataset and Hierarchical Tree Framework for Implicit Reasoning. *arXiv preprint arXiv:2503.07018*.

Liu, J.; Zhu, Y.; Wang, S.; Wei, X.; Min, E.; Lu, Y.; Wang, S.; Yin, D.; and Dou, Z. 2024. LLMs + Persona-Plug = Personalized LLMs. *CoRR*, abs/2409.11901.

Lu, J.; An, S.; Lin, M.; Pergola, G.; He, Y.; Yin, D.; Sun, X.; and Wu, Y. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Ma, Z.; Dou, Z.; Zhu, Y.; Zhong, H.; and Wen, J.-R. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 555–564.

Maharana, A.; Lee, D.-H.; Tulyakov, S.; Bansal, M.; Barbieri, F.; and Fang, Y. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13851–13870.

Ong, K. T.-i.; Kim, N.; Gwak, M.; Chae, H.; Kwon, T.; Jo, Y.; Hwang, S.-w.; Lee, D.; and Yeo, J. 2024. Towards Lifelong Dialogue Agents via Relation-aware Memory Construction and Timeline-augmented Response Generation. *arXiv preprint arXiv:2406.10996*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867.

Team, Q. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; et al. 2024a. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450.

Wang, Q.; Fu, Y.; Cao, Y.; Wang, S.; Tian, Z.; and Ding, L. 2025. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 130193.

Wang, Z.; Li, Z.; Jiang, Z.; Tu, D.; and Shi, W. 2024b. Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4891–4906.

Wu, B.; Wang, W.; Li, H.; Li, Y.; Yu, J.; and Wang, B. 2025a. Interpersonal Memory Matters: A New Task for Proactive Dialogue Utilizing Conversational History. *arXiv preprint arXiv:2503.05150*.

Wu, D.; Wang, H.; Yu, W.; Zhang, Y.; Chang, K.-W.; and Yu, D. 2025b. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The Thirteenth International Conference on Learning Representations*.

Xu, J.; Szlam, A.; and Weston, J. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5180–5197.

Yuan, R.; Sun, S.; Li, Y.; Wang, Z.; Cao, Z.; and Li, W. 2025. Personalized Large Language Model Assistant with Evolving Conditional Memory. In *Proceedings of the 31st International Conference on Computational Linguistics*, 3764–3777.

Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024a. CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. In

Findings of the Association for Computational Linguistics ACL 2024, 13947–13966.

Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Chen, G.; Li, J.; Wu, X.; Zhiyi, Z.; Xiao, Q.; et al. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10859–10885.

Zhang, K.; Kang, Y.; Zhao, F.; and Liu, X. 2024b. LLM-based Medical Assistant Personalization with Short-and Long-Term Memory Coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2386–2398.

Zhang, K.; Kim, Y.; and Liu, X. 2024. Personalized llm response generation with parameterized memory injection. *arXiv preprint arXiv:2404.03565*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213.

Zhang, T.; Zhang, X.; Zhao, J.; Zhou, L.; and Jin, Q. 2024c. ESCoT: Towards Interpretable Emotional Support Dialogue Systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13395–13412.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19724–19731.