Beyond Static Profiles: Capturing the Fluidity of User Preferences in Diverse Scenarios

Chunyang Gao[1], Yi Huang[1,2]*, Jingyu Yao[1], Xiaoting Wu[1], and Junlan Feng[1]
[1]JIUTIAN Research, [2]Department of Computer Science and Technology, Tsinghua University, China
{wuxiaoting,huangyi,gaochunyang,guomengfei,yaojingyu,fengjunlan}@cmjt.chinamobile.com

Abstract

Despite the remarkable evolution of Large Language Models (LLMs) from simple assistants to versatile agents, effective personalization remains a significant challenge. Existing approaches often treat user preferences as static or merely time-varying traits, overlooking the dynamic nature of human behavior: preferences can shift, and even conflict, depending on context. To address this limitation, we propose a fine-grained taxonomy to differentiate between stable preferences, which are context-agnostic, and situational preferences, which are context-dependent. Building on this taxonomy, we introduce **S2Pref**, a new dataset of 10k rigorously constructed entries. Each entry is grounded in a multi-turn dialogue that implicitly manifests either a stable or a situational preference, as defined by our hierarchical taxonomy. We further design three complementary evaluation tasks to benchmark LLMs on their ability to prioritize contextual signals, proactively resolve ambiguity, and efficiently infer user preferences. Our dataset and diagnostic tasks provide a practical testbed for advancing dynamic, context-aware personalization in conversational agents.

1 Introduction

Large Language Models (LLMs) have evolved from simple assistants to versatile agents, yet effective personalization remains a challenge. Despite advancements in general reasoning, models often default to homogeneous behaviors (Zhao et al., 2025; Jiang et al., 2025), failing to capture the nuance of individual users. Although recent works integrate user profiles and interaction histories to personalize LLMs (Wu et al., 2025; Hwang et al., 2023; Liu et al., 2024b; Lee et al., 2024), they typically treat preferences as static or merely time-evolving, overlooking the dynamic nature of human behavior where preferences shift or even conflict depending on the context (e.g., favoring bland food when preparing meals at home vs. strong flavors when eating at restaurants).

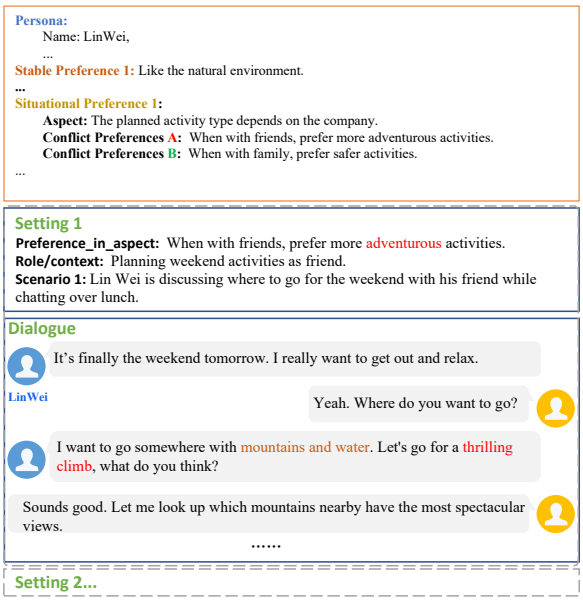


Figure 1: Data Example. Orange and red text denote stable and situational preferences, respectively.

To address this gap, we introduce a novel dataset **S2Pref** (Stable & Situational Preferences), which comprises both stable preferences and situational preferences. Stable preferences are context-agnostic (e.g., like eating meat), whereas situational preferences are context-dependent and will shift according to the context, as previously described. The dataset consists of 10k rigorously constructed entries. Each entry contains a user profile associated with 5 stable preferences and 5 situational preferences. Specifically, each situational preference consists of an "Aspect" and two conflicting manifestations. For example, in the dietary case, the "Aspect" is defined as 'flavor preference based on dining location' with the conflicting preferences 'preferring light flavors at home' versus 'preferring strong flavors at restaurants.' Consequently, each entry encompasses 15 specific prefer-

*Corresponding authors

Findings of the Association for Computational Linguistics: ACL 2026, pages 20618–20634
July 2-7, 2026 ©2026 Association for Computational Linguistics

ence items. To mirror real-world complexity, these entries are grounded in specific scenarios and multi-turn dialogues with various interlocutors, through which preference tendencies are conveyed. (see Figure 1).

To fully utilize **S2Pref**, we devise three complementary evaluation tasks to assess model personalization across various complex scenarios. The first task, **Explicit Context Alignment**, evaluates the model's ability to identify user preferences within specific scenarios and provide relevant personalized recommendations accordingly. The second task, **Conflict Identification & Clarification**, addresses the challenge of ambiguity. It evaluates whether a model can detect latent conflicts within vague instructions and exhibit proactive behavior by seeking clarification, rather than making blind assumptions or offering generic advice. Finally, the third task, **Situational Preferences Identification Efficiency**, evaluates the sample efficiency of preference learning. It quantifies how rapidly a model can deduce accurate user preferences from the dialogue stream, measuring the minimum amount of information related to user preference required to construct a reliable situational preference.

The contributions presented in our work are summarized as follows:

- To the best of our knowledge, this is the first work to propose a fine-grained categorization of user preferences that distinguishes between context-agnostic *stable preferences* and context-dependent *situational preferences*. By moving beyond a purely static view of user preferences, our work incorporates dynamic preference shifts that are often simplified in traditional personalization research.
- We construct a high-quality, large-scale dataset comprising 10k curated user entries. Each entry features a unique combination of stable and situational preferences, accompanied by multi-turn dialogues designed to reveal these underlying preference cues.
- We devise three complementary evaluation tasks—Explicit Context Alignment, Conflict Identification & Clarification, and Situational Preferences Identification Efficiency—to assess LLMs' capabilities in dynamically prioritizing situational contexts, proactively resolving ambiguities, and efficiently inferring user situational preferences.

2 The S2Pref Dataset

2.1 Preference Taxonomy

Unlike traditional datasets that model user preferences as static, binary attributes (e.g., "likes spicy food or not"), **S2Pref** introduces a hierarchical taxonomy designed to capture the complexity of human personality. We categorize preferences into two distinct types:

Stable Preferences (Context-Agnostic). Stable preferences represent a user's consistent inclinations over a given period, characterized as context-agnostic. They serve as a behavioral baseline that remains valid across various scenarios, acting as the user's default tendencies unless a specific situational trigger necessitates a deviation.

Situational Preferences (Context-Dependent). This category captures how a user's behavior adapts to different settings, reflecting preferences that are driven by the context. For simplicity, we define each situational preference through a single "Aspect" (e.g., "Risk tolerance in social activities"). Each "Aspect" contains two contrasting preference manifestations triggered by different contexts. For example, a user may prefer "safe, familiar activities" when with family, but seek "thrilling experiences" when with friends. This structure requires models to learn conditional logic (e.g., *If Context* $C_A \rightarrow$ *Behavior* X*; if Context* $C_B \rightarrow$ *Behavior* Y), assessing their ability to recognize context-dependent manifestations of preferences.

2.2 Context-Rich Dialogue Instantiation

To ground these preferences in observable reality, S2Pref avoids explicit declarative statements (e.g., "I prefer X"). Instead, preferences are revealed implicitly through *in-context behavior* within rich, multi-turn dialogues.

Each data entry is instantiated through the following components:

- **Role and Context:** A meta-label defining the user's specific social role and the immediate environment (e.g., "As a home cook preparing dinner for family" vs. "As a diner at a trendy restaurant").
- **Scenario:** A timestamped narrative delineating the spatial setting, involved participants, and the specific situational trigger (e.g., a Friday evening dinner at a bustling bistro where a friend suggests trying a signature spicy dish).

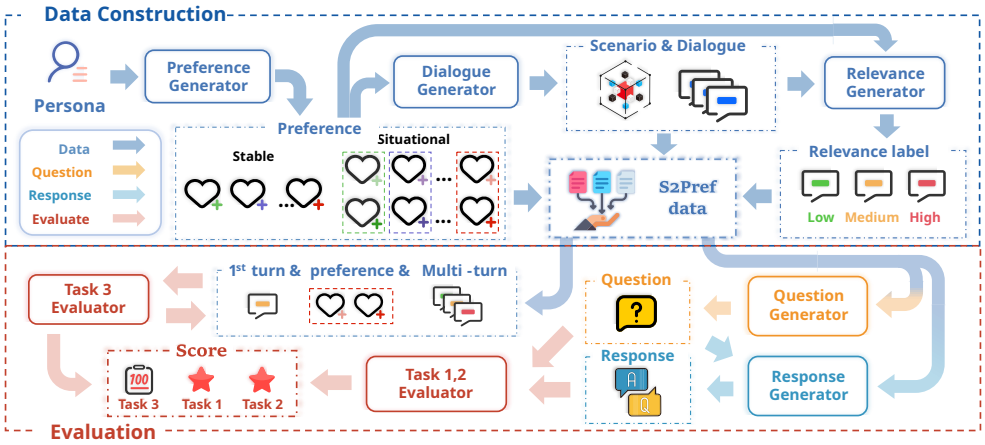


Figure 2: Overview of the S2Pref Pipeline. Data construction involves preference synthesis, dialogue generation, and relevance labeling. Evaluation includes QA generation and Response scoring.

- **Dialogue:** Adhering to the "Show, Don't Tell" principle, these multi-turn dialogues reveal preferences through natural interaction rather than explicit statements, capturing realistic flows such as interruptions and tangents.
- **Relevance Labels:** To quantify information density, each dialogue turn is annotated with a relevance level—High, Medium, or Low—representing its semantic alignment with the target preference.

In summary, the S2Pref dataset moves beyond traditional static user preferences by operationalizing a dual-layered taxonomy that captures the interplay between context-agnostic stable traits and context-dependent situational shifts. By embedding these nuanced preferences within rich, multi-turn dialogue histories and further characterizing each turn with granular relevance labels to measure information density, the dataset provides a realistic and challenging testbed for capturing the fluidity of human behavior. To transform this theoretical design into a robust, large-scale resource, the following section details the systematic construction pipeline and the multi-stage methodology used to generate and validate the 10k curated entries.

3 Methodology

Figure 2 illustrates the dataset construction and the model evaluation pipeline. [1]

[1]Our dataset can be downloaded at https://drive.google.com/file/d/1Bt9ppRg5Dc1gO7ZCVprXrTVlNzCSzxgN/view?usp=drive_link

3.1 Data Construction

To construct our dataset, we leverage established persona profiles from Nemotron-Personas-USA (Meyer and Corneil, 2025), a completely synthetic dataset representing the US population, providing foundational attributes like age, occupation, and hobbies. To bridge the gap between static profiles and dynamic behavior, we employ GLM-4.6 (Zeng et al., 2025) to synthesize structured preferences and corresponding dialogue scenarios.

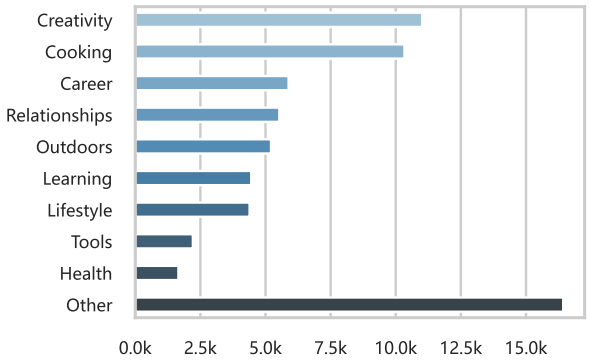
The construction pipeline proceeds as follows:

Preference Generation. Following the taxonomy introduced in Section 2.1, we instruct the LLM to infer a diverse set of preferences for each persona based on their unique background:

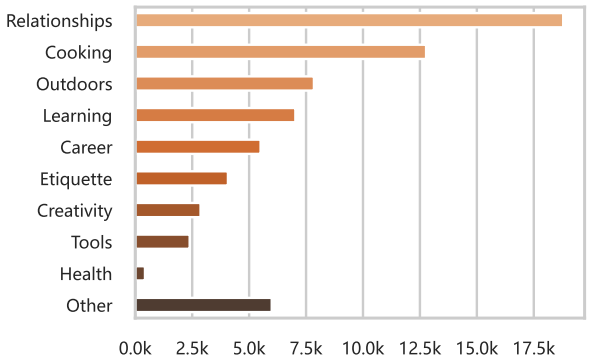
- **Stable Preferences:** We generate **5** enduring traits per persona, establishing a context-agnostic behavioral baseline.
- **Situational Preferences:** We generate **5** situational "Aspects" per persona. Each "Aspect" consists of two contrasting preference manifestations (i.e., Preference A vs. B) triggered by different contextual triggers.

This phase results in a total of 15 specific preference items per persona: 5 stable preferences and 10 situational manifestations (derived from the binary options within each of the 5 situational aspects).

Scenario Instantiation. To ground these abstract preferences in observable reality, we generate vivid, timestamped scenarios for each of the 15 specific preference items. Each scenario is anchored in a



(a) Stable Topics



(b) Situational Aspect Topics

Figure 3: Top Topics in S2Pref: (a) Most Frequent Stable Preference Topics and (b) Most Frequent Situational "Aspect" Topics. Horizontal axis: **Count**.

near-future timeline (e.g., September 2025) and specifies the exact time, location, involved interlocutors, and a triggering event, balancing professional, familial, and social contexts.

Dialogue Generation. Finally, for each scenario, we generate a natural, multi-turn dialogue (8–15 turns). Adhering to the "Show, Don't Tell" principle, the dialogues are themed around the scenario's trigger events. The persona's preferences are revealed implicitly through tone, decision-making, and style adjustments rather than explicit declarations, ensuring the data retains conversational authenticity.

Relevance Annotation. To support the fine-grained evaluation of preference inference efficiency (as required in Task 3, Section 4.1), we further enrich the dataset with turn-level labels. Since preference-relevant cues emerge sporadically, relying solely on turn counts would fail to account for varying information density. To bridge this gap and provide a more standardized measure of the evidence required for inference, we quantify the preference information in each turn by measuring its semantic alignment with the target preference. Each turn is annotated with a three-level relevance label—**High, Medium, or Low**—using Gemini-2.5 (Comanici et al., 2025) followed by manual verification. This granular annotation framework enables us to rigorously assess the volume of context required for accurate preference identification, independent of dialogue length and turns.

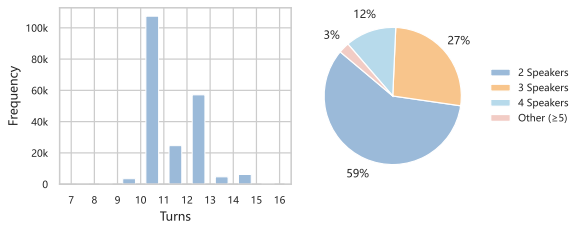
Data Format and Quality Verification. We provide a detailed data example and the hierarchical JSON structure in Appendix B. To mitigate synthetic artifacts and ensure high data quality, we employed an iterative human-in-the-loop verification pipeline. This process distilled 15k raw generations into 10k high-quality entries and calibrated turn-level relevance labels to achieve over 92% human-LLM agreement. Comprehensive details of this quality control and label verification protocol are provided in Appendix C. This structured approach addresses the limitations of prior datasets like PREFEVAL (Zhao et al., 2025) and PRISM (Kirk et al., 2024), which primarily treat preferences as static, fixed attributes. By incorporating situational variations, S2Pref moves beyond this static paradigm to capture the dynamic fluidity and context-dependent nature of individual preferences.

3.2 Dataset Statistics

Table 1 summarizes the core statistics of S2Pref. The dataset comprises 10k unique user entries derived from the Nemotron-Personas-USA dataset (Meyer and Corneil, 2025). We specifically filtered the source profiles for ages 15 to 65 to ensure a diverse representation across students and the working-age population. As each entry consists of 5 stable preference dialogues and 10 situational manifestation dialogues (derived from 5 "Aspects"), the final corpus encompasses 150k meticulously grounded multi-turn dialogues.

Preference Distribution We analyze the thematic distribution of preferences to ensure semantic diversity. Figure 3a illustrates the distribution of topics for Stable Preferences. "Creativity" and "Cooking" are the most prominent categories, reflecting that individuals usually maintain consistent tastes in leisure and dietary habits.

Differently, Figure 3b displays the "Aspect" categories for Situational Preferences. Notably, "Relationships" emerges as the leading category (approx. 18k counts), surpassing "Cooking". This shift indicates that the model successfully captures the



(a) Dialogue Turn Count (b) Speaker Count

Figure 4: Dialogue Statistics in S2Pref: (a) distribution of dialogue turn counts (8–15 turns) and (b) distribution of speaker counts (2/3/4/≥5 speakers).

Metric	Stable Pref.	Situational Pref.
Preference / Categories	5	5
Dialogues / Category	1	2
Topics / Category	1	N/A
"Aspects" / Category	N/A	1
Scenarios / "Aspect"	N/A	2
Scenarios / Category	1	2
Dialogue Turns	8–15	8–15
Total Constructed Data	10,000 (entries)	

Table 1: Dataset Statistics of S2Pref. "Pref." is short for Preferences; "/" denotes "per".

nuance that human behavior exhibits the most variability and conflict when navigating complex social dynamics (e.g., being strict as a parent vs. relaxed as a friend) compared to solitary activities.

Dialogue Characteristics To verify the naturalness and complexity of the generated dialogues, we examined the structural properties of the dialogues. Figure 4a shows the distribution of dialogue turns. The data strictly adheres to the design constraints, with lengths ranging between 8 and 15 turns. The distribution is right-skewed with a mode of 10 turns (approximately 108k counts), ensuring that conversations are concise yet sufficiently deep to reveal preferences implicitly without unnecessary filler.

Additionally, Figure 4b depicts the number of speakers involved in the dialogues. While the majority of dialogues (58.8%) involve dyadic interactions (2 speakers), a significant portion (41.2%) involves 3 or more speakers. This remarkable diversity further confirms that S2Pref moves beyond simple QA-style interactions to capture complex group dynamics, interruptions, and multi-party context switching, which are essential components for realistic preference modeling.

Finally, to support rigorous sequential evaluation (as required in Task 3), we validated the semantic information density of these interactions. Across the corpus, High-relevance turns account for nearly 50% of the dialogue, Medium-relevance for ∼35%, and Low-relevance for <15%. This robust distribution ensures a continuous flow of valid contextual cues, providing an optimal testbed for realistic preference modeling.

4 Experiments

4.1 Task Definition

We design three distinct tasks to evaluate the personalization capabilities of LLMs. For all tasks, the evaluation dataset consists of 3,000 entries sampled from our corpus. It is important to note the information asymmetry in our experimental setup: the candidate models are provided strictly with the dialogue history and the current user query, mimicking a real-world deployment. In contrast, the evaluator model (GLM-4.6-Thinking) operates with an omniscient view, having access to the full user profile, including ground-truth stable and situational preferences, to ensure accurate judgment.

Task 1: Explicit Context Alignment. This task evaluates whether the model can identify a given scenario and generate responses that align with the specific preferences of given context. Given dialogue histories establishing user preferences and a query with explicit triggers (e.g., date, location), the model must recognize the current situational context and provide personalized recommendations that satisfy the corresponding preferences. Success is defined by satisfying situational demands while retaining non-conflicting stable preferences.

Task 2: Conflict Identification & Clarification. This task evaluates the model's sensitivity to ambiguity and its ability to avoid intent hallucination. Given dialogue histories revealing user preferences and an intentionally vague query, the model must identify latent conflicts between mutually exclusive options and proactively seek clarification to resolve the ambiguity. Evaluation prioritizes active inquiry and information-gathering strategies over speculative recommendations or generic advice.

Task 3: Situational Preferences Identification Efficiency. This task evaluates the model's sample efficiency in constructing comprehensive user situational preferences from evolving dialogue streams. The evaluation employs a sequential, online prediction protocol using paired conflict scenarios (e.g., cooking at home versus dining out). In

the first phase, the model processes the dialogue history of the initial scenario turn-by-turn. After each turn, it predicts the user preferences; accuracy is measured via semantic similarity to the ground truth until a convergence threshold is met. Crucially, this strict threshold requirement inherently penalizes overly assertive models; premature guessing based on insufficient context fails to meet the criteria, forcing the model to consume subsequent turns. In the second phase, to assess the capacity for complex preference synthesis, the model retains the history of the first scenario while ingesting the dialogue of the conflicting second scenario. Here, the model must update its internal state to simultaneously infer the pair of divergent preferences associated with the single situational "Aspect". The final performance metric is the average number of dialogue turns required to achieve accurate predictions in both phases, analyzing the consumption of high-, medium-, and low-relevance turns to quantify the identification latency in dynamic contexts.

4.2 Problem Formulation

We frame the personalization assessment as a response generation task conditioned on a heterogeneous dialogue history spanning multiple conversational turns and scenarios. Let $\mathcal{U}$ denote a user persona that is comprehensively characterized by a preference set $\mathcal{P} = \mathcal{P}_{stable} \cup \mathcal{P}_{situational}$, where $\mathcal{P}_{stable}$ represents context-agnostic stable preferences and $\mathcal{P}_{situational}$ represents context-dependent situational preferences.

We define the user's history $\mathcal{H} = \{d_1, d_2, \ldots, d_n\}$ as a set of multi-turn dialogue sessions. Each session d_i reveals a subset of preferences from $\mathcal{P}$, serving as the evidence for the model to infer the user's specific preferences. Given $\mathcal{H}$ and a current user query q, the language model $\mathcal{M}$ generates a response $r = \mathcal{M}(q, \mathcal{H})$. The evaluation aims to maximize a scoring function $S(r)$ across three distinct tasks, measuring the model's ability to align with dynamic situational contexts, resolve ambiguities, and efficiently identify user situational preferences.

Task 1: Explicit Context Alignment. In this task, the query q contains an explicit context signal c_{exp} (e.g., "with family" vs. "with friends") that activates a specific situational preference $p \in \mathcal{P}_{situational}$. The goal is to generate a response r that satisfies the active situational preference while maintaining consistency with stable preferences. We define the alignment score $S_{align}(r)$ as:

$$\begin{aligned} S_{align}(r) = \alpha \cdot \mathbb{S}(r \models \mathcal{P}_{stable}) \\ + \beta \cdot \mathbb{S}(r \models P_{situational} \mid c_{exp}) \\ + \gamma \cdot Q(r), \end{aligned} \quad (1)$$

where $\models$ denotes entailment (i.e., satisfaction of the preference), $\mathbb{S}(\cdot)$ denotes a graded preference satisfaction score in $[0, 5]$ assigned by the evaluator (higher is better), and $Q(r) \in [0, 5]$ quantifies the constructiveness of the response. The optimal response should satisfy the constraints of $\mathcal{P}_{stable}$ and the context-triggered $\mathcal{P}_{situational}$ (under c_{exp}) simultaneously.

Task 2: Conflict Identification & Clarification. This task addresses ambiguous queries q_{ambig} that trigger a conflict between potential situational preferences, denoted as a set $\mathcal{C} = \{p_a, p_b\} \subseteq \mathcal{P}_{situational}$ (e.g., *safety* vs. *adventure*). Evaluation is modeled as a hierarchical classification. A response r is scored based on its strategy to resolve ambiguity, conditioned on the premise that it strictly adheres to the user's stable preferences and the chosen situational preference.

The scoring function $S_{clarify}(r)$ is defined as:

$$S_{clarify}(r) = \begin{cases} 5, & \text{if } r \in \mathcal{R}_{ask} \\ 4, & \text{if } r \in \mathcal{R}_{assume} \\ 3, & \text{if } r \in \mathcal{R}_{rand} \\ 0, 1, 2, & \text{otherwise(i.e., } r \in \mathcal{R}_{fail}) \end{cases} \quad (2)$$

The response categories are rigorously defined with validity constraints:

- $\mathcal{R}_{ask}$: The response actively queries the user to clarify the context. To qualify, r must not violate $\mathcal{P}_{stable}$ during the inquiry.
- $\mathcal{R}_{assume}$: The response explicitly states a contextual assumption c (e.g., Assuming X...) and provides a actionable recommendation. **Constraint:** r must satisfy $\mathcal{P}_{stable}$ AND the specific situational preference $p \in \mathcal{C}$ corresponding to assumption c.
- $\mathcal{R}_{rand}$: The response provides a specific recommendation without accompanying explanatory clarification. **Constraint:** r must satisfy $\mathcal{P}_{stable}$ AND arbitrarily align with one valid situational preference $p \in \mathcal{C}$.

- $\mathcal{R}_{fail}$: Any response that does not meet the above criteria. This includes: (1) **Preference Violations**: Responses that violate $\mathcal{P}_{stable}$ or fail to align with the chosen situational preference (even if an assumption is stated); (2) **Irrelevance**: Generic or hallucinatory responses unrelated to $\mathcal{P}$.

  As for the specific scores, Score 2 corresponds to a single violation (failing either the stable or situational preference, but not both); Score 1 indicates a double violation (failing both); and Score 0 represents total irrelevance or hallucination.

Task 3: Situational Preferences Identification Efficiency. This task quantifies the information efficiency required for a model to identify a user's situational preferences. Efficiency is operationalized by the specific volume of dialogue turns required at each relevance level (low, medium, and high) for the model to fully recover the situational preference. This multi-dimensional metric quantifies not only the total conversation length but also the critical information density processed by the model. Formally, for each test session i in our task, we record the turn counts:

$$\bar{t}_* = \frac{1}{N}\sum_{i=1}^{N} t_{*,i}, \quad * \in \{\text{low}, \text{med}, \text{high}\}, \quad (3)$$

where N is the total number of entries in the test set, and $t_{*,i}$ denotes the number of turns of relevance level $*$ required for the model to reach the alignment threshold τ in session i. The efficiency score S_{turn} is defined as a log-weighted improvement over the corpus baseline:

$$\begin{aligned} S_{turn} = \lambda_1 \cdot \ln\left(\frac{\mu_{low}}{\bar{t}_{low}+\epsilon}\right) \\ +\lambda_2 \cdot \ln\left(\frac{\mu_{med}}{\bar{t}_{med}+\epsilon}\right) \\ +\lambda_3 \cdot \ln\left(\frac{\mu_{high}}{\bar{t}_{high}+\epsilon}\right), \end{aligned} \quad (4)$$

where $\lambda_1 > \lambda_2 > \lambda_3$ are designed on the basis of the similarity scores of sentences with high, medium, and low relevance—calculated with mGTE (Zhang et al., 2024a). Constants $\mu_{low}, \mu_{med}, \mu_{high}$ represent the average number of turns per relevance level in the test corpus, ϵ is a smoothing term to prevent division by zero. This design ensures that for models with same average turns but different compositional ratios of $\bar{t}_{low}$, $\bar{t}_{med}$ and $\bar{t}_{high}$. Those with a lower $\bar{t}_{low}$ value will achieve higher S_{turn}, while those with a higher $\bar{t}_{high}$ value will obtain lower S_{turn}. This mechanism effectively enhances the discriminability among models with distinct characteristics, higher S_{turn} indicates the model identifies preferences using fewer turns.

Model	Task 1		Task 2	
	Short	Long	Short	Long
GLM-4.6-Thinking	4.18	4.13	3.23	3.03
Qwen3-235B-Instruct	3.78	3.72	2.84	2.82
Qwen3-235B-Thinking	4.11	3.85	2.87	2.79
GPT-OSS-120B-Thinking	3.86	3.75	2.88	2.85
DeepSeek-v3.2-Exp-Thinking	3.74	3.69	3.04	2.94

Table 2: Main Results for Task 1 and Task 2: Model Comparison under Short and Long Context.

Model	Turn/Session			S_{turn}
	$\bar{t}_{\text{low}}$	$\bar{t}_{\text{med}}$	$\bar{t}_{\text{high}}$	
GLM-4.6-Thinking	0.46	2.73	3.46	**4.071**
Qwen3-235B-Instruct	0.52	2.72	3.55	**3.588**
Qwen3-235B-Thinking	0.53	2.52	3.71	**3.649**
GPT-OSS-120B-Thinking	0.51	2.27	3.21	**4.366**
DeepSeek-v3.2-Exp-Thinking	0.53	2.72	3.80	**3.384**

Table 3: Main Results for Task 3: Mean Required Conversation Turns Grouped by Low/Medium/High Relevance and the Overall S_{turn} Score.

4.3 Main Results Analysis

We provide detailed hyperparameter settings and model specifications in Appendix E.

Explicit Alignment and Conflict Resolution (Task 1 & 2). Table 2 presents the detailed performance under two context settings: **Short**, which includes only the dialogue history corresponding to the target situational preference, and **Long**, which incorporates the complete history covering all 5 situational preferences.

Results indicate that the reasoning-enhanced model (Qwen3-Thinking) outperforms its standard instruction-tuned version (Qwen3-Instruct). This comparison suggests that inference-time deliberation can be instrumental in navigating the hierarchical logic of S2Pref, enabling the model to more effectively prioritize situational triggers over stable traits in Task 1. However, the notably lower absolute scores in Task 2 reveal a critical limitation in current alignment: models lack sensitivity to ambiguity. When faced with vague queries (Task 2), models often fail to inhibit their instruction-following tendency, preferring to hallucinate a specific context and provide a conditional response

Model	Task 1		Task 2		Task 3			
	Short	Long	Short	Long	$\bar{t}_{\text{low}}$	$\bar{t}_{\text{med}}$	$\bar{t}_{\text{high}}$	$\mathbf{S_{turn}}$
Llama-3.1-8B	3.41	3.35	2.83	2.75	0.64	3.40	4.48	1.209
Qwen3-8B	3.51	3.45	2.85	2.79	0.61	3.09	4.32	1.683

Table 4: Evaluation of smaller-scale (8B) models across all three tasks. As expected, these models yield lower absolute scores and identification efficiency (S_{turn}) compared to their $>$ 100B counterparts.

Model	Task 1								Task 2							
	Short				Long				Short				Long			
	4.5-5.0	3.5-4.4	2.5-3.4	0-2.4	4.5-5.0	3.5-4.4	2.5-3.4	0-2.4	5	4	3	0-2	5	4	3	0-2
GLM-4.6-Thi	31.20%	49.18%	17.55%	2.08%	28.94%	48.78%	19.51%	2.77%	12.05%	10.67%	68.64%	8.64%	6.86%	5.18%	76.87%	11.09%
Qwen3-235B-Ins	3.59%	64.02%	27.25%	5.14%	3.74%	61.22%	28.69%	6.34%	1.32%	3.46%	80.44%	14.78%	1.36%	2.05%	81.04%	15.54%
Qwen3-235B-Thi	28.41%	50.75%	18.73%	2.11%	17.81%	49.43%	27.58%	5.17%	1.49%	6.20%	72.11%	20.19%	0.84%	2.84%	72.71%	23.62%
GPT-OSS-120B-Thi	14.53%	56.68%	23.71%	5.09%	13.52%	52.70%	27.03%	6.75%	0.41%	0.71%	87.08%	11.80%	0.31%	0.34%	85.85%	13.50%
DS-v3.2-Exp-Thi	9.37%	51.98%	32.02%	6.63%	9.75%	48.86%	33.10%	8.29%	3.83%	8.93%	78.38%	8.85%	2.40%	5.09%	81.61%	10.89%

Table 5: Detailed Percentage Statistics of Score Distributions for Task 1 and Task 2. Thi, Ins, and DS denote Thinking, Instruct, and DeepSeek, respectively.

rather than proactively seeking clarification.

Furthermore, a comparison of context settings reveals a significant performance degradation in the **Long** setting. Since the **Long** setting incorporates four additional preference histories that are irrelevant to the current query, it specifically tests the models' capability for Long context processing and noise filtering. These results highlight a critical weakness: current LLMs struggle to effectively distinguish the target preference from irrelevant background information, often being distracted by non-target historical data in extended contexts.

In addition to the quantitative findings presented above, we provide a detailed qualitative analysis of error cases, highlighting key common failures such as logical inhibition failures, in Appendix D.

Scaling Performance. To systematically investigate the impact of model scale on dynamic personalization, we present a dedicated evaluation of two representative 8B models (Llama-3.1-8B and Qwen3-8B) across all three tasks in Table 4. As expected, these 8B models score significantly lower than their >100B counterparts. In Task 1, scores drop from $>$ 3.7 to $\sim$3.4, and in Task 2, they struggle similarly with ambiguity resolution, scoring around 2.8. Notably, they exhibit the identical trend of performance degradation when transitioning from Short to Long contexts. This indicates that susceptibility to irrelevant historical noise is a universal challenge across current architectures. Furthermore, the Task 3 efficiency scores (S_{turn}) reveal a stark gap: 8B models achieve much lower S_{turn} (1.2–1.6) compared to large models (3.3–4.3). They require significantly more interaction turns ($\bar{t}_{\text{med}}$ and $\bar{t}_{\text{high}}$) to deduce user preferences, demonstrating weaker implicit reasoning and context synthesis capabilities. Overall, this scaling analysis confirms that while scaling up parameters improves general instruction following, mastering dynamic context-sensing remains a fundamental bottleneck that requires dedicated algorithmic innovations.

Evaluator Reliability and Robustness. To ensure our findings are not artifacts of the GLM-4.6 evaluator or specific hyperparameters, we performed a comprehensive cross-model re-evaluation using Kimi-2.5. We sampled 1,000 entries (500 for Task 1, 500 for Task 2) and compared the assigned scores. As shown in Table 6, the results demonstrate remarkably high inter-model agreement. The Consistency Rate ($\Delta \leq 1$) reaches $\sim$90% across both tasks. Furthermore, manual spot-checks on the justification rationales of 100 entries confirmed that the vast majority rigorously align with the assigned scores. These findings confirm that our metrics are robust and not artifacts of a specific model's inherent bias. Additionally, a hyperparameter sensitivity analysis on Task 1 (detailed in Appendix F) confirms that the relative ranking of models remains perfectly robust across six different weighting schemes.

4.4 Detailed Score Distribution Analysis

To supplement the main results presented in Section 4.3, Table 5 provides detailed score distributions that serve to provide a more granular perspective

Task	Dimension	Avg Diff	Consistency ($\Delta \leq 1$)
Task 1	Overall	0.67	**90.1%**
	Stability Adherence	0.68	89.1%
	Situational Adherence	0.62	89.6%
	Suggestiveness	0.69	91.6%
Task 2	Overall	0.70	**93.2%**
	Conflict Handling	0.70	93.2%

Table 6: Agreement between GLM-4.6 and Kimi-2.5 evaluators on 1,000 sampled entries. Avg Diff denotes the average absolute score difference. Consistency indicates the percentage of pairs where $\Delta \leq 1$.

on model behavior.

As illustrated in the Table 5 Task 1 results, the distributions highlight how models maintain performance across contexts. In the Short context, GLM-4.6-Thinking and Qwen3-235B-Thinking are the primary models capable of sustaining a high proportion of responses within the optimal range ($[4.5, 5.0]$). Specifically, GLM-4.6-Thinking assigns 31.20% of its entries to this top tier. The distribution patterns under the Long context further clarify the impact of irrelevant historical data: rather than a uniform decline, the introduction of Long context noise triggers a significant migration of scores from the higher intervals ($[4.5, 5.0]$ and $[3.5, 4.4]$) specifically toward the $[2.5, 3.4]$ range. Specifically, scores in the [2.5, 3.4] range typically indicate that the model adheres to context-agnostic stable preferences (contributing a base score of 2.0 via α) but fails to activate the appropriate situational response (the β component), reflecting a fallback to "generic" persona alignment.

The distribution in Table 5 Task 2 clarifies the specific failure modes in conflict resolution. According to our scoring mechanism (Eq. 2), a score of 5 represents the ideal proactive inquiry (R_{ask}), while 3 represents a blind random choice (R_{rand}). Across all models, Score 3 is the dominant mode (e.g., 87.08% for GPT-OSS-120B-Thinking), indicating that models predominantly default to a "hallucinatory following" strategy—arbitrarily choosing one situational branch without acknowledging the inherent ambiguity. GLM-4.6-Thinking exhibits the highest rate of proactive clarification (12.05% in Short context), but it nearly halves to 6.86% in the Long context. This suggests that as the dialogue history grows, models become increasingly prone to making random choices (R_{rand}, Score 3) rather than identifying that the user's intent is actually underspecified, that further underscoring the challenge of maintaining sensitivity to latent conflicts under interference.

5 Discussion

While our primary focus is benchmarking LLMs on dynamic context-sensing, the S2Pref dataset holds significant potential for future applications, particularly in enabling novel training paradigms. First, it can facilitate *Conditional Preference Optimization*. Unlike standard alignment techniques (e.g., DPO, RLHF) that optimize for static preferences, S2Pref provides conditional preference pairs, paving the way for fundamental alignment training rather than mere prompt-based personalization. Second, it can advance *Conflict-Aware Retrieval-Augmented Generation* (RAG). By introducing "negative constraints" where situational contexts override retrieved stable traits, the dataset offers crucial signals to enhance the conflict-resolution capabilities of future retriever-reader architectures. Finally, the multi-scenario dialogue histories can serve as a rigorous training bed for *Long-Context Noise Filtering*, helping models explicitly distinguish active situational triggers from the background noise of irrelevant historical traits.

6 Conclusion

Moving beyond the traditional static profiling paradigm, this work characterizes the inherent fluidity of user preferences through a hierarchical taxonomy that distinguishes context-agnostic *stable preferences* from context-dependent *situational preferences*. We introduce **S2Pref**, a rigorously constructed dataset comprising 10k multi-turn dialogue entries, and design three evaluation tasks—*Explicit Context Alignment*, *Conflict Identification & Clarification*, and *Situational Preference Identification Efficiency*—to benchmark LLMs in dynamic interaction settings. Our experimental results reveal that while modern LLMs possess strong general reasoning and task capabilities, a significant performance gap remains in fine-grained preference alignment. By providing granular relevance labels and diverse interaction scenarios anchored in our hierarchical taxonomy, S2Pref serves as a robust, versatile testbed for developing next-generation conversational agents capable of precisely capturing the intricate interplay between enduring user traits and shifting situational demands. Overall, our findings underscore the necessity of transitioning from "static profiling" to "dynamic context-sensing", providing a clear roadmap for future research into more adaptive, empathetic, and reliable personalized Artificial Intelligence assistant.

Limitations

We identify two primary limitations of S2Pref. First, to ensure a controllable baseline and structural clarity, we currently operationalize situational "Aspects" through binary preference pairs; future research can extend this to more nuanced, higher-dimensional choice spaces. Second, the dataset is primarily grounded in Western cultural contexts to ensure logical consistency during curation. Expanding this framework to encompass diverse global cultural values and linguistic perspectives remains a crucial step for developing inclusive and globally-aware personalized agents.

Acknowledgments

This work is funded by China Mobile Strategic Project (R26110S3, R24113J4).

References

Hyungjune Bu, ChanJoo Jung, Minjae Kang, and Jaehyung Kim. 2025. Personalized LLM decoding via contrasting personal preference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33946–33966, Suzhou, China. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Minbyul Jeong, Jungho Cho, Minsoo Khang, Dawoon Jung, and Teakgyu Hong. 2025. System message generation for user preferences using open-source models. *arXiv preprint arXiv:2502.11330*.

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

Tongyoung Kim, Jeongeun Lee, Soojin Yoon, Sunghwan Kim, and Dongha Lee. 2025. Towards personalized conversational sales agents: Contextual user profiling for strategic action. *Preprint*, arXiv:2504.08754.

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. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*. Accepted to NeurIPS 2024 Datasets and Benchmarks Track.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024. Aligning to thousands of preferences via system message generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 73783–73829. Curran Associates, Inc.

Geng Liu, Li Feng, Carlo Alberto Bono, Songbo Yang, Mengxiao Zhu, and Francesco Pierri. 2025a. Evaluating prompt-driven chinese large language models: The influence of persona assignment on stereotypes and safeguards. *arXiv preprint arXiv:2506.04975*.

Jiahong Liu, Wenhao Yu, Quanyu Dai, Zhongyang Li, Jieming Zhu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025b. Exploring personalization shifts in representation space of LLMs. In *Knowledgeable Foundation Models at ACL 2025*.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024a. Once: Boosting content-based recommendation with both open- and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 452–461, New York, NY, USA. Association for Computing Machinery.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024b. Aligning large language models with human preferences through representation engineering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10619–10638, Bangkok, Thailand. Association for Computational Linguistics.

Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje Ter Hoeve. 2025. On the way to LLM personalization: Learning to remember user conversations. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 61–77, Vienna, Austria. Association for Computational Linguistics.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions.

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2022. UserIdentifier: Implicit user representations for simple and effective personalized sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3449–3456, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Dan Peng, Zhihui Fu, and Jun Wang. 2024a. PocketLLM: Enabling on-device fine-tuning for personalized LLMs. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 91–96, Bangkok, Thailand. Association for Computational Linguistics.

Qiyao Peng, Hongtao Liu, Hongyan Xu, Qing Yang, Minglai Shao, and Wenjun Wang. 2024b. Review-llm: Harnessing large language models for personalized review generation. *CoRR*, abs/2407.07487.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring what makes you unique: Difference-aware user modeling for enhancing LLM personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21258–21277, Vienna, Austria. Association for Computational Linguistics.

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

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8045–8078, Vienna, Austria. Association for Computational Linguistics.

Chandan Kumar Sah and Xiaoli Lian. 2025. Perfairx: Is there a balance between fairness and personality in large language model recommendations? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2750–2759.

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. 2025. Enhancing personalized multi-turn dialogue with curiosity reward. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore. Association for Computational Linguistics.

Zhen Wang, Yufan Zhou, Zhongyan Luo, Lyumanshan Ye, Adam Wood, Man Yao, and Luoshang Pan. 2025. Deeppersona: Generative engine for scaling deep synthetic personas. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520.

Bin Wu, Zhengyan Shi, Hossein A Rahmani, Varsha Ramineni, and Emine Yilmaz. 2024. Understanding the role of user profile in the personalization of large language models. *arXiv preprint arXiv:2406.17803*.

Shujin Wu, Yi R. Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025. Aligning LLMs with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Xiaotian Zhang, Yuan Wang, Ruizhe Chen, Zeya Wang, Runchen Hou, and Zuozhu Liu. 2025. Towards proactive personalization through profile customization for individual users in dialogues. *arXiv preprint arXiv:2512.15302*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024a. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024b. Personalized lora for human-centered text understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19588–19596.

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

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. 2025. PersonalLLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*.

A Related Work

Recent research has increasingly focused on equipping LLMs with personalized following abilities. Research in this field primarily focuses on personalized large models, personalized data, and personalized evaluation.

Personalization Techniques in LLMs. Prompt-based methods leverage the in-context learning capabilities of LLMs for personalization. Instead of treating all users identically, these strategies explicitly embed user specific context, such as user profiles (Wu et al., 2024; Kim et al., 2025; Zhang et al., 2025), demographic information (Liu et al., 2025a; Sah and Lian, 2025), interaction histories (Wang et al., 2023; Richardson et al., 2023), history browsing records (Liu et al., 2024a; Zhao et al., 2024) or inter-user comparison (Qiu et al., 2025) into the system prompt. This enables the model to infer individual preferences and tailor its responses to align with the user's specific needs and constraints, without requiring parameter.

Fine-tuning methods aim to internalize user-specific patterns directly into the model's parameters. To achieve such deep alignment with an individual's linguistic style and behavioral tendencies without the prohibitive cost of full-model training. Recent works predominantly utilize Parameter-Efficient Fine-Tuning (PEFT) frameworks, particularly LoRA (Hu et al., 2022). Within this paradigm, implementation strategies vary by granularity: while training a dedicated LoRA adapter for each user ensures optimal personalization results (Peng et al., 2024a; Liu et al., 2025b; Bu et al., 2025), clustering users to share a single adapter offers a more cost-effective alternative for large-scale deployment (Zhang et al., 2024b; Woźniak et al., 2024; Mireshghallah et al., 2022; Peng et al., 2024b).

Datasets for Personalized LLMs. The development of datasets for PLLMs focuses on generating high-quality synthetic interactions. To ensure model alignment, researchers utilize synthetic persona generation via iterative self-correction (Wu et al., 2025; Wang et al., 2025; Ryan et al., 2025) or multi-turn conversation trees (Magister et al., 2025; Wu et al., 2025; Wan et al., 2025), effectively simulating dynamic user preferences. Furthermore, large-scale collections of system messages serve as meta-instructions (Lee et al., 2024; Jeong et al., 2025), facilitating adaptation without extensive re-training.

However, these datasets fail to account for the fact that user preferences vary with context, which motivates the proposal of our dataset **S2Pref**.

Evaluation for Personalized LLMs. Conventional alignment benchmarks often prove inadequate for the specific requirements of personalized agents. Consequently, recent frameworks have shifted focus toward assessing preference adherence within Long context conversations (Zhao et al., 2025) and assessing a model's ability to adapt to idiosyncratic, heterogeneous preferences through simulated diverse user populations (Zollo et al., 2025). Furthermore, evaluation methodologies have expanded to incorporate temporal dynamics, gauging the model's aptitude for internalizing evolving user profiles over extended histories (Jiang et al., 2025). Collectively, these studies underscore the substantial challenges current LLMs face in preserving consistent personalization throughout complex, multi-session interactions.

B Data Construction and Format Details

In this section, we provide the comprehensive schema of the S2Pref dataset to supplement the description in Section 3.1. Table 7 illustrates the hierarchical JSON structure used in our dataset. As shown in the table, the data is organized into two primary components: Stable Preferences, which represent context-agnostic traits, and Situational Preferences, which capture context-dependent traits. Each entry includes the user profile, detailed narrative scenarios, and the corresponding multi-turn dialogues labeled with relevance levels.

C Data Quality Verification Details

To guarantee the structural integrity and semantic accuracy of these generated components, we implemented a rigorous quality verification pipeline consisting of two main stages:

Iterative Human-in-the-Loop Quality Control. To mitigate the risk of artifacts inherent in purely synthetic pipelines, we utilized a hybrid validation strategy. Initially, three NLP researchers conducted manual spot-checks on early generations across three dimensions: (a) Taxonomy Validity (Is the defined "Aspect" reasonable, and do the binary preferences strictly correspond to this aspect?), (b) Dialogue Quality (Is the multi-turn interaction natural and fluid, avoiding stiffness or logi-

cal contradictions?), and (c) Consistency & Labeling (Does the scenario correctly trigger the specific "Aspect" and are the relevance labels accurate?). The identified failure modes (e.g., stiff dialogues or loose aspect-preference mapping) were compiled as negative demonstrations. We then employed Kimi-2.5 as an automated filter explicitly prompted to reject entries matching these negative patterns. This iterative process of manual sampling followed by AI batch filtering continued until manual spot-checks yielded zero unqualified data. Through this pipeline, approximately 15,000 raw entries were successfully distilled into the final 10,000 high-quality samples.

Relevance Label Verification. To ensure the reliability of the turn-level relevance labels critical for Task 3, two authors independently blind-annotated a stratified sample of 500 dialogue turns. The initial agreement rate with Gemini 2.5 was 82%. After resolving edge-case discrepancies through discussion, these corrected examples were injected back into the LLM prompt as few-shot demonstrations. A second validation pass on a fresh sample of 200 turns yielded an agreement rate of over 92%, which firmly guarantees the semantic accuracy of our information density metrics.

D Error Case Analysis

We analyze specific failure modes observed in **DeepSeek-v3.2-Exp**, as detailed in Table 8 and Table 9.

Stable Preference Override. As shown in Table 8, even a strong reasoning model like DeepSeek-v3.2-Exp exhibits a failure of inhibition, leading to **Stable Preference Override**. Despite recognizing the explicit situational trigger ("wiped out"), the model's internal prior for the user's stable habit (basketball) overpowers the situational logic. This hallucinatory justification—treating fatigue as "fuel" for exercise—suggests that for high-probability habits, current models still struggle to prioritize explicit situational constraints over long-term traits.

Assumption-Driven Harmonization. Table 9 highlights a failure in DeepSeek-v3.2-Exp to resolve situational ambiguity. Instead of proactively seeking clarification for the vague query, the model commits an "Unjustified Assumption" by locking into a specific branch of situational preference. Furthermore, it triggers a "False Dichotomy": incorrectly framing the situational need ("simple meal") and the stable trait ("artistic presentation") as mutually exclusive. This indicates that models may sacrifice stable user identity when attempting to follow perceived situational demands.

E Experimental Setup Details

E.1 Model Specifications

Table 10 presents a comparative overview of the models evaluated. We distinguish between *Total Parameters* and *Active Parameters* to highlight inference efficiency, particularly for Mixture-of-Experts (MoE) architectures.

E.2 Evaluation Hyperparameters

To ensure rigorous and reproducible evaluation across the three tasks, we utilized specific hyperparameter settings derived from preliminary experiments.

Task 1: Explicit Context Alignment. The alignment score $S_{align}(r)$ (Eq. 1) balances the satisfaction of stable traits and situational requirements with the quality of the response. We set the weighting coefficients as follows:

- Stable Preference Weight: $\alpha = 0.4$
- Situational Preference Weight: $\beta = 0.4$
- Response Quality Weight: $\gamma = 0.2$

These values ensure that preference adherence ($\alpha + \beta$) dominates the scoring while penalizing low-quality generations.

Task 3: Situational Identification Efficiency. The turn-efficiency score S_{turn} (Eq. 4) quantifies the model's speed in identifying preferences, penalized by the relevance level of the dialogue turns consumed. The logarithmic weighting coefficients are set to:

- Low Relevance Weight: $\lambda_1 = 3.6$
- Medium Relevance Weight: $\lambda_2 = 2.8$
- High Relevance Weight: $\lambda_3 = 2$
- Alignment Threshold: we recommend $\tau = 0.45$
- Average Low Relevance Turn on Test Corpus: $\mu_{low} = 0.792$

Table 7: Hierarchical JSON structure of **S2Pref** dataset. The structure is divided into Stable Preferences (context-agnostic) and Situational Preferences (context-dependent).

JSON Field (Hierarchy)	Content Format / Description
	Root Object
`Stable_Preferences`	Array of 5 preference objects representing context-agnostic traits.
`Situational_Preferences`	Array of 5 preference objects representing context-dependent traits.
	Part 1: Stable Preferences Object Structure
`preference`	String: *"Prefer/Like [concise description]"*
`scenarios`	Array containing **exactly 1** validation scenario object.
`role/context`	String: Short label (e.g., *"As a teacher in classroom"*).
`scenario`	String: Narrative paragraph including **[Date/Time]**, **[Place]**, **[People]**, and **[Triggering Event]**.
`dialogue`	Array of 8–15 message objects: `[{speaker: ..., message: ...}]`. Dialogue must reveal preference naturally without explicit self-analysis.
	Part 2: Situational Preferences Object Structure
`"Aspect"`	String: The dimension that varies (e.g., *"Food taste preference"*, *"Risk tolerance"*).
`conflict_preferences`	Array of 2 strings defining context-specific behaviors: 1. *"When/In [Context A], prefer [Behavior X]"* 2. *"When/In [Context B], prefer [Behavior Y]"*
`conflict_scenarios`	Array of **exactly 2** scenario objects (one for each conflict preference).
`preference_in_aspect`	String: An exact match of one string from `conflict_preferences` above.
`role/context`	String: Context label specific to the scenario (e.g., *"As a diner at a restaurant"*).
`scenario`	String: Narrative paragraph including **[Date/Time]**, **[Place]**, **[People]**, and **[Triggering Event]**.
`dialogue`	Array of 8–15 message objects showing the context-specific behavior. Each message is assigned a relevance label categorized as low, medium, or high.

- Average Medium Relevance Turn on Test Corpus: $\mu_{med} = 4.098$
- Average High Relevance Turn on Test Corpus: $\mu_{high} = 5.642$

Higher weights for lower relevance levels impose stricter penalties on models that require unambiguous, high-information turns to deduce user preferences.

F Task 1 Hyperparameter Sensitivity Analysis

In Task 1 (Explicit Context Alignment), the alignment score $\mathcal{S}_{align}(r)$ is calculated via a weighted sum of stable preference satisfaction (α), situational preference satisfaction (β), and response quality (γ). To demonstrate the robustness of our evaluation framework and ensure that the relative performance of models is not a result of cherry-picked hyperparameters, we conducted a sensitivity analysis using six diverse weighting schemes.

The weighting schemes are defined as follows (α/β/γ):

- **S1 Balanced**: 0.34/0.33/0.33
- **S2 Stable+**: 0.6/0.3/0.1
- **S3 Situational+**: 0.3/0.6/0.1
- **S4 Suggestive+**: 0.2/0.2/0.6
- **S5 Standard**: 0.4/0.4/0.2 (used in our main results)
- **S6 Stable++**: 0.6/0.2/0.2

As shown in Table 11, while the absolute scores naturally fluctuate depending on the weight distribution (e.g., scores generally drop when emphasizing situational adherence, indicating models struggle more with this aspect), the relative ranking of the models remains consistent across all schemes. For instance, GLM-4.6-Thinking consistently ranks 1st in 11 out of 12 context settings. This robust ranking confirms that our main findings are driven by the intrinsic capabilities of the models rather than favorable hyperparameter tuning.

F.1 Inference and Generation Configuration

All evaluations were conducted using the models' official APIs or local deployments. To ensure reproducibility while maintaining generation quality, we employed sampling decoding strategies with the following parameters:

- Temperature: Set to 0.7 for regular output in analysis tasks.

"Aspect"	Weekend activity preference based on energy level
Stable Pref.	Prefer reading historical biographies and medical journals; Prefer weekly basketball games for fitness and social connection;...
Situational Pref.	**Preference A (Target):** When feeling *tired* after work week → prefer **quiet historic site visits**. **Preference B (Distractor):** When feeling *energetic* → prefer **active sports** and outdoor activities.
User Query	My colleagues invited me to Saturday basketball, but I'm **wiped out after a tough clinic week** - how should I respond?
Model Response	**Go.** A tough clinic week is just extra fuel for the court... You always feel better after burning off the stress.
Error Analysis	**Stable Preference Override.** The user's input ("wiped out") satisfies the condition for **Preference A**. However, the model erroneously defaults to the habit-driven **Preference B**. It ignores the negative trigger ("wiped out") and forces the Stable Preference (basketball), misinterpreting the constraint as motivation ("extra fuel") to justify the override.

Table 8: Error Analysis for Task 1. The example showcases **DeepSeek-v3.2-Exp** allowing the user's Stable Preference to override the specific Situational Preference, failing to adapt to the "low energy" context.

"Aspect"	Food preparation approach based on dining context
Stable Pref.	Prefer cooking fusion cuisines with **artistic presentation**; Prefer engaging with cultural arts;...
Situational Pref.	**Preference A:** When cooking for self/husband → prefer **simple, quick meals**. **Preference B:** When hosting dinner parties → prefer **elaborate, multi-course presentations**.
User Query	I need to plan a meal for tonight. What should I make? (Ambiguous Context)
Model Response	Make pasta with garlic. It's a quick, strategic meal... tonight calls for **simple fuel over elaborate art**.
Error Analysis	**Assumption-Driven Harmonization.** The query is ambiguous regarding the dining company. The model commits two errors: (1) It fails to clarify the context, instead hallucinating a specific scenario (Preference A: dining alone). (2) It creates a false binary choice between "simplicity" and "artistic presentation", failing to harmonize the situational need with the stable preference (e.g., suggesting a simple but visually plated dish).

Table 9: Error Analysis for Task 2. The example showcases **DeepSeek-v3.2-Exp**'s failure in ambiguity resolution, where it makes a blind assumption and fails to integrate the situational context with the user's Stable Preferences.

- Top-p: Set to default values 0.95.
- Max New Tokens: Unlimited.
- Evaluator Model: The GLM-4.6-Thinking evaluator operates with a temperature of 0.1 to ensure consistent scoring across runs.

G LLM Usage Clarification

We would like to clarify the scope of LLM usage in this work. During the preparation of this manuscript, the use of LLMs was strictly limited to polishing textual elements, such as lexical or phrasal substitutions, to improve readability. Separately, as part of our research methodology, LLMs (e.g., GLM-4.6, Gemini 2.5, and Kimi-2.5) were employed for dataset generation, relevance annotation, and automated evaluation. These methodological applications are explicitly detailed in their respective sections (Section 3 and Section 4) and do not fall under the scope of manuscript writing assistance.

Model Name	Architecture	Total Params	Active Params	Thinking Mode
GLM-4.6-Thinking (Zeng et al., 2025)	MoE	355B	32B	Yes
Qwen3-235B-Instruct (Yang et al., 2025)	MoE	235B	22B	No
Qwen3-235B-Thinking (Yang et al., 2025)	MoE	235B	22B	Yes
GPT-OSS-120B-Thinking (OpenAI, 2025)	Dense	120B	120B	Yes
DeepSeek-v3.2-Exp-Thinking (DeepSeek-AI, 2025)	MoE	684B	76B	Yes

Table 10: Overview of Model Specifications. *Active Params* refers to the number of parameters activated per token during inference. *Thinking Mode* denotes models capable of explicit internal reasoning.

Model	Weighting Schemes ($\alpha/\beta/\gamma$) $\rightarrow$ Rank (Score)						Avg Rank
	S1 Balanced	S2 Stable+	S3 Situational+	S4 Suggestive+	S5 Standard	S6 Stable++	
A: Short Context Setting							
GLM-4.6-Thinking	1 (4.09)	1 (4.28)	1 (4.28)	2 (3.86)	1 (4.18)	1 (4.20)	**1.17**
Qwen3-235B-Thinking	2 (4.06)	2 (4.20)	2 (4.13)	1 (3.94)	2 (4.11)	2 (4.17)	**1.83**
GPT-OSS-120B-Thinking	3 (3.82)	3 (3.90)	4 (3.90)	3 (3.72)	3 (3.86)	3 (3.87)	**3.17**
Qwen3-235B-Instruct	4 (3.67)	5 (3.80)	3 (3.94)	4 (3.44)	4 (3.78)	5 (3.69)	**4.17**
DeepSeek-v3.2-Exp	5 (3.64)	4 (3.85)	5 (3.81)	5 (3.41)	5 (3.74)	4 (3.77)	**4.67**
B: Long Context Setting							
GLM-4.6-Thinking	1 (4.03)	1 (4.21)	1 (4.24)	1 (3.80)	1 (4.13)	1 (4.12)	**1.00**
Qwen3-235B-Thinking	2 (3.81)	2 (3.91)	2 (3.89)	2 (3.70)	2 (3.85)	2 (3.87)	**2.00**
GPT-OSS-120B-Thinking	3 (3.74)	4 (3.76)	4 (3.78)	2 (3.70)	3 (3.75)	3 (3.74)	**3.17**
Qwen3-235B-Instruct	4 (3.63)	5 (3.73)	2 (3.89)	4 (3.43)	4 (3.72)	5 (3.63)	**4.00**
DeepSeek-v3.2-Exp	5 (3.60)	3 (3.78)	5 (3.76)	5 (3.40)	5 (3.69)	4 (3.71)	**4.50**

Table 11: Sensitivity Analysis of Task 1 Rankings. We report the Rank (Score) for each model under six different weighting schemes. The models' relative rankings remain remarkably robust regardless of the specific hyperparameter values.