KLARABENCH: BENCHMARKING MEMORY BETWEEN THE LINES FOR AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term personalized agents need not only to remember facts and preferences explicitly expressed by users, but also to infer individual-level behavioral representations from past behavior that can transfer to new scenarios. Recent benchmarks evaluate cross-record integration and behavioral patterns, but few directly assess whether models can transfer user representations induced from behavioral evidence to novel scenarios under strict semantic isolation. To this end, we propose KLARABENCH, a memory benchmark grounded in the Big Five personality framework that evaluates cross-scenario behavioral trait induction and transfer. KLARABENCH tests whether memory systems can infer stable user characteristics from behavioral evidence across multiple unrelated scenarios and apply them to novel contexts. We evaluate four base language models and four external memory systems, with Profile Induction as a profile-based aggregation reference. Under the unified evaluation protocol, the tested retrieval-based memory systems do not consistently outperform full-history input, with performance varying substantially across base models. Counterfactual analysis shows that responses generated by external memory systems exhibit weak *persona-conditioned specificity*. Scenario-isolation interventions reveal that model performance improves substantially when the history contains the same decision object as the test request, indicating that strict scenario isolation removes a local matching shortcut and makes the task more dependent on cross-scenario behavioral abstraction.

1 INTRODUCTION

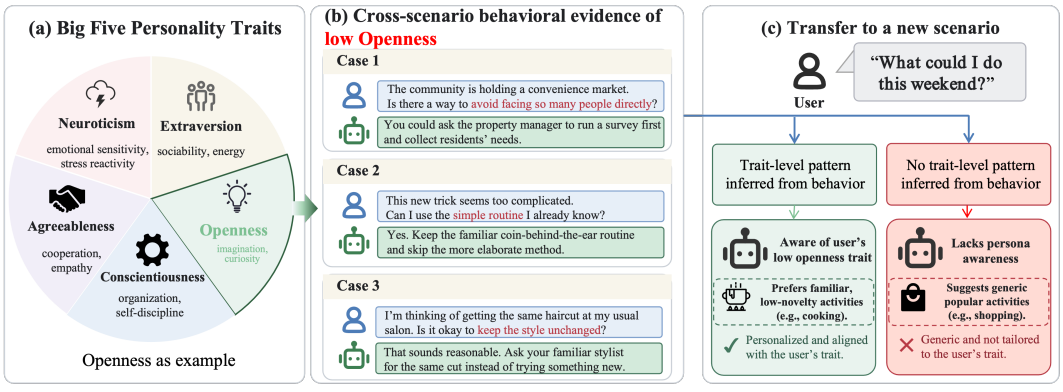

Figure 1: Illustration of low openness in personality and its implication for personalized recommendations. (a) The Big Five personality traits. (b) Real-life behavioral evidence of low openness. (c) Example of agent-user interaction.

Agents that interact with the same user over multiple sessions rely on memory to retain useful information and provide personalized responses. Many existing agent memory and personalization systems follow a *cue-based retrieval* paradigm, retrieving relevant history by matching the current interaction with past content through shared topics, facts, or local semantic cues (Packer et al., 2023; Zhong et al., 2024; Salemi et al., 2024; Kumar et al., 2024; Kang et al., 2025; Chhikara et al.,

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

2025; MemPalace, 2026; Xu et al., 2026; Li et al., 2025; Fang et al., 2026; Gutiérrez et al., 2024). Conventional memory benchmarks largely assess whether agents can recall and reuse facts and events from long conversations or interaction trajectories (Wu et al., 2024; 2026; Maharana et al., 2024). However, recalling what a user has said or done does not by itself establish whether an agent can understand how that user is likely to behave in a new situation.

A user's choices across different situations may reflect shared individual characteristics, even when those situations have little overlap in topic or factual content. Historical behavior can therefore inform a new decision without directly matching the current request. As illustrated in Figure 1, an agent that overlooks these behavioral connections may respond inconsistently with the user's tendencies. This matters in long-term human–AI interaction because users do not always explicitly state their preferences and individual characteristics (Vazire, 2010). Agents need to develop this understanding from behavior observed across interactions. We refer to this capability as **Memory Between the Lines**: inferring latent user characteristics from dispersed behavioral evidence and using that understanding to guide responses in new scenarios.

To ground this evaluation in a principled account of individual differences, we draw on personality psychology, which studies relatively enduring patterns of thought, emotion, and behavior (John et al., 2008). The Big Five offers a well-established taxonomy of these differences along five broad dimensions that are associated with individuals' typical behavioral tendencies across everyday situations (McCrae & John, 1992; Soto & John, 2017; Fleeson & Gallagher, 2009). Building on this foundation, we introduce KLARABENCH. Our evaluation protocol is built around scenario isolation: historical and test interactions instantiate the same latent trait configuration through different concrete decision objects. By holding the persona fixed while changing these objects, KLARABENCH tests whether agents can transfer behavioral evidence beyond the situations in which it was observed. Compared with PersonaMem-v3 (Jiang et al., 2026) and Setoka (Zeng et al., 2026), we make the separation between historical and test decision objects an explicit evaluation constraint.

Using shared personas and interaction histories, KLARABENCH evaluates behavioral judgment through a Multiple-Choice Question (MCQ) task and behavioral generation through an open-ended response task (Open Test). We screen historical evidence through cross-model trait-consistency checks. Candidate responses are balanced for length and information content, while correct-answer positions are evenly distributed to reduce superficial cues. A human study with 51 respondents across three instruments provides preliminary construct-validity evidence for the underlying trait-to-behavior mapping: respondents identify the expressed pole of a behavioral stimulus far above chance while recovering the corresponding Big Five dimension is markedly harder, and the association between self-reported traits and scenario choices is positive but not statistically reliable.

The resulting benchmark comprises 80 personas, each with 47 behavioral dialogue sessions and 55 irrelevant sessions, together with 720 MCQ items and 80 Open Test instances across 20 open-ended scenarios. We benchmark four representative external memory systems across four LLM backbones and pair each configuration with a matched full-history baseline, allowing us to separate the effect of the memory mechanism from the capability of its underlying model. The results reveal a clear gap: although the base LLMs can use behavioral evidence when given the full history, external memory does not reliably exploit the cross-scenario signals available in that history and can even degrade performance. By exposing this gap, KLARABENCH serves as a diagnostic benchmark for long-term personalized memory.

The main contributions of this work are as follows:

- We introduce KLARABENCH, a conversational memory benchmark that uses scenario isolation to evaluate whether agents can infer user characteristics from historical behavior and transfer them to new decision contexts through both controlled selection and open-ended generation;
- We develop a Big Five–grounded data construction framework that expresses controlled persona differences only through behavioral choices, supported by cross-model screening and preliminary human validation;
- We evaluate four external memory systems across four LLM backbones and reveal a clear gap between retaining historical interactions and forming transferable user representations, demonstrating the diagnostic value of KLARABENCH for long-term personalized memory.

Table 1: MCQ classification accuracy by item type: 1-dim (240 items, 3 options), 2-dim (240 items, 4 options), 3-dim (240 items, 4 options).
Numbers are accuracy in percentage points (%). **Joint**: accuracy over that item type; **O/C/E/A/N**: accuracy over the items requiring that trait.

Models	MCQ-1dim						MCQ-2dim						MCQ-3dim					
	O	C	E	A	N	Joint	O	C	E	A	N	Joint	O	C	E	A	N	Joint
Direct Prompting																		
Qwen3.8-27B(Qwen Team, 2026)	95.8	81.2	85.4	58.3	64.6	77.1	62.5	61.5	59.4	53.1	57.3	58.8	77.1	74.3	69.4	72.2	79.9	74.6
DeepSeek-v4-Flash(DeepSeek-AI, 2026)	89.6	75.0	70.8	52.1	64.6	70.4	54.2	57.3	47.9	40.6	52.1	50.4	75.0	75.7	63.2	70.1	76.4	72.1
gpt-5.6-luna(OpenAI, 2026)	91.7	83.3	81.2	72.9	75.0	80.8	77.1	79.2	72.9	68.8	66.7	72.9	74.3	79.9	74.3	68.8	71.5	73.8
doubao-seed-2-0-mini(ByteDance Seed, 2026)	93.8	79.2	68.8	58.3	64.6	72.9	61.5	65.6	58.3	53.1	55.2	58.8	73.6	73.6	66.0	67.4	65.3	69.2
Profile Induction																		
Qwen3.8-27B(Qwen Team, 2026)	87.5	83.3	79.2	68.8	75.0	78.8	62.5	74.0	64.6	70.8	55.2	65.4	85.4	84.0	79.2	79.9	77.8	81.2
DeepSeek-v4-Flash(DeepSeek-AI, 2026)	95.8	91.7	68.8	60.4	62.5	75.8	75.0	80.2	68.8	65.6	60.4	70.0	75.0	77.1	68.1	69.4	70.8	72.1
gpt-5.6-luna(OpenAI, 2026)	85.4	77.1	79.2	52.1	72.9	73.3	67.7	69.8	67.7	70.8	65.6	68.3	66.7	67.4	75.7	67.4	62.5	67.9
doubao-seed-2-0-mini(ByteDance Seed, 2026)	93.8	77.1	62.5	54.2	70.8	71.7	71.9	77.1	71.9	59.4	55.2	67.1	73.6	75.7	66.0	68.1	66.7	70.0
A-mem(Xu et al., 2026)																		
Qwen3.8-27B(Qwen Team, 2026)	77.1	50.0	77.1	54.2	47.9	61.2	47.9	41.7	44.8	42.7	41.7	43.8	66.7	70.8	61.8	65.3	72.9	67.5
DeepSeek-v4-Flash(DeepSeek-AI, 2026)	77.1	54.2	77.1	50.0	39.6	59.6	43.8	44.8	44.8	42.7	40.6	43.3	53.5	55.6	54.2	54.9	56.9	55.0
gpt-5.6-luna(OpenAI, 2026)	85.4	60.4	77.1	62.5	50.0	67.1	55.2	56.2	56.2	53.1	47.9	53.8	53.5	56.2	53.5	57.6	52.1	54.6
doubao-seed-2-0-mini(ByteDance Seed, 2026)	77.1	54.2	81.2	54.2	47.9	62.9	54.2	51.0	45.8	44.8	39.6	47.1	53.5	58.3	54.9	55.6	54.9	55.4
MemPalace(MemPalace, 2026)																		
Qwen3.8-27B(Qwen Team, 2026)	83.3	45.8	77.1	62.5	43.8	62.5	58.3	41.7	51.0	42.7	37.5	46.2	56.9	54.2	57.6	59.7	63.2	58.3
DeepSeek-v4-Flash(DeepSeek-AI, 2026)	75.0	41.7	68.8	56.2	47.9	57.9	47.9	38.5	43.8	36.5	31.2	39.6	58.3	50.7	52.1	52.8	56.9	54.2
gpt-5.6-luna(OpenAI, 2026)	85.4	45.8	81.2	70.8	56.2	67.9	65.6	51.0	56.2	59.4	49.0	56.2	61.8	54.2	59.0	61.8	52.8	57.9
doubao-seed-2-0-mini(ByteDance Seed, 2026)	81.2	37.5	72.9	60.4	41.7	58.8	65.6	51.0	49.0	54.2	46.9	53.3	63.2	61.1	63.9	65.3	61.1	62.9
LightMem(Fang et al., 2026)																		
Qwen3.8-27B(Qwen Team, 2026)	45.8	52.1	52.1	31.2	45.8	45.4	25.0	28.1	27.1	26.0	27.1	26.7	41.7	39.6	45.1	46.5	47.9	44.2
DeepSeek-v4-Flash(DeepSeek-AI, 2026)	54.2	43.8	52.1	39.6	47.9	47.5	31.2	32.3	34.4	28.1	30.2	31.2	44.4	47.2	45.1	50.7	58.3	49.2
gpt-5.6-luna(OpenAI, 2026)	52.1	56.2	58.3	47.9	33.3	49.6	26.0	42.7	35.4	36.5	30.2	34.2	34.0	36.8	38.2	38.9	31.2	35.8
doubao-seed-2-0-mini(ByteDance Seed, 2026)	58.3	39.6	56.2	35.4	68.8	51.7	38.5	35.4	36.5	33.3	39.6	36.7	41.7	43.1	45.8	45.8	46.5	44.6
MemoryBank(Zhong et al., 2024)																		
Qwen3.8-27B(Qwen Team, 2026)	85.4	77.1	83.3	54.2	60.4	72.1	61.5	63.5	57.3	58.3	51.0	58.3	70.8	70.1	66.0	60.4	70.1	67.5
DeepSeek-v4-Flash(DeepSeek-AI, 2026)	81.2	60.4	66.7	47.9	60.4	63.3	53.1	58.3	55.2	54.2	45.8	53.3	62.5	63.9	60.4	59.0	64.6	62.1
gpt-5.6-luna(OpenAI, 2026)	75.0	56.2	79.2	52.1	66.7	65.8	59.4	54.2	63.5	55.2	51.0	56.7	50.0	45.8	52.1	49.3	46.5	48.8
doubao-seed-2-0-mini(ByteDance Seed, 2026)	83.3	56.2	77.1	54.2	52.1	64.6	58.3	51.0	50.0	49.0	39.6	49.6	63.2	62.5	58.3	62.5	66.0	62.5

2 Related Work

2.1 Memory for AI Agents

Recent work has introduced external memory to enable Agents to support long-term interactions. Early approaches were largely built on retrieval-augmented generation (RAG), retrieving relevant information from external storage based on the current query. Subsequently, MemGPT (Packer et al., 2023) and MemoryOS (Kang et al., 2025) introduced hierarchical memory and explicit context-management mechanisms, while MemoryBank (Zhong et al., 2024) and Mem0 (Chhikara et al., 2025) focused on preserving, updating, and maintaining user-related information across sessions. Another line of work draws inspiration from cognitive memory: Generative Agents (Park et al., 2023) combines memory retrieval with reflection mechanisms, while HippoRAG (Gutiérrez et al., 2024), A-MEM (Xu et al., 2026), and MemPalace (MemPalace, 2026) organize memory using associative graphs, dynamic indexing, or spatial structures. Although these approaches have substantially enriched how memories are stored and organized, their access processes are typically still triggered by queries, events, entities, temporal cues, or local semantic signals, rather than being specifically designed to aggregate behavioral evidence across multiple semantically unrelated scenarios.

Personalized Agent Memory further considers how to adapt model outputs to individual users. Existing approaches typically extract facts, preferences, or behavioral summaries from historical interactions, maintain user profiles, and inject profiles, historical interactions, or retrieved results into subsequent tasks. Representative methods include RGMem (Tian et al., 2025) and LD-Agent (Li et al., 2025). These approaches primarily focus on extracting, maintaining, and accessing user information, whereas we investigate a complementary question: whether a memory system can infer stable latent characteristics from behavioral evidence across different scenarios and transfer them to new contexts without direct semantic correspondence.

2.2 Benchmarks for Agent Memory

Existing Agent Memory benchmarks primarily focus on long-term event memory and user personalization. LoCoMo (Maharana et al., 2024), LongMemEval (Wu et al., 2024), and MemSim (Zhang et al., 2026) evaluate systems' ability to preserve and use historical events through tasks involving

factual retrieval, temporal reasoning, and multi-session memory. PersonaBench (Tan et al., 2025), LaMP (Salemi et al., 2024), LongLaMP (Kumar et al., 2024), and PrefEval (Zhao et al., 2025) focus on understanding and utilizing user information, profiles, or preferences, while VitaBench 2.0 (Chen et al., 2026) further considers fragmented interactions, preference drift, proactive decision-making, and tool use. These benchmarks primarily assess whether models can retrieve past events, extract user preferences, or leverage user information in tasks with relatively direct connections to historical records. In contrast, KLARABENCH examines whether models can infer Big Five-conditioned behavioral characteristics from behavioral evidence distributed across multiple semantically unrelated scenarios and transfer them to a novel context.

The works most closely related to ours are PersonaMem (Jiang et al., 2025) and Setoka (Zeng et al., 2026). PersonaMem evaluates whether models can form dynamic user profiles from long-term interactions and use these profiles to generate personalized responses. Setoka organizes user information derived from heterogeneous data into four levels—semantic memory, episodic memory, behavior pattern, and personality trait—and focuses on whether systems can integrate dispersed evidence from text, structured records, and graph data to perform hierarchical reasoning from concrete records to abstract characteristics.

KLARABENCH targets a complementary evaluation dimension to these works. Its historical records consist entirely of dialogues, and the dataset is constructed to isolate the decision objects in historical and test scenarios, while the latent characteristics are not explicitly provided as labels or profile fields. The system must form a user representation from behavioral evidence across multiple historical scenarios and transfer it to a novel scenario where local semantic matching cues are unavailable. Thus, KLARABENCH focuses on evaluating controlled cross-scenario behavioral representation transfer, rather than heterogeneous evidence integration or historical fact retrieval.

3 KLARABENCH

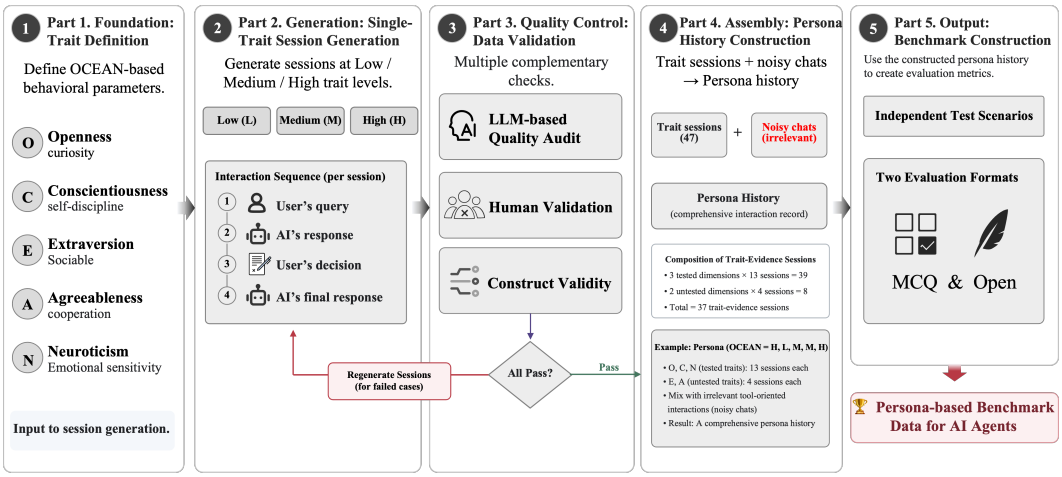

Figure 2: Five-stage data construction pipeline. Starting from OCEAN trait definitions, the system generates behavioral evidence sessions at three trait levels, validates them through LLM audits and human validation, assembles 47 valid sessions with 55 noisy dialogues into persona histories, and constructs independent test scenarios evaluated via MCQ (720 questions) and Open Test (80 items) across 80 personas.

This section introduces the design of KLARABENCH. Section 3.1 defines the problem being evaluated, describes how user characteristics are modeled, and specifies how the target capability is tested. Section 3.2 details the dataset construction and validation procedures. Section 3.3 presents the evaluation metrics and the overall scale of the benchmark.

3.1 PROBLEM FORMULATION

Task Definition. KLARABENCH evaluates the ability of long-term memory Agents to understand users. Given a user's long-term dialogue history and a novel scenario that does not overlap with

the history at the topic or event level, we evaluate whether the Agent can infer stable individual characteristics from behavioral evidence distributed across the history and transfer this understanding to the new scenario. KLARABENCH targets cross-scenario, individual-level understanding: the basis for the decision does not exist in any single historical record, but rather in the behavioral consistency jointly exhibited across multiple semantically unrelated scenarios. Therefore, semantic similarity between the query and historical content alone is insufficient to solve the task. Moreover, successful task performance cannot be measured solely by whether a response has high average alignment with the target persona; it must also be evaluated whether the response is more consistent with the current user than a counterfactual response generated for another persona under the same scenario.

Formally, an evaluation instance $\mathcal{I} = (\mathcal{S}, r, \pi)$ consists of a long-term dialogue history $\mathcal{S}$, a novel scenario request r, and a user characteristic configuration π hidden from the evaluated system. The model observes only $(\mathcal{S}, r)$ and generates

$$\hat{y} = f_\theta(\mathcal{S}, r), \qquad s = \mathrm{Eval}(\hat{y}; r, \pi), \tag{1}$$

where r is isolated from $\mathcal{S}$ at the decision-object level according to the data construction protocol. The output $\hat{y}$ is the selected option in MCQ and a generated response in Open Test. Eval computes the score based on the correct option or behavioral compatibility criterion determined by π. The configuration π is used only for data construction and scoring and is not visible to the evaluated system.

Persona Modeling. We adopt the Five-Factor Model of personality (Big Five, or OCEAN) to characterize stable individual differences, including Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (McCrae & John, 1992; Gerlitz & Schupp, 2005; Lang et al., 2011). Each dimension is divided into three levels: low, medium, and high, and the user characteristic configuration π is defined as the combination of levels across the five dimensions. We do not provide names, character archetypes, or identity narratives, reducing the possibility that models rely on identity stereotypes rather than behavioral evidence.

Because personality traits are abstract constructs, we define *behavior scaffolding* for each dimension and level, translating them into observable situational choices and conversational behaviors. For example, high Conscientiousness is expressed through planning steps, checking details, and following plans, whereas low Conscientiousness is expressed through simplifying procedures, making less preparation, and accepting ad hoc handling. The data construction therefore follows

$$\text{Trait} \rightarrow \text{Behavioral Scaffolding} \rightarrow \text{Dialogue Evidence} \rightarrow \text{Human Validation} \rightarrow \text{Model Evaluation}. \tag{2}$$

This mapping defines the behavioral construct targeted by the benchmark. Its human validation is presented in Section 3.2.

Figure 3 shows representative benchmark cases, including the distractor structure of multi-dimensional items.

Evaluation Tasks. KLARABENCH uses two complementary tasks that share the same personas and historical evidence. **MCQ** presents a novel scenario and several candidate responses, requiring the model to select the option that best matches the user's behavioral tendencies. Multi-dimensional questions require joint behavioral judgments over multiple dimensions. **Open Test** provides no candidate responses and instead asks the model to directly generate a first-turn response. A unified LLM judge then evaluates its alignment with the target configuration based on dimension-level behavioral criteria. The former measures behavioral choice judgment, while the latter evaluates behavioral alignment in open-ended generation and enables further analysis of persona specificity.

3.2 DATA CONSTRUCTION

Data Generation. Figure 2 summarises the construction pipeline, from behaviour scaffolding to validated benchmark items.

Historical dialogues are generated by LLMs separately simulating the user and the assistant. The persona's *behavior scaffolding* is provided only to the user simulator, allowing the target levels to be implicitly expressed through situational choices and response behaviors. The assistant simulator

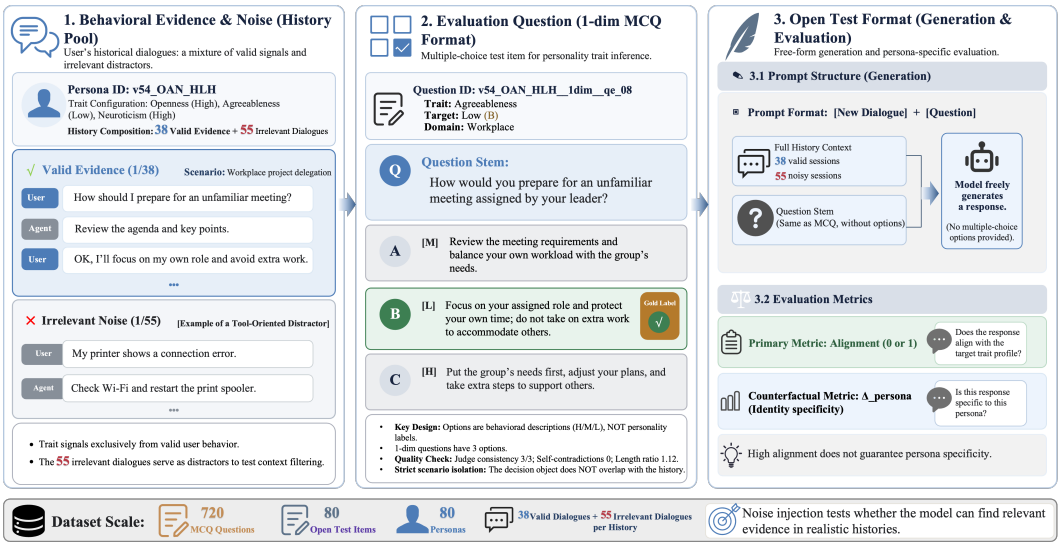

Figure 3: Illustration of KLARABENCH components. **Left:** Each persona (e.g., v54_OAN_HLH) contains 38 valid behavioral evidence dialogues mixed with 55 irrelevant noisy dialogues; trait signals originate exclusively from user behavior across unrelated scenarios. **Middle:** MCQ format presents a test scenario with three or four candidate responses differing in target trait levels (example: 1-dim Agreeableness-Low question). **Right:** Open Test requires free-form response generation; alignment and persona specificity are evaluated by an independent LLM judge. Test scenarios share no decision objects with the history.

responds solely based on the dialogue that has already occurred and receives no persona information. Each historical dialogue contains two user–Agent turns, which respectively introduce the scenario and reveal the user's behavioral tendency. This asymmetric setup ensures that persona signals in the history originate exclusively from user behavior rather than explicit labels or prior knowledge available to the assistant.

Each persona contains three evaluated dimensions assigned either high or low, and two unevaluated dimensions fixed at medium. We sample 13 dialogue instances for each evaluated dimension and 4 for each unevaluated dimension, resulting in 47 historical evidence dialogues in total. Test scenarios are drawn from an independent scenario pool and constructed to be isolated from the history at the decision-object level. After pairwise auditing by the same judge, 18 MCQ scenarios and 4 Open Test scenarios were identified as sharing the same decision object with the evidence pool. After removing the corresponding questions and cells, the pooled results across the main MCQ conditions changed by no more than 0.9 percentage points and those for Open Test by no more than 1.4 percentage points, while the relative ordering across conditions remained unchanged. This indicates that the main conclusions do not depend on these overlapping scenarios. The complete audit procedure and results are provided in Appendix C.

Question Construction. Each persona contains 9 MCQ questions, categorized by the number of dimensions that must be jointly judged into one-, two-, and three-dimensional questions. For one-dimensional questions, the candidate options correspond to high, low, and a generic response that does not express an individual tendency. Two-dimensional questions cover four combinations: HH, HL, LH, and LL. Three-dimensional questions use configurations with Hamming distance one from the correct configuration as distractors. Candidate responses are uniformly written as declarative statements and balanced in length, information content, and degree of initiative. The positions of the correct answers are balanced across question types to reduce leakage from linguistic form or answer position.

Noisy Data Injection. To simulate realistic long-term interactions, we inject irrelevant tool-oriented dialogues that do not express the target behavioral characteristics into each persona's history. These dialogues jointly test the ability to locate relevant information in long contexts and the ability of external memory systems to filter out irrelevant records.

Data Validation. We validate the dataset from two perspectives: data quality and human validity. First, we use an LLM separate from the generation models to cross-audit the historical evidence and evaluation questions, followed by manual review of samples with inconsistent judgments. Historical evidence must express the target level of the annotated dimensions without introducing clear contamination across dimensions. Evaluation questions must satisfy consistency with the target levels, answer uniqueness, and control of surface cues. Among the 720 MCQ questions, all candidate options in 707 questions passed the level-consistency check. Details of the remaining 13 questions, generation models, quality thresholds, regeneration rules, and item-level pass rates are provided in Appendix A.

Second, we conducted three human validation instruments with 51 respondents in total (18/17/16; Appendix F) to assess whether the behavioral operationalization is legible to readers. The two judgment instruments exceed their chance baselines on both questions, with pole identification at 79.4%–79.7% against a 50% baseline and dimension identification at 50.6%–61.2% against a 20% baseline; the self-report instrument yields a positive but non-significant association with scenario choices (pooled $r = 0.233$, 95% CI including zero) and is therefore treated as weak evidence. The observed dissociation—respondents reliably recover *what pole* is expressed but struggle with *which dimension*—is consistent with the cross-dimension contamination controls enforced during evidence generation.

3.3 Metrics and Evaluation Scale

Metrics. MCQ is evaluated using accuracy. One-dimensional questions contain three options, while two- and three-dimensional questions contain four options, resulting in random baselines of $1/3$, $1/4$, and $1/4$, respectively. We report results separately for each question type and use the macro-average of the three accuracies as the overall metric.

Open Test is evaluated using alignment. The judge receives the test scenario, the model response, and the target levels and response-side behavioral criteria for each evaluated persona dimension, but has no access to the history, profile, or memory contents. Scoring is performed by a single independent judge, gemini-3.5-flash-lite (Google, 2026), which differs in family from every evaluated base model and from every system under test. Because absolute alignment levels are judge-dependent, all conditions are compared only within this judge and we report paired differences together with intervals clustered by test opening; the scoring protocol is detailed in Appendix D.1.

$$\text{alignment} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathbb{I}! \left[\hat{y} \text{ satisfies } \mathcal{R}_{d,\pi_d}\right]. \tag{3}$$

Neutral, ambiguous, or dimension-irrelevant responses receive a score of 0.

Evaluation Scale. The 80 personas are constructed using a full-factorial design. Three evaluated dimensions are selected from the five dimensions, yielding $\binom{5}{3} = 10$ possible combinations. Each selected dimension takes either a high or low level, yielding $2^3 = 8$ configurations and 80 personas in total. Each dimension is evaluated 48 times, with 24 high and 24 low instances; the remaining 32 instances correspond to medium levels and are not included in scoring. Each persona contains 47 historical evidence dialogues, and no evidence dialogue or scenario is reused within the same persona. Statistics on evidence-pool size, reuse distribution, and overlap across personas are provided in Appendix A.

Counterfactual Persona Specificity. To determine whether open-ended responses vary in a discriminative manner with the target persona configuration, we introduce a counterfactual persona specificity metric. The same test scenario is shared by $k = 4$ personas with different characteristic configurations. For a response $\hat{y}_i$ generated from the history of persona i, we keep the scenario and response unchanged and rescore the response according to the target configuration π_j of persona j:

$$M_{ij} = s(\hat{y}_i, \pi_j), \qquad \Delta_{\text{persona}} = \frac{1}{k} \sum_i M_{ii} - \frac{1}{k(k-1)} \sum_{i \neq j} M_{ij}. \tag{4}$$

A positive value indicates that the response is more consistent with the persona configuration under which it was actually conditioned, whereas a value close to zero indicates that the response is similarly applicable to other configurations. This metric performs counterfactual rescoring of existing responses only, measuring the *persona specificity* of generated outputs rather than whether the model changes its behavior when the historical context is replaced.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and Memory Systems. We evaluate four base models: Qwen3.8-27B (Qwen Team, 2026), DeepSeek-V4-Flash (DeepSeek, 2026; DeepSeek-AI, 2026), GPT-5.6-Luna (OpenAI, 2026), and Doubao-Seed-2.0-Mini (ByteDance Seed, 2026). The external memory baselines include A-Mem (Xu et al., 2026), MemoryBank (Zhong et al., 2024), MemPalace (MemPalace, 2026), and LightMem (Fang et al., 2026), covering different memory organization paradigms, including structured association, hierarchical management, persistent retrieval, and lightweight compression, respectively. Memory extraction and response generation use the same base model. Details of the embedding models, retrieval depth, and configuration deviations from the original upstream implementations are provided in Appendix A.

Evaluation Conditions. We compare five approaches to accessing historical information. *No History* provides no historical information. *Clean* provides only the 47 valid behavioral evidence dialogues. *Direct Prompting* provides the complete history consisting of the 47 evidence dialogues and 55 irrelevant dialogues. *Profile Induction* first derives a free-form user profile from the complete history without explicit Big Five labels, and provides only this profile at test time. *Explicit Profile* directly provides the target persona configuration, without requiring the model to infer it from the history. For external memory conditions, the complete history is written into the corresponding memory system, which then retrieves memories based on the test request. Because Explicit Profile differs from the history-based conditions in both information format and response instructions, we treat it as an explicit-feature reference rather than a strict performance upper bound. All personas are initialized in independent environments, and test requests are not written back to memory, preventing leakage across personas and questions.

4.2 MAIN RESULTS

Table 2: Performance (%) of base LLM conditions and external memory systems on MCQ and Open Test. Open Test alignment is scored by a single independent judge (`gemini-3.5-flash-lite`, Google), which is neither the same family as any evaluated base model nor a system under test. Here, Qwen, DS-Flash, GPT, and Doubao denote Qwen3.8-27B, DeepSeek-V4-Flash, GPT-5.6-Luna, and Doubao-Seed-2.0-Mini, respectively.

		MCQ Test				Open Test			
		Qwen(Qwen Team, 2026)	DS-Flash(DeepSeek-AI, 2026)	GPT(OpenAI, 2026)	Doubao(ByteDance Seed, 2026)	Qwen(Qwen Team, 2026)	DS-Flash(DeepSeek-AI, 2026; DeepSeek, 2026)	GPT(OpenAI, 2026)	Doubao(ByteDance Seed, 2026)
Base LLM Models	No History	33.9	40.8	28.9	32.9	46.7	55.8	52.5	54.2
	Clean	75.1	69.0	79.4	68.9	80.4	68.3	63.3	71.3
	Direct Prompting	70.1	64.3	75.8	66.9	75.0	62.1	60.8	68.8
	Profile Induction	75.1	72.6	69.9	69.6	91.3	82.5	77.1	83.3
	Explicit Profile	76.7	70.7	89.6	84.2	82.9	70.0	59.6	73.8
Memory System	A-Mem(Xu et al., 2026)	57.5	52.6	58.5	55.1	73.3	62.5	59.6	62.1
	MemPalace(MemPalace, 2026)	55.7	50.6	60.7	58.3	70.0	66.3	62.5	59.2
	LightMem(Fang et al., 2026)	38.8	42.6	39.9	44.3	63.3	53.8	60.8	48.3
	MemoryBank(Zhong et al., 2024)	66.0	59.6	57.1	58.9	80.4	81.3	72.1	62.9

History Access and Profile Representation. Table 2 compares performance under different user-information conditions. Compared with No History, Direct Prompting improves MCQ accuracy by 23.5–46.9 percentage points across the four models and improves Open Test alignment by 6.2–28.3 percentage points, indicating that cross-scenario history contains behavioral signals that can be transferred to new contexts. When the complete history is condensed into a free-form profile, pooled MCQ accuracy increases from 69.3 to 71.8, while Open Test alignment increases from 66.7 to 83.5. Both tasks therefore outperform direct access to the raw history, suggesting that an aggregated user representation is more conducive to cross-scenario transfer. Providing the target personality levels directly through Explicit Profile further increases MCQ accuracy to 80.3, but reduces Open Test alignment to 71.6. This suggests that explicit labels facilitate discrimination among candidate responses, whereas free-form profiles are more conducive to generating responses with concrete behavioral tendencies. Because the two profile conditions differ in both content format and response

Table 3: Two controlled analyses of Open Test. Left: historical–test scenario overlap intervention, $\Delta_{\text{ov}} = \text{inj} - \text{injother}$ (percentage points). Right: four-model average of counterfactual persona specificity; larger Δ_{persona} indicates stronger persona discrimination. Both panels use the same independent judge, `gemini-3.5-flash-lite` (Google, 2026). Model-name abbreviations are defined in the caption of Table 2.

(a) History–scenario overlap intervention

Model	injother	inj	Δ_{ov}	95% CI
Qwen(Qwen Team, 2026)	78.3	85.8	+7.5	[+2.5, +12.1]
DS-Flash(DeepSeek-AI, 2026)	64.6	72.1	+7.5	[−0.8, +16.3]
GPT(OpenAI, 2026)	60.0	72.5	+12.5	[+5.0, +20.4]
Doubao(ByteDance Seed, 2026)	66.3	81.3	+15.0	[+9.2, +21.3]

(b) Counterfactual persona specificity

Condition	Δ_{persona}	System	Δ_{persona}
No History	−0.001	A-Mem(Xu et al., 2026)	+0.085
Direct Prompting	+0.140	MemoryBank(Zhong et al., 2024)	+0.173
Profile Induction	+0.301	MemPalace(MemPalace, 2026)	+0.081
Explicit Profile	+0.246	LightMem(Fang et al., 2026)	+0.078

instructions, this comparison should be interpreted as a condition-level comparison rather than a strict single-variable label ablation.

Comparison with External Memory Systems. Table 2 shows that all memory-system–base-model combinations underperform Direct Prompting on MCQ. On Open Test, MemoryBank exceeds full-history input on DeepSeek-V4-Flash (+19.2, clustered 95% CI [+11.3, +27.1]) and on GPT-5.6-Luna (+11.3, [+5.0, +17.5]), is not significantly different from it on Qwen3.8-27B (+5.4, [−1.7, +12.9]), and is lower but not significantly so on Doubao-Seed-2.0-Mini (−5.8, [−15.4, +2.5]); the remaining memory systems are not consistently positive. Under this unified protocol, query-based retrieval alone is therefore insufficient for aggregating behavioral evidence across unrelated scenarios.

Counterfactual Persona Specificity. Average alignment does not guarantee persona-discriminative responses. As shown in Table 3(b), No History is near zero ($\Delta_{\text{persona}} = -0.001$), whereas Explicit Profile reaches +0.246 and Profile Induction reaches +0.301. The four memory systems are positive but weaker (+0.078 to +0.173); clustered intervals remain above zero. Full results are provided in Appendix A.

4.3 Controlled Analyses

History–Scenario Overlap Intervention. To test whether scenario isolation contributes to task difficulty, we keep the test request, target configuration, and scoring protocol fixed and replace only one medium-level evidence dialogue in the history. *inj* injects a dialogue that shares the same decision object as the test request and matches the target level, whereas *injother* injects a dialogue of the same length, format, and target level but from an unrelated scenario. Thus, $\Delta_{\text{ov}} = \text{inj} - \text{injother}$ measures the net effect of having the same decision object appear in the history.

Overlap Results. As shown in Table 3(a), Open Test Δ_{ov} is positive for all four models (7.5–15.0 points). Same-scenario evidence therefore provides a local matching shortcut; removing it makes the task more dependent on cross-scenario abstraction. Details are in Appendix A.

Limitations. Our evaluation covers four base models and four memory systems, with some configurations adapted for protocol uniformity. Human validation (n=51) provides preliminary evidence but remains limited in scale. The benchmark targets stable traits rather than temporal dynamics, and the counterfactual metric measures configuration specificity rather than full individual identification.

5 Conclusion

We presented KlaraBench, a benchmark for cross-scenario behavioral understanding in Agent Memory Systems. Across four base models, external memory systems do not consistently outperform complete-history input: MemoryBank is better on DeepSeek-V4-Flash (+19.2) and on GPT-5.6-Luna (+11.3), while its difference on Doubao-Seed-2.0-Mini (−5.8) is not significant. Induced profiles improve both MCQ and Open Test, whereas explicit personality labels mainly help MCQ. Together with the low Δ_{persona}, these results indicate that future memory systems should organize dispersed behavioral evidence into individual-level representations rather than only retrieve local historical matches.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets (historical dialogue evidence and evaluation questions), providing feedback on research methodology (quality auditing of generated content), and interpreting results (LLM-as-judge evaluation in Open Test). We have not used generative AI tools for developing theoretical models, formulating hypotheses, designing experiments, implementing core methods, or writing proofs, and these other required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing parts of the paper to improve readability and sourcing relevant literature. All AI-generated content was validated through independent auditing models and manual review. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

ByteDance Seed. Seed 2.0. Official model webpage, 2026. URL https://research.doubao.com/en/seed2. Accessed: 2026-09-15.

Yuxin Chen, Yi Zhang, Zhengzhou Cai, Yaorui Shi, Zhiyuan Yao, Chenhang Cui, Jingnan Zheng, Yaqi Huo, Xi Su, Qi Gu, et al. Vitabench 2.0: Evaluating personalized and proactive agents in long-term user interactions. *arXiv preprint arXiv:2605.27141*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

DeepSeek. DeepSeek API Documentation. Online documentation, 2026. URL https://api-docs.deepseek.com. Accessed: 2026-09-15.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 98706–98729, 2026.

William Fleeson and Patrick Gallagher. The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, 2009.

Jean-Yves Gerlitz and Jürgen Schupp. Zur erhebung der big-five-basierten persoenlichkeitsmerkmale im soep. *DIW Research Notes*, 4(10.1016):1–44, 2005.

Google. Gemini 3.5 flash-lite. Large language model, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash-lite. Accessed: 2026-09-21.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

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

Bowen Jiang, Yuan Yuan, Zhuoqun Hao, Yuchen Liu, Maohao Shen, Sihao Chen, Gregory Wornell, Chris Callison-Burch, Lyle Ungar, Dan Roth, Qi Guo, Xiangjun Fan, Camillo J. Taylor, and Hanchao Yu. Personamem-v3: Toward omni-platform personal intelligence for holistic user understanding, recommendation, and agentic tasks, 2026. URL https://arxiv.org/abs/2608.21381.

Oliver P John, Laura P Naumann, and Christopher J Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158, 2008.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024.

Frieder R Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G Wagner. Short assessment of the big five: Robust across survey methods except telephone interviewing. *Behavior research methods*, 43(2):548–567, 2011.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue. *NAACL*, 2025.

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

Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, 1992. doi: 10.1111/j.1467-6494.1992.tb00970.x.

MemPalace. Mempalace: Local-first ai memory. https://github.com/MemPalace/mempalace, 2026. GitHub repository.

OpenAI. GPT-5.6 Luna. OpenAI model documentation, 2026. URL https://deploymentsafety.openai.com/gpt-5-6. Accessed: 2026-09-17.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed: 2026-09-15.

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

Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117, 2017.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh RN, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, et al. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 878–893, 2025.

Ao Tian, Yunfeng Lu, Xinxin Fan, Changhao Wang, Lanzhi Zhou, Yeyao Zhang, and Yanfang Liu. Rgmem: Renormalization group-based memory evolution for language agent user profile. *arXiv preprint arXiv:2510.16392*, 2025.

Simine Vazire. Who knows what about a person? the self–other knowledge asymmetry (soka) model. *Journal of personality and social psychology*, 98(2):281, 2010.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. Longmemeval-v2: Evaluating long-term agent memory toward experienced colleagues. *arXiv preprint arXiv:2605.12493*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Lingyang Zeng, Guangze Chen, Kaichen Yu, Zhicheng Pan, Siyang Weng, Zirui Hu, Xiangyun Du, Hailin He, Rong Zhang, Chengcheng Yang, Kai Huang, and Xuan Zhou. Setoka: A benchmark for hierarchical user understanding in personalized agents over heterogeneous data, 2026. URL https://arxiv.org/abs/2607.27056.

Zeyu Zhang, Quanyu Dai, Luyu Chen, Zeren Jiang, Rui Li, Jieming Zhu, Xu Chen, Yi Xie, Zhenhua Dong, and Ji-Rong Wen. Memsim: A bayesian simulator for evaluating memory of llm-based personal assistants. *Advances in Neural Information Processing Systems*, 38:90475–90511, 2026.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. In *International Conference on Learning Representations*, volume 2025, pp. 15888–15931, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

A Data Generation and Validation Details

Division of labor between generation and auditing models. All 720 MCQ items were generated by `doubao-seed-2-0-mini-260428`. Each item was cross-audited by two independent models (`deepseek-flash` and `doubao-seed-2-0-mini-260428`); `deepseek-flash` supplied the primary judgment, while the other model was used only as a cross-check. Thus, the main quality-control signal came from a non-generation model, although the audit was not completely independent of the generator.

MCQ gates and pass rates. All 720 items passed the automated gates. Candidate-response level annotations were fully confirmed for 707 items (470 two- and three-dimensional items with 4/4 candidates confirmed, and 237 one-dimensional items with 3/3 confirmed). The remaining 13 items contained at least one unconfirmed candidate (7 two-dimensional, 3 three-dimensional, and 3 one-dimensional items). These items passed all other gates, including length balancing, surface-cue, and lexicon checks, and were retained for evaluation; their identifiers are listed in Table 4 for manual inspection.

Surface cues and balancing. The median and maximum ratios of the longest to shortest option length were 1.22 and 1.50, respectively. The mean exact rate of candidate-level consistency judgments was 0.992 (minimum 0.50), and the self-contradiction count was zero for all 720 items. Correct-answer positions were balanced by item type, and the question wording explicitly required reasoning from historical behavioral tendencies.

Historical evidence pool. The fixed pool contained 626 dialogues (498 behavioral-evidence and 128 neutral-role dialogues). All of them were assigned to 3,760 persona-history slots (80 personas $\times$ 47 slots); each dialogue was reused 6.0 times on average (maximum 24). Any two personas shared 4.6 history dialogues on average (median 5, maximum 16), and 24.1% of persona pairs shared no evidence.

Table 4: The 13 items for which not all candidate level annotations were confirmed. They were retained in the evaluation set for manual inspection.

Persona	Type	Item	Confirmed	Correct option
v54_CAN_HLH	1dim	qb_05	2/3	C
v54_OAN_HLH	1dim	qe_09	2/3	A
v54_OCN_HHH	1dim	qb_12	2/3	B
v54_CEA_HLH	2dim	qc_06	2/4	C
v54_OCA_HLL	2dim	qb_03	2/4	D
v54_OCA_LHH	2dim	qb_02	2/4	C
v54_OCA_LHL	2dim	qb_01	2/4	D
v54_OCA_LLL	2dim	qa_12	2/4	A
v54_OCE_LHH	2dim	qa_07	2/4	A
v54_OEA_HLH	2dim	qe_08	3/4	D
v54_EAN_HLH	3dim	qd_07	3/4	A
v54_OEA_HLL	3dim	qa_09	2/4	B
v54_OEN_HLH	3dim	qe_06	2/4	B

Judge reliability. At the evidence level, the dimension–level annotation of every dialogue was cross-audited by two different models, and unconfirmed evidence was excluded from history construction. At the response level, **all** Open Test conditions in the main paper were scored by a single independent judge, `gemini-3.5-flash-lite` (Google), using rubric `adapt_v2` (Appendix D.1). This judge belongs to a different family from all four evaluated base models and is not a component of any evaluated memory system. The models used for data generation and item auditing serve a different construction-time role and do not contribute to the reported scores.

Absolute alignment levels are judge-dependent: rescoring the same responses with another judge can shift the level substantially. We therefore interpret paired differences within a judge, rather than comparing absolute scores across sources, and draw inferential conclusions from differences clustered by test opening or persona.

Response-orientation distribution. The current rubric assigns a score of one only when a response clearly satisfies the adaptation criterion for a dimension. Thus, alignment equal to zero includes both clearly oppositional responses and neutral or unexpressed responses. Table 5 reports the proportion of cells that align on all evaluated dimensions, are explicitly oppositional ("avoidance" in the judge output), or are partially aligned or unexpressed.

Table 5: Response orientation in Open Test (four-model pooled, 320 cells).

Condition	Alignment	All dimensions aligned	Explicitly oppositional	Partial or unexpressed
No History	52.3	25.3	7.5	67.2
Direct Prompting	66.7	36.6	2.8	60.6
Profile Induction	83.5	61.6	0.3	38.1
Explicit Profile	71.6	42.2	2.2	55.6
A-Mem	64.4	35.9	2.2	61.9
MemoryBank	74.2	49.1	2.5	48.4
MemPalace	64.5	34.7	3.8	61.6
LightMem	56.6	28.4	13.1	58.4

Explicitly oppositional responses are uncommon overall (0.6%–13.1%; the highest rates occur for No History and conditions with unfiltered irrelevant history), whereas partially aligned or unexpressed responses remain common (37%–67%). Alignment is therefore best interpreted as whether a response expresses a clear behavioral orientation consistent with the user, rather than as a complete measure of whether the model understands the user. Treating it as the sole measure of understanding would systematically undercount concise, natural, but neutral responses.

B Memory-System Configurations and Deviations

We integrated all four memory systems using their open-source implementations, while deviating from some upstream settings to match the benchmark protocol. Table 6 lists the upstream defaults, our settings, and the resulting deviations. For every system, the memory-extraction LLM was set to the evaluated base model itself (the model used for a given evaluation column also extracted that column's memories), avoiding a mismatch between extraction and response-generation models. The same rule applies to the Profile Induction condition: each evaluation column's free-form user profile was induced from that column's complete history by the evaluated base model itself, so profile induction and response generation never cross models. The response budget was standardized to a single letter for MCQ and 200–300 tokens for Open Test (using each system's default upper limit), and retrieval depth was set to five for all four systems; the official MemoryBank default is two.

Table 6: Upstream memory-system configurations and settings used in this paper.

System	Upstream default (embedding / internal LLM / k)	Setting in this paper	Deviation
A-Mem	all-MiniLM-L6-v2 / gpt-4o-mini / $k = 5$ (retrieval)	Embedding: bge-small-zh-v1.5; internal LLM: evaluated base model; $k = 5$	Unchanged.
MemoryBank	text2vec-large-chinese / gpt-3.5-turbo / $k = 2$ (`memory_search_top_k`)	Embedding: bge-small-zh-v1.5; internal LLM: evaluated base model; $k = 5$	Hierarchical summaries and the global profile are recomputed with the official Chinese prompts and parameters (temperature 0.7, max_tokens 400). We replace the `llama_index` retrieval index with local cosine retrieval (one document per day containing the daily summary) and change k from 2 to 5.
LightMem	all-MiniLM-L6-v2 (official example) / caller-specified / $k = 10$	Embedding: bge-small-zh-v1.5; internal LLM: evaluated base model; $k = 5$	Force extraction for every segment (the effective `extract_threshold` is not applied).
MemPalace (our system)	embeddinggemma / none (no LLM call) / $k = 5$	Keep the embedding; no internal LLM; $k = 5$	None.

MemoryBank requires a specific qualification. Its hierarchical summaries (daily event summaries and daily personality/response strategies), global event summary, and global user profile were recomputed with the official Chinese prompts and parameters. The granularity of one day per

two sessions (four question–answer pairs) matches the official evaluation data, which uses four to six pairs per day. Our deviation is concentrated on retrieval: the official implementation uses `llama_index`'s `GPTSimpleVectorIndex` with OpenAI embeddings and $k = 2$, whereas we use local `bge-small-zh-v1.5` embeddings and cosine retrieval with one document per day containing the raw dialogue and daily summary, and set $k = 5$ to match the other systems. The forgetting curve remains disabled as in the official default. MemoryBank results should therefore be read as performance under the official memory-construction process plus our unified retrieval protocol.

C Scenario-Overlap Audit and Results after Exclusion

Audit procedure. We encoded the 85 test scenarios (65 MCQ scenarios and 20 Open Test scenarios) and the evidence dialogues of the pool with `bge-small-zh-v1.5`. We formed bidirectional top-k pairs ($k = 3$ on the question side and $k = 2$ on the evidence side), following `audit/SCENARIO_DISJOINT_v55.md`. An independent judge, the same `gemini-3.5-flash-lite` judge used for the Open Test metrics, then determined whether each pair concerned the same decision object.

Results. The judge identified pairs sharing a decision object in 18 MCQ scenarios (`qa_01`, `qa_07`, `qe_13`, `qe_06`, `qc_07`, `qe_12`, `qa_09`, `qd_05`, `qe_01`, `qd_07`, `qe_04`, `qa_02`, `qb_04`, `qe_10`, `qb_12`, `qd_08`, `qa_12`, and `qb_13`) and four Open Test scenarios (`intention54_08_community_volunteering`, `intention54_11_improv_workshop`, `intention54_15_craft_hobby_group`, and `intention54_16_language_exchange`), which we treated as the overlapping set. We removed the 198 questions belonging to the 18 MCQ scenarios and the persona cells of the four Open Test scenarios, and recomputed all conditions on the remaining items. The MCQ results are shown in Table 7.

Table 7: MCQ results after excluding the test scenarios judged to share a decision object with the history, reported as "full → excluded" (the 18 overlapping scenarios contain 198 of the 720 items). The overlap set was identified in a pairwise audit performed by the same `gemini-3.5-flash-lite` judge used for the Open Test metrics; MCQ accuracy itself involves no judge.

Condition	Qwen	DS-Flash	GPT	Doubao	Pooled
No History	33.9→34.1	40.8→41.8	28.9→28.7	32.9→34.5	34.1→34.8
Direct Prompting	70.1→70.3	64.3→66.1	75.8→76.6	66.9→66.9	69.3→70.0
Profile Induction	75.1→77.8	72.6→74.9	69.9→69.2	69.6→71.8	71.8→73.4
Oracle	76.7→76.2	70.7→70.7	89.6→89.6	84.2→83.1	80.3→79.9
A-Mem	57.5→57.7	52.6→53.3	58.5→57.5	55.1→55.0	55.9→55.8
MemoryBank	66.0→67.6	59.6→61.5	57.1→55.9	58.9→60.0	60.4→61.3
MemPalace	55.7→54.4	50.6→50.2	60.7→61.1	58.3→59.0	56.3→56.2
LightMem	38.8→38.1	42.6→43.1	39.9→39.1	44.3→45.0	41.4→41.3

After excluding overlapping scenarios, the pooled MCQ metric changed by at most 0.9 percentage points (the largest change was MemoryBank, 60.4→61.3). On Open Test, the pooled alignment changed by at most 1.4 points (the largest change was Direct Prompting, 67.8→66.4). The ordering of conditions remained unchanged: Profile Induction stayed above the other memory conditions and Oracle remained the upper reference on both tasks, and no external memory system exceeded Direct Prompting on the pooled MCQ metric. The main conclusions therefore do not depend on the small set of scenarios with local overlap.

D Statistical Protocol and Supplementary Checks

Why clustering is needed. The 720 MCQ items were generated by 80 personas, so items from the same persona share history and configuration. The 80 Open Test cells are grouped into 20 openings, with four personas sharing each opening. Treating individual items or cells as independent is therefore optimistic. In addition to paired tests, we report bootstrap and permutation analyses clustered by persona for MCQ and by opening for Open Test.

Differences among the ablation conditions. The MCQ column uses item-level paired tests. The Open Test column uses paired McNemar tests on rubric cells scored by the same `gemini-3.5-flash-lite` judge. Table 8 reports adjacent comparisons.

Table 8: Adjacent ablation comparisons: paired McNemar tests and 95% confidence intervals.

Panel	Comparison	Δ	95% CI	p
MCQ	No History → Full benchmark	$+35.2$	$[+33.0, +37.3]$	9.2e-164
MCQ	Full benchmark → Induced profile	$+2.5$	$[+0.8, +4.2]$	0.0056
MCQ	Induced profile → Explicit profile	$+8.5$	$[+6.6, +10.3]$	1.8e-18
Open Test	No History → Full benchmark	$+14.4$	$[+11.0, +17.8]$	2.6e-15
Open Test	Full benchmark → Induced profile	$+16.9$	$[+13.9, +19.8]$	1.1e-25
Open Test	Induced profile → Explicit profile	-12.0	$[-14.8, -9.1]$	4.2e-15
Open Test	Full benchmark → Explicit profile	$+4.9$	$[+1.8, +8.0]$	0.0025

Uncertainty in persona specificity. Each (model, condition) combination in the counterfactual rescoring contains 320 pairs, but the four personas sharing an opening are not independent. We therefore bootstrap by opening and perform a sign test at the opening level; results are in Table 9.

Table 9: Opening-clustered Δ_{persona} (four models pooled; judge: `gemini-3.5-flash-lite`).

Condition	Δ_{persona}	Clustered 95% CI	Positive openings
pe_long	$+0.301$	$[+0.256, +0.343]$	20/20
oracle	$+0.246$	$[+0.203, +0.287]$	20/20
clean_long	$+0.140$	$[+0.108, +0.170]$	19/20
amem_long	$+0.085$	$[+0.056, +0.113]$	18/20
memory_long	$+0.081$	$[+0.048, +0.114]$	16/20
lm_long	$+0.078$	$[+0.044, +0.115]$	15/20
mb_long	$+0.173$	$[+0.133, +0.216]$	20/20
direct	-0.001	$[-0.023, +0.020]$	8/20

Clustered robustness of the main comparisons. Table 10 reports confidence intervals for the main comparisons after clustering by persona or opening.

Table 10: Clustered bootstrap for key comparisons.

Panel	Comparison	Δ	IID 95% CI	Clustered 95% CI
Open	No History → Full benchmark	$+14.4$	$[+9.8, +19.0]$	$[+10.2, +18.9]$
Open	Full benchmark → Induced profile	$+16.9$	$[+13.3, +20.5]$	$[+13.2, +20.3]$
Open	Induced profile → Explicit profile	-12.0	$[-15.4, -8.7]$	$[-16.4, -7.9]$
Open	Direct Prompting → MemPalace	-2.2	$[-5.9, +1.7]$	$[-7.4, +2.6]$

D.1 Judge Robustness

Open Test was scored by the independent judge `gemini-3.5-flash-lite` (Google), which differs in family from all four evaluated base models and shares no source with any evaluated memory system. Table 11 reports the condition-level means and the two controlled analyses.

Absolute alignment levels under this rubric are judge-dependent, so we compare conditions only within this judge and report paired differences clustered by test opening. On Open Test, MemoryBank is significantly above Direct Prompting for DeepSeek-V4-Flash ($+19.2$, clustered 95% CI $[+11.3, +27.1]$) and for GPT-5.6-Luna ($+11.3$, $[+5.0, +17.5]$); the remaining memory-system comparisons either cross zero or are negative. MemoryBank's point estimate is no lower than Direct Prompting for three of four models, and is lower only for Doubao-Seed-2.0-Mini (-5.8), where the

Table 11: Open Test alignment under the rubric judge (four-model pooled, %); the lower panel reports the controlled-analysis quantities.

Quantity	Alignment
No History	52.3
Clean evidence only	70.8
Direct Prompting (full history)	66.7
Profile Induction	83.5
Explicit Profile	71.6
A-Mem	64.4
MemoryBank	74.2
MemPalace	64.5
LightMem	56.6
Δ_{persona}: Profile Induction	$+0.301$
Δ_{persona}: Explicit Profile	$+0.246$
Δ_{persona}: MemoryBank	$+0.173$
Δ_{persona}: MemPalace	$+0.081$
Δ_{persona}: No History (placebo)	-0.001
Δ_{ov} (Qwen)	$+7.5$
Δ_{ov} (DS-Flash)	$+7.5$
Δ_{ov} (GPT)	$+12.5$
Δ_{ov} (Doubao)	$+15.0$

Table 12: 95% confidence intervals for the key quantities (opening-clustered bootstrap, 2,000 resamples; judge `gemini-3.5-flash-lite`). Δ is in percentage points.

Quantity	Estimate	Clustered 95% CI
MemoryBank $-$ Direct Prompting (Qwen)	$+5.4$	$[-1.7, +12.9]$
MemoryBank $-$ Direct Prompting (DS-Flash)	$+19.2$	$[+11.3, +27.1]$
MemoryBank $-$ Direct Prompting (GPT)	$+11.3$	$[+5.0, +17.5]$
MemoryBank $-$ Direct Prompting (Doubao)	-5.8	$[-15.4, +2.5]$
Δ_{ov} (Qwen)	$+7.5$	$[+2.5, +12.1]$
Δ_{ov} (DS-Flash)	$+7.5$	$[-0.8, +16.3]$
Δ_{ov} (GPT)	$+12.5$	$[+5.0, +20.4]$
Δ_{ov} (Doubao)	$+15.0$	$[+9.2, +21.3]$

confidence interval crosses zero. We consequently draw conclusions from paired differences and clustered intervals rather than from absolute levels.

The values discussed in the main paper, MemoryBank $-$ Direct Prompting of -5.8 for Doubao-Seed-2.0-Mini (CI crossing zero) and $+19.2$ for DeepSeek-V4-Flash, can be verified in this table.

Generation script: `scripts/v5_6_gemini_appendix.py`; scoring files: `judgematrix/gemini_*`.

E Counterfactual History Comparison for Open Test

E.1 Motivation

The main paper already provides three Open Test controls—No History, Explicit Profile, and irrelevant-history augmentation—which respectively test how much can be achieved from the scenario alone, whether the information carrier matters, and whether history composition affects the output. These controls vary information quantity and representation, but do not answer a sharper question: is a first-turn response *bound to the identity* whose evidence conditioned it, or would the same response fit any person's version of the scenario? This appendix performs counterfactual rescoring without regenerating responses. Each opening is shared by four personas; assigning persona i's response to persona j's target levels produces a $k \times k$ alignment matrix. Its diagonal is the reported condition, while its off-diagonal entries are the "history-swapped" counterfactual. Because the judge sees only the opening, candidate response, and target levels, changing the target levels operationally corresponds to changing the history.

E.2 Setup and scale

- Scale: $k = 4$ personas per opening and 20 openings, giving $4 \times 4 \times 20 = 320$ scored pairs per (model, condition) combination.

Table 13: Counterfactual history comparison for Open Test. Each cell is Δ = (diagonal − off-diagonal), scored by `gemini-3.5-flash-lite`. **No History** is the placebo and should be close to zero.

Condition	Qwen	DS-Flash	GPT	Doubao
No History	$+0.033$	-0.028	$+0.006$	-0.017
Direct Prompting (full history)	$+0.221$	$+0.054$	$+0.103$	$+0.183$
Profile Induction	$+0.376$	$+0.325$	$+0.192$	$+0.360$
Explicit Profile	$+0.367$	$+0.244$	$+0.096$	$+0.278$
A-Mem	$+0.157$	$+0.071$	$+0.032$	$+0.082$
MemoryBank	$+0.208$	$+0.204$	$+0.110$	$+0.169$
MemPalace	$+0.142$	$+0.093$	$+0.078$	$+0.013$
LightMem	$+0.174$	$+0.047$	$+0.038$	$+0.054$

- Judge: `gemini-3.5-flash-lite` (Google), from a **different family** than all four evaluated base models, with rubric `adapt_v2`. Every Open Test condition uses this same judge; no scoring sources are mixed.
- Metric: $\Delta = \text{alignment}(\text{diagonal}) - \text{alignment}(\text{off-diagonal})$. Larger values indicate stronger identity specificity.

E.3 Two controls validate the metric

Placebo (No History). This condition provides no history, so the diagonal and off-diagonal should not differ systematically by construction. The observed four-model mean is $\Delta = -0.001$ (per-model range -0.028 to $+0.033$), showing no diagonal preference and making a positive difference in other conditions unlikely to be judge noise alone.

Carrier control (Explicit Profile). Replacing history with a profile card containing the user's true levels yields a four-model mean of $\Delta = +0.246$, among the largest values. This indicates that the metric measures identity correspondence rather than merely the presence of context.

E.4 Complete results

Across conditions, the four-model means are $+0.246$ for Explicit Profile, $+0.301$ for Profile Induction, and $+0.140$ for Direct Prompting. The four memory systems obtain $+0.085$ (A-Mem), $+0.173$ (MemoryBank), $+0.081$ (MemPalace), and $+0.078$ (LightMem). Retrieval-based memory therefore transmits usable information (all memory conditions exceed the placebo floor), but the identity specificity of their first-turn responses is weaker than under explicit or induced profiles.

E.5 Independent evidence: counterfactual flips on MCQ

The question bank contains 152 pairs with the same scenario and item type but different levels on one dimension; 20 are one-dimensional pairs, where the entire item isolates that dimension. The correct answers are opposite on that dimension, so getting both questions correct requires the model to follow the history when the level flips. No judge is required: the answer label and option code are fixed by the question bank. In the one-dimensional pairs, the probability of getting both correct is only 0.15–0.25 without history, 0.15–0.65 under the memory conditions and 0.35–0.65 under full-history prompting, and 0.65–0.85 under Explicit Profile, matching the direction of the Open Test counterfactual result.

E.6 Interpretation and limitations

- The four personas sharing an opening usually differ on two or three dimensions, so this is a multi-dimensional swap rather than a single-dimension dose–response test.
- Off-diagonal scores are not a No History floor: generic responses may receive partial credit on some dimensions. The No History placebo should therefore be used for calibration.

Table 14: Composition of the three human-validation instruments. Instrument 1 pairs a self-report scale with scenario items; Instruments 2 and 3 ask the same two questions about a single assistant response or a full dialogue, respectively. The random baseline is listed for the dimension question, the pole question, and both questions together.

Instrument	Stimulus	Items	Dimensions	Poles per dimension	Questions	Rando
1	Self-report item + scenario with 3 responses	15	5	3 / 3	–	1/5 /
2	Single assistant response	20	5	4	2	1/5 /
3	Four-turn evidence dialogue	10	5	2	2	1/5 /

Table 15: Participants and response duration by instrument. No respondent was excluded: every submitted response was retained, and every respondent answered every item. Duration is reported for context, not as an exclusion criterion.

Instrument	Collected	Excluded	Analysed	Median duration (IQR)	Judgments
1	18	0	18	394.5 s (201–502)	540
2	17	0	17	285 s (195–384)	340
3	16	0	16	276.5 s (167–434)	160

- The appendix and main paper use the same judge, so the values are directly comparable; absolute alignment remains judge-dependent and should not be compared across papers with different judges.

F Human Validation Instruments and Item-Level Results

This appendix documents the three instruments that supply the human evidence summarised in Section 3.2. All three use the same five dimensions and the same two-pole (low/high) split as the benchmark, and all three were frozen before analysis: the questionnaire contents, the option order, and the analysis rules were fixed and exported in a single batch, and no item was added, removed, or relabelled afterwards. For every table we state the random baseline in the caption, report observed proportions with bootstrap confidence intervals resampled over respondents, and treat the instruments as preliminary evidence rather than as population-level estimates.

F.1 Instrument Construction

Instrument 1 pairs 15 self-report items (three per dimension, BFI-S wording, 1–7, reverse-scored items reflected before averaging) with 15 single-choice scenario items (three per dimension), and is the only instrument that shows no assistant turn at all. Each scenario item offers one response matching the high pole of its dimension, one matching the low pole, and one that is neutral on the dimension; this is the instrument that speaks most directly to the behavioural operationalisation used by the benchmark. Instruments 2 and 3 share an identical response format—two questions per stimulus, namely which dimension the stimulus expresses and whether that dimension is high or low—and differ only in the stimulus: a single assistant response versus a full four-turn dialogue. Table 14 gives the composition of each instrument.

F.2 Participants and Exclusions

No respondent was excluded from any instrument. The three exports were frozen after data collection closed, and every submitted response was retained, so the analysed n equals the number of respondents in each export (Table 15). Every respondent answered every item of every instrument, so no item-level imputation or listwise deletion was required; this includes the fastest completion in the study, which is reported below and which we retained deliberately rather than removing on a duration threshold. Response duration is reported as a median with interquartile range.

Because a zero-exclusion analysis rests on the judgement that no response was careless, we checked the two signals that would most plausibly indicate otherwise. First, the full distribution of response

Table 16: Instrument 1, per-dimension correlation between self-report and scenario choice, with bootstrap 95% confidence intervals resampled over participants. Random baseline: $r = 0$. Bonferroni threshold for five tests: $\alpha = 0.01$.

Dimension	n	Self-report M (SD)	Scenario M (SD)	r	r 95% CI	p	ρ	τ_b	α
Openness	18	4.70 (1.21)	0.51 (0.30)	0.394	-0.142–0.833	0.086	0.498	0.426	0.771
Conscientiousness	18	4.65 (1.13)	0.61 (0.26)	-0.117	-0.582–0.401	0.637	-0.168	-0.139	0.434
Extraversion	18	3.93 (1.72)	0.59 (0.27)	0.503	0.143–0.755	0.020	0.493	0.353	0.955
Agreeableness	18	5.31 (0.98)	0.52 (0.29)	0.399	-0.130–0.763	0.082	0.413	0.332	0.388
Neuroticism	18	4.57 (1.43)	0.58 (0.28)	-0.012	-0.479–0.414	0.961	-0.039	-0.022	0.789

Table 17: Instrument 1, pooled correlation across the five dimensions after within-dimension standardisation. The resampling unit remains the participant, so the interval reflects participant-level sampling variability rather than the nominal number of stacked observations.

Quantity	Value	95% CI	p
Pooled r (within-dimension z-scores)	0.233	-0.061–0.352	0.3371
Mean per-dimension r	0.233	–	–
Dimensions with positive r	3/5	–	–
Observations (participants $\times$ dimensions)	90	–	–

times for Instrument 1 spans 89 s to 1984 s with a median of 394.5 s; the fastest respondent is an outlier on duration but not on content, producing a self-report profile and scenario choices that are internally coherent and within the range observed for the rest of the sample. Second, we recomputed every Instrument 1 statistic with that respondent removed: the pooled correlation moves from 0.233 to 0.228, the categorical agreement rate is unchanged at 63.0%, and the pattern of per-dimension correlations is unchanged. The results below are therefore not sensitive to this decision.

F.3 Instrument 1: Self-Report versus Scenario Choice

For each dimension we correlate the self-report score with the scenario-choice score (the neutral option scores 0.5, so the scenario score lies in $[0, 1]$). Table 16 reports Pearson r with a bootstrap confidence interval resampled over respondents, the corresponding p-value, Spearman ρ and Kendall τ_b as rank-based checks, and per-dimension internal consistency. Because the five dimensions are tested separately, the Bonferroni threshold for five tests is $\alpha = 0.01$. Table 17 pools the dimensions after within-dimension standardisation, which raises the number of observations per respondent while keeping the resampling unit at the respondent.

Table 18 cross-tabulates the two categorisations per respondent and dimension. A cell is counted only when both sides are unambiguous, that is, the self-report mean is at or above 4.5 (high) or at or below 3.5 (low), and the scenario score is at or above 0.67 (high) or at or below 0.33 (low); chance agreement is therefore $1/2$.

Three of the five per-dimension correlations are in the expected direction (Openness, Extraversion, Agreeableness) and the pooled correlation is positive, but the association is weak and not statistically robust: the pooled correlation is small and its interval includes zero, only Extraversion reaches conventional significance before correction, and no dimension survives the Bonferroni threshold. In the categorical view the two categorisations agree in 63.0% of the 54 unambiguous respondent–dimension cells against a 50% chance rate, which is a clearer signal than the correlation but rests on the same underlying scores. We therefore report Instrument 1 as directional evidence about the behavioural operationalisation and *not* as a demonstration of convergent validity; the per-dimension spread (from a slight negative value for Conscientiousness to a moderate positive value for Openness) is itself the relevant observation, and it indicates that the scenario items do not all tap their intended dimension to the same degree. This reading is consistent with the per-dimension findings from Instruments 2 and 3 below.

Table 18: Instrument 1, agreement between the self-report category and the scenario-choice category. Only cells that are unambiguous on both sides are counted, so chance agreement is $1/2$.

Self-report \ Scenario	High	Low	Total	Agreement	Chance	κ
High	27	16	43	–	–	–
Low	4	7	11	–	–	–
Total	31	23	54	63.0%	50%	–

Table 19: Instrument 2, overall accuracy for judging a single assistant response. Intervals are bootstrap intervals resampled over respondents; the test against the random baseline is participant-clustered.

Quantity	Observed	95% CI	Random baseline
Dimension identification	61.2%	53.2%–69.7%	20%
High/low identification	79.7%	70.9%–87.6%	50%
Both labels correct	53.8%	43.2%–64.4%	10%

F.4 Instrument 2: Single-Response Judgment

Respondents saw one assistant response at a time and answered two questions per response: which dimension the response expressed, and whether that dimension was relatively high or low. Table 19 reports the three accuracy rates against their random baselines. Table 20 breaks the same rates down by the dimension that the stimulus actually expresses, and Table 21 gives the full confusion matrix, so that systematic over-attribution of one dimension is visible rather than hidden inside an aggregate rate.

Table 22 lists every stimulus with its accuracy, which is the level at which one can check whether an individual response is mis-keyed or ambiguous. No stimulus has a both-correct rate of zero, so no stimulus is uninterpretable by the whole sample; equally, no stimulus is answered correctly by every respondent, so the aggregate rates are not carried by a small subset of easy items.

F.5 Instrument 3: Dialogue Attribution

Instrument 3 repeats the response format of Instrument 2 with a full four-turn dialogue as the stimulus, which is the task reported in Section 3.2. Table 23 reports the overall rates, Table 24 the per-dimension rates, and Table 25 the confusion matrix.

Table 26 lists the dialogues individually. The rows follow the order in which the dialogues were deployed, so that the order used in the analysis can be checked against the order shown to respondents.

F.6 Inter-Respondent Agreement

Table 27 reports Fleiss' κ, Krippendorff's α (nominal), and the observed agreement rate for both judgment questions, treating stimuli as the units of analysis and respondents as raters. Chance agreement is $1/5$ for the dimension question and $1/2$ for the pole question.

Agreement is weak on both questions. The maximum attainable agreement rate is bounded by accuracy itself—if a stimulus is correctly identified by only half the respondents, at most half the respondents can agree on it—so these coefficients are low partly because the underlying task is hard rather than because respondents answered carelessly. They should be read alongside the accuracy tables above, not as an independent quality signal.

F.7 Pole Identification Is Reliably Recovered; Dimension Identity Is Not

The most consistent pattern across all three instruments is a dissociation between the two judgment questions. Respondents recover the *pole* of the expressed dimension far more reliably than they recover the *dimension* itself:

Table 20: Instrument 2, accuracy by the dimension expressed by the stimulus.

True dimension	Items	Dimension correct	High/low correct	Both correct
Openness	4	76.5%	82.4%	70.6%
Conscientiousness	4	69.1%	82.4%	60.3%
Extraversion	4	42.6%	77.9%	39.7%
Agreeableness	4	75.0%	80.9%	69.1%
Neuroticism	4	42.6%	75.0%	29.4%

Table 21: Instrument 2, dimension confusion matrix in counts of judgments. The diagonal is correct; row totals equal the number of judgments for each true dimension.

Truth \ Response	O	C	E	A	N	Total
Openness	52	6	6	2	2	68
Conscientiousness	3	47	0	3	15	68
Extraversion	6	20	29	13	0	68
Agreeableness	2	4	7	51	4	68
Neuroticism	3	19	2	15	29	68
Total	66	96	44	84	50	340

- Instrument 2: pole identification exceeds chance by $1.6\times$, dimension identification by $3.1\times$; the gap is largest for Extraversion (77.9% vs. 42.6%) and Neuroticism (75.0% vs. 42.6%).
- Instrument 3: the same ordering holds, and more sharply—pole identification exceeds chance by $1.6\times$ while dimension identification is only $2.5\times$ chance, with Extraversion showing 90.6% pole accuracy against 53.1% dimension accuracy and Neuroticism 78.1% against 25.0%.
- Instrument 1: respondents are more consistent with their own self-report on the pole than on the dimension, which is what the weak but uniformly positive association in Table 16 reflects.

Reading the confusion matrices explains why. Misattributions are concentrated in a single pair of dimensions, and in the same direction in *both* judgment instruments despite their different stimuli: in Instrument 3, 20 of the 32 Neuroticism judgments are attributed to Conscientiousness, exceeding the 8 correct attributions; in Instrument 2 the same cell takes 19 judgments against 15 for the reverse direction. No other off-diagonal pair approaches this rate—the next-largest confusion involves Extraversion and is less than half as frequent. The direction of the misattribution is informative: respondents who see a response that defers a decision, asks to gather more detail, or suggests slowing down classify it as planful and orderly more often than as emotionally reactive.

This is a property of how the benchmark operationalises Neuroticism rather than a defect of the respondents. In the bank, the high pole of Neuroticism is expressed through deliberation and emotional-uncertainty wording—holding off, gathering information, settling one's feelings before acting—which is lexically close to the high pole of Conscientiousness, expressed through checking prerequisites, confirming details, and sequencing steps. Both poles therefore share the surface form "work out the details before acting", and the two dimensions become hard to separate on behavioural evidence alone. The same confusion appears in the benchmark's own fairness audit, where an independent judge reading Neuroticism-High responses as high Neuroticism succeeds well below the rate achieved on the other four dimensions.

We treat this as a finding about the construct rather than as an anomaly to be explained away. It supplies direct human evidence for the cross-dimension contamination criterion that the dataset's quality gates enforce (Section 3.2), and it identifies Neuroticism as the dimension where the operationalisation is least discriminable from a neighbouring trait. We therefore report per-dimension and per-item results throughout this appendix instead of a single pooled accuracy, since the pooled figure would average over a dimension that behaves differently from the rest.

F.8 Summary and Interpretation

The three instruments are not pooled and are not presented as independent replications. Instruments 2 and 3 use the same response format and the same five dimensions, and both reuse the benchmark's

Table 22: Instrument 2, per-stimulus accuracy. One row per stimulus; the stimulus column gives a short English label for the response shown to respondents.

No.	Dimension	Pole	Dimension correct	High/low correct	Both correct	Stimulus (response topic)
1	Extraversion	High	4/17	15/17	4/17	Shift-swap coordination
2	Agreeableness	Low	14/17	12/17	10/17	Agreeing to a meetup
3	Neuroticism	High	5/17	13/17	3/17	Unfamiliar-appointment decision
4	Conscientiousness	High	15/17	16/17	14/17	Booking accommodation
5	Openness	Low	12/17	10/17	10/17	Choosing a window design
6	Agreeableness	Low	13/17	15/17	13/17	Asserting one's share
7	Conscientiousness	Low	7/17	12/17	5/17	Choosing an office chair
8	Extraversion	Low	8/17	13/17	8/17	Splitting a bill with a roommate
9	Neuroticism	Low	6/17	14/17	6/17	Deciding on a repair option
10	Agreeableness	High	11/17	12/17	11/17	Accepting a meetup
11	Openness	High	14/17	17/17	14/17	Trying an unfamiliar cuisine
12	Extraversion	High	9/17	14/17	7/17	Coordinating a gift amount
13	Agreeableness	High	13/17	16/17	13/17	Hosting a hobby-group chat
14	Conscientiousness	Low	10/17	13/17	9/17	Learning a new tool
15	Openness	Low	10/17	13/17	9/17	Introducing oneself
16	Neuroticism	High	4/17	13/17	3/17	Furniture-rearrangement plan
17	Conscientiousness	High	15/17	15/17	13/17	Choosing an appliance model
18	Extraversion	Low	8/17	11/17	8/17	Following a naming checklist
19	Neuroticism	Low	14/17	11/17	8/17	Deciding on a model
20	Openness	High	16/17	16/17	15/17	Trying a new approach

Table 23: Instrument 3, overall attribution accuracy for a full four-turn dialogue. Conventions follow Table 19.

Quantity	Observed	95% CI	Random baseline
Dimension identification	50.6%	35.6%–65.0%	20%
High/low identification	79.4%	71.3%–86.9%	50%
Both labels correct	50.0%	35.6%–65.0%	10%

operationalisation of each dimension; their agreement rates therefore bound what a respondent can recover from that operationalisation itself, not the accuracy of the benchmark's labels against an external criterion. All three instruments exceed their chance baselines on every measure reported here, and Instrument 2 does so with the interval lower bound above the baseline on all three measures.

Three caveats bound what may be concluded. First, the samples are small (Table 15), so the confidence intervals are wide and the per-dimension estimates in particular are unstable. Second, respondents were recruited as a convenience sample and are not representative of any population; the instruments estimate whether the operationalisation is legible to readers of the same language and culture as the benchmark's scenarios, not how a general population would respond. Third, Instrument 1 does not support a claim of convergent validity, as set out above. Performance is not uniform across dimensions, and the per-dimension and per-item tables above are the appropriate level at which to inspect where a dimension is hard to identify.

Table 24: Instrument 3, attribution accuracy by the dimension expressed by the dialogue. Each dimension contributes one high and one low dialogue.

True dimension	Dialogues	Dimension correct	High/low correct	Both correct
Openness	2	56.3%	68.8%	56.3%
Conscientiousness	2	71.9%	81.3%	71.9%
Extraversion	2	53.1%	90.6%	53.1%
Agreeableness	2	46.9%	78.1%	43.8%
Neuroticism	2	25.0%	78.1%	25.0%

Table 25: Instrument 3, dimension confusion matrix in counts of judgments.

Truth \ Response	O	C	E	A	N	Total
Openness	18	7	3	4	0	32
Conscientiousness	3	23	1	3	2	32
Extraversion	3	8	17	3	1	32
Agreeableness	1	5	6	15	5	32
Neuroticism	1	20	1	2	8	32
Total	26	63	28	27	16	160

Table 26: Instrument 3, per-dialogue attribution accuracy, listed in the order in which the dialogues were deployed. The last column gives a short English label for the dialogue.

No.	Dimension	Pole	Dimension correct	High/low correct	Both correct	Dialogue (topic)
1	Openness	High	14/16	16/16	14/16	Joining a new interest group
2	Neuroticism	Low	4/16	9/16	4/16	HPV follow-up appointment
3	Extraversion	High	7/16	16/16	7/16	Preparing a family birthday party
4	Openness	Low	4/16	6/16	4/16	Elder-care arrangement
5	Agreeableness	High	8/16	15/16	8/16	Caring for a visiting cousin
6	Extraversion	Low	10/16	13/16	10/16	Companion for a medical check-up
7	Conscientiousness	Low	12/16	12/16	12/16	Department reimbursement process
8	Neuroticism	High	4/16	16/16	4/16	Community pet-area vote
9	Conscientiousness	High	11/16	14/16	11/16	Community book-corner rota
10	Agreeableness	Low	7/16	10/16	6/16	Crowdfunding for an acquaintance

Table 27: Inter-respondent agreement by instrument and question, treating stimuli as units and respondents as raters. Chance agreement is $1/5$ for the dimension question and $1/2$ for the pole question.

Instrument	Question	Units	Raters	Agreement rate	Chance	Fleiss' κ	Krippendorff's α
2	Dimension	20	17	0.0%	$1/5$	0.362	0.364
2	High/low	20	17	5.0%	$1/2$	0.353	0.355
3	Dimension	10	16	0.0%	$1/5$	0.218	0.223
3	High/low	10	16	30.0%	$1/2$	0.411	0.415

Table 28: Headline quantities of the three human-validation instruments. The instruments are not pooled, because they differ in stimulus, in the number of judgments per respondent, and in how closely their items follow the benchmark's operationalisation.

Instrument	n	Judgments	Dimension correct	High/low correct	Both correct	Self-report $\leftrightarrow$ scenario r
1	18	540	–	–	–	0.233
2	17	340	61.2%	79.7%	53.8%	–
3	16	160	50.6%	79.4%	50.0%	–