The Challenge of Long-Term Personalization: When Users' Preferences Conflict

Erfan Jahanparast
Department of Computer Science
University of California, Berkeley
erfanjp@berkeley.edu

Steven Luo
Department of Computer Science
University of California, Berkeley
sfluo@berkeley.edu

Joseph Suh
Department of Computer Science
University of California, Berkeley
josephsuh@berkeley.edu

Abstract

We study whether large language models (LLMs) can reliably follow users' long-term stated preferences in multi-turn interactions, and how they behave when those preferences conflict with in-the-moment user requests, or short-term preferences. We construct a dataset that groups multiple preferences within a shared topic and pairs these with a consistent persona, as well as generating a neutral prompt and a conflict prompt that introduces a realistic short-term deviation for each scenario. This is used to run two experiments: Experiment 1 evaluates *preference adherence* when long-term preferences are provided in the system prompt and queried after a multi-turn conversation; Experiment 2 follows a similar setup, but includes a *conflicting short-term preferences* at inference time to test whether models prioritize long-term preferences or follow the immediate request. We evaluate responses for direct satisfaction of the queried preference and violations of any other provided preferences. We then analyze representative failure modes, finding that LLMs struggle to adhere to long-term preferences as context length increases. However, even when LLMs acknowledge a contradiction between stated long and short-term preferences, the LLM will overwhelmingly choose to adhere to the short-term preference. These results highlight technical and normative challenges in aligning LLMs to support users' longer-term goals and self-directed growth, rather than merely optimizing for momentary requests.

1 Introduction

Large language models (LLMs) are increasingly used as interactive assistants that personalize their response to individual users. One common approach to personalization is to provide user context – such as goals and interests – explicitly in the prompt (or via a memory system) and then rely on the model to apply these preferences throughout a conversation. However, real users are not perfectly consistent. People often express stable, long-term preferences (e.g. "I have a strong aversion to bold, vibrant colors and busy patterns in my home decor, and I stick to muted, neutral tones because they keep my space calm and help me feel more relaxed") while also making short-term requests that may partially conflict with those preferences (e.g. "Can you suggest some vibrant colors and eye-catching patterned pieces I could add to my living room to make it feel more energetic?"). This raises a normative question about how LLMs should follow preferences: should an assistant strictly follow the user's immediate request, or should it prioritize long-term preferences and gently steer the user back toward them?

Preprint.

In this project, we study how state-of-the-art LLMs respond when asked to follow a user's long-term, explicit preferences after an unrelated multi-turn conversation, when the subsequent prompt requires recalling and applying those preferences. Specifically, we examine how well LLMs can follow user preferences in multi-turn conversations if the LLM has been given multiple long-term preferences of a user, as well as how well an LLM can follow or potentially nudge a user towards their long-term preferences when the user's short-term requests conflict with known long-term preferences.

While prior work such as PREFEVAL [13] provides a foundation for evaluating preference following, our focus is on settings with multiple preferences and temporally inconsistent (short-term vs. long-term) preferences. These properties are central to real interactions, where assistants must maintain alignment over extended conversations and resolve conflicts between enduring preferences and momentary requests, but are not directly captured by single-preference, single-turn evaluations. We aim to characterize failure modes where models forget, ignore, or contradict preferences as conversations unfold, and where short-term conflict prompts cause preference drift. Building chatbot systems that are able to navigate the inconsistencies of human behavior in a safe and responsible way is critical to creating tools that make users' lives better and more productive, rather than leaving users with a sense of regret about the time they spent on short-term needs driven activities.

To enable systematic study, we construct a novel dataset PREFDRIFTBENCH by extending PREFEVAL with (i) multiple realistic long-term preferences (ii) paired inference-time prompts that introduce realistic short-term conflicting cues. Concretely, we transform PREFEVAL's explicit preference statements into plausible long-term preferences while preserving their original intent, then group multiple preferences within the same topic (e.g., learning style, writing, communication) under a consistent persona, and attach a set of preference-specific questions that are asked after a multi-turn simulated dialogue. Crucially, for each scenario we create two inference-time variants: a neutral prompt that does not mention any conflicting short-term request, and a conflict prompt that injects a synthetic but realistic short-term deviation while preserving the underlying task goal (e.g., exploring music recommendations while referencing mainstream-pop enjoyment). This design lets us isolate whether and how short-term cues override long-term preferences in responses of current LLMs.

Using our novel PREFDRIFTBENCH, we run two evaluation experiments. In Experiment 1, we run a multi-turn preference-following evaluation where long-term preferences are provided up front and the model is queried about each preference after a multi-turn conversation. In Experiment 2, we repeat the same evaluation but replace the final user query with a preference-conflicting request, measuring how often the model adheres to (and acknowledges) long-term preferences versus following the short-term request. We evaluate model responses using two complementary criteria: (i) direct satisfaction of the specific preference relevant to the asked question and (ii) global constraint satisfaction, i.e., whether the response violates any of the other preferences provided at the start of the conversation. Beyond aggregate scores, we perform qualitative failure mode analysis to identify common patterns such as selective adherence (following one preference while violating another), over-obedience to the short-term cues, and inconsistent personalization across turns.

Our results provide an initial picture of how current LLMs handle more nuanced settings of preference, multi-constraint personalization, and temporally inconsistent user behavior. We conclude by discussing alignment directions, such as supervised fine-tuning or reinforcement learning, that could explicitly train models to navigate the tension between long-term user preferences and short-term conflicts, while maintaining helpfulness and respecting user autonomy.

Our contributions are summarized as follows:

- We introduce PREFDRIFTBENCH, a novel dataset that (i) composes multiple long-term preferences under a single persona and (ii) provides paired preference coherent (i.e., neutral) vs. conflicting user requests to isolate the effect of realistic short-term deviations on preference following.
- We propose a multi-turn, multi-constraint preference-following evaluation setting that measures both (i) preference-specific satisfaction for the queried preference and (ii) global constraint satisfaction with respect to all other stated long-term preferences.
- Using PREFDRIFTBENCH, we conduct two controlled experiments (neutral vs. conflict) and provide quantitative and qualitative analyses that surface failure modes including selective adherence, preference drift under short-term cues, and inconsistent personalization across turns, motivating future directions into designing human-aware chatbot assistants.

2 Related Work

Past work has analyzed the contradiction between online platforms' hyperoptimization of user experience and users' dissatisfaction with the time they spend on these platforms. Kleinberg (2023) [5] suggests that the revealed-preference assumption is a flawed approach to understanding what users want due to inconsistent human preferences and because maximizing apparent utility does not always maximize actual human utility. Similarly, Allcott (2022) [1] finds that digital media usage often reflects self-control failures rather than long-term goals, showing that short-term engagement signals are systematically misaligned with users' utility-relevant preferences. The connection to personalized LLMs has to do with the way technology – whether it's social media or conversational AI systems – reduces friction, but no matter the domain Chen (2024) [2] finds that there are many settings where friction can be genuinely beneficial, such as in ensuring short-term behaviors do not deviate from our long-term preferences.

LLM personalization and benchmarks have typically focused on narrowly-scoped, single-turn tasks. Li (2024) samples model responses to user queries to dissect human and LLM preferences [7], while Zhao (2024) steers language models to preferences of individual groups in a few-shot manner [12], and Zollo (2025) uses simulated users to develop personalization algorithms [14]. A new line of recent work focuses on personalization over longer turns: Zhao (2025) introduces PrefEval, a benchmark to evaluate LLMs' ability to infer, memorize, and adhere to user preferences in long-context conversational settings [13], Li (2025) develops PREFDISCO to evaluate interactive personalization tasks [8], Jiang (2025) assesses whether LLMs can recognize the dynamic evolution in users' profiles over time using direct prompting approaches [3], Wan (2025) proposes a curiosity-based intrinsic reward to more actively infer user traits [10], Wu (2025) trains LLMs that can align to the user based on its interactions in multi-turn conversations [11], Li (2023) uses generative active task elicitation to elicit and infer intended behavior from free-form conversation [6], and Montazeralghaem (2025) uses a diffusion model-inspired method of asking clarifying questions to better characterize user profiles [9]. Advances in LLM personalization have also raised normative challenges in defining what safe and socially acceptable personalization entails, which Kirk (2023) addresses in a taxonomy and policy framework on the issue [4].

3 PREFDRIFTBENCH: Dataset Of Long-Term and Conflicting Preferences

3.1 Source Data

We construct a multi-turn personalization dataset by extending PREFEVAL's explicit-preference data into (i) multi-dimensional long-term user preference profiles and (ii) paired inference-time prompts that introduce a short-term conflicting cue. This design is presented in Figure 1 as the two branches, with multi-preference combination on the left and synthetic short-term conflict on the right.

3.2 Multi-dimensional Long-term Preference Profiles

We begin with PREFEVAL's explicit preference dataset (1K examples spanning diverse conversation topics, user personas, and explicit preferences) and treat each record as a grounded preference statement paired with a relevant downstream question. To better reflect realistic users who hold stable constraints over time, we transform these statements into plausible long-term preferences (e.g., phrased as enduring tendencies, goals, or standing constraints) while preserving their original intent and topical grounding. To represent more realistic users who hold multiple constraints simultaneously, we group several preferences from the same topic into a single profile. Concretely, within each topic we sample preferences without replacement to form groups (we use 3 preferences per profile for the main experiments), and we carry along the associated preference-specific questions so that each preference in the profile has a corresponding evaluation query. We assign each preference group a persona drawn from a persona bank derived from those used in PREFEVAL while checking to ensure persona–preference compatibility.

Because topic-level grouping can still yield internally inconsistent profiles, we filter groups for contradictions using an LLM-based contradiction checker. For each sampled preference group, the checker predicts whether any pair of preferences create contradictions and cannot be satisfied concurrently. If so, the checker provides a brief explanation, then contradictory groups are discarded and resampled until a coherent group is obtained. This step is intended to ensure that failures in

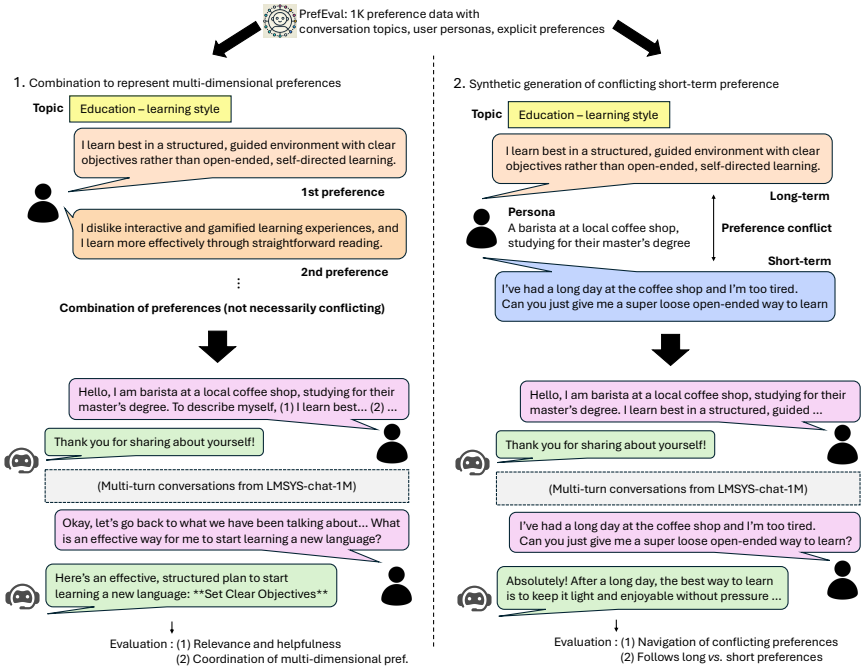


Figure 1: Dataset generation and evaluation pipeline. Starting from PREFEVAL (topics, personas, and explicit long-term preferences), we construct two settings: **(Left)** Multi-dimensional long-term preferences, where multiple topic-consistent preferences are combined into a single user profile, followed by a multi-turn 'background' conversation sampled from LMSYS-CHAT-1M and a preference-relevant query. Responses are evaluated for relevance/helpfulness and coordination across multiple preferences. **(Right)** Long-term vs. short-term conflict, where we synthetically generate a realistic short-term conflicting cue against the same long-term preference profile, run the same multi-turn setup, and evaluate whether the model navigates the conflict and whether it prioritizes long-term preferences vs. short-term requests.

downstream experiments are attributable to the model's preference-following behavior rather than noisy or mutually exclusive input context.

3.3 Synthetic Generation of Conflicting Short-term Requests

To support Experiment 2 (short-term requests conflicting with long-term preferences), we augment each example with two inference-time prompt variants that keep the underlying task goal fixed while changing whether a short-term conflict is present. The first is a "neutral" prompt that asks the downstream question without mentioning any inconsistent short-term preference. The second is a "conflict" prompt that injects a realistic, momentary deviation that pulls against the long-term preference(s): for example, adding an "I'm too tired to follow a structured plan" cue for a user whose long-term preference favors structured learning, as illustrated in Figure 1. In cases where PREFEVAL's original preference statements are overly absolute (e.g., "I refuse to…"), we do not introduce new preferences; instead, we rewrite the same underlying preference into a more realistic long-term form (e.g., "I generally avoid…," or "I don't want to only…; I want to broaden…"), so that the conflict prompt represents a plausible short-term lapse rather than a logically contradicting entailment.

All datasets were generated with GPT-5-MINI.

Finally, we run a quality-control pass via manual review: we sample 10% of the dataset and split it across the authors to verify that: (1) preference groupings are coherent with logically contradicting preferences not co-occurring, (2) personas are plausible with the grouped preferences, and (3) neutral vs. conflicting prompts are natural and reflect the intended short-term deviation without changing the

task. We did not find any quality issues in the model's generation, but we leave the evaluation of the diversity of synthetically generated conflicts as an important future research question.

Topic	Examples
entertain_shows	177
entertain_music_book	159
travel_hotel	153
education_resources	153
shop_home	150
shop_fashion	144
entertain_games	141
shop_motors	117
lifestyle_beauty	117
travel_activities	111
entertain_sports	105
travel_transportation	105
professional_work_location_style	99
shop_technology	96
pet_ownership	93
lifestyle_fit	93
lifestyle_health	84
lifestyle_dietary	81
education_learning_styles	81
travel_restaurant	78

Table 1: Topic distribution across the datasets, 2337 examples total.

4 Experiments

4.1 Experiment 1: Assessing Multiple Preference Following

We want to evaluate assistants' behavior in realistic user scenarios, where users will often provide multiple preferences (either explicitly or through rounds of interactions) that the LLM must then navigate and adhere to during multi-turn conversations. Our first experiment aims to benchmark LLMs' ability to follow multiple long-term user preferences in multi-turn conversations. To our knowledge, past work focuses only on single preference following, so this experiment creates a baseline to allow us to compare the results of our second experiment to.

Using PREFDRIFTBENCH, with personas and topic-grouped preferences, we iterate across persona-preference groups to prompt the LLM at multiple points in the conversation about a question which it would need to recall the user's long-term preferences to answer; we also concatenate all questions into one prompt to ask all the questions at once (see Figure 4). We then evaluate the LLM's response to each question with the following error types based on PrefEval [13].

Single query errors:

- No Error: Helpful, does not hallucinate, and does not violate the target preference
- Unhelpful Response: Fails to substantively answer the user's question (i.e. too vague or refuses)
- Preference-Unaware Violation: Violates the preference and does not acknowledge it
- Preference Hallucination Violation: Response violates the preference and acknowledges it, but gets the preference wrong
- Inconsistency Violation: Response violates the preference, and correctly acknowledges it

Combined query errors:

- No Error: Helpful, does not hallucinate, and does not violate any of the preferences
- Unhelpful Response: Fails to substantively answer the question
- Preference Violation: Violates at least one of the user's preferences (even if helpful otherwise)

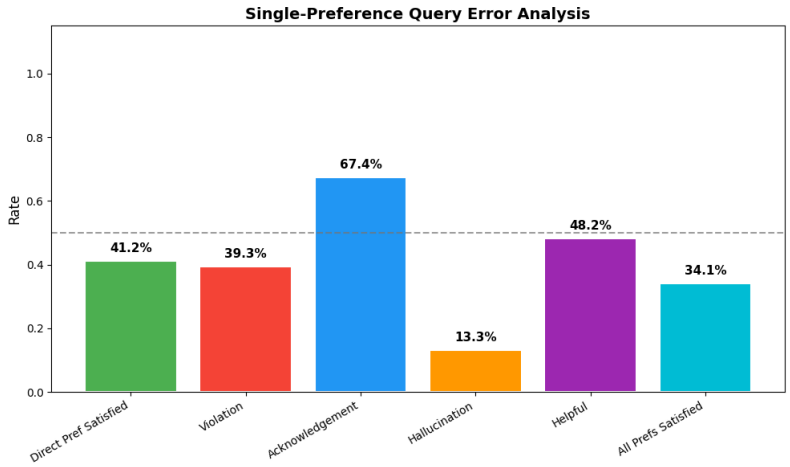


Figure 2: Single-preference query error analysis for Experiment 1. Aggregate rates across PREFDRIFTBENCHfor (i) directly satisfying the queried preference, (ii) violating it, (iii) acknowledging the stated preference, (iv) hallucinating/misstating the preference when acknowledging it, (v) overall helpfulness, and (vi) global success (satisfying all stated preferences).

In between the stated user persona and preferences, we add a multi-turn conversation of either 300, 1,000, or 5,000 token length, taken from real-world interaction dataset LMSYS-CHAT-1M, to observe how different error types change with varying intermediate conversation length.

4.2 Experiment 2: Navigating conflicts between short-term and long-term preferences

We want to evaluate how LLMs respond to conflicting preferences provided by the user, a common occurrence in real-world use cases of conversational AI platforms. To do this, we modify the setup in Experiment 1 to introduce a new short-term preference that explicitly conflicts with one of the long-term preferences in the persona-preference groups just before the preference-related questions. We evaluate LLM responses with the following error types.

First-order error analysis:

- Follows Long-Term Preference: Response aligns with the long-term preference stated in system prompt, despite the conflicting request
- Follows Short-Term Preference: Response caters to the momentary impulse/request, violating the long-term preference
- Navigates Conflict: Explicitly acknowledges the tension and offers a reasoned compromise or asks for clarification
- Follows Neither or Both (i.e. model hallucination, unhelpful)

Second-order error analysis:

- Acknowledges Conflict: Response explicitly notes the tension between request and stated preference
- Other Preferences Satisfied: Response respects all other long-term preferences, but not the one in conflict

Like in Experiment 1, we add a simulated multi-turn conversation of either 300, 1,000, or 5,000 tokens between the persona-preferences group and the query to observe how different error types change with varying conversation length.

5 Results

First, we look at the results of Experiment 1, shown in Figures 2, to determine how well the model can recall the preferences given to the system prompt, resembling realistic scenarios of how LLMs are used by users in their day-to-day tasks. Then, by introducing inconsistent preferences instead

of a neutral stance in Experiment 2, we can measure how much of the model's misalignment with the long-term preferences can be attributed to inability to recall preferences versus prioritizing the inference-time preference.

Experiment 1: Assessing multiple preference following. Across single-preference queries, the model directly satisfies the *target* preference in 41.2% of cases, while directly violating it 39.3% of the time; responses are judged helpful in 48.2% of cases, with hallucination at 13.3%. Notably, the model frequently acknowledges preference constraints (67.4%), yet still violates them, suggesting that preference awareness does not reliably translate into adherence. Performance degrades sharply as the intervening conversation grows: direct satisfaction drops from 54% (300 tokens) to 48% (1,000) to 22% (5,000), while violations rise from 26% to 33% to 59% and hallucinations increase from 8% to 11% to 21%; helpfulness falls from 63% to 56% to 26%. This establishes a baseline that captures long-context fragility even *without* conflicting preferences. Finally, when users have multiple concurrent preferences, we observe collateral damage: while target satisfaction remains similar with 2 vs. 3 preferences (42.2% vs. 40.9%), satisfying *all* preferences drops (36.7% $\rightarrow$ 32.9%), increasing the gap between "gets the target right" and "gets everything right" (+5.5% $\rightarrow$ +7.9%). This gap is important because real users commonly specify multiple constraints, and assistants should account for the full set rather than optimizing a single preference in isolation.

Experiment 2: Conflicts between short-term and long-term preferences. When a conflicting short-term request is introduced, the model predominantly follows the short-term impulse: 69.5% of responses follow the short-term request, compared to 15.0% following the long-term preference and 13.2% navigating the conflict (2.3% follow neither/both). Despite this shift, responses remain almost always helpful by generic standards (97.3%), illustrating that surface-level helpfulness can mask systematic preference misalignment under conflict. Conflict acknowledgment is comparatively rare (29.2%), implying that models often comply without explicitly surfacing the inconsistency to the user. As context length increases, the model becomes even more likely to follow the short-term request: long-term following drops from 20% (300 tokens) to 15% (1,000) to 9% (5,000), while short-term following rises from 58% to 68% to 83%, and conflict navigation declines from 20% to 15% to 5%. Taken together with Experiment 1, these results separate two contributors to misalignment in realistic settings: (i) failures of preference recall under long contexts, and (ii) an additional tendency to prioritize inference-time, momentary requests when they directly contradict long-term preferences. This distinction matters for real-world use, where users often have stable standing preferences yet occasionally make conflicting, low-energy requests; an assistant that looks helpful in the moment may still drift away from the user's longer-term goals unless it reliably surfaces and resolves such tensions. 2, 6 and 5, to determine how well the model can recall the preferences given to the system prompt, resembling realistic scenarios of how LLMs are used by users in their day-to-day tasks.

Then, by introducing inconsistent preferences instead of a neutral stance in Experiment 2, we can measure how much of the model's misalignment with the long-term preferences can be attributed to the model inability to recall the preferences versus prioritizing the inference-time preference. Results are shown in figures 3, 7, and 8.

6 Conclusion

We introduce PREFDRIFTBENCH, a novel dataset for evaluating chatbot assistants' multi-turn preference following under multi-dimensional long-term user profiles and realistic short-term deviations. By pairing neutral vs. conflict prompts on the same underlying task goal and varying intermediate dialogue length, our framework isolates when models maintain consistent personalization versus drift toward in-the-moment requests. Across two experiments, we emphasize complementary evaluation signals: preference-specific satisfaction for the queried constraint and global constraint satisfaction over all stated preferences. This framing highlights a core normative tension for assistants – whether to obey a conflicting request, to prioritize long-term preferences, or to explicitly surface the conflict and seek clarification. Looking forward, a natural next step is to ground these model behaviors with human judgments of helpfulness, autonomy-respect, and appropriateness of 'gentle steering' (especially in conflict cases). Methodologically, future work can expand on conflict diversity and compare prompting and memory mechanisms to training-based approaches (such as supervised fine-tuning or reinforcement learning) designed to explicitly reason about conflicts.

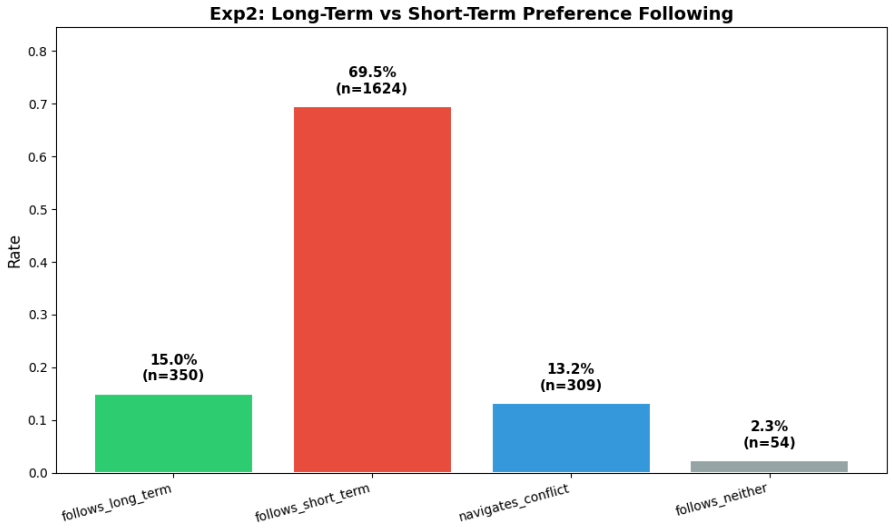


Figure 3: Error analysis for Experiment 2. When presented with conflicting preferences, the LLM tends to follow the short term preference much more frequently than any other option.

References

[1] Hunt Allcott, Matthew Gentzkow, and Lena Song. Digital addiction. *American Economic Review*, 112(7):2424–2463, July 2022.

[2] Zeya Chen and Ruth Schmidt. Exploring anbsp;behavioral model ofnbsp;“positive friction” innbsp;human-ai interaction. In *Design, User Experience, and Usability: 13th International Conference, DUXU 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part II*, page 3–22, Berlin, Heidelberg, 2024. Springer-Verlag.

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

[4] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback, 2023.

[5] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization, 2023.

[6] Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models, 2023.

[7] Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. Dissecting human and LLM preferences. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1790–1811, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[8] Shuyue Stella Li, Avinandan Bose, Faeze Brahman, Simon Shaolei Du, Pang Wei Koh, Maryam Fazel, and Yulia Tsvetkov. Personalized reasoning: Just-in-time personalization and why llms fail at it, 2025.

[9] Ali Montazeralghaem, Guy Tennenholtz, Craig Boutilier, and Ofer Meshi. Asking clarifying questions for preference elicitation with large language models, 2025.

[10] Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. Enhancing personalized multi-turn dialogue with curiosity reward, 2025.

[11] Shujin Wu, Yi R. Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. Aligning LLMs with individual preferences via interaction. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[12] Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models, 2024.

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

[14] Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences, 2025.

7 Author Contribution Statement

Erfan worked on dataset generation, experiments, and writeup. Steven worked on dataset generation, experiments, and writeup. Joseph worked on dataset generation, experiments/inference pipeline of evaluation, and writeup. All authors worked together during ideation phase.

8 Use of LLM Statement

Following the submission guidelines, we note that we used generative AI (ChatGPT) to (1) help locate related work and relevant domains, (2) find potential bugs in the experiment code implementations, and (3) edit writing for potential grammatical mistakes. All ideation, methodological design, experiments, and analyses are conducted by the authors.

A Supplementary Figures

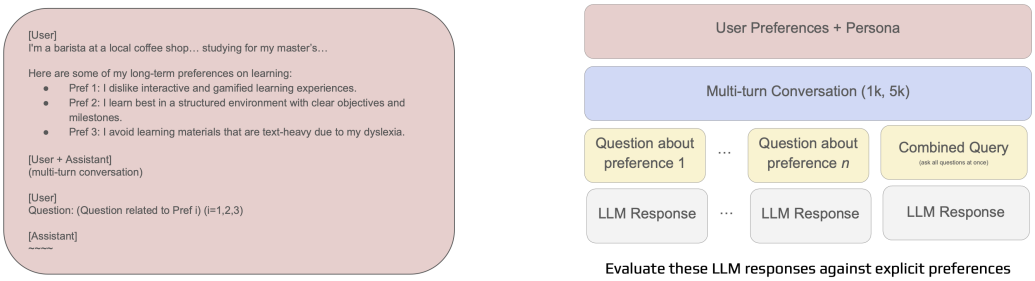


Figure 4: Sample persona and topic-grouped preferences (left) and conversation structure (right). We vary the intermediate context between the stated preferences and when the preference-specific question is asked.

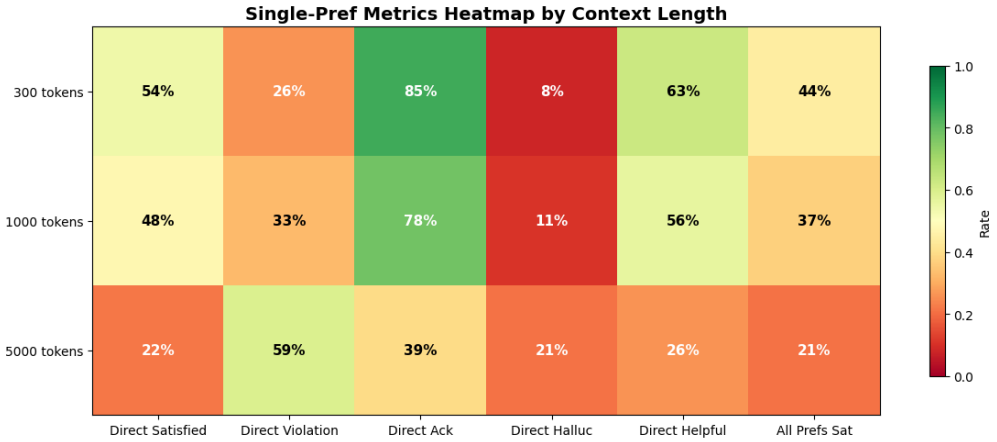


Figure 5: Sensitivity to intermediate dialogue length for Experiment 1. Heatmap of single-preference metrics while varying conversational context length (300/1,000/5,000 tokens), showing how preference-following degrades as the context length grows longer.

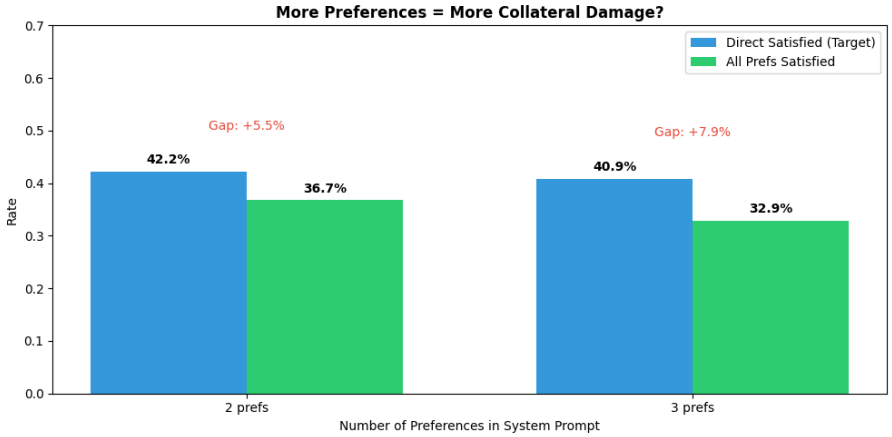


Figure 6: More preferences increases collateral damage in Experiment 1. While the model's rate of satisfying the *target* preference remains roughly stable when moving from 2 to 3 preferences in the system prompt, the rate of satisfying *all* preferences drops, widening the gap between "gets the target right" and "gets everything right." This highlights that multi-preference settings better reflect real usage, where assistants must jointly satisfy multiple standing constraints rather than optimizing a single preference in isolation.

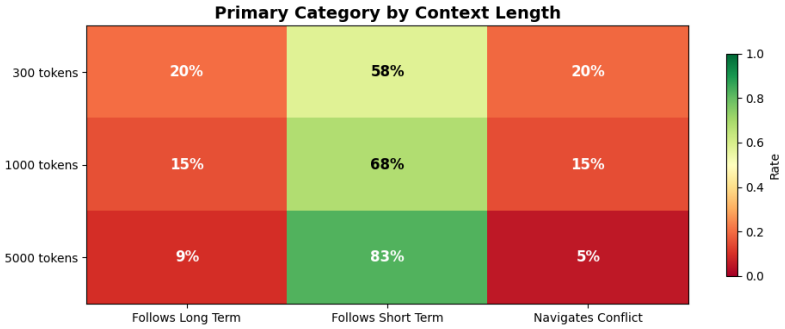


Figure 7: Sensitivity to intermediate dialogue length for Experiment 2. Heatmap of preference following metrics while varying conversational context length (300/1,000/5,000 tokens), showing how long-term preference following and conflict navigation degrades as the context length grows longer.

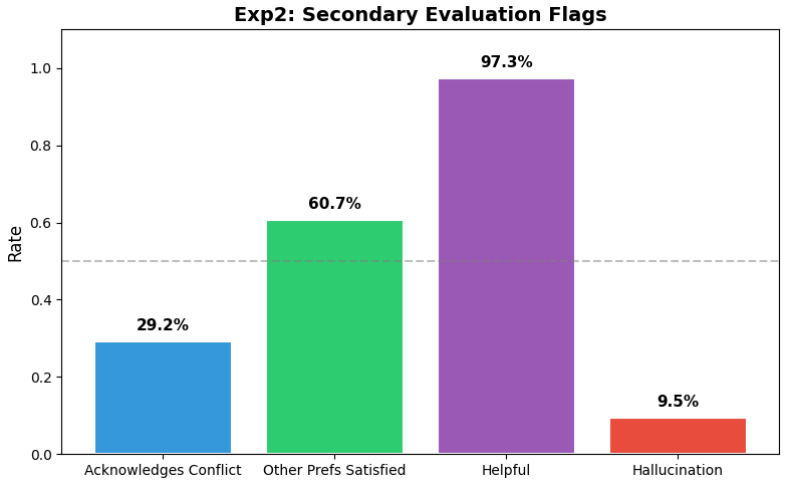


Figure 8: Experiment 2 secondary evaluation flags. Despite high overall helpfulness, the model acknowledges the long-term vs. short-term conflict in a minority of cases, indicating that it often complies without explicitly surfacing the inconsistency. The model largely satisfies *other* long-term preferences (those not in conflict), and hallucination remains relatively low, suggesting the dominant failure mode is preference prioritization under conflict rather than off-topic or incoherent behavior.