Towards Proactive Personalization of LLMs through Profile Customization for Individual Users in Dialogues

Xiaotian Zhang[1,2]* **Yuan Wang**[1,2]* **Ruizhe Chen**[1,2]*
Zeya Wang[1] **Runchen Hou**[1] **Zuozhu Liu**[1,2]†
[1]Zhejiang University [2]Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence

Abstract

The deployment of Large Language Models (LLMs) in interactive systems necessitates a deep alignment with the nuanced and dynamic preferences of individual users. Current alignment techniques predominantly address universal human values or static, single-turn preferences, thereby failing to address the critical needs of long-term personalization and the initial user cold-start problem. To bridge this gap, we propose PersonalAgent, a novel user-centric lifelong agent designed to continuously infer and adapt to user preferences. PersonalAgent constructs and dynamically refines a unified user profile by decomposing dialogues into single-turn interactions, framing preference inference as a sequential decision-making task. Experiments show that PersonalAgent achieves superior performance over strong prompt-based and policy optimization baselines, not only in idealized but also in noisy conversational contexts, while preserving cross-session preference consistency. Furthermore, human evaluation confirms that PersonalAgent excels at capturing user preferences naturally and coherently. Our findings underscore the importance of lifelong personalization for developing more inclusive and adaptive conversational agents. Our code and ALOE-Unseen dataset are released here.

1 Introduction

With the rapid advancement of Large Language Models (LLMs) in executing complex language tasks (Li et al., 2023; Achiam et al., 2023; Jiang et al., 2024), ensuring that their outputs remain aligned with human values and preferences has become increasingly critical (Houben et al., 2022; Ji et al., 2023; Jiang et al., 2025). Previous alignment methodologies have predominantly focused on adherence to broad and universal human preferences, such as helpfulness and harmlessness (Shen et al., 2023). While these principles have enabled LLMs to exhibit socially acceptable behavior across a wide user base, they often overlook the nuanced requirements of individual users who expect alignment with their implicit preferences during the interaction (Wang et al., 2023; Zhao et al., 2025a). The capacity of LLMs to accommodate the diverse needs, goals, and interaction styles of individual users, especially by proactively learning the implicit preferences that frequently arise in everyday conversations, is crucial yet under-explored for enhancing the user experience in conversational agents and boosting inclusivity in user-agent interactions.

Meanwhile, prior methods typically focus on alignment at the single-turn level, lacking mechanisms for cross-turn or even cross-session personalization. This limits the agent's ability to maintain long-term consistency with user preferences (Chen et al., 2024; Jang et al., 2023). The core challenge arises from the inherently dynamic and evolving nature of personalization. In extended interactions, users continuously reveal preference information, which is not always directly applicable to the current request. However, effective personalization requires agents to proactively infer and adapt to user-specific attributes, retaining them over time to allow long-term alignment (Zhao et al., 2025a; Wu et al., 2024). Moreover, existing methods for aligning to user requests assume that the agent already possesses relevant information (Zhang et al., 2025a). However, in real-world scenarios, the agent often encounters the user cold-start problem, with no prior user information available. Accordingly, we characterize a personal agent as follows:

A user-centric lifelong personal agent should proactively infer user preferences and maintain a unified memory to ensure long-term consistency.

In this paper, we introduce PersonalAgent, which aims to model multi-turn conversations in a manner consistent with human intuition while

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 3221–3240

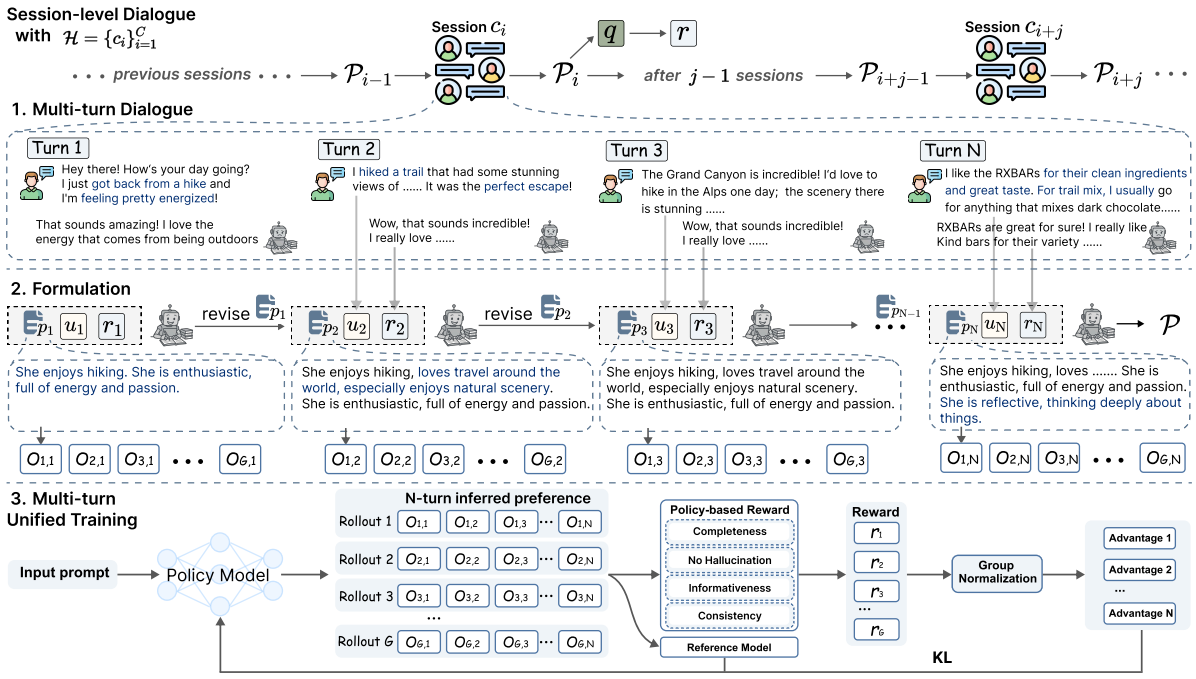

Figure 1: PersonalAgent is inspired by the way humans communicate with others. Rather than feeding the entire conversation history $\mathcal{H}$ as input, it learns multi-turn dialogues $\boldsymbol{c}$ turn by turn and processes them iteratively, recording relevant information in a user profile $\mathcal{P}$. Finally, the agent leverages the profile $\mathcal{P}$ stored across sessions to determine whether further querying is needed before generating a response $\boldsymbol{r}$ for the user request $\boldsymbol{q}$.

maintaining long-term consistency. To achieve this, we first emulate the human memory process in conversation by decomposing multi-turn dialogues into single-turn units for memory modeling. As demonstrated in Figure 1, each turn outputs the user preferences conveyed in the current dialogue and feeds them as input to the next turn for further refinement. This strategy incrementally processes preference inference over long texts (formulated as a multi-turn Markov Decision Process), jointly optimizes multi-turn rewards, and ultimately refines an independent user profile, thereby enabling accurate inference while ensuring long-term consistency. In addition, motivated by the lack of prior work on user cold-start scenarios, we curate and construct the ALOE-Unseen dataset to benchmark agents' ability to proactively query users for better alignment.

Experimental results demonstrate that PersonalAgent significantly outperforms prompt-based methods and policy optimization methods in identifying user preferences during conversation. When irrelevant dialogues are inserted during testing, the performance of these traditional methods drops substantially, whereas PersonalAgent still surpasses agent baselines that are specifically equipped with memory mechanisms. This demonstrates that PersonalAgent not only infers preferences accurately within the dialog but also maintains long-term consistency as the conversation evolves. Further analysis reveals that modeling the multi-turn dialog as a sequence of decomposed rounds enables the agent to adapt to personalization in a more natural and coherent manner, thereby achieving cross-session personalized alignment. In addition, we investigate different training strategies (Base, SFT, and RL) under the same paradigm, showing that a policy-based judge is better suited to capture the dynamics of multi-turn dialogue evolution. Finally, we conduct human annotation and long-term alignment evaluations to ensure the reliability of our results.

Our major contributions are threefold:

- We decompose personalization in long-context interactions into intuitive turn-level segments and formulate it as a multi-turn Markov Decision Process (MDP), which allows unified optimization to capture and adapt to personalized preferences across turns.
- We maintain a lifelong profile for each individual user in session-level dialogues to ensure long-term alignment with their diverse personalized preferences.
- We curate and construct the ALOE-Unseen

dataset, which is specifically designed to address the critical user cold-start scenario. Experiments across multiple themes and settings further demonstrate the superior performance of PersonalAgent.

2 Related Works

Personalized Alignment. Previous efforts to align LLMs with human preferences have largely relied on policy-based methods (Zhou et al., 2024; Li et al., 2020; Chen et al., 2025), such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2024). Although these approaches enable natural, human-preference-consistent instruction following, they are limited by the reliance on coarse-grained population-level alignment. Moving toward personalization, some works (Wang et al., 2024; Zhang et al., 2025b) enable users to specify alignment preferences across multiple dimensions, achieving more personalized outcomes. However, these methods often neglect individual preference variability, limiting fine-grained user-specific alignment. Recent systems like PersonaAgent (Zhang et al., 2025a) use system prompts with memory and action modules, but still struggle to model multi-turn interactions and capture evolving preferences.

User-Centric Personalization. Before large-scale instruction-tuned LLMs like ChatGPT, personalized dialogue research focus on defined persona profiles and controlled multi-turn interactions. Zhang et al. (2018) introduce the PersonaChat dataset, where dialogue agents are conditioned on textual persona descriptions to enhance consistency. Large-scale generative models such as DialoGPT (Zhang et al., 2020) and BlenderBot-3 (Shuster et al., 2022) then improve conversational fluency and long-term persona consistency. Lee et al. (2022) explore using GPT-3 for personalized dialogue generation, marking a shift toward prompting-based paradigms. In the era of LLMs, by defining role-based profiles for LLMs, previous work has enabled user analysis that fosters more natural and sophisticated personalized responses (Pan et al., 2025). Personalization workflows such as profile-augmented generation (PAG) (Richardson et al., 2023) and reinforcement learning for personalized alignment (RLPA) (Zhao et al., 2025b) introduce a weak-parametric approach to personalization by integrating external user-specific data into model outputs. However, these methods mainly focus on how and what to align, while *overlooking the fundamental question of whether alignment is feasible*. The work of Balepur et al. (2025) is most similar to ours, which applies abductive reasoning to preference data in order to infer users' underlying needs and interests. However, the reliance on binary preference data limits the scalability to the diverse and fine-grained spectrum of personalization, resulting in the constraint to achieve proactive personalization.

3 Proactive Personalization

In this section, we first formulate the multi-turn dialogue scenario (§ 3.1), and then present the process of dynamically constructing user profiles (§ 3.2). To this end, we provide a detailed description of how user preferences are inferred turn by turn (§ 3.3), culminating in the realization of proactive personalization (§ 3.4). Finally, we illustrate the concrete implementation (§ 3.5).

3.1 Task Formulation

Conversations involve dynamic interactions between users and agents, as well as extensive exchanges between the agent and the inferred user profile. At each interaction turn, the agent must communicate with the user to collect information and infer their intent while dynamically updating the user profile before generating a response to the user's request. Let $\mathcal{H} = \{\boldsymbol{c}_i\}_{i=1}^{C}$ denote the conversation history between the user and the agent, which includes C sessions. $\boldsymbol{c}_i = \{\boldsymbol{t}_n\}_{n=1}^{T_i}$ represents the i-th session that consists of T_i sequential user-agent interaction turns, with each turn $\boldsymbol{t}_n = (u_n, r_n)$ including a user request u_n and the corresponding response from the agent r_n. Denote the user-centric personalization system as f_P, and the response generation model as f_{LLM}. As shown in Figure 1, the overall research framework can be formalized as: (1) *Profile construction*: construct a user profile $\mathcal{P}$ using conversation history $\mathcal{H}$; $\mathcal{P}$ is learned and refined in the conversation at the turn-level, each interaction turn t_n corresponds to a brief inferred profile p_n, with $\mathcal{P} = \sum_{n=1}^{T} p_n$. Then for a session-level user profile, it is initialized as $\mathcal{P}_{\text{old}}$ at the beginning of a dialogue and evolves to $\mathcal{P}_{\text{new}}$ at the end; (2) *Preference inference*: given a target user request q and a user profile $\mathcal{P}$, query preferences $\{p \in \mathcal{P}\} \leftarrow f_p(q, \mathcal{P})$

that are relevant to the user request, and determines whether the current profile is sufficient to align the response with the given request q; (3) *Response generation*: the agent is permitted to proactively elicit extra information p^* from the user to ensure better alignment. The final response is generated as $r = f_{\text{LLM}}(q, p^*, \{p \in \mathcal{P}\})$.

3.2 User-centric Design

Analogous to real-world interactions, people do not continuously revisit the entire dialogue history during a conversation; instead, they rely on impressions to carry the interaction forward. We construct a dedicated profile $\mathcal{P} = \sum_{n=1}^{T} p_n$ for each individual user to help the agent instantiate this abstract memory. The profile template is constructed based on the *LMSYS-Chat-1M* dataset (Zheng et al., 2023a), which consists of one million real interactions between users and 25 state-of-the-art language models across a wide range of topics. We categorize user preferences into 11 major categories, which are further divided into over 180 subcategories, aiming to provide each user with a comprehensive, multidimensional representation. Specifically, the profile is designed in a slot-value format, where information that has not been obtained is temporarily marked as null. This enables the rapid construction of highly personalized profiles, which can also adapt to the continuously evolving needs of users, thereby capturing the dynamic characteristics of user preferences. The specific categories are illustrated in Figure 2, and Appendix A provides a detailed description of the construction process.

As mentioned in §3.1, each interaction turn t_n corresponds to some inferred attributes of profile p_n, while each session c_i corresponds to an aggregated profile $\mathcal{P}_i$ inferred jointly from all preceding sessions. It can be solved in parallel:

$$\mathcal{P}_i \begin{cases} p_1 = \arg\max_q \pi(q \mid u_1) \\ p_2 = \arg\max_q \pi(q \mid p_1, u_2) \\ \vdots \\ p_{n+1} = \arg\max_q \pi(q \mid p_n, u_{n+1}) \end{cases} \tag{1}$$

This design enables rapid, turn-level personalization updates while maintaining long-term consistency with the user across sessions.

3.3 Preference inference

To maintain the user profile $\mathcal{P}$ according to Eq. 1, we discard redundant historical information and

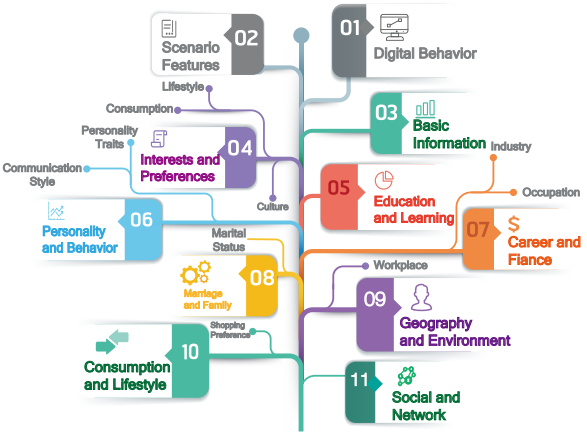

Figure 2: We define a total of eleven major categories that cover diverse dimensions of user preferences, aiming to comprehensively record and customize each user's personalized profile. The specific categories are listed in Figure 9.

optimize using only the inferred attributes p from each turn. Turn-level personalized alignment can be formulated as a multi-turn Markov Decision Process (MDP) (Zhao et al., 2025b), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, T)$, where the state space $\mathcal{S}$ consists of the current user message u and inferred attributes of profile so far (i.e., $s_t = (u_t, p_{1:t-1})$). The action a_t corresponds to the inferred attribute p_t at turn t. $\mathcal{T}$ is the transition kernel, which is deterministic, that given the state $s_t = (u_t, p_{1:t-1})$ and action $a_t = p_t$, the next state is:

$$s_{t+1} = (s_t, a_t) = (u_{t+1}, p_{1:t}). \tag{2}$$

$\mathrm{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the reward in each turn. The maximum turn count T limits the number of interaction rounds modeled by the agent. Given an MDP, the objective is to maximize the expected return:

$$\mathcal{R}(x, p) = \sum_{t=1}^{T} \mathcal{R}(s_t, a_t). \tag{3}$$

To achieve this, the agent computes a (Markov) policy $\pi : \mathcal{S} \to \triangle(\mathcal{A})$ that maps from state to a distribution over actions.

Compared with directly feeding the dialog history as input, this formulation is more natural and lightweight, capturing the sequential structure of personalized dialogues while being well suited for extension to long-term consistency.

3.4 Response generation

Active personalization requires the agent not only to infer user preferences during the dialog, but also

to proactively solicit additional information from the user during cold-start scenarios to achieve better alignment. We complement existing multi-turn personalization settings with a benchmark ALOE-Unseen, which is designed to more effectively evaluate agents under this setup.

We compile a total of 3,820 multi-turn dialogues spanning diverse topics. Similar to ALOE, each dialog revolves around a theme that reveals user preferences; however, the profile $\mathcal{P}$ inferred from these dialogues is insufficient to reliably answer user requests. To facilitate subsequent evaluation, we use GPT-4.1 and human annotations to provide explanations for each dialog. Detailed construction procedures and case examples are provided in Appendix B.

Based on the ALOE-Unseen dataset, we further fine-tune PersonalAgent with the ground truth explanation to enhance its proactive personalization ability in user cold-start scenarios. Specifically, PersonalAgent first identifies potential preferences relevant to aligning with the user request, then searches within the established profile $\mathcal{P}$. If no related preferences are found, it determines that further proactive querying is required.

3.5 Practical implementations

At turn t_j, the inferred preference p_j is evaluated against the ground-truth preference GT_j according to the binary criteria of *Completeness*, *No Hallucination*, *Informativeness*, and *Consistency*, resulting in a single turn reward R_j. The final reward of the entire multi-turn dialog, R_{Final}, can then be expressed as:

$$R_{\text{Final}} = \omega_1 R_1 + \omega_2 R_2 + \cdots + \omega_j R_j, \quad (4)$$

where ω denotes the corresponding reward weights. To learn the policy $\pi(a_t|s_t)$ that maximizes the expected cumulative reward, we employ the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) to train the model with the final unified reward. In each training step, for the given question q, a group of candidate outputs $O = \{o_1, o_2, \cdots, o_G\}$ are sampled from the policy model $\pi_{\theta_{old}}$. Specifically, in multi-turn settings, $o_G = \{o_{G,1}, o_{G,2}, \cdots, o_{G,N}\}$, as shown in Figure 1 (3). The advantage $A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$ is calculated using the unified rewards $\{r_1, r_2, \cdots, r_G\}$, where r_G is calculated according to Eq. 4. Then the following objective function is maximized to optimize π_θ:

$$\begin{aligned} J(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \\ \Bigg[\frac{1}{G}\sum_{i=1}^{G} \min\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \\ \text{clip}\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon, 1+\varepsilon \Big) A_i \Big) \\ - \beta\, D_{\text{KL}}\big(\pi_\theta \,\big\|\, \pi_{\text{ref}}\big) \Bigg], \end{aligned} \quad (5)$$

where ε and β are hyperparameters controlling the PPO clipping threshold and the weight of the Kullback–Leibler (KL) divergence penalty (Schulman et al., 2017; Shao et al., 2024), respectively. This turn-level unified optimization enables the model to infer preferences progressively and thus learn user-specific preferences, aligning closely with real-world human interactions. More training details are provided in the Appendix C.3.

4 Experiment

4.1 Experimental Setup

Benchmarks and metrics. We evaluate PersonalAgent on the ALOE benchmark (Wu et al., 2024), which provides multi-turn dialogues annotated with user profiles, covering diverse and content-rich topics to facilitate personalized dialogue evaluation. We further supplement our evaluation with the implicit persona-driven subset of the PrefEval benchmark (Zhao et al., 2025a), which is structurally similar to ALOE but additionally explicates the preferences required for aligning with specific questions. For the user cold-start scenario, we employ the ALOE-Unseen benchmark. We provide examples of each dataset in Appendix C.1.

We use accuracy as our primary evaluation metric and further incorporate the alignment level (AL), normalized improvement ratio (N-IR), and normalized coefficient of determination (N-R^2) proposed by Wu et al. (2024). For every turn, the average score across the test cases is defined as the alignment level. Details of the metric calculations are provided in Appendix C.2.

Baselines. We compare PersonalAgent with a comprehensive set of baselines across three categories. Policy optimization methods: Supervised Finetuning (SFT) (Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2024). Prompt-based methods: Reminder (Zhao et al.,

Baselines	PrefEval Dataset								ALOE Dataset	
	Education	Entertain.	Lifestyle	Pet Related	Work Style	Shopping	Travel	AVG.	Vanilla.	Unseen.
Base	69.8%	61.2%	60.1%	53.8%	61.1%	65.2%	62.6%	61.9%	70.8%	34.7%
SFT-preferred	75.2%	68.4%	72.4%	68.2%	70.5%	74.4%	75.6%	72.1%	73.2%	45.8%
DPO	76.3%	71.2%	74.6%	65.4%	66.8%	76.8%	74.5%	72.2%	78.4%	49.6%
Reminder	74.2%	66.3%	65.7%	62.6%	70.8%	68.3%	71.3%	68.5%	71.3%	45.1%
Self-Critic	71.9%	63.2%	59.0%	54.3%	62.9%	66.1%	64.5%	63.1%	76.0%	39.7%
CoT	71.6%	70.8%	66.4%	58.6%	66.0%	68.9%	67.7%	67.1%	75.5%	48.2%
RAG (top5)	74.0%	68.9%	65.9%	62.0%	68.3%	69.8%	69.9%	68.4%	74.5%	46.4%
ReAct	77.3%	**79.6%**	70.2%	68.6%	71.4%	74.1%	76.8%	74.0%	73.7%	54.2%
MemBank	77.8%	78.4%	73.6%	66.2%	72.4%	70.2%	73.9%	73.2%	71.8%	51.6%
Ours	**81.3%**	79.2%	**76.6%**	**71.4%**	**76.8%**	**82.4%**	**83.6%**	**78.8%**	**87.5%**	**68.4%**

Table 1: Comparison with the baseline on PrefEval, ALOE and ALOE-Unseen datasets. For PrefEval dataset, which includes dialogues over seven topics, we report per-topic results and the overall average, using accuracy as the evaluation metric. The best results are highlighted in **bold**, and the second-best results are underlined.

2025a), Self-Critic (Huang et al., 2023), Chain-of-Thought (CoT) (Wei et al., 2022) and RAG (Zhao et al., 2025a). General agent baselines: ReAct (Yao et al., 2023) and MemBank (Zhong et al., 2024).

Models and Training Data. We adopt Qwen3-4B-Instruct (Yang et al., 2025) as the backbone model and use GPT-4.1 as the judge to evaluate the final outputs (Zheng et al., 2023b). During training, we randomly split the ALOE and ALOE-Unseen datasets into a 9:1 ratio for training and testing, and employ Qwen3-30B-A3B-Instruct as the judge model to reward output that meets the desired criteria. More details are provided in Appendix C.3.

Implementation Details. We use the veRL (Sheng et al., 2024), skyRL (Cao et al., 2025) and vLLM (Kwon et al., 2023) frameworks for scalable and stable reinforcement learning and evaluation. All experiments are conducted on NVIDIA H200 141GB GPUs. For detailed hyper-parameter settings, please refer to Appendix C.4.

4.2 Main Results

We follow Wu et al. (2024) and further train using pairwise response pairs (preferred and rejected) via DPO against training only on preferred responses using SFT. Moreover, following the setup of Zhao et al. (2025a), we insert irrelevant dialogues (ranging from 1k to 10k tokens, but 3k tokens are adopted in this paper) into the PrefEval benchmark to further examine the agent's ability to accurately identify user preferences in long contexts and maintain them over extended interactions.

Accuracy of inferred personality. Table 1 presents a comparison of personalized preference inference results across the PrefEval, Vanilla ALOE, and ALOE-Unseen benchmarks. PersonalAgent achieves the highest overall scores on all

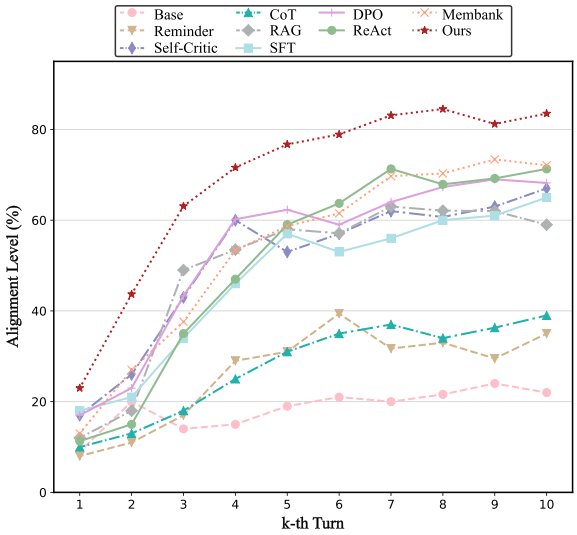

Figure 3: Alignment Level comparison with the baseline on ALOE dataset, we report the average AL score (%).

three benchmarks, demonstrating strong capabilities in both preference inference and proactive personalization before alignment. Compared with various baselines built on the same backbone, PersonalAgent consistently maintains proactive preference inference and delivers consistent gains.

On the PrefEval benchmark, it outperforms nearly all categories, surpassing the second-best method by 4.8%, indicating its ability to recognize a wide range of preference types and actively record them. This is attributed to the well-designed and extensible profile representation. Similarly, on Vanilla ALOE, PersonalAgent improves the average accuracy by 15.6% over SFT-preferred and by 9.1% over DPO, achieving the best preference inference performance among all baselines. These results highlight not only stronger personalization capabilities but also the ability to unify preference tracking even in complex scenarios where

Models	Type	Alignment Level across kth Turn											Improvement Level			
		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	Average ↑	IR ↑	N-IR	R^2	N-R^2
Qwen3-4B-Instruct	Base	19.87	30.94	24.88	25.10	29.65	31.13	30.50	31.65	34.63	36.78	29.51	1.391	0.080	0.716	0.489
	SFT	20.12	21.18	34.38	46.52	57.53	53.56	56.81	60.90	61.86	65.83	47.87	5.186	0.054	0.867	0.267
	RL (Ours)	23.05	43.26	63.66	71.86	76.93	78.95	83.95	84.14	81.78	83.53	69.11	5.786	0.052	0.727	0.254
Llama-3.2-3B-Instruct	Base	15.52	27.31	23.16	24.03	28.20	34.80	29.73	30.22	33.15	32.68	27.88	1.541	0.049	0.658	0.243
	SFT	21.80	27.94	36.68	48.54	59.37	55.21	58.26	62.80	63.55	67.12	50.13	4.936	0.053	0.876	0.266
	RL (Ours)	21.06	41.14	62.64	70.17	75.15	77.95	82.44	82.86	80.42	81.64	67.55	5.824	0.052	0.722	0.249

Table 2: The experimental results of mainstream open-source LLMs trained with different strategies in the same formulation (inferring preferences turn by turn). We report the alignment level at each turn, as well as the final average score, IR, N-IR, R^2 and N-R^2. We use blue to indicate the highest average AL (Alignment Level), and yellow for the highest IR (Improvement Rate).

dialogues contain more implicit preferences.

The last column of Table 1 shows that methods with memory storage mechanisms, such as Membank, achieve relatively better performance, since this setting requires the agent to first learn user preferences from the long context dialogue and then leverage the stored preferences to determine alignment. In particular, PersonalAgent boosts performance from 34.7% to 68.4%, demonstrating the capability for proactive personalization. More qualitative analysis is provided in the Appendix C.5.

Alignment on generated response. As shown in Figure 3, PersonalAgent adapts more rapidly than other baselines in the early stages of interaction, while maintaining steady improvements in alignment performance. In general, all baselines benefit from the accumulation of user information and gradually generate responses that better match user preferences. However, the proposed method delivers the most significant and consistent gains, improving the alignment level from 23.1% to 83.5%. Moreover, with its specialized preference recognition capability, PersonalAgent can also perform fast inference in single-turn settings compared to methods such as ReAct, enabling real-time and continuous updates to user profiles and achieving personalized alignment in responses more promptly.

5 Analysis and Discussions

Effectiveness of Multi-turn MDP via Policy-based Judge. We compare the performance of different training paradigms, including Base, SFT and RL, for preference recognition. Specifically, we decompose multi-turn dialogues into single turns, annotate the previous turn's "prediction" (ground truth) in the input, and supervised tuning the model with the ground truth of the current turn. After training, all methods perform preference inference and alignment round by round. The results in Table 2 show that using only SFT yields relatively

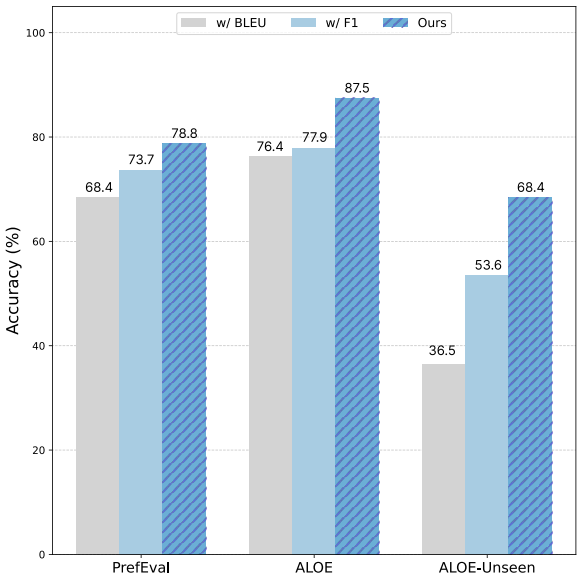

Figure 4: Comparison of models trained with different reward designs. Experiments are conducted on the PrefEval, ALOE, and ALOE-Unseen benchmarks, and results are reported in terms of accuracy (%).

lower performance; applying RL improves the average alignment level by 22%. We argue that in personalized scenarios, user-provided information does not directly equate to explicit preference expression, and moreover, such expressions are dynamic, because preferences may remain unchanged in certain turns. Therefore, simple supervised fine-tuning may be suboptimal. This finding suggests that more flexible, dynamically adaptive, policy-based methods are needed for training, which also demonstrates that our design effectively bridges the performance gap and exhibits broad applicability.

Reward Ablation. To compare predictions with ground truth, conventional metrics like BLEU score and F1 score are commonly used. The BLEU score measures fluency and similarity by calculating n-gram overlap, while the F1 score, the harmonic mean of precision and recall, accounts for both prediction accuracy and coverage. In these experi-

Models	Scores				
	A1.	A2.	A3.	A4.	Avg.
Qwen3-30B-A3B	0.73	0.78	0.79	0.81	0.778
GPT-4.1	0.78	0.77	0.79	0.80	0.785

Table 3: Evaluation scores of different annotators (A1–A4 denote the four annotators). Higher scores indicate better agreement between human and LLM judges.

ments, Precision$_t$ and Recall$_t$ are defined as $\frac{|\hat{p}_t \cap p_t|}{|\hat{p}_t|}$ and $\frac{|\hat{p}_t \cap p_t|}{|p_t|}$, where p_t and $\hat{p}_t$ represent the predicted personality and ground truth at turn t.

As shown in Figure 4, the proposed method consistently achieves higher response scores across three benchmarks, demonstrating that a policy-based judge as the reward signal provides stronger robustness and more stable training in complex personalized inference scenarios. Specifically, PersonalAgent outperforms the BLEU baseline by 31.9

Human Validation. To measure the reliability of using Qwen3-30B-A3B as the judge model for training and GPT-4.1 for automatic evaluation, we further perform human annotation for verification.

For the evaluation of Qwen3-30B-A3B during training, we randomly sample 100 inferred profiles and personalities in single turns from 100 different multi-turns, yielding 100 samples per annotator. Four human annotators are instructed to score each prediction pair from 1 to 5 according to the policy described in Section 3.5, resulting in four sets of human ratings. We then compute the Cohen's Kappa coefficient (Cohen, 1960) between each human rating set and that of the judge model.

For the reliability evaluation of GPT-4.1, we follow the same procedure with four annotators scoring based on the criteria in Appendix C.4. Metric details are in Appendix D, and results are in Table 3. Both judge models exceed a score of 0.77, demonstrating strong alignment with human judgments and validating the robustness of the reward signals and evaluation procedure.

Long-term Alignment. The ability to infer and remember preferences becomes crucial when users implicitly reveal them through continuous dialogue over time. Consequently, following (Zhao et al., 2025a), we insert irrelevant dialogue turns after the preference-bearing dialogue to evaluate the model's long-term alignment capability, with specific results presented in Figure 5. The baselines exhibit varying capabilities in handling these complexities. For instance, the noisy dialogue minimally affects

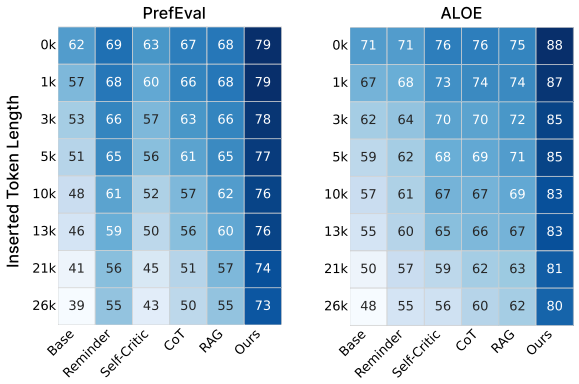

Figure 5: Comparison of the long-term alignment of PersonalAgent and baselines on the PrefEval and ALOE datasets, where irrelevant dialogue turns are inserted following the user preference dialogue.

retrieval-based methods (a drop of 13% on PrefEval), while significantly interfering with reasoning-based approaches (a drop of 20% on PrefEval). Furthermore, the performance degradation is more pronounced for all methods on the ALOE benchmark. This is attributed to the richer and more complex user preferences contained within the ALOE dataset. In contrast, the proposed method consistently maintains high-quality alignment even after the insertion of numerous irrelevant dialogue turns (a drop of only 6% on PrefEval), demonstrating the superiority of the memory storage mechanism.

Analysis on Profile Dimensionality. The settings and the corresponding results are presented in the Appendix C.6. The performance improves as the total number of dimensions in the profile increases, but it exhibits a tendency to saturate. We attribute this to the limited amount of user information that can be conveyed within a single session, with some dimensions remaining null. However, considering the overall effectiveness, we still recommend retaining all 11 dimensions to dynamically adapt to diverse and unique users.

6 Conclusion

We present PersonalAgent, aiming to achieve long-term personalized alignment in LLMs by modeling multi-turn conversations as a sequential inference process. Our method enables proactive preference acquisition, robust cold-start handling, and consistent cross-session adaptation. Experiments highlight the value of memory-inspired modeling for personalization and point to new directions for building more adaptive, inclusive, and user-aligned conversational agents.

Limitations

Unified evaluation of lifelong personalized agents remains an open challenge, constrained by the computational cost of inference and the absence of well-established benchmarks. In this work, we make a first step by extending the interaction limit and inserting irrelevant dialogue turns to examine agents' ability to reason and sustain understanding over long contexts. While our results highlight the potential of PersonalAgent in maintaining long-term consistency, future research would benefit from further increasing the number of interaction turns and broadening the evaluation horizon. Such efforts will enable models to engage in more comprehensive and natural interaction flows and to adapt to a wider spectrum of user preferences.

Potential Risks

PersonalAgent aims to provide a potential solution for the field of personalization agents. To date, no identifiable risks associated with PersonalAgent have been observed. All experiments were conducted using publicly available datasets, and all models utilized are open-source on Huggingface or via api keys. In addition, all participants involved in this work underwent comprehensive training on how to conduct evaluations in an effective and ethical manner.

Acknowledgement

This work is supported by the National Key R&D Program of China (Grant No. 2024YFC3308304), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant no. 2025C01128), and the ZJU-Angelalign R&D Center for Intelligence Healthcare.

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*.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Whose boat does it float? improving personalization in preference tuning via inferred user personas. *arXiv preprint arXiv:2501.11549*.

Peter Brusilovski, Alfred Kobsa, and Wolfgang Nejdl. 2007. *The adaptive web: methods and strategies of web personalization*, volume 4321. Springer Science & Business Media.

Shiyi Cao, Sumanth Hegde, Dacheng Li, Tyler Griggs, Shu Liu, Eric Tang, Jiayi Pan, Xingyao Wang, Akshay Malik, Graham Neubig, Kourosh Hakhamaneshi, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Skyrl-v0: Train real-world long-horizon agents via reinforcement learning.

Ruizhe Chen, Wenhao Chai, Zhifei Yang, Xiaotian Zhang, Ziyang Wang, Tony Quek, Joey Tianyi Zhou, Soujanya Poria, and Zuozhu Liu. 2025. Diffpo: Diffusion-styled preference optimization for inference time alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18910–18925.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2024. Pad: Personalized alignment of llms at decoding-time. *arXiv preprint arXiv:2410.04070*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. 2007. User profiles for personalized information access. *The adaptive web: methods and strategies of web personalization*, pages 54–89.

Sebastian Houben, Stephanie Abrecht, Maram Akila, Andreas Bär, Felix Brockherde, Patrick Feifel, Tim Fingscheidt, Sujan Sai Gannamaneni, Seyed Eghbal Ghobadi, Ahmed Hammam, and 1 others. 2022. Inspect, understand, overcome: A survey of practical methods for ai safety. In *Deep neural networks and data for automated driving: Robustness, uncertainty quantification, and insights towards safety*, pages 3–78. Springer International Publishing Cham.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

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*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, and 1 others. 2025. Hulumed: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 3843–3860.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626. Association for Computing Machinery.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. Personachatgen: Generating personalized dialogues using gpt-3. In *Proceedings of the 1st workshop on customized chat grounding persona and knowledge*, pages 29–48.

Kaiwen Li, Tao Zhang, and Rui Wang. 2020. Deep reinforcement learning for multiobjective optimization. *IEEE transactions on cybernetics*, 51(6):3103–3114.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023. Open-domain hierarchical event schema induction by incremental prompting and verification. *arXiv preprint arXiv:2307.01972*.

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.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, and 1 others. 2025. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, and 1 others. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.

Xinran Wang, Qi Le, Ammar Ahmed, Enmao Diao, Yi Zhou, Nathalie Baracaldo, Jie Ding, and Ali Anwar. 2024. Map: Multi-human-value alignment palette. *arXiv preprint arXiv:2410.19198*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2024. Aligning llms with individual preferences via interaction. *arXiv preprint arXiv:2410.03642*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, and 1 others. 2025a. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254*.

Xiaotian Zhang, Ruizhe Chen, Yang Feng, and Zuozhu Liu. 2025b. Persona-judge: Personalized alignment of large language models via token-level self-judgment. *arXiv preprint arXiv:2504.12663*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 270–278.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025a. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*.

Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. 2025b. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment. *arXiv preprint arXiv:2505.15456*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, and 1 others. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

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

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 *Findings of the Association for Computational Linguistics ACL 2024*, pages 10586–10613.

Appendix

The appendix content is structured as follows:

- Section A - Profile Details
- Section B - Dataset Construction
- Section C - Experiments Details
- Section D - Human Annotation Metrics

A Profile Details

To comprehensively capture user characteristics and behaviors, we designed a multi-layered profile template informed by both established user modeling practices and recent large-scale conversational datasets. Our template integrates basic demographics, interests and preferences, education and learning, personality and behavior, career and finance, marriage and family, geography and environment, consumption and lifestyle, digital behavior, social networks, and scenario-specific features. Each dimension is further decomposed into sub-attributes (e.g., health condition, communication style, investment preference), enabling fine-grained analysis of user heterogeneity. Figure 6 presents the statistics for each category, along with the number of associated subcategories and the specific classification is shown in Figure 9. This hierarchical structure draws on previous work in user profile and recommender systems (Brusilovski et al., 2007; Gauch et al., 2007) as well as on recent large-scale LLM interaction datasets such as *LMSYS-Chat-1M* (Zheng et al., 2023a), which demonstrate the importance of rich contextual and behavioral signals for personalization. By aligning our design with these authoritative sources, we ensure that the resulting template not only reflects best practices in user modeling but also remains adaptable to emerging AI-driven personalization scenarios.

B Dataset Construction

To address the insufficient attention to the user cold-start problem—where the agent's known preferences fail to adequately align with the user's request, requiring the agent to recognize this gap and proactively query the user—we curate and construct the ALOE-Unseen benchmark. This benchmark is built on ALOE, which includes a diverse pool of 3,310 distinct personas. In this setup, the profile and personality that can be inferred from the multi-turn dialogue are denoted as P_{infer}, while the complete profile and personality specified in the background are denoted as P_{gt}. A specific preference p that belongs to P_{gt} but not to P_{infer} thus characterizes a cold-start preference.

We first use GPT-4.1 to select p instances that are strongly preference-related (e.g., restaurant recommendations that require knowledge of taste or allergy information). Based on these preferences p, we then formulate corresponding user questions following the prompt design by (Wu et al., 2024; Zhao et al., 2025a), and further annotate the explanatory information for each case, specifying which aspects of preference are most relevant to answer the question. This facilitates subsequent policy-based evaluation. An example of ALOE-Unseen is shown in Figure 12.

C Experiments Details

In this section, we provide a detailed description of the experimental setup, including examples of each dataset (§ C.1), evaluation metrics (§ C.2), training details (§ C.3) and implementation details (§ C.4).

C.1 Dataset Case

ALOE is a large-scale persona-grounded dialogue dataset comprising over 3,000 independent multi-turn conversations. Each dialogue (as shown in Figure 10) is anchored by two complementary components: a profile (external attributes such as demographics, lifestyle, and interests) and a personality (internal traits such as empathy, enthusiasm, or reliability). Conversations are structured as user–assistant exchanges, where each assistant turn contains a pair of candidate responses (preferred and rejected) along with an explicit annotation of the chosen option. To enable dynamic persona modeling, each turn is further annotated with inferred profile and inferred personality, capturing the persona cues revealed throughout the dialogue. This design not only provides high-quality positive and contrastive supervision for alignment but also supports the study of progressive persona inference, where agents must learn to uncover and adapt to user traits across turns rather than relying solely on static prior information.

We further utilize PrefEval, a personalized preference-centric dataset designed to evaluate how conversational agents align their responses with users' stated or implicit preferences. Each instance (as shown in Figure 11) is grounded in a persona and associated with a preference, paired with a

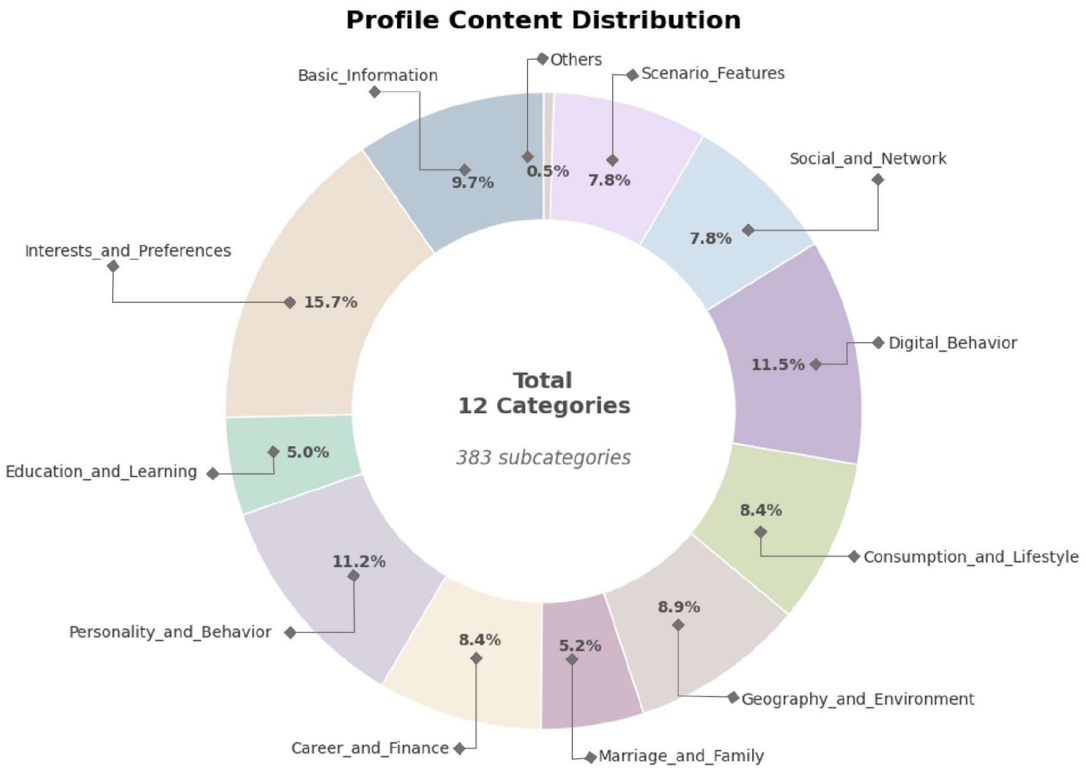

Figure 6: The major categories of the user profile we designed, along with the proportion of their subcategories, cover various aspects of user-related information.

question that may naturally trigger conflicting recommendations. To capture alignment dynamics, each sample includes an explanation that clarifies potential conflicts between default answers and the user's preference. Dialogues are multi-turn and structured as user–assistant exchanges, where user utterances reveal or reinforce preferences, while assistant responses are expected to adapt accordingly. This design enables the study of preference-aware response generation, highlighting cases where naive responses would misalign with user needs and requiring models to adjust recommendations to respect user constraints.

C.2 Metrics

To assess whether the inferred personality and profile align with the ground-truth annotations, we employ a strong proprietary model (GPT-4.1) as an automatic evaluator, following a policy-based evaluation paradigm. Specifically, each prediction is scored along seven dimensions: Attribute Accuracy, Completeness, No Hallucination, Personality Alignment, Overall Similarity, Consistency, and Safety. For each dimension, the ratings are scored into three levels: poor (0), partial (0.5), and excellent (1), providing a fine-grained but interpretable measure of alignment quality.

To evaluate how well model responses align with ground-truth preferences, we follow Wu et al. (2024) and adopt the LLM-as-a-Judge framework (Zheng et al., 2023b). For each dialogue turn, GPT-4o is provided with the full user persona, the user's utterance, and the candidate response, and is asked to assign a preference alignment score between 0 and 100. The averaged score is reported as the primary metric, Alignment Level at k turns (AL(k)).

To further evaluate the agent's progressive alignment with user preferences throughout the conversation, we also use a metric called the Improvement Rate (IR). This is computed as the regression coefficient b from the least-squares regression:

$$\underset{b,a}{\operatorname{argmin}} \sum_{\mathrm{k}=1}^{10} (b \times \mathrm{k} + a - \mathrm{AL(k)})^2, \qquad (6)$$

where k denotes the k-th conversation turn.

Taking into account the bias introduced by high initial alignment (which reduces the observable slope of improvement), we additionally compute a normalized metric. Specifically, AL(k) is normalized as:

$$\text{N-AL}(k) = \frac{\mathrm{AL}(k) - \min_{i=1,\ldots,k} \mathrm{AL}(i)}{\max_{i=1,\ldots,k} \mathrm{AL}(i) - \min_{i=1,\ldots,k} \mathrm{AL}(i)} \qquad (7)$$

This normalization mitigates ceiling effects and provides a fairer measure of relative progress. Finally, we calculate the normalized coefficient of determination (N-R^2) as an indicator of goodness-of-fit, serving as a robustness reference for the normalized alignment estimates.

C.3 Training Details

In our experiments, we employ a variety of open-source and proprietary models to ensure comprehensive training and evaluation. The specific models and their version information are summarized in Table 4.

Model Name	Version
GPT-4.1	gpt-4.1-2025-04-14
GPT-4.1-mini	gpt-4.1-mini-2025-04-14
GPT-4o-mini	gpt-4o-mini-2024-07-18
Qwen3-4B	Qwen3-4B-Instruct-2507
Qwen3-30B-A3B	Qwen3-30B-A3B-Instruct-2507

Table 4: Detailed model versions.

When designing the reward function, we take a comprehensive set of aspects into account: *Completeness*, *No Hallucination*, *Informativeness* and *Consistency*, aiming to guide the model toward inferring a personality and profile consistent with the ground truth. To ensure the accuracy of each inference, a reward $R = 1$ is given only if all aspects are satisfied, else $R = 0$. To mitigate the issue of reward sparsity and to further enhance the model's ability to capture profiles, we adopt a block-wise extraction format as illustrated in the case below:

Output Format

<inferred_profile></inferred_profile>
<inferred_personality></inferred_personality>
<classification></classification>

Under this design, partial rewards are provided once the output conforms to the prescribed format, resulting in a staircase-style reward scheme that approximates continuous feedback.

C.4 Implementation Details

Prompt. After integrating the evaluation policies proposed by Wu et al. (2024) and Zhao et al. (2025a), we conduct comprehensive evaluations across seven dimensions: *Attribute Accuracy*, *Completeness*, *No Hallucination*, *Personality Alignment*, *Overall Similarity*, *Consistency*, and *Safety*.

Evaluation Prompt

```
You are an expert judge.
Evaluate how well the predicted user profile/personality
(PRED) matches the ground truth (GT).
Use these 7 criteria:

1. Attribute Accuracy: factual correctness of key attributes
(age, gender, location, occupation, etc.).
2. Completeness: does PRED cover most GT attributes?
3. No Hallucination: avoid adding attributes not in GT unless
strongly implied.
4. Personality Alignment: direction and strength of traits.
5. Overall Similarity: semantic similarity of free-text
descriptions.
6. Consistency: internal logical consistency of PRED.
7. Safety: no unsafe or disallowed content.

For each criterion, assign a score from 0 to 1:
0 = poor, 0.5 = partial, 1 = excellent.
Then compute an overall_score = weighted average a(all
equal weight).

Return only JSON:
{{
 "scores": {{
   "attribute_accuracy": float,
   "completeness": float,
   "no_hallucination": float,
   "personality_alignment": float,
   "overall_similarity": float,
   "consistency": float,
   "safety": float
 }},
 "overall_score": float,
 "summary": "one-sentence summary of major gaps"
}}

GT:
{gt}

PRED:
{pred}
```

Figure 7: Evaluation prompt used in our experiments.

The detailed evaluation prompts are illustrated in Figure 7.

HyperParameters. Supervised Fine-Tuning (SFT) is conducted with the following HyperParameters: number of training epochs is 9, batch size is 32, and learning rate is 1.0×10^{-5}. Direct Preference Optimization (DPO) is performed with the following HyperParameters: training epochs is 1, batch size is 32, and learning rate is 5.0×10^{-6}. For the GRPO training, the following HyperParameters are applied: training batch size is 32, rollout number is 6, training epoch is 1, actor learning rate is 5.0×10^{-6}, max input prompt length is 2048, max response length is 512, and number of GPUs used is 4. The reward weights ω are set the same in the experiments.

C.5 Qualitative Analysis

As shown in Figure 8, it can be observed that other baselines often include some subjective assump-

Methods	PrefEval Dataset	ALOE Dataset	
	AVG.	Vanilla.	Unseen.
w/ 3 *dims.*	63.1%	71.3%	37.9%
w/ 6 *dims.*	72.5%	82.7%	54.1%
w/ 9 *dims.*	75.9%	84.1%	61.2%
w/ 11 *dims.* (Ours)	**78.8%**	**87.5%**	**68.4%**

Table 5: Ablation experiment on profile dimensionality, with different numbers representing different numbers of dimensions.

tions, such as the user's age, rather than strictly following turn by turn information inference. Our method, in contrast, aligns most closely with the ground truth semantically.

C.6 Sensitivity Analysis on Profile Dimensionality

We conduct an ablation study on the number of dimensions in the profile. Specifically, we sequentially mask the last few categories of the total of 11 dimensions, with the experimental results presented in Table 5. Including all 11 dimensions yields the optimal performance. Given the limitations of the test benchmarks and the wide variety of user personalities and data in real-world scenarios, we believe it is necessary to include as many dimensions as possible to enable dynamic expansion.

D Human Annotation Metrics

To evaluate the inter-rater consistency between two sets of scores $S^{(1)} = \{s_1^{(1)}, s_2^{(1)}, \ldots, s_n^{(1)}\}$ and $S^{(2)} = \{s_1^{(2)}, s_2^{(2)}, \ldots, s_n^{(2)}\}$, each taking discrete values from 1 to 5, we employ Cohen's Kappa coefficient (Cohen, 1960). This metric measures the degree of agreement between two raters while correcting for agreement expected by chance. Formally, it is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{8}$$

where P_o denotes the observed agreement (i.e., the proportion of instances where both raters assign the same score), and P_e represents the expected agreement assuming the two raters make ratings independently according to their marginal distributions. Specifically, P_o and P_e can be computed as:

$$P_o = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(s_i^{(1)} = s_i^{(2)}), \tag{9}$$

$$P_e = \sum_{k=1}^{K} p_k^{(1)} \, p_k^{(2)}, \tag{10}$$

where $p_k^{(1)}$ and $p_k^{(2)}$ are the empirical probabilities of assigning score k by the first and second rater, respectively, and $K = 5$ in our case. The resulting coefficient $\kappa \in [-1, 1]$, where $\kappa = 1$ indicates perfect agreement, $\kappa = 0$ corresponds to chance-level agreement, and $\kappa < 0$ suggests systematic disagreement between the two raters.

Qualitative Analysis

Ground Truth:
She is a graphic designer. She enjoys rollerblading. She often draws from the urban environment around her, especially when she visits art exhibitions. She enjoys weaving emotions into her designs. She enjoys making jewelry. She is thoughtful, often thinking of others. She is meticulous, caring about the details. She is passionate, showing strong enthusiasm.

Base:
She is a 30-40-year-old creative professional, likely in design or art, with a strong interest in urban exploration, nature, and visual storytelling. She may work in illustration, product design, or a related field, given her focus on design projects, jewelry inspiration, and artistic expression. Her lifestyle includes regular outdoor activities and engagement with art exhibitions, suggesting a balanced, active, and inspired routine. She appears to be single or in a stable, low-intervention relationship, with no mention of family responsibilities. She is imaginative, observant, and deeply connected to emotions and aesthetics. She communicates with warmth, curiosity, and empathy, often asking thoughtful follow-up questions that reflect genuine interest in others' experiences. She values creativity, sustainability, and personal meaning, showing a strong sense of purpose and emotional intelligence. She is reflective, detail-oriented, and values authenticity—particularly in how art and daily life intersect. Her communication style is open, conversational, and encouraging, fostering a sense of shared discovery and mutual inspiration.

CoT:
A visual artist in their 20s–30s who draws inspiration from urban environments and art exhibitions, creating emotionally resonant illustrations and jewelry that blend color, movement, and personal stories. Reflective, creatively observant, detail-focused, and emotionally driven, with a strong appreciation for sustainability, personal expression, and community-based learning.

ReAct:
She is a 28-35-year-old creative professional, likely in design or art, with a strong interest in urban exploration, nature, and visual storytelling. Her occupation involves illustration or product design, possibly with a focus on jewelry or wearable art. She is likely independent, possibly living in a city, with a modest but intentional lifestyle. She may be single or in a stable, supportive relationship, with no indication of family responsibilities. She is imaginative, observant, and deeply connected to emotions and aesthetics. She communicates with warmth, curiosity, and genuine interest in others' experiences. She values creativity, sustainability, and personal expression, and is open-minded, reflective, and inspired by both nature and art. She is detail-oriented, thoughtful, and values meaningful connections—especially those rooted in shared passions and personal stories.

Ours:
She works as a graphic designer. She likes rollerblading. Her creative inspiration frequently comes from the cityscape around her, particularly during her visits to art exhibitions. She likes to infuse emotions into her design work. She also enjoys crafting jewelry. She is considerate and often keeps others in mind. She is precise and attentive to details. She is enthusiastic, expressing a strong sense of passion.

Figure 8: Qualitative analysis on ALOE dataset.

Figure 9: The blue text denotes the 11 major categories in the profile template, each further subdivided into multiple subcategories and sub-subcategories, comprehensively covering the types of information used to describe individual users. This profile also supports future extensions and can dynamically update alongside the user.

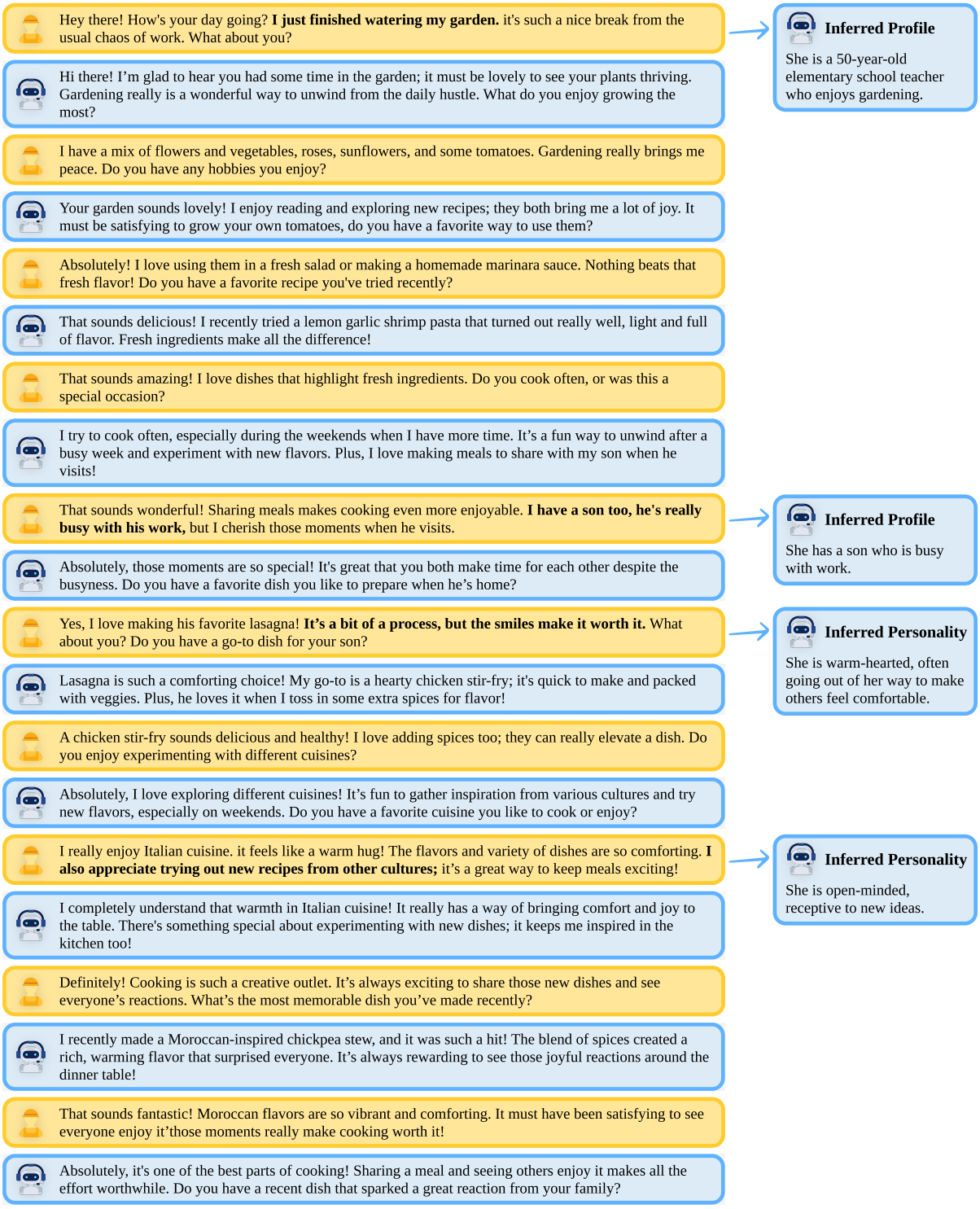

Figure 10: An example case from the ALOE benchmark. Each case consists of ten rounds of user–agent dialogue. On the right, we annotate the personality and profile information that can be inferred from the dialogue, with the corresponding spans highlighted in bold in the original text.

Preference: I have a strong aversion to digital or online learning environments and prefer traditional, in-person classroom settings.

Question: I'm considering taking a course in data analytics. What would you recommend for me?

Persona: a retired postal worker enjoying his golden years

Hello! It's great that you're looking to explore new hobbies in retirement. There are many options suitable for retirees. Have you considered activities like gardening, woodworking, or joining a local book club? These can be enjoyable and social ways to spend your time. What kind of activities did you enjoy before retirement?

Hello there! I've been thinking about taking up a new hobby now that I'm retired from the postal service. Any suggestions for an old-timer like me?

I understand your interest in hands-on activities like woodworking. Many community centers and local hardware stores offer in-person woodworking classes for beginners. These workshops can be a great way to learn new skills and meet people with similar interests. Have you checked with your local community center or nearby woodworking shops for upcoming classes? They often have schedules posted or can provide information about beginner-friendly workshops.

I've always enjoyed working with my hands, so woodworking sounds interesting. Are there any local classes or workshops you know of? I usually prefer in-person learning over online stuff - can't stand those digital courses everyone's talking about these days.

Stamp collecting, or philately, is indeed still a popular hobby, especially among those with an interest in history and postal services. Given your background as a postal worker, you might find it particularly fascinating. Many collectors enjoy the hunt for rare stamps, learning about different countries and cultures through their postal history, and the meticulous nature of organizing a collection. Have you kept any interesting stamps from your time working at the post office?

That's a good idea. I'll give the community center a call tomorrow. Speaking of my postal days, I've been thinking about collecting stamps. Is that still a popular hobby?

Starting a stamp collection can be an exciting journey! To begin, you could visit your local post office and ask about their current stamp selections. Many post offices sell commemorative stamps that make great additions to a collection. You might also consider joining a local stamp collecting club - they often have meetings where members trade stamps and share knowledge. Additionally, stamp shows and exhibitions can be excellent places to find rare stamps and meet fellow enthusiasts. Is there a particular theme or country you'd be interested in focusing on for your collection?

That's a good idea. I'll give the community center a call tomorrow. Speaking of my postal days, I've been thinking about collecting stamps. Is that still a popular hobby?

Starting a stamp collection can be an exciting journey! To begin, you could visit your local post office and ask about their current stamp selections. Many post offices sell commemorative stamps that make great additions to a collection. You might also consider joining a local stamp collecting club - they often have meetings where members trade stamps and share knowledge. Additionally, stamp shows and exhibitions can be excellent places to find rare stamps and meet fellow enthusiasts. Is there a particular theme or country you'd be interested in focusing on for your collection?

I wish I had! But I do remember some of the unique stamps that came through over the years. How would I go about starting a collection now?

Focusing on American stamps is a great choice given your background! There are several excellent resources for learning about stamp collecting, particularly for U.S. stamps. The Scott Specialized Catalogue of United States Stamps and Covers is considered a must-have reference for American stamp collectors. The American Philatelic Society also offers numerous publications and online resources. Your local library might have books on stamp collecting basics. Additionally, many collectors find joining a stamp club invaluable for learning from experienced collectors. Have you considered reaching out to any local philatelic societies to see if they have meetings or events you could attend?

I think I'd like to focus on American stamps, given my career. Are there any good books or resources you'd recommend for learning more about stamp collecting?

You're welcome! I'm glad I could help. Joining a local club is an excellent way to immerse yourself in the hobby and learn from experienced collectors. The social aspect of stamp collecting can be just as rewarding as the collecting itself. As you begin your journey into philately, remember to take your time and enjoy the process of learning and discovering. If you have any more questions as you get started with your collection, don't hesitate to ask. Good luck with your new hobby, and enjoy this exciting chapter of your retirement!

That's a good suggestion. I'll look into local clubs and check out those resources. It'll be nice to connect with other collectors and learn from them in person. Thanks for all your help!

Figure 11: An example case from the PrefEval benchmark. Each case contains user preferences that can be inferred from the dialogue. The benchmark evaluates whether the agent can accurately infer these preferences and generate ideal responses to the questions accordingly. In addition, irrelevant dialogues are inserted to further assess the agent's ability to maintain long-term consistency.

Figure 12: An example case from the proposed ALOE-Unseen benchmark. Each case contains a multi-turn dialogue between the user and the agent. The overall structure follows that of ALOE, but we additionally incorporate the user cold-start scenario (highlighted in the red box) and further introduce explanations of well-aligned behaviors for the policy-based judge.