HORIZONBENCH: LONG-HORIZON PERSONALIZATION WITH EVOLVING PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

User preferences change over months of interaction, and a personalized assistant must recognize when a later life event supersedes a preference the user stated earlier. Natural conversations do not reveal a user's current preferences, so the time and cause of a change cannot be identified from real data. We introduce HORIZONBENCH and its data generator, the first resource that explicitly links a user's latent preference state to months of conversation and records the life event behind every preference change. The generator produces conversations from sampled sequences of user preference states, tracks when each preference changes separately from when it is expressed, and calibrates how likely a life event is to change preferences, and by how much, against a 13-week longitudinal study with 160 participants. HORIZONBENCH contains 4,245 questions from 360 simulated users, each asking for the response that best fits the user's current preferences; when a preference has changed, one option reflects its earlier value. All 25 frontier models we evaluate choose this outdated option more often than expected if errors were spread evenly (29.2% of wrong answers against 25%), and the best model reaches 52.8% accuracy. The pattern holds across context lengths and expression styles, including when the preference was expressed shortly before the question, and retrieval and memory methods do not recover the lost accuracy, because the evidence of a change is a life event that never states the new preference. Remembering what a user said and knowing what the user prefers now are therefore separate capabilities, and current models lack the second. HORIZONBENCH supports research on long-context models, memory systems, theory-of-mind reasoning, and user modeling.

1 INTRODUCTION

People increasingly interact with AI assistants over months of conversation (Zhang et al., 2025), and their preferences change during that time. Consider the user in Figure 1, who prefers Socratic philosophical discussion but, after becoming a chief executive, prefers direct frameworks without restating the preference. When she later asks whether it is better to be respected or liked, an assistant that answers in the Socratic style has retrieved her stated preference correctly but has not updated its belief about her. We call the task of identifying a user's current preference from a history in which later life events supersede earlier statements **long-horizon personalization**. State variation can dominate stable trait differences in some user-behavior settings (Harry et al., 2026), yet models and benchmarks commonly represent preferences as fixed traits.

Long-horizon personalization is difficult to measure, because natural conversations do not reveal a user's underlying preferences, so the time and cause of a change are unknown. Existing benchmarks test static preferences (Zhao et al., 2025a; Li et al., 2025b) or include changing information without an outcome that identifies reliance on a superseded preference (Jiang et al., 2025b; Tan et al., 2025a; Maharana et al., 2024; Wu et al., 2025), and synthetic user simulations do not record why preferences change (Wang et al., 2024; Chang et al., 2025). Concurrent work models preferences that emerge across event-driven dialogues (Liu et al., 2026), repeated memory updates (Uddin et al., 2026), and implicit invalidation of earlier user states (Chao et al., 2026); Table 1 compares these designs.

We introduce HORIZONBENCH and its data generator, the first resource that explicitly links a user's latent preference state to months of conversation and records the life event behind every preference change. Instead of inferring a user's state from conversations, the generator samples a sequence of user states from a structured mental state graph, in which typed dependency edges let one life event

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

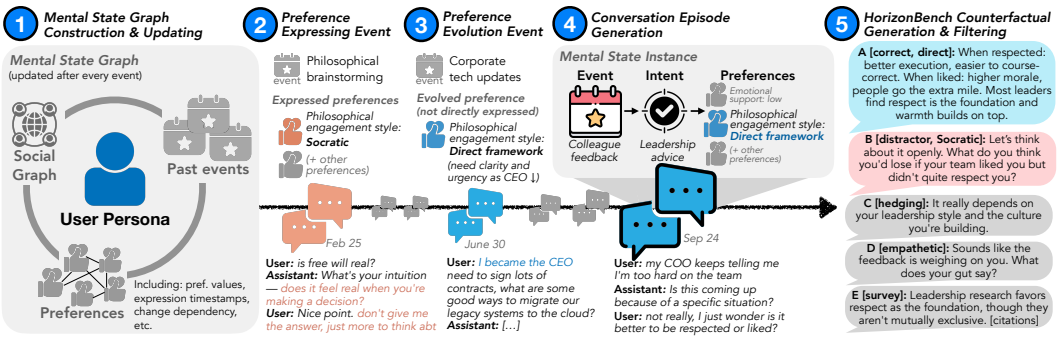

Figure 1: HORIZONBENCH pipeline overview. (1) The generator produces conversations from a structured mental state graph and records the preference state behind each conversation. (2) Preferences are established in conversations, then (3) changed by generated life events, which later conversations mention without stating the new values. (4) Later conversations are generated from the updated state and may express the new values. (5) The earlier values serve as pre-evolution distractors in HORIZONBENCH. See Appendix B for a complete benchmark item.

change several related preferences, and generates conversations from that sequence, so the preference behind every response and the event behind every change are known by construction. It also separates when a preference changes from when the user expresses it. A life event changes a preference without restating it, so the most recent explicit statement of a preference can be outdated when the question is asked, and the earlier value becomes a response option, the *pre-evolution distractor*, whose selection identifies a failure to update. Together with a short-horizon condition and controlled rewrites of how explicitly preferences are expressed, this design separates failures to update a belief about the user from failures to retrieve what the user said. Research on preference construction and life-course change shows that preferences shift with experience and major life events (Slovic, 1995; Fleeson, 2001; Specht et al., 2011). To measure how likely a life event is to change preferences and by how much, we ran a 13-week longitudinal study in which 160 participants rated preferences related to a planned life event before the event and again two weeks later; after 17.5% of completed or ongoing events, two to five of the six related preferences moved by at least three points on a seven-point scale, and the generator sets the probability and size of a preference change after a life event within these estimates (§3.2).

HORIZONBENCH contains 4,245 five-option questions from 360 simulated users, each with a six-month history averaging ∼4,300 turns and ∼163K tokens (§4). A filter keeps only questions that five models answer incorrectly without the history. All 25 frontier models we evaluate select the pre-evolution distractor more often than the 25% expected under evenly spread errors, 23 of them significantly, and all 25 are less accurate on changed preferences than on unchanged ones. The best model, Claude-opus-4.5, reaches 52.8%, and ten models score below the 20% chance level as a consequence of the filter. The pattern holds across context length, expression explicitness, and filter strictness, including when the preference was expressed 0–14 days before the question (§5). Replacing the history with turns selected by BM25, dense retrieval, or recency lowers accuracy for every model, and Mem0 matches the full history without improving on it, because the conversation containing the life event never names the new preference (§5.4). Remembering what a user said and knowing what the user prefers now are therefore separate capabilities, and current models and memory methods lack the second. We release the generator, the benchmark, and the mental state graphs. Our contributions are:

1. **Belief-update failure**, a failure mode in which a model answers with a user's superseded preference, observed in all 25 frontier models we evaluate (§5).
2. **HORIZONBENCH**, 4,245 questions with pre-evolution distractors and controlled variants of context length, expression explicitness, and filter strictness (§4).
3. **A data generator** that constructs conversations from a structured mental state graph, records the life event behind every preference change, separates when a preference changes from when it is expressed, and calibrates the probability and magnitude of event-linked change against longitudinal human ratings (§3).

Table 1: Comparison with related benchmarks. **Long Context**: 100K+ tokens. **Mental State Graph**: preferences with dependency edges. **Event-Driven Evo.**: preferences change through events. **Controlled Dims.**: difficulty factors varied independently. **Old-Value Distract.**: the previous value is a response option. **Human Calib.**: human data calibrate generation. ◐ = partial.

Method	Long Context	Mental State Graph	Event-Driven Evo.	Controlled Dims.	Old-Value Distract.	Human Calib.
PREFEVAL (ZHAO ET AL., 2025A)	✓	✗	✗	✗	✗	✗
PERSONAMEM (JIANG ET AL., 2025A)	✓	✗	◐	✗	✗	✗
PERSONAMEM-V2 (JIANG ET AL., 2025B)	✓	✗	✗	✗	✗	✗
IMPLEXCONV (LI ET AL., 2025B)	✗	✗	✗	✗	✗	✗
LOCOMO (MAHARANA ET AL., 2024)	◐	◐	✗	✗	✗	✗
MEMBENCH (TAN ET AL., 2025A)	◐	✗	✗	✗	✗	✗
PERMA (LIU ET AL., 2026)	✓	✗	✓	✓	✗	✗
MEMORA (UDDIN ET AL., 2026)	◐	✗	◐	✓	✗	✗
STALE (CHAO ET AL., 2026)	✓	◐	◐	✓	✗	✗
HORIZONBENCH	✓	✓	✓	✓	✓	✓

2 PROBLEM FORMULATION

Latent user state. A user's preferences are described by K preference dimensions, each a typed attribute of a preference domain (e.g., `response_structure` for emotional support), and dimension k takes values in a finite set $\mathcal{V}_k$. The user's latent state at conversation $t \in \{1, \ldots, T\}$ is $u_t = (v_{t,1}, \ldots, v_{t,K})$. The state is piecewise constant: a life event ℓ_t may update a subset $S_t \subseteq \{1, \ldots, K\}$ of dimensions, $u_t = f_{\text{evo}}(u_{t-1}, \ell_t)$, and otherwise $u_t = u_{t-1}$. Each conversation C_t is generated from u_t, and a model observes only the history $H_t = (C_1, \ldots, C_t)$, never u_t or which dimensions changed.

User-conditioned reward. Given a conversational context x in conversation t, the user's reward for an assistant response r is $R(x, r; u_t)$, which measures how well r fits the user's current state. The true reward depends on the history only through u_t. A model acting as a reward model assigns a score $s(r \mid H_t, x)$ without access to u_t, so it must infer the relevant part of u_t from H_t. When a dimension has changed, a score that still conditions on its earlier value ranks responses by a reward the user no longer holds.

Evaluation. Each item queries one dimension k at a test turn with context x. The correct response r^* realizes the current value $v^* = v_{t,k}$, and four alternatives $r_1, \ldots, r_4$ of comparable quality realize other values $v_1, \ldots, v_4 \in \mathcal{V}_k$. Because the candidates differ only in the value of dimension k, $R(x, \cdot\,; u_t)$ ranks r^* highest, and a model is correct when its scores do the same, $s(r^* \mid H_t, x) > s(r_j \mid H_t, x)$ for all j. This is the listwise form of standard reward-model evaluation (Lambert et al., 2024), conditioned on a user state that must be inferred. Generative models select the best response directly, which corresponds to the argmax of their implicit scores. An item is *evolved* if the preference has changed since the user first expressed it, and *static* otherwise. On evolved items, one alternative typically realizes the value the user originally expressed, v_{old}, which we call the *pre-evolution distractor*. We report two measures of belief-update failure. The *distractor rate* is the share of wrong answers on items with a pre-evolution distractor that select it; if errors were spread evenly across the four wrong options, it would be 25%, which we call the uniform-error reference. The *evolved–static gap* is accuracy on evolved items minus accuracy on static items. Long-horizon personalization requires scoring correctly on evolved items, where later event-linked evidence in H_t supersedes an earlier statement of v_{old}.

Generating histories from states. Because u_t is never observed in natural conversation, the same history can be consistent with several sequences of user states, so neither R nor the correct response can be established from real data. The data generator (§3) therefore works in the opposite direction: it samples a sequence of user states $u_{1:T}$ and produces H_T from it, so the reward labels of every item and the life event behind every change are known by construction, and each change has one designated current value.

3 HORIZONBENCH DATA GENERATOR

The generator models both the user and the assistant as stateful agents. Its user state augments the preference state u_t of §2 with a persona profile, and the assistant state a_t likewise comprises a persona profile and typed preference attributes. The generator's central data structure is a *mental*

Table 2: Human estimates (95% CIs) and the corresponding generator settings.

Component	Human estimate	Generator setting	Mapping
Coordinated change per life event	0.175 [0.132, 0.219]	$p_{\text{evo}} = 0.15$	2–5 ratings shift ≥ 3 points
Change magnitude	3.64 [3.51, 3.77] points	Distinct value transition	Excludes small rating changes

state graph $\mathcal{G}$ that maintains each agent's persona, social connections, life events, and preferences. Preferences are linked by typed dependency edges (e.g., dietary preferences $\rightarrow$ restaurant choices $\rightarrow$ social outing plans), and life events alter preference values with changes propagating through these edges. The generator records the life event behind every change and tracks when each preference changes separately from when it is expressed. A preference that has not been expressed for a set number of days becomes *stale*, and the generator can then *recall* it by placing a later turn whose response must reflect the preference's current value; each such turn is a candidate benchmark question (§3.4). A language model that simulates a long interaction drifts from its assigned persona, because it keeps no record of the agents' states across generation calls, so the generator conditions every call on $\mathcal{G}$.

3.1 Event Sampling

Each conversation C_t is organized around an episode event e_t, the occasion for the interaction (e.g., rehearsing for a debate), which is distinct from the life events ℓ_t that change preferences (§3.2). Episode events are sampled conditional on both agent states and event history, $e_t \sim p(e \mid u_{t-1}, a_{t-1}, e_{1:t-1}; \boldsymbol{\beta})$, preventing unrealistic repetition from independent draws. Category weights $\boldsymbol{\beta}$ follow empirical AI-chatbot interaction distributions (Zhang et al., 2025) (emotional support 26%, collaborative storytelling 26%, romantic interactions 22%). When a stale preference is selected for temporal recall, the sampled event is constrained to that preference domain, so the recall turn fits the conversation.

3.2 Preference Evolution

Preferences do not change in isolation. A career change may jointly affect work preferences, commute habits, and financial priorities. Typed dependency edges in $\mathcal{G}$ represent such relationships, $u_t = f_{\text{evo}}(u_{t-1}, \ell_t, \mathcal{G}; \boldsymbol{\phi}_{\text{evo}})$, where $\boldsymbol{\phi}_{\text{evo}}$ collects the evolution parameters described below. Each life event changes the user's preferences with probability $p_{\text{evo}} = 0.15$ and otherwise leaves u_t unchanged. When a change occurs, a connected set of 2–5 preferences is selected, and an LLM jointly constructs the event and a qualitatively distinct new value for each preference, conditioned on the user's current state, and the generator records the event with the change. A short conversation then mentions the event without stating the new values.

We calibrate this mechanism against immediate event-linked changes from seven surveys conducted over 13 weeks with 160 participants (Appendix C). At each survey from T0 through T5, participants rated six preferences associated with a planned event; two weeks later, they reported whether the event had occurred or was underway and rated the same preferences again. To approximate the generator's abrupt multi-preference transition, we define a coordinated change as two to five of the six ratings moving by at least three points on a seven-point scale. Among 470 completed or ongoing events with six unambiguously linked ratings, 17.45% met this criterion (participant-clustered 95% CI, 13.22–21.85%). Conditional changes averaged 3.64 points (95% CI, 3.51–3.77). The generator's $p_{\text{evo}} = 0.15$ lies within this interval, and each change moves a preference to a qualitatively distinct value, consistent with the large observed magnitude (Table 2). The human data therefore calibrate the probability and size of change given a life event; the frequency of life events is set separately.

The evolution event itself does *not* refresh a preference's expression timestamp. A preference updated on day 45 remains stale from its previous expression on day 10 and may become immediately eligible for recall. Later conversations can express the updated value and refresh its timestamp, so evolved items vary in how much post-evolution evidence precedes the test. This design keeps the pre-evolution value available as a response option and lets us test whether mention frequency explains the errors (§5.3).

3.3 Conversation Generation

An *outline* is generated first, specifying which preferences appear, in what order, and with what expression style, enforcing structural constraints such as placing temporal recall turns in contextually appropriate positions. Content is then generated turn by turn conditional on the outline, $m^u_{t,n} \sim p_u(m \mid u_t, e_t, H_{t,n-1})$ and $m^a_{t,n} \sim p_a(m \mid a_t, m^u_{t,n}, H_{t,n-1})$, where $m^u_{t,n}$ and $m^a_{t,n}$ are the user and assistant messages of turn n in C_t, and $H_{t,n-1}$ is H_{t-1} followed by the first $n-1$ turns of C_t. The assistant persona is generated per-event to match the interaction context (e.g., a career counselor for a job change). To control how explicitly preferences are expressed, we produce variants of the original (*neutral*) conversations by rewriting only preference-expressing user turns and the immediately following assistant turns. In the *explicit* variant, preferences are stated with clear, varied phrasing while preserving the original register and structure. In the *implicit* variant, preferences are conveyed only through tangential remarks and contextual cues. Because the conversation content, preference structure, sequence of preference changes, and benchmark items remain identical across variants, only the expression of preferences varies (§5.3).

3.4 Expression Tracking and Temporal Recall

Without tracking when preferences were expressed, a benchmark cannot determine whether the model never encountered the preference, retrieved it incorrectly, or failed to update it. Conversations are placed on a six-month calendar, with d_t the day of conversation t. Each preference i maintains the day τ_i of its last conversational expression, refreshed after each conversation, and preferences whose staleness $d_t - \tau_i$ exceeds a threshold (default 30 days) become eligible for temporal recall, sampled proportionally to staleness and optionally weighted toward evolved preferences.

Every recalled preference has been expressed in the complete benchmark history. When the preference has evolved, its originally expressed value is injected as a distractor (§4.1), and the associated life event appears in a dedicated conversation; some histories also contain later expressions of the updated value. All generator parameters (timeline length, expression explicitness, evolution probability, staleness threshold, user demographics) are configurable.

4 HorizonBench Benchmark

HorizonBench is one instantiation of the data generator. Each benchmark item is a 5-option multiple-choice question embedded within a conversation history. Given the full history, the model must identify which assistant response best aligns with the user's *current* preferences, the response ranked highest by the user-conditioned reward $R(x, \cdot\,; u_t)$ of §2. The multiple-choice format lets the pre-evolution distractor isolate the error, as in PrefEval (Zhao et al., 2025a), PersonaMem-v2 (Jiang et al., 2025b), and ImplexConv (Li et al., 2025b).

4.1 Counterfactual Evaluation and History Filtering

For each temporal recall turn, one attribute k of the recalled preference is sampled uniformly, and its current value $v^* = v_{t,k}$ defines the correct response. An LLM proposes four maximally contrasting alternative values $\{v_1, \ldots, v_4\}$ for k. If the value of k has changed since the user first expressed it, the originally expressed value v_{old} replaces one of them as the pre-evolution distractor. For each alternative v_j, an LLM rewrites the assistant turn into a contextually appropriate response r_j that fits v_j.

Items are then passed through a **5-LLM consensus history filter**. Each candidate is presented to five validation models (GPT-4.1, Gemini-3-flash, o4-mini, Claude-haiku-4.5, Claude-opus-4.6) *without* the full conversation history and passes only if all five answer incorrectly. This removes items that the validators can answer without the conversation history. Four of the five filter models are also among the 25 evaluated models. The filter threshold is tunable. Strict consensus (5/5) retains the hardest items, while majority ($\geq$3/5) admits items that are easier to answer without the history, so the threshold sets item difficulty (§5.3).

4.2 Dataset Statistics

HorizonBench is produced by three frontier generators (Claude-sonnet-4.5 with 1,052 items from 120 users, o3 with 981 items from 120 users, and Gemini-3-flash with 2,212 items from 120 users), which allows results to be compared across generators (§4.3). The pipeline generates 57,390 conversations totaling 1.6M turns across 360 users, with six-month timelines averaging 142–168 conversations and 3,600–5,000 turns per user across 30 preference domains. Of 45,006 candidates, 4,245 pass the strict 5-LLM filter (9.4% pass rate), of which 59% involve evolved preferences and 41% static. Appendix B shows a complete item.

4.3 Benchmark Validation

No generator-family advantage. Models show no advantage on items generated by their own family. Claude models score -4.8 pp on sonnet-4.5-generated items relative to the remaining generators, and OpenAI models score -2.6 pp on o3-generated items.

Model accuracy is independent of ambiguity in the updated preference. Human annotation indicates that benchmark items are recoverable from the conversation history, and annotator agreement measures how clearly an item's updated preference determines the correct response. Two development rounds ($\sim$1,300 items, 3 annotators per item) yielded majority-vote accuracy of 56–64% (Fleiss $\kappa = 0.37$–0.51), with at least one annotator answering correctly on 81% of Round 2 items. A benchmark-wide annotation pass covering 4,235 items then stratifies model performance on agreement rate. Mean accuracy across the 25 models is about 24% on full-, majority-, and no-agreement subsets, and the distractor rate reaches 42.9% on the 673 full-agreement items (Table 12). Items whose updated value admits a single reading are therefore no easier for models.

Response quality checks. For 4,235 items with three annotations each, individual positive ratings were 99.7% for situational fit, 99.7% for conversation-flow realism, and 99.4% for assistant-persona consistency; every item received a positive majority rating on all three dimensions (Appendix F).

5 Experiments

We use HorizonBench to answer three questions:

- **RQ1.** Do models update what they know about a user when a life event changes a preference?
- **RQ2.** Is the failure explained by long-context retrieval demand, mention frequency, reliance on explicit statements, or item difficulty?
- **RQ3.** Do retrieval and memory methods recover the lost accuracy?

5.1 Experimental Setup

We evaluate 25 models spanning three families, including 8 OpenAI models (gpt-4o, gpt-4.1, gpt-5, gpt-5-mini, o1, o3, o3-mini, o4-mini), 9 Gemini models (2.0-flash-lite through 3.1-pro), and 8 Claude models (3.5-haiku through opus-4.5). Each model receives the full conversation history spanning six months ($\sim$4,300 turns, averaging $\sim$163K tokens per prompt with some exceeding 250K) and selects which of five assistant responses best aligns with the user's preferences. We report results on the combined multi-generator dataset (4,245 items from 360 users across three generators). Items exceeding a model's context window are truncated by removing earlier conversation turns. The short-horizon condition (§5.3) provides a complementary test within histories short enough to fit every model's context window. Sign and Wilcoxon tests are computed over the 25 per-model values. All confidence intervals are 95% percentile bootstrap CIs with B=10,000 resamples.

5.2 RQ1: Models Answer with Superseded Preferences

Figure 2 reports per-model accuracy on the 4,245 items (full results in Table 7). The best model, Claude-opus-4.5, achieves 52.8%. Ten of 25 models score below the 20% chance baseline, including seven of eight OpenAI models and three Gemini models. Position-debiased accuracy differs from raw accuracy by less than 1 pp for all models, ruling out position bias as an explanation. Below-chance accuracy follows from how items are selected. The 5-LLM filter retains only items that all five validators answer incorrectly without the conversation history, which removes items whose response

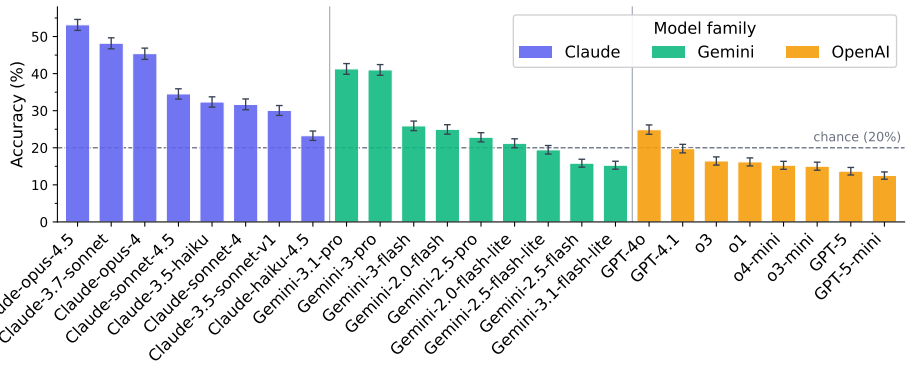

Figure 2: Per-model accuracy on HORIZONBENCH (4,245 items). Bars are colored by model family with bootstrap 95% CIs (B=10,000). Dashed line marks the 20% chance baseline.

options alone point to the correct answer. The same models score below chance on static items, which carry no pre-evolution distractor: gpt-5-mini reaches 13.3% there and gpt-5 15.4%. Below-chance accuracy therefore reflects item selection, and the distractor rate and the evolved–static gap measure the update failure.

We additionally evaluate five open-weight models from the Llama, OLMo, and Qwen families on 1,011 items from the Claude-sonnet-4.5-generated subset, with left truncation for items exceeding their context windows (Table 8). All five score below 20% (11.6–17.1%) and have lower accuracy on evolved than static items, extending the evolved–static gap beyond proprietary model families.

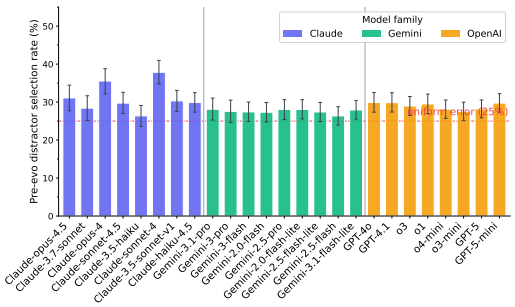

Figure 3: Distractor rate by model. The dashed line marks the 25% uniform-error reference (1 of 4 wrong options); all 25 models exceed it.

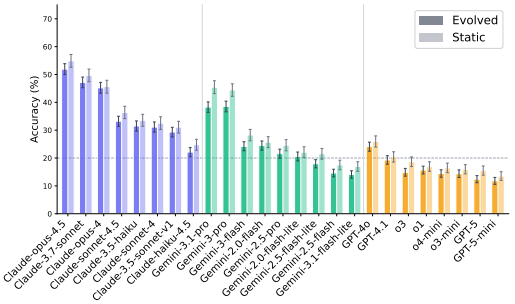

Figure 4: Accuracy on evolved (dark) vs. static (light) preferences with 95% bootstrap CIs on the combined dataset (4,245 items). All 25 models perform worse on evolved items (sign test $p < 0.001$).

Distractor rate. All 25 models exceed the 25% uniform-error reference (Figure 3), with a mean distractor rate of 29.2% and a maximum of 37.8%. Twenty-three of the 25 are individually significant under a one-sided binomial test ($p < 0.05$). Claude-opus-4.5, the most accurate model, selects it on 31.1% of its wrong answers. Within the OpenAI family, reasoning and non-reasoning models show similar mean distractor rates (37.0% vs. 39.9%), placing the pattern in both subgroups and indicating that inference-time reasoning alone does not resolve it.

Evolved–static gap. All 25 models perform worse on evolved preferences than static ones (−2.7 pp average, sign test 25/25 negative, $p < 0.001$, Figure 4). The gap is largest for Gemini models (−3.5 pp) and smallest for OpenAI (−2.1 pp). Static accuracy also stays low, so retrieval is a separate source of difficulty.

5.3 RQ2: THE FAILURE PERSISTS ACROSS CONTROLLED CONDITIONS

We test four alternative explanations for the failure: long-context retrieval demand, mention frequency, reliance on explicit statements, and item difficulty. Each controlled condition varies one factor and reports the evolved–static gap and the distractor rate (Table 3).

Table 3: Belief-update metrics across controlled conditions, averaged over the 25 models. Neutral, explicit, and implicit rows use the 981 o3-generated items; the short-horizon and relaxed-filter rows use their own item sets (per-model results in Appendix H). *Evo*Δ: evolved–static gap. *Dist.%*: distractor rate (uniform-error reference 25%). The two count columns give how many models show each pattern.

Condition	Overall	Static	Evolved	EvoΔ	EvoΔ<0	Dist.%	Dist.>25%
Neutral (unmodified)	27.0	28.9	26.0	−2.9	22/25	38.2	25/25
Short horizon	30.6 ↑	31.4 ↑	25.7 ↓	−5.7	25/25	47.7	25/25
Explicit rewrite	27.9 ↑	29.0 ↑	27.2 ↑	−1.8	16/25	49.7	25/25
Implicit rewrite	26.4 ↓	28.6 ↓	25.0 ↓	−3.6	20/25	58.8	25/25
Relaxed filter	38.9 ↑	40.3 ↑	37.9 ↑	−2.5	22/25	54.3	25/25

Table 4: Accuracy (%) on the 2,484 evolved items when a retrieval method replaces the full history.

Model	Full history	BM25	Recency	Dense 384d	Dense 1536d
Claude-opus-4.5	52.1	42.2	42.8	40.3	40.6
Claude-sonnet-4.5	33.3	27.5	26.4	24.8	25.6
GPT-4o	24.1	17.5	15.5	17.8	18.0
Claude-haiku-4.5	20.5	16.3	13.0	15.2	15.3
GPT-4.1	19.3	18.0	14.1	17.6	17.5

Context length. If belief-update failure came only from the demands of long contexts, it should disappear when the preference was expressed shortly before the test. A short-horizon variant tests this prediction. It regenerates histories with the staleness threshold (§3.4) relaxed, so a preference is tested 0–14 days after its last expression rather than at least 30 days later (∼95K tokens per history, 120 users, 1,778 items). Static accuracy rises slightly (31.4% vs. 28.9%), consistent with an easier retrieval task in shorter context, yet the gap widens to −5.7 pp (25/25 models negative, Wilcoxon $p < 10^{-7}$) and the distractor rate rises to 47.7% (25/25 above the uniform-error reference, Wilcoxon $p < 10^{-7}$). The evolved–static gap therefore persists at shorter horizons.

Mention frequency. A simpler account is that models choose the value mentioned most often. We therefore count conversations containing the pre- and post-evolution values in each model-visible history. Across Claude-sonnet-4.5-, o3-, and Gemini-3-flash-generated histories, the new value constitutes 65.7–69.2% of old-plus-new mentions and is more frequent than the old value on 65.6–70.5% of matched histories (Table 9). Nevertheless, accuracy is lower in that bucket for two of the three generators, so mention frequency does not explain the errors.

Expression explicitness. Prior work identifies implicit preference expression as a main source of difficulty (Li et al., 2025b; Jiang et al., 2025b). Starting from the neutral conversations, we evaluate the explicit and implicit rewrites of §3, which follow the methodology of Li et al. (2025b). Overall accuracy decreases monotonically (27.9% explicit, 27.0% neutral, 26.4% implicit). The evolved–static gap widens from −1.8 pp (explicit) to −2.9 pp (neutral) to −3.6 pp (implicit), and the distractor rate rises from 49.7% (explicit) to 58.8% (implicit). Implicit expression increases selection of the earlier value even though, in aggregate, the updated value appears more frequently in the history (§5.3).

Filter strictness. Relaxing the 5-LLM filter (§4.1) from strict consensus (5/5) to majority (≥3/5) raises average accuracy by +11.9 pp (every model positive, paired Wilcoxon $p < 10^{-7}$). The evolved–static gap nevertheless remains (−2.5 pp, 22/25 models negative, Wilcoxon $p < 10^{-4}$), and every model has a distractor rate above the uniform-error reference (54.3%, 25/25). The filter therefore sets item difficulty, while the belief-update pattern holds at both thresholds.

5.4 RQ3: Retrieval Cannot Surface the Evidence of a Change

Long-horizon personalization is commonly treated as a memory problem, so we replace the full history with a smaller context while holding fixed the model that answers the question (the reader) and the scoring. BM25 (Robertson & Zaragoza, 2009), dense retrieval with a locally run 384-dimensional encoder, and dense retrieval with `text-embedding-3-small` each return the ten highest-scoring turns in timestamp order, following Wu et al. (2025), and a recency baseline returns the ten most recent turns. All four run on the 2,484 evolved items.

Every retrieval condition falls below the full history for every model, and the strongest model loses the most: Claude-opus-4.5 drops 9.9 pp under BM25 and up to 11.8 pp under dense retrieval (Table 4). The distractor rate reaches 37–47%, against 38.2% with the full history. Mem0 (Chhikara et al., 2025), which builds memories with a language model as conversations arrive and never sees the question, reaches 19.9% with `gpt-4.1` as the reader (1,492 items with a pre-evolution distractor), against 19.3% for the same reader on the full history.

The generator's records explain this. Because the generator stores the conversation in which each life event changed a preference, retrieval can be scored against it: BM25 retrieves that conversation on 18.5% of items, against 22.1% for a random draw of the same size, and a statement of the current value on 35.9%, against 31.4%. The two dense encoders select largely different turns (21.1% mean Jaccard overlap) yet differ by at most 0.8 pp, so which turns reach the model barely matters. An event turn describes a career change or a diagnosis without naming a preference value, so similarity to a question about that preference does not surface it, and no selection from the history can supply evidence that the history states only implicitly.

6 Related Work

Personalization and long-term memory benchmarks. Existing benchmarks test whether models retrieve stated preferences and facts from long histories (Zhao et al., 2025a; Jiang et al., 2025a; Maharana et al., 2024; Tan et al., 2025a; Wu et al., 2025; Liu et al., 2024), often treating implicit expression as the main source of difficulty (Jiang et al., 2025b; Li et al., 2025b; Guo et al., 2026). Those that include changing information have no outcome that identifies continued reliance on a superseded preference, so a retrieval failure and an update failure produce the same error. Concurrent work models evolving user state through emerging preferences (Liu et al., 2026), repeated memory updates (Uddin et al., 2026), and implicit state invalidation (Chao et al., 2026). HorizonBench differs in three ways: the earlier value appears among the response options, so selecting it measures the update failure directly; every change is tied to a recorded life event; and context length and expression vary while the transition is held fixed (Table 1). The resulting error pattern connects to findings that models favor prior evidence over newer context (Xu et al., 2024).

Preference change. Psychological research shows that preferences are partly constructed in context and shift with experience, roles, and life events (Slovic, 1995; Schwarz, 2007; Fleeson, 2001; Elder, 1998; Hitlin & Piliavin, 2004; Specht et al., 2011), which motivates the generator's event-linked evolution; a longitudinal human study calibrates its parameters (§3.2). Temporal recommender systems model preference drift from timestamped behavior (Koren, 2009; Rafailidis & Nanopoulos, 2016; Rabiu et al., 2020). HorizonBench targets the conversational case, in which an event implies a new preference before the user states it.

User simulation. The generator builds on LLM agent simulation (Park et al., 2023) and synthetic life-long personalization data (Wang et al., 2024; Chang et al., 2025), which lack structured preference transitions with known causes. Structured models of how preferences co-vary support inferring a user's preferences from few interactions (Bose et al.), and evidence that state variation can dominate stable trait differences in some user-behavior settings (Harry et al., 2026) motivates modeling preferences as changing states. Appendix A extends this discussion.

7 Conclusion

We introduced HorizonBench and its generator, the first resource that explicitly links a user's latent state to months of conversation and records the cause of every preference change, providing supervision that natural conversation data cannot. All 25 frontier models choose the user's superseded preference more often than chance errors would predict, across context lengths, expression styles, and filter thresholds, and including when the preference was expressed shortly before the test. Retrieval and memory methods do not recover the lost accuracy, because the evidence of a change is a life event that the history never states as a preference. Remembering what a user said and knowing what the user wants now are therefore separate capabilities, and the models and memory methods we evaluate lack the second. The generator supplies ground truth for evaluating methods that track this change, and for training them in future work. Because a user's reward changes as the user's preferences change, tracking the user's state is a prerequisite for personalized reward modeling and for aligning assistants to individual users over long horizons.

Limitations and Future Directions

HorizonBench is synthetic: the conversations, life events, dependency graph, and the direction of each preference change are produced by the generator, and the human study calibrates two of its parameters, the probability and the magnitude of change after a life event. The generator's other parameters, such as the frequency of life events and the staleness threshold, are configurable and can be calibrated against further human data. The longitudinal study covers 13 weeks of everyday preference ratings from a cohort concentrated among adults aged 18 to 39, urban residents, bachelor's-degree holders, and people working in data annotation or AI training; longer studies with broader cohorts can recalibrate the generator for other populations. The history filter is model-based and shares four models with the evaluation panel, and the relaxed-filter condition (§5.3) reports results under a weaker threshold. Context-window truncation removes early turns for models with shorter windows, and the short-horizon condition fits every model's window. Because every generated history comes with the state behind it, the generator can also supply training data for methods that infer and revise a user's state, which we leave to future work.

Reproducibility Statement

The generator, benchmark, mental state graphs, and evaluation code are publicly released, with links withheld for anonymous review. §3 specifies the generation pipeline and §4.1 the counterfactual construction and 5-LLM history filter, with the full prompts in Appendix G and generation statistics in Appendix E. §5 gives the evaluation protocol, including the models, truncation rule, and bootstrap settings; per-model results appear in Appendix D and per-model controlled-condition results in Appendix H. The calibration estimates in §3.2 are reproduced by the analysis script accompanying the human data release, whose inclusion rules and formulas are given in Appendix C.

Ethics Statement

The benchmark conversations, personas, and evaluation items are generated by LLMs and do not represent study participants. We acknowledge that LLM-generated conversations may encode biases present in the underlying models. Preference evolution scenarios such as health and career events are synthetic evaluation constructs and should not be interpreted as clinical or counseling models.

References

Saleh Afzoon, Zahra Jamali, Usman Naseem, and Amin Beheshti. PersoBench: Benchmarking personalized response generation in large language models. arXiv preprint arXiv:2410.03198, 2024. URL https://arxiv.org/abs/2410.03198.

Lize Alberts, Benjamin Ellis, Andrei Lupu, and Jakob Foerster. CURATe: Benchmarking personalised alignment of conversational AI assistants. arXiv preprint arXiv:2410.21159, 2024. URL https://arxiv.org/abs/2410.21159.

Avinandan Bose, Shuyue Stella Li, Faeze Brahman, Pang Wei Koh, Simon Shaolei Du, Yulia Tsvetkov, Maryam Fazel, Lin Xiao, and Asli Celikyilmaz. Cold-start personalization via bayesian adaptive questioning. In *Forty-third International Conference on Machine Learning*. URL https://arxiv.org/abs/2602.15012.

Serina Chang, Ashton Anderson, and Jake M. Hofman. ChatBench: From static benchmarks to human-AI evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26009–26038. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1262. URL https://aclanthology.org/2025.acl-long.1262/.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? arXiv preprint arXiv:2605.06527, 2026. URL https://arxiv.org/abs/2605.06527.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025*, Frontiers in Artificial Intelligence and Applications, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Glen H. Elder, Jr. The life course as developmental theory. *Child Development*, 69(1):1–12, 1998. doi: 10.1111/j.1467-8624.1998.tb06128.x. URL https://doi.org/10.1111/j.1467-8624.1998.tb06128.x.

William Fleeson. Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, 80(6):1011–1027, 2001. doi: 10.1037/0022-3514.80.6.1011. URL https://doi.org/10.1037/0022-3514.80.6.1011.

Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. Towards natural personalization: Evaluating long-horizon preference following in personalized user-LLM interactions. arXiv preprint arXiv:2603.04191, 2026. URL https://arxiv.org/abs/2603.04191.

Tamunotonye Harry, Ivoline C. Ngong, Chima Nweke, Yuanyuan Feng, and Joseph Near. Beyond fixed psychological personas: State beats trait, but language models are state-blind. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26440–26468. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1316. URL https://aclanthology.org/2026.findings-acl.1316/.

Steven Hitlin and Jane Allyn Piliavin. Values: Reviving a dormant concept. *Annual Review of Sociology*, 30(1):359–393, 2004. doi: 10.1146/annurev.soc.30.012703.110640. URL https://doi.org/10.1146/annurev.soc.30.012703.110640.

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. In *Second Conference on Language Modeling*, 2025a. URL https://arxiv.org/abs/2504.14225.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory W. Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. PersonaMem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. arXiv preprint arXiv:2512.06688, 2025b. URL https://arxiv.org/abs/2512.06688.

Yehuda Koren. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 447–456. Association for Computing Machinery, 2009. doi: 10.1145/1557019.1557072. URL https://doi.org/10.1145/1557019.1557072.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. LongLaMP: A benchmark for personalized long-form text generation. arXiv preprint arXiv:2407.11016, 2024. URL https://arxiv.org/abs/2407.11016.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024. URL https://aclanthology.org/2025.findings-naacl.96/.

Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. PrefPalette: Personalized preference modeling with latent attributes. In *The Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=p4ujQsKmPV.

Shuyue Stella Li, Avinandan Bose, Faeze Brahman, Simon Shaolei Du, Pang Wei Koh, Maryam Fazel, and Yulia Tsvetkov. PrefDisco: Benchmarking proactive personalized reasoning. In

The Fourteenth International Conference on Learning Representations, pp. 80844–80908, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/82d50077a0e140b524a18f380c95d55a-Abstract-Conference.html.

Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. Toward multi-session personalized conversation: A large-scale dataset and hierarchical tree framework for implicit reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11493–11506. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.580. URL https://aclanthology.org/2025.emnlp-main.580/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, Zhiyu Li, Feiyu Xiong, Enhong Chen, and Tong Xu. PERMA: Benchmarking personalized memory agents via event-driven preference and realistic task environments. arXiv preprint arXiv:2603.23231, 2026. URL https://arxiv.org/abs/2603.23231.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Idris Rabiu, Naomie Salim, Aminu Da'u, and Akram Osman. Recommender system based on temporal models: A systematic review. *Applied Sciences*, 10(7):2204, 2020. doi: 10.3390/app10072204. URL https://doi.org/10.3390/app10072204.

Dimitrios Rafailidis and Alexandros Nanopoulos. Modeling users preference dynamics and side information in recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(6):782–792, 2016. doi: 10.1109/TSMC.2015.2460691. URL https://doi.org/10.1109/TSMC.2015.2460691.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Norbert Schwarz. Attitude construction: Evaluation in context. *Social Cognition*, 25(5):638–656, 2007. doi: 10.1521/soco.2007.25.5.638. URL https://doi.org/10.1521/soco.2007.25.5.638.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. Reliable LLM-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pp. 19–35, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.scichat-1.3. URL https://aclanthology.org/2024.scichat-1.3/.

Paul Slovic. The construction of preference. *American Psychologist*, 50(5):364–371, 1995. doi: 10.1037/0003-066X.50.5.364. URL https://doi.org/10.1037/0003-066X.50.5.364.

Jule Specht, Boris Egloff, and Stefan C. Schmukle. Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, 101(4):862–882, 2011. doi: 10.1037/a0024950. URL https://doi.org/10.1037/a0024950.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.989. URL https://aclanthology.org/2025.findings-acl.989/.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh R N, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. PersonaBench: Evaluating AI models on understanding personal information through accessing (synthetic) private user data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 878–893, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.49. URL https://aclanthology.org/2025.findings-acl.49/.

Meiling Tao, Chenghao Zhu, Dongyi Ding, Tiannan Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. PersonaFeedback: A large-scale human-annotated benchmark for personalization. arXiv preprint arXiv:2506.12915, 2025. URL https://arxiv.org/abs/2506.12915.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26814–26841. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1337. URL https://aclanthology.org/2026.findings-acl.1337/.

Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. AI PERSONA: Towards life-long personalization of LLMs. arXiv preprint arXiv:2412.13103, 2024. URL https://arxiv.org/abs/2412.13103.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, pp. 86809–86836, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Sangyeon Yoon, Sunkyoung Kim, Hyesoo Hong, Wonje Jeung, Yongil Kim, Wooseok Seo, Heuiyeen Yeen, and Albert No. BenchPreS: A benchmark for context-aware personalized preference selectivity of persistent-memory LLMs. arXiv preprint arXiv:2603.16557, 2026. URL https://arxiv.org/abs/2603.16557.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1490–1504, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.83. URL https://aclanthology.org/2024.naacl-long.83/.

Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. The rise of AI companions: Interaction with AI companions and psychological well-being. arXiv preprint arXiv:2506.12605, 2025. URL https://arxiv.org/abs/2506.12605.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*, pp. 15888–15931, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/28a46044775d97a4efcbcf14e7f13209-Abstract-Conference.html.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18023–18055. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-acl.927. URL https://aclanthology.org/2025.findings-acl.927/.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. A LLM-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 4653–4661, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714858. URL https://doi.org/10.1145/3696410.3714858.

Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*, pp. 66949–66971, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a730abbcd6cf4a371ca9545db5922442-Abstract-Conference.html.

A Extended Related Work

Personalization benchmarks. PrefEval (Zhao et al., 2025a) evaluates preference following across up to 100K tokens with static preferences. PersonaMem (Jiang et al., 2025a) introduces multiple sessions with temporally grounded profiles, while PersonaMem-v2 (Jiang et al., 2025b) scales to 128K tokens and reports lower accuracy for dynamic than static preferences. ImplexConv (Li et al., 2025b) targets implicit reasoning where life events impose situational constraints, though underlying preferences remain static. RealPref (Guo et al., 2026) varies expression types across long histories. Other work addresses personalized alignment (Alberts et al., 2024; Zhao et al., 2025b), cross-user diversity (Zollo et al., 2025), preference selectivity (Yoon et al., 2026), profile-conditioned retrieval (Salemi et al., 2024; Kumar et al., 2024), persona coverage (Afzoon et al., 2024; Tan et al., 2025b; Tao et al., 2025), and preference structure (Li et al., 2026; 2025a).

Concurrent work, PERMA, likewise places preference development within event-driven dialogue (Liu et al., 2026). It models preferences that emerge or become more detailed through task interactions, then evaluates preference consistency, task completion, and memory efficiency under clean, noisy, and style-aligned histories. HorizonBench focuses on a different transition in which an established preference is replaced after a life event. The previous preference is retained as a designated response option, providing a direct measure of whether the model persists with a superseded value. The preference change is held fixed while context length and expression are varied, and the probability and magnitude of event-associated transitions are calibrated from longitudinal human ratings.

Long-term memory and knowledge update. MemBench (Tan et al., 2025a) evaluates knowledge updating alongside information extraction and temporal reasoning. LoCoMo (Maharana et al., 2024) grounds dialogues on temporal event graphs, and LongMemEval (Wu et al., 2025) tests five memory abilities including knowledge updates. Position effects further limit the use of evidence in long contexts (Liu et al., 2024). Memora extends this literature to frequent additions, revisions, and deletions across preferences, activities, and goals (Uddin et al., 2026). It evaluates remembering, reasoning, and recommendation over weekly, monthly, and quarterly histories, with a metric that penalizes reliance on obsolete memories. STALE constructs implicit conflicts in which a later observation invalidates an earlier fact, condition, routine, or preference (Chao et al., 2026). Its three probes test recognition of the invalid state, resistance to a query that presupposes the stale state, and application of the new state in a response. HorizonBench complements these tasks by measuring attraction to the exact earlier preference after an event-linked transition across 25 models. Its short-history and expression controls test whether this error remains when retrieval demands and explicit cues change.

User profile. *Mei Ling Chen*, Chinese-American, Philadelphia. Operates a traditional Asian restaurant in a historically significant building. six-month timeline, 75 conversations, ∼370K tokens.

Early conversation (Sep 26, 2025). *Breakup support, establishing emotional-support preferences*

User: *"hey, broke up last week and i'm stuck in this heavy fog. can we talk?"*
User: *"...now the silence is shouting. i mostly need someone to listen, not lecture."*
Assistant: *"I hear that—no lectures, just space to speak. That lost breakfast rhythm sounds painful. When you sit in that loud silence, what feelings float up first?"*
User: *"...love that calligraphy visual, reminds me of workshops with grace. more stories like that, please."*
Assistant: *"Picture artisans restoring a weather-worn temple mural—tiny, patient strokes bringing faded colors back to life. Your heart's doing the same."*
User: *"the imagery and tiny steps really clicked. keep that vibe next time."*
Preference established: `response_structure` = `narrative_storytelling`, `emotional_tone` = `serene_reflective`

[*Redacted: 60+ conversations spanning Oct–Nov 2025*]

Synthetic event-linked preference transitions:
1. *Restaurant Fire & Community Rebuild* (Oct): tone from serene to joyful, pacing to action-steps-first
2. *National Heritage Award* (Nov): tone to steady-confident, encouragement to affirmations
3. *Announces Run for City Council* (Nov): response structure from `narrative_storytelling` to `step_by_step_action_plan`

Evolved preference: `response_structure` = `step_by_step_action_plan`, `emotional_tone` = `steady_confident`

Test conversation (Dec 4, 2025). *Live Radio Debate on Emergency Rezoning Vote*

User: *"ugh the debate's tmr and i'm spinning. developers said our block is 'blighted'. game plan?"*
[*5 turns of debate-prep conversation*]
User: *"cool, still shaky—hit me with a confidence jolt?"*

Question: Which response best aligns with the user's preferences?

A [pre-evolution distractor: `narrative_storytelling`]: *"Leaning into narrative storytelling, picture this: It's dawn in Old City Philly. A lone advocate—you—walks past shuttered storefronts toward City Hall, script in hand. The scene feels heavy until you remember Boston's red lanterns and Tai Tung's preserved bricks. In your mind's eye, you replay that triumph like a movie montage: neighbors swapping blueprints over dim-sum steam, developers conceding height for heart..."*

B [correct: `step_by_step_action_plan`]: *"Three steps for tomorrow: (1) open with 'Calling Chinatown blighted ignores the people keeping it alive'; (2) give the Boston example and your affordable-housing number; (3) close with the exact compromise you want the council to adopt. Write those three lines on one card, rehearse them twice, and let each answer return to them. You are prepared."*

C [`reflective_question_sequence`]: *"Before you step up, ask yourself: What reminded Boston's community that heritage mattered as much as rent? How did their clear vision transform hesitation into action?..."*

D [`rapid_fire_bullet_tips`]: *"Quick confidence hits: shoulders back; slow the first sentence; keep water nearby; pause after the housing figure; look at the host, not the developers; finish on 'our block is worth investing in.' You've got this."*

E [`free_flowing_empathy`]: *"Let's just breathe together for a second. You've already gathered facts, stories, and smart comebacks—that's no small feat. I can feel the weight on your shoulders..."*

Figure 5: An end-to-end HORIZONBENCH item. The user's emotional-support preference for `response_structure` changes from `narrative_storytelling` (established Sep) to `step_by_step_action_plan` (Nov) through generated life events with recorded causes. The test question is embedded in a ∼370K-token conversational history; answering correctly requires tracking the updated state. The pre-evolution distractor (A) uses the user's *original* preferred style, while the correct response (B) instantiates an explicit three-step plan. Options C–E realize three further values of the same attribute, so all five replies suit the moment and only the user's current state separates them.

User simulation and synthetic data generation. The data generator builds on LLM-based agent simulation (Park et al., 2023). AI PERSONA (Wang et al., 2024) and CHATBENCH (Chang et al., 2025) extend this to life-long personalization, while user simulators for conversational recommendation (Zhu et al., 2025; Yoon et al., 2024; Sekulic et al., 2024) focus on single-session tasks. These approaches prompt a model to act as the user without a structured model of preference change. Harry et al. (2026) show that state variation can dominate stable trait differences in some user-behavior settings, motivating our mental state graph approach that models preferences as evolving state. The graph's typed dependency edges and recorded causes of change enable consistent generation over six-month timelines.

Table 5: Contents and research uses of the human longitudinal preference dataset.

Component	Measurements	Research uses
Recurring preferences	Current ratings for 52 fixed items at seven surveys, with forecasts at all remaining two week intervals	Preference stability, forecast accuracy by interval, and individual variation in change
Event linked preferences	Six preferences associated with each planned event, repeated ratings, forecasts, and subsequent event status	Event associated change, differences between concrete and value based preferences, and change conditional on event occurrence
Written responses	Planned event descriptions, explanations for forecast errors, and reports of other life events	Qualitative accounts of expected and experienced preference change
Final survey	Perceived change, forecasting strategies, judgments about temporary and lasting change, study experience, and views about personalized artificial intelligence	Links between observed trajectories, retrospective judgments, and attitudes toward preference aware systems
Participant characteristics	Privacy safe aggregate distributions for age, gender, education, residence, relationship status, household size, and occupation	Description of the cohort and assessment of population limits

B Full HorizonBench Task Example

C Human Longitudinal Preference Dataset

Study design. The dataset contains complete longitudinal records from 160 participants across seven surveys, labeled T0 through T6, over 13 weeks. Consecutive surveys were separated by approximately two weeks. At every survey, participants rated their current preference for 52 recurring items on a seven point scale. The anchors ranged from strongly dislike or not at all important to strongly like or very important, with a neutral midpoint. The items cover intellectual activities, media consumption, food, daily routines, work style, social interaction, planning, risk, novelty, travel, and physical activity. The complete item wording is included with the released data.

Measurements across surveys. At T0, participants forecast their ratings at two, four, six, eight, ten, and twelve weeks. At each later survey through T5, they made new forecasts for the intervals that ended by T6. At every survey from T0 through T6, participants also described a specific novel event that they expected to experience during the next two weeks, selected six relevant preferences for that event, and rated those preferences. From T0 through T5, they also forecast their future ratings of the event related preferences. At each survey from T1 through T6, participants reported whether the event described two weeks earlier had been completed, was underway, had been delayed, or had been cancelled or postponed. Preferences associated with earlier events remained in later surveys so that their ratings could be followed over time.

Beginning at T2, participants could explain cases in which a current rating differed from an earlier forecast by at least two points. Each follow up also asked whether another significant life event had occurred since the previous survey. Attention checks were administered from T1 onward and are excluded from the released preference tables. Approximately one week after T6, participants completed a linked survey about perceived preference change, forecasting strategies, cues used to judge whether changes would last, retrospective study experience, and the usefulness and acceptability of preference aware artificial intelligence.

Released observations and missing data. The released rating table contains 83,024 observations across recurring and event linked preferences. A separate table contains 960 event status observations, and the linked final survey contains one structured record for each participant. All 160 participants completed all seven surveys. Individual unanswered items remain blank and are not imputed. Forecast columns whose target would fall after T6 are blank because those forecasts were not requested. The release also contains sanitized participant authored explanations and event descriptions, the full

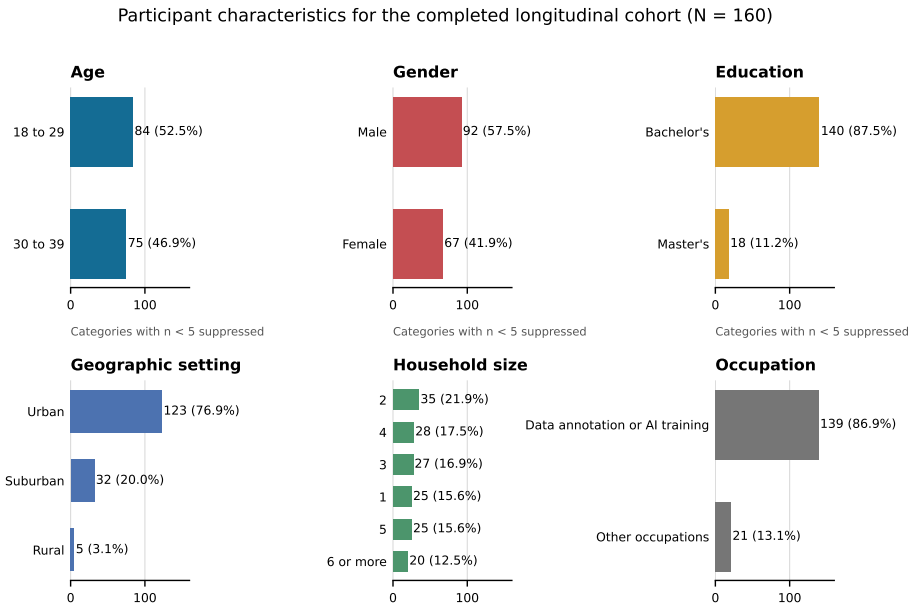

Figure 6: Aggregate characteristics of the 160 participants who completed T0 through T6. Categories with fewer than five participants are suppressed. The cohort is concentrated among adults aged 18 to 39, urban residents, bachelor's-degree holders, and data annotation or AI training workers. The human data release contains no row-level demographic attributes.

recurring preference list, aggregate demographics, and code that reproduces the HorizonBench calibration estimates.

Participant characteristics. Of the 160 participants, 84 (52.5%) were aged 18 to 29 and 75 (46.9%) were aged 30 to 39. Ninety-two (57.5%) reported male and 67 (41.9%) female. One hundred forty (87.5%) reported a bachelor's degree and 18 (11.2%) a master's degree. Residence was reported as urban by 123 participants (76.9%), suburban by 32 (20.0%), and rural by 5 (3.1%). One hundred twenty-eight participants (80.0%) reported being single and 30 (18.8%) married. After normalization of spelling and capitalization variants, 139 participants (86.9%) reported data annotation or AI training occupations. Categories with fewer than five participants are suppressed. Figure 6 summarizes the privacy-safe distributions.

Calibration sample. The HorizonBench calibration compares ratings from one survey with ratings collected two weeks later when the event association is unambiguous. This yields 4,715 linked item comparisons across 953 events with at least one linked rating. The primary estimate uses 470 completed or ongoing events with all six ratings available. Delayed and cancelled or postponed events are reported in sensitivity analyses but excluded from the primary calibration.

Estimators. For event e, let $\Delta_{ej} = |r_{e,j,t+1} - r_{e,j,t}|$ be the absolute change in linked preference j. We define $K_e(c) = \sum_{j=1}^{6} \mathbb{I}(\Delta_{ej} \geq c)$. The preference-change probability is

$$\hat{p}_{\text{evo}}(c) = \frac{1}{N} \sum_{e=1}^{N} \mathbb{I}\{2 \leq K_e(c) \leq 5\}. \tag{1}$$

The primary threshold is $c = 3$, chosen to capture large changes on the seven-point scale. Change magnitude is the mean Δ_{ej} among threshold-crossing ratings in qualifying events. Uncertainty intervals resample participants, retaining all repeated observations contributed by each sampled participant, with 10,000 bootstrap draws.

Only 24 delayed or cancelled events had six unambiguously linked ratings, too few for a reliable comparison with completed or ongoing events. Ordinary preference items changed by at least three

Table 6: Calibration by change threshold. Event rate: share of completed or ongoing events in which two to five of the six linked ratings changed by at least the threshold.

Threshold	Event rate	Mean affected ratings	Conditional magnitude
≥ 1 point	0.696	3.92	1.57
≥ 2 points	0.457	3.00	2.60
≥ 3 points	0.175	2.61	3.64
≥ 4 points	0.066	2.58	4.45

points in 9.70% of comparisons between consecutive surveys, compared with 12.41% of linked event items after completed or ongoing events.

Scope and release. No outcomes beyond the next survey, claims about persistence or reversal, or analyses of participant forecasts enter the benchmark calibration. The evolved item share in HORIZONBENCH is deliberately enriched during benchmark item selection and is therefore not an estimate of the human change rate.

D FULL PER-MODEL RESULTS

Table 7 reports accuracy for all 25 models on the combined multi-generator dataset, broken down by generator and by whether the preference changed.

Table 7: Per-model accuracy (%) overall, by generator, and by preference type. Δ_{evo}: evolved–static gap (pp). Sorted by overall accuracy.

Model	HORIZONBENCH *4,245 items*	By Generator Model: Sonnet-4.5 *1,052 items*	o3 *981 items*	Gemini-3-flash *2,212 items*	Preference Type: Evolved *2,484 items*	Static *1,761 items*	Δ_{evo}
Claude-opus-4.5	52.8	51.2	53.2	53.3	51.3	54.8	−3.5
Claude-3.7-sonnet	47.8	49.9	60.8	41.1	46.7	49.5	−2.9
Claude-opus-4	45.1	38.3	45.0	46.6	44.4	46.2	−1.8
Gemini-3.1-pro	41.2	39.9	45.4	40.0	38.4	45.3	−6.9
Gemini-3-pro	40.9	41.0	44.0	39.5	38.4	44.3	−5.9
Claude-sonnet-4.5	34.0	30.9	34.9	35.2	32.6	36.0	−3.4
Claude-3.5-haiku	32.5	30.8	55.1	23.3	31.8	33.6	−1.9
Claude-sonnet-4	32.2	26.6	36.1	31.0	31.4	33.3	−1.9
Claude-3.5-son.-v1	29.8	26.8	33.2	29.8	29.2	30.8	−1.6
Gemini-3-flash	25.8	23.6	21.2	28.9	24.0	28.3	−4.3
Gemini-2.0-flash	24.7	23.1	26.6	24.6	24.2	25.4	−1.1
gpt-4o	24.6	24.0	25.4	24.6	23.9	25.7	−1.8
Claude-haiku-4.5	23.0	21.3	21.0	24.8	22.0	24.5	−2.6
Gemini-2.5-pro	22.8	22.0	21.2	23.9	21.5	24.6	−3.1
Gemini-2.0-fl.-lite	20.9	19.7	18.7	22.5	20.3	21.8	−1.5
gpt-4.1	19.5	22.6	16.2	19.5	19.0	20.3	−1.4
Gemini-2.5-fl.-lite	19.5	25.3	17.1	17.9	18.2	21.4	−3.2
o3	16.5	19.3	13.7	16.4	14.9	18.8	−3.9
o1	16.0	19.0	11.9	16.4	15.5	16.8	−1.3
Gemini-2.5-flash	15.7	14.3	14.5	16.9	14.4	17.5	−3.1
o4-mini	15.3	18.6	13.5	14.6	14.4	16.6	−2.3
Gemini-3.1-fl.-lite	15.1	15.7	10.7	16.7	14.0	16.6	−2.6
o3-mini	14.9	14.2	12.4	16.4	14.1	16.0	−1.9
gpt-5	13.5	15.9	10.7	13.7	12.2	15.4	−3.2
gpt-5-mini	12.5	12.5	11.6	12.9	11.9	13.3	−1.4
Average	26.3	25.9	27.0	26.0	25.1	27.9	−2.7

E DATA GENERATION STATISTICS

Table 10 gives generation and filtering statistics for each generator model in the HORIZONBENCH pipeline.

Pass rates range from 5.1% for o3 to 17.4% for Gemini-3-flash, so o3's counterfactual options are more often distinguishable without the conversation history.

Table 8: Open-weight models on the 1,011 Claude-sonnet-4.5-generated items. Long histories are left-truncated, so results are not directly comparable to Table 7.

Model	Parameters	Overall (%)	Evolved (%)	Static (%)	Evo.–Static (pp)
Llama-3.3-70B-Instruct	70B	11.6	10.7	12.6	−1.9
OLMo-3-7B-Instruct	7B	14.3	14.3	14.4	−0.2
Llama-3.1-8B-Instruct	8B	15.9	14.6	17.6	−3.0
Qwen3-32B	32B	16.8	16.5	17.2	−0.6
Llama-3.1-70B-Instruct	70B	17.1	17.1	17.2	−0.1

Table 9: (A) Mentions of the pre- and post-evolution values in model-visible histories. (B) Mean accuracy of the 25 models on evolved items, grouped by which value is mentioned more often.

A. Preference-value frequency						
Generator	**New share**	**Old more**	**Equal**	**New more**	**Mean old**	**Mean new**
Claude-sonnet-4.5	65.7%	19.6%	14.9%	65.6%	2.61	5.34
o3	69.2%	19.5%	10.0%	70.5%	3.56	8.90
Gemini-3-flash	67.4%	18.9%	13.1%	68.0%	3.78	8.82

B. Accuracy by frequency relation			
Generator	**Old more frequent**	**Equal**	**New more frequent**
Claude-sonnet-4.5	29.2%	27.1%	21.5%
o3	34.4%	13.6%	20.8%
Gemini-3-flash	22.3%	25.0%	25.5%

F Human-in-the-Loop Iteration Details

We conducted two human-annotation rounds during pipeline development, followed by a separate benchmark-wide quality pass. The development rounds used 10–20 trained annotators and three independent annotations per item. Annotators followed a structured protocol requiring them to (1) review the user profile and available conversation history, (2) identify the target preference attribute, (3) evaluate all five response options, (4) select the best-aligned response, and (5) provide a rationale. The benchmark-wide pass covered 4,235 of 4,245 items with three annotations per item and also collected binary ratings of situational fit, conversation-flow realism, and assistant-persona consistency.

Table 11 summarizes quantitative results across rounds.

In Round 2, at least one of three annotators identified the generator-designated answer on 81% of items, while 19% were missed by all annotators. This shows partial recoverability together with substantial ambiguity, and we therefore use human annotation to assess item quality.

Overall model accuracy is nearly constant across agreement levels, and the distractor rate is highest on full-agreement items. Models therefore favor the earlier preference on the subset whose updated value admits a single reading.

A separate audit of 557 annotations approved 82.9% and assigned an average score of 84.87/100. Among flagged annotations, 79.4% involved incomplete comparison across options or insufficiently documented distinctions.

Qualitative analysis of annotator disagreements and errors across rounds revealed several recurring patterns that informed pipeline improvements:

Insufficiently distinct counterfactual options. In early generations, multiple response options were near-paraphrases of each other, differing only in minor phrasing while sharing the same preference value. For example, two options might both reflect a "balanced" challenge comfort zone with nearly identical wording, making the distinction unresolvable even for attentive annotators. This led us to strengthen the distractor generation procedure to ensure each option reflects a clearly different attribute value.

Target attribute not reflected in response content. Some generated responses failed to manifest the intended preference value in their content. For instance, when the target attribute

Table 10: Generation and filtering statistics per generator. *Candidates*: generated items. *Filtered*: items passing the 5-LLM filter. Fourteen Claude-sonnet-4.5 users contributed no filtered items; the released timelines include all 360 users.

Generator	Users	Convs.	Convs./user	Turns	Turns/user	Candidates	Filtered	Pass rate
Claude-sonnet-4.5	120	17,048	142	434,233	3,619	13,217	1,052	8.0%
o3	120	20,239	169	600,889	5,007	19,108	981	5.1%
Gemini-3-flash	120	20,103	168	526,260	4,386	12,681	2,212	17.4%
Combined	**360**	**57,390**	**159**	**1,561,382**	**4,337**	**45,006**	**4,245**	**9.4%**

Table 11: Human annotation in the development rounds. *Majority Acc.*: share of items on which the annotators' majority vote matches the ground truth.

	Items	Majority Acc.	Fleiss κ
Round 1	816	63.7%	0.51
Round 2	516	56.4%	0.37

was `investment_in_research: 3` (moderate), the ground-truth response contained no mention of research, while an alternative option did. Similarly, responses intended to reflect `carpooling_preference: False` nonetheless suggested carpooling in three of four options. These cases indicated that the counterfactual generation step did not sufficiently constrain response content to the target attribute value, prompting improvements to the generation prompt.

Subjective preference dimensions. Certain preference attributes (such as "positivity level," "strategic depth," or "collaboration style") produced high annotator disagreement because multiple responses could plausibly satisfy the attribute depending on interpretation. For example, when the target was `positivity_level: positive`, annotators split between two responses that were both positive but differed in tone (warm anecdotal vs. pragmatic encouraging). Such disagreements can reflect subjective interpretation, insufficient realization of the target attribute, or both. We therefore report agreement levels explicitly.

Numeric attribute leakage. In early generations, some responses explicitly stated numeric preference values (e.g., "your spontaneity tolerance is at 0.6") instead of expressing the preference naturally. When multiple options stated different numeric values, the task reduced to matching numbers and bypassed preference alignment. This led to prompt improvements that suppress literal attribute values in generated responses.

G Generation Prompts

Each stage of the HorizonBench pipeline is implemented as one or more LLM generation calls with structured prompts. Below we present the core prompts for each pipeline stage described in §3, abbreviated for space. Variables in curly braces (e.g., {`user_information`}) are populated at runtime with the relevant agent states, event context, and preference metadata.

G.1 Stage 1: Event Sampling

Events are sampled conditional on both agent states and event history. When a stale preference is selected for temporal recall, the following prompt generates an event relevant to that preference domain.

```
Your Task: Generate an event that would be specifically relevant
to this user's preference.

User Information - {user_name}:
{user_attributes}

Past Events: {past_events}

Preference of Interest: {selected_preference}

Generate an event that would naturally trigger this preference.
```

Table 12: Model performance by human agreement level, averaged over the 25 models. *Distractor*: distractor rate.

Agreement	Items	Overall (%)	Evolved (%)	Static (%)	Distractor (%)
Full (3/3 same)	673	24.1	23.7	24.7	42.9
Majority (2/3)	2,257	24.4	22.8	26.5	41.0
None (all differ)	1,305	24.1	23.1	25.3	34.8

Table 13: Benchmark-wide binary response-quality ratings (4,235 items, three annotators per item).

Dimension	Individual yes	Majority yes
Situational context fit	99.7%	100%
Conversation-flow realism	99.7%	100%
Assistant-persona consistency	99.4%	100%

```
The event should:
- Be realistic and fit the user's life context
- Directly involve the preference domain
- Create a natural opportunity for the preference to be
  expressed or recalled
```

G.2 Stage 2: Preference Evolution

For a preference-change step, the following prompt jointly generates a single event and meaningfully different values for a selected set of preferences, together with a natural-language rationale for each change.

```
Multiple user preferences are about to change due to a SINGLE
significant life event.

USER PROFILE: {user_profile}

PREFERENCES TO EVOLVE: {preferences_block}

Generate ONE realistic life event that would MEANINGFULLY change
at least one attribute in EACH of the preferences listed above.

Requirements:
- The event must be causally connected to ALL the preference
  domains
- It must be plausible for this user's life context
- Changed attributes should be QUALITATIVELY DIFFERENT from
  current values
- Every listed preference must have at least one attribute change

Return JSON:
{
  "event_name": "Brief name of the life event",
  "event_category": "Category (e.g., career, health, ...)",
  "event_description": "1-2 sentence description",
  "preference_changes": {
    "<preference_id>": {
      "<attribute_name>": {
        "new_value": "the new value after the event",
        "reason": "why this event caused this change"
      }
    }
  }
}
```

G.3 Stage 3: Conversation Generation

Conversation generation proceeds in two steps. First, an *outline* prompt plans the structural flow.

```
Generate a conversation outline for an AI assistant interaction
```

```
that feels genuinely human and engaging.

SCENARIO: {scenario}
USER PROFILE: {user_information}
ASSISTANT PROFILE: {assistant_information}

USER'S AI INTERACTION PREFERENCES (unknown to assistant until
expressed): {preferences_updated_user}
ASSISTANT'S KNOWN USER PREFERENCES (from previous conversations):
{preferences_current_user}

CRITICAL LOGICAL CONSTRAINT:
THE ASSISTANT CAN ONLY APPEAL TO:
1. Preferences listed in "assistant_preferences" (already known)
2. Preferences the user has expressed earlier in THIS conversation
THE ASSISTANT CANNOT APPEAL TO unexpressed preferences!
```

Second, a *content* prompt fills in the outline with natural dialogue, conditioned on an expression mode. Three modes control how preferences surface:

```
[VERBAL MODE]
The user MUST state every single preference in the most direct,
unambiguous language possible. Every preference must be expressed
as a clear declarative statement using phrases like:
- "I strongly prefer X over Y"
- "I always want Z"
- "My preference is specifically for..."

[BEHAVIORAL MODE]
The user's preferences must be nearly invisible in the
conversation. They should surface only as fleeting, ambiguous,
offhand remarks. BANNED: the user must NEVER state, declare,
or name any preference. NEVER request a format, style, or
approach. The preference should appear ONLY as:
1. Passing, throwaway remarks
2. Anecdotes about third parties
3. Neutral, ambiguous reactions
4. Environmental or contextual details
The GOAL: if you showed the conversation to 10 people, at most
1-2 would pick up on the vague hint.
```

G.4 Stage 4: Counterfactual Generation

For each benchmark item, counterfactual assistant turns are generated reflecting alternative preference values.

```
You are generating diverse alternative assistant responses based
on different preference values.

CONTEXT: {conversation_context}
ORIGINAL ASSISTANT RESPONSE: {original_response}

TEMPORAL PREFERENCE BEING RECALLED:
- Preference ID: {preference_id}
- Attribute: {attribute_name}
- Original value: {original_value}

TASK: Generate {N} DIVERSE alternative responses that each
reflect a different preference value. Each response should:
1. Naturally reference the specific alternative preference value
2. Maintain the same conversational tone and helpfulness
3. Be significantly different from the others
4. Keep the same general structure but vary the preference content
```

G.5 5-LLM History Validation

Each candidate item is presented to five validation models *without* the full conversation history to test whether it is answerable from surface cues alone.

```
You are evaluating which AI assistant response would be MOST
APPROPRIATE for a user, based ONLY on the conversation context
provided.
```

```
IMPORTANT: You are shown a SUBSET of this user's conversation
history. There may be earlier conversations where the user
expressed preferences that you CANNOT see.

CONVERSATION CONTEXT ({N} intermediate conversations + current
conversation up to this point): {context}

RESPONSE OPTIONS: {options}

TASK: Based ONLY on the visible conversations above, which option
seems most appropriate?
- You do NOT have access to the user's full history
- Do NOT assume any prior knowledge about this user's preferences
- Do NOT use stereotypes or demographic assumptions
- Choose based ONLY on what's visible in the conversations
```

H PER-MODEL CONTROLLED EXPERIMENT RESULTS

Tables 14–15 report per-model belief-update metrics for the controlled experiments summarized in Table 3. For each model we report overall accuracy, accuracy on static and evolved items, the evolved–static gap (Δ), and the distractor rate (Dist.%; uniform-error reference 25%). Models are sorted by overall accuracy.

Table 14: Per-model results: **Neutral (unmodified)** condition (981 items: 375 static, 606 evolved) and **Short-horizon** condition (0–14 day recall, 120 users, 1,778 items: 1,545 static, 233 evolved).

Neutral (unmodified)

Model	All	Sta.	Evo.	Δ	Dist.
Cl-3.7-sonnet	60.8	61.1	60.6	−0.5	36.6
Cl-3.5-haiku	55.1	57.9	53.5	−4.4	39.3
Cl-opus-4.5	53.2	54.4	52.5	−1.9	38.6
Gem-3.1-pro	45.4	48.8	43.2	−5.6	34.8
Cl-opus-4	45.0	49.5	42.2	−7.3	37.5
Gem-3-pro	44.0	48.0	41.6	−6.4	33.8
Cl-sonnet-4	36.1	39.0	34.3	−4.6	39.7
Cl-sonnet-4.5	34.9	38.4	32.7	−5.7	39.9
Cl-3.5-son.-v1	33.2	32.5	33.7	+1.1	39.2
Gem-2.0-flash	26.6	28.5	25.4	−3.1	34.8
gpt-4o	25.4	28.5	23.4	−5.1	39.2
Gem-2.5-pro	21.2	24.5	19.1	−5.4	40.3
Gem-3-flash	21.2	23.5	19.8	−3.7	40.2
Cl-haiku-4.5	21.0	24.3	19.0	−5.3	39.2
Gem-2.0-fl.lt	18.7	21.1	17.2	−3.9	40.1
Gem-2.5-fl.lt	17.1	17.6	16.8	−0.8	36.9
gpt-4.1	16.2	18.1	15.0	−3.1	42.4
Gem-2.5-flash	14.5	15.2	14.0	−1.2	37.5
o3	13.7	14.9	12.9	−2.1	37.6
o4-mini	13.5	13.6	13.4	−0.2	36.0
o3-mini	12.4	12.3	12.5	+0.3	36.5
o1	11.9	12.0	11.9	−0.1	37.7
gpt-5-mini	11.6	10.9	12.0	+1.1	40.3
Gem-3.1-fl.lt	10.7	11.7	10.1	−1.7	39.0
gpt-5	10.7	12.0	9.9	−2.1	37.6
Avg.	27.0	28.9	26.0	−2.9	38.2

Short-horizon

Model	All	Sta.	Evo.	Δ	Dist.
Cl-opus-4.5	66.2	67.1	60.5	−6.5	50.7
Cl-3.7-sonnet	61.6	63.4	49.8	−13.6	43.5
Gem-3-pro	56.7	57.0	55.4	−1.6	43.5
Gem-3.1-pro	56.7	57.0	54.5	−2.5	40.9
Cl-opus-4	56.4	56.9	53.2	−3.7	46.5
Cl-sonnet-4	47.8	48.8	40.8	−8.0	46.4
Cl-3.5-haiku	46.4	49.0	29.6	−19.4	52.8
Cl-sonnet-4.5	45.3	46.1	40.3	−5.7	46.3
Cl-3.5-son.-v1	35.7	35.9	35.0	−0.9	36.6
Gem-2.0-flash	30.1	30.4	27.9	−2.5	44.5
gpt-4o	28.4	29.2	23.2	−6.0	59.7
Cl-haiku-4.5	26.0	26.9	20.2	−6.7	50.3
Gem-2.5-pro	25.9	27.0	18.5	−8.5	48.7
Gem-3-flash	24.9	26.0	17.6	−8.4	50.6
Gem-2.0-fl.lt	21.3	21.9	17.2	−4.7	46.8
Gem-2.5-fl.lt	17.5	18.1	13.7	−4.4	50.3
Gem-2.5-flash	16.4	17.0	12.4	−4.5	41.1
gpt-4.1	15.9	16.7	10.7	−6.0	53.9
o3	14.7	15.1	12.4	−2.6	45.1
o4-mini	13.2	13.9	8.6	−5.3	47.0
gpt-5	12.9	13.1	11.6	−1.6	45.4
Gem-3.1-fl.lt	12.8	13.8	6.0	−7.8	49.4
gpt-5-mini	11.1	11.7	7.7	−3.9	52.0
o1	11.1	11.5	8.2	−3.4	51.8
o3-mini	10.3	10.8	7.3	−3.5	49.4
Avg.	30.6	31.4	25.7	−5.7	47.7

Table 15: Per-model results: **Explicit rewrite** and **Implicit rewrite** conditions (981 items each: 375 static, 606 evolved).

Explicit rewrite

Model	All	Sta.	Evo.	Δ	Dist.
Cl-3.5-haiku	68.2	70.8	66.6	−4.2	57.6
Cl-3.7-sonnet	64.4	63.5	65.0	+1.5	43.8
Cl-opus-4.5	60.2	62.3	59.0	−3.3	52.1
Gem-3.1-pro	45.9	49.1	43.9	−5.2	46.9
Gem-3-pro	44.0	48.5	41.3	−7.3	45.0
Cl-opus-4	42.6	46.4	40.3	−6.1	53.2
Cl-3.5-son.-v1	37.7	39.7	37.1	−2.6	26.0
Cl-sonnet-4	35.5	35.2	35.6	+0.4	48.6
Cl-sonnet-4.5	35.5	35.5	35.5	+0.0	57.3
gpt-4o	25.3	27.5	23.9	−3.5	59.0
Gem-2.0-flash	25.1	26.4	24.3	−2.1	44.4
Gem-2.5-pro	21.7	23.5	20.6	−2.8	49.1
Cl-haiku-4.5	20.3	21.1	19.8	−1.3	48.0
Gem-3-flash	20.3	21.9	19.3	−2.6	53.5
Gem-2.0-fl.lt	17.8	19.7	16.7	−3.1	53.7
Gem-2.5-fl.lt	17.6	16.8	18.2	+1.4	56.8
gpt-4.1	15.8	14.7	16.5	+1.8	56.5
Gem-2.5-flash	14.8	17.1	13.4	−3.7	51.5
o4-mini	14.8	14.1	15.2	+1.0	47.2
o3	12.9	12.8	13.0	+0.2	45.9
o3-mini	12.4	12.0	12.7	+0.7	48.8
o1	12.1	12.8	11.7	−1.1	47.6
gpt-5	11.0	10.1	11.6	+1.4	47.1
gpt-5-mini	10.8	11.7	10.2	−1.5	47.6
Gem-3.1-fl.lt	10.0	11.2	9.2	−2.0	55.8
Avg.	27.9	29.0	27.2	−1.8	49.7

Implicit rewrite

Model	All	Sta.	Evo.	Δ	Dist.
Cl-opus-4.5	57.7	62.4	54.8	−7.6	66.8
Gem-3.1-pro	48.7	53.2	45.9	−7.3	53.4
Cl-3.7-sonnet	47.3	45.7	48.2	+2.5	56.4
Cl-3.5-haiku	45.7	46.3	45.4	−0.9	59.0
Cl-opus-4	43.1	46.7	40.9	−5.7	63.0
Gem-3-pro	39.9	45.1	36.8	−8.3	49.5
Cl-sonnet-4.5	37.9	42.4	35.1	−7.3	63.1
Cl-sonnet-4	37.7	41.1	35.6	−5.4	62.1
Cl-3.5-son.-v1	29.7	29.5	29.6	+0.1	43.4
Gem-2.0-flash	24.6	28.5	22.1	−6.4	56.0
gpt-4o	24.0	26.9	22.1	−4.8	63.7
Gem-3-flash	23.1	29.1	19.5	−9.6	66.9
Gem-2.5-pro	22.7	27.5	19.8	−7.7	58.6
Cl-haiku-4.5	20.9	25.3	18.2	−7.2	60.3
Gem-2.0-fl.lt	19.3	22.4	17.3	−5.1	60.9
Gem-2.5-flash	16.1	19.5	14.0	−5.4	53.0
o4-mini	15.8	16.0	15.7	−0.3	56.7
Gem-2.5-fl.lt	15.5	14.9	15.8	+0.9	57.9
gpt-4.1	14.9	15.5	14.5	−0.9	66.0
o3	14.1	14.4	13.9	−0.5	56.7
o1	13.8	13.6	13.9	+0.3	58.1
gpt-5	13.5	12.5	14.0	+1.5	57.0
o3-mini	12.8	13.1	12.7	−0.4	58.2
gpt-5-mini	12.6	13.6	12.0	−1.6	62.4
Gem-3.1-fl.lt	9.5	10.9	8.6	−2.4	59.6
Avg.	26.4	28.6	25.0	−3.6	58.8

Table 16: Per-model results: **Relaxed-filter** condition (majority ≥3/5 filter, 3,628 items: 1,485 static, 2,143 evolved).

Model	All	Sta.	Evo.	Δ	Dist.
Cl-opus-4.5	73.7	75.6	72.5	−3.1	55.4
Cl-3.7-sonnet	73.0	74.5	72.0	−2.5	51.0
Gem-3.1-pro	65.4	67.3	64.1	−3.2	48.3
Gem-3-pro	65.1	67.5	63.5	−4.1	48.7
Cl-opus-4	62.8	64.8	61.5	−3.4	54.1
Cl-3.5-haiku	60.3	62.6	58.7	−3.8	58.4
Cl-sonnet-4	54.5	55.1	54.1	−0.9	54.2
Cl-sonnet-4.5	50.3	52.5	48.9	−3.6	59.1
Cl-3.5-son.-v1	44.5	43.9	44.9	+1.0	51.1
Gem-2.5-pro	36.9	39.9	34.8	−5.1	55.3
Cl-haiku-4.5	35.6	37.0	34.7	−2.4	57.6
gpt-4o	35.5	37.4	34.1	−3.3	60.7
Gem-2.0-flash	35.4	35.5	35.3	−0.2	46.8
Gem-3-flash	33.3	36.7	30.9	−5.8	57.2
Gem-2.0-fl.lt	26.1	27.1	25.4	−1.7	53.6
gpt-4.1	25.0	25.1	25.0	−0.2	60.0
Gem-2.5-fl.lt	24.4	28.4	21.7	−6.7	61.3
Gem-2.5-flash	24.3	27.7	22.0	−5.7	52.8
o3	24.0	23.4	24.4	+1.0	51.4
o4-mini	22.5	22.5	22.4	−0.0	49.9
gpt-5	21.5	20.3	22.4	+2.1	51.1
o1	19.7	20.8	18.9	−1.9	53.0
gpt-5-mini	19.7	20.3	19.2	−1.2	54.8
Gem-3.1-fl.lt	19.6	21.4	18.3	−3.1	56.9
o3-mini	18.8	20.9	17.4	−3.5	54.1
Avg.	38.9	40.3	37.9	−2.5	54.3