AlpsBench: An LLM Personalization Benchmark for Real-Dialogue Memorization and Preference Alignment

Jianfei Xiao
jianfeixiao@mail.ustc.edu.cn
University of Science and Technology of China
Hefei, China

Xiang Yu
yux661988@gmail.com
University of Science and Technology of China
Hefei, China

Chengbing Wang
wwq197297@mail.ustc.edu.cn
University of Science and Technology of China
Hefei, China

Wuqiang Zheng
qqqqqzheng@gmail.com
University of Science and Technology of China
Hefei, China

Xinyu Lin
xylin1028@gmail.com
National University of Singapore
Kent Ridge, Singapore

Kaining Liu
unraveller@mail.ustc.edu.cn
University of Science and Technology of China
Hefei, China

Hongxun Ding
hongxunding02@gmail.com
University of Science and Technology of China
Hefei, China

Yang Zhang
zyang1580@gmail.com
National University of Singapore
Kent Ridge, Singapore

Wenjie Wang
wenjiewang96@gmail.com
University of Science and Technology of China
Hefei, China

Fuli Feng
fulifeng93@gmail.com
University of Science and Technology of China
Hefei, China

Xiangnan He*
hexn@ustc.edu.cn
University of Science and Technology of China
Hefei, China

Abstract

As Large Language Models (LLMs) evolve into lifelong AI assistants, LLM personalization has become a critical frontier. However, progress is currently bottlenecked by the absence of a gold-standard evaluation benchmark. Existing benchmarks either overlook personalized information management that is critical for personalization or rely heavily on synthetic dialogues, which exhibit an inherent distribution gap from real-world dialogue. To bridge this gap, we introduce **AlpsBench**, **A**n **L**LM **P**er**S**onalization benchmark derived from real-world human–LLM dialogues. AlpsBench comprises 2,500 long-term interaction sequences curated from WildChat, paired with human-verified structured memories that encapsulate both explicit and implicit personalization signals. We define four pivotal tasks—personalized information *extraction*, *updating*, *retrieval*, and *utilization*—and establish protocols to evaluate the entire lifecycle of memory management. Our benchmarking of frontier LLMs and memory-centric systems reveals that: (i) models struggle to reliably extract latent user traits; (ii) memory updating faces a performance ceiling even in the strongest models; (iii) retrieval accuracy declines sharply in the presence of large distractor pools; and (iv) while explicit memory mechanisms improve recall, they do not inherently guarantee more preference-aligned or emotionally resonant responses. AlpsBench aims to provide a comprehensive framework to accelerate research toward truly personalized AI assistants.

CCS Concepts

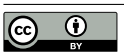
• **Information systems** → **Personalization**.

Keywords

LLM Personalization Benchmark; Structured Personalized Memory; Implicit User Preferences

ACM Reference Format:
Jianfei Xiao, Xiang Yu, Chengbing Wang, Wuqiang Zheng, Xinyu Lin, Kaining Liu, Hongxun Ding, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2026. AlpsBench: An LLM Personalization Benchmark for Real-Dialogue Memorization and Preference Alignment. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3805712.3808634

*Corresponding author.



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808634

1 Introduction

Recent advancements in LLMs (*e.g.,* long-context understanding and self-evolution) have demonstrated promising results across different domains [45, 46, 52], opening up significant potential for LLMs to shift from instant general-purpose tools to lifelong evolving AI assistants. Despite that mainstream LLMs are proficient in

Table 1: Comparison between AlpsBench and existing LLM personalization benchmarks. "Real" denotes whether the benchmark is built upon real-world data. Columns "T1" to "T4" represent whether the benchmarks contain corresponding testing dimensions.

Benchmark	Real	T1	T2	T3	T4: Utilization Dimensions				
		Ext.	Upd.	Retr.	PA	PF	VRA	CF	EI
LaMP	●	○	○	○	◐	◐	○	○	○
PersonalLLM	○	○	○	○	◐	●	○	○	○
EQ-Bench	○	○	○	○	○	○	○	○	●
PersoBench	○	○	○	○	●	●	○	●	○
PersonaFeedback	○	○	○	○	●	●	○	●	○
LoCoMo	○	◐	○	●	◐	○	○	○	○
LongMemEval	○	●	●	●	◐	○	○	○	○
PersonaLens	○	○	○	◐	●	●	○	●	○
HaluMem	○	●	●	◐	◐	○	○	○	○
PersonaMem v2	○	●	●	◐	●	●	○	○	○
AlpsBench	●	●	●	●	●	●	●	●	●

Category: Memory-free Preference Alignment, Memory-aware Preference Alignment, Ours.
●: Fully Supported / Real Data, ◐: Partially Supported, ○: Not Supported / Synthetic Data.
Ext. = Memory Extraction, **Upd.** = Memory Updating, **Retr.** = Memory Retrieval, **PA** = Persona Awareness, **PF** = Preference Following, **VRA** = Virtual-Reality Awareness. **CF** = Constraint Following. **EI** = Emotional Intelligence.

addressing generic tasks, they still struggle to accommodate heterogeneous users' needs, potentially hurting the user experience in daily use of AI [22, 35]. In the light of this, enabling LLM personalization becomes crucial to achieve lifelong personal intelligence, strongly attracting and motivating both academia [31, 42, 54, 60] and industry (*e.g.,* OpenAI [28], Google [18], and Anthropic [3]) to explore personalizing LLM responses for different individuals.

Facilitating LLM personalization techniques crucially relies on high-quality evaluation benchmarks. Existing benchmarks can be broadly categorized into two groups as shown in Table 1:

- ***Memory-free*** benchmarks (*e.g.,* LaMP [35], PersonalLLM [65]) focus on the alignment between LLMs' final output and user preference in downstream tasks, such as personalized email generation. Despite the effectiveness in assessing personalization in response, they overlook the process of personalized information governance (*e.g.,* memory extraction, update, retrieval), which is a key component in understanding and leveraging user preference in personalized responses.
- ***Memory-aware*** benchmarks explicitly incorporate memory evaluation based on synthetic human-LLM dialogues along multiple dimensions (*e.g.,* long-term history memorization [23, 47] and emotional preference alignment [29, 33]). Nonetheless, they suffer from a critical distributional gap between simulated and real-world data, causing two limitations. 1) Lack of diverse conversations, where LLM-synthesized dialogues tend to be homogeneous [14], failing to capture the natural conversational diversity (see Figure 1). 2) Lack of implicit expressions. In real-world interactions, users often convey personalized information implicitly. However, synthetic dialogues can be overly explicit and simplistic (see Figure 2), failing to generalize to real-world scenarios.

In this light, it is imperative to build benchmarks based on real-world user conversational data that capture both conversational diversity and implicit personalization signals. Moreover, instead of simply verifying LLMs' final output, it is essential to systematically

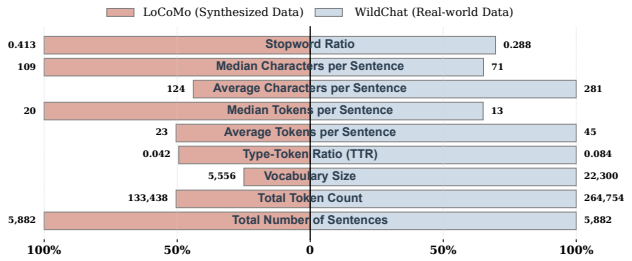


Figure 1: Comparison between synthesized and real data.

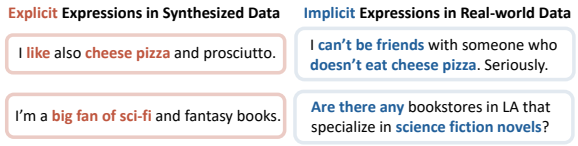


Figure 2: Examples of the synthesized and the real-world human expressions.

evaluate the process of understanding, storing, and utilizing personalized information. Hence, we propose a comprehensive evaluation framework with four core tasks.

- *Task 1: Personalized information* ***extraction***, which evaluates the ability to extract personalized information from dialogues. Given the raw conversational data, the evaluated LLMs are asked to distill them into structured memory containing personalized information (refer to Section 2.1 for details). In the paper, we report semantic-matching F1 with LLM-as-a-Judge against ground-truth memory, while additional diagnostic metrics are available on the leaderboard.
- *Task 2: Personalized information* ***update***, which assesses the capacity to track the dynamics of user information and update structured memory accurately. Given both a historical and a new dialogue, the evaluated LLMs are asked to output updated memory and the memory manipulation (*i.e.,* retention, addition, and modification). In the paper, we report semantic-matching F1 for updated memories, while action-level accuracy and other diagnostic metrics are available on the leaderboard.
- *Task 3: Personalized information* ***retrieval***, which measures the ability to retrieve relevant personalized information. Given a user query and a candidate set of memories, the evaluated LLMs are asked to select the relevant memory for generating a personalized response. Recall is adopted as the evaluation metric.
- *Task 4: Personalized information* ***utilization***, which examines the alignment of user preference in LLMs' response. We break down the task into measurement of five LLM capabilities, including persona awareness, preference following, virtual-reality awareness, constraint following, and emotional intelligence for a comprehensive alignment of dimensions (refer to Section 2.4 for detailed task designs and evaluation metrics).

To construct a real-world comprehensive benchmark, we develop a four-step pipeline. *Step 1*: Data collection. We leverage the real-world WildChat dataset [58] and collect long-term dialogues[1]. *Step 2*: Memory extraction and filtering. We design a structured memory

[1]WildChat is distributed under the Open Data Commons Attribution License (ODC-BY). We acknowledge WildChat as the source dataset and make our reprocessed derivative data available on Hugging Face under ODC-BY.

taxonomy to extract personalized information from dialogues and filter redundant or low-signal samples. *Step 3*: Task construction. Based on the collected dialogues and extracted memories, we construct evaluation data for four tasks, including evaluation queries and corresponding ground truth. *Step 4*: Human verification and quality control. Human annotators then verify the constructed memories, task instances, and LLM-as-a-Judge alignment to ensure benchmark reliability. This pipeline can be continuously updated with evolving real-world human-LLM conversations, enabling a dynamic benchmark.

Building upon the above pipeline, we introduce **AlpsBench**, a comprehensive benchmark for evaluating LLM personalization, with the following key features: 1) AlpsBench consists of 2,500 high-quality real-world human–LLM dialogues with 6 to 249 turns. 2) It covers four tasks for personalized information, *i.e.,* extraction, update, retrieval, and utilization. 3) It assesses multidimensional capabilities of memory utilization, *i.e.,* persona awareness, preference following, virtual-reality awareness, constraint following, and emotional intelligence. We also release a public leaderboard[2] to facilitate continuous benchmarking from the community.

We benchmark frontier LLMs, including general-purpose and memory-oriented LLMs, and derive several key findings: (i) models exhibit limited reliability in extracting latent user traits; (ii) memory updating reaches a performance plateau even for the strongest models; (iii) retrieval accuracy degrades substantially in the presence of large distractor sets; and (iv) explicit memory mechanisms do not inherently ensure stronger preference alignment or emotional resonance in responses. We release our code at https://github.com/ThisIsCosine/AlpsBench.

The key contributions of this work are summarized as follows:

- We introduce AlpsBench, a dynamic and extensible LLM personalization benchmark built on real-world human–LLM conversations with human-verified structured memory.
- We propose a systematic evaluation framework comprising four core tasks (*i.e.,* personalized information extraction, update, retrieval, and utilization) to benchmark personalization ability.
- Extensive evaluations on AlpsBench reveal that existing models remain limited in reliable long-term personalization and context-aware utilization, highlighting substantial opportunities for future research in personalized LLMs.

2 AlpsBench Evaluation Framework

To rigorously evaluate the personalization capabilities of AI assistants, we introduce AlpsBench, a comprehensive benchmark designed to simulate real-world personalization scenarios. In this section, we provide a detailed description of the task design for AlpsBench (*cf.* Figure 3), which comprises four core tasks specifically designed to assess different facets of personalized AI assistants.

2.1 Task 1: Personalized Information Extraction

A core capability of personalized AI assistants lies in their proficiency in transforming raw conversational data into structured, high-quality memories. This capability reflects the assistants' depth of understanding regarding historical user preferences and key information. Motivated by the need to assess this capability, we propose a *Personalized Information Extraction* task:

$$\mathcal{H} \xrightarrow{\text{AI assistant}} [\hat{\mathcal{M}}_1, \hat{\mathcal{M}}_2, \ldots], \tag{1}$$

where the AI assistant is tasked with extracting structured user information from the original user dialogue history ($\mathcal{H}$). The resulting output consists of a set of memories ($\hat{\mathcal{M}}_1, \hat{\mathcal{M}}_2, \ldots$), each containing the following attributes:

- **Memory ID:** A unique identifier for the memory entry.
- **Memory Type:** Explicit or implicit preference.
- **Label:** The taxonomic category of the memory.
- **Value:** The exact statement extracted from the original dialogue.
- **Confidence:** The confidence level of the extraction.

The quality of the extractions (*i.e.,* memories) is evaluated by comparing them against human-annotated ground truth ($\mathcal{M}$). Specifically, we use an LLM-as-a-Judge protocol to semantically match predicted and human-annotated memories one-to-one, computing F1 to capture nuances beyond lexical similarity.

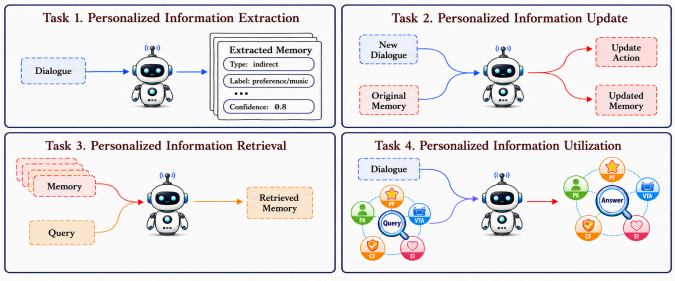


Figure 3: The evaluation tasks of AlpsBench.

2.2 Task 2: Personalized Information Update

The dynamic evolution of user preferences requires personalized AI assistants to possess a sophisticated memory-updating mechanism. This process hinges on three fundamental capabilities: (i) the robustness in filtering dialogue noise, (ii) the ability to add new user preferences, and (iii) the logic for resolving preference conflicts by reasonably updating outdated memories with new information. Thus, we introduce the *Personalized Information Update* task:

$$(\mathcal{M}, \mathcal{H}_{new}) \xrightarrow{\text{AI assistant}} (\hat{\mathcal{M}}_{new}, \hat{\text{Act}}), \tag{2}$$

where the task requires the assistant to output updated memories $\hat{\mathcal{M}}_{new}$ and categorize each update according to an action type $\hat{\text{Act}} \in \{\text{Retention, Addition, Modification}\}$, corresponding to the aforementioned capabilities. In this paper, we apply the same semantic matching protocol as in Task 1 and report F1 between updated memories and human-annotated ground truth as the primary metric. Action-level accuracy and other diagnostic metrics are available on the leaderboard. The updated memories can also be leveraged for further analysis of the AI assistants (*cf.* Subsection 4.1.2).

2.3 Task 3: Personalized Information Retrieval

The ability to retrieve relevant memories is a cornerstone of an AI assistant's personalization. Accurately identifying useful insights from a user's vast memory pool significantly enhances the AI assistant's capacity to provide tailored responses. In this light, we

[2]https://misshsiaoo.github.io/Alps_Bench/.

introduce the *Personalized Information Retrieval* task:

$$(\mathcal{M}_{\text{pos}}, \{\mathcal{M}_{\text{neg},i}\}_{i=1}^{K}, Q) \xrightarrow{\text{AI assistant}} \hat{\mathcal{M}}_{\text{pos}}, \tag{3}$$

where the AI assistant is expected to retrieve the most relevant memory ($\hat{\mathcal{M}}_{\text{pos}}$) based on the user query (Q) from a set of candidate memories, which consists of one positive sample ($\mathcal{M}_{\text{pos}}$) labeled by humans and K randomly sampled negative samples $\{\mathcal{M}_{\text{neg},i}\}_{i=1}^{K}$. We use the widely recognized metric of recall to report the evaluation results of this task.

2.4 Task 4: Personalized Information Utilization

The ability to utilize users' personal information to address real-world needs is fundamental to personalized AI assistants. To systematically evaluate this capability, we introduce the *Personalized Information Utilization* task, formulated as follows:

$$(\mathcal{H}, Q) \xrightarrow{\text{AI assistant}} \hat{\mathcal{R}}, \tag{4}$$

where the assistant is tasked to produce a response $\hat{\mathcal{R}}$ to a user query Q based on dialogue history $\mathcal{H}$.

Based on the characteristics of users' real-world needs, we design five core evaluation dimensions for this task:

Persona Awareness (PA). This dimension assesses whether the assistant correctly recalls and applies explicit user attributes, such as educational background. Users often provide explicit persona information across multiple sessions, requiring the AI assistant to seamlessly integrate this data into its responses.

Preference Following (PF). This dimension measures whether the assistant can capture implicit user preferences within dialogue history and align its responses accordingly. Beyond explicit persona information recall, it tests inductive reasoning to align responses with potential tastes and habits—a critical capability for personalized recommendation and planning.

Virtual-Reality Awareness (VRA). This dimension evaluates whether the assistant can distinguish real user information from role-play or fictional content, ensuring that in-character data does not contaminate real-world assistance. For example, if a middle-aged professional previously role-played as a student, the assistant should disregard student's attributes when drafting his/her resume.

Constraint Following (CF). This dimension examines whether the assistant respects constraints expressed by the user in prior interactions. There may be instances where the user explicitly requests that certain information be excluded or specific conditions be met. The assistant should consistently follow these guidelines to ensure alignment with the user's expectations.

Emotional Intelligence (EI). This dimension assesses whether the assistant uses user dialogue history to provide emotionally appropriate responses. An effective AI assistant enables differentiated emotional responses, such as offering encouragement to resilient users and reassurance to more sensitive users.

In this task, we adopt LLM-as-a-Judge to evaluate AI assistants. For each dimension (except for Emotional Intelligence), the judge determines whether the AI assistant's response meets the specified criteria and assigns a binary score (0 or 1) accordingly. Since Emotional Intelligence requires a more nuanced evaluation, we follow prior evaluation practice [44], which uses the LLM as a reward model to assign a score on a 1–5 scale.

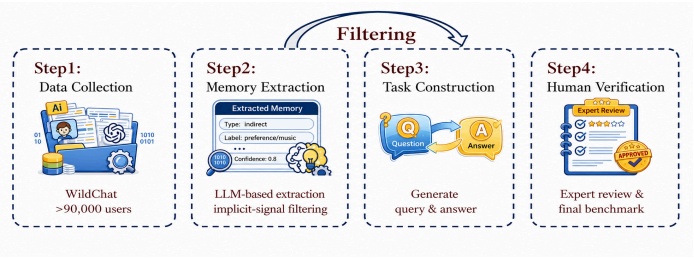


Figure 4: The four-step benchmark construction pipeline based on real human–LLM dialogues.

3 Curation Pipeline for AlpsBench

In this section, we describe the four-step curation pipeline of AlpsBench, as depicted in Figure 4.

- **Step 1: Data Collection**. Personalized AI assistants are intended to serve actual human needs. Therefore, to effectively test their capabilities, the design of the personalization benchmark should be grounded in the real-world human-AI interactions. To achieve this, we use the WildChat [58] dataset, which contains approximately 90,000 users' interaction records covering a broad range of authentic conversational scenarios. Considering longer dialogues inherently carry richer contextual dependencies and cross-temporal information, we further refined the data by selecting dialogues with more interaction turns. Each dialogue in this subset contains between 6 and 249 turns of interaction.
- **Step 2: Memory Extraction and Filtering**. To enable a systematic evaluation of an AI assistant's ability to process and utilize personalized information, we first employ the *DeepSeek-v3.2 reasoning model* [6] to automatically extract structured information from raw user dialogue histories and format it into memory representations, as illustrated in Section 2.1.

  Considering the substantial computational cost of running AI assistants and the expense of high-quality human verification (*cf.* Step 4), we further perform data filtering based on high-level semantic categories in the structured memories (*i.e.,* the *label* attribute). Specifically, we subsample instances within each category, ensuring that the number of samples per category does not exceed a predefined upper bound. This strategy effectively removes large amounts of semantically redundant content.

  Moreover, to enhance the challenge and discriminative power of the benchmark, we prioritize retaining users whose memories contain implicit information while performing category-based filtering. Specifically, we retain a user u if there exists at least one memory $\mathcal{M}$ with the type attribute labeled as *implicit*. Formally:

$$U_{\text{retained}} = \{u \in U \mid \exists \mathcal{M} \in \mathcal{M}_u : \text{type}(\mathcal{M}) = \textit{implicit}\}, \tag{5}$$

  where U is the set of all users. $\mathcal{M}_u$ is the set of memories associated with user u. Ultimately, this process yields 2,500 user samples that constitute the final benchmark.
- **Step 3: Task Construction**. Based on the collected dialogue histories, we construct evaluation queries and corresponding ground truth for the four tasks described in Section 2. Selecting a memory as a target memory ($\mathcal{M}_{tgt}$) and using the original user dialogue ($\mathcal{H}$) as context, we employ *GPT-5.2* [27] to synthesize task-specific queries and ground-truth answers. This process can be formulated as:

$$(\mathcal{H}, \mathcal{M}_{tgt}, [*]) \xrightarrow{\text{GPT-5.2}} (\hat{Q}, \hat{\mathcal{A}}) \xrightarrow{\text{Human}} (Q, \mathcal{A}) \tag{6}$$

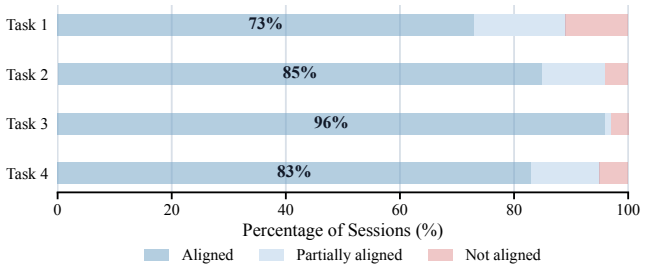
where [∗] represents task-specific auxiliary inputs. In the case of ***Task 2***, its [∗] represents target update strategies—specifically Retention, Addition, and Modification—to guide the generation of update-oriented queries. For ***Task 3***, its [∗] represents randomly selected memories as distractors (negative samples) to enhance the realism and complexity of the retrieval tasks.

The output of this process consists of two main elements: Q and $\mathcal{A}$. Specifically, for ***Task 1***, $\hat{Q}$ consists of system instructions for memory extraction, while $\hat{\mathcal{A}}$ represents the $\mathcal{M}_{tgt}$ itself. For ***Task 2***, $\hat{Q}$ corresponds to the user's new dialogue history along with the system's instruction (*e.g.*, "Please update memory") and $\hat{\mathcal{A}}$ reflects the applied update strategy. For ***Task 3***, $\hat{Q}$ is a practical user inquiry related to the $\mathcal{M}_{tgt}$, with $\hat{\mathcal{A}}$ as the $\mathcal{M}_{tgt}$. Finally, for ***Task 4***, $\hat{Q}$ is a user query related to the $\mathcal{M}_{tgt}$ topic, and $\hat{\mathcal{A}}$ is the generated response addressing that query.

• **Step 4: Human Verification and Quality Control**. We verify AlpsBench from two aspects: whether the constructed benchmark data provide reliable ground truth, and whether the LLM-as-a-Judge evaluator is aligned with human judgments.

Task Construction Quality. We first extracted candidate memories from user dialogues and asked annotators to verify the gold memory entries for Task 1. These verified memories then served as the basis for constructing the remaining benchmark instances, including update-oriented dialogues and actions for Task 2 and query-answer pairs for Tasks 3 and 4, yielding 2,500 verified instances for each of the four tasks. All benchmark instances were independently examined by two experienced annotators under a double-blind setting. The two annotators showed strong inter-annotator agreement: for Tasks 1 and 2, inter-annotator memory-set F1 scores reached 0.763 and 0.844, confidence Spearman correlations were 0.868 and 0.735, and Cohen's κ values for memory type, label, and time scope ranged from 0.734 to 0.927. For Tasks 3 and 4, approximately 92% of instances were deemed high-quality. For instances where annotators disagreed, an additional round of manual adjudication and correction was performed. These adjudicated instances were then used to form the final benchmark dataset.

LLM-as-a-Judge Alignment. To support scalable evaluation, we further compare the LLM-based judge with human expert judgments. As shown in Figure 5, the judge is highly consistent with human preferences across all four tasks, with alignment rates of 96% for Task 3, 85% for Task 2, 83% for Task 4, and 73% for Task 1. These results indicate that the automated evaluation reflects human-centric assessment while remaining scalable to the full benchmark.



Fully aligned: LLM decision consistent with both annotators. Partially aligned: consistent with one annotator. Not aligned: inconsistent with both.

Figure 5: Alignment between the LLM-based judge and human experts across four tasks.

4 Experiments

• **Evaluated Models**. For **general-purpose LLMs**, we include **open-source models** (DeepSeek-v3.2 in thinking mode [6], Qwen3-max [51], Llama-4 Maverick [24]) and **closed-source models** (GPT-5.2 [27], GPT-4.1-mini [26], Gemini-3-Flash-preview [8], Claude-Sonnet-4.5 [2]). For **memory-oriented systems**, we implemented MemoryOS [15], EverMemOS [10], A-Mem [50], LightMem [7], MemOS [19], Mem0 [5], and Mem0g [5]. All memory systems use **GPT-4.1-mini** as the backbone LLM. For Task 3 in the general-model setting, we additionally evaluate two standard retrieval baselines for memory models, **nltk+BM25** [32, 41] and **all-MiniLM-L6-v2** [37]. These classic retrieval methods are widely used as external retrievers in memory systems, making them suitable baselines for Task 3. • **Implementation Details**. We use GPT-5.2 [27] as the unified generator for query generation, and DeepSeek-v3.2 [6] as the judge model for automatic evaluation.

4.1 Performance on AlpsBench

4.1.1 Task 1: Extraction Result Analysis. **General-purpose LLMs remain limited in personalized information extraction.** As shown in Table 2, the best Task 1 F1 score is achieved by DeepSeek-v3.2 (0.6742), closely followed by GPT-5.2 (0.6720), while Llama-4 Maverick remains substantially lower (0.3772). As shown in Figure 6, a precision–recall decomposition further indicates that extraction performance is not governed by a single capability. Qwen3-max obtains the highest precision (0.8104) but a relatively lower recall (0.6190), suggesting a conservative extraction behavior; in contrast, Claude-Sonnet-4.5 and Gemini-3 Flash achieve much higher recall (0.8373 and 0.8175) with lower precision. These results show that current LLMs face a persistent trade-off between extracting broader personalized memory sets and maintaining extraction precision.

The direct–indirect split provides a more diagnostic view of this limitation. As shown in Figure 6, direct memories consistently achieve higher recall than indirect memories across all evaluated models. The gap is relatively small for stronger reasoning models such as DeepSeek-v3.2 (0.5386 vs. 0.4716) and Qwen3-max (0.4399 vs. 0.3586), but becomes severe for comparatively weaker reasoning models: GPT-4.1-mini drops from 0.5526 direct recall to only 0.0032 indirect recall, and Llama-4 Maverick drops from 0.1862 to 0.0337. This pattern suggests that explicit user information can often be captured through surface-level matching, whereas implicit personalized information requires stronger inference over conversational context.

Memory-oriented systems improve coverage but still face precision–recall trade-offs. As shown in Table 3, most memory-oriented systems substantially improve recall over the GPT-4.1-mini backbone, suggesting that persistent memory pipelines are effective at preserving a broader set of user-related information. However, this coverage often comes at the cost of precision. EverMemOS achieves the highest recall (0.9340), but its precision remains only 0.0500, leading to a low F1 score of 0.0900. A-Mem and LightMem exhibit the same trade-off: they achieve recall above 0.80, but their low precision leads to limited F1, indicating that higher coverage

Table 2: Tasks 1–4. Experimental evaluation results of general-purpose LLMs.

Model	Task 1 Extraction	Task 2 Update	Task 3 Retrieval					Task 4 Utilization						
			Number of Distractor Memories					PA	PF		VRA	CF	EI	
			100	300	500	700	1000		Gen.	Int.			EN	CN
GPT-5.2	0.6720	**0.7793**	0.9519	0.9400	0.9249	0.9138	0.8945	0.5684	0.7043	**0.7680**	0.6640	0.9083	3.42	3.90
GPT-4.1-mini	0.5373	0.6214	0.9110	0.8578	0.8134	0.8010	0.7700	0.4112	0.5012	0.5240	0.2600	0.7896	2.79	2.92
DeepSeek-v3.2	**0.6742**	0.7543	**0.9728**	**0.9687**	**0.9606**	0.9333	**0.9495**	0.5912	0.6499	0.6120	0.3540	**0.9419**	3.66	**4.00**
Gemini-3 Flash	0.6487	0.7686	0.9642	0.9585	0.9484	**0.9427**	0.9440	**0.6889**	**0.7524**	0.7052	0.3000	0.8328	3.49	3.58
Llama-4 Maverick	0.3772	0.6824	0.8958	0.6290	0.5849	0.5531	0.5275	0.2789	0.1820	0.3080	0.7340	0.8506	2.48	2.38
Claude-Sonnet-4.5	0.6541	0.7543	0.9691	0.9258	0.9437	0.9184	0.9048	0.5649	0.5828	0.5498	**0.8360**	0.9271	3.10	3.05
Qwen3-max	0.6570	0.7190	0.9402	0.8986	0.8671	0.8216	0.7943	0.6246	0.6981	0.6574	0.4060	0.8267	**3.68**	3.84

Task 1/2: semantic-matching F1. **Task 3**: recall under each distractor-memory setting. **Task 4**: binary scores for PA/PF/VRA/CF and 1–5 scores for EI. **PA** = Persona Awareness, **PF** = Preference Following, **VRA** = Virtual-Reality Awareness, **CF** = Constraint Following, **EI** = Emotional Intelligence.

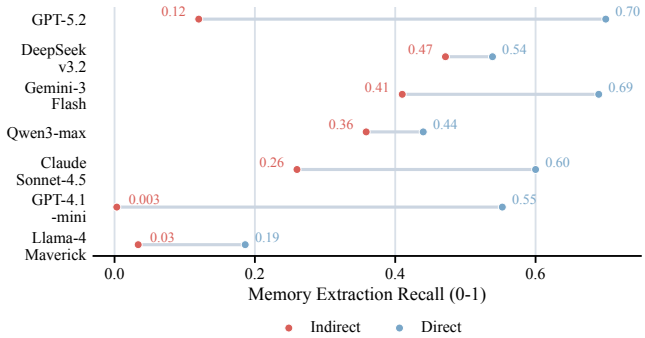


Figure 6: Task 1. Direct and indirect memory extraction recall across evaluated models.

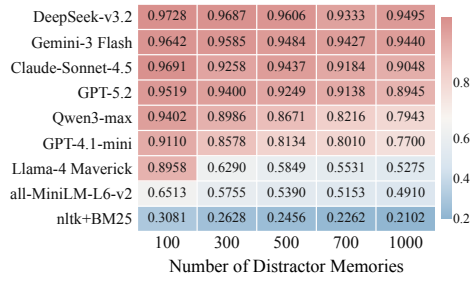
Table 3: Task 1. Extracted information matching results of memory-oriented systems.

	GPT-4.1 -mini	Ever-MemOS	A-Mem	MemOS no-dial	Mem0	Mem0g	MemOS	LightMem
Precision	0.3044	0.0500	0.1370	0.2850	**0.4460**	0.4010	0.3260	0.1200
Recall	0.3581	**0.9340**	0.8830	0.8500	0.6920	0.7230	0.8240	0.8120
F1	0.2970	0.0900	0.2250	0.3890	**0.4570**	0.4460	0.4330	0.1850

no-dial: MemOS without dialogue context. **Bold**: best. Underline: above GPT-4.1-mini.

is obtained at the cost of many false positives. In contrast, Mem0 provides the best overall balance, achieving the highest precision (0.4460) and F1 (0.4570), followed by Mem0g (0.4460 F1) and MemOS (0.4330 F1). These results indicate that simply expanding memory coverage is insufficient for personalized information extraction; effective memory systems must also filter and consolidate stored information to avoid introducing noisy or weakly personalized memories.

4.1.2 Task 2: Update Result Analysis. **General-purpose LLMs show a narrower but still visible performance gap on personalized information updating.** As shown in Table 2, GPT-5.2 achieves the best Task 2 F1 (0.7793), followed closely by Gemini-3 Flash (0.7686), DeepSeek-v3.2 (0.7543), and Claude-Sonnet-4.5 (0.7543). Compared with Task 1, the top-performing models are more tightly clustered, suggesting that the update task remains challenging but is less sharply differentiating among frontier LLMs under the current evaluation setting. Based on these results, Task 2 suggests that memory updating remains an unresolved challenge: even the strongest models achieve only moderate F1, and the tightly clustered top scores indicate that integrating evolving user information into persistent memory is difficult across frontier LLMs.



Models are ordered by their average retrieval score across distractor settings.

Figure 7: Task 3. Heatmap of retrieval scores across evaluated models as the number of distractor memories increases.

4.1.3 Task 3: Retrieval Result Analysis. **The sensitivity of general-purpose LLMs to distractor memories varies significantly, with the performance gap between models expanding as distractor density increases.** As evidenced by **Figure 7**, all seven evaluated LLMs maintain high and relatively uniform retrieval performance in low-noise environments (100 distractors), with scores tightly clustered between 0.8958 and 0.9728. In high-noise settings (1000 distractors), the strongest evaluated LLMs remain robust: Gemini-3 Flash and DeepSeek-v3.2 both achieve scores above 0.94, demonstrating strong resilience in retrieval tasks. However, given the inherent limitations of context window size and the escalating token costs associated with long-context processing, improving performance in realistic, long-horizon multi-turn scenarios requires a more lightweight and efficient retrieval architecture.

Current semantic-level retrieval methods are largely inadequate for personalized memory. There is a stark performance disparity between frontier LLMs and the retrieval components typically integrated into memory frameworks. Classical statistical methods such as *nltk+BM25* and semantic embedding baselines like *all-MiniLM-L6-v2* degrade substantially as distractor density scales, with retrieval performance dropping to 0.2102 and 0.4910, respectively. This suggests that "plug-and-play" retrieval modules are fundamentally inadequate for the complexities of long-term personalization. Future research should prioritize the development of logic-aware retrieval layers capable of bringing LLM-level reasoning to the memory retrieval process without the prohibitive costs of full-context processing.

4.1.4 Task 4: Utilization Result Analysis. **Existing LLMs struggle to perform well across all dimensions of personalized memory utilization.** From the results shown in Table 2, no single

Table 4: Task 4. Utilization performance comparison of memory-oriented systems.

Method	PA	PF		VRA	CF	EI	
		Gen.	Int.			EN	CN
GPT-4.1-mini	0.4112	0.5012	0.5240	0.2600	0.7896	2.79	2.87
Grd. Mem.	-	-	-	-	-	3.21	3.33
MemoryOS	0.3895	0.5195	0.5458	0.1460	0.7872	2.23	2.26
A-Mem	0.2895	0.5422	0.4303	**0.2240**	**0.8497**	1.98	1.95
EverMemOS	**0.7246**	**0.7888**	**0.6494**	0.1667	0.7265	**2.68**	**2.73**
Mem0	0.2561	0.4123	0.3825	0.1640	0.7988	1.88	2.02
Mem0g	0.2684	0.4091	0.3904	0.1960	0.7927	1.87	1.91
MemOS	0.2596	0.4943	0.4502	0.1440	0.8354	2.05	2.08
LightMem	0.2527	0.4411	0.4025	0.1579	0.8113	1.79	1.83

Grd. Mem. denotes presenting the query together with the ground-truth memories to the LLM.

general-purpose LLM consistently dominates all utilization dimensions. Gemini-3 Flash performs best on Persona Awareness (0.6889) and general Preference Following (0.7524), while GPT-5.2 leads on interactive Preference Following (0.7680). Claude-Sonnet-4.5 achieves the strongest Virtual–Reality Awareness (0.8360), DeepSeek-v3.2 performs best on Constraint Following (0.9419) and Chinese Emotional Intelligence (4.00), and Qwen3-max leads English Emotional Intelligence (3.68). This fragmented pattern suggests that personalized agentic competence is multidimensional: models that are strong at aligning with user preferences are not necessarily the most reliable at distinguishing real from virtual memories, following constraints, or producing emotionally appropriate responses.

While memory systems enhance model capabilities, they also introduce "personalization bias." From the results shown in Table 4, memory systems provide clear gains on some utilization dimensions, but the gains are selective. EverMemOS substantially improves Persona Awareness over the GPT-4.1-mini backbone (0.7246 vs. 0.4112) and also improves both general and interactive Preference Following (0.7888 vs. 0.5012; 0.6494 vs. 0.5240). However, these improvements do not transfer uniformly. All evaluated memory systems underperform the backbone on Virtual–Reality Awareness, with the best memory-system score reaching only 0.2240 compared with the backbone's 0.2600. Emotional Intelligence also consistently drops: even the best memory-augmented system reaches only 2.68 in English and 2.73 in Chinese, below the backbone's 2.79 and 2.87. Constraint Following shows a more mixed pattern, where A-Mem performs best (0.8497) but EverMemOS falls below the backbone (0.7265 vs. 0.7896).

These results suggest that current memory systems can make models more persona-aware and preference-sensitive, but may also introduce a form of personalization bias: once retrieved memories are injected, models can over-rely on them, confuse hypothetical or virtual information with real user attributes, and produce less emotionally adaptive responses. Future memory systems should therefore move beyond increasing memory capacity alone and place more emphasis on memory filtering, reliability verification, context grounding, and response-level control, so that personalization improves user alignment without sacrificing factuality, reality awareness, or emotional quality.

5 Related Work

LLM Personalization. LLM personalization research [20, 38, 43, 54, 61] primarily unfolds along three dimensions [9, 21, 30, 48, 55, 56, 59]. Personalized Prompting retrieves user-specific context without modifying parameters, evolving from simple storage to agentic architectures: Mem0 [5] and MemInsight [34] establish persistent memory layers; A-MEM [50] and Nemori [25] introduce autonomous decision-making; LightMem [7] employs a cognitive "Sensory-Short-Long" architecture with sleep-time updates; while MemOS [19] and MemoryOS [15] elevate memory to system-level resources, and EverMemOS [10] utilizes Engram-inspired lifecycle management. Personalized Adaptation focuses on parameter-efficient fine-tuning to internalize user patterns. Moving beyond PLoRA [53]'s specific adapters, RecLoRA [64] and iLoRA [16] leverage Mixture-of-Experts with dynamic routing for shifting preferences, while OPPU [39] balances privacy via collaborative training. Finally, Personalized Alignment optimizes training objectives: MORLHF [49] improves RL via multi-objective rewards, MODPO [63] integrates alignment into DPO, and Personalized Soups [11] merges parameters at inference for diverse preferences.

LLM Personalization Benchmarks. Existing benchmarks can be broadly categorized into two groups: Memory-free and Memory-aware. The first category, represented by LaMP [35], LongLaMP [17], and LaMP-QA [36], focuses on aligning generated content with user preferences in downstream tasks via retrieved static history. Similarly, PersonalLLM [65], PersonaFeedback [40], and PersoBench [1] evaluate adherence to pre-defined profiles. However, by treating user history as fixed context, these methods overlook the critical process of personalized information governance (*e.g.,* memory extraction, retrieval, and update). In contrast, Memory-aware benchmarks incorporate multi-dimensional assessments of memory capabilities. Early works like EmoBench [33] and EQ-Bench [29] targeted emotional alignment, while LoCoMo [23] and LongMemEval [47] target long-term memorization. Specific diagnostics such as PrefEval [57] and HaluMem [4] further address preference consistency and memory hallucinations. Recently, agentic simulations like PersonaLens [62], PersonaMem [12], and its successor PersonaMem v2 [13] have been proposed to track profile evolution. Despite these advances, their reliance on synthetic human-LLM dialogues creates a significant distributional gap from real-world data, resulting in homogeneous conversations [14] that lack natural diversity and fail to capture the implicit expressions characteristic of authentic human interactions.

6 Conclusion and Future Work

In this paper, we presented AlpsBench, a benchmark designed to evaluate the complete lifecycle of LLM personalization using real-world dialogue data. By leveraging long-term human-LLM interactions and expert-verified structured memories, AlpsBench enables a granular diagnosis of LLM performance across four key dimensions: personalized information extraction, update, retrieval, and utilization. Extensive experiments highlight several critical challenges for current LLMs, including difficulties in interpreting implicit user information, handling preference drift, and maintaining retrieval reliability under heavy interference. Going forward, we will continuously maintain and update the dataset, making it a dynamic benchmark to evaluate emerging LLMs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U25B2071 and 62525211).

References

[1] Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. 2024. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198* (2024).

[2] Anthropic. 2025. *Claude Sonnet 4.5 System Card.* Technical Report. Anthropic. https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf [Online; accessed Feb. 12, 2026].

[3] Anthropic. 2026. The Assistant Axis: Situating and Stabilizing the Character of Large Language Models. https://www.anthropic.com/research/assistant-axis. [Online; accessed Feb. 13, 2026].

[4] Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. 2025. Halumem: Evaluating hallucinations in memory systems of agents. *arXiv preprint arXiv:2511.03506* (2025).

[5] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413* (2025).

[6] DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *CoRR* abs/2512.02556 (2025).

[7] Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. 2025. Lightmem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866* (2025).

[8] Google DeepMind. 2025. *Gemini 3 Flash Model Card.* Technical Report. Google DeepMind. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf [Online; accessed Feb. 2026].

[9] Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A Survey on Personalized Alignment - The Missing Piece for Large Language Models in Real-World Applications. In *ACL (Findings) (Findings of ACL, Vol. ACL 2025)*. Association for Computational Linguistics, 5313–5333.

[10] Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, et al. 2026. EverMemOS: A Self-Organizing Memory Operating System for Structured Long-Horizon Reasoning. *arXiv preprint arXiv:2601.02163* (2026).

[11] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564* (2023).

[12] Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225* (2025).

[13] Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. 2025. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688* (2025).

[14] Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *NeurIPS*.

[15] Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI Agent. In *EMNLP*. Association for Computational Linguistics, 25961–25970.

[16] Xiaoyu Kong, Jiancan Wu, An Zhang, Leheng Sheng, Hui Lin, Xiang Wang, and Xiangnan He. 2024. Customizing language models with instance-wise lora for sequential recommendation. *Advances in Neural Information Processing Systems* 37 (2024), 113072–113095.

[17] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016* (2024).

[18] Amy Armento Lee, Narayan Hegde, Nina Deliu, Emily Rosenzweig, Arun Suggala, Sriram Lakshminarasimhan, Qian He, John Hernandez, Martin Seneviratne, Rahul Singh, et al. 2025. A Personalized Exercise Assistant using Reinforcement Learning (PEARL): Results from a four-arm Randomized-controlled Trial. *arXiv:2508.10060* (2025).

[19] Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, et al. 2025. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724* (2025).

[20] Xinyu Lin, Pengyuan Liu, Wenjie Wang, Yicheng Hu, Chen Xu, Fuli Feng, Qifan Wang, and Tat-Seng Chua. 2026. Bringing Reasoning to Generative Recommendation Through the Lens of Cascaded Ranking. *arXiv:2602.03692* (2026).

[21] Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528* (2025).

[22] Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025. A Survey of Personalized Large Language Models: Progress and Future Directions. *CoRR* abs/2502.11528 (2025).

[23] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *ACL (1)*. Association for Computational Linguistics, 13851–13870.

[24] Meta AI. 2025. Llama 4: Multimodal Intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. [Online; accessed Feb. 12, 2026].

[25] Jiayan Nan, Wenquan Ma, Wenlong Wu, and Yize Chen. 2025. Nemori: Self-organizing agent memory inspired by cognitive science. *arXiv preprint arXiv:2508.03341* (2025).

[26] OpenAI. 2025. GPT-4.1. https://openai.com/index/gpt-4-1/. [Online; accessed Feb. 12, 2026].

[27] OpenAI. 2025. *GPT-5.2 System Card.* Technical Report. OpenAI. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf [Online; accessed Feb. 2026].

[28] OpenAI. 2025. The Power of Personalized AI. https://openai.com/global-affairs/the-power-of-personalized-ai/. [Online; accessed Feb. 13, 2026].

[29] Samuel J. Paech. 2023. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models. *CoRR* abs/2312.06281 (2023).

[30] Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025. Latent Inter-User Difference Modeling for LLM Personalization. In *EMNLP*. Association for Computational Linguistics, 10599–10617.

[31] Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring What Makes You Unique: Difference-Aware User Modeling for Enhancing LLM Personalization. In *ACL (Findings) (Findings of ACL, Vol. ACL 2025)*. Association for Computational Linguistics, 21258–21277.

[32] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389.

[33] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *ACL (1)*. Association for Computational Linguistics, 5986–6004.

[34] Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. MemInsight: Autonomous Memory Augmentation for LLM Agents. In *EMNLP*. Association for Computational Linguistics, 33136–33152.

[35] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *ACL (1)*. Association for Computational Linguistics, 7370–7392.

[36] Alireza Salemi and Hamed Zamani. 2025. LaMP-QA: A Benchmark for Personalized Long-form Question Answering. In *EMNLP*. Association for Computational Linguistics, 1139–1159.

[37] sentence-transformers. 2025. all-MiniLM-L6-v2 Sentence Embedding Model. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. [Online; accessed Feb. 12, 2026].

[38] Wentao Shi, Xiangnan He, Yang Zhang, Chongming Gao, Xinyue Li, Jizhi Zhang, Qifan Wang, and Fuli Feng. 2024. Large language models are learnable planners for long-term recommendation. In *SIGIR*. 1893–1903.

[39] Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning. In *EMNLP*. Association for Computational Linguistics, 6476–6491.

[40] Meiling Tao, Chenghao Zhu, Dongyi Ding, Tiannan Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025. PersonaFeedback: A Large-scale Human-annotated Benchmark For Personalization. *arXiv preprint arXiv:2506.12915* (2025).

[41] Wiebke Wagner. 2010. Steven Bird, Ewan Klein and Edward Loper: Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit - O'Reilly Media, Beijing, 2009, ISBN 978-0-596-51649-9. *Lang. Resour. Evaluation* 44, 4 (2010), 421–424.

[42] Chengbing Wang, Yang Zhang, Wenjie Wang, Xiaoyan Zhao, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2025. Think-While-Generating: On-the-Fly Reasoning for Personalized Long-Form Generation. *CoRR* abs/2512.06690 (2025).

[43] Chengbing Wang, Yang Zhang, Wenjie Wang, Xiaoyan Zhao, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2026. Think-While-Generating: On-the-Fly Reasoning for Personalized Long-Form Generation. (2026).

[44] Chengbing Wang, Wuqiang Zheng, Yang Zhang, Fengbin Zhu, Junyi Cheng, Yi Xie, Wenjie Wang, and Fuli Feng. 2026. PERM: Psychology-grounded Empathetic Reward Modeling for Large Language Models. *arXiv preprint arXiv:2601.10532* (2026).

[45] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Trans. Mach. Learn. Res.* 2024 (2024).

[46] Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuan-Jing Huang, and Zhongyu Wei. 2025. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. In *Proceedings of the 31st international conference on computational linguistics*. 3310–3328.

[47] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *ICLR*. OpenReview.net.

[48] Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A. Rossi, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Namyong Park, Sungchul Kim, Huanrui Yang, Subrata Mitra, Zhengmian Hu, Nedim Lipka, Dang Nguyen, Yue Zhao, Jiebo Luo, and Julian J. McAuley. 2024. Personalized Multimodal Large Language Models: A Survey. *CoRR* abs/2412.02142 (2024).

[49] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. In *NeurIPS*.

[50] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110* (2025).

[51] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *CoRR* abs/2505.09388 (2025).

[52] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. In *NeurIPS*.

[53] You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024. Personalized LoRA for Human-Centered Text Understanding. In *AAAI*. AAAI Press, 19588–19596.

[54] Yang Zhang, Wenxin Xu, Xiaoyan Zhao, Wenjie Wang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2025. Reinforced Latent Reasoning for LLM-based Recommendation. *CoRR* abs/2505.19092 (2025).

[55] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. 2025. Personalization of Large Language Models: A Survey. *Trans. Mach. Learn. Res.* 2025 (2025).

[56] Zeyu Zhang, Yang Zhang, Haoran Tan, Rui Li, and Xu Chen. 2025. Explicit vs implicit memory: Exploring multi-hop complex reasoning over personalized information. *arXiv preprint arXiv:2508.13250* (2025).

[57] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *ICLR*. OpenReview.net.

[58] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *ICLR*. OpenReview.net.

[59] Xiaoyan Zhao, Ming Yan, Yilun Qiu, Haoting Ni, Yang Zhang, Fuli Feng, Hong Cheng, and Tat-Seng Chua. 2025. SteerX: Disentangled Steering for LLM Personalization. *CoRR* abs/2510.22256 (2025).

[60] Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2025. NextQuill: Causal Preference Modeling for Enhancing LLM Personalization. *CoRR* abs/2506.02368 (2025).

[61] Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2025. Nextquill: Causal preference modeling for enhancing llm personalization. In *arXiv:2506.02368*.

[62] Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025. PersonaLens: A Benchmark for Personalization Evaluation in Conversational AI Assistants. *arXiv preprint arXiv:2506.09902* (2025).

[63] Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. In *ACL (Findings) (Findings of ACL, Vol. ACL 2024)*. Association for Computational Linguistics, 10586–10613.

[64] Jiachen Zhu, Jianghao Lin, Xinyi Dai, Bo Chen, Rong Shan, Jieming Zhu, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Lifelong personalized low-rank adaptation of large language models for recommendation. *arXiv preprint arXiv:2408.03533* (2024).

[65] Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. 2025. PersonalLLM: Tailoring LLMs to Individual Preferences. In *ICLR*. OpenReview.net.