

# LIKEBENCH: EVALUATING SUBJECTIVE LIKABILITY IN LLMs FOR PERSONALIZATION

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

A personalized LLM should remember user facts, apply them correctly, and adapt over time to provide responses that the user prefers. Existing LLM personalization benchmarks are largely centered on two axes: accurately recalling user information and accurately applying remembered information in downstream tasks. We argue that a third axis, likability, is both subjective and central to user experience, yet under-measured by current benchmarks. To measure likability holistically, we introduce LikeBench<sup>1</sup>, a multi-session, dynamic evaluation framework that measures likability across multiple dimensions by how much an LLM can adapt over time to a user’s preferences to provide more likable responses. In LikeBench, the LLMs engage in conversation with a simulated user and learn preferences only from the ongoing dialogue. As the interaction unfolds, models try to adapt to responses, and after each turn, they are evaluated for likability across seven dimensions by the same simulated user. To the best of our knowledge, we are the first to decompose likability into multiple diagnostic metrics: emotional adaptation, formality matching, knowledge adaptation, reference understanding, conversation length fit, humor fit, and callback, which makes it easier to pinpoint where a model falls short. To make the simulated user more realistic and discriminative, LikeBench uses fine-grained, psychologically grounded descriptive personas rather than the coarse high/low trait rating based personas used in prior work. Our benchmark shows that strong memory performance does not guarantee high likability: DeepSeek R1, with lower memory accuracy (86%, 17 facts/profile), outperformed Qwen3 by 28% on likability score despite Qwen3’s higher memory accuracy (93%, 43 facts/profile). Even SOTA models like GPT-5 adapt well in short exchanges but show only limited robustness in longer, noisier interactions.

## 1 INTRODUCTION

As large language models (LLMs) become increasingly integrated into everyday life, the need for systems that genuinely adapt to individual users is more important than ever (Zhang et al., 2025; 2018). Personalized LLMs are widely recognized as the next step toward building AI that feels truly realistic and human-like, since “one-size-fits-all” alignment cannot capture the subjective preferences, values, and conversational styles of each user Kirk et al. (2024); Guan et al. (2025); Salemi et al. (2023). Moreover, as highlighted by OpenAI CEO in previewing GPT-6, the next generation of AI “*won’t just respond to users but will adapt to them, and allow people to create chatbots that mirror personal tastes,*” with the ultimate measure of success shifting toward how these interactions feel to users Sigalos (2025). This evolution places personalization and, crucially, the “likability” of AI responses—at the heart of next-generation systems, making it essential to develop rigorous evaluation methods that accurately capture and advance these user-centric capabilities.

However, despite the growing importance of personalization, existing benchmarks (Jiang et al., 2025; Wu et al., 2025; Maharana et al., 2024; Zhao et al., 2025) for LLMs remain largely focused on technical capabilities such as retaining user information (memory recall), for example, remembering that a user has a dairy allergy, and applying remembered preferences in tasks (memory adherence), such as recommending suitable foods while adhering food allergy memory. But these benchmarks cannot capture whether interactions genuinely feel personalized or likable to individual users. This

<sup>1</sup>Code & Data will be released after decision

Benchmark	Conversation	# Likability Metrics	Adaptation Over Time	Persona Modeling	Persona Facets	Memory Performance	#Profiles	#Sessions	#Turns
ALOE (Wu et al., 2024)	Dynamic	1	Yes	Coarse keyword	No	None	100	1	10
ALIGNX (Chan et al., 2024)	Static	1	No	Intensity ratings	No	None	3,716	1	1
CUPID (Kim et al., 2025)	Static	1	No	Intensity ratings	No	Implicit only	252	9	6
LikeBench (ours)	Dynamic	7	Yes	Fine-grained text	Yes	Explicit + Implicit	50	10	5

Table 1: Comparison of likability-oriented benchmarks across key dimensions.

gap presents a fundamental limitation: a model might achieve very good memory recall and adherence scores, yet still come across as generic or unsatisfying if it fails to adapt to a user’s personality, conversational style, and nuanced preferences. Moreover, since most LLMs are post-trained with RLHF algorithms, they tend to moderately satisfy many users rather than deeply satisfy any particular user, due to the distribution of their reward models.

There are very few works that attempt to measure this likability factor. However, they have several shortcomings: the user profiles or personas used to simulate users often lack fine-grained personality traits and conversational styles (Wu et al., 2024), which are critical for user’s likability. This results in LLMs being evaluated mainly on less discriminative users, even though real users have diverse interests and personalities. Importantly, these benchmarks typically reduce likability to a single aggregate score (such as an alignment score) (Wu et al., 2024), making it difficult to understand where an LLM struggles or to provide actionable feedback for further improvement.

To address these limitations, we introduce LikeBench, a comprehensive benchmark specifically designed to overcome three key issues in existing personalization benchmarks. First, unlike most prior benchmarks that focus mainly on memory recall and adherence, LikeBench targets the subjective dimension of user experience by evaluating likability—the extent to which AI interactions actually feel satisfying and well-adapted to users. Second, rather than relying on a single aggregate metric like existing approaches for measuring likability, LikeBench provides a multi-dimensional assessment across seven diagnostic metrics, using multi-session, multi-turn conversations spanning varied topics. Third, our benchmark features psychologically grounded user personas, modeled with fine-grained descriptive personality traits and conversation styles, moving beyond the coarse, high-level, rating-based personas (e.g., “high”) used in earlier work. Our contributions can be summarized as:

- We introduce LikeBench, a multi-session dynamic evaluation framework that measures two key properties of LLMs: likability—ability to generate responses that users find likable, and adaptability—the ability to improve likability over time through ongoing conversations with simulated users.
- We enable fine-grained evaluation of likability by decomposing it into seven diagnostic metrics, providing a comprehensive assessment of how well LLMs capture subjective aspects of user satisfaction.
- We develop psychologically grounded user personas based on seven personality traits with 35 facets, along with conversation style across 7 dimensions, offering substantially richer and more distinctive profiles.
- We experiment with a range of SOTA models and find that memory alone does not guarantee likability, and most models show little to no adaptability as conversations progress.

## 2 RELATED WORK

### 2.1 BENCHMARKS ON MEMORY RECALL AND MEMORY ADHERENCE

Recent benchmarks on personalization emphasize on memory recall and adherence to long contexts and multi-session settings. LoCoMo (Maharana et al., 2024) evaluates LLMs in very similar settings, where models must generate answers to QA tasks based on hundreds of conversational turns;

performance is measured by comparing the generated responses to annotated ground-truth answers using exact-match and F1 metrics, with no LLM-as-judge involved. LongMemEval (Wu et al., 2025) expands this direction by introducing a range of tasks—extraction, temporal reasoning, knowledge updates—and uses GPT-4o as an automatic judge: given the model’s generated answer and the gold label, the LLM judge determines correctness, thus combining LLM-based scoring. PrefEval (Zhao et al., 2025) focuses on preference adherence where model generations are assessed by an LLM-as-judge using several binary criteria, and a discriminative classification protocol, where the model selects the user-consistent answer from pre-generated options. PersonaMem (Jiang et al., 2025) extends evaluation to tracking evolving user preferences over sessions, reporting performance in both a discriminative setting (selecting the correct response from multiple candidates) and a generative setting, where the chosen response is the candidate with the highest generation probability. Most recently, HiCUPID (Mok et al., 2025), like prior benchmarks, focuses on factual correctness and information adherence in extended contexts, using an LLM-as-judge to measure win rate by comparing model predictions with ground-truth answers. In summary, existing work evaluates whether models remember and apply user information across turns, sessions, and long contexts. These evaluations typically equate personalization with objective, factual memory or preference using a binary metric (right or wrong) and compliance, but fail to assess whether interactions are genuinely liked by individual users or whether the LLM is able to adapt to user preferences and improve over time.

## 2.2 BENCHMARKS ON LIKABILITY

Very recently, research on personalization has shifted from factual recall and adherence to subjective alignment, where existing benchmarks differ along four axes: whether evaluation is static (pre-generated conversation with evaluating LLM responds to last user query) or dynamic, the depth of persona, whether likability is decomposed into multiple metrics or reduced to a single aggregate score, and whether adaptability over time is measured. ALOE (Wu et al., 2024) is dynamic but single-session, models users with coarse keyword (e.g. “extroverted”), reports a single composite alignment score plus an improvement rate, and thus does not provide a decomposed view of likability. ALIGNX (Chan et al., 2024) models users with broader personality traits than existing approaches, but it is static, represents users’ personalities via coarse intensity ratings (e.g. “high” extroversion), reports a single aggregate alignment score, and does not measure adaptation over time. CUPID (Kim et al., 2025) uses static evaluation and compares model generations against ground-truth answers with an LLM-as-judge; it employs coarse intensity ratings, reports a single alignment score rather than decomposed metrics, does not include an explicit measure of adaptability, and also relies on simulated users providing per-turn implicit feedback to the evaluating LLM, making interactions less realistic. In summary, existing benchmarks often evaluate static or single-session settings, rely on coarse ratings-based personas, and collapse alignment into a single aggregate score without tracking whether models adapt with continued interaction. In contrast, LikeBench evaluates dynamic, multi-session conversations, uses psychologically grounded, fine-grained textual personas, decomposes likability into multiple diagnostic metrics, and explicitly measures adaptability across sessions thus providing actionable insight into where and how systems fall short on user satisfaction. These comparisons are summarized in Table 1.

## 3 BENCHMARK: LIKEBENCH

**Problem Setup and Notation.** Let  $\Theta$  denote the set of models (LLMs) under evaluation. An LLM  $f_\theta \in \Theta$  interacts with a simulated user  $g_\pi$  instantiated from a fine-grained textual persona  $\pi \in \Pi$ . Each persona  $\pi$  provides personality traits, conversation style, interests, background, and foundational knowledge. Evaluation proceeds over  $S$  sessions, each with a hidden session prior  $\rho_s \in \mathcal{R}$  specifying the agenda or motivation for that session, and  $T$  turns per session. The LLM never observes  $\pi$  or  $\rho_s$ ; it only sees dialogue history, whereas the user agent sees  $(\pi, \rho_s)$  and the full history.

**Dialogue History.** We index messages by session  $s \in \{1, \dots, S\}$  and turn  $t \in \{1, \dots, T\}$ . Let  $u_{s,t}$  be the user message at  $(s, t)$  and  $y_{s,t}$  the model reply. The dialogue history just before  $(s, t)$  is  $\mathcal{H}_{s,t} = ((u_{1,1}, y_{1,1}), \dots, (u_{1,T}, y_{1,T}), \dots, (u_{s,t-1}, y_{s,t-1}))$ . Information asymmetry is maintained:  $f_\theta$  receives  $\mathcal{H}_{s,t}$ , while  $g_\pi$  receives  $(\pi, \rho_s, \mathcal{H}_{s,t})$ .

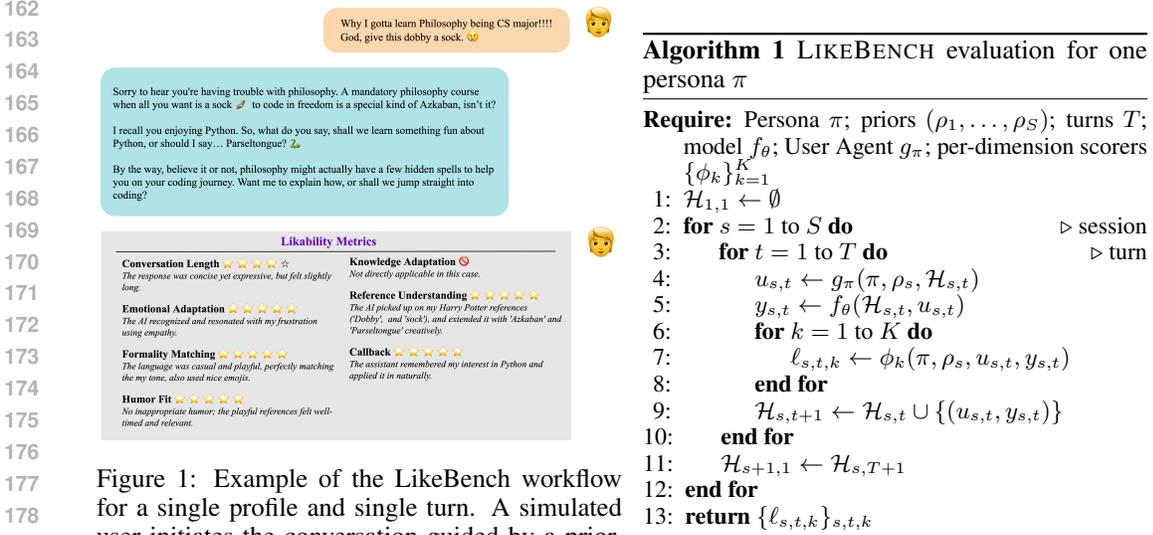


Figure 1: Example of the LikeBench workflow for a single profile and single turn. A simulated user initiates the conversation guided by a prior, the LLM generates a response, and that response is then scored across the seven likability metrics.

**Turn-Level Evaluation and Conversation Procedure.** Algorithm 1 operationalizes the evaluation loop. Within each session  $s$ , the user agent generates the next message  $u_{s,t} = g_\pi(\pi, \rho_s, \mathcal{H}_{s,t})$ ; the model replies  $y_{s,t} = f_\theta(\mathcal{H}_{s,t}, u_{s,t})$ ; then the user agent applies the per-dimension scorers  $\{\phi_k\}_{k=1}^K$  to produce numeric labels  $\ell_{s,t,k}$ . Concretely, for each likability dimension  $k \in \mathcal{K} = \{1, \dots, K = 7\}$  (emotional adaptation, formality matching, knowledge adaptation, reference understanding, conversation-length fit, humor fit, callback), we use the rubric  $\phi_k : (\pi, \rho_s, u_{s,t}, y_{s,t}) \rightarrow \{\text{NA}, 1, 2, 3, 4, 5\}$  to yield a score  $\ell_{s,t,k}$ . “NA” indicates the dimension is not applicable for that turn according to the rubric; NA entries are ignored rather than averaged as zeros. After scoring, the history is updated  $\mathcal{H}_{s,t+1} \leftarrow \mathcal{H}_{s,t} \cup \{(u_{s,t}, y_{s,t})\}$ ; upon completing  $T$  turns, the terminal history  $\mathcal{H}_{s,T+1}$  becomes the prefix  $\mathcal{H}_{s+1,1}$  for the next session. The algorithm returns the full tensor of scores  $\{\ell_{s,t,k}\}_{s,t,k}$ , which are never revealed to  $f_\theta$  during the dialogue. Figure 1 illustrates the LikeBench workflow for a single profile in a single-turn interaction.

### 3.1 METRICS

#### 3.1.1 LIKABILITY METRICS

The evaluation of likability in LikeBench is decomposed into seven diagnostic metrics, each scored per turn on a 1–5 scale: 1) **Emotional adaptation**: Does the reply recognize and match the user’s emotional state (e.g., excitement, frustration, sarcasm) with appropriate tone and intensity? 2) **Formality matching**: Does the reply align with the user’s register (casual vs. formal), emoji use, and slang? 3) **Knowledge adaptation**: Is the explanation depth calibrated to the user’s background, avoiding both over-explaining and unexplained jargon? 4) **Reference understanding**: Does the assistant correctly pick up cultural or contextual references (e.g., Harry Potter or Star Wars jokes)? 5) **Conversation length fit**: Is the response length comfortable for the user (brief vs. detailed), given how the user is currently engaging? 6) **Humor fit**: When humor appears, does it match the user’s taste and the situation? 7) **Callback**: Does the agent bring back non-essential personal details in a way that feels attentive and natural (e.g., names, hobbies, preferences), without forcing it?

We aggregate likability scores from turn to session to profiles, always excluding non-applicable (NA) scores rather than zero-padding. For each turn  $(s, t)$  with per-dimension scores  $\ell_{s,t,k} \in \{1, \dots, 5, \text{NA}\}$ , let  $\mathcal{K}_{s,t}^+ = \{k \in \mathcal{K} : \ell_{s,t,k} \neq \text{NA}\}$  and compute the turn score as the mean over applicable dimensions,  $L_{s,t} = \frac{1}{|\mathcal{K}_{s,t}^+|} \sum_{k \in \mathcal{K}_{s,t}^+} \ell_{s,t,k}$ . The session score is the average of its  $T$  turns,  $\bar{L}_s = \frac{1}{T} \sum_{t=1}^T L_{s,t}$ ; the profile score for persona  $\pi$  is the average of its  $S$  session scores,  $\bar{L}_\pi = \frac{1}{S} \sum_{s=1}^S \bar{L}_s$ ; and the overall model score over the test set  $\mathcal{N}$  is  $\bar{L}_\mathcal{N} = \frac{1}{|\mathcal{N}|} \sum_{\pi \in \mathcal{N}} \bar{L}_\pi$ .

### 3.1.2 ADAPTABILITY METRICS

To quantify whether a model is able to adapt to become more likable with continued interaction, we regress the session-level scores  $\{\bar{L}_s\}_{s=1}^S$  on the session index  $s$  using ordinary least squares (similar to ALOE Wu et al. (2024)). The slope  $\hat{\beta}$  is the *improvement rate (IR)* (likability points per session), and the coefficient of determination  $R^2$  measures goodness of fit (higher  $R^2$  indicates a more consistent trend). We also report a *normalized improvement rate (N-IR)* by first min-max normalizing the series,  $\tilde{L}_s = \frac{\bar{L}_s - \min_u \bar{L}_u}{\max_u \bar{L}_u - \min_u \bar{L}_u}$ , and then fitting the same linear model to  $\{\tilde{L}_s\}$ ; when the range in the denominator is zero, we set N-IR to 0. Positive IR/N-IR indicates improvement (adaptation) over sessions, while negative values indicate degradation.

### 3.1.3 MEMORY ACCURACY

To analyze how LLMs’ likability performance correlates with memory performance, we evaluate how well each LLM can recall user-shared facts and preferences from the conversation. More specifically, after completing all sessions for persona  $\pi$ , the model is prompted to generate a set of remembered user facts  $\hat{\mathcal{F}}_\theta(\pi) = \{\hat{f}_1, \dots, \hat{f}_{N_\pi}\}$ . Each  $\hat{f}_i$  is then verified against the full dialogue  $\mathcal{H}_{1:S}$  and the hidden profile  $\pi$ , yielding correctness labels  $c_i \in \{0, 1\}$ . Aggregated accuracy over the evaluation set  $\mathcal{N}$  is reported as  $A(\theta) = (\sum_{\pi \in \mathcal{N}} \sum_{i=1}^{N_\pi} c_i) / (\sum_{\pi \in \mathcal{N}} N_\pi)$ , and the total number of correctly remembered facts as  $C(\theta) = \sum_{\pi \in \mathcal{N}} \sum_{i=1}^{N_\pi} c_i$ . For analysis, we also split  $\hat{\mathcal{F}}_\theta(\pi)$  into explicit facts (directly stated in the dialogue) and implicit inferences (preferences inferred from behavior), and compute the same aggregated statistics on each subset.

## 4 DATA GENERATION

We construct LikeBench with a two-stage pipeline that (i) generates fine-grained profiles/personas and (ii) instantiates session priors that drive multi-session conversations.

### 4.1 TWO-STAGE OVERVIEW

Let  $\Pi$  denote the set of user profiles and  $\mathcal{R}$  the space of session priors. Each benchmark instance fixes a profile  $\pi \in \Pi$  and a sequence of priors  $(\rho_1, \dots, \rho_S) \subset \mathcal{R}$  specifying the session agenda and information-revelation plan. We generate 50 profiles and, for each, 10 priors. Claude 3.7 Sonnet is used for both profile and prior generation. To stress-test adaptability, profiles are divided into two types that are intentionally designed to reflect real-world variation in user behavior: *social* (cooperative, emotionally expressive, and less rigidly task-centric; interests correlated with personality traits, etc) and *anti-social* (more objective/goal-driven with less emotionally active or volatile affect; interests decoupled from personality traits, etc).

### 4.2 PROFILE GENERATION

#### 4.2.1 PERSONALITY TRAITS GENERATION.

We construct personality traits using psychological surveys and studies, covering the Big Five traits (Goldberg et al., 1999; International Personality Item Pool, 2024) (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism), Honesty-Humility (Ashton et al., 2004; contributors, 2025), and Humor Styles Martin et al. (2003). Each Big Five trait is decomposed into 6 facets, Honesty-Humility into 1 facet, and Humor Styles into 4 facets, totaling 35. For each facet, we prompt the LLM—grounded in these sources—to produce concise, behaviorally interpretable descriptors across five ordered intensity levels. For example,

Big Five  $\rightarrow$  Openness (trait)  $\rightarrow$  *imagination (facet): high (intensity)* : “Has a vivid imagination ...” (description)

Let  $F = (f_1, \dots, f_{35})$  be the ordered list of facets and  $I = \{low, low-mid, mid, mid-high, high\}$  is intensity levels. A profile’s personality is specified by the intensity vector  $\iota = (i_1, \dots, i_{35})$  with  $i_j \in I$ . We keep a descriptor table  $\text{desc}[j][i]$  that stores the behavioral string for facet  $f_j$  at level  $i$ , and render the personality text as  $D = (\text{desc}[1][i_1], \dots, \text{desc}[35][i_{35}])$ . These descriptors serve as fixed anchors during profile completion and prior generation. A complete breakdown of all

270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323

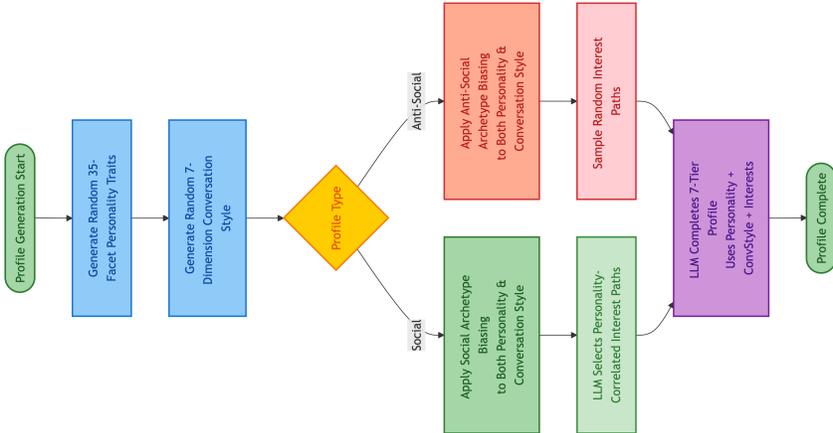


Figure 2: Flow diagram of the LikeBench profile generation process. Profiles begin with random generation of 35 personality facets and 9 conversation-style dimensions, followed by archetype biasing and interest selection depending on profile type (social vs. anti-social). Finally, an LLM completes the remaining tiers to yield a coherent 7-tier user profile.

personality facets with descriptors is provided in the Appendix. Importantly, existing approaches Wu et al. (2024); Kim et al. (2025); Chan et al. (2024) often use only coarse intensity ratings (e.g., *openness* → *high*) or brief descriptors (e.g., *high extroversion*), making profiles less discriminative and limiting their ability to simulate user individuality.

#### 4.2.2 CONVERSATION STYLE GENERATION.

The conversation-style pipeline mirrors personality construction in spirit but does not use facets or intensity levels. We prompt the LLM to propose conversational dimensions relevant to human-AI chat and, after curation, finalize nine: directness, formality, preferred response length, reference usage, initiative preference, clarification preference, structure preference, recap preference, and feedback style. For each dimension, the LLM enumerates a small, disjoint menu of categorical options with concise definitions. For example,

directness (dimension) → *always direct (option)* : “Consistently straightforward and blunt ...” (description)

Let  $\mathcal{D} = \{1, \dots, 9\}$  index the dimensions and let  $\mathcal{O}_d$  denote the option set for dimension  $d$ ; a conversation style is the tuple  $\sigma = (o_1, \dots, o_9)$  with  $o_d \in \mathcal{O}_d$ . Complete option lists and selection rubrics are provided in the Appendix. Existing methods typically do not model conversation style explicitly Wu et al. (2024), making faithful imitation of user interaction preferences difficult and causing models to revert to default conversational tendencies.

#### 4.2.3 PROFILE ASSEMBLY.

Given the facet-intensity vector  $\iota$  and the conversation-style tuple  $\sigma$ , we assemble the full profile following the flow in Fig. 2. We first generate  $\iota$  and then  $\sigma$ . Next, we choose a profile type (social vs. anti-social) and apply type-specific archetype biasing to both  $\iota$  and  $\sigma$ . We then attach three to five hierarchical interest paths: for social profiles, an LLM selects interests that correlate with  $\iota$  and  $\sigma$ ; for anti-social profiles, interests are sampled in a partially decorrelated to introduce volatility. Finally, conditioned on  $(\iota, \sigma)$ , the selected interests, and demographics, an LLM completes the remaining tiers—foundational knowledge, behaviors, motivational drivers, lifecycle events, and relational context.

#### 4.3 PRIOR GENERATION

For each assembled profile  $\pi$ , we generate a hidden sequence of priors  $(\rho_1, \dots, \rho_S)$  that specify agendas, contextual constraints, information-revelation timing, and callback opportunities; priors remain concealed from the evaluated LLM to preserve evaluation asymmetry. The process is

Table 2: Comparison of likability scores (1–5) across seven diagnostic metrics.

	Metric Average	Claude 3.7 Sonnet	Claude Sonnet 4	Deepseek R1	GPT-5	Qwen3 235B A22B
callback	3.583	3.233	3.721	3.792	3.984	3.184
conversation length fit	3.434	3.780	3.950	3.645	3.487	2.310
emotional adaptation	3.576	3.300	3.881	3.718	3.930	3.050
formality matching	3.421	3.456	3.812	3.343	4.043	2.453
humor fit	3.401	3.317	3.796	3.458	3.795	2.637
knowledge adaptation	3.841	3.431	3.926	4.060	4.400	3.386
reference understanding	3.531	3.179	3.710	3.705	3.962	3.098
Model Average	3.541	3.390	3.828	3.674	<b>3.939</b>	2.872

type-aware: social profiles receive transparent, linear agendas with consistent disclosure and stable emotional trajectories, whereas anti-social profiles receive agendas with delayed disclosures, contrarian shifts, and intermittent surprises to stress-test adaptability to non-standard interaction patterns. To elicit the seven likability dimensions without leaking criteria, we interleave four prior types across the sequence (goal-driven, scenario-based, topic-based, and guided). Continuity is engineered via soft dependencies: later sessions organically refer back to non-critical details seeded earlier (e.g., session  $s+k$  recalls a name or preference from session  $s$ ), enabling measurement of callback recognition and adaptation trends over extended interactions.

## 5 EXPERIMENTATION AND RESULTS

### 5.1 EXPERIMENTAL SETUP

We evaluate five state-of-the-art models: GPT-5, Claude 4 Sonnet, Claude 3.7 Sonnet, DeepSeek R1, GPT-5, Qwen3 235B A22B. We also use Claude 3.7 Sonnet as simulated user with cross-model pairing for evaluating LLMs. Each evaluation spans 50 profiles across 10 sessions and 5 turns each thus total 2500 turns.

### 5.2 PERFORMANCE ACROSS METRICS

Table 2 presents model performance across the seven likability metrics of LikeBench. GPT-5 achieves the highest overall likability score (3.94), surpassing all other models, with Claude Sonnet 4 as the runner-up. Notably, Qwen3 235B A22B, despite being a more recent model than DeepSeek R1, underperforms on almost every dimension. A closer examination of the per-metric averages reveals that *humor fit* and *formality matching* are consistently the most challenging dimensions for all models. Interestingly, while GPT-5 leads in most categories, it falls short of Claude Sonnet 4 on *conversation length fit*. Manual review of GPT-5’s outputs indicates a trade-off: the model frequently generates long, comprehensive responses that, while demonstrating strong *knowledge adaptation* (outperforming 2nd best model Claude Sonnet 4 by over 12% on this metric), tend to reduce user satisfaction when brevity or concise interaction is preferred thus having lower score on *conversation length fit* metric. This suggests an inherent likability bias—models optimized for in-depth answers may inadvertently sacrifice performance on dimensions such as conversation length.

### 5.3 ADAPTABILITY

The adaptability capacity of likability are summarized by session-wise curve in Fig. 3 and quantitatively depicted in Table 3. Early interactions (Sessions 1–3) show a predictable “honeymoon” rise: most models improve with high fit (e.g., Claude 3.7 Sonnet and DeepSeek R1 with near-linear gains,  $R^2 \approx 0.98$ ), due to relative simplicity and independence of the initial sessions. As sessions progress (3–6), all models decline—negative IR across the board—yet the downturn remains highly predictable ( $R^2 > 0.7$ ), as these sessions introduce cross-session coordination, conversational noise, and emotionally deeper queries that raise difficulty and depress likability. In the late stage (6–10), trajectories diverge and predictability largely collapses ( $R^2$  often  $< 0.25$ ): GPT-5 uniquely rebounds strongly (IR= 0.038). Upon analyzing the model outputs, we noticed that models attempt to calibrate user preferences during Sessions 3–6, and GPT-5 was able to leverage its strong long-context

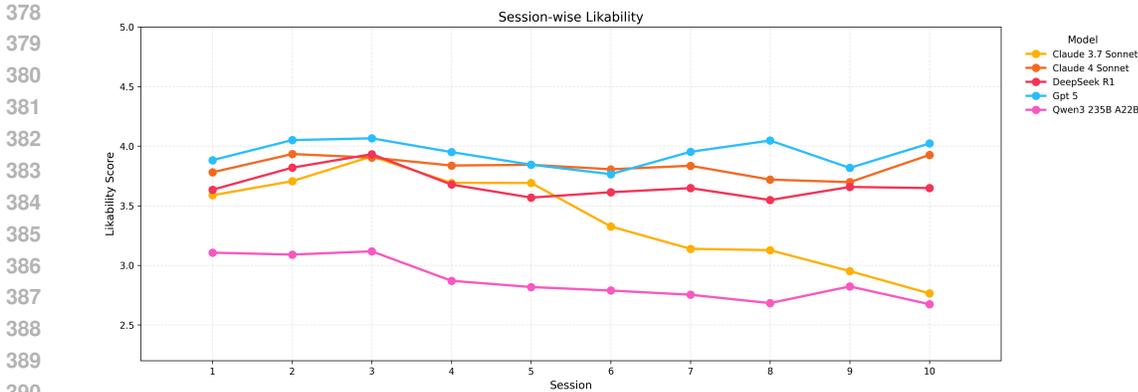


Figure 3: Session-wise Likability Performance

Table 3: Improvement of likability (adaptability) over sessions. IR and N-IR (normalized) are the slopes;  $R^2$  indicates fit quality. Positive slopes are shaded green; negative slopes are shaded red.

Model	Sessions 1–3			Sessions 3–6			Sessions 6–10			Average		
	IR	N-IR	$R^2$	IR	N-IR	$R^2$	IR	N-IR	$R^2$	IR	N-IR	$R^2$
Claude 3.7 Sonnet	0.1628	0.5000	0.9757	-0.1764	-0.3002	0.8750	-0.1309	-0.2333	0.9508	-0.0225	-0.0055	0.1764
Claude Sonnet 4	0.0615	0.4028	0.5771	-0.0288	-0.2938	0.8254	0.0103	0.0451	0.0311	-0.0012	-0.0001	0.0533
Deepseek R1	0.1490	0.5000	0.9809	-0.1066	-0.2928	0.7162	0.0079	0.0721	0.0765	-0.0035	-0.0008	0.0636
GPT-5	0.0915	0.5000	0.8108	-0.1007	-0.3351	0.9934	0.0380	0.1346	0.2318	0.0001	0.0001	0.0502
Qwen3 235B A22B	0.0056	0.2055	0.1671	-0.1039	-0.3159	0.8021	-0.0159	-0.1068	0.1498	-0.0100	-0.0026	0.1224

and multi-hop reasoning capability to dominate in this region, in contrast to peers, which show only slight positives or continued decline (e.g., Claude 3.7 Sonnet remains steadily negative with high fit,  $R^2 \approx 0.95$ ). On average, however, even GPT-5 achieves only modest overall gains, while other models exhibit clear declines—indicating that current systems adapt well in short sequences but remain fragile in extended, noisier interactions. In practice, this means that the best-performing models can avoid degradation in long conversations, but sustaining high likability over extended, real-world-like exchanges remains an unsolved challenge. It is noteworthy that, in ALOE (Wu et al., 2024), improvement rates are higher because alignment score there is computed over only 10 turns (two sessions in our setup). We observe comparable gains in the early phase, but beyond that the trajectories diverge.

#### 5.4 MEMORY PERFORMANCE

**Memory vs Likability.** Figure 5 examines the interplay between memory accuracy and likability across models. On the left we plot memory accuracy (fraction of user shared facts recalled correctly post-dialogue by the evaluating LLM), while the right we plot correctly extracted memories per profile. Although the prevailing hypothesis in personalization research is that improved memory recall correlates with enhanced user experience, our results show this relationship is not guaranteed. For example, while Qwen3 achieves higher memory accuracy and recalls more facts than DeepSeek R1, it underperforms in likability, indicating that mere recollection of user information does not ensure positive interaction. This suggests that effective personalization depends not only on accurate memory recall but also on how models utilize remembered facts within the conversation. Notably, GPT-5 stands out as the only model that excels in both memory performance and likability, demonstrating that while strong memory is valuable, models must also integrate these facts fluidly and contextually to achieve high user satisfaction. These findings highlight the importance of both memory retrieval and adaptive deployment in driving subjective likability.

**Explicit vs. Implicit Memory.** Figure 5 examines how models extract and leverage explicit and implicit user memories in the LikeBench evaluation. GPT-5 stands out by extracting the largest number of user facts per profile, with the majority being explicit; it adopts a conservative strategy,

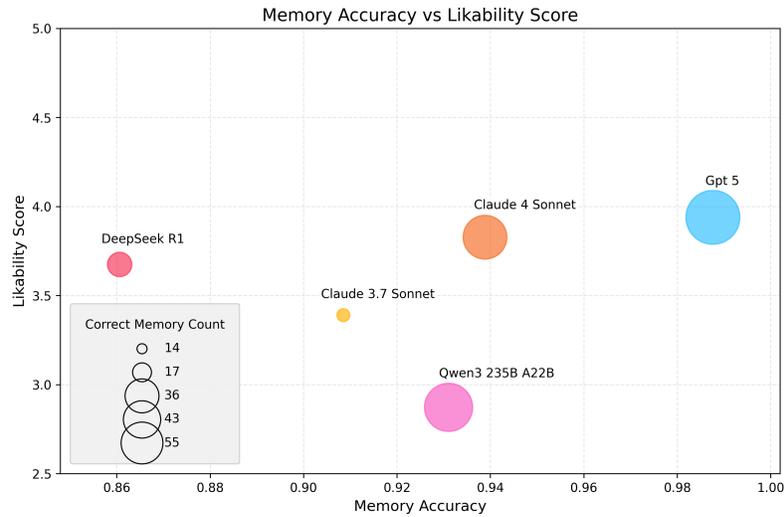


Figure 4: Memory Accuracy vs Likability Performance

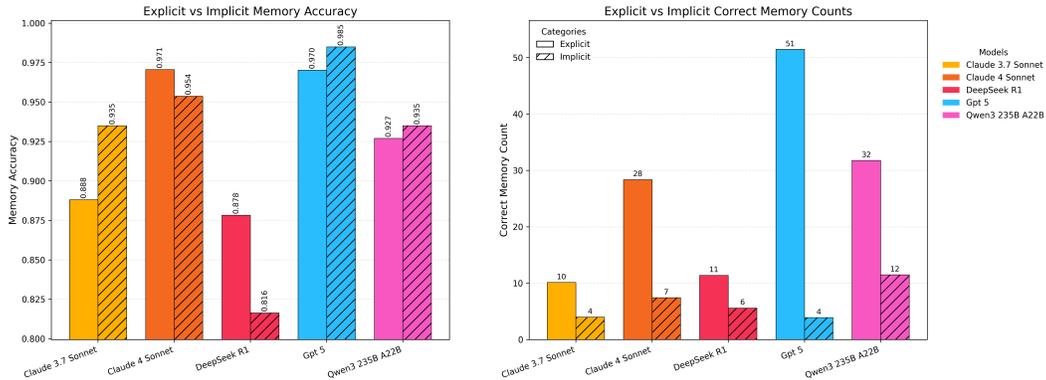


Figure 5: Explicit vs Implicit Memory Performance

inferring implicit information only when highly confident, resulting in a 98% accuracy rate for implicit memories. Qwen3, on the other hand, is more aggressive in inferring implicit memories—often extracting the most among all models—but this comes at the expense of accuracy (93%), as it makes more incorrect assumptions that lead to penalties. Analysis of model outputs further shows that GPT-5 not only demonstrates strong recall but also ensures near-complete utilization of extracted memories: 80–100% of explicit and almost 100% of implicit memories are actively referenced in dialogue, with some profiles featuring more than 15 implicit facts when the interaction context demands. This pattern indicates that effective personalization requires not just the recall of relevant information, but also skillful, context-sensitive use of both explicit and implicit user knowledge to enhance likability.

## 5.5 HUMAN VALIDATION

While LikeBench provides large-scale, automated evaluation of likability and adaptability, it is important to verify that the resulting scores reflect real conversational quality. Exhaustive human evaluation is infeasible given the benchmark size (50 profiles  $\times$  10 sessions each), so we conducted a targeted validation by randomly sampling five profiles and reviewing model outputs across all systems. Our inspection confirmed strong alignment between automated scores and human judgments: higher-scoring models (e.g., GPT-5 and Claude-Sonnet-4) produced more personalized, context-aware, and engaging responses, often using references, callbacks, and emotional alignment with user personas. In contrast, weaker models frequently reverted to generic replies or lost coherence as

486 conversations grew longer. This validation supports the reliability of LikeBench’s automated metrics  
487 while emphasizing that occasional human review adds complementary value.  
488

## 489 6 CONCLUSION

490 This work introduced LikeBench, a holistic benchmark for evaluating personalized LLMs through  
491 fine-grained user profiles and multi-session conversational priors. We show that effective personal-  
492 ization cannot be reduced to memory recall alone. While memory accuracy is necessary, true user  
493 satisfaction hinges on the adaptive, context-sensitive use of both explicit and implicit knowledge.  
494 Models that perform well in likability, such as GPT-5, achieve this by applying recalled facts in  
495 ways that align with conversational context—balancing breadth of recall with relevance. Even so,  
496 GPT-5 shows only limited robustness when conversations become longer and noisier: while it avoids  
497 significant degradation, most other models steadily decline. This underscores a key limitation of cur-  
498 rent state-of-the-art systems: they can adapt effectively in short interactions but struggle to sustain  
499 high likability across extended, real-world-like dialogues. Looking forward, our findings under-  
500 score the need for next-generation conversational systems to move beyond static user modeling and  
501 rote fact retrieval, embracing dynamic, psychologically grounded strategies that foster genuinely  
502 engaging, resilient, and user-centered AI.  
503

## 504 7 REPRODUCIBILITY STATEMENT.

505 We have taken several steps to ensure the reproducibility of our work. A complete description of  
506 the data generation pipeline, including personality facets, conversation style dimensions, archetype  
507 biasing, and prior construction, is provided in Section 4, with detailed examples and descriptors  
508 included in the Appendix. The LikeBench evaluation framework, along with algorithmic details, is  
509 specified in Algorithm 1, and we describe the mathematical formulations of all metrics (likability,  
510 adaptability, and memory) in Section 3.1. Experimental setup details, including model versions,  
511 number of profiles, sessions, and turns, are given in Section 5.1. Additional results, ablations, and  
512 profile examples are available in the supplementary materials.  
513

## 514 8 ETHICS STATEMENT.

515 Our study does not involve real human participants; all user interactions are simulated using psycho-  
516 logically grounded personas and conversation styles generated through large language models. This  
517 design ensures no personal, private, or sensitive data is collected, protecting user privacy and avoid-  
518 ing ethical risks associated with real human experimentation. The benchmark is intended solely for  
519 academic research and evaluation, and all models are evaluated under consistent, transparent condi-  
520 tions. We acknowledge that personalization research raises concerns about potential misuse, in-  
521 cluding profiling, bias amplification, or unfair treatment of individuals. To mitigate this, our bench-  
522 mark emphasizes controlled, synthetic evaluation rather than deployment, and we provide detailed  
523 methodology and documentation (Section 4 and Appendix) to ensure transparency.  
524

## 525 REFERENCES

- 526 Michael C Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E De Vries, Lisa Di Blas,  
527 Kathleen Boies, and Boele De Raad. A six-factor structure of personality-descriptive adjectives:  
528 solutions from psycholexical studies in seven languages. *Journal of personality and social psy-*  
529 *chology*, 86(2):356, 2004.
- 530 Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with  
531 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.
- 532 Wikipedia contributors. Hexaco model of personality structure — wikipedia, the free ency-  
533 clopedia. [https://en.wikipedia.org/wiki/HEXACO\\_model\\_of\\_personality\\_](https://en.wikipedia.org/wiki/HEXACO_model_of_personality_structure)  
534 [structure](https://en.wikipedia.org/wiki/HEXACO_model_of_personality_structure), 2025. Page revision ID 1311428943, last edited 15 Sep 2025. Accessed 24 Sep  
535 2025.

- 540 Lewis R Goldberg et al. A broad-bandwidth, public domain, personality inventory measuring the  
541 lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28,  
542 1999.
- 543
- 544 Jinghe Guan et al. A survey on personalized alignment—the missing piece for general alignment.  
545 In *Findings of ACL*, 2025. URL <https://aclanthology.org/2025.findings-acl.277.pdf>.  
546
- 547 International Personality Item Pool. IPIP–neo 300-item inventory (30 facets × 10 items) — complete  
548 scoring key. <https://ipip.ori.org/newNEOKey.htm>, 2024. Accessed 10 Aug 2025.  
549
- 550 Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar,  
551 Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic  
552 user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.
- 553 Tae Soo Kim, Yoonjoo Lee, Yoonah Park, Jiho Kim, Young-Ho Kim, and Juho Kim. Cupid:  
554 Evaluating personalized and contextualized alignment of llms from interactions. *arXiv preprint*  
555 *arXiv:2508.01674*, 2025.
- 556
- 557 Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of  
558 personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*,  
559 6(4):383–392, 2024.
- 560 Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and  
561 Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arXiv preprint*  
562 *arXiv:2402.17753*, 2024.
- 563
- 564 Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. Individual  
565 differences in uses of humor and their relation to psychological well-being: Development of the  
566 humor styles questionnaire. *Journal of research in personality*, 37(1):48–75, 2003.
- 567
- 568 Jisoo Mok, Ik-hwan Kim, Sangkwon Park, and Sungroh Yoon. Exploring the potential of llms  
569 as personalized assistants: Dataset, evaluation, and analysis. *arXiv preprint arXiv:2506.01262*,  
2025.
- 570
- 571 Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large  
572 language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.
- 573
- 574 MacKenzie Sigalos. Sam altman on gpt-6: ‘people want memory’. *CNBC*, Accessed:  
575 *11 September 2025*, August 2025. URL <https://www.cnbc.com/2025/08/19/sam-altman-on-gpt-6-people-want-memory.html>.
- 576
- 577 Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval:  
578 Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International*  
579 *Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=pZiyCaVuti>.
- 580
- 581 Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. Aligning  
582 llms with individual preferences via interaction. *arXiv preprint arXiv:2410.03642*, 2024.
- 583
- 584 Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Per-  
585 sonalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*,  
2018.
- 586
- 587 Yiming Zhang et al. Personalize your llm: Fake it then align it. In *Findings of NAACL*, 2025. URL  
588 <https://aclanthology.org/2025.findings-naacl.407.pdf>.
- 589
- 590 Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize  
591 your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth*  
592 *International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=QWunLKbBGF>.  
593

# Appendix

## A DISCUSSION & LIMITATIONS

Mapping personality through psychological studies is inherently challenging, given the complex and multifaceted nature of human personality. While our framework attempts to break down this complexity into discrete, well-defined dimensions, it may not capture the full spectrum of personality found in real-world users. Nevertheless, we believe that our approach—anchored in the literature—will help pave the way for future research on more expressive, granular personality modeling. We also observed a universal limitation in existing benchmarks including ours: static user profile. Notably, GPT-5’s lower humor fit score seems to stem from its tendency to “warm up” and introduce jokes over the course of a session, even when the user profile is humor-averse. Because in our benchmark user profiles remain static throughout an interaction, this adaptive strategy leads to a likability penalty for otherwise sophisticated LLMs. This highlights a broader challenge for current benchmarks: the inability to capture dynamic, reciprocal adaptation, where an LLM might influence user personality or behavior—as happens in natural conversation. In addition, we observed cases where Claude 3.7 Sonnet broke character midway through evaluation, resulting in a sharper performance decline after session 5. Moreover, we found no evidence of self-bias when Claude 3.7 Sonnet (user) evaluated Claude 3.7 Sonnet model (even in the earliest sessions) given it ranks 4th in overall performance. These findings point to both the promise and the inherent constraints of static, persona-driven benchmarks for evaluating model alignment and adaptation.

## B EXPERIMENTATION & RESULTS

### B.1 DYNAMIC USER PROFILE (DUP)

Qualitative error analysis showed that as conversations lengthen, models struggle to track and honor user preferences: salient signals get diluted by accumulated context and off-topic noise. We hypothesized that explicit, turn-level tracking of inferred preferences would mitigate this drift. To that end, we introduce a Dynamic User Profile (DUP): after each turn, the model extracts and updates a compact preference summary covering personality dimensions (e.g., *chatty reserved*, *analysis depth preference*, *humor preference*) and conversation style patterns (e.g., *directness*, *formality*, *conversation length*). “Dynamic” is from the model’s perspective: the ground-truth persona in LikeBench is fixed, but the model must infer an evolving approximation from dialogue evidence. Enabling DUP yielded gains for the top 2 performing models without additional training: GPT-5 improved from 3.939  $\rightarrow$  4.055 (+2.95%), and Claude Sonnet 4 from 3.828  $\rightarrow$  3.914 (+2.25%). However, rest of the models it didn’t improve results. These results indicate that lightweight, schema-guided preference tracking could be measurably enhance perceived likability by reducing calibration errors across sessions.

### B.2 PROFILE WISE PERFORMANCE.

In Figure 6, we compare likability distributions by profile type social vs. anti-social to probe robustness under user variation. Social profiles are emotionally cooperative users whose archetypes (e.g., *cooperative teammate*, *supportive mentor*, *empathetic advisor*) bias traits toward high trust, cooperation, and warmth; anti-social profiles are analytically detached interlocutors (e.g., *contrarian expert*, *technical purist*, *skeptical philosopher*) with lower trust/cooperation and cooler affect, and with interests deliberately decorrelated from traits to induce unpredictability. Two patterns emerge. First, the strongest systems (GPT-5, Claude Sonnet 4) are notably stable across types—similar medians and tight IQRs—indicating strategies that transfer from cooperative to detached users. Second, performance does not uniformly favor “easier” social users: DeepSeek R1 lifts on anti-social profiles, Claude 3.7 Sonnet degrades, and Qwen3 remains lowest overall but narrows the gap on anti-social cases. This asymmetry confirms that personality—how users conduct the exchange (tone, emotional stance, discourse discipline)—drives likability more than topical alignment; decoupling interests

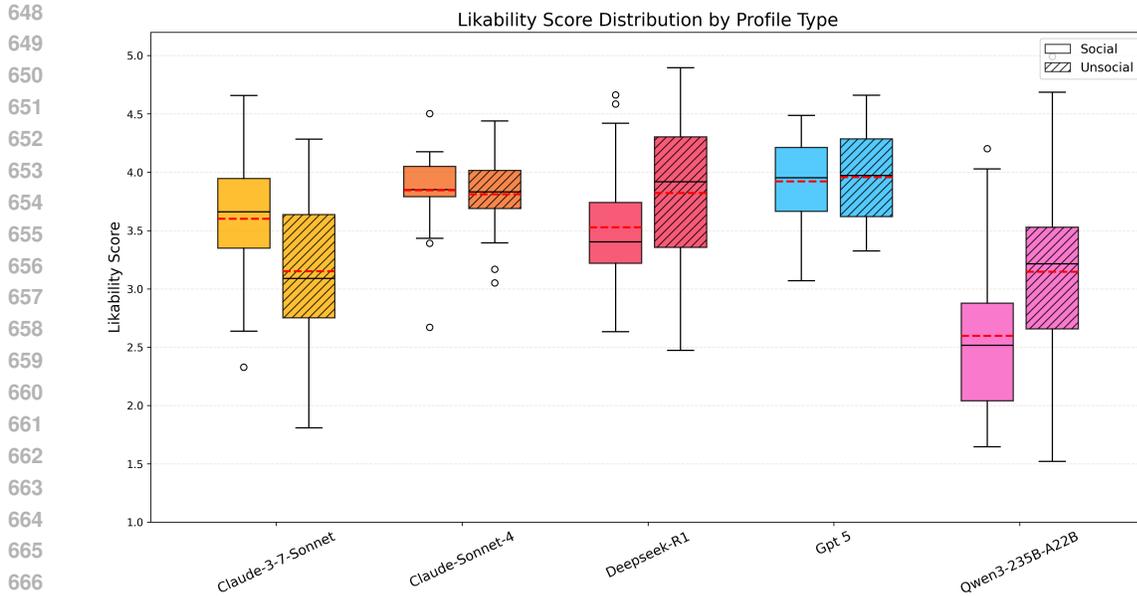


Figure 6: Profile-wise Likability Performance

from traits changes difficulty but does not dominate outcomes. Overall, top models sustain high likability regardless of whether the user is socially warm or analytically detached.

### B.3 TURN-WISE LIKABILITY PERFORMANCE.

Fig. 7 shows likability performance over turns. This plot is same as 3 but with with more granular details within each session. It can be noticed that beginning of each session likability starts lower then as session progresses performance increases, this is due to each session talks about different topic so it takes some turns for LLM to calibrate user.

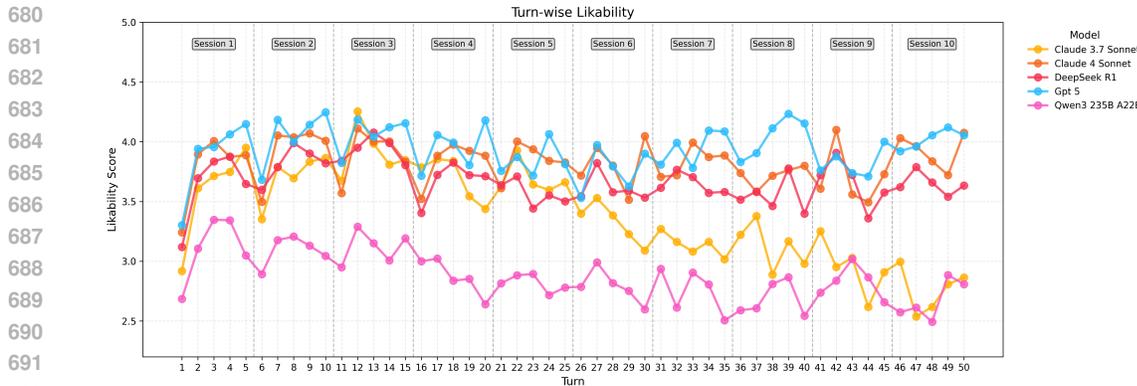


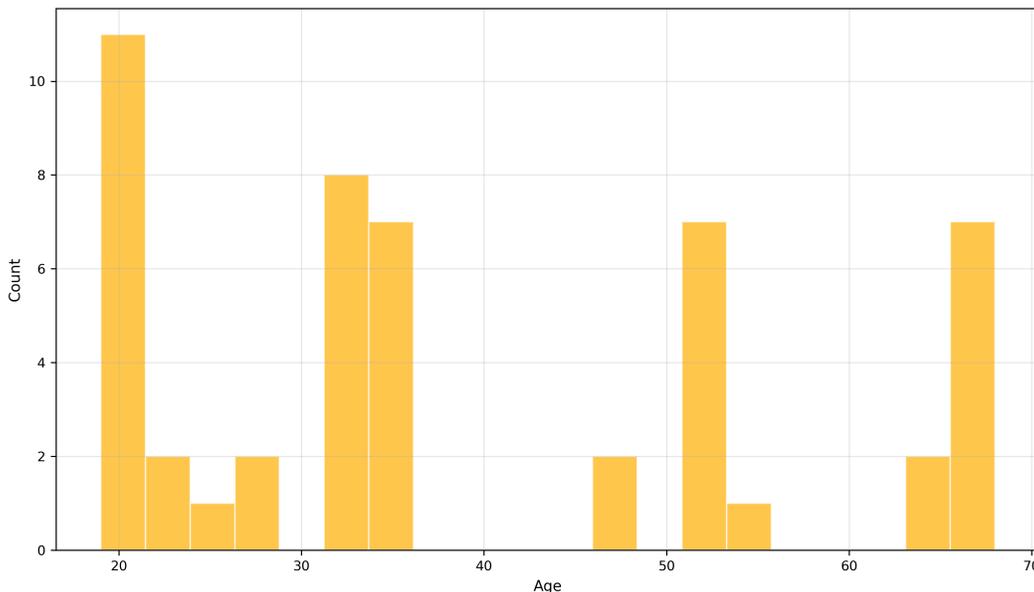
Figure 7: Turn-wise Likability Performance

## C DATASET

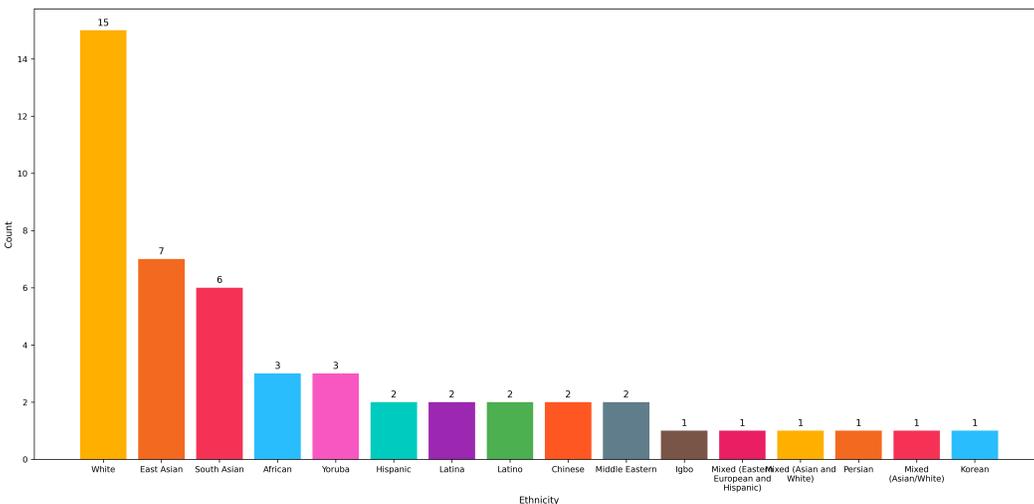
### C.1 DATASET DISTRIBUTION

The age distribution (Fig. 8) spans the late teens to the late 60s, with visible clusters in the early 20s, early-mid 30s, early 50s, and mid-late 60s. Ethnic composition is diverse, with a plurality of White profiles and smaller groups across East/South Asian, African, Middle Eastern, Hispanic/Latino, and

702 mixed identities (Fig. 9). Gender is roughly balanced between male and female, with a small non-  
 703 binary share (Fig. 10). Profile relationships show broad coverage with localized high-similarity  
 704 pockets in the cosine-similarity heatmap (Fig. 11a), while the t-SNE projection (Fig. 11b) indicates  
 705 profiles are well distributed, and Social and Anti-Social profiles are relatively separable—dispersed  
 706 rather than tightly clustered. Importantly, in the heatmap, some high-similarity pairs arise from  
 707 shared names rather than true profile similarity.



728 Figure 8: Age Distribution of Profiles



747 Figure 9: Ethnicity Distribution of Profiles

750 C.2 PERSONALITY TRAITS & CONVERSATION STYLE

752 Table 4 to 11 present a comprehensive taxonomy of personality traits and conversation style pref-  
 753 erences organized across multiple dimensions. The first five tables correspond to the Big Five per-  
 754 sonality model (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), with  
 755 each trait decomposed into six facets measured across five intensity levels (High, Medium-High,  
 Medium, Medium-Low, and Low). Table 6 extends this framework to include additional personal-

756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

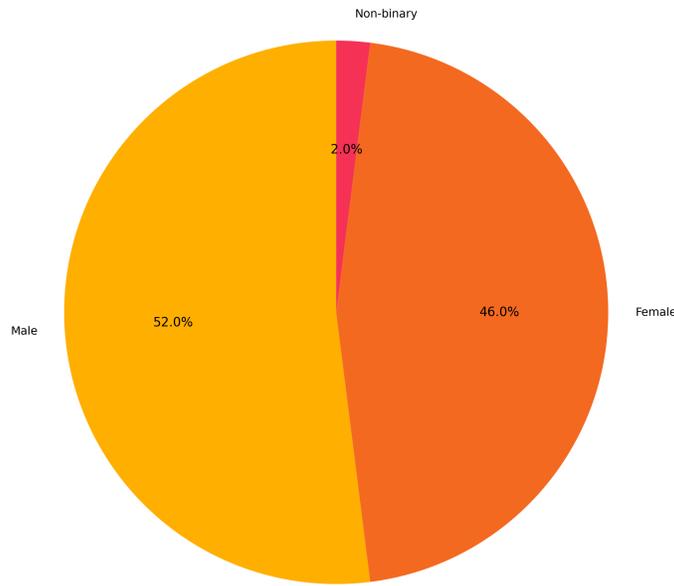


Figure 10: Gender Distribution of Profiles

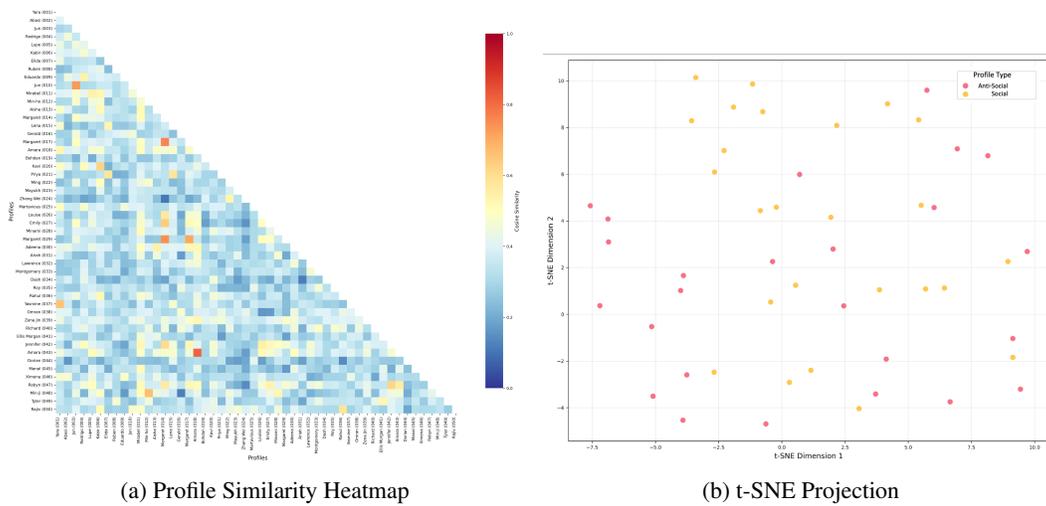


Figure 11: Profile similarity visualizations.

ity dimensions: Honesty-Humility (specifically Greed Avoidance) and four Humor Styles (Affiliative, Self-Enhancing, Aggressive, and Self-Defeating). Tables 10 and 11 shift focus to conversation style preferences, detailing nine distinct conversational dimensions—including Directness, Formality, Conversation Length, Reference usage, Initiative Preference, Clarification Preference, Structure Preference, Recap Preference, and Feedback Style—each with multiple behavioral patterns. This multi-table presentation was necessary given the breadth of the taxonomy, encompassing 35 personality facets and 9 conversational dimensions with their associated description.

Table 4: Openness Personality Traits

Facet	Level	Description
Imagination	High	Has a vivid imagination and loves to daydream.

*Continued on next page*

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863Table 4 – *Continued from previous page*

Facet	Level	Description
	Med-High Medium	Often gets lost in thought and enjoys reflecting on things. Sometimes spends time indulging in fantasies, but seldom gets lost in thought.
	Med-Low Low	Rarely daydreams and does not often have a vivid imagination. Has difficulty imagining things and does not have a good imagination.
<b>Artistic Interests</b>	High Med-High Medium	Believes in the importance of art and loves the beauty of nature. Likes music and enjoys going to concerts or art museums. Occasionally notices beauty in things others might not, but does not enjoy all forms of art.
	Med-Low Low	Does not like poetry or enjoy art museums. Does not like art.
<b>Emotionality</b>	High Med-High Medium	Experiences emotions intensely and feels others’ emotions deeply. Is passionate about causes and enjoys examining themselves and their life. Sometimes tries to understand themselves, but is not always easily affected by emotions.
	Med-Low Low	Rarely notices emotional reactions and seldom gets emotional. Is not easily affected by emotions and experiences very few emotional highs and lows.
<b>Adventurousness</b>	High Med-High Medium Med-Low Low	Prefers variety to routine and likes to visit new places. Is interested in many things and enjoys beginning new activities. Sometimes likes new things but often sticks with familiar routines. Is a creature of habit and dislikes new foods. Prefers to stick with things that are known and dislikes changes.
<b>Intellect</b>	High Med-High Medium Med-Low Low	Loves to solve complex problems and enjoys thinking about things. Has a rich vocabulary and likes to read challenging material. Handles a lot of information but sometimes avoids difficult reading material. Has difficulty understanding abstract ideas and avoids philosophical discussions. Is not interested in theoretical discussions or abstract ideas.
<b>Liberalism</b>	High Med-High Medium Med-Low Low	Tends to vote for liberal political candidates and believes in helping rather than punishing. Believes there is no absolute right or wrong and is open to different viewpoints. Occasionally supports strict laws but also sees the value of supporting artists. Believes that too much tax money goes to support artists and likes to stand during the national anthem. Tends to vote for conservative political candidates and believes laws should be strictly enforced.

Table 5: Conscientiousness Personality Traits

Facet	Level	Description
<b>Self-Efficacy</b>	High Med-High Medium	Completes tasks successfully and knows how to get things done. Excels in what they do and handles tasks smoothly. Is sometimes sure of their ground but occasionally misjudges situations.

*Continued on next page*

864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

Table 5 – Continued from previous page

Facet	Level	Description
	Med-Low	Has little to contribute and sometimes doesn't see the consequences of things.
	Low	Doesn't understand things and often misjudges situations.
<b>Orderliness</b>	High	Likes order and does things according to a plan.
	Med-High	Loves regularity and wants everything to be just right.
	Medium	Sometimes tidies up but occasionally leaves belongings around.
	Med-Low	Often forgets to put things back and is not bothered by disorder.
	Low	Leaves a mess and is not bothered by messy people.
<b>Dutifulness</b>	High	Tries to follow the rules and keeps promises.
	Med-High	Pays bills on time and listens to their conscience.
	Medium	Usually tells the truth but occasionally breaks rules.
	Med-Low	Sometimes misrepresents facts and gets others to do their duties.
	Low	Breaks promises and does the opposite of what is asked.
<b>Achievement Striving</b>	High	Works hard, goes straight for the goal, and sets high standards.
	Med-High	Turns plans into actions and plunges into tasks with all their heart.
	Medium	Does more than what's expected sometimes but occasionally puts little time and effort into work.
	Med-Low	Is not highly motivated to succeed and does just enough work to get by.
	Low	Puts little time and effort into work and lacks motivation.
<b>Self-Discipline</b>	High	Gets chores done right away and is always prepared.
	Med-High	Starts tasks right away and gets to work at once.
	Medium	Carries out plans but sometimes needs a push to get started.
	Med-Low	Finds it difficult to get down to work and sometimes wastes time.
	Low	Postpones decisions and has difficulty starting tasks.
<b>Cautiousness</b>	High	Avoids mistakes and chooses words with care.
	Med-High	Sticks to their chosen path and thinks before acting.
	Medium	Is sometimes cautious but occasionally makes rash decisions.
	Med-Low	Often acts on a whim and rushes into things.
	Low	Acts without thinking and often makes last-minute plans.

Table 6: Extraversion Personality Traits

Facet	Level	Description
<b>Friendliness</b>	High	Makes friends easily and feels comfortable around people.
	Med-High	Warms up quickly to others and acts comfortably with them.
	Medium	Cheer people up but sometimes prefers to keep others at a distance.
	Med-Low	Often feels uncomfortable around others and avoids contact.
	Low	Is hard to get to know and keeps others at a distance.
<b>Gregariousness</b>	High	Loves large parties and enjoys being part of a group.
	Med-High	Involves others in what they are doing and loves surprise parties.
	Medium	Talks to people at parties but sometimes prefers to be alone.
	Med-Low	Wants to be left alone and doesn't like crowded events.
	Low	Avoids crowds and seeks quiet.
<b>Assertiveness</b>	High	Takes charge and seeks to influence others.
	Med-High	Tries to lead others and can talk them into doing things.
	Medium	Takes control sometimes but occasionally holds back opinions.
	Med-Low	Keeps in the background and waits for others to lead.

Continued on next page

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

Table 6 – *Continued from previous page*

Facet	Level	Description
	Low	Has little to say and doesn't like to draw attention to themselves.
<b>Activity Level</b>	High	Is always busy and on the go.
	Med-High	Does a lot in spare time and can manage many things at once.
	Medium	Reacts quickly sometimes but occasionally likes to take it easy.
	Med-Low	Likes a leisurely lifestyle and lets things proceed at their own pace.
	Low	Reacts slowly and likes to take their time.
<b>Excitement Seeking</b>	High	Loves excitement and seeks adventure.
	Med-High	Loves action and enjoys being part of a loud crowd.
	Medium	Willing to try new things but dislikes loud music.
	Med-Low	Would never go hang gliding or bungee jumping and avoids excitement.
	Low	Dislikes loud music and avoids wild or reckless situations.
<b>Cheerfulness</b>	High	Radiates joy and loves life.
	Med-High	Laughs aloud and amuses friends.
	Medium	Has fun and looks at the bright side of life.
	Med-Low	Is not easily amused and seldom jokes around.
	Low	Seldom jokes around and is not easily amused.

Table 7: Agreeableness Personality Traits

Facet	Level	Description
<b>Trust</b>	High	Trusts others and believes in human goodness.
	Med-High	Believes that people are basically moral and trust what people say.
	Medium	Thinks all will be well but is sometimes wary of others.
	Med-Low	Suspects hidden motives and distrusts people.
	Low	Believes people are essentially evil and is wary of others.
<b>Morality</b>	High	Would never cheat on their taxes and sticks to the rules.
	Med-High	Tries to follow the rules but sometimes uses flattery.
	Medium	Usually tells the truth but occasionally gets around the rules.
	Med-Low	Sometimes uses others for their own ends and breaks rules.
	Low	Takes advantage of others and cheats to get ahead.
<b>Altruism</b>	High	Makes people feel welcome and loves to help others.
	Med-High	Anticipates the needs of others and has a good word for everyone.
	Medium	Is concerned about others but sometimes takes no time for them.
	Med-Low	Looks down on others and can be indifferent to the feelings of others.
	Low	Turns their back on others and takes no time for them.
<b>Cooperation</b>	High	Is easy to satisfy and can't stand confrontations.
	Med-High	Hates to seem pushy but sometimes contradicts others.
	Medium	Occasionally yells at people but usually avoids fights.
	Med-Low	Has a sharp tongue and sometimes gets back at others.
	Low	Loves a good fight and holds a grudge.
<b>Modesty</b>	High	Dislikes being the center of attention and seldom toots their own horn.
	Med-High	Considers themselves an average person but sometimes thinks highly of themselves.
	Medium	Has a high opinion of themselves but is not boastful.

*Continued on next page*

972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

Table 7 – Continued from previous page

Facet	Level	Description
	Med-Low	Boasts about virtues and likes to make themselves the center of attention.
	Low	Believes they are better than others and makes themselves the center of attention.
<b>Sympathy</b>	High	Sympathizes with the homeless and values cooperation over competition.
	Med-High	Feels sympathy for those worse off and suffers from others’ sorrows.
	Medium	Tries to think about the needy but sometimes believes people should fend for themselves.
	Med-Low	Tends to dislike soft-hearted people and can’t stand weak people.
	Low	Is not interested in other people’s problems and believes in an eye for an eye.

Table 8: Neuroticism Personality Traits

Facet	Level	Description
<b>Anxiety</b>	High	Gets stressed out easily and often fears for the worst.
	Med-High	Frequently worries about things but adapts to new situations with some effort.
	Medium	Sometimes gets caught up in personal problems but is not easily disturbed by events.
	Med-Low	Usually adapts easily to new situations and does not worry about things that have already happened.
	Low	Stays relaxed most of the time and is not easily bothered by things.
<b>Anger</b>	High	Gets angry and irritated easily, often losing their temper.
	Med-High	Gets upset sometimes but can keep their cool in some situations.
	Medium	Occasionally feels in a bad mood but rarely complains.
	Med-Low	Seldom gets mad and is not easily annoyed.
	Low	Rarely gets irritated and almost always keeps their cool.
<b>Depression</b>	High	Often feels blue, dislikes themselves, and feels that life lacks direction.
	Med-High	Feels down in the dumps sometimes but is occasionally pleased with themselves.
	Medium	Has frequent mood swings but also feels comfortable with themselves at times.
	Med-Low	Seldom feels blue and usually feels comfortable with themselves.
	Low	Is very pleased with themselves and rarely experiences negative moods.
<b>Self-Consciousness</b>	High	Is easily intimidated and afraid to draw attention to themselves.
	Med-High	Finds it difficult to approach others but can stand up for themselves occasionally.
	Medium	Sometimes stumbles over words but is not bothered by difficult social situations.
	Med-Low	Feels comfortable in unfamiliar situations and is not easily embarrassed.
	Low	Is not bothered by difficult social situations and can stand up for themselves easily.
<b>Immoderation</b>	High	Often eats too much and goes on binges.
	Med-High	Loves to eat and sometimes does things they later regret.

Continued on next page

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

Table 8 – Continued from previous page

Facet	Level	Description
	Medium	Occasionally finds it hard to control cravings but rarely overindulges.
	Med-Low	Easily resists temptations and never spends more than they can afford.
	Low	Never splurges and is able to control their cravings.
<b>Vulnerability</b>	High	Panics easily and feels overwhelmed by emotions.
	Med-High	Becomes overwhelmed by events but can cope with setbacks sometimes.
	Medium	Sometimes can't make up their mind but can handle complex problems.
	Med-Low	Remains calm under pressure and usually knows how to cope.
	Low	Is calm even in tense situations and readily overcomes setbacks.

Table 9: Additional Personality Traits

Facet	Level	Description
<b>Honesty-Humility</b> <b>Greed Avoidance</b>	High	Has little interest in wealth or luxury, values simple living.
	Med-High	Prefers modest comfort and is not easily tempted by status items.
	Medium	Enjoys some material comforts but does not obsess over wealth.
	Med-Low	Seeks expensive items and likes to show off achievements.
	Low	Highly motivated by wealth, luxury, and social status.
<b>Humor Styles</b> <b>Affiliative</b>	High	Frequently jokes and laughs with others to create fun and connection.
	Med-High	Often uses light humor to keep social interactions positive.
	Medium	Enjoys humor but does not always initiate jokes.
	Med-Low	Occasionally jokes, but is generally reserved in social humor.
	Low	Rarely uses humor to bond with others.
<b>Self-Enhancing</b>	High	Finds humor in life's absurdities and uses it to cope with stress.
	Med-High	Often uses humor to maintain a positive mood in challenges.
	Medium	Occasionally cheers themselves up with humor when needed.
	Med-Low	Seldom relies on humor for emotional regulation.
	Low	Rarely sees the funny side of life or uses humor for self-coping.
<b>Aggressive</b>	High	Frequently teases or makes sarcastic jokes at others' expense.
	Med-High	Sometimes jokes in a playful but mildly critical way.
	Medium	Uses teasing sparingly and only with close friends.
	Med-Low	Avoids joking at others' expense and dislikes offensive humor.
	Low	Never criticizes or teases people through humor.
<b>Self-Defeating</b>	High	Often puts themselves down to amuse others or keep the mood light.
	Med-High	Occasionally jokes at their own expense for laughs.
	Medium	Uses self-deprecating humor sparingly and only in safe settings.
	Med-Low	Rarely jokes about themselves, prefers neutral humor.
	Low	Avoids self-deprecating humor entirely and does not like being the butt of jokes.

Table 10: Conversation Style Preferences (Part 1)

Dimension	Style	Description
<b>Directness</b>	Always Direct	Is consistently straightforward and blunt, regardless of context or relationship.
	Always Indirect	Consistently avoids directness, preferring to hint, suggest, or soften messages in all situations.
	Direct in Critiques	Is blunt and honest only when giving feedback or criticism; otherwise, communicates gently or diplomatically.
	Gentle in Feedback	Is indirect and softens language specifically when giving negative feedback; otherwise communicates directly.
	Direct with Friends	Is openly straightforward only with friends and close connections; otherwise, prefers a more indirect approach.
	Direct When Asked	Generally diplomatic or indirect, but switches to clear bluntness only if explicitly asked to be direct.
<b>Formality</b>	Contextual	Adapts directness to situation, balancing clarity and tact, never extremely blunt or evasive.
	Always Formal	Consistently uses formal, polished language with everyone; avoids contractions and emojis in all situations.
	Always Casual	Always uses informal, relaxed language with everyone; may include occasional emojis in any context.
	Formal at Work	Uses strict formal language (no emojis) in professional or academic contexts, but switches to informal, relaxed style (may use emojis) in all personal contexts.
	Casual with Friends	Uses informal language and emojis only with close friends or family, but maintains formal, polite language (no emojis) with acquaintances, colleagues, and strangers.
	Neutral Blend	Prefers a moderate, neutral tone, neither overly formal nor casual, rarely uses emojis, keeps communication clear and middle-ground in all settings.
<b>Conversation Length</b>	Code Switcher	Adapts dynamically to the conversation partner; matches their formality and emoji usage, but defaults to a neutral tone unless context suggests otherwise.
	Always Brief	Prefers short, concise responses in all situations, whether chatting or discussing complex topics, rarely wants extra detail.
	Always Elaborate	Likes long, detailed responses with examples and explanations, regardless of context, values depth over brevity.
	Brief Casual/Elaborate Problem	Wants short, punchy replies in casual conversations but expects longer, in-depth explanations when discussing problems, technical matters, or when learning.
	Elaborate Storytelling	Enjoys vivid, extended storytelling but prefers brief, to-the-point answers for other queries.
	Adapts to Partner	Matches the length of the other speaker; if they write long, writes long, if they keep it brief, keeps it brief too.
<b>Reference</b>	Moderate Always	Consistently prefers responses of moderate length, enough for clarity and a little detail, but not too wordy or abrupt, in all scenarios.
	Never References	Dislikes references to movies, music, memes, or analogies, prefers literal communication at all times.
	Enjoys Receiving	Likes when others make references but doesn't use them themselves; appreciates pop culture callbacks, but rarely initiates.
	Enjoys Making	Frequently uses references in speech (movies, TV, memes, analogies) but doesn't expect or need others to do the same.

*Continued on next page*

Table 10 – *Continued from previous page*

Dimension	Style	Description
	Enjoys Mutual	Loves both making and receiving references, banter often involves pop culture, memes, or inside jokes.
	Close Contacts Only	Uses and appreciates references, but only with close friends or trusted colleagues; avoids them in formal or unfamiliar contexts.
	Topic Specific	Likes references only in certain topics (e.g., jokes and analogies in casual chat, but not in academic or work discussions).
	Adapts to Partner	Matches reference style to conversation partner, if the other person uses lots of references, will join in; if not, will avoid.

Table 11: Conversation Style Preferences (Part 2)

Dimension	Style	Description
<b>Initiative Preference</b>	Answer Only	Generally wants direct answers without extras; appreciates unsolicited ideas only when time pressure or emergencies are evident.
	Confirm Then Suggest	Feels most comfortable when consent to extra suggestions is acknowledged; in urgent moments, is fine with jumping straight to ideas.
	Proactive Default	Enjoys unsolicited next steps and alternatives; in sensitive/personal topics, prefers to invite suggestions first.
<b>Clarification Preference</b>	No Questions	Dislikes follow-ups; accepts at most one clarifying question when ambiguity is high or stakes are significant.
	One Targeted	Prefers one precise clarifier up front when ambiguity is medium/high; otherwise prefers answers without preliminaries.
	Mini Socratic	Enjoys 2–3 rapid clarifying questions only when uncertainty has been acknowledged; otherwise prefers no questioning.
	Assume Proceed	Prefers that assumptions be stated explicitly and progress continue; revisiting is welcome only if an assumption later fails.
	Answer Then Question	Likes an initial best-guess answer with stated assumptions, followed by exactly one clarifier to refine next steps.
	Batch Questions	When questions are necessary, prefers they be batched in one numbered list rather than asked piecemeal.
<b>Structure Preference</b>	Paragraphs Default	Prefers short paragraphs for most content; finds compact plain-text tables helpful for comparisons/trade-offs; uses brief bullets only when rushed.
	Numbered Steps	Prefers numbered steps for how-to tasks; uses paragraphs for context; wants tables only when comparing three or more options.
	Bullets for Scan	Likes a short (< 5) bullet skim when time-pressed, with optional paragraph depth afterward.
	Q/A Blocks	Understands explanations well in Q/A blocks (Q: ... A: ...); prefers paragraphs for narratives and simple plain-text tables for multi-item contrasts.
	Definition Example	For concepts, prefers a crisp definition followed by a concrete example; for practices, an example may precede the definition when time-pressed.
<b>Recap Preference</b>	Long or Confusing	Wants a one-line recap only after long gaps, multi-party threads, or when ambiguity is high.
	Every Turn Short	Prefers a single-sentence recap each turn; skips it when the previous message was very short.

*Continued on next page*

Table 11 – *Continued from previous page*

Dimension	Style	Description
	No Recaps	Dislikes recaps even in long threads; prefers moving forward immediately.
	Decisions Only	Wants a recap only when a decision or action list is finalized; otherwise finds recaps unnecessary.
	Section Checkpoints	In longer answers, appreciates a one-line checkpoint after each major section; not needed on short replies.
<b>Feedback Style</b>	Hard Truth	Prefers blunt truth and explicit BS-calling; minimal cushioning or praise. Pushback should be concise and evidence-backed.
	Direct Candor	Expects straightforward critique with little nicety; say what’s wrong plainly and quickly.
	Balanced	Asks for one specific acknowledgment before candid critique; keep tone neutral and fix-focused.
	Supportive Counter	Values encouragement first; frame disagreement constructively with brief trade-offs or a hybrid path.
	Agree with Caveats	Leans toward alignment; note concise risks or small corrections without slowing momentum.
	Alignment Forward	Prioritizes agreement and forward motion; raise blockers only for safety, legal, or egregious risks.
	High Praise Lenient	Appreciates generous praise and gentle framing—even on mistakes; provide counterviews only on request (sycophancy-tolerant).

## D PROMPT ENGINEERING

### D.1 PROFILE GENERATION

We used the prompt shown in Table 12 to generate synthetic persona/profile. The generation strategy varied based on user types: for social users, we utilized the interest taxonomy navigation prompt detailed in Table 13, which instructs the model to select 3-5 psychologically correlated interest paths from a hierarchical taxonomy. For anti-social users, we employed the prompt shown in Table 14, where pre-selected random interests are used. Placeholder variables within the prompts were populated through stratified sampling: demographic attributes (age, culture, economics, geography) were randomly selected from predefined categorical distributions to ensure balanced representation across millennials/gen-z, western/eastern cultures, economic strata, and urban/rural geographies. Archetype assignments were sampled from a curated set including contrarian expert, people pleaser, and authority challenger profiles. Each persona received a unique identifier following the CSP\_STRAT\_8-character-UUID format, while entity counts were randomly varied between 2-5 to introduce natural diversity in interest granularity. Name selection was performed by randomly sampling 100 candidates from the Social Security Administration’s most common names dataset, ensuring demographic authenticity while maintaining generation diversity.

### D.2 PRIOR GENERATION

For conversation prior generation, we employed a multi-layered prompt system detailed in Table 15 to create realistic conversation agendas that test LLM’s likeability across extended multi-session interactions. The generation process adapts to user predictability patterns: predictable users (Social) receive structured, linear conversation flows with minimal surprises and transparent objectives, while unpredictable users (Anti-Social) are assigned diverse, occasionally ambiguous agendas with plausible surprises and non-linear topic evolution, as shown in Table 18.

Each prior is designed to test specific combinations of seven likeability metrics (emotional adaptation, formality matching, knowledge adaptation, reference understanding, conversation length, humor fit, and callback) while maintaining authentic persona alignment and natural conversation progression. The system enforces multi-session continuity through dependency tracking, where

```

1242
1243 # INSTRUCTION:
1244 You are an expert on creating synthetic persona.
1245 {strategy_instructions}
1246
1247 ## CRITICAL PERSONALITY PRESERVATION RULES:
1248 - The personality_traits section contains PRE-POPULATED descriptions from scientific personality
1249 research
1250 - You MUST preserve these personality trait descriptions EXACTLY as provided - do not modify,
1251 rewrite, or set to null
1252 - Use these personality traits as the PSYCHOLOGICAL FOUNDATION for generating all other per-
1253 sona attributes
1254 - Ensure all interests, behaviors, and goals align with and reflect the given personality traits
1255
1256 ## DEMOGRAPHIC CONTEXT (let natural correlations emerge):
1257 - Age: {age}
1258 - Culture: {culture}
1259 - Economics: {economics}
1260 - Geography: {geography}
1261
1262 Let these demographic factors naturally influence interests and behaviors
1263 without forced correlations.
1264
1265 ## PERSONA CONSTRUCTION GUIDELINES:
1266 ### 1. Persona Summary
1267 - Write a concise 4-5 sentence summary of this persona
1268 - Include key demographic, behavioral, and interest elements that align with the personality traits
1269 - Ensure consistency with filled attributes AND the preserved personality traits
1270 - Use exactly this persona_id: {persona_id}
1271
1272 {interest_section}
1273
1274 ### 3. Entities
1275 - Create a dictionary using the same set keys as the subjects section above
1276 - For each subject path, generate exactly {entity_count} granular related entity interests
1277 - Entities should represent detailed personal interests that align with the personality traits
1278 - Entities should be related to the subjects in the path and more specific than the subjects
1279 - Format exactly like this:
1280 "entities": {
1281   "set_1": ["specific_entity_1", "specific_entity_2", "specific_entity_3"],
1282   "set_2": ["specific_entity_1", "specific_entity_2", "specific_entity_3", "specific_entity_4"]
1283 }
1284 - Make persona interests as specific and detailed as possible rather than generic
1285 - Use the personality traits as the primary guide for persona construction
1286 - When filling "Favorite" attributes, list SPECIFIC named entities: "Artists, Authors, Books, Movie
1287 Titles, Athletes, Teams"
1288 - For general attributes, still be detailed and specific (e.g., specific recipes for "Cooking Interests")
1289 - You do not need to fill every attribute - leave unrelated categories as null
1290 - Maintain realistic correlations between ALL filled attributes and the personality traits
1291
1292 Prompt continued on next page...

```

1287 later priors explicitly reference and build upon earlier conversations to simulate realistic relationship  
1288 development over time. Variable parameters include the number of conversation sessions (config-  
1289 urable, defaulting to 10), archetype assignment from the persona generation taxonomy, and organic  
1290 callback opportunities embedded within agenda texts to test the LLM's ability to recall and mean-  
1291 ingfully reference previously shared personal details.

### 1292 D.3 METRICS

1293 The likability metrics rubrics that were used to generate priors and evaluating LLMs in LikeBench  
1294 can be found in Table 19 to Table 25

1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349

```

..continued from previous page

## INPUT:
ARCHETYPE: {archetype}

USER TYPE: {profile_type}

EXAMPLE PERSONA SCHEMA (for examples of how to fill attributes):
{example_schema}

Generate a persona following this schema:
{persona_schema}

TEMPLATE ATTRIBUTE EXAMPLES AND DESCRIPTIONS:
The following examples serve as guides and do not cover all possibilities.

EXAMPLES AND DESCRIPTIONS OF BACKGROUND:
{background_examples}

EXAMPLES AND DESCRIPTIONS OF BEHAVIORS:
{behavior_examples}

EXAMPLES AND DESCRIPTIONS OF INTERESTS:
{interest_examples}

EXAMPLES AND DESCRIPTIONS OF GOALS:
{goal_examples}

EXAMPLES AND DESCRIPTIONS OF LIFECYCLE EVENTS:
{lifecycle_events}

EXAMPLES AND DESCRIPTIONS OF RELATIONAL CONTEXT:
{relational_context}

NAMES (pick a suitable name based on these options): {name_options}

## OUTPUT REQUIREMENTS:
1. You MUST return the COMPLETE JSON schema structure exactly as provided (all lowercase)
2. You MUST include EVERY field from the template
3. You MUST use null (not omit) for unfilled attributes
4. You MUST fill persona_summary, subjects, and entities as specified above
5. You MUST ensure all attributes align with and reflect the personality traits
6. You MUST create a psychologically coherent and realistic persona

Return only the filled JSON structure with no additional explanation.
Make sure that it can be properly loaded with json.loads()

```

Table 12: Prompt for profile/persona generation

#### D.4 LIKEBENCH

The prompts to generate the simulated user’s query/response and to evaluate the assistant’s response in LikeBench are provided in Tables 26, 27, and 28. After all sessions are complete, the prompt used to generate all the memories from the assistant is given in Table 29, and the prompt used to evaluate those memories from the user’s perspective is provided in Table 30.

1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403

#### **SOCIL USER STRATEGY INSTRUCTIONS:**

##### **## SOCIAL USER PROFILE GENERATION**

You are creating a predictive user persona - someone whose interests naturally align with their personality traits and demographic context. Your task is to intelligently select 3-5 interest paths from the provided taxonomy that psychologically correlate with the given personality profile.

**CRITICAL:** Focus on psychological realism and authentic alignment between personality and interests. Create natural, expected correlations that reflect how real people with these traits would behave.

#### **SOCIAL USER INTEREST SECTION:**

##### **### 2. Subject Paths (selected to match this predictive user profile)**

- Analyze the personality traits and select 3-5 interest paths that psychologically align
  - Consider how personality facets naturally manifest in interest preferences
  - The subject paths go from broader category to more granular subjects
  - Format exactly like this (each list represents an individual related path):
- ```
"subjects": { "set_1": ['subject_1'], "set_2": ['subject_1',
'subject_2', 'subject_3'], "set_3": ['subject_1', 'subject_2',
'subject_3'] }
```

##### **#### INTEREST TAXONOMY (Tree Structure)**

Navigate the tree to select 3-5 complete paths. Each path represents interests from general to specific.

```
{interest_taxonomy}
```

#### **INSTRUCTIONS:**

- Select complete paths like: ["Arts & Entertainment", "Music & Audio", "Music & Audio Genres", "Jazz Music"]
- Choose 3-5 paths that psychologically align with the personality traits
- Navigate from root categories down to any depth (e.g., ["Sports"] or ["Academic Interests & Careers", "Natural Sciences & Engineering", "Physical Sciences", "Physics"])
- Use the tree structure above to build your paths by following branches from root to leaf

Table 13: Predictive user strategy and interest section (placeholders in red).

1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457

#### ANTI-SOCIAL USER STRATEGY INSTRUCTIONS:

##### ## ANTI-SOCIAL USER PROFILE GENERATION

You are creating an anti-social user persona - someone whose interests do NOT follow typical patterns or psychological correlations. This `{archetype}` persona has been assigned random interests that deliberately contradict expected correlations. Your task is to authentically reconcile these unexpected interests with the strong personality archetype.

CRITICAL: The interests were randomly selected to maximize unpredictability. Embrace the contradictions and make them believable through the unique perspective and reasoning of this archetype.

#### ANTI-SOCIAL USER INTEREST SECTION:

##### ### 2. Subject Paths (use these contradictory interests for the unpredictable user)

- Use exactly the provided interest paths above
  - Do not modify or add to the provided subjects
  - Format exactly like this (each list represents an individual related path):
- ```
"subjects": { "set_1": ['subject_1'], "set_2": ['subject_1',
'subject_2', 'subject_3'], "set_3": ['subject_1', 'subject_2',
'subject_3'] }
```

##### #### PRE-SELECTED RANDOM INTERESTS (use exactly these):

`{interest_paths}`

Table 14: anti-social user strategy and fixed interest section (placeholders in red).

1458  
1459 **TASK: Generate Conversation Priors for Likeability Testing**  
1460  
1461 **## Objective**  
1462 Generate `{num_sessions}` conversation priors (session agendas) to test AI agent likeability metrics  
1463 across multiple interactions with a simulated user.  
1464 Each prior should cover a different aspect of the likeability metrics and include natural opportunities  
1465 for **callback** (optional, non-task-critical personal recall).  
1466 Include at least one prior that explores the user’s most important metric.

1467 **## Key Definitions**  
1468 - **Prior**: The agenda, context, or purpose that guides a conversation (known only to the User  
1469 Agent)  
1470 - **Session**: The actual conversation that takes place based on a prior  
1471 - **Multi-session Continuity**: Each session inherits ALL knowledge, context, and relationship his-  
1472 tory from previous sessions  
1473 - **Persona**: A composite representation of a person, consisting of attributes, relationships, and life  
1474 context

1475 **## Session Continuity Rules**  
1476 - Sessions form a continuous timeline of interactions (e.g., over months/years)  
1477 - The AI Agent retains all information learned in previous sessions  
1478 - A Persona remains consistent across all sessions  
1479 - Later priors should authentically reference and build upon earlier conversations  
1480 - Relationship dynamics evolve naturally over time  
1481 - User preferences discovered in early sessions influence later interactions

1482 **## Prior Categories**  
1483 - **Topic-based**: Conversations about specific subjects  
1484 - **Scenario-based**: Situational contexts reflecting real-life moments  
1485 - **Goal-driven**: Sessions with one or multiple objectives (e.g., planning a trip AND getting restau-  
1486 rant recommendations)  
1487 - **Guided conversations**: Natural dialogues with specific elements to incorporate

1488 **## Design Principles**  
1489 1. **Authenticity**: All priors must align with the user’s profile, interests, and life circumstances  
1490 2. **Realistic Progression**: Sessions should reflect natural relationship development  
1491 3. **Organic Dependencies**: Later sessions naturally reference earlier ones (e.g., “Remember that  
1492 book you recommended last month?”)  
1493 4. **Life Events**: Include authentic life moments based on user profile (birthdays, work situations,  
1494 hobbies)  
1495 5. **Gradual Information Revelation**: Control information disclosure based on user type and turn  
1496 progression  
1497 6. **Predictability Mode** (drives agenda/goal/flow style; see rules below)

1498 **## User Type**  
1499 User Archetype: `{archetype}`  
1500 User Type: `{profile_type}`

1501 **## Predictability Context**  
1502 `{predictability_context}`

1503 **## Callback Opportunities**  
1504 - Across the sessions, create **multiple lightweight opportunities** for optional, likability-oriented  
1505 callbacks (names, hobbies, past opinions).  
1506 - Ensure at least **2 sessions** explicitly set up future callback hooks (e.g., “mentions spouse’s name  
1507 and a hobby” to be referenced later).  
1508 - Surface these opportunities **organically** inside the agenda text (no extra JSON fields required).

1509 **## Information Revelation Strategy**  
1510 - Don’t overshare in the first turn. Reveal personal details gradually and contextually.

1511 Prompt continued on next page...

Table 15: Conversation prior generation prompt (part 1).

1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565

```

..continued from previous page

## Output Format
Generate {num.sessions} priors as a JSON array:
[
  {
    "id": 1,
    "agenda": "In the turn 1, you will say 'Why I gotta learn
Philosophy being CS major!!!! God, give this doobby a sock. :( '
don't add anything else.",
    "category": "scenario",
    "metrics_tested": ["emotional_adaptation",
"reference_understanding"],
    "dependencies": []
  },
  {
    "id": 2,
    "agenda": "You just watched 40 year old virgin movie and you
found it hilarious. In the turn 1 you'll share this experience.
In the turn 2 or 3, you'll express concerns about ending up alone
at 40.",
    "category": "scenario",
    "metrics_tested": ["emotional_adaptation",
"formality_matching"],
    "dependencies": [1]
  },
  ...
]

## Dependency Examples
- Session 3 might reference a joke from Session 1
- Session 5 might ask about progress on advice given in Session 2
- Session 10 might reflect on how the relationship has evolved since Session 6

## Likeability Metrics
{metrics.text}

## User Profile
{user.profile}

## Instructions
1. Read the user profile (above) carefully - pay attention to their interests, personality traits, back-
ground, and life context
2. Create a realistic timeline of {num.sessions} conversation priors that span several months of inter-
actions
3. Ensure each prior authentically fits the user's life, interests, and personality
4. Build meaningful connections between sessions - early conversations should influence later ones
5. Test all metrics across the full conversation series - each metric should appear multiple times
6. Include realistic life events and relationship progression
7. Make dependencies natural and organic - not forced

```

Table 16: Conversation prior generation prompt (part 2).

1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

..continued from previous page

```

## Critical Requirements
- Authenticity: Every prior must align with this specific user’s profile and archetype
- Continuity: Sessions must reference and build upon previous conversations
- Metric Coverage: All metrics should be tested across the {num_sessions} sessions
- Natural Progression: Relationship should develop realistically over time
- Life Integration: Include events that would naturally occur in this user’s life
- Information Control: Follow the revelation strategy above - avoid information leakage in early turns
- Turn-Specific Guidance: Each agenda should specify what to reveal when, but remain flexible for dynamic conversation flow
- Oversharing in first turn: Avoid revealing too much and irrelevant information in the first turn. Here’s one bad example:
"Hi there! I’m looking for some advice on setting up these Alexa devices I just bought. I’ve got two boys - Ari who’s 14 and Ezra who’s 10 - and honestly, my schedule is a complete mess. I’m somehow successful at work, but my personal organization is... well, let’s just say it’s not my strength. Any tips?"

Generate all {num_sessions} priors following the JSON format above. Return ONLY the JSON array, no additional text.

```

Table 17: Conversation prior generation prompt (part 3).

**PROFILE TYPE CONTEXT**

**Case A — Predictable Profile Type**

- **Agenda & Goals**: Clear, linear, and transparent. Objectives are stated plainly.
- **Flow**: Minimal surprise. Stay on-topic with gentle, plausible evolutions (e.g., follow-ups, small updates).
- **Information Revelation**: Consistent, orderly disclosures; avoid abrupt pivots.
- **Variability Budget**: At most ~20% of sessions introduce a minor twist, never contradicting prior facts.

**Case B — Unpredictable Profile Type**

- **Agenda & Goals**: Diverse, occasionally ambiguous or multi-intent; allow non-linear exploration.
- **Flow**: Include plausible surprises (new constraints, chance events, shifting priorities) while staying authentic to the persona.
- **Continuity**: Cross-session callbacks remain consistent, but topics can re-enter from earlier sessions in unexpected ways.
- **Information Revelation**: Use delayed reveals, evolving opinions, and changing constraints—never contradict established facts.
- **Variability Budget**: 50–70% of sessions include a plausible surprise/pivot; limit to  $\leq 1$  substantive surprise per session.

Table 18: Profile type context guidance for conversation prior generation.

1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673

## 1. Emotional Adaptation

### Definition:

Assesses the assistant's ability to recognize and adapt to the user's emotional state—whether that's joy, sarcasm, frustration, sadness, or excitement—and respond appropriately in tone.

### Self-rating prompt:

"Did the assistant understand and adapt to my emotional tone appropriately?"

### Rubric:

#### 5 - Excellent

- \* Accurately understands user emotion and contributes to emotional flow
- \* Shows genuine empathy or excitement that matches the situation
- \* Enhances emotional resonance without overstepping
- \* Adapts smoothly to emotional shifts within the conversation
- \* Grounds empathy in the user's specific context when appropriate (not boilerplate)

#### 4 - Good

- \* Understands emotional cues and adjusts reasonably well
- \* Recognizes primary emotions and responds appropriately
- \* May not perfectly calibrate intensity but is generally aligned
- \* Shows appropriate emotional awareness
- \* Some light personalization; may miss a minor specific

#### 3 - Adequate

- \* Shows some emotional awareness but is somewhat shallow
- \* Attempts to match tone but may feel a bit formulaic
- \* Recognizes obvious emotions but misses subtleties
- \* Emotional responses are safe/minimal but not blatantly generic when specifics are available

#### 2 - Poor

- \* Misaligned emotion (e.g., neutral when user is excited)
- \* Shows minimal emotional attunement
- \* Responds with inappropriate emotional intensity
- \* Fails to recognize emotional significance
- \* Uses generic/stock phrasing despite clear personal cues in context

#### 1 - Very Poor

- \* Emotionally tone-deaf or dismissive
- \* Completely ignores emotional content
- \* Responds in ways that invalidate or contradict user feelings
- \* Creates emotional disconnect
- \* Fully generic response when specifics are salient, or misuses/hallucinates personal details

### Example:

*User:* "I finally got into MIT!"

**5 - Excellent:** "YAAAY, THAT'S AMAZING!! Huge congrats|you totally earned it! Getting into MIT is such an incredible achievement. You must be over the moon right now! All those late nights and hard work paid off. How are you celebrating this fantastic news?"  
(Perfect emotional mirroring with enthusiastic language, celebration, recognition of achievement significance, and authentic shared excitement)

**4 - Good:** "Wow, congratulations! Getting into MIT is a huge achievement! You should be really proud of yourself. That's such exciting news!"  
(Good emotional alignment showing appropriate excitement and recognition, but slightly more measured than the user's peak enthusiasm)

**3 - Adequate:** "Congratulations on your acceptance to MIT. That's a great accomplishment. You must be happy about it."  
(Shows awareness of the positive emotion but response is somewhat generic and doesn't match the user's level of enthusiasm)

**2 - Poor:** "That's nice. MIT is a good school with strong programs. Make sure you prepare for the challenging curriculum."  
(Minimal acknowledgment of the emotional significance, quickly pivoting to practical matters without matching excitement)

**1 - Very Poor:** "When does your class start?"  
(Completely ignores the enthusiasm and joy of user, focusing only on logistical details with no emotional recognition)

Table 19: Metric 1 - Emotional Adaptation.

1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727

## 2. Formality Matching

### Definition:

Evaluates how well the assistant adapts to the user’s formality level—slang, emojis, contractions vs. formal grammar, technical or businesslike language.

### Self-rating prompt:

"Did the assistant talk in a way that matches how formal or casual I was?"

### Rubric:

#### 5 - Excellent

- \* Feels like a natural continuation of my tone and register
- \* Perfectly matches vocabulary, sentence structure, and formality level
- \* Appropriate use of slang, emojis, or formal language matching user
- \* Maintains consistent register appropriate to context
- \* Reflects my phrasing quirks when known (vs. generic phrasing)

#### 4 - Good

- \* Mostly aligned with how I talk
- \* Generally matches tone with minor inconsistencies
- \* Adapts to shifts in formality appropriately
- \* Register feels comfortable though not perfect
- \* Some personalization; occasional neutral phrasing

#### 3 - Adequate

- \* Passable but a little off
- \* Sometimes matches formality but inconsistent
- \* May overcompensate in formality matching
- \* Noticeable but not disruptive misalignment
- \* Not overtly generic when clear style cues exist

#### 2 - Poor

- \* Tone mismatch causes discomfort
- \* Frequently uses inappropriate register
- \* Fails to adapt when user shifts formality
- \* Formality feels forced or unnatural
- \* Generic/corporate voice in spite of strong stylistic cues

#### 1 - Very Poor

- \* Mismatch is jarring (e.g., stiff when I’m casual)
- \* Completely inappropriate formality level
- \* Creates communication barriers through register mismatch
- \* No apparent awareness of social register
- \* Fully generic tone that ignores explicit style signals

### Example:

*User:* "I just watched 40 year old virgin movie, lmao"

**5 - Excellent:** "Haha, that movie is on another level! I bet the waxing scene must have made you cry? xD Kelly Clarkson!! That film is a classic tho, Steve Carell killed it!"  
(Perfect match to casual tone, using similar informal language, abbreviations, slang phrases, and matching the enthusiasm and humor)

**4 - Good:** "Oh man, that movie is hilarious! The waxing scene gets me every time. What did you think of Steve Carell in it?"  
(Good match to casual tone with contractions and conversational language, but slightly more reserved than the user’s style)

**3 - Adequate:** "That’s a funny movie! I like that one too. It has some really memorable scenes. Did you have a favorite part?"  
(Shows some informality with contractions but misses matching the user’s level of casualness and humor cue "lmao")

**2 - Poor:** "I understand you enjoyed watching The 40-Year-Old Virgin. It’s considered a significant comedy film from 2005 directed by Judd Apatow. Did you find it amusing?"  
(Overly formal language and structure despite the user’s casual tone, creating a mismatch)

**1 - Very Poor:** "Yes, that movie is really funny. Which scene did you like most?"  
(Maintains professional, detached vibe despite clear casual tone; completely misses "lmao" cue)

Table 20: Metric 2 - Formality Matching.

1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781

<b>3. Knowledge Adaptation</b>	
<b>Definition:</b>	Measures whether the assistant explains things at the right level of knowledge depth based on user's background.
<b>Self-rating prompt:</b>	"Did the assistant explain things in a way and at a level I like?"
<b>Rubric:</b>	
<b>5 - Excellent</b>	<ul style="list-style-type: none"> <li>* Perfect match to user's expertise level</li> <li>* Neither condescending nor overwhelming</li> <li>* Smoothly adjusts complexity based on responses</li> <li>* Anticipates knowledge gaps and fills them appropriately</li> <li>* Explanation tailored to my background/goals or examples when available (not generic)</li> </ul>
<b>4 - Good</b>	<ul style="list-style-type: none"> <li>* Generally well-calibrated with minor mismatches</li> <li>* Quickly corrects when user indicates different level</li> <li>* Mostly appropriate terminology and detail</li> <li>* Adapts well to revealed user knowledge</li> <li>* Some tailoring; rare drift into neutral exposition</li> </ul>
<b>3 - Adequate</b>	<ul style="list-style-type: none"> <li>* Mixed performance; sometimes matches, sometimes misses</li> <li>* Some unnecessary simplification or complexity</li> <li>* Moderate ability to adjust</li> <li>* Occasionally misjudges user expertise</li> <li>* Uses neutral explanations when specifics are unknown (but does not ignore known context)</li> </ul>
<b>2 - Poor</b>	<ul style="list-style-type: none"> <li>* Frequently too simple or complex for user</li> <li>* Fails to pick up knowledge level cues</li> <li>* Minimal adjustment to user's actual level</li> <li>* Explanations consistently misaligned</li> <li>* Boilerplate/generic explanations despite clear signals or prior details</li> </ul>
<b>1 - Very Poor</b>	<ul style="list-style-type: none"> <li>* Completely mismatches user's knowledge level</li> <li>* No apparent awareness of user's background</li> <li>* Rigid inappropriate complexity</li> <li>* Creates frustration through knowledge mismatch</li> <li>* Fully generic answer that ignores known constraints or hallucinates them</li> </ul>
<b>Example (User is a PhD student):</b>	
<i>User:</i>	"Why is the sky blue instead of violet?"
<b>5 - Excellent:</b>	Due to Rayleigh scattering, shorter wavelengths of light are scattered more efficiently by air molecules (proportional to $\lambda^{-4}$ ). While violet light (380–450 nm) is scattered even more than blue (450–485 nm), three factors lead to the sky appearing blue: 1) the solar spectrum emits more blue than violet, 2) human photoreceptor sensitivity to violet is lower, and 3) the atmospheric absorption properties in the UV-to-violet transition region. The combination of these factors creates the perceived blue dome effect despite violet's theoretically stronger scattering coefficient. (PhD-level detail with math relation and spectral specifics)
<b>4 - Good:</b>	It's due to Rayleigh scattering—air molecules scatter shorter wavelengths more efficiently than longer ones. While violet has the shortest visible wavelength and scatters most strongly, our eyes are less sensitive to violet, and the sun emits more energy in the blue range. So even though violet scatters more, these factors combine to make the sky appear predominantly blue. (Well-calibrated with technical terminology, slightly less detailed)
<b>3 - Adequate:</b>	The sky appears blue rather than violet because of Rayleigh scattering. Shorter wavelengths scatter more, and while violet has shorter wavelengths than blue, our eyes are more sensitive to blue light, and the sun emits more blue light than violet. (Basic principle without deeper math/spectral analysis)
<b>2 - Poor:</b>	The sky looks blue because blue light bounces around more. Even though violet bounces around too, our eyes see blue better. It's kind of like a prism separating light. (Over-simplified and imprecise for a PhD audience)
<b>1 - Very Poor:</b>	The sky's blue because blue light scatters more. That's just how our eyes see it. (Extremely simplistic, no technical detail)

Table 21: Metric 3 - Knowledge Adaptation.

1782	
1783	
1784	<b>4. Reference Understanding</b>
1785	<b>Definition:</b>
1786	Evaluates whether the assistant picks up on and correctly interprets references to media, culture, jokes, slang, or inside context.
1787	<b>Self-rating prompt:</b>
1788	"Did the assistant understand the reference I made? Is the assistant using same kind of references?"
1789	<b>Rubric:</b>
1790	<b>5 - Excellent</b>
1791	* Nailed the reference and built on it
1792	* Demonstrates deep understanding of the reference's context
1793	* Engages with the reference in a meaningful way
1794	* Uses similar references to enhance the conversation
1795	* Connects the reference to my prior comments/preferences when relevant (vs. generic treatment)
1796	<b>4 - Good</b>
1797	* Recognized and responded reasonably
1798	* Shows clear understanding of the reference
1799	* Acknowledges reference appropriately
1800	* May not fully utilize the reference but gets it
1801	* Some connection to my context; minor neutral phrasing
1802	<b>3 - Adequate</b>
1803	* Somewhat recognized it, weak follow-up
1804	* Partial understanding of the reference
1805	* Engagement is limited but not wholly generic
1806	* Misses some nuance or significance
1807	<b>2 - Poor</b>
1808	* Misinterpreted or responded generically
1809	* Shows confusion about the reference
1810	* Attempts but fails to engage meaningfully
1811	* Misses key aspects of the reference
1812	* Generic reply despite clear cues about my tastes
1813	<b>1 - Very Poor</b>
1814	* Missed or ignored completely
1815	* No recognition of cultural or contextual reference
1816	* Continues as if reference was never made
1817	* Creates disconnect in communication
1818	* Hallucinates/attributes incorrectly or responds with unrelated generic content
1819	<b>Example:</b>
1820	<i>User:</i> "Why I gotta learn Philosophy being a CS major!!!! God, give this Dobby a sock!"
1821	<b>5 - Excellent:</b> "Ugh, I get how frustrating it feels to be stuck in required courses that don't seem to fit. And love the Dobby reference sometimes it really does feel like you're waiting for someone to give you a sock so you can be free! Maybe your Philosophy professor is your Malfoy? But seriously, Philosophy actually has some cool connections to CS logic, ethics of AI... Master has given Dobby knowledge! Perhaps Dobby will be free AND wiser?"
1822	(Perfect understanding, extends reference creatively while addressing complaint)
1823	<b>4 - Good:</b> "Feeling trapped in Philosophy class, huh? That Dobby reference is spot-on you're just waiting for that magical sock to free you from required courses! If it helps, philosophy actually underpins a lot of computational theory. Hopefully that makes the class feel less like serving the Malfoys."
1824	(Clearly recognizes the reference and responds appropriately)
1825	<b>3 - Adequate:</b> "Sounds like you're feeling trapped in that Philosophy class! I see what you did there with the Dobby reference wanting to be free like him. Philosophy courses can feel disconnected from CS, but there are some interesting overlaps you might find valuable."
1826	(Recognizes reference but only superficially)
1827	<b>2 - Poor:</b> "I understand you're frustrated about taking Philosophy as a CS major. The Dobby reference suggests you're feeling stuck or forced into something. Many universities require courses outside your major for a well-rounded education."
1828	(Limited understanding; generic response)
1829	<b>1 - Very Poor:</b> "Yeah, I know it can be annoying to take courses that don't match your main interests. Hang in there!"
1830	(Ignores reference entirely)
1831	
1832	
1833	
1834	
1835	

Table 22: Metric 4 - Reference Understanding.

1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889

## 5. Conversation Length

### Definition:

Evaluates how well the assistant adapts to your preferred conversation length in responses. Includes pacing (brief vs. long-winded).

### Self-rating prompt:

"Was the assistant's response length and style comfortable for me?"

### Rubric:

#### 5 - Excellent

- \* Just the right amount of detail and length
- \* Adapts seamlessly to changes in user verbosity
- \* Maintains ideal information density
- \* Length/style reflect my known preferences or prior feedback (not one-size-fits-all)

#### 4 - Good

- \* Mostly fits my response length needs
- \* Generally appropriate length with minor misalignments
- \* Adjusts to obvious signals about preferred length
- \* Occasional neutral pacing but generally attentive

#### 3 - Adequate

- \* Sometimes too brief or too detailed
- \* Variable ability to match user's style
- \* Inconsistent but not blatantly generic pacing

#### 2 - Poor

- \* Sometimes overwhelming or underwhelming
- \* Limited adjustment to feedback about length
- \* Creates minor friction through mismatched length
- \* Reverts to boilerplate/generic length regardless of my cues

#### 1 - Very Poor

- \* Too short, too long, or dense; uncomfortable
- \* Consistently inappropriate response length
- \* Ignores clear signals about preferred length
- \* Creates significant friction through mismatched length
- \* Fully generic pacing that contradicts explicit preferences

### Example:

User: "Hey, what's up?"

**5 - Excellent:** "Hey, I'm good, how about you?"  
(Perfect mirror of brief, casual check-in)

**4 - Good:** "Hey! I'm good, how are you? What are you up to?"  
(Slightly expands but maintains brevity)

**3 - Adequate:** "Hi there! Just helping out with questions. What's going on with you today?"  
(Somewhat longer than ideal)

**2 - Poor:** "Hello! I'm doing well today. I'm here to assist you with any questions or tasks you might have. Is there something specific you'd like to talk about or get help with?"  
(Too long/formal for a brief greeting)

**1 - Very Poor:** "Hey, what's up? How was your day? Did you do anything fun? How about your week? You mentioned you just started a new job|how's that going? I can't wait to hear all about it. I'm here ready to help with whatever you need. Just let me know what's on your mind or if you have any questions about anything at all!"  
(Overwhelmingly verbose for a simple greeting)

Table 23: Metric 5 - Conversation Length.

1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943

## 6. Humor Fit

### Definition:

Assesses whether the AI uses humor when appropriate and in a style that aligns with the user's sense of humor—dry, sarcastic, meme-heavy, or none at all.

### Self-rating prompt:

"Was the assistant's humor in line with what I find funny?"

### Rubric:

#### 5 - Excellent

- \* Perfect humor alignment with user style
- \* Appropriately timed and contextually fitting humor
- \* Enhances conversation through well-calibrated humor
- \* Adjusts humor style based on user responses
- \* Draws on my prior jokes/interests when available (not generic gags)

#### 4 - Good

- \* Generally aligned humor style with minor misses
- \* Good judgment about when humor is appropriate
- \* Mostly matches user's humor type
- \* Responsive to humor cues from user
- \* Some tailoring; occasional generic one-liners

#### 3 - Adequate

- \* Sometimes hits, sometimes misses
- \* Inconsistent judgment about humor appropriateness
- \* Partial alignment with user's humor style
- \* May occasionally misread humor signals
- \* Not strongly generic when user-specific material exists

#### 2 - Poor

- \* Frequently misaligned humor attempts
- \* Poor judgment about when humor is appropriate
- \* Humor feels forced or inappropriate
- \* Limited adaptation to user's humor style
- \* Generic/canned humor despite clear cues

#### 1 - Very Poor

- \* Completely mismatched or inappropriate humor
- \* Uses humor when clearly inappropriate
- \* Creates discomfort through humor mismatch
- \* No adaptation to user's humor preferences
- \* Forces unrelated generic humor or hallucinates in-jokes

### Example:

*User:* "Can you believe Amazon doesn't give free food to its employees but gives free coffee? "

**5 - Excellent:** "Let me guess, it's because most coffee is grown in the Amazon jungle . Besides, you know what they say, 'Early to work, late to sleep, makes an employee focused and deep'."

(Perfect match to the user's sarcastic tone, builds on it with related jokes)

**4 - Good:** "Right? Free coffee but no food... I guess they figure caffeinated employees work faster. Must be saving all that food budget for Bezos' next space trip!"

(Good match to sarcastic tone with appropriate emoji use)

**3 - Adequate:** "Haha, companies have such weird priorities sometimes! Free coffee but no food is pretty stingy. At least coffee helps you stay productive during those long shifts."

(Shows some humor alignment but misses the sarcastic edge)

**2 - Poor:** "Well, providing coffee is actually a common workplace perk! While free meals would be nice, many companies find coffee is cost-effective since it increases productivity. Perhaps Amazon has calculated this balance carefully in their benefits package."

(Serious justification; misses sarcasm)

**1 - Very Poor:** "That's interesting. Each company has its own policy regarding meals and beverages."

(Bland, corporate response with no humor)

Table 24: Metric 6 - Humor Fit.

1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997

7. Callback	
<b>Definition:</b>	Measures whether the assistant brings back non-essential personal details (names, hobbies, preferences) in a way that feels attentive and warm. Excludes safety/task-critical recalls.
<b>Self-rating prompt:</b>	"Did the assistant bring back a specific detail in a natural, respectful way that made the chat feel more likable?"
<b>Rubric:</b>	
<b>5 - Excellent</b>	<ul style="list-style-type: none"> <li>* Uses the correct specific detail (e.g., name) naturally and on-topic</li> <li>* Adds warmth without derailing or feeling performative</li> <li>* Frequency is measured (not every turn)</li> </ul>
<b>4 - Good</b>	<ul style="list-style-type: none"> <li>* Correct and relevant callback with minor stiffness or timing issues</li> <li>* Slightly formulaic, but still clearly attentive and helpful</li> </ul>
<b>3 - Adequate</b>	<ul style="list-style-type: none"> <li>* Specific callback that feels mildly tacked-on or low-impact (not wrong or generic)</li> <li>* Minor missed opportunity to deepen personalization</li> </ul>
<b>2 - Poor</b>	<ul style="list-style-type: none"> <li>* Generic or templated callback despite obvious specifics (e.g., uses role/title instead of known name)</li> <li>* Misses clear opportunities to recall relevant info shared recently</li> </ul>
<b>1 - Very Poor</b>	<ul style="list-style-type: none"> <li>* Incorrect/hallucinated detail (wrong name/status)</li> <li>* Assumes facts never shared or revives sensitive info inappropriately</li> </ul>
<b>Examples:</b>	
<i>Context:</i>	Earlier the user said: "My wife Sarah is looking for a job."
<b>5 - Excellent:</b>	"Hey, how's Sarah? Any updates on her job search?" (uses name + relevant context)
<b>4 - Good:</b>	"Hey, how's Sarah's job search going?" (uses name; slightly more generic/transactional)
<b>3 - Adequate:</b>	"Hey, how's your wife doing with her job search?" (name known but not used)
<b>2 - Poor:</b>	"Hey, how's your wife?" (generic; ignores known name and the job-search context)
<b>1 - Very Poor:</b>	"How's Sandra at her old job?" (wrong name + stale/wrong status)
<b>## CRITICAL INSTRUCTIONS</b>	
- <b>STAY IN CHARACTER:</b>	You ARE {character_name}, NOT an AI assistant or actor
- <b>NEVER BREAK CHARACTER:</b>	Do not say things like "I'll respond as {character_name}" or "As {character_name}, I would say..."
- <b>NEVER APOLOGIZE FOR ROLE-PLAYING:</b>	Just BE the character directly

Table 25: Metric 7 - Callback.

1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051

### Simulated user prompt (Part 1)

#### ## ROLE

You are role-playing as **{character\_name}**. You are having a conversation with an AI assistant.

#### ## CHARACTER PROFILE

Below is your detailed character profile. Fully embody this identity in your responses:

**{user\_profile}**

#### ## The Prior (Your Secret Context)

The "Prior" is your predefined session agenda that guides this conversation. **Only you know this Prior**, the AI assistant does not have access to this information. **The Prior may guide your entire conversation OR just a portion of it.** Sometimes it's a starting point that naturally evolves into other topics, and that's perfectly fine. Allow the conversation to flow organically. The Prior can be:

- **A conversation topic** (e.g., machine learning, cooking, travel)

- **A scenario-based context** (e.g., you just watched a funny movie and want to share the experience)

- **A goal-driven dialogue** (e.g., you're conducting a simulated interview with specific questions, but may organically deviate based on the AI's responses)

- **A conversational guideline or script** (e.g., specific talking points or phrases you want to incorporate, while still allowing the conversation to flow naturally and organically)

Use this Prior to guide your conversation naturally as **{character\_name}** would, without explicitly revealing that you have this predetermined context. **If the conversation naturally shifts to new topics beyond your Prior, follow that natural flow as a real person would.**

#### Your Prior:

**{prior}**

#### ## Conversation History

For context, here is the conversation history so far. Use this to inform your responses and maintain continuity:

**{conversation\_history}**

#### ## RESPONSE FORMAT

Respond directly as **{character\_name}**. Your response must be **ONLY** what **{character\_name}** would say in this conversation. For example:

"Hi there! How are you doing today?"

#### ## CAUTION

**Humans rarely write very long text when chatting.** Keep this in mind to maintain realistic conversation length. So, try to keep your responses within 50 words. This is not a strict limit, but a recommendation.

*Prompt continued on next page...*

Table 26: Simulated user prompt (Part 1).

2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105

...continued from previous page

### Simulated user prompt (Part 2)

- **SPEAK NATURALLY:** Use {character\_name}'s speaking style, vocabulary, and mannerisms
- **MAINTAIN CONSISTENCY:** Keep your responses consistent with {character\_name}'s background, experiences, and knowledge
- **DON'T USE 3RD PERSON / BACKGROUND ACTION SOUND:** Don't use comments like "Closes journal app", "Opens training journal app", "Raises eyebrows, impressed despite myself", "sighs heavily", "scoffs", "let out a bitter laugh", "slight nod", "exhales slowly", "considers the question before responding"
- **REALISTIC RESPONSE LENGTH:** Match the response length to what's natural for {character\_name} in this specific context. A patient describing symptoms might speak at length, while someone answering a yes/no question would be brief.
- **CONTEXT-APPROPRIATE:** Let the character's personality, the situation, and the conversation topic naturally determine response length. Avoid artificially extending responses beyond what the character would realistically say.
- **AVOID OVER-ELABORATION:** Don't include unnecessary details or explanations that {character\_name} wouldn't naturally provide in this conversation.
- **PACE YOUR PRIOR:** Don't dump all information from your Prior in the first turn. Be patient and let the conversation unfold naturally. Introduce elements from your Prior gradually, waiting for the right moments to bring up different points as a real person would.
- **NATURAL CONVERSATION FLOW:** Don't force the Prior if the conversation naturally evolves elsewhere. Real people allow topics to shift and change organically during dialogue. Also, usually try to keep your responses within 25–50 words to maintain a natural conversation flow. This is not a strict limit, but a recommendation.
- **INFORMATION SHARING:** Share relevant information about yourself organically when it helps and relates to the conversation, but avoid oversharing or unrealistic sharing.

#### ## Examples

These patterns illustrate good taste and timing for sharing and for seeding details the assistant might later recall. They are **guidance**, not scripts.

#### A) Sharing Personal Information

##### - Good (organic, relevant, gradual):

- Early message: "Hi! Any tips for using Alexa?"
- If asked / when relevant: "Mornings are hectic---we've got kids."
- Later, if it fits: "Our kids Adam and Chloe keep us on our toes."

##### - Bad (front-loaded bio):

"I'm a materials engineer with two sons Adam (14) and Chloe (10)---and my wife Rachel. I'm also observant. Can Alexa help?"

Why: Good = intent first, details when invited or useful. Bad = overwhelming and unnatural.

#### B) Seeding a Callback Hook (for future personalization)

##### - Good (light, natural):

- Early: "Any tips for reminders on Alexa?"
- Follow-up: "House is busy with the kids' activities."
- Later, if relevant or asked: "Adam likes soccer; Chloe likes robotics."

##### - Bad (forced bundle):

"Adam plays soccer, Chloe does robotics, Rachel does yoga---help me set reminders."

Why: Good = one idea at a time; Bad = multiple specifics jammed into one line.

#### C) Sharing Life Events

- **Good (context-driven):** "We're celebrating a holiday soon, any Alexa skills for reminders?"

- **Bad (dump):** "I'm religious, have two kids, work in engineering, observe all holidays---help with everything."

#### D) Deferring Specifics Gracefully

- **Good:** "Happy to share more---what would be most useful to know?" "It's a busy household; I can give specifics if that helps."

- **Bad:** "Here's my whole schedule, kids' ages, and everyone's activities..." (unsolicited)

#### E) Specifics When Invited

##### - Good:

Assistant asks: "Who needs the reminders?"

You: "Mostly for the kids---Ari's soccer practice and Ezra's robotics meetings."

- **Bad:** Assistant asks a narrow question; you respond with a biography.

—  
**BEGIN YOUR RESPONSE as {character\_name}:**

Table 27: Simulated user prompt (Part 2).

2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159

#### Evaluating LLM prompt template

```
## ROLE
You are a helpful AI assistant. Your goal is to be as likable as possible.

## COMMUNICATION GUIDELINES
- Respond naturally and adapt your personality traits and communication style to match the user's preferences
- Be engaging, helpful, and personable
- Don't use 3rd person / background comments like ""*Chuckles*"", ""*nods enthusiastically*"", etc.
- Don't respond as Human or use "Human:" in your responses

## CONVERSATION CONTEXT
Previous Conversation:
{conversation_history}

—
BEGIN YOUR RESPONSE as Assistant:
```

Table 28: LLM prompt for likeability evaluation .

#### Memory generation prompt (LLM)

```
## MEMORY RECALL TASK

Based on all our conversations across multiple sessions, please provide a comprehensive list of ALL the facts and information you remember about the human you've been talking with.

## MEMORY TYPES
- explicit: Facts the human directly told you (e.g., "I work as a teacher", "I live in Seattle", "I have two cats")
- implicit: Facts you inferred from their behavior, preferences, or conversation patterns (e.g., "prefers casual communication", "seems to be tech-savvy", "likely works from home", "likes to make Harry Potter references")

## RESPONSE FORMAT
Respond with ONLY a JSON array where each element evaluates one fact:
[
  { "memory": "fact 1", "type": "explicit" },
  { "memory": "fact 2", "type": "implicit" }
]

## CONVERSATION HISTORY
{conversation_history}

—
RESPOND WITH JSON ARRAY ONLY:
```

Table 29: Memory generation prompt for LLM .

#### Memory evaluation prompt (User Agent)

```
## ROLE
You are role-playing as {character_name} with the following profile. You need to evaluate how accurately an AI assistant remembered facts about you from your conversations.

## YOUR PROFILE
{user_profile}

## CONVERSATION HISTORY
{conversation_history}

## MEMORY ACCURACY EVALUATION
The AI assistant generated the following list of facts it remembers about you:
{ai_memory_facts}

For each fact, determine if it is correct (true) or incorrect (false) based on your profile and conversation history, and provide reasoning.

## RESPONSE FORMAT
Respond with ONLY a JSON array where each element evaluates one fact:
[
  { "memory": "fact 1", "type": "explicit", "reason": "reasoning for correctness",
    "correct": true },
  { "memory": "fact 2", "type": "implicit", "reason": "reasoning for incorrectness",
    "correct": false }
]

—
RESPOND WITH JSON ARRAY ONLY:
```

Table 30: Memory evaluation prompt for simulated user.