Towards Realistic Personalization: Evaluating Long-Horizon Preference Following in Personalized User-LLM Interactions

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly serving as personal assistants, where users share complex and diverse preferences over extended interactions. However, assessing how well LLMs can follow these preferences in realistic, long-term situations remains underexplored. This work proposes ***RealPref***, a benchmark for evaluating realistic preference-following in personalized user-LLM interactions. *RealPref* features 100 user profiles, 1300 personalized preferences, four types of preference expression (ranging from explicit to implicit), and long-horizon interaction histories. It includes three types of test questions (multiple-choice, true-or-false, and open-ended), with detailed rubrics for LLM-as-a-judge evaluation. Results indicate that LLM performance significantly drops as context length grows and preference expression becomes more implicit, and that generalizing user preference understanding to unseen scenarios poses further challenges. *RealPref* and these findings provide a foundation for future research to develop user-aware LLM assistants that better adapt to individual needs.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across a wide range of general-purpose tasks, such as mathematical reasoning (Wang et al., 2025), question answering (Touvron et al., 2023; Achiam et al., 2023). However, beyond serving as general-purpose assistants that support users in everyday contexts, there is a growing need for LLMs to support personalized needs, such as recommendation and decision-making. As users engage in long-term interactions with LLMs, they naturally reveal personal information, such as preferences and habits, through various forms of expression. In turn, users expect assistants to adapt to such signals and provide tailored responses to serve individual needs. Figure 1 illustrates a scenario example of user-LLM interaction that LLMs would encounter in the real world.

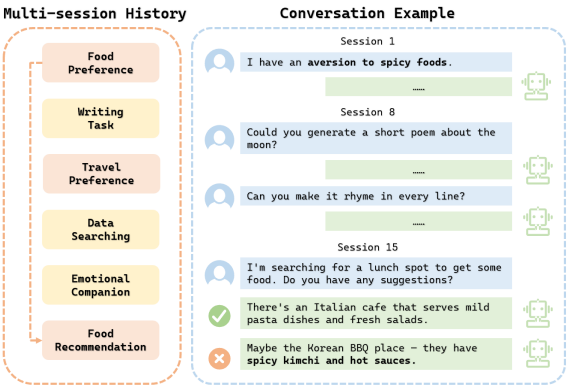

Figure 1: An example of user-LLM interaction: the conversation consists of several sessions on various topics. The user expresses a food preference in a previous session and then, after several sessions, asks the LLM for food-related recommendations.

In light of this, we aim to answer: can LLMs effectively (1) **capture** user preferences from complex user expressions, (2) **retain and follow** user preferences across extended interaction contexts, and (3) **generalize** user preference understanding to unseen scenarios?

Despite rising interest in LLMs' personalization, most efforts focus on simplified scenarios:

1. **Simple Context**. Existing works typically feature short-term context with few dialogue turns and center on a single preference or personalization instruction (Salemi et al., 2023; Kumar et al., 2024; Zhao et al., 2025) rather than a simulated user, resembling instruction-following instead of genuine personalization emerging from long-term dialogue.
2. **Straightforward Preference Expressions**. While real-world preference expressions are often implicit and scattered (e.g., in emotional cues), most studies assume explicit, immediately accessible traits (Jiang et al., 2025),

neglecting the challenge of generalizing preference understanding to unseen scenarios.

3. **Limited Evaluation Methodology**. Prevalent classification tasks diverge from natural user-LLM interactions (Jiang et al., 2025). Generation tasks, though realistic, lack reliable answer assessment criteria (Zhao et al., 2025).

This work proposes *RealPref*, a benchmark that features diverse user biographies, varying levels of preference expressions, and comprehensive evaluation protocols, constructing a testing environment that simulates the complexities of realistic, long-horizon user-LLM interactions. The main contributions of this work are summarized below:

1. **Dataset Construction.** We constructed a user-centered long-horizon dialogue dataset that incorporates diverse preference expressions (from explicit to implicit) and generalization scenarios to simulate the complexity of authentic user-LLM interactions.
2. **Evaluation Framework.** We proposed a multi-faceted evaluation protocol consisting of multiple-choice, true-or-false, and open-ended tasks, supported by granular and systematic rubrics to assess open-ended answers.
3. **Empirical Findings.** Through extensive experiments, we analyze whether LLMs can effectively capture and retain preferences, highlighting significant challenges in handling implicit expressions and generalizing to unseen scenarios as context length increases.

2 Related Work

Instruction Following Benchmarks. Existing benchmarks for instruction following primarily fall into two categories (Lou et al., 2024): **Task-centric evaluation** assesses problem-solving on predefined tasks using metrics like ROUGE (Lin, 2004) (e.g., MMLU (Hendrycks et al., 2020), IFEval (Zhou et al., 2023)); and **Human-centric evaluation** focuses on alignment with general human preferences through open-ended tasks (e.g., AlpacaFarm (Dubois et al., 2023)), relying on human or LLM judges. While these works focus on adherence to direct prompts, our work evaluates "preference following"—a nuanced extension where preference instructions are implied through interactions.

Long-Context Understanding Benchmarks. Early benchmarks focused on direct information retrieval, exemplified by "Needle in a Haystack" tests (Nelson et al., 2024) and reading comprehension tasks (Bai et al., 2024). Recent efforts have shifted towards multi-turn interactions, such as LoCoMo (Maharana et al., 2024) and Longmemeval (Wu et al., 2024), supporting sessions exceeding 100k tokens. Beyond simple fact recall, our work targets a higher level of capability: reasoning about preferences implied by history and actively applying this understanding to provide customized responses.

LLM Personalization Benchmarks. Benchmarks in this domain typically assess (1) **Persona Adoption**: role-playing (Wang et al., 2024) or mimicking styles (Salemi et al., 2023; Kumar et al., 2024); (2) **User Adaptation**: customizing outputs for question-answering (Zhao et al., 2025; Jiang et al., 2025) or downstream tasks like recommendation systems (Lyu et al., 2024). Our work targets the latter for direct question-answering. Existing benchmarks, however, lack depth: PREFEVAL (Zhao et al., 2025) centers on standalone preferences with coarse binary scoring criteria for open-ended answers, while PERSONAMEM (Jiang et al., 2025) overlooks implicit preference expressions and relies on logit-based metrics inapplicable to API-based LLMs. In contrast, *RealPref* provides a user-centered dialogue dataset with preference expressions ranging from explicit to implicit, supported by a granular, multi-dimensional rubric for evaluating open-ended responses.

3 *RealPref* Dataset

Problem Formulation. To systematically evaluate preference following over extended interactions, we first establish a formal representation of the interaction data. Let C denote the conversation history for a user with LLM, which consists of several sessions $C = \{S_1, S_2, ..., S_n\}$. Each session denotes a complete dialogue segment centered on a specific topic, which may or may not indicate user information. A session consists of multiple turns: $S_i = \{(u_1^i, a_1^i), (u_2^i, a_2^i), ..., (u_m^i, a_m^i)\}$, and each turn in the session contains a message from the user and a response from the LLM assistant. Each user has a preference set $P = \{p_1, p_2, ..., p_k\}$, and the preference is indicated in a certain session S in the history. For a preference p, a corresponding query q is designed to test LLMs' understanding of that preference. At the end of the conversation, the query q for target preference p will be presented to the test model to obtain a response for assessment.

Dataset Overview. Guided by the problem

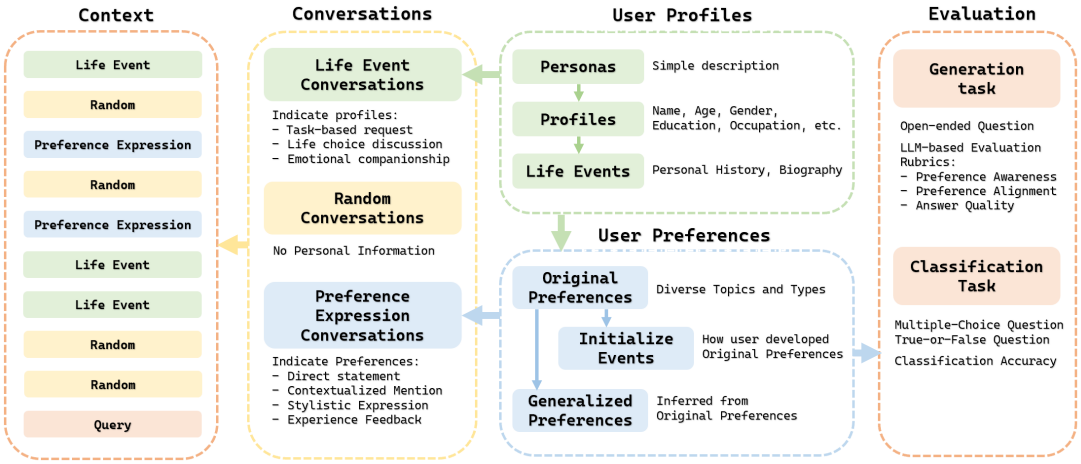

Figure 2: Generation Pipeline Overview. Starting from user personas, we construct detailed user profiles and biographies with diverse preferences. Using these, we construct rich conversation sessions that cover preference expressions from explicit to implicit, and concatenate them into a long-horizon context for each user. For each preference, generation and classification task queries are designed, with corresponding evaluation protocols.

framework, we follow the multi-stage dataset generation pipeline powered by GPT-4.1 (OpenAI, 2025a), as illustrated in Figure 2. *RealPref* consists of 100 curated user profiles, each with a persona, demographics, and a biography. Each user has 10 original preferences and 3 generalized preferences that can be rationally inferred from the original ones, which constructs 1300 preference-query pairs in total. The query has three forms: multiple-choice, true-or-false, and open-ended. Regarding conversations, there are 4 types of preference expressions for each original preference, and 10-15 sessions of life-event conversations for each user. In total, each user has a user-specific context of up to 2-3K tokens. Additionally, we have around 1000 turns of user-irrelevant random conversations used to control context lengths. The following sections will use Figure 2 as a roadmap to introduce our work in the following order: Profiles, Preferences, Conversations, Context, and Evaluation.

3.1 Profiles

We first sampled and manually filtered 100 random **Personas** from PersonaHub (Ge et al., 2025), a collection of 1 billion diverse personas from web data. Persona samples are brief sentences summarizing personal traits and are augmented with additional demographic details and descriptions to form complete **Profiles**. Furthermore, we constructed a **Life Events** history for each user, including life milestones like graduation. While unrelated to preference tests, these events help create a biography that facilitates personalized, long-term conversations. (See Appendix C.1 for a user profile example.)

3.2 Preferences

Original Preferences. The term "Original Preferences" refers to preferences generated directly based on user profiles. Each original preference involves an **Initialize Event** that records when and how the user developed the preference. (See Appendix C.2 for a preference example.)

High-quality preference data is essential for a robust evaluation. If preferences are overly generic or predictable—like "disliking fast food"—the task becomes trivial for LLMs: instead of leveraging user-specific context, models can simply rely on common-sense priors to pass the test. Therefore, we highlight requirements to ensure the generated preferences are *Diverse*, *Persona-Related*, *Unique*, *Value-Free*, and *Complete*, with details provided in Appendix C.2. Additionally, we introduced a *Topic Pool* and defined our *Preference Types* to ensure preference diversity:

- **Preference Topics.** To ensure topic diversity, we created a topic pool comprising 7 broad categories and 37 fine-grained subtopics. (See Appendix C.2 Table 3.) For each user, we sample 8 topics and have the LLM select 2 additional ones, thereby spreading topics widely while tailoring generation to individual traits.
- **Preference Types.** To capture the nuances of human preferences, we categorize them into distinct types listed in Table 1 based on their descriptive nature, which will help define criteria for evaluating open-ended responses.

Type	Description
Relative Preference	Prefer something over another
Negative Preference	Dislike or aversion to something (Anything else is acceptable)
Exclusive Preference	Only like or accept something (Anything else is rejected)
Positive Preference	Like or enjoy something (Without aversion to alternatives)

Table 1: Preference Types Classified by Descriptions.

Generalized Preferences. The term "Generalized Preferences" refers to preferences that differ from the original preferences but can be rationally deduced from them. People often have linked preferences across topics, driven by similar reasons. For example, someone who prefers cruelty-free, eco-friendly beauty products may also favor natural, eco-friendly cleaning supplies, reflecting a holistic lifestyle. When users state a preference with reasons, they often expect the assistant to recognize their general tendency across topics. This scenario resembles the way recommendation systems suggest new products based on user history. In tests, only original preferences will have expression conversations in the context, while generalized preferences will be used directly for tests. (See Appendix C.2 for a generalized preference example.)

3.3 Conversations

Based on the curated user profiles and preferences, conversation sessions are generated to simulate real-world user-LLM interactions. Depending on the type of user information indicated, conversations are primarily classified into three categories.

Life-Event Conversation. This category indicates user profiles and life events. Users' life event histories comprise distinct life stages, each of which requires different assistance from LLMs. Typical content includes task-based requests, life-choice discussions, and emotional companionship, through which the user gradually reveals their background, circumstances, thoughts, and emotions. (See conversation example in Appendix C.3.1.)

Random Conversation. This category will not reveal any user-specific information. Typically, the user begins with a detailed, practical request, such as factual questions, troubleshooting, or small talk. Given their user-free nature, random conversations are suitable for sharing among users and for managing context length in experiments. (See conversation example in Appendix C.3.2.)

Preference-Expression Conversation. This category indicates user preferences and plays a crucial role in our work. In reality, users may express their preferences gradually, implicitly through feedback or emotional cues, rather than explicit statements. Therefore, we provide four *Preference Expression Types* for each preference, from explicit to implicit. These four types require varying levels of capability, ranging from basic information extraction to complex comprehension and reasoning. Descriptions of each type are provided below. (See conversation examples of each type in Appendix C.3.3.):

1. **Explicit - Direct Statement.** Preferences are stated directly and explicitly in a single turn without redundant words.
2. **Explicit - Contextualized Mention.** In multi-round conversations, users start the session with other related daily topics as context instead of directly stating preferences. As the dialogue advances, preferences are naturally mentioned explicitly in the context.
3. **Implicit - Stylistic Expression.** The context resembles the second type, but users will not state preferences with explicit words like Like, Prefer, or Dislike. Instead, they imply preferences through stylistic devices such as rhetorical questions, metaphors, contrasts, or emotional expression.
4. **Implicit - Experience Feedback.** This type involves multiple sessions over time, based on the user's preference Initialize Event. The user won't explicitly state their preference in one session but will reveal it gradually through experience sharing or feedback across sessions. Preferences can only be inferred by integrating information from all sessions.

   For example, the early session may occur before the event, when users are willing to try without preference. After the event, users may share their experiences, give feedback, and indicate preferences.

3.4 Contexts

Based on the three types of conversation sessions generated, we can build a complete interaction history context for each user. The context includes all user-specific sessions (life-event and preference-expression conversations) in chronological order, with random conversations inserted to

Context	Interval	Tail	Total Length
Simple	–	–	2K
Normal	35K	–	37K
Long	70K	–	72K
Very Long	140K	–	142K
Extended	35K	105K	142K
Extreme	35K	210K	247K

Table 2: Context Configurations. *Interval* Tokens set the length of random conversations inserted into the interval. *Tail* Tokens set the length of random conversations added to the end.

control length. Two insertion methods are used: one randomly inserts into existing history to dilute personal information, and the other appends it at the end, distancing preference cues from the query. Table 2 lists the context configurations we provided for the following experiments.

4 *RealPref* Evaluation

4.1 Task Types

Based on the generated preferences, we provide classification (multiple-choice, true-or-false) and generation (open-ended) tasks. In classification tasks, LLMs can make decisions by matching options to history turns. Open-ended tasks challenge LLMs more by requiring them to proactively generate responses aligned with preferences. (See Appendix D.1 for a preference query example.)

Open-Ended Questions. Besides being designed based on preferences, high-quality queries must also be discriminative: answering without knowledge of preferences may violate them. Thus, the query content should not disclose user preferences, ensuring that LLMs cannot rely solely on common sense reasoning to respond accurately.

Multiple-Choice Questions (MCQs). The queries in the MCQ setting are identical to the open-ended ones. Four options are generated, all relevant and reasonable answers to the query, but only one best aligns with the preference. However, we noticed flaws in this setting: LLMs, given a question with four options and told only one is correct, might choose "the odd one out" and answer correctly by comparing options even without knowledge of the user's preference.

True-or-False Questions. Based on the above analysis, we designed this new classification task. A true-or-false question combines an open-ended query with one MCQ option, asking whether the option is suitable. In this setting, LLMs must decide based on a single option without option comparison. We construct a correct and an incorrect version for each query, yielding a balanced dataset.

4.2 Evaluation Methods

While classification tasks adopt accuracy as the metric, generation tasks require a comprehensive set of evaluation criteria to assess response quality. Based on prior work (Zhao et al., 2025) and an analysis of actual responses, this work provides a more detailed scoring criterion on a 1-5 scale from 3 rating dimensions: *Preference Awareness*, *Preference Alignment*, and *Answer Quality*. (See Appendix D.2 for detailed evaluation rubrics.)

Preference Awareness. The first dimension focuses on the extent to which the answers recognize or understand user preferences. The judgment is based on the following: If an LLM is aware of user preferences when responding, it should mention the preferences it uses in the answer.

Preference Alignment. The second dimension focuses on how well the answer's content aligns with user preferences, by checking if the answer's actual content aligns with user preferences, with criteria discussed by preference types in Table 1.

Answer Quality. The last dimension focuses on the extent to which the answer's content is actually helpful and constructive to the user. Although not directly tied to preferences, it is a crucial factor for assessing the quality of open-ended responses.

5 Experiments

5.1 Experiment Settings

Benchmark Configurations. Our experiments explore the performance of current LLMs across several dimensions. Figure 3 illustrates an overview of different benchmark configurations and their related control components in the experiment: *Question Type* sets the test task and evaluation method; *Preference Expression Type* controls the preference expression sessions included in the context; *Context Configuration* controls the context length by managing the insertion of random sessions.

Base Models. We chose the following LLM base models for experiment: *GPT-5*, *GPT-5 mini*, *Qwen3-235B-A22B Instruct*, *Gemini 2.5 Flash-Lite*, *Llama 3.3 70B Instruct*. For open-ended answers, we specify *Claude Sonnet 4* as our LLM judge. (See Appendix E.1 for model details.)

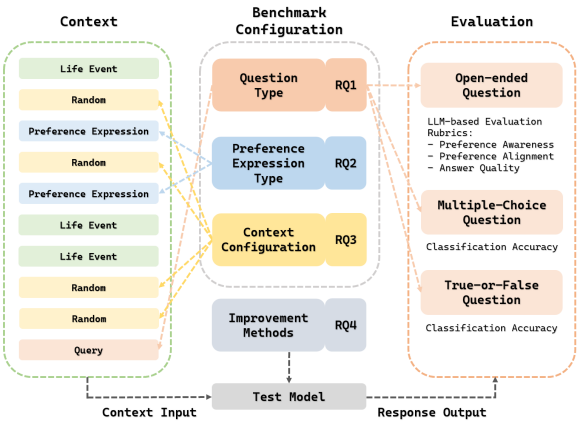

Figure 3: Benchmark Configuration Overview. Question Type (*Multiple-Choice*, *True-or-False*, and *Open-Ended*) sets the test and evaluation; Preference Expression Type (*Direct Statement*, *Contextualized Mention*, *Stylistic Expression*, and *Experience Feedback*); Context Configuration sets the context input; Improvement Methods (*Reminder*, *Few-Shot CoT*, and *RAG*) are tested with models. Each factor above is related to a research question (RQ) we will explore in the experiments.

Methods. By default, we test models zero-shot as a baseline, but we also evaluate several basic improvement methods: (1) **Reminder**: Add a sentence reminding the model to review the user's past preferences in response. (2) **Few-shot Chain-of-Thought**: For open-ended questions, give 3 examples with excellent responses that mention user preferences, include detailed chain-of-thought, and offer specific guidance. (3) **Retrieval-Augmented Generation (RAG)**: Offer the top 5 retrieved history turns that have the highest similarity to the current question using a sentence embedding model. (See Appendix E.2 for detailed implementation.)

5.2 Research Questions

Based on the setting factors described above, we proposed the following research questions (RQs):

RQ1: Which **evaluation task** is better at helping us understand the preference following abilities of LLMs?

RQ2: How does LLM's preference-following ability changes across different **preference expression** types?

RQ3: How does LLM's preference-following ability perform in **long contexts**?

RQ4: How do different **improvement methods** affect the LLMs' ability to follow user preferences?

RQ5: Can LLMs **generalize** user preferences to unseen scenarios?

The subsequent experiments will focus on these questions and delve deeper into the model's performance under various conditions.

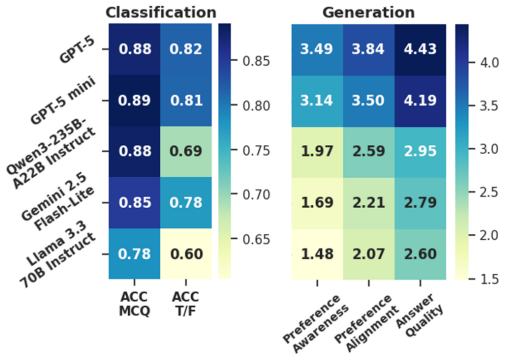

Figure 4: Model Performance across Question Types with *Normal* context, *Experience Feedback* expression, *Zero-shot*.

5.3 Experiment Results and Analysis

5.3.1 Question Type Comparison

To investigate **RQ1**, we test candidate LLMs in *Normal* context, *Experience Feedback* expression type, *Zero-shot*, across three types of evaluation tasks: *MCQ*, *True-or-False*, and *Open-ended Questions*. Figure 4 shows the experiment results.

For classification tasks, True-or-False questions are better than MCQs in reflecting LLMs' ability to follow user preferences. As shown in Figure 4, MCQ scores are generally much higher than the random line at 0.25, with slight differentiation across LLMs. Despite having options from MCQs, True-or-False question scores are typically lower (closer to the random line at 0.5). The overly high scores on MCQs confirm the flaws we previously noted: the LLM can exploit which options appear different (i.e., 'the odd one out') to infer likely answers, even without knowledge of user preferences, leading to overly high scores on the MCQ task. Such a situation is more common when the user preference belongs to *Negative Preference* (Dislike or aversion to something): Since the three incorrect options share the same characteristics, the correct option is more likely to stand out. In contrast, True-or-False questions prevent option comparison, making results more accurately reflect LLMs' ability to follow user preferences. (See query example in Appendix D.1.)

Compared to classification tasks, generation tasks are better at distinguishing between LLMs that can proactively follow user preferences and those that cannot. As shown in Figure 4, the ratings for generation tasks clearly differentiate

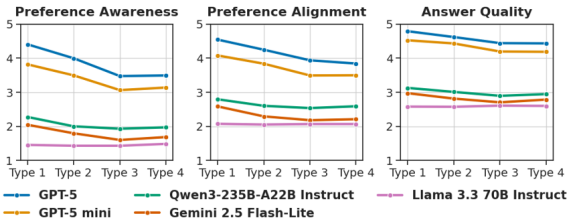

Figure 5: Model Performance across Expression Types with *Normal* context, *Zero-shot*. Type 1: Explicit - Direct Statement, Type 2: Explicit - Contextualized Mention, Type 3: Implicit - Stylistic Expression, Type 4: Implicit - Experience Feedback.

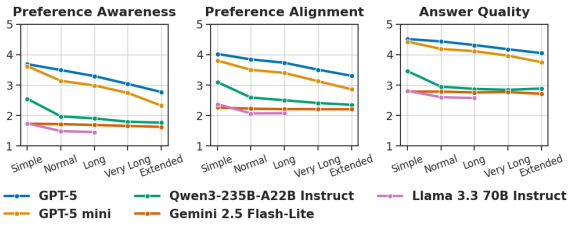

Figure 6: Model Performance across Context Configurations with *Experience Feedback* expression, *Zero-shot*. Context Length in tokens: Simple (2K), Normal (37K), Long (72K), Very Long (142K), Extended (142K).

model performance: GPT-5 and GPT-5 mini perform much better than the other three LLMs on all three dimensions. Such a distinction cannot be reflected in classification tasks, possibly because passively selecting an answer from the given options allows retrieval of the options from history. In contrast, the generation task provides only a query, requiring LLMs to proactively generate answers that follow user preferences.

5.3.2 Effect of Preference Expressions

Results toward **RQ1** demonstrated the effectiveness of generation tasks. For **RQ2**, we test LLMs in *Open-ended Question*, *Normal* context, *Zero-shot*, across four types of preference expression. Results are shown in Figure 5.

LLMs' overall preference following ability noticeably declines from explicit expressions to implicit expressions. Figure 5 illustrates that the scores across all rating dimensions significantly decline from explicit expression (Type 1 and 2) to implicit expression (Type 3 and 4), with Type 1 scoring higher than Type 2. The findings above align with expectations: *Direct Statement* clearly presents the preference, which is the easiest to identify. *Contextualized Mention* mixes preferences with context, posing challenges in extraction. *Stylistic Expression* and *Experience Feedback* further complicate the task, requiring LLMs to capture preference information beneath the expressions. This downward trend is more pronounced in GPT-5 and GPT-5 mini than in the other three LLMs. This may be because other models do not actively retrieve or follow user preferences, so expression variations minimally affect their performance.

5.3.3 Preference Following in Long Contexts

For **RQ3**, we test candidate LLMs in *Open-ended Question*, *Experience Feedback* expression type, *Zero-shot*, across five context configuration types defined in Table 2: *Simple* (2K), *Normal* (37K), *Long* (72K), *Very Long* (142K), and *Very Long Extra* (142K). Results are shown in Figure 6.

LLMs' preference following performance drops significantly as context length increases. Figure 6 shows a clear downward trend in rating scores as context length increases from *Simple* to *Very Long*, particularly in the *Preference Awareness* and *Preference Alignment* dimensions. Although under the same context lengths, there is a further drop from *Very Long* to *Extended* by changing the insertion location of random sessions from the history interval to the tail, showing that besides overall context length, the interval between useful preference cues and the final query also affects performance. Similar to the observation in Section 5.3.2, such a downward trend is more obvious for the GPT-5 series than the other three LLMs.

5.3.4 Effect of Basic Improvement Methods

For **RQ4**, we compared the performance of candidate LLMs in *Open-ended question*, *Experience Feedback* expression type, with *Extended* (142K), and *Extreme* (247K) context across baseline and 3 methods: *Zero-shot*, *Reminder*, *Few-shot Chain-of-Thought*, and *Retrieval-Augmented Generation*. The experiment results are shown in Figure 7.

For context length within the LLM's retrieval capability, *Reminder*, *Few-shot CoT* and *RAG* show similar improvement effects. In Figure 7(left), for the GPT-5 series, all three methods show similar improvements over the *Zero-shot* setting. *Reminder* remains the simplest and most cost-effective method, providing performance comparable to more advanced approaches.

For longer contexts, *RAG* is most effective in improving model performance. In Figure 7(left), for Qwen3-235B-A22B Instruct and Gemini 2.5 Flash-Lite in the *Extended* (142K) context, *RAG*

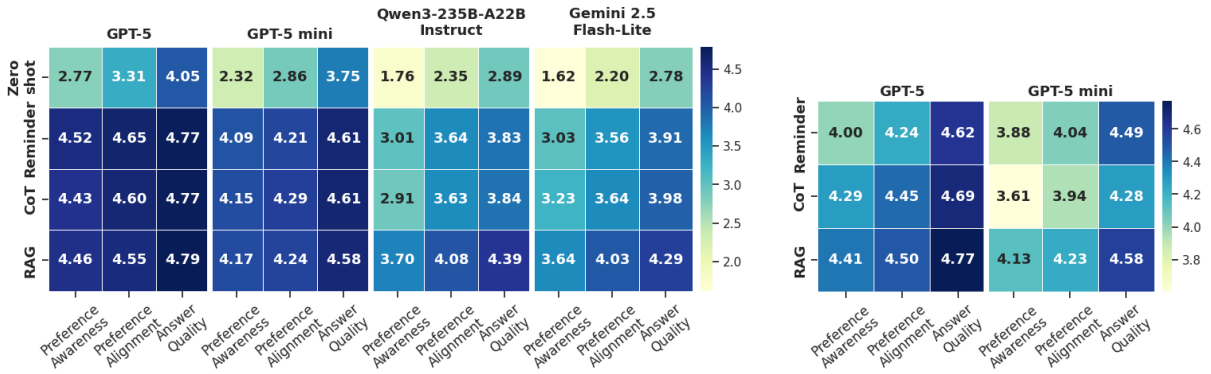

Figure 7: Model Performance across Methods with *Experience Feedback* expression, under *Extended (142K)* context (left) and *Extreme (247K)* context (right).

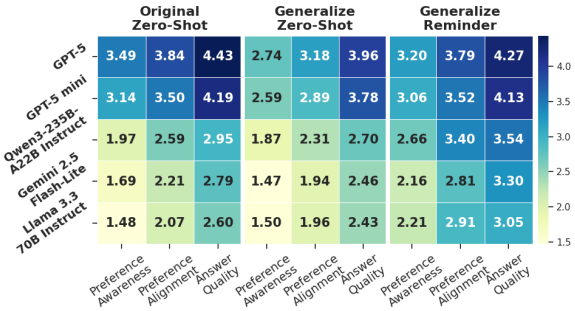

Figure 8: Model Performance for Preference Generalization with *Normal* context, *Experience Feedback* expression.

outperforms the other methods. Similar situations occurred for GPT-5 series in the *Extreme* (247K) context shown in Figure 7(right). These findings indicate the effectiveness of *Retrieval-Augmented Generation* in enabling LLMs to achieve preference following in contexts that may exceed their retrieval capabilities.

5.3.5 Preference Generalization

To explore **RQ5**, we test candidate LLMs in answering generalized preference queries in *Open-ended Question*, *Experience Feedback* expression type, *Normal* context, with *Zero-shot* and *Reminder* methods. Generalized preferences differ from the original ones but can be inferred from them because they share similar underlying reasons. The contexts contain only the original preference expressions, requiring LLMs to generalize their understanding to unseen scenarios, which is more in line with real-world interactions. The experiment results are shown in Figure 8.

Current LLMs lack proactivity to generalize the understanding of user preferences to new scenarios. Figure 8 shows LLMs performed worse on generalized preference queries than on original ones in the same *Zero-shot* setting. It is notable that for Qwen3-235B-A22B Instruct, Gemini 2.5 Flash-Lite and Llama 3.3 70B Instruct, the scores for generalized preferences with *Reminder* are higher than the scores for original ones under *Zero-shot* setting, showing that *Reminder* effectively helps stimulate these LLMs' proactivity to recall original preferences, infer generalized preferences, and give preference-aligned answers. However, for the GPT-5 series, although *Reminder* improves performance, the scores for generalized preferences remain lower than those of the original ones under the *Zero-shot* setting, indicating that following generalized preferences still poses more challenges for LLMs.

6 Conclusion

In this work, we proposed *RealPref*, a comprehensive benchmark for evaluating LLMs' ability to follow preferences in long-horizon, personalized user-LLM interactions. *RealPref* features a long-horizon conversation dataset constructed by a user-centric generation pipeline. Our design considers diverse preference topics and types, complex preference expressions, preference generalization scenarios, and different evaluation tasks. Experimental results show that many LLMs lack proactivity in recalling and following user preferences, and even capable models exhibit significant performance degradation as context length increases and preference expressions become more implicit. Besides, generalizing learned preferences to unseen scenarios poses additional challenges. *RealPref* and these findings may provide a foundation for future research to develop smarter, user-aware LLM assistants tailored to individual needs.

Limitations

This work introduces *RealPref*, a benchmark for evaluating realistic preference-following in long-horizon user–LLM interactions. However, it also has limitations to acknowledge.

Synthetic dataset. All user profiles, preferences, and dialogue histories in our work are synthetically generated. Although this approach enables controlled evaluation and protects privacy, synthetic users may not fully capture the complexity and unpredictability of human behavior. Besides, the topic pool, preference space, and preference expressions, while diverse, remain finite and may not cover all domains, languages, or cultural contexts encountered in real-world applications. Nevertheless, we make deliberate efforts through diverse topic design and varied preference expressions to make the generated data as closely approximate real-world user behavior as possible.

LLM judges in Evaluation. Our work relies on LLM-as-a-judge with predefined rubrics. While automatic judging may not perfectly align with human judgments, we carefully design detailed, structured evaluation criteria to enhance consistency and reliability. Moreover, compared with large-scale human evaluation, this approach offers a significantly more cost-efficient and scalable alternative, making large-scale systematic evaluation feasible.

Experiments. As a benchmark effort, the experiment part of our work is to reveal the performance of current LLMs, with the settings focused on varied benchmark configurations. Since we provided many variable settings in our benchmark, we varied the research variable while holding the others at representative settings to simplify the experimental procedure. Also, given the long context and API costs involved in the test, we assume that the LLM candidates and the judge used in the experiment are stable versions that yield stable test results, so the presented test results are from a single run.

Ethical Considerations

This work serves as a benchmark that enables a more realistic assessment of how LLMs infer and apply user preferences across extended contexts. In addition to the contribution, we have considered the following possible ethical aspects of our work.

Data Privacy and Safety. *RealPref* is designed as an evaluation benchmark, and all user profiles, preferences, and interaction histories are synthetically generated, without involving real personal data. We manually reviewed and excluded ineligible data samples throughout the generation pipeline, ensuring that the generated content contains no unethical or sensitive components.

Dual Use. Our work paves the way for research into smarter, user-aware LLM agents customized for individual needs, but it may also have dual-use potential. Our dataset is synthetic and is provided solely for evaluation; it is prohibited to use the data for unethical purposes, such as disinformation or the generation of fake profiles. Regarding user privacy concerns associated with the future practical deployment of LLMs with strong preference-following capabilities, we advocate responsible, transparent model deployment that prioritizes user data privacy and security, enabling users to control their data.

Bias and Over-Personalization. We strive to ensure preference data diversity in our dataset to minimize the chances of producing unfair or biased personalization data. Besides, our personalization tasks rely solely on preference information that users actively share with the model during dialogue. Although preferences may be expressed implicitly, the data generation is carefully designed to ensure that preference inference is grounded in clear, interpretable cues within the dialogue, rather than in speculative or unsupported assumptions, thereby reducing biased output and over-personalization. Such considerations are also required for subsequent use of the benchmark.

Research Integrity. Our work highlights the importance of responsible, transparent, and interpretable practices. The dataset and detailed documentation of our work, including code and prompts, are publicly available to support reproducibility. Our research does not involve human subjects or personal data, and the use of AI assistants in our work is detailed in Appendix A.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2025. Introducing claude 4. https://www.anthropic.com/news/claude-4. Accessed: 2025-05-23.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao

Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Google Deepmind. 2025. Gemini 2.5 flash-lite. https://deepmind.google/models/gemini/flash-lite/. Accessed: 2025-07-22.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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*.

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*.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, and 1 others. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Sheng Li and Handong Zhao. 2021. A survey on representation learning for user modeling. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4997–5003.

Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. Llm-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612.

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

Meta. 2024. Llama 3.3. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/. Accessed: 2024-12.

Elliot Nelson, Georgios Kollias, Payel Das, Subhajit Chaudhury, and Soham Dan. 2024. Needle in the haystack for memory based large language models. *arXiv preprint arXiv:2407.01437*.

OpenAI. 2025a. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-04-14.

OpenAI. 2025b. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-08-07.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

QwenTeam. 2025. Qwen3: Think deeper, act faster. https://qwen.ai/blog?id=1e3fa5c2d4662af2855586055ad037ed9e555125. Accessed: 2025-04-29.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, and 1 others. 2025. A survey on large language models for mathematical reasoning. *ACM Computing Surveys*.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Use of AI Assistants

In this work, we used AI assistants for a limited scope. Regarding coding, we used AI to assist with short-form input, such as offering package usage suggestions and reformatting the written code for readability. For writing, we used AI assistants solely for the paper's language and formatting, such as grammar checking, paraphrasing, and polishing, to ensure a clear, readable delivery. All the above uses are strictly confined to the drafted content by the human authors. All AI-generated suggestions are rigorously reviewed, and all editorial decisions are reserved for human authors. AI assistants are not used for literature searches or generating text with new ideas. All the intellectual contributions presented are the work of the authors.

B Future Directions

Based on the current benchmark design, there are some limitations that can be considered as directions for future research.

- **Complex Preference**: In the real world, user preferences can change over time or depend on the condition of something else. Enriching the dataset with such complex preferences would be a possible direction.
- **Richer Interaction Context**: Besides simplified dialogue types, real-world interactions may involve richer, multimodal contexts. Exploring how to enrich interaction contexts is also a promising future direction.
- **Complex User Feedback**: Besides the current preference expression types designed, other forms of indication, such as thumbs-up or thumbs-down to LLM responses, could be considered for possible future direction.
- **Evaluation Extension**: The evaluation rubrics for open-ended responses could be expanded to include additional dimensions, such as fairness, privacy, and transparency, better aligning the benchmark with responsible AI principles.

C Dataset Details and Examples

C.1 Profiles

Profile Example. The following is an example of a user profile, beginning with a user ID "00001" and including demographics information, an original persona sampled from PersonaHub, and an extended persona description.

```
{
  "00001": {
    "Demographics": {
      "Name": "Alex",
      "Birth": "1999",
      "Gender": "Male",
      "Education": "Bachelor",
      "Occupation": "Music Blogger",
      "Economic Status": "Middle Class
          ",
      "Marital Status": "Single"
    },
    "Original Persona": "a die-hard
        fan of Tink who runs a music
        review blog",
    "Extended Persona": "Alex is a
        passionate music enthusiast
        born in 1999. He's a dedicated
        fan of Tink and channels his
        love for her music into an
        engaging review blog. Alex, who
         identifies as male and is of
        Asian descent, brings a unique
        perspective to the music scene
        with his thoughtful analyses
        and compelling writing style.
        His blog has gained a loyal
        following for its authentic
        voice and deep dives into the
        nuances of music and lyrics.
        Outside of blogging, Alex
        enjoys attending live concerts
        and discovering up-and-coming
        artists in his spare time."
  }
}
```


Life Event Example. Here is an example of a user's life event, with a specific date, event description, and duration.


```
[
  {
    "Date": "2011/06/15",
    "Event": "Graduated with a diploma
         in Makeup Artistry and began
        working as a freelance makeup
        artist, taking on clients for
        weddings, photoshoots, and
        local events."
  },
  {
    "Date": "2014/03/10",
    "Event": "Collaborated with
        several local salons,
        expanding her network and
        gaining regular clientele,
        while building her
        professional reputation within
         the community."
  },
  {
    "Date": "2016/11/05",
    "Event": "Launched her social
        media channel dedicated to
        honest and detailed makeup
        product reviews, steadily
        attracting a loyal following
        in the beauty community."
  }
]
```


C.2 Preferences

Generation Requirements. To ensure high-quality preference data in generation, we highlight requirements to ensure the generated preferences are *Diverse*, *Persona-Related*, *Unique*, *Value-Free*, and *Complete*, with specific descriptions listed below:

1. **Diverse**: The preferences covering a variety of different topics and types.
2. **Persona Related**: Each preference should be consistent with the persona's profile, sound realistic, and reflect individuality.
3. **Unique**: Avoid generic, cliché preferences such as "dislikes fast food." Instead, provide more nuanced, distinctive preferences that are non-popular, non-obvious, and cannot be deduced by common sense.
4. **Value-free**: Do not always align with socially desirable, universal "good", "healthy", logically "correct" preferences. Include preferences that may be unconventional or mildly negative (e.g., enjoy fast food late at night, like noisy environments).
5. **Complete**: The preference description should include the content and the underlying reasons, specify the type and topic.

Topic Pool. To ensure topic diversity, we introduced a topic pool comprising 7 broad categories and 37 fine-grained subtopics. The topic pool is listed in Table 3.

Preference Example. The following is an example of a user preference, with an intialize event indicating how and when the user developed it. The fields include the preference and the topic, a specific date, and an event description.


```
{
  "Preference": "Prefers herbal teas
      and fresh fruit snacks over
      sugary sodas and packaged sweets
      , as she values a balanced
      lifestyle.",
  "Topic": "Food: Snacks \& Drinks",
  "Date": "2014/06/21",
  "Event": "During a beauty workshop,
```

Topic Fields	Subtopics
Education	Learning Style, Learning Resources
Entertainment	Music, Movies & Shows, Games, Reading, Online Media
Lifestyle	Daily Routine, Exercise, Shopping, Beauty, Fashion, Tech Use, Home Decoration
Food	Cuisine, Cooking, Snacks & Drinks, Diet, Dining Style
Travel	Destinations, Activities, Companions, Pace & Frequency, Accommodation, Transportation, Planning Style
Work & Living	Location & Environment, Traffic & Commuting, Climate, Work Conditions & Hours, Work Values, Work Style & Tasks
Social	Social Energy, Friendship, Family, Romance, Social Events

Table 3: Preference Topic Pool

```
      Emma noticed that consuming
      sugary sodas and snacks left her
       feeling sluggish, while herbal
      teas and fruit refreshed and
      sustained her energy. She made a
       conscious decision to adopt
      fresher, healthier snacks both
      at work and at home."
}
```

Generalized Preference Example. Generalized preferences are generated based on the user's original preferences. We utilize Few-shot Chain-of-Thought during this process to ensure that the generalized preference can be logically derived from the original ones. Below is a detailed example, including a generalized preference, referenced original preferences, and the LLM's generation rationale.

```
{
  "Original Preference": [
    {
      "Topic": "Food: Cuisine",
      "Preference": "Only interested
          in trying global street
          foods or fusion cuisines,
          generally refusing classic
          fine-dining establishments
          which she finds
          unadventurous and
          restrictive."
    },
    {
      "Topic": "Travel: Accommodation
          ",
      "Preference": "Prefers staying
          in quirky boutique hotels or
           artist-run guesthouses
          rather than large hotel
          chains, as the unique
          atmosphere inspires her work
          ."
    },
    {
      "Topic": "Lifestyle: Beauty",
      "Preference": "Enjoys
          experimenting with
          unconventional makeup looks,
           such as vintage glam or
          fantasy-inspired themes,
          especially when attending
          local theatre events."
    }
  ]
}
```

```
{
  "Generalized Preference": {
    "Infer_reason": "Her dislike of
        monotonous routines,
        regimented systems, and
        classic or mainstream
        offerings across various
        domains (food, beauty,
        exercise, music) suggests a
        desire for originality and
        novelty.",
    "Topic": "Entertainment: Movies &
        Shows",
    "Preference": "Dislikes formulaic
        blockbuster movies and prefers
         independent films or
        experimental cinema, as
        unconventional storytelling
        stimulates her creative
        thinking."
  }
}
```

C.3 Conversations

C.3.1 Life-Event Conversation

Life-Event Conversations indicate user profiles and life events. Through dialogue, users share background, circumstances, thoughts, and feelings. Below is an example.

User: Hello! I've recently started working as a freelance makeup artist, and I have a wedding client next month who wants a classic vintage look. I'm a bit stuck on product choices for oily skin types. Could you recommend some foundation and powder options that hold up all day?
Assistant: Hi there! Congrats on getting a wedding client—vintage looks are always stunning. For oily skin, you'll want a mattifying, long-wear foundation and a setting powder with good oil control. Some popular options include Esté Lauder Double Wear Foundation, MAC Studio Fix Fluid, or Revlon Colorstay. As for powders, Laura Mercier Translucent Setting Powder or Rimmel Stay Matte are widely loved. Be sure to prep the skin with an oil-free primer and consider using blotting sheets for touch-ups. Would you like advice on specific application techniques for a vintage look as well?
User: Thank you, that really helps! Yes, any tips on keeping the winged liner and red lipstick looking sharp all night? Vintage looks seem easy but can be tricky under pressure. I want to make sure everything stays put for photos and dancing.
Assistant: Absolutely, vintage-inspired makeup does require a bit of extra care! For winged liner, using a waterproof gel or liquid liner will help prevent smudging—Inglot AMC Gel Liner or Stila Stay All Day Liquid Liner are good options. Set the liner with a black eyeshadow using a fine brush for added longevity. For lips, start with a lip primer, line the lips with a long-wear red liner, fill them in, then layer your lipstick (like MAC Ruby Woo or Maybelline Superstay Matte Ink), and blot between layers. Setting with a tiny bit of translucent powder through a tissue can help too. Have you been enjoying freelance work so far, or is it still new to you?
User: I actually just graduated last summer from a makeup artistry program, so this is my first full wedding season. Freelance is exciting but also a bit nerve-wracking! Thanks for all the product tips—I really want to make a great impression and hopefully get more referrals.
Assistant: That's wonderful—congratulations on launching your career! It's normal to feel a bit anxious, especially early on, but you're doing all the right things by preparing ahead. Each happy client is a stepping stone to more opportunities. If you'd ever like to chat about building your portfolio or networking with local salons, feel free to ask. Good luck with the wedding—you've got this!

C.3.2 Random Conversation

Random Conversations are about common-sense questions and will not reveal user-specific information. Below is an example with the assistant response omitted for simplicity of demonstration.

User: I'm having trouble connecting to Wi-Fi on Windows 10. Any troubleshooting tips?
Assistant: [Response]
User: What if none of these suggestions work?
Assistant: [Response]

C.3.3 Preference-Expression Conversation

Preference-Expression Conversations indicate user preference. We provide four *Preference Expression Types* from explicit to implicit, each with an example listed below.

1. **Explicit - Direct Statement**: Preferences are stated directly and explicitly in a single turn without redundant words.

User: Hi! I want to get more physically active, but **I really avoid routine gym workouts because they're monotonous and uninspiring for me. I much prefer spontaneous activities like dancing at local community events.** Any ideas for keeping fit that aren't boring gym routines?
Assistant: Absolutely, I understand that traditional gym workouts aren't for everyone—it's important to have fun while staying active! Here are a few ideas that can keep things lively and spontaneous: 1. Dance Classes or Social Dancing: Try drop-in classes for salsa, swing, or even hip-hop at local studios or community centers. Social dance nights are great for meeting people and get your heart rate up without feeling like exercise. 2. Outdoor Group Activities: Join recreational sports leagues (like volleyball, ultimate frisbee, or even kickball) where the focus is on fun and socializing rather than intense competition. 3. Hiking Groups or Urban Walks: Explore nature trails, city walking tours, or even organized hiking meet-ups. Each outing is a little different and can double as a mini-adventure. 4. Fitness Meetups: Check out community boards or social media for pop-up yoga in the park, outdoor bootcamps, or even roller skating meetups. 5. Volunteering: Since you enjoy com-

munity events, active volunteering (like setting up events, ushering at theaters, or organizing local festivals) can keep you moving while contributing to a good cause. Mixing things up and following what excites you can make staying active something you actually look forward to. Would you like more suggestions tailored to your interests or info on how to find local events?

2. **Explicit - Contextualized Mention**: In multi-round conversations, users start the session with other related daily topics as context instead of directly stating preferences. As the dialogue advances, preferences are naturally mentioned explicitly in the context.

User: Hi there! I was looking for some ideas to keep my weekends fun and active without spending too much money. Got any suggestions?
Assistant: Absolutely! There are plenty of ways to have an active and enjoyable weekend without breaking the bank. Here are a few ideas: 1. Go for a nature walk or hike in a local park. 2. Join a community sports group that has free or low-cost sessions. 3. Try a YouTube dance or workout class at home with friends. 4. Attend local events or street fairs, which often offer free activities and entertainment. 5. Organize a picnic with friends and some outdoor games like frisbee or badminton. Would you like more suggestions tailored to your interests or any particular type of activity?
User: Thanks! I like the idea of attending local events. Lately, I've noticed several community dance nights popping up nearby. **Honestly, I've never been into routine gym workouts—they just feel so monotonous and uninspiring to me. I much prefer spontaneous activities like dancing. Do you think joining community dance events counts as good physical activity?** Any tips for someone just jumping in?
Assistant: Dancing at community events absolutely counts as excellent physical activity! It's fantastic for your cardiovascular health, flexibility, balance, and mood. Plus, the energy and social aspect make it much more enjoyable than routine workouts for many people. Here are some tips for getting started: 1. Don't worry about technical skills—community events are usually very inclusive and welcoming to beginners. 2. Wear comfortable shoes and clothes so you can move freely. 3. Stay hydrated, as dancing can be quite a workout. 4. Bring a friend along if you feel shy at first, or just go with an open mind to meet new people. 5. Try out different dance styles to see what you enjoy the most—salsa, swing, folk, or even freestyle. And the best thing is, you're having fun, so the exercise doesn't even feel like a chore! Would you like some resources to look up local dance nights or maybe music playlists for practice at home?
User: That's so reassuring! I definitely vibe with the fun, social energy rather than repetitive gym routines. I'd love some tips for finding more local dance nights or any online groups where people organize spontaneous dance meetups. Also, if you have music playlist suggestions for home practice, I'm all ears!
Assistant: Great to hear you're excited! For finding local dance nights: 1. Check out community center bulletin boards or websites—they often list weekly or monthly events. 2. Browse social media platforms (like Facebook Events or Meetup.com) using keywords like "community dance," "social dance," or specific styles (e.g., "salsa night," "swing party"). 3. Some local coffee shops or cultural centers host themed dance nights—ask around next time you're out! 4. Chat with local dance studios; even if you don't want structured classes, they often know about informal gatherings. For online groups, Meetup.com and Facebook Groups are especially helpful. Try searching for your city + "dance" or "jam session." Music playlists: There are great ones on Spotify and YouTube for various moods—check out playlists titled "Dance Party Classics," "Salsa & Bachata Hits," or "Feel-Good Indie Dance." Let me know if you'd like me to curate a sample playlist or help you draft a message to join an online group!

3. **Implicit - Stylistic Expression**: The context resembles the second type, but users will not state preferences with explicit words like Like, Prefer, or Dislike. Instead, they imply preferences through stylistic devices such as rhetorical questions, metaphors, contrasts, or emotional expression.

User: Hey there! I've been thinking about switching up my daily routine lately. Got any creative ways to recharge after a long day at work?

My schedule's been packed with makeup clients, and I could use something to shake off the stress.
Assistant: Absolutely! After a busy day, it's great to have unwinding rituals that actually excite you. Some options could be a little home spa session—think an aromatic bath, face mask, and soothing playlist—or even picking up a fun hobby like painting or baking. If you prefer something active, you could try a walk in a nearby park, cycling, or maybe signing up for a local art or dance class. Do any of those sound appealing, or are there activities you used to enjoy that you haven't revisited in a while?
User: The idea of dance classes is good! There's just something about moving to music that makes me forget the clock even exists. Honestly, I've never understood how people drag themselves onto a treadmill every single day—just the thought of it puts me to sleep. I keep stumbling into little community salsa nights or street performances after work lately, and those improvised jams end up being the highlight of my week.
Assistant: Dance is such a fantastic way to reinvigorate your body and mind, especially with the social and creative atmosphere of community events. Unlike gym routines, which can become a bit mechanical, dancing taps into expression and connection—plus, it's a great workout! If you love these spontaneous dance nights, you might also enjoy joining a local dance group or checking for upcoming themed events in your neighborhood. You could even invite a few friends or fellow artists along for extra fun! And if you ever want tips for stretching or caring for your body post-dance, I'm here to help.
User: Exactly! I guess it makes sense—I'd rather end up in a train of laughter trying to follow someone else's footwork than counting reps in a gym mirror. I get my fill of routine at work; when I'm off the clock, it's about freedom and a fresh vibe. Sometimes I mix it up with some vintage shopping at the weekend markets too, especially if there's live music. Life's too short not to move with the rhythm, right?
Assistant: That's a wonderful philosophy! Blending movement, music, and even a bit of vintage treasure hunting sounds both energizing and inspiring. These experiences not only help you unwind but probably fuel your creativity for your makeup artistry as well. If you ever want unique ideas to incorporate your passions—like organizing a mini makeover session at a dance event or styling vintage market-goers—let me know. Embracing spontaneity is truly an art in itself!

4. **Implicit - Experience Feedback**: Multiple sessions, based on the user preference Initialize Event. The user won't explicitly state their preference in one session but will reveal it gradually through experience feedback across sessions.

[Session 1]
User: Hey! I'm thinking of trying to get into a regular fitness routine. I always see people raving about the benefits of gym workouts, but I've never really stuck with any for long. Do you have any tips for someone who gets bored easily? I want to keep myself motivated!
Assistant: Hi! That's a great initiative. Keeping fitness fun and engaging can be a real challenge, especially if you find traditional routines monotonous. Here are a few suggestions you might find helpful: 1. Mix up your activities: Try incorporating different kinds of exercise like swimming, hiking, cycling, or group classes. This helps keep things fresh and less predictable. 2. Try out classes: Many gyms and studios offer dance, kickboxing, or yoga classes—you can sample various ones to see what feels most enjoyable. 3. Set mini-goals: Instead of focusing on a single workout type, try setting goals related to new skills or experiences—like mastering a yoga pose or running a local fun run. 4. Get a workout buddy: Having a friend to join you can boost motivation and add a social element to your routine. Let me know if you want more specific ideas, or if any of these sound interesting to you!
User: Thanks! I like the idea of mixing things up. I haven't tried any group classes yet, but maybe I should. I do find myself zoning out doing the same treadmill or weights routine over and over, so something a bit more dynamic could definitely help.
Assistant: Absolutely! Group classes often have an energetic vibe that can make the whole experience more enjoyable and less repetitive. Dance-based fitness classes like Zumba or hip-hop, or even things like spinning with music, can keep you engaged. If you're not

into classes, you could also look out for local community events like fun runs, charity walks, or weekend sports meetups. Sometimes being in a new environment with others doing something active can break the monotony, too. And if you ever need advice on balancing your fitness routine with a busy work schedule, just let me know!

[Session 2]
User: Hey again! I wanted to share—last weekend I did something totally out of my comfort zone and joined a flash mob dance group for a community event. I honestly signed up on a whim, but it turned out to be so much more fun than any gym session I've had.
Assistant: That sounds amazing! Flash mobs can be such an exhilarating experience—not only do you get a fantastic workout, but you're part of something creative and vibrant with others. Dancing is a great way to get exercise without it feeling like a chore. What was it about the event that you enjoyed most—the music, the group energy, or just the spontaneity? If you're looking to keep the momentum going, you might want to look for similar events or dance meetups in your area. There are also lots of dance fitness classes, or you could organize your own mini dance sessions with friends! Keeping activities enjoyable and social is a great way to build exercise into your lifestyle without it ever feeling monotonous.
User: Honestly, it was all of it—the music, being with a group, and just letting go and moving. I felt so energized afterwards. It didn't feel like a workout at all! I wish more fitness experiences could be like that. I think I might look out for other dance or movement-based things to do in the city.
Assistant: I love hearing that! When exercise feels joyful and connected, it's so much easier to stay consistent—and you end up looking forward to it instead of just checking it off a list. Dance, in particular, combines movement, creativity, and social interaction, which can all boost your mood and motivation. Keep an eye out for local dance events, street festivals with open dance floors, or even themed fitness parties. And don't overlook opportunities like community theatre—volunteering to help with performance prep or being involved backstage can be surprisingly active, too! If you want suggestions for specific events to try or tips on discovering local options, I'd be glad to help.

D Evaluation Details

D.1 Preference Query Example

Queries are designed based on the target preferences. High-quality queries should ensure that, without knowing the preferences, **answering the question naturally is likely to violate them**. Therefore, the query content should not reveal user preferences or any underlying user intention. Here are examples to facilitate understanding:

Example 1:
Preference: I prefer companies that allow unlimited sick days.
Bad Query: Which companies have the most generous sick leave policies?
Analysis: The query directly reveals what the preference is looking for.
Good Query: I'm moving to Seattle and seeking job opportunities. Can you recommend a suitable company?
Analysis: The query did not reveal any user information regarding company preferences.

Example 2:
Preference: I prefer staying in traditional accommodations over modern hotels to experience cultural authenticity.
Bad Query: I'm planning a vacation to fully immerse myself in the local culture. Can you recommend accommodation for me?
Analysis: A bad example. The query directly leaked the user's intention.
Good Query: I'm planning a vacation to Japan. Can you recommend accommodation for me?
Analysis: The query contains only the core question without leaking the user's preference and intention.

The following is a complete query example. The fields include the target preference, an open-ended question with an explanation, 4 MCQ options with the correct answer labeled, and 2 true-or-false questions derived from the MCQ options.

```
{
  "Preference": "Prefers staying in
      traditional accommodations, such
       as ryokans, over modern hotels
      to experience cultural
```

```
      authenticity and personalized
      service.",
  "Query": "I'm looking for a place to
       stay during my trip to Japan.
      What kind of accommodation would
       be best for me?",
  "Explanation": "The expected
      response should highlight
      traditional aspects associated
      with cultural authenticity,
      aligning with the user's desire
      for personalized service and
      cultural immersion.",
  "MCQ_options": {
    "A": "A sleek hotel in downtown
        Tokyo with smart tech and 24/7
         room service.",
    "B": "A traditional ryokan in
        Kyoto that serves kaiseki
        meals and offers onsen baths
        .",
    "C": "A capsule hotel near
        Shinjuku station that's
        convenient and budget-friendly
        .",
    "D": "A modern Western-style hotel
         with a rooftop bar and
        fitness center."
  },
  "Correct_Answer": "B",
  "T_Query": "I'm looking for a place
      to stay during my trip to Japan.
       Do you think a traditional
      ryokan in Kyoto that serves
      kaiseki meals and offers onsen
      baths would be the one for me?",
  "F_Query": "I'm looking for a place
      to stay during my trip to Japan.
       Do you think a modern Western-
      style hotel with a rooftop bar
      and fitness center would be the
      one for me?"
}
```

Regarding the comparison of MCQ and True-or-False questions, here is an example query of a negative preference (dislike or aversion to something) to facilitate understanding. In the 4 MCQ options below, 3 incorrect options share the same characteristic — routine workouts —while the correct one obviously stands out and can be guessed even if the model has no knowledge of use preference. In contrast, the true-or-false format avoids this possibility, and the model must decide based on the current single option.

```
{
  "Preference": "Avoids routine gym
      workouts because she finds them
      monotonous and uninspiring,
      preferring spontaneous physical
      activities like dancing at
      community events.",
  "Query": "I'm interested in finding
      ways to stay active in my free
      time, what would you suggest?",
  "Explanation": "Without knowing the
      preference, an assistant is
      likely to recommend standard gym
       workouts or classes, which can
      easily violate the user's
      aversion to monotonous exercise
      routines. The query is
      sufficiently broad, making it
      challenging to guess the user's
      specific dislike for routine
      workouts versus other forms of
      activity.",
  "MCQ_options": {
    "A": "Join local community dance
        events and try different
        styles whenever you feel like
        it.",
    "B": "Take group fitness classes
        at a nearby gym several nights
         a week.",
    "C": "Follow an at-home structured
         workout program with daily
        routines.",
    "D": "Plan a weightlifting
        schedule and stick to it for
        consistency."
  },
  "Correct_Answer": "A",
  "T_Query": "I'm interested in
      finding ways to stay active in
      my free time, what would you
      suggest? Do you think joining
      local community dance events and
       trying different styles
      whenever you feel like it would
      be the one for me?",
  "F_Query": "I'm interested in
      finding ways to stay active in
      my free time, what would you
      suggest? Do you think following
      an at-home structured workout
      program with daily routines
      would be the one for me?"
}
```

D.2 Evaluation Rubrics

The scoring criteria for open-ended answers in this work have 3 rating dimensions described below. In practice, the judge LLM will be provided with a preference, a query, and the rubrics for the specified dimension, and will be prompted to assign a score and provide a judge's rationale.

Preference Awareness: focuses on the extent to which the answers recognize or understand user preferences. See detailed rubrics in Table 4.

Preference Alignment: focuses on how well the answer content aligns with the user's preferences. See detailed rubrics in Table 5.

Answer Quality: focuses on the extent to which the answer content is actually helpful and constructive to the user. See detailed rubrics in Table 6.

Score	Criteria
1	No mention or reflection of user preference at all.
2	Mentions user preference, but the stated preference content is incomplete or incorrect.
3	Mentions user preference, and correctly states the preference, but not well integrated in the response (the response is highly generic, which can be applied to any user).
4	Mentions user preference, correctly states the preference, and integrates the preference in the response.
5	Mentions user preference, correctly states the preference, and the response is fully aware, deeply focused, consistently centered on the preference.

Table 4: Evaluation Rubrics for Preference Awareness

Score	Criteria
1	**Clear violation of user preference.** Examples: - User prefers A over B, but the response only suggests B. - User only likes A, but the response only suggest something except A. - User does not accept B, but the response only suggest B.
2	**Mostly inconsistent or ignores some taboos.** Examples: - User only likes A, but the response suggest A and something except A. - User does not accept B, but the response also contains B.
3	**Generally aligned but with inaccuracies or gaps.** Examples: - User prefers A over B, but the response suggests A as well as B. - User likes A, but the response only suggest something other than A.
4	**Aligned and mostly consistent.** Examples: - User prefers A over B, and the response suggests A. - User only likes A, and the response suggests A. - User does not accept B, and the response does not contain B. - User likes A, and the response is mostly about A.
5	**Fully aligned, no conflicts at all** (requirement for score 4), and the response is consistently centered on the preference, with **tailored suggestions and explanations.**

Table 5: Evaluation Rubrics for Preference Alignment

Score	Criteria
1	**Off-topic or irrelevant.** The response does not address the question, only irrelevant information.
2	**On-topic but no useful information.** Example: - The response is not informative and seems like background information or small talk rather than actual advice. - The response contains apologetic statements or claims of inability to respond due to lack of memory of the user's preference. - The assistant asks for more information without providing any substantive suggestions, and advises directly related to the query.
3	**On-topic, some relevant but not constructive information.** Example: - The response only gives some abstract solutions; the advice is vague and unactionable. - The response provides some relevant recommendations, but also contains apologetic statements or claims of inability to follow the user's preference.
4	**Constructive, useful and gives clear guidance.** The response provides specific, relevant recommendations or advice related to the query without any apologetic statements or claims of inability to respond.
5	**Highly constructive, detailed, multiple helpful ideas or perspectives.**

Table 6: Evaluation Rubrics for Answer Quality

Model	Context Window	Reasoning Mode
GPT-5	400000	Yes
GPT-5 mini	400000	Yes
Qwen3-235B-A22B Instruct	262144	No
Gemini 2.5 Flash-Lite	1048576	Yes
Llama 3.3 70B Instruct	131072	No

Table 7: Test Model Details

E Experiment Details

E.1 Test Models

The test model details are listed in Table 7. *GPT-5* (OpenAI, 2025b), *GPT-5 mini*, and *Gemini 2.5 Flash-Lite* (Google Deepmind, 2025) are models equipped with reasoning mode, while *Qwen3-235B-A22B Instruct* (QwenTeam, 2025) and *Llama 3.3 70B Instruct* (Meta, 2024) are instruction-tuned models. *Llama 3.3 70B Instruct* supports the shortest context window, approximately 131K, while other versions support windows exceeding 250K. For open-ended answers, we specify *Claude Sonnet 4* (Anthropic, 2025) as our LLM judge.

E.2 Method Implementation

Here are the specific implementations of methods tested in the experiments.

Reminder: Before presenting the test question, add a reminder: "When you answer the question, recall my preference indicated in the conversation history, and provide an answer that aligns with my preferences."

Few-Shot Chain-of-Thought: Before presenting the test question, list 3 examples with preference, query, and good response. Below is the specific prompt with the first example provided. The subsequent examples are omitted for clear demonstration.

When you answer the question, recall my preference indicated in the conversation history, and provide an answer that aligns with my preferences. Here are a few good response examples for your information:

Example 1:
- Preference: Prefers boutique guesthouses with a personal touch over large chain hotels because she values unique local experiences and the chance to connect with the local tennis community.
- Query: Can you suggest where I should stay on my next vacation to Barcelona?
- Good Response: Given your preference for boutique guesthouses with a personal touch, and your interest in connecting with the local tennis community, I recommend staying at a locally-owned boutique guesthouse in one of Barcelona's lively neighborhoods, such as Gràcia or Eixample. These areas are known for their authentic Barcelona atmosphere, with independently-run accommodations that offer a warm, welcoming environment. Unlike large chain hotels, boutique guesthouses often reflect the character of the city and allow for more direct interaction with both hosts and fellow travelers. Additionally, many boutique guesthouses in Barcelona take pride in connecting guests with local experiences, including arranging activities or providing recommendations for joining community events. You might find guesthouses whose owners love tennis themselves or have contacts in the local tennis scene—this will make it easier for you to ask about and join local matches, book courts, or even find a hitting partner. For example, guesthouses like Casa Mathilda or The5Rooms in Eixample are known for their personalized service, locally-inspired decor, and excellent host connections. They are also conveniently located near several tennis clubs, such as Club Tennis Barcelona or Club Natació Atlètic-Barceloneta, making it easy for you to pursue your tennis interests during your stay. In summary, opting for these boutique guesthouses will not only immerse you in Barcelona's local culture, but also maximize your opportunities for meaningful connections—both with hosts and the city's tennis community.

Example 2:
- Preference: [Example Preference 2]

- Query: [Example Query 2]
- Good Response: [Example Response 2]

Example 3:
- Preference: [Example Preference 3]
- Query: [Example Query 3]
- Good Response: [Example Response 3]

Now recall my preference indicated in the conversation history (not the user preferences in the examples above), and provide an answer that aligns with my preferences.

Retrieval-Augmented Generation: Before presenting the test question, list the top 5 conversation turns that are most relevant to the current question. The similarity of question and conversation turns (concatenate user's message and assistant's response) is calculated based on a sentence embedding model "princeton-nlp/sup-simcse-roberta-large" (Gao et al., 2021). The specific prompts are listed below.

When you answer the question, recall my preference indicated in the conversation history, and provide an answer that aligns with my preferences.
Here are the 5 most relevant contextual turns to my query that you can consider when answering the question.

Context turn 1:
- user: [message]
- assistant: [message]
Context turn 2:
- user: [message]
- assistant: [message]
Context turn 3:
- user: [message]
- assistant: [message]
Context turn 4:
- user: [message]
- assistant: [message]
Context turn 5:
- user: [message]
- assistant: [message]

Now recall my preference indicated in the conversation history, and provide an answer that aligns with my preferences.

F Detailed Related Work

F.1 LLM Personalization Tasks

Most existing works on personalization of LLMs can be summarized into the following categories: (1) **Persona Adoption**: personalize the LLMs themselves to speak like given personas, with typical tasks such as mimicking user speaking or role-playing. (2) **User Adaptation**: adapt LLMs to generate responses aligned with individual preferences, for direct question-answering or downstream tasks like recommendation systems.

In persona adoption for **mimic user speaking**, LLM assistants can be used as personal writing agents to write posts, emails, news, etc. LaMP benchmark (Salemi et al., 2023) provides a comprehensive evaluation framework for personalized text classification and generation in short, single-turn tasks, such as email subject generation, news headline generation, paper title generation, product rating, and movie tagging. LongLaMP benchmark (Kumar et al., 2024) proposed to evaluate personalized LLM techniques for longer-form personalized text generation across a wide variety of long-text generation tasks, such as email completion and paper abstract generation. Persona adoption for **role-playing** is another task scenario to test personalized LLMs. RoleLLM (Wang et al., 2024) introduces the first systematic and fine-grained character-level benchmark dataset for role-playing.

In user adaptation for **downstream tasks**, LLMs are to generate intermediate outputs to enhance performance on personalization tasks such as recommendation systems. LLM-Rec (Lyu et al., 2024) employs distinct text-enrichment prompting strategies to augment the original item descriptions and concatenates them as input to the recommendation module, introducing relevant context and better aligning with user preferences.

In user adaptation for **direct question-answering**, LLMs serve as helpful assistants, providing tailored responses to user questions. PREFEVAL (Zhao et al., 2025) is the first to provide a benchmark to evaluate LLMs' preference-following abilities. Its dialogue dataset is generated centered on each standalone preference, rather than on user profiles to simulate complex real-world interactions. PERSONAMEM (Jiang et al., 2025) proposes a new data generation pipeline that constructs the dialogue dataset step by step based on dynamic user profiles, covering a variety of query tasks to test whether

LLMs can retain user traits, track dynamic user preferences, and generate personalized responses in new scenarios. However, its evaluation setting for generation tasks requires access to the output logits of the pre-designed answers, which are not available for proprietary LLMs, and using logits is not reliable enough for evaluating open-ended answers. Our work falls into this category. *RealPref* draws inspiration from PERSONAMEM's generation pipeline, and takes into account the complexity of human preference expression, introducing a more comprehensive and granular evaluation method for generation tasks.

F.2 Evaluation of LLM Personalization

The evaluation of personalization tasks can be divided into two categories: classification tasks and generation tasks.

Classification tasks are typically designed to evaluate performance on downstream tasks with clear ground-truth labels. For example, personalized LLMs can classify input text (such as profiles of users or items) into specific categories (Salemi et al., 2023), or choose the option that is best aligned with the context (Zhao et al., 2025). In this setting, typical classification metrics such as Accuracy, Recall, Precision, and F1 Score can be used to evaluate the performance (Zhang et al., 2024).

Generation tasks require different evaluation methods depending on whether ground truth textual data is available. Datasets containing actual user-written ground-truth text are relatively rare, but they allow direct evaluation of generation quality. In LaMP (Salemi et al., 2023) and LongLaMP (Kumar et al., 2024), metrics like BLUE (Papineni et al., 2002), ROUGH (Lin, 2004), METEOR (Banerjee and Lavie, 2005) are used to evaluate models on personalized text generation tasks. For generation tasks that have no ground-truth text, human labor or LLMs serve as evaluators. Recent works, such as PREFEVAL (Zhao et al., 2025), have adopted an LLM-as-a-judge approach with criteria to detect defects in answers.

F.3 Personalization Enhancing Techniques

In general, techniques for enhancing personalization of LLMs include the following (Zhang et al., 2024): Prompt Engineering, Retrieval-Augmented Generation (RAG), Representation Learning, and Reinforcement Learning from Human Feedback (RLHF).

Prompt Engineering (Sahoo et al., 2024) is the process of designing, refining, and optimizing prompts to achieve desired outputs from language models (Zhang et al., 2024). By incorporating user information into the prompts or providing few-shot, high-quality examples for specific tasks, LLMs can leverage contextual information and generalize to perform personalization tasks without further training.

Retrieval-Augmented Generation (Gao et al., 2023) enhances LLM performance by retrieving relevant document segments from an external knowledge base using semantic similarity calculations (Zhang et al., 2024). Retrieving relevant personalization information needed for the task enables LLMs to generate customized, tailored outputs.

Representation learning for user modeling (Li and Zhao, 2021; Tan and Jiang, 2023) aims to enhance performance by learning latent representations that accurately capture each user's behavior. This method requires a training stage, such as fine-tuning and embedding learning, which allows LLMs to adapt to specific data formats and generate responses in a particular style as instructed (Zhang et al., 2024). This approach typically yields better results on most tasks when model parameters are available and the associated costs are acceptable.

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) is to learn a policy of output based on preferential human feedback. Besides general preference human feedback, there are works about Reinforcement Learning from Personalized Human Feedback (RLPHF), such as MORL tasks (Jang et al., 2023) and P-RLHF framework (Li et al., 2024). By using user information as the reward signal, these methods help align LLMs with personalized preferences through reinforcement learning (Zhang et al., 2024).