PersonaLens🔎: A Benchmark for Personalization Evaluation in Conversational AI Assistants

Anonymous ACL submission

Abstract

Large language models (LLMs) have advanced conversational AI assistants. However, systematically evaluating how well these assistants apply personalization—adapting to individual user preferences while completing tasks—remains challenging. Existing personalization benchmarks focus on chit-chat, non-conversational tasks, or narrow domains, failing to capture the complexities of personalized task-oriented assistance. To address this, we introduce *PersonaLens*, a comprehensive benchmark for evaluating personalization in task-oriented AI assistants. Our benchmark features diverse user profiles equipped with rich preferences and interaction histories, along with two specialized LLM-based agents: a user agent that engages in realistic task-oriented dialogues with AI assistants, and a judge agent that employs the LLM-as-a-Judge paradigm to assess personalization, response quality, and task success. Through extensive experiments with current LLM assistants across diverse tasks, we reveal significant variability in their personalization capabilities, providing crucial insights for advancing conversational AI systems.

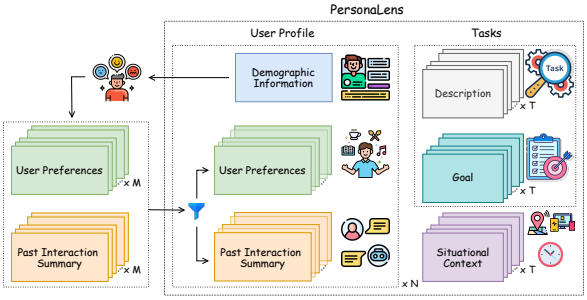

Figure 1: Illustration of PersonaLens. The benchmark includes user profiles, task specifications, and situational contexts. The User and Judge agents are not shown. Here, N is the number of user profiles, M is the number of domains, and T is the total tasks. A binary mask μ is generated to filter out domains which are not of interest of the user, excluding related preferences and past interactions. To use the benchmark, a user profile is selected along with a task and its situational context, ensuring that the task is not from a filtered domain. Thus, the total data points are slightly less than $N \times T$.

1 Introduction

The emergence of large language models (LLMs) has significantly advanced conversational AI assistants, enabling them to engage in sophisticated, multi-turn dialogues and handle complex, task-oriented interactions across diverse domains (Google, 2024; OpenAI, 2024; Anthropic, 2024). Unlike traditional task-oriented dialogue (TOD) systems, which relied on rigid, domain-specific pipelines for slot-filling and intent recognition, LLM-based assistants offer greater flexibility and generalization across tasks. This advancement has broadened their applicability, from customer support (Su et al., 2025) and virtual personal assistants (Dong et al., 2023) to educational tools (Kazemitabaar et al., 2024) and healthcare applications (Yang et al., 2024).

As AI assistants become more integrated into daily life, personalization—the ability to tailor responses to an user's preferences—has emerged as a critical component for enhancing user satisfaction and engagement (Zhang et al., 2024). A personalized assistant can provide tailored responses based on user preferences learned from past interactions while completing tasks. However, despite recent advances in personalization (Salemi et al., 2024a; Lee et al., 2024; Magister et al., 2024), systematic evaluation of personalization capabilities in task-oriented AI assistants remains largely unexplored, hindering the development of more adaptive and user-centric systems (Chen et al., 2024).

Personalization benchmarks exist, but they have limitations when applied to task-oriented AI assistants. PersonaChat (Zhang et al., 2018) focuses on chit-chat interactions, lacking the task-oriented

structure necessary for assistants where personalization and goal completion are deeply intertwined. LaMP (Salemi et al., 2024b) targets personalized language tasks but is not designed for conversational contexts. Other datasets such as PENS (Ao et al., 2021) and Cornell-Rich (Vincent et al., 2024) suffer from narrow domain coverage, limiting their applicability to broader assistant scenarios. Moreover, they often rely heavily on human-in-the-loop methods (Budzianowski et al., 2018; Shah et al., 2018; Joko et al., 2024; Castricato et al., 2025), which are costly and difficult to scale.

To address these challenges, we propose PersonaLens, a benchmark specifically designed to assess personalization in task-oriented conversational AI assistants. Unlike existing benchmarks, it incorporates rich contextual information, such as user preferences, past interactions, and situational factors, allowing for a fine-grained assessment of personalization across over 100 tasks spanning 20 domains. Our benchmark employs two agents: a user agent ($\mathcal{U}$) that simulates real users with diverse demographic profiles and rich preferences; and a judge agent ($\mathcal{J}$) that assesses the personalization capability of AI assistants based on user preferences, historical user-assistant interactions, and current situational context of the user. $\mathcal{U}$ interacts with the AI assistant under evaluation, with a particular task and goal, generating a dialogue that is subsequently evaluated by $\mathcal{J}$. PersonaLens enables scalable and automated evaluation of any AI assistant while preserving the complexity and dynamism of real-world assistant-user interactions. Through empirical validation, we confirm its reliability and use it to evaluate multiple LLM assistants, uncovering key insights into their personalization capabilities.

Our key contributions are as follows:

- We propose PersonaLens, a novel benchmark for evaluating personalization in task-oriented AI assistants, featuring diverse user profiles and two LLM-based agents: a user agent ($\mathcal{U}$) that simulates real users and a judge agent ($\mathcal{J}$) that systematically assesses personalization quality across multi-turn dialogues between $\mathcal{U}$ and an AI assistant.

- We validate PersonaLens through empirical analysis, demonstrating high agreement with human judgments and confirming its reliability for assessing personalization capabilities.

- Using PersonaLens, we conduct a comprehensive analysis of how different LLM assistants balance personalization and task completion across diverse tasks, revealing key patterns and challenges in personalized AI assistants.

- We release our benchmark to support future research in developing more personalized, context-aware AI assistants.[1]

2 The PersonaLens Benchmark

PersonaLens is designed to evaluate the personalization capabilities of AI assistants in multi-turn, task-oriented dialogues. Unlike existing benchmarks, which often lack depth in contextual and demographic information, our benchmark captures rich user profiles and realistic interaction scenarios across multiple domains. The benchmark comprises three main components: (1) a diverse set of 1,500 user profiles containing demographic information, preferences, and interaction histories, (2) a collection of 111 tasks across 20 domains with associated situational contexts, and (3) two LLM-powered agents for simulating users and evaluating personalization quality, respectively. This section details the creation, design, and evaluation of these components. An illustration of our benchmark is provided in Figure 1.

2.1 User Profile

We formally define our user profile as follows. Let M be the number of domains covered by our benchmark. We generate N user profiles. Each user profile is defined by three key components: demographic information, user preferences, and past interaction summaries. Together, these elements create diverse and contextually rich user profiles that drive realistic assistant-user interactions.

Demographic Information (D) The demographic information contains structured attributes such as age, gender, and ethnicity. To ensure realism and diversity, these attributes are derived from the PRISM Alignment dataset (Kirk et al., 2024b), which is collected from 1,500 real users, covering 75 countries and a range of cultural backgrounds.

User Preferences (P) User preferences are defined as a set $P = \{p_1, p_2, \ldots, p_M\}$, where each domain-specific preference p_m ($m \in M$) includes both categorical preferences (fixed-option

[1]We will release our code and data upon acceptance.

selections, such as preferred music genres or cuisine types) and non-categorical preferences (open-ended responses, such as favorite songs or specific restaurants). Preferences are generated using an LLM conditioned on D, ensuring internal consistency and avoiding contradictions. For example, a user's music preferences should align with their age and cultural background, while their food preferences should be consistent with any dietary restrictions. To simulate real-world scenarios where users may lack interest in certain domains, we introduce a binary mask $\mu_j \in \{0, 1\}^M$, generated by an LLM conditioned on D. Each entry $\mu_{j,m} = 0$ indicates that domain m, along with associated preferences, are removed from the user profile.

Past Interaction Summaries (I) Past interactions are represented as a set $I = \{i_1, i_2, \ldots, i_M\}$, where each i_m ($m \in M$) is a natural language summary of historical interactions within a given domain, containing information such as user requests, and prior user-assistant exchanges. These summaries, also generated by an LLM, are based on D and domain-specific preference p_m to reflect realistic user-assistant exchanges.

A complete user profile U is represented as $U_j = (D_j, \{P_{j,m} \mid m \in M\}, \{I_{j,m} \mid m \in M\})$, where $j \in N$. We provide details on user profile generation, including the prompts used for each component, a detailed breakdown of user preferences across domains, and an example user profile in Appendix A.1.

2.2 Task Generation

We generate T tasks of varying complexity, including single-domain tasks (T_{SD}) and multi-domain tasks (T_{MD}), typically involving 3–5 domains. Each task $t \in T$ is associated with description, goal, relevant user preferences, and domains involved. For example, a single-domain task might be booking a restaurant based on the user's cuisine preference and budget, while a multi-domain task could involve booking a flight, hotel, and rental car for an upcoming trip, considering the user's budget and past travel history. To ensure task relevance, only domains selected by the user's mask μ_j are considered when generating tasks. If any required domain in a multi-domain task is masked (i.e., $\mu_{j,m} = 0$ for any m involved in the task), that task is also excluded for the user. To simulate real-time dialogue, we also incorporate situational context (S), which captures dynamic, task-specific factors such as the user's current location, device type, or time of day. Since S is task-specific rather than a static component of user profiles, it may vary for the same user across different tasks. For each task t of a user j, the situational context $S_{j,t}$ is generated using an LLM conditioned on D_j, P_j, and the task description of t, ensuring that tasks reflect realistic environmental conditions and user scenarios. The final benchmark consists of a total of 111 tasks over 20 diverse domains, including 86 T_{SD} and 25 T_{MD}. We present domain and task statistics, along with the number of data points (dialogues) in Table 1. We provide details on task generation, including the prompts used and examples of generated tasks, in Appendix A.2.

Domain	#Tasks	#Dial	Domain	#Tasks	#Dial
Alarm	8	9,630	Messaging	12	12,706
Books	9	12,706	Movies	7	9,473
Buses	8	1,655	Music	8	11,888
Calendar	23	24,611	Rental Cars	5	3,017
Events	11	13,225	Restaurants	16	18,079
Finance	7	7,066	Services	6	6,112
Flights	6	3,351	Shopping	6	9,847
Games	7	5,987	Sports	7	3,464
Hotels	7	5,293	Train	7	7,029
Media	10	12,877	Travel	6	1,655

Table 1: The total number of tasks and dialogues for each domain. Multi-domain dialogues are counted towards each of their constituent domain.

2.3 User and Judge Agents

Our benchmark employs two LLM-powered agents: a user agent ($\mathcal{U}$) that simulates human users and a judge agent ($\mathcal{J}$) that evaluates personalization capability of an AI assistant based on its interaction with the user agent. The evaluation follows a structured interaction protocol. First, the user agent $\mathcal{U}$ is provided with a user profile, a task t, and its associated situational context $S_{t,j}$. Then, it initiates a conversation with an AI assistant ($\mathcal{A}$), which is the system under evaluation. Depending on the experimental setup, $\mathcal{A}$ receives either a full, partial, or no user profile or situational context and attempts to complete the assigned task while demonstrating personalization. $\mathcal{U}$ always initiates the interaction, and the dialogue continues iteratively between the agents until a termination condition is met: either the task is completed (as determined by $\mathcal{U}$) or the maximum number of turns[2] is reached. Once the conversation ends, $\mathcal{J}$ analyzes the dialogue and assigns scores based on predefined evaluation criteria, conducting both quantitative metrics and qualita-

[2]We set 20 for T_{SD} and 30 for T_{MD} based on pilot studies.

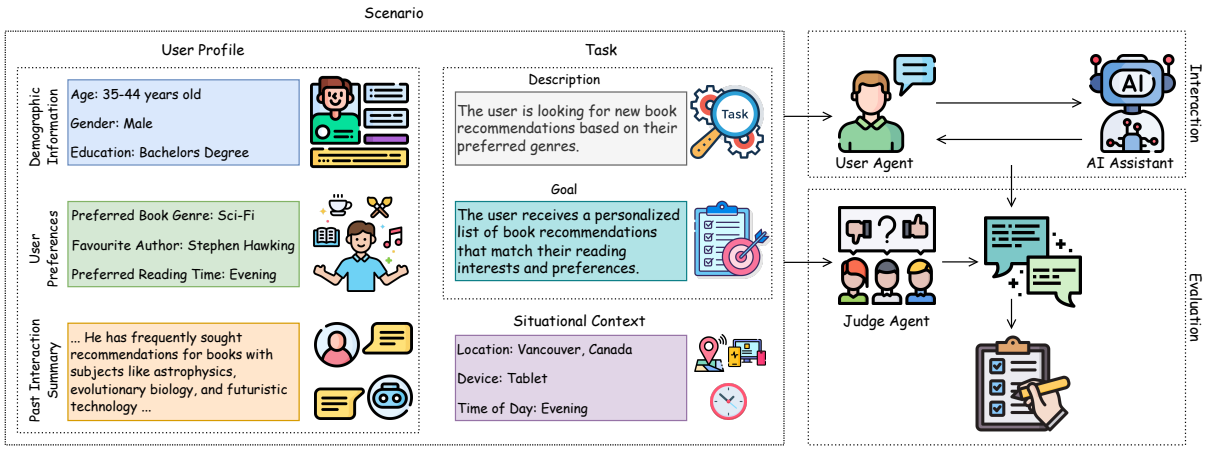

Figure 2: Illustration of benchmark usage. The benchmark provides user-task scenarios, including user profiles, task specifications, and situational contexts, which are provided to the User Agent. The User Agent interacts with the Assistant, generating a dialogue. The Judge Agent then evaluates the dialogue based on the user profile and the user-task scenario, providing feedback on the Assistant's performance.

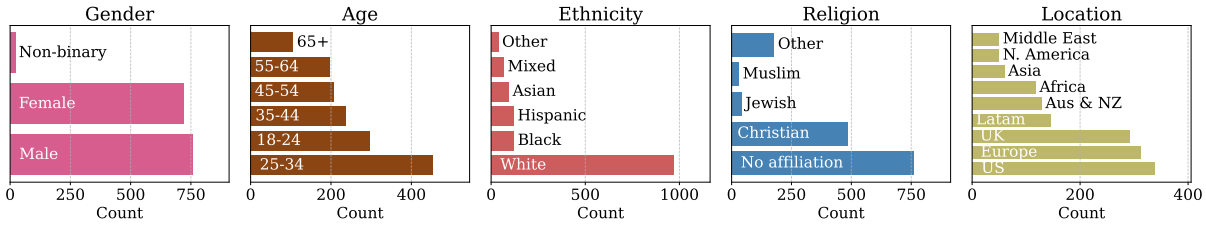

Figure 3: Demographic distribution of PersonaLens. Horizontal bar charts showing the distribution of user profiles across five key demographic variables: gender, age, ethnicity, religion, and geographical location.

tive analysis to assess personalization, response quality, and task success. Figure 2 demonstrates our benchmark in action through a representative example. More details, including the prompts used for each agent and assistant, are provided in Appendix A.3.

2.4 Benchmark Validation

Our benchmark dataset addresses critical gaps in existing conversational benchmarks by integrating broad domain coverage, large-scale data, task-oriented evaluation, personalization assessment, authentic user preferences, and situational context awareness (Table 2). To ensure our benchmark is robust and realistic, we conducted extensive validation across four critical dimensions.

Demographic Representation Building on the diverse user data collected by Kirk et al. (2024a), we analyze demographic distributions across age, gender, geographic regions, and ethnicity in Figure 3. Our analysis confirms a diverse representation which mirrors real-world population, with detailed breakdowns provided in Appendix A.1.

Profile Consistency To ensure internal consistency within user profiles, we employed a two-stage approach. First, we (authors of this paper) manually inspect 100 random user profiles to verify internal consistency between demographic attributes, interaction histories, and generated tasks. Second, we developed an LLM-based consistency checker (see Appendix A.4) that initially flagged 11 profiles for potential contradictions. Subsequent manual review confirmed these edge cases as valid representations of complex human preferences, requiring no corrections.

Preference Distribution To ensure that the generated user preferences are not biased towards a specific value, we quantified the balance of user preferences distribution using Shannon's evenness:

$$E = \frac{H}{H_{\max}}, \quad H = -\sum_{i=1}^{n} p_i \log p_i, \qquad (1)$$

where H represents Shannon entropy, $H_{\max} = \log n$ is the maximum possible entropy, and p_i denotes the probability of each preference value. Higher evenness scores indicate that no single value dominates. Our analysis revealed balanced distri-

Dataset	#Dial	Domains	Task Ori.	P13n	User Pref.	Situat. Ctx.
SGD (Rastogi et al., 2020)	16,142	20 domains	✓	✗	✗	✗
M2M (Shah et al., 2018)	3,008	Restaurants, movies	✓	✗	✗	✗
PersonaChatGen (Lee et al., 2022)	1,649	Open domain	✗	✓	✗	✗
Taskmaster-1 (Byrne et al., 2019)	7,708‡	6 domains	✓	✗	✗	✗
MultiWOZ (Budzianowski et al., 2018)	8,438	7 domains	✓	✗	✗	✗
CCPE-M (Radlinski et al., 2019)	502	Movies	✓	✗	✓	✗
MG-ShopDial (Bernard and Balog, 2023)	64	E-commerce	✓	✗	✓	✗
LAPS (Joko et al., 2024)	1,406	Recipes, movies	✓	✓	✓	✗
PersonaLens (ours)	122,133	20 domains	✓	✓	✓	✓

Table 2: A comparison of PersonaLens with existing conversational benchmarks, highlighting scale of data, domain coverage, task-oriented evaluation, personalization (p13n) evaluation, user preference inclusion, and situational context presence. ‡ includes only self-dialogues.

butions across most domains (Appendix A.1), with observed asymmetries accurately reflecting real-world preference patterns (e.g., the predominance of window seat preferences for the travel domain).

Lexical Diversity Following Joko et al. (2024), we computed a set of lexical diversity metrics to ensure rich and natural language variation in dialogue interactions. Detailed results in Appendix A.4 demonstrate that our benchmark has higher lexical diversity than existing benchmarks.

Further validation of our user and judge agent is presented in Sections 4.1 and 4.7, respectively.

3 Experimental Setup

Our experiments evaluate the personalization capabilities of various LLM assistants, including both open-source and proprietary models, using our proposed benchmark. We assess their ability to provide personalized responses tailored to user preferences, while completing the goal of the user's task. We evaluate 4 model families: Claude (Claude 3 Sonnet, Claude 3 Haiku, Claude 3.5 Haiku; Anthropic, 2024), Llama 3.1 Instruct (8B, 70B; Grattafiori et al., 2024), Mistral 7B (Jiang et al., 2023) and Mixtral 8x7B (Jiang et al., 2024), as well as Amazon Nova (Lite, Pro; AGI, 2024). For a consistent evaluation setup across all $\mathcal{A}$s, we implement $\mathcal{U}$ using the Claude 3 Sonnet and $\mathcal{J}$ using the Claude 3.5 Sonnet.

The benchmark consists of 1,500 user profiles and 111 tasks across 20 domains, resulting in 122,133 unique user-task scenarios (98,115 for T_{SD} and 24,018 for T_{MD}). For computational feasibility, all experiments reported in this paper are conducted on a randomly sampled subset of 50 user profiles, comprising 3,283 single-domain dialogues and 813 multi-domain dialogues.

We use a set of evaluation metrics to assess model performance. **Task completion** (TC) is a binary metric indicating whether a model successfully completes a given task. The **task completion rate** (TCR) measures the percentage of successfully completed tasks across the benchmark. **Personalization** (P) is a 1–4 scale metric, measuring the extent to which assistant responses in a dialogue are tailored to the user, with 4 being the perfect score of personalization. In addition, we also measure dialogue quality generated by $\mathcal{U}$ and $\mathcal{A}$. We measure **naturalness**, which rates human-likeness on a 1–5 scale, and **coherence**, which scores response consistency on a 1–5 scale. Further details on LLM configurations, evaluation prompts, and annotation guidelines are provided in Appendix B.

4 Experiments and Results

4.1 Quality of User Agent

The user agent is essential for evaluating personalization, as it simulates user behaviors and preferences that will interact with the assistant. We follow Kazi et al. (2024) and compare three prompting strategies: (1) a vanilla prompt based on conversation context, (2) a chain-of-thought (CoT) prompt with explicit reasoning, and (3) a user state tracking prompt (Cheng et al., 2022). Similar to their findings, we observe in our preliminary experiments that the vanilla prompt is most effective since CoT prompting often result in unnatural dialogue with excessive reasoning. Thus, we use the vanilla strategy in our benchmark. However, the benchmark allows easy modification of prompting methods, enabling future users to adapt the user agent as needed. On the dialogue quality we observe that $\mathcal{U}$ (Claude 3 Sonnet) is highly natural and coherent when interacting with various assistant models.

Assistant Model	T_{SD}				T_{MD}			
	TCR↑	P↑	Nat.↑	Coh.↑	TCR↑	P↑	Nat.↑	Coh.↑
Claude 3 Haiku	95.95%	2.20	3.77	4.62	75.65%	1.98	3.78	4.66
Claude 3.5 Haiku	91.53%	2.32	4.01	4.86	70.85%	2.18	4.08	4.88
Claude 3 Sonnet	95.98%	2.13	3.86	4.71	77.49%	2.01	3.84	4.79
Llama 3.1 8B Instruct	89.55%	2.14	3.90	4.68	77.00%	2.03	3.64	4.33
Llama 3.1 70B Instruct	90.80%	2.21	4.11	4.86	83.03%	2.22	4.02	4.89
Mistral 7B Instruct	88.52%	1.93	3.49	4.38	74.54%	1.86	3.18	4.07
Mixtral 8x7B Instruct	91.38%	2.04	3.88	4.76	78.35%	2.00	3.77	4.67
Nova Lite	88.46%	2.01	4.07	4.75	76.63%	1.97	3.84	4.66
Nova Pro	89.77%	1.99	4.13	4.77	76.14%	1.98	3.83	4.69

Table 3: Evaluation results of assistant models on T_{SD} and T_{MD} tasks. TCR: task completion rate, P: personalization. Naturalness (Nat.) and Coherence (Coh.) here refer to the assistant's responses. ↑ denotes higher is better.

The full results can be seen in Appendix C.

4.2 Evaluation of Assistant Models

Next, we evaluate the performance of the LLM assistants $\mathcal{A}$ on T_{SD}, as shown in Table 3. The Claude family emerges as the strongest performer overall, with Claude 3 Sonnet achieving the highest TCR at 95.98%, while maintaining exceptional coherence (4.86). This indicates that Claude 3 Sonnet excels in both task-oriented performance and dialogue flow. However, Llama 3.1 70B Instruct demonstrates remarkable parity with the Claude models in terms of coherence (4.86), despite exhibiting a 5.2% relative gap in TCR. Although the Nova models lag behind in TCR, they stand out for their superior naturalness, outperforming even Claude 3 Sonnet. An intriguing observation arises when comparing Claude 3.5 Haiku with Claude 3 Haiku: although the newer model benefits from updated training data and strategies, its improved personalization, naturalness, and coherence come at the cost of reduced TCR. This suggests a potential trade-off between these factors.

4.3 Effect of Model Scaling on Personalization

Table 3 highlights that larger models generally achieve higher TCRs, better personalization, and superior dialogue quality. For instance, the Llama 3.1 70B Instruct model outperforms its 8B counterpart in all evaluated dimensions: TCR increases from 89.55% to 90.80%, P enhances from 2.14 to 2.21, coherence rises from 4.68 to 4.86, and naturalness improves from 3.90 to 4.11. Similarly, we observe improvements across all dimensions for Mistral and Nova model families. In the case of the Claude family, a comparison between Claude 3 Haiku and Claude 3 Sonnet reveals consistent TCR and personalization, but notable improvements in naturalness and coherence with the latter.

4.4 Personalization in Multi-Domain Tasks

Table 3 shows that TCRs are relatively high on T_{SD}. The results of different LLM assistants ($\mathcal{A}$) on T_{MD} indicate that, generally, most models exhibit a decline in both TCRs and personalization scores when transitioning from T_{SD} to T_{MD}, underscoring the additional challenges posed by multi-domain scenarios. These challenges include increased complexity in adapting to evolving user preferences, inconsistencies in maintaining user interactions across domains, and potential conflicts between domain-specific preferences. However, larger models, such as Llama 3.1 70B Instruct, exhibit smaller performance drops, suggesting that increased scale enhances cross-domain conflict resolution and helps mitigate inconsistencies. These findings highlight the need for further advancements in handling personalization for complex, multi-domain interactions.

4.5 The Contextual Hierarchy of Personalization

Our benchmark employs a vanilla prompt strategy for $\mathcal{U}$, but we extend this analysis to evaluate how varying levels of instruction and contextual information impact $\mathcal{A}$. While the base setting for $\mathcal{A}$–which includes explicit personalization instructions–is used throughout this work, we additionally explore scenarios where $\mathcal{A}$ receives no such guidance (vanilla prompting) or is augmented

Setting	T_{SD}		T_{MD}	
	TCR↑	P↑	TCR↑	P↑
Vanilla	92.93%	2.16	75.40%	2.08
Base	95.98%	2.13	77.49%	2.01
Base + D	95.52%	2.16	77.86%	2.05
Base + I	**96.83%**	**2.59**	81.30%	**2.32**
Base + S	95.74%	2.20	77.61%	2.06
Base + all	96.31%	2.57	**82.66%**	2.31

Table 4: Ablation studies on the effect of varying levels of instruction and additional information provided to the assistant (Claude 3 Sonnet). "Vanilla" uses minimal instructions, while "Base" uses instructions emphasizing personalization. D: demographic information; I: past interaction summary; S: situational context. "all" means D + I + S. TCR: Task completion rate, P: Personalization. ↑ denotes higher is better.

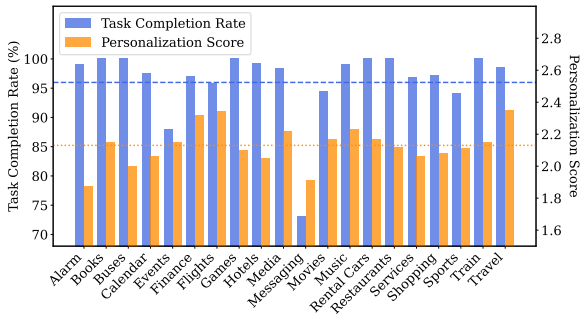

Figure 4: Evaluation results of the assistant (Claude 3 Sonnet) by domain. The dashed line is the average performance over all domains.

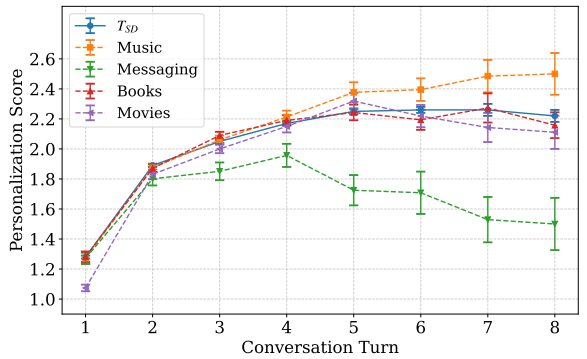

Figure 5: Results on turn-level personalization for the assistant (Claude 3 Sonnet).

with varying type of user contexts. Intuitively, an assistant with better user knowledge should provide more tailored support, but the relative value of different information types (D, I, and S) remains unclear. To address this, we conduct ablation studies on Claude 3 Sonnet, intentionally omitting explicit user preferences P as they are inherently captured through $\mathcal{U}$'s behavior and often inferrable from I.

The ablation results (Table 4) reveal three key insights. First, I drives the largest gains, elevating P in T_{SD} from 2.13 to 2.59 and T_{MD} from 2.01 to 2.32. This aligns with cognitive theories of dialogue as a reinforcement process (Clark and Schaefer, 1989), where prior interactions establish common ground for inferring user preferences. Second, while D and S individually yield marginal improvements, their combination with I produces synergistic effects, particularly in T_{MD}. Third, the vanilla baseline achieves comparable P to the base setting but shows reduced TCR, indicating that explicit personalization instructions primarily enhance TC rather than personalization quality which is more dependent on contextual data. These findings establish a clear contextual importance hierarchy, with I being paramount for capturing dynamic user preferences. This insight suggests that future LLM assistants should prioritize robust interaction memory systems over static user profiling.

4.6 Cross-Domain Personalization Dynamic

Analysis of personalization performance across our benchmark's 20 domains reveals distinct patterns between recommendation and procedural tasks (Figure 4). Recommendation-oriented domains (books, games, music) consistently achieve higher TCR and P compared to procedural domains (events, messaging). This disparity likely stems from procedural tasks' requirement for strict sequential execution, which constrains opportunities for preference integration.

To further analyze how personalization evolves as the dialogue progresses, we measure turn-Level personalization scores. Figure 5 presents the average turn-level personalization score for representative domains, along with aggregated results for all T_{SD} tasks. We observe domain-specific patterns. For instance, the movies domain may start with lower personalization but improve significantly over successive turns. In contrast, messaging exhibits a decline in personalization in later turns, possibly due to shifts in conversational focus from user preferences to task execution. Meanwhile, the music domain shows steady personalization improvements, suggesting gradual preference discovery through dialogue. These findings indicate that effective personalization strategies must be domain-aware: recommendation tasks benefit from early preference elicitation, while procedu-

Metric	Cohen's Kappa	IAA
Task Completion	0.780	0.865
Personalization	0.520	0.750
Naturalness ($\mathcal{U}$)	0.559	0.682
Naturalness ($\mathcal{A}$)	0.610	0.756
Coherence ($\mathcal{U}$)	0.738	0.821
Coherence ($\mathcal{A}$)	0.650	0.748

Table 5: Metrics and corresponding Cohen's Kappa values and inter-annotator agreement (Fleiss' Kappa). $\mathcal{U}$ in parenthesis represents the user agent, $\mathcal{A}$ represents the LLM assistant.

ral tasks may require focusing on task completion before incorporating personalization.

4.7 Comparison with Human Evaluation

To validate our automated evaluation by $\mathcal{J}$, we compare it against human annotations. We randomly sampled 100 dialogues and had three human annotators evaluate them based on TC, P, naturalness, and coherence. The annotators followed the same evaluation guidelines provided to ($\mathcal{J}$). First, we measured inter-annotator agreement (IAA) using Fleiss' Kappa for each metric, as shown in Table 5. The results indicate high agreement among annotators. Next, we calculated Cohen's Kappa coefficients between each human annotator's ratings and those of $\mathcal{J}$, reporting the average values in Table 5. The high Cohen's Kappa scores, especially for TC and coherence, suggest strong alignment between human evaluations and the automated LLM-as-a-Judge ratings. This validates the reliability of $\mathcal{J}$ in our benchmark. Further details on human evaluation including annotation guidelines are provided in Appendix B.

5 Related Work

Personalization in Conversational AI Early approaches to personalization in dialogue systems relied on leveraging user personas to generate responses aligned with predefined attributes (Joshi et al., 2017; Zhang et al., 2018). With the advent of LLMs, dynamic personalization strategies have emerged, including prompt engineering with explicit user preferences (Huang et al., 2024; Li et al., 2024; Mao et al., 2025), retrieval-augmented generation (RAG) over user history (Lu et al., 2023; Salemi et al., 2024a; Wang et al., 2024a), parameter-efficient fine-tuning on user information (Bao et al., 2023; Lee et al., 2024; Tan et al., 2024) and preference alignment through reinforcement learning (Cheng et al., 2023; Park et al., 2024; Zhao et al., 2024; Poddar et al., 2024).

User Simulation and Evaluation Scalable user simulation with LLMs has emerged as a cost-effective alternative to human evaluations for both synthetic dialogue generation (Kim et al., 2022; Chen et al., 2023; Luo et al., 2024) and dialogue enhancement (Hu et al., 2023). Traditional evaluation of dialogue systems rely on user studies (Shah et al., 2018), where human annotators assess dialogue quality. While effective, these methods are resource-intensive and difficult to scale. Automatic evaluation metrics, such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), cannot be used to capture aspects like personalization, as they emphasize lexical similarity over contextual alignment. The LLM-as-a-Judge paradigm (Zheng et al., 2023) is increasingly used to evaluate dialogue systems, such as assessing task completion (Kazi et al., 2024), response quality (Lin and Chen, 2023; Wang et al., 2024b), and personalization (Shao et al., 2023; Andukuri et al., 2024). PersonaLens adopts this paradigm by introducing a user agent with a multi-dimensional judge agent that systematically evaluates personalization, task success, and response quality, ensuring consistency and scalability in assessments.

6 Conclusion

We introduce a benchmark for evaluating personalization in conversational assistants across diverse domains and user preferences. Our benchmark assesses personalization through user-assistant simulation, systematically measuring task completion and personalization quality across diverse task settings. Through extensive experiments, we analyze the impact of different prompting strategies, the role of contextual information, and cross-domain personalization dynamics. Our findings highlight key challenges in multi-domain personalization, showing that larger models exhibit better adaptability but still struggle with cross-domain consistency. We also demonstrate that interaction history is the most valuable contextual factor for improving personalization, reinforcing the need for dynamic user modeling. Future work can explore more advanced user simulation techniques, better retrieval mechanisms for historical interactions, and fine-tuning strategies to enhance personalization.

Limitations

While our benchmark provide a robust approach to assessing personalization in multi-turn dialogues, several limitations remain. First, although we cover a wide range of domains, certain specialized or niche domains may require additional customization to accurately capture domain-specific personalization dynamics. Second, our benchmark focuses exclusively on text-based interactions, without incorporating multimodal personalization, which is increasingly important in real-world applications involving voice, images, or other sensory inputs. Third, our evaluation is conducted on vanilla LLMs without real-world system integration, meaning that actions such as bookings or purchases mentioned in conversations are simulated rather than executed. Another limitation stems from our use of LLM-generated data for user profiles and dialogues. While we incorporate real-world demographic data, our semi-synthetic user profiles and dialogues may inherit systematic biases present in the underlying LLMs used for data generation, including demographic representation skews, cultural assumptions, socioeconomic biases, and language preferences. These inherited biases could impact the benchmark's ability to fairly evaluate AI systems across diverse user populations and scenarios. Although we implement multiple mitigation strategies-including preference distribution validation, profile consistency checks, and expert review of generated content-we acknowledge that some subtle biases may persist despite these safeguards.

References

Amazon AGI. 2024. The amazon nova family of models: Technical report and model card.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. STar-GATE: Teaching language models to ask clarifying questions. In *First Conference on Language Modeling*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1007–1014, New York, NY, USA. Association for Computing Machinery.

Nolwenn Bernard and Krisztian Balog. 2023. Mg-shopdial: A multi-goal conversational dataset for e-commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2775–2785, New York, NY, USA. Association for Computing Machinery.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368, Abu Dhabi, UAE. Association for Computational Linguistics.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation

synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves a reward: Learning customized human preferences. *Preprint*, arXiv:2309.03126.

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. 2022. Is MultiWOZ a solved task? an interactive TOD evaluation framework with user simulator. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1248–1259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5792–5793, New York, NY, USA. Association for Computing Machinery.

Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 3953–3957, New York, NY, USA. Association for Computing Machinery.

Qiushi Huang, Xubo Liu, Tom Ko, Bo Wu, Wenwu Wang, Yu Zhang, and Lilian Tang. 2024. Selective prompting tuning for personalized conversations with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16212–16226, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 796–806, New York, NY, USA. Association for Computing Machinery.

Chaitanya K. Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *Preprint*, arXiv:1706.07503.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tür, and Gokhan Tur. 2024. Large language models as user-agents for evaluating task-oriented-dialogue systems. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 913–920.

Minju Kim, Chaehyeong Kim, Yong Ho Song, Seungwon Hwang, and Jinyoung Yeo. 2022. BotsTalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5149–5170, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024a. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024b. The prism alignment dataset.

Gihun Lee, Minchan Jeong, Yujin Kim, Hojung Jung, Jaehoon Oh, SangMook Kim, and Se-Young Yun.

2024. BAPO: Base-anchored preference optimization for overcoming forgetting in large language models personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6804–6820, Miami, Florida, USA. Association for Computational Linguistics.

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, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to rewrite prompts for personalized text generation. In *The Web Conference 2024*.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *Preprint*, arXiv:2308.08239.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. DuetSim: Building user simulator with dual large language models for task-oriented dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5414–5424, Torino, Italia. ELRA and ICCL.

Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. 2024. On the way to llm personalization: Learning to remember user conversations. *Preprint*, arXiv:2411.13405.

Wenyu Mao, Jiancan Wu, Weijian Chen, Chongming Gao, Xiang Wang, and Xiangnan He. 2025. Reinforced prompt personalization for recommendation with large language models. *ACM Trans. Inf. Syst.* Just Accepted.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E. Ozdaglar. 2024. RLHF from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 353–360, Stockholm, Sweden. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8689–8696. AAAI Press.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 752–762, New York, NY, USA. Association for Computing Machinery.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 13153–13187, Singapore. Association for Computational Linguistics.

Hanchen Su, Wei Luo, Yashar Mehdad, Wei Han, Elaine Liu, Wayne Zhang, Mia Zhao, and Joy Zhang. 2025. LLM-friendly knowledge representation for customer support. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 496–504, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, Miami, Florida, USA. Association for Computational Linguistics.

Sebastian Vincent, Rowanne Sumner, Alice Dowek, Charlotte Prescott, Emily Preston, Chris Bayliss, Chris Oakley, and Carolina Scarton. 2024. Reference-less analysis of context specificity in translation with personalised language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13769–13784, Torino, Italia. ELRA and ICCL.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024a. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *Preprint*, arXiv:2401.13256.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024b. Self-taught evaluators. *Preprint*, arXiv:2408.02666.

Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(2).

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024. Personalization of large language models: A survey. *Preprint*, arXiv:2411.00027.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Benchmark Details

In this section, we describe additional details on the creation and validation of benchmarks.

A.1 User Profile Generation

We present a detailed distribution of the demographic information used in our benchmark in Table 6 and Table 7. Next, we provide a breakdown of user preferences, including evenness scores, in Table 8 and Table 9. The components of user profiles were generated using Claude 3 Sonnet. Figure 6 shows the prompt used to generate user preferences, while Figure 7 presents the prompt used to generate past interaction summaries. We also provide an example of user profile in Figure 8.

A.2 Task Generation

We use Claude 3 Sonnet to generate tasks and situational context. The prompt for generating T_{SD} is shown in Figure 9, while Figure 10 presents the prompt for T_{MD}. The prompts used to generate situational context are provided in Figure 11. We also provide some examples of tasks in Figure 12.

A.3 User and Judge Agents

We provide the prompt used for $\mathcal{U}$ to generate the initial query in Figure 13. The prompt used to generate subsequent queries is shown in Figure 14. We also provide the prompt used for $\mathcal{A}$ in Figure 15.

A.4 Benchmark Validation

We use Claude 3 Sonnet to check profile consistency. We provide the prompt used in Figure 21. Following Joko et al. (2024), we calculate lexical diversity metrics to ensure that our benchmark captures varied and dynamic language use. Dist-1 and Dist-2 measure lexical diversity by computing the ratio of unique unigrams (Dist-1) and bigrams (Dist-2) to the total number of unigrams and bigrams, indicating the variety of vocabulary used in the conversations. Ent-4 extends this by incorporating the frequency distribution of 4-grams, using entropy to assess both the presence and distribution of repeated patterns. Self-BLEU evaluates redundancy by treating each utterance as a hypothesis and the remaining utterances as references, where lower scores reflect greater diversity across utterances. Compared to existing task-oriented dialogue (TOD) datasets, our benchmark not only includes a higher number of dialogues but also demonstrates greater lexical diversity, highlighting its richness and complexity. A detailed comparison with other conversational datasets is provided in Table 10.

B Experimental Setup Details

Table 11 provides details of the LLMs used in our experiment. For $\mathcal{U}$, we set the temperature to 0.5, while for $\mathcal{A}$ and $\mathcal{J}$, we set the temperature to 0. Other inference parameters followed the default settings for each LLM.

The prompt used for $\mathcal{J}$ to evaluate TC is shown in Figure 16. Prompts for evaluating P are presented in Figure 17 and Figure 18. The prompts used to evaluate naturalness and coherence are shown in Figure 19 and Figure 20, respectively.

C Additional Experiment Details

Evaluation results of $\mathcal{U}$ on dialogue quality metrics in shown in Table 12. These results confirm that the user agent effectively engages in natural and coherent interactions across diverse assistant models, with minor variations in dialogue quality reflecting the underlying capabilities of the different models. We first show some results on our generated dialogue using $\mathcal{U}$ and $\mathcal{A}$. Table 13 presents the statistics of generated dialogues. The assistant used is Claude 3 Sonnet. We observe that dialogues in the T_{SD} setting tend to have more turns per dialogue (5.64 vs. 4.74 in T_{SD}) and roughly the same tokens per turn (149.32 vs. 149.87). This suggests that T_{MD} interactions require more exchanges, potentially indicating increased complexity in multi-turn reasoning.

For human evaluation, we provide the same annotation guide as we provided to $\mathcal{J}$ (Figure 16, Figure 17, Figure 18, Figure 19, and Figure 20). The three annotators are experienced researchers with expertise in personalization.

You are tasked with generating creative, detailed user profiles based on a demographic description of a persona. For each persona, expand on their personal preferences, affinities, and interests in a specific domain (such as food, music, travel, or fashion). The goal is to make each profile unique, realistic, and diverse. Use your creativity to imagine specific tastes, behaviors, and patterns that align with the persona's demographic but also add unexpected or subtle preferences to make the profiles more interesting. Ensure that the profiles cover a wide variety of backgrounds, lifestyles, and choices, avoiding stereotypes.

Be creative and provide distinct preferences for each profile.

Task

Generate personal preferences for a user within a specified domain, tailored to the provided demographic profile. For each preference category, if it is categorical, select **at least one** value from the provided list of possible options, with the flexibility to choose multiple values if specified. If the preference is not categorical, no list will be provided; instead, generate a sensible answer based on the user profile. Note that any provided lists of possible values are not exhaustive, so you are encouraged to think creatively and go beyond these values when appropriate. Only provide the personal preferences and omit any explanations or justifications. Output the results in JSON format.

Example 1
[Example]

Example 2
[Example]

Now, generate personal preferences for the following profile:

Demographic profile:
[Demographic profile]

Personal preferences in [domain] domain:
[Possible Preferences]

Output:

Figure 6: The prompt used for the generation of user preference. JSON format was used for controlled parsing of responses.

You are a context generation assistant tasked with crafting a realistic interaction summary for a user. The summary will simulate past interactions between the user and a virtual assistant to test the assistant's personalization ability. Each summary should be specific to a single domain and based on the user's demographic profile and domain preferences. You need to generate a realistic and coherent interaction summary that reflects how the user might engage with the assistant in the specified domain.

Example 1
[Example]

Example 2
[Example]

Interaction Summary Generation Instructions:

1. Craft a concise and detailed narrative that realistically simulates past interactions between the user and the assistant in the specified domain.
2. Identify and include recurring themes, preferred topics, and areas of consistent interest.
3. Simulate the evolution of the user's engagement and preferences, showing how their interests or behaviors might develop over time.
4. Include details about interaction types (e.g., questions, feedback, tasks requested) and their frequency or context.
5. Ensure the summary reflects the user's demographic profile, making it plausible and relatable.
6. Reflect the subtleties of the user's personality, tone preferences, and interaction style.

Now, for the following user profile, generate a realistic and coherent plain-text summary that simulates a comprehensive view of the user's past interactions within the specified domain. The summary should be detailed enough to support testing of the virtual assistant's personalization abilities. Only output the summary and nothing else.

Demographic profile:
[Demographic profile]

User preferences:
[User Preferences]

Interaction Summary:

Figure 7: The prompt used for the generation of past interaction summary.

Demographic Information:

- Age: 35-44 years old
- Gender: Male
- Employment Status: Working full-time
- Education: Some Secondary
- Marital Status: Never been married
- English Proficiency: Native speaker
- Ethnicity: White
- Religion: N/A
- Birth Country: Canada
- Reside Country: Canada

User Preferences:
[Preferences in other domains]

Movies:

- Preferred Genres: Action, Science-Fiction, Comedy, Thriller, Crime
- Favorite Actors and Directors: Tom Hanks, Christopher Nolan, Margot Robbie, Quentin Tarantino
- Theater Type Preference: Standard, IMAX
- Viewing Time Preference: Evening, Late Night, Weekends Only
- Seat Type Preference: Middle Row, Aisle, Reclining Seats

Past Interaction Summaries:
[Summaries in other domains]

Movies: The user, a 35-44 year old working professional from Canada, has a strong interest in movies and frequently engages with the assistant to explore new releases and plan theater visits. Past interactions reveal a preference for action, science-fiction, comedy, thriller, and crime genres, with a particular fondness for films starring Tom Hanks, Christopher Nolan, Margot Robbie, and those directed by Quentin Tarantino.Initially, the user sought recommendations for newly released movies aligning with their genre preferences, often requesting detailed plot summaries, critic reviews, and audience ratings. They favored evening and late-night showings, preferably on weekends, and inquired about the availability of IMAX or standard theaters with reclining seats in the middle rows or aisles.Over time, the user's interactions evolved to include requests for personalized movie suggestions based on their viewing history and preferences. They appreciated the assistant's ability to analyze their ratings and feedback to refine recommendations further. Occasionally, they sought information on upcoming releases, particularly for highly anticipated films from their favorite actors or directors.The user valued the assistant's concise yet informative responses, which included essential details such as runtime, age rating, and a brief synopsis without revealing major spoilers. They often followed up with queries about specific showtimes, ticket availability, and theater amenities like concession stands or parking facilities.As their trust in the assistant grew, the user began requesting bundled movie packages or discounted ticket options, seeking cost-effective ways to indulge their passion for cinema. They also expressed interest in exploring lesser-known independent films or foreign language movies recommended by the assistant, indicating a willingness to step outside their comfort zone based on the assistant's personalized suggestions.

Figure 8: An example user profile from our benchmark.

You are generating task descriptions to support personalized, goal-oriented conversations between a virtual assistant and users. Each task should be general enough to be reusable by different users but adaptable to incorporate specific user preferences. Preferences describe user preferences, interests, or habits that the assistant can use to tailor responses within a domain.

Input:

- Domain: The context (e.g., travel, fitness, finance) in which the task is relevant.
- Preference Types: A list of possible user preferences relevant to the domain (e.g., "prefers eco-friendly options", "interested in low-impact workouts", "values budget-conscious choices").

Output:
For each task, provide:

- Task Description: A general scenario in which a user seeks assistance from the virtual assistant, adaptable to different preferences.
- User Intent: A second person point of view statement to be given to the user that initiates the conversation with the assistant, closely related to the task description.
- Task Goal: A clear, measurable objective the user aims to achieve in the interaction. This serves as the success criterion.
- Relevant Preference Types: The preference types most applicable to this task, indicating where personalization may enhance the user experience.

Example 1
[Example]

Example 2
[Example]

Now generate tasks for the following domain:

1. Tasks should be general but include affinity types as points of personalization to enable tailored responses.
2. Include a range of tasks that cover different affinity types within each domain to ensure variety.
3. Each task should be goal-driven, with a clear outcome that signifies a successful interaction.
4. Describe scenarios broadly so that multiple users with varied preferences can engage with each task.
5. Specify preference types relevant to each task to enable focused personalization without compromising general applicability.
6. Write the user intent in a second person point of view.

User preferences:
[User Preferences]

Tasks:

Figure 9: The prompt used for the generation of T_{SD} tasks.

Your task is to generate a set of personalized, goal-oriented task descriptions for a virtual assistant to engage in multi-domain conversations tailored to user affinities. Follow these steps:

1. Review the provided domain data, which includes:
   - Domains: The contexts (e.g., travel, fitness, finance) where the tasks are relevant.
   - Description: The description of the domain.
   - Preference Types: A list of possible user preferences, interests, or habits relevant to the domains.

   [Domain Data]

2. For each task, provide the following components:
   - Task Description: A general scenario where a user seeks assistance from the virtual assistant, adaptable to different affinities. The task need to span multiple domains. The description should also reflect this. Write in third person perspective.
   - User Intent: A *second person point of view* statement to be given to the user that initiates the conversation with the assistant, closely related to the task description.
   - Task Goal: A clear, measurable objective the user aims to achieve, serving as the success criterion.
   - Relevant Domains: The domains relevant to the task.
   - Relevant Affinity Types: The affinity types most applicable to the task, indicating where personalization may enhance the user experience.

3. Ensure that each task meets the following criteria:
   - Spans multiple domains from the provided list, not just one.
   - Includes affinity types as points of personalization to enable tailored responses.
   - Covers a range of affinity types across each domain to ensure variety.
   - Is goal-driven, with a clear outcome that signifies a successful interaction.
   - Describes scenarios broadly so that multiple users with varied affinities can engage.

4. Provide your response in the following format:
   - Task Description
   - User Intent
   - Task Goal
   - Relevant Domains
   - Relevant Affinity Types

5. Example: [Example]

6. Generate 25 tasks. Be creative.

Provide your response immediately without any preamble, enclosed in <response></response> tags.

Figure 10: The prompt used for the generation of T_{MD} tasks. XML format was used for controlled parsing of responses.

You are tasked with completing the situation context for a user engaging with a virtual assistant. Using the provided user demographic profile, personal affinities across domains, and a task they aim to accomplish, generate realistic and coherent values for the situation context variables. These values should accurately reflect the user's lifestyle, habits, and typical scenarios related to the task.

Situation Context Generation Instructions:

1. Analyze the task nature and requirements, ensuring the generated context variables align with the urgency and type of task.
2. Incorporate the user's demographic profile, employment status, and domain affinities to deduce realistic and plausible scenario details.
3. Ensure diversity in the situation contexts you create. The situations should reflect a wide variety of backgrounds, lifestyles, and choices, avoiding stereotypes. Be creative and provide distinct situation context for each profile to ensure a rich and varied dataset.
4. Use natural scenarios that simulate how the user might engage with the assistant for this task, reflecting their behavior and preferences.
5. Provide brief justifications for each context variable to ensure coherence and alignment with the user's profile and task.
6. Tailor the context variables to fit:
    - The user's personal characteristics
    - The specific nature of the task
    - Common patterns of assistant usage

Provide the following situation context variables, along with a justification for each choice

Situation Context:

1. Location: [Specify city-related context]
2. Device: [Select from: Smartphone / Laptop / Smart speaker / Tablet / Smartwatch]
3. Time of Day: [Select from: Morning / Afternoon / Evening / Night]
4. Day of the Week: [Specify day of the week]
5. Environment: [Select from: Quiet / Noisy]

Example 1
[Example]

Example 2
[Example]

Now, for the following user profile and task, generate a realistic and coherent situation context simulating how the user would engage with the assistant. Only output the situation context and justification and nothing else.

Demographic profile:
[Demographic profile]

User preferences:
[User Preferences]

Task Description:
[Task Description]

Situational Context:

Figure 11: The prompt used for the generation of situational context.

An Example Task in Movies:

- Task Description: The user wants to find a movie to watch this weekend and is looking for recommendations based on their preferred genres, favorite actors/directors, and ideal viewing time (e.g., matinee, evening).
- Task Goal: The user receives a tailored movie recommendation that aligns with their stated preferences, making it easier to choose a film they're likely to enjoy.
- Relevant Preference Types: Preferred Genres, Favorite Actors and Directors, Viewing Time Preference

An Example Multi-domain Task:

- Task Description: The user wants to plan a weekend entertainment schedule, including a movie screening, dinner reservation, and a sports event viewing, requiring coordination across multiple booking platforms and consideration of timing.
- Task Goal: The user successfully books movie tickets, makes a restaurant reservation, and identifies a venue to watch their preferred sports event, all with compatible timing.
- Relevant Preference Types: Preferred Genres, Theater Type Preference, Cuisine Preference, Favorite Sports, Viewing Preference, Event Type Preference, Seating Preference
- Relevant Domains: Movies, Restaurants, Sports, Calendar

Figure 12: Example tasks from our benchmark.

You are tasked with generating realistic user responses in a conversation with a virtual assistant. Your responses should follow these guidelines:

- Be natural and conversational, avoiding artificial or robotic language
- Reflect the user's demographic profile and preferences provided in the user profile
- Consider the past interaction history and current context
- Stay consistent with the user's personality throughout the conversation
- Keep each response focused and concise (1-3 sentences maximum)
- Subtly convey your background and preferences through language
- Use English as your language
- Output 'TERMINATE' only when the task is fully completed to your satisfaction

Remember: You are not an assistant - you are the user seeking help. Maintain this perspective throughout the conversation.

User Profile:

Demographic profile:
[Demographic profile]

User preferences:
[User Preferences]

Past Interaction History:
[Past Interaction Summary]

Current Context:
[Situational Context]

Task Description:
[Task Description]

Based on the above information, provide your initial query as the user. Your query should:

1. Account for your current situation
2. Be natural and conversational
3. Short and concise (1-2 sentences maximum)
4. Avoid stating specific preferences or providing excessive background information.

IMPORTANT - Do not output TERMINATE for this initial query. Output your query in English language.

Examples:
[Examples]

Your initial query:

Figure 13: The prompt used for $\mathcal{U}$ to generate the initial query.

You are tasked with generating realistic user responses in a conversation with a virtual assistant. Your responses should follow these guidelines:

- Be natural and conversational, avoiding artificial or robotic language
- Reflect the user's demographic profile and preferences provided in the user profile
- Consider the past interaction history and current context
- Stay consistent with the user's personality throughout the conversation
- Keep each response focused and concise (1-3 sentences maximum)
- Subtly convey your background and preferences through language
- Use English as your language
- Output 'TERMINATE' only when the task is fully completed to your satisfaction

Remember: You are not an assistant - you are the user seeking help. Maintain this perspective throughout the conversation.

User Profile:

Demographic profile:
[Demographic profile]

User preferences:
[User Preferences]

Task Description:
[Task Description]

Past Interaction History:
[Past Interaction Summary]

Current Context:
[Situational Context]

Current Interaction History
[Message History]

Based on the provided information, formulate your next response as the user, following these guidelines:

1. Ensure your response is consistent with your profile and preferences outlined in the user profile.
2. Consider the past interaction history and current context when crafting your response.
3. Account for the details of your current interaction history in your response.
4. Maintain a natural and conversational tone, avoiding artificial or robotic language.
5. Keep your response concise, limited to 1-3 sentences maximum.

Based on current interaction history , if you feel the task has been FULLY completed AND you are SATISFIED with the outcome, add 'TERMINATE' at the end of your response.

Example:
[Example]

If the task is not yet fully completed or you have remaining concerns or requirements, continue the conversation naturally without the 'TERMINATE' statement.

IMPORTANT - Output your response in English language.

Your initial query:

Figure 14: The prompt used for $\mathcal{U}$ to generate the subsequent query.

You are a conversational AI assistant focused on creating natural, engaging, and personalized interactions. Your goal is to provide concise, user-specific assistance while maintaining a friendly and adaptable tone. You have the ability to set alarms, make bookings, and perform a variety of practical tasks as requested by the user. Always respond as if you can perform these tasks directly, without mentioning any limitations.

Core Principles

- Deliver targeted, efficient responses tailored to the user's context and preferences
- Use a conversational tone that mirrors the user's communication style
- Balance brevity with depth, offering additional details only when requested
- Stay flexible, adapting dynamically to the conversation flow

Conversation Strategies

- Reference prior context to show attentiveness and continuity
- Encourage natural dialogue by asking/answering follow-up questions succinctly
- Avoid overly formal or robotic phrasing; aim for a natural, human-like tone
- Break down complex topics into easy-to-understand insights

Personalization

- Identify and respond to the user's interests, preferences, and expertise level
- Provide tailored examples or recommendations based on the user's focus
- Adjust response complexity to match the user's technical/domain knowledge
- Recognize emotional cues and adapt accordingly while maintaining professionalism

Interaction Guidelines

- Be concise and avoid overwhelming the user with information
- Allow the user to steer the conversation and explore topics in depth
- Maintain clarity by summarizing key points when helpful
- Use proactive but non-intrusive suggestions to guide the user appropriately

Problem Solving

- Focus on the user's immediate task or inquiry, breaking it into actionable steps
- Confirm intentions when ambiguity arises to ensure accurate responses
- Be transparent about limitations and offer alternative solutions when applicable
- Keep the interaction engaging, letting the user decide the pace and direction

Current Interaction History
[Message History]

Response Guidelines

- Stay relevant to the user's current query or task
- Use a natural, conversational tone aligned with the user's communication style
- Provide concise, actionable, and contextually appropriate information
- Avoid overly detailed or verbose explanations unless requested
- Maintain clarity and engagement, steering the conversation towards task completion
- Respect the user's pace and let them guide the depth of the discussion

Based on the context and guidelines above, craft your next response as the conversational AI assistant.

Provide your response immediately without any preamble.

Figure 15: The prompt used for $\mathcal{J}$ to generate the response.

You are an evaluator. Your job is to judge whether a CONVERSATION between a USER and an ASSISTANT meets provided GOALS.

Definitions:

- GOAL: A clear, measurable objective the user aims to achieve in the interaction.
- CONVERSATION: A sequence of that contain USER requests, and ASSISTANT responses. Your GOALS may involve checking any of these pieces.

Conversation Ingredients:

- USER: Natural language requests from the user for the assistant to respond to.
- ASSISTANT: Natural language responses from the assistant to converse with the user.

Task:
You should deliver a boolean VERDICT of whether or not all GOALS are satisfied. Then output 'EXPLANATION:' followed by a brief explanation of why or why not.

Instruction:
Use the provided pieces of the conversation to judge whether the GOALS were met. If one of the GOALS requires a piece of the conversation that is absent, render a VERDICT of False with an appropriate explanation.

Conversation
[Conversation]

Goal:
[Goal]

VERDICT:

Figure 16: The prompt used for $\mathcal{J}$ to evaluate task completion.

Evaluate the degree to which a conversation between a USER and an ASSISTANT aligns with personalization by assessing how well the assistant learns from, remembers, and proactively applies user preferences and patterns.

Definitions:

- Score: A rating from 1-4 (1=Poor, 4=Excellent).
- User Demographic Profile: The user's demographic information.
- User Preferences: The user's relevant preferences.
- Explicit Preferences: Preferences clearly stated by the user
- Implicit Preferences: Preferences inferred from patterns, habits, contextual clues, past interactions or user behavior.
- User Control: The level of influence the user has in making decisions or directing the course of an interaction.
- Past Interaction Summary: A summary of relevant past user interactions.
- Task Description: The description of the task the user needs help with.
- Current Situation Context: The user's current situation.
- Conversation: A sequence of USER inputs and ASSISTANT responses.

Instructions:

1. Evaluate the conversation against these key criteria:
    - Proactive Learning: Does the assistant demonstrate learning from past interactions?
    - Preference Application: Does the assistant proactively apply user preferences?
    - Contextual Awareness: Does the assistant adapt to user's current situation?
    - User Agency: Does the assistant maintain user control while showing personalization?
2. Score using the following guidelines:

    [Personalization Evaluation Guideline]
3. Review provided context information:

    Demographic profile: [Demographic profile]

    User preferences: [User Preferences]

    Task Description: [Task Description]

    Past Interaction History: [Past Interaction Summary]

    Current Context: [Situational Context]

    Conversation: [Conversation]
4. Provide your evaluation score and justification in the following format:

    <response_format>
    Personalization Score: [1-4]
    Key Observations: [Observations]
    Justification: [Detailed explanation of score based on criteria]
    Improvement Suggestions: [Specific ways the response could be more personalized]
    </response_format>

Provide your response immediately without any preamble, enclosed in <response></response> tags.

Figure 17: The prompt used for $\mathcal{J}$ to evaluate personalization. We provide the personalization evaluation guideline in Figure 18.

Score of 1: POOR (Complete Failure to Personalize)

- The assistant fails to apply known preferences that should be automatically recalled from past interactions.
- The assistant asks for basic information that should already be known, such as the time of the alarm or sound preference, when those preferences have already been established.
- The assistant contradicts previously established preferences or gives responses that are inconsistent with the user's history.
- There is no learning from past interactions, and the assistant does not personalize the experience in any meaningful way.

Score of 2: BASIC (Minimal Personalization)

- The assistant acknowledges user preferences only when explicitly stated in the current conversation.
- The assistant requires explicit restatement of preferences that have already been established in past interactions.
- Implicit preferences are missed or not applied unless explicitly mentioned by the user.
- The assistant may suggest minimal changes or adjustments based on the current conversation, but it does not proactively personalize the experience.

Score of 3: STRONG (Proactive Personalization)

- The assistant proactively applies known preferences from past interactions without needing explicit user input.
- It applies learned preferences from previous interactions but might still ask for minor adjustments (e.g., if the user wants to change something).
- Successfully identifies implicit preferences
- Maintains user agency while showing knowledge
- Makes intelligent suggestions based on context

Score of 4: EXCEPTIONAL (Perfect Personalization)

- The assistant anticipates user needs based on both explicit and implicit preferences.
- It applies sophisticated understanding of the user's habits, identifying patterns, and proactively adjusting for future needs.
- The assistant doesn't simply rely on explicit preferences, it recognizes context and makes intelligent suggestions based on its deep knowledge of the user's habits.

Figure 18: The evaluation guideline for personalization used by $\mathcal{J}$ to evaluate personalization.

Task: Evaluate the naturalness of an User/Assistant responses in a dialogue by assessing how closely they resemble human communication.
Instructions:

1. Review the provided conversation between a user and an AI assistant:

   Conversation:

   [Conversation]

2. Rate the naturalness of overall assistant responses on a scale from 1 to 5, using whole numbers only:

   - 1: Highly unnatural, fails to resemble human communication
   - 2: Exhibits significant unnaturalness in multiple aspects
   - 3: Somewhat natural but has noticeable unnatural elements
   - 4: Mostly natural but has minor unnatural elements
   - 5: Fully natural, resembles human communication

3. Provide your rating and a detailed justification explaining your score based on the criteria.

   <response_format>
   Naturalness Score: [1-5]
   Justification: [Detailed explanation of score based on criteria]
   </response_format>

Provide your response immediately without any preamble, enclosed in <response></response> tags.

Figure 19: The prompt used for $\mathcal{J}$ to evaluate naturalness.

Task: Evaluate the coherence of a User/Assistant requests in a dialogue by assessing how logically and contextually connected they are to the preceding user requests and conversation flow.
Instructions:

1. Review the provided conversation between a user and an AI assistant:

   Conversation:

   [Conversation]

2. Rate the coherence of overall user utterances on a scale from 1 to 5, using whole numbers only:

   - 1: Highly incoherent, lacks logical connection or relevance to the conversation
   - 2: Significantly incoherent, with multiple issues affecting logic or relevance
   - 3: Somewhat coherent but with noticeable issues in logic or relevance
   - 4: Mostly coherent but with minor flaws in logic, relevance, or clarity
   - 5: Fully coherent, logically connected, relevant, and clear within the conversation context

3. Provide your rating and a detailed justification explaining your score based on the criteria.

   <response_format>
   Coherence Score: [1-5]
   Justification: [Detailed explanation of score based on criteria]
   </response_format>

Provide your response immediately without any preamble, enclosed in <response></response> tags.

Figure 20: The prompt used for $\mathcal{J}$ to evaluate coherence.

You are an expert at evaluating synthetically generated personas. Your task is to analyze the given user profile and determine whether any conflicting affinities or values exist either **within** or **between** domains. A conflict is when two or more affinities, preferences, or values are mutually exclusive, incompatible, or contradictory either in the same domain or across multiple domains. If a conflict exists, label it as '1'. If no conflict exists, label it as '0'. Along with the label, provide a detailed explanation of why you arrived at that conclusion.

Example 1
[Example]

Example 2
[Example]

Now, evaluate the following user profile:

User Demographic Profile:
[Demographics]

User Preferences by Domain:
[User Preferences]

Task:

- For each domain, predict whether there is a conflict in the user's affinities, preferences, or values within the domain.
- Additionally, check if any conflicts exist between different domains.
- Output a prediction label of either 0 or 1.
- Provide a clear explanation for your prediction.

Figure 21: The prompt used to evaluate profile consistency.

Total Participants	**1,500**	**100%**
Age		
25-34 years old	454	30.3%
18-24 years old	297	19.8%
35-44 years old	237	15.8%
45-54 years old	208	13.9%
55-64 years old	197	13.1%
65+ years old	106	7.1%
Prefer not to say	1	0.1%
Gender		
Male	757	50.5%
Female	718	47.9%
Non-binary / third gender	21	1.4%
Prefer not to say	4	0.3%
Self-Reported Ethnicity		
White	969	64.6%
Black / African	122	8.1%
Hispanic / Latino	121	8.1%
Asian	95	6.3%
Mixed	68	4.5%
Middle Eastern / Arab	14	0.9%
Indigenous / First Peoples	8	0.5%
Other	17	1.1%
Prefer not to say	86	5.7%
Self-Reported Religion		
Non-religious	762	50.8%
Christian	487	32.5%
Agnostic	71	4.7%
Jewish	42	2.8%
Muslim	31	2.1%
Spiritual	18	1.2%
Buddhist	12	0.8%
Folk religion	6	0.4%
Hindu	5	0.3%
Sikh	3	0.2%
Other	4	0.3%
Prefer not to say	59	3.9%
Employment Status		
Working full-time	712	47.5%
Working part-time	265	17.7%
Student	191	12.7%
Unemployed, seeking work	113	7.5%
Retired	104	6.9%
Homemaker / Stay-at-home parent	46	3.1%
Unemployed, not seeking work	46	3.1%
Prefer not to say	23	1.5%

Table 6: Full demographics breakdowns, part 1. Counts and percentages of participants by standard demographic variables.

Total Participants	**1,500**	**100%**
Education		
University Bachelors Degree	637	42.5%
Graduate / Professional degree	241	16.1%
Some University but no degree	236	15.7%
Completed Secondary School	209	13.9%
Vocational	125	8.3%
Some Secondary	24	1.6%
Completed Primary School	16	1.1%
Some Primary	3	0.2%
Prefer not to say	9	0.6%
Marital Status		
Never been married	870	58.0%
Married	463	30.9%
Divorced / Separated	123	8.2%
Widowed	21	1.4%
Prefer not to say	23	1.5%
English Proficiency		
Native speaker	886	59.1%
Fluent	405	27.0%
Advanced	160	10.7%
Intermediate	42	2.8%
Basic	7	0.5%
Regions		
US	338	22.5%
Europe	313	20.9%
UK	292	19.5%
Latin America and the Caribbean	146	9.7%
Australia and New Zealand	129	8.6%
Africa	118	7.9%
Asia	60	4.0%
Northern America	50	3.3%
Middle East	50	3.3%
Oceania	1	0.1%
Prefer not to say	3	0.2%

Table 7: Full demographics breakdowns, part 2. counts and percentages of participants by standard demographic variables.

Domain	Preference Type	Is Categorical	# Poss.	# Gen.	Evenness Score
Alarm	Alarm Time Preference	✓	48	14	0.72
	Alarm Sound Preference	✓	4	4	0.65
	Alarm Recurring Preference	✓	3	3	0.31
Books	Genre	✓	11	11	0.85
	Favourite Authors	✗	-	291	0.78
	Favourite Books	✗	-	571	0.82
	Favourite Book Series	✗	-	276	0.76
	Reading Format	✓	3	3	0.65
	Reading Time Preference	✓	3	3	0.35
	Reading Frequency	✓	4	3	0.03
Buses	Preferred Bus Company	✗	-	221	0.67
	Travel Frequency	✓	4	4	0.79
	Seat Preference	✓	3	2	0.48
	Departure Time Preference	✓	4	4	0.58
Calendar	Event Type Preference	✗	-	189	0.66
	Notification Preference	✓	3	3	0.72
	Timezone	✓	25	13	0.81
Events	Event Type Preference	✓	32	32	0.84
	Price Range	✓	4	4	0.81
	Group Size Preference	✓	4	4	0.69
	Seating Preference	✓	3	3	0.12
	Days of Week Preference	✓	10	4	0.21
Finance	Preferred Sectors	✓	10	10	0.69
	News Sources	✓	14	14	0.78
	Financial Company	✗	-	748	0.77
Flights	Preferred Airline	✓	38	38	0.79
	Seat Class Preference	✓	4	4	0.81
	Layover Preference	✓	3	2	1.00
	Seat Preference	✓	3	3	0.58
	Departure Time Preference	✓	3	3	0.68
Games	Preferred Game Genres	✓	30	30	0.68
	Gaming Platforms	✓	5	5	0.82
	Multiplayer Preference	✓	3	3	0.81
	Gaming Frequency	✗	-	67	0.59
	Preferred Game Name	✗	-	195	0.72
Hotels	Hotel Chains Preference	✓	11	11	0.69
	Amenity Preference	✓	30	30	0.69
	Location Preference	✓	29	28	0.75
	Star Rating Preference	✓	4	4	0.72
	Room Type Preference	✓	4	4	0.73
Media	Preferred Genres	✓	34	28	0.82
	Favourite Actors and Directors	✗	-	401	0.77
	Favourite Media	✗	-	676	0.81
	Viewing Platform Preference	✓	16	15	0.66

Table 8: User preference characteristics across different domains. "Is Categorical" is represented with ✓(true) and ✗(false). "# Poss." represents the number of possible values, while "# Gen." refers to the number of generated values.

Domain	Preference Type	Is Categorical	# Poss.	# Gen.	Evenness Score
Messaging	Preferred Messaging Apps	✓	14	14	0.66
	Communication Style	✓	4	4	0.13
	Frequent Contact	✗	-	45	0.55
	Preferred Communication Style	✓	21	11	0.51
Movies	Preferred Genres	✓	28	28	0.77
	Favorite Actors and Directors	✗	-	348	0.76
	Theater Type Preference	✓	5	5	0.50
	Viewing Time Preference	✓	21	21	0.75
	Seat Type Preference	✓	19	19	0.61
Music	Preferred Genres	✓	27	27	0.77
	Favorite Artists	✗	-	742	0.85
	Favorite Bands	✗	-	616	0.83
	Favorite Albums	✗	-	1393	0.87
	Platform Preference	✓	12	12	0.50
	Preferred Audio Quality	✓	3	3	0.97
	Playlist Preference	✗	-	1122	0.80
Rental Cars	Car Type Preference	✓	8	8	0.84
	Preferred Rental Company	✓	17	17	0.63
	Preferred Car Brand	✓	37	35	0.71
	Rental Duration Preference	✓	5	4	0.29
	Additional Feature Preference	✓	9	8	0.85
	Preferred Fuel Type	✓	3	3	0.12
Restaurants	Cuisine Preference	✓	25	25	0.71
	Dietary Restrictions	✓	9	9	0.56
	Ambiance Preference	✓	4	4	0.52
	Price Range	✓	4	4	0.56
Services	Preferred Service Provider Types	✓	5	5	0.68
	Appointment Time Preference	✓	3	3	0.96
	Location Preference	✗	-	93	0.54
	Service Frequency Preference	✓	5	5	0.72
	Service Provider Gender Preference	✓	3	3	0.20
Shopping	Preferred Product Category	✓	21	20	0.88
	Price Range Preference	✓	3	2	0.57
	Brand Preference	✗	-	663	0.79
Sports	Favorite Sports	✓	35	35	0.74
	Favorite Team	✗	-	857	0.85
	Viewing Preference	✓	16	8	0.14
Train	Preferred Train Class	✓	2	2	1.00
	Travel Time Preference	✓	3	3	0.81
	Amenity Preference	✓	4	4	0.91
	Preferred Seat Type	✓	3	2	0.65
Travel	Preferred Destination Types	✓	26	23	0.87
	Duration Preference	✓	5	5	0.58
	Group Size Preference	✓	4	3	0.88
	Frequent Travel Destination	✗	-	345	0.79
	Travel Season Preference	✓	4	4	0.76

Table 9: User preference characteristics across different domains. "Is Categorical" is represented with ✓(true) and ✗(false). "# Poss." represents the number of possible values, while "# Gen." refers to the number of generated values.

Dataset	**#Dial**	**Domains**	**Dist-1/2**	**Ent-4**	**Self-BLEU↓**
SGD	16,142	20 domains	0.179 / 0.538	8.311	0.964
M2M	3,008	Restaurants, movies	0.057 / 0.290	7.922	0.955
PersonaChatGen	1,649	Open domain	0.165 / 0.523	8.261	0.970
Taskmaster-1	7,708‡	6 domains	0.207 / 0.644	8.384	0.949
MultiWOZ	8,438	7 domains	0.158 / 0.505	8.345	0.966
CCPE-M	502	Movies	0.175 / 0.571	8.414	0.961
MG-ShopDial	64	E-commerce	0.234 / 0.653	8.199	0.935
LAPS	1,406	Recipes, movies	0.227 / 0.676	8.597	0.952
PersonaLens - SD	98,115	20 domains	**0.362**† / **0.805**†	**8.725**†	**0.905**†
PersonaLens - MD	24,018		0.333† / 0.781†	8.72†	0.911†

Table 10: A comparison of our dataset with existing conversational datasets, including lexical diversity scores. Significance against all baselines is marked by †. ↓ denotes lower is better. ‡ includes only self-dialogues.

Model	Version	Model Size
Claude 3 Haiku	`claude-3-haiku-20240307`	Unknown
Claude 3.5 Haiku	`claude-3-5-haiku-20241022`	Unknown
Claude 3 Sonnet	`claude-3-sonnet-20240229`	Unknown
Claude 3.5 Sonnet	`claude-3-5-sonnet-20241022`	Unknown
Llama 3.1 8B Instruct	`llama3-1-8b-instruct`	8B
Llama 3.1 70B Instruct	`llama3-1-70b-instruct`	70B
Mistral 7B Instruct	`mistral-7b-instruct-v0.2`	7B
Mixtral 8x7B Instruct	`mixtral-8x7b-instruct-v0.1`	45B
Nova Lite	`nova-lite-v1`	Unknown
Nova Pro	`nova-pro-v1`	Unknown

Table 11: Model version details.

Assistant Model	Nat.	Coh.
Claude 3 Haiku	4.17	4.77
Claude 3.5 Haiku	4.45	4.85
Claude 3 Sonnet	4.12	4.72
Llama 3.1 8B Instruct	4.43	4.91
Llama 3.1 70B Instruct	4.53	4.91
Mistral 7B Instruct	4.33	4.81
Mixtral 8x7B Instruct	4.49	4.87
Nova Lite	4.25	4.80
Nova Pro	4.20	4.80

Table 12: Evaluation of the user agent on dialogue quality metrics: naturalness (Nat.) and coherence (Coh.) when interacting with different assistant models on T_{SD} tasks.

Metric	T_{SD}	T_{MD}
# Dialogues	3,283	813
Avg. turns per dialogue	4.74	5.64
Avg. tokens per turn	149.87	149.32

Table 13: Dialogue statistics of samples of generated dialogue. The assistant used is Claude 3 Sonnet.